Outsourcing of Secure k-Nearest Neighbours Interpolation Method

Muhammad Rifthy Kalideen
Department of Islamic Studies
South Eastern University of Sri Lanka
Oluvil, Sri Lanka, 32360

Bulent Tugrul
Department of Computer Engineering
Ankara University
Ankara, Turkey, 06100

***Abstract*—Cloud computing becomes essential in these days for the enterprises. Most of the large companies are moving their services and data to the cloud servers which offer flexibility and efficiency. Data owner (DO) hires a cloud service provider (CSP) to store its data and carry out the related computation. The query owner (QO) sends a request which is crucial for its future plans to the CSP. The CSP computes all necessary calculations and returns the result back to the QO. Neither the data nor query owners want to reveal their private data to anyone. k-Nearest Neighbour (k-NN) interpolation is one of the essential algorithms to produce a prediction value for an unmeasured location. Simply, it finds k number of nearest neighbours around the query point to produce an output. Oblivious RAM (ORAM) has been used to protect the privacy in cloud computing. In our work, we will perform the k-NN method using the kd-tree and ORAM without revealing both the data-owner's and query owner's confidential data to each other or to third parties. The proposed solution will be analysed to ensure that it provides accurate and reliable predictions while preserving the privacy of all parties.**

Keywords*—Cloud computing; k-Nearest neighbour; spatial interpolation

I. Introduction

Governments, companies and institutions save data for a variety of reasons. Some examples of such data are medical, insurance, banking and geographic. Data should be analysed to obtain useful information. Data mining methods are conducted to extract useful information from the stored data. The biggest concern in data mining is privacy. Two major problems may arise during data mining applications. First one, the data required for mining may have been collected by two different institutions. When executing the data mining algorithm on the combined databases, each party does not want to share its private data or laws may prohibit data sharing. The second issue is that the calculated/resulting data must be transferred to others for further activities. Data transfer should ensure the privacy of the data owner and produce the correct output. The banking sector is one of the examples regarding the privacy of data while running data mining algorithm. Nowadays, multiple different banks need to merge their data to mine for further activities. Due to law and privacy policies, the data cannot be pooled in one place. In these cases, solutions that provide data privacy are preferred instead of traditional data mining algorithms [1], [2].

Information Technology moved to a new era in this decade known as cloud computing. Cloud computing provides access to resources such as servers, storage and applications. Cloud computing is divided into different models according to the services they offer [3].

a) Software as a Service (SaaS)
b) Platform as a Service (PaaS)
c) Infrastructure as a Service (IaaS)
d) Data as a Service (DaaS)
e) Database as a Service (DBaaS)

All models have the following characteristics: shared infrastructure, dynamic provisioning, network access and managed metering. Furthermore, deployment of the cloud computing is divided into four categories. They are private, community, public and hybrid cloud. Some of the challenges associated with cloud computing are privacy, lack of standards, constantly evolving, and compatibility concerns [3].

Spatial data should also be analysed to find concealed patterns and properties. Spatial interpolation methods (SIM) try to analyse and interpret spatial variability using statistics, mathematics, and geographic assumptions. Retrieving the values for the unknown new points from the well-known stored points without a big deviation from the original value is called as spatial interpolation. There are about 42 different spatial interpolation methods such as Nearest Neighbour (NN), Triangular Irregular Network (TIN), Inverse Distance Weighted (IDW), Kriging, etc. These methods are divided into three categories [4].

a) Non-geostatistical methods
b) Geostatistical methods
c) Combined methods

A. Problem Description

There are three parties involved in this scheme. The first party is the data owner which collects data for a specific subject. The second party is the cloud service provider which stores data and conducts necessary computations in place of the data owner. CSP can be a public cloud like Google, Amazon, etc. The CSP also can provide services like SaaS, PaaS, IaaS, DaaS, and DbaaS. Final party is the query owner. Query owner can be an organization, a company or a single user. Database (referred as D hereafter) which is stored in CSP in an encrypted form with the key owned by DO. Data stored in D are points (e.g. location) and the requests are geometric operations like finding a point or calculating nearest neighbours. Stored database on CSP is encrypted with the public key (pk) of DO.

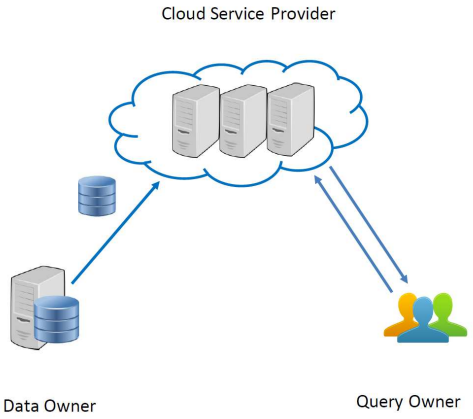

Fig. 1. The system model includes a data owner, query owner and a cloud service provider.

The public and private key pair of DO are generated using a homomorphic encryption scheme.

(i) DO has a spatial database D.
(ii) DO hires a CSP to store the D in encrypted form.
(iii) QO sends a request to CSP to get a prediction value for an unmeasured location.
(iv) k-NN spatial interpolation method is employed to predict the value by the CSP
(v) Finally, CSP sends the prediction value to the QO.

Data stored by DO

$$\begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ . & . & . \\ . & . & . \\ . & . & . \\ x_k & y_k & z_k \end{bmatrix}$$

Data stored by CSP

$$\begin{bmatrix} \xi_{pk}(x_1) & \xi_{pk}(y_1) & \xi_{pk}(z_1) \\ \xi_{pk}(x_2) & \xi_{pk}(y_2) & \xi_{pk}(z_2) \\ . & . & . \\ . & . & . \\ . & . & . \\ \xi_{pk}(x_m) & \xi_{pk}(y_m) & \xi_{pk}(z_m) \end{bmatrix}$$

In this framework, nobody wants to reveal their private data to others. In other words, DO does not want to reveal his database D to neither CSP nor C. Additionally, both DO and CSP cannot learn the query point which is the private data of QO. The secure outsourcing of k-NN method will be utilized under these data privacy conditions.

B. Security Definition

We consider semi-honest (honest-but-curious) model which means that each party follows the steps of the proposed protocol in the right manner, but tries to obtain clues about the data of the other party from the messages sent to it.

The system model includes a data owner, query owner and a cloud service provider (Fig. 1).

This paper is constructed as follows: Background of this work explained in Section 2. We proposed a solution in Section 3. Next we analyse the solution in Section 4 and Section 5 concludes the paper and suggests the future work.

II. Related Work

The increase in the amount and variety of the data causes the costs of the institutions to increase day by day. This situation necessitates transferring the storage and computation complexity of institutions from client-server architecture to cloud infrastructure. However, the concern that data privacy may be violated is seen as the biggest obstacle in front of the institutions [5]. Health and financial data are protected by laws in many countries. Conscious or unconscious violations of these laws cause both financial and prestige losses of institutions [6]. Protocols that provide data privacy between parties will abolish such concerns.

Encrypting data before transferring to cloud servers provides privacy. However, encryption does not allow some calculations to be made on joint data. Secure multi-party computations (SMC) initiated by Yao [7] are able to calculate a common function using private inputs of two or more parties. Yao employs garbled circuits to reach his objective. On the other hand, other protocols use secret sharing [8], [9]. Many applications that require privacy have been solved with SMC protocols. The first real application of SMC proposed by Bogetoft et al. [10]. They implemented a secure auction system between many sellers and buyers. Andrychowicz et al. [11] presented how SMC protocols can adapt essential Bitcoin properties to their solutions.

Fix and Hodges [12] suggested a non-parametric method for classification algorithms which has since been known as k-nearest neighbours. It has a significant role in machine learning, image processing and spatial analysis. The k-NN is a method that is frequently used in the analysis of minerals in the soil, determining the position of the interior space, and in image processing applications. In machine learning applications, k-NN is used as a classifier method. Joachims [13] used Support Vector Machines and many other classification algorithms to categorize texts. There are various spatial interpolation methods used in geostatistics. Xin et al. [14] implemented a spatial interpolation method based k-NN and IDW to analyse soil nutrient.

III. Background

A. k-NN Method

There are many algorithms used to find similarity metrics in data mining [15]. One of the most commonly used algorithms is k-NN. Because it is easy to apply, fast to train new data and effective for large amounts of data. Generally, points are represented as distance metrics in the k-NN algorithm. It finds the k number of nearest points for a given point using several distance metrics like Euclidean, Manhattan and so on. The steps of k-NN algorithm are as follows:

(i) Compute the distances from all training points to the query point.
(ii) Choose the k nearest points around the query point.
(iii) Class labels are used to predict the class of query point or average of k points is calculated for interpolation methods.

The value of k should be decided carefully. A small k value takes immediate measurements into account. If large k values are selected, it is possible to include the locations where the similarity is low in the calculation process. Therefore, the optimal k value is one of the research topics. kd-tree is one of the binary trees used to represent the spatial data in a multidimensional space. Points are divided in two ways to build a kd-tree. First one by a median of the x-axis. A vertical line divides the points into two halves. Half of the points are on the left side of the line and another half on the right side. A horizontal line is drawn by the median of the y-axis to divide the points in two. One of the halves is in top the line and remaining in the bottom.

B. Oblivious RAM (ORAM)

In a client-server architecture model, a client wants to perform an operation on the data which is stored in a server where the client does not want to share any information about the operation. Because it does not trust the server. Because of this neither the stored data nor the access pattern of the data would be revealed to the server. Goldreich and Ostrovsky [16] suggest a solution to overcome this called as ORAM. ORAM works as follows: all the data were stored as blocks (N blocks). These blocks are encrypted before sending to the server. If the client wants to access one of the stored blocks (i^{th} block), client access all the blocks one by one and when it gets the i^{th} block, it works on it. Then re-encrypts all the block and send back to the server. The server cannot identify any common patterns to identify i^{th} block if it accessed by the client several times.

The very first idea of constructing hierarchical data-structure modelling for the RAM was introduced by Goldreich and Ostrovsky [16]. The idea is ORAM stored in buckets, size of the buckets increases in a geometrical series. Smallest bucket in the top and it increases in downwards. In worst case scenario it needs $O(nlog^2n)$. A tree-based memory structure ORAM is suggested by Shi et al. [17].

The structure of constructing the tree is:

(i) The client wants to store n number of array blocks $[M_1, M_2 \dots M_n]$ in a server.
(ii) Size of a block B = $log^c n$ where c is a constant.
(iii) Binary tree with n leaves and the height of $logn$ is constructed to store the memory.
(iv) A node in the tree considered as a bucket and a bucket contains Z number of blocks.
(v) A path $p(i)$ in the tree belongs to block i.
(vi) Check each and every bucket which is a match to M_i. If it is not matched, it will be discarded.

The accessed block needs to be stored again in the memory. There is a possibility of leakage of information if the block stored in the previous place when reading the block again and again. To avoid this situation the block is re-encrypted and send it back to the top node.

C. Homomorphic Encryption (HE)

A special type of cryptography that provides mathematical operation on cipher-texts called homomorphic encryption. Many researchers suggested several algorithms for homomorphic encryption likes Paillier cryptosystem [18], Goldwasser-Micali [19], ElGamal [20], Boneh-Goh-Nissim [21] encryption schemes and so on. Paillier encryption system provides two algebraic operations on encrypted text. They are additive and multiplicative operation. Assume that ξ_{pk} is the encryption function and ξ_{qk} is the decryption function where pk is the public key and the private key is qk in this scenario. To retrieve x from $\xi_{pk}(x)$ you need to know about qk, without any knowledge about sk nobody cannot decrypt the encryption function.

Homomorphic Addition: $D_{sk}((\xi_{pk}(a) \cdot \xi_{pk}(b)) mod N^2) = (a + b) mod N$

Homomorphic Multiplication: $D_{sk}((\xi_{pk}(a)^b) mod N^2) = (a \cdot b) mod N$

where a and b are plain-texts, N is a product of two large prime numbers.

IV. Proposed Solution

As explained above, there are three parties involved in the outsourcing of secure k-NN interpolation method: Data Owner, Cloud Service Provider and Query Owner. The database which stores all measured data is possessed by DO. CSP provides storage and computation services. QO is interested in the query point where it may plan their future investments. Our solution follows the steps explained below (Fig. 2):

(i) First, DO builds a kd-tree based on the points and corresponding measurements in encrypted form. It sends the kd-tree to CSP which acts as an ORAM server.
(ii) QO sends its query point to CSP. However, query owners must send their inputs in encrypted form to hide coordinates where they need prediction values.
(iii) CSP processes the query and determines the k-NN points around the query coordinate. CSP cannot compute the prediction value because all data are encrypted with the key of DO. Therefore, CSP calculates $\prod_{i=1}^{k} \xi(z_i) = \xi(z_1) \cdot \xi(z_2) \cdots \xi(z_k)$ which equals $\sum_{i=1}^{k} z_i$ in decrypted form and sends to DO.
(iv) DO has the required key to decrypt the coming value from CSP. It decrypts and gets the prediction value. The prediction value, alone does not reveal information about query coordinate. Therefore, DO cannot learn the coordinate where the query owner is interested.
(v) DO encrypts the prediction value with the public key of the query owner to hide the prediction value from CSP and sends to CSP.
(vi) CSP forwards the prediction value in cipher-text form to the query owner.
(vii) Query owner has the corresponding private key to open cipher-text and gets the value where it needs a prediction.

V. Analysis of Proposed Solution

A. Supplementary Cost Analysis

DO provides all functionalities like storing the data and computing the prediction value for the location requested by a query owner in the traditional spatial interpolation architecture. However, the private data of both parties are at risk. On

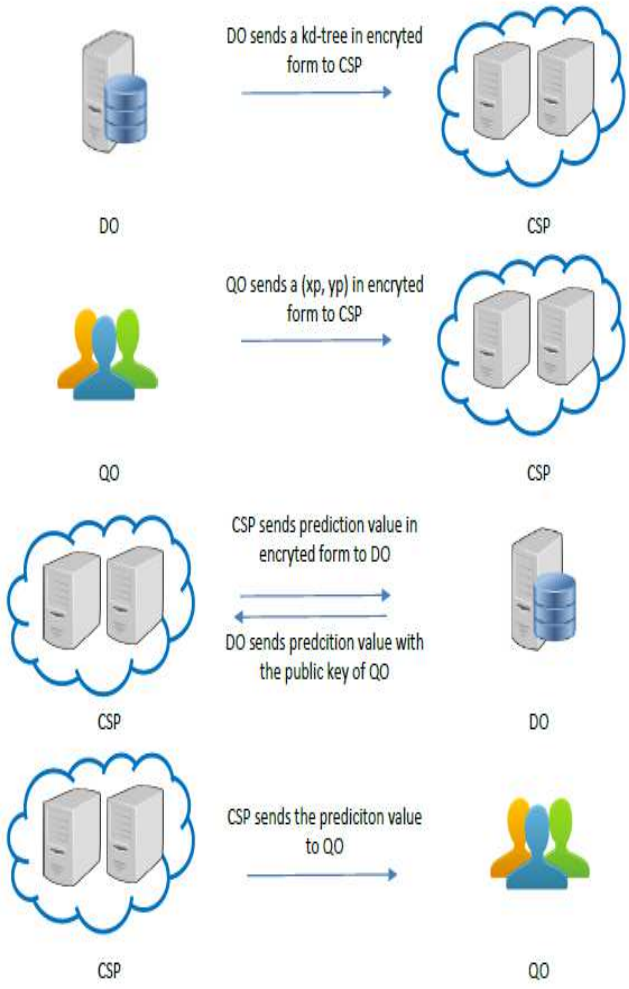

Fig. 2. Framework of secure k-NN spatial interpolation method.

the other hand, our idea of using privacy-preserving scheme protects the privacy of both parties. Besides, our work cause additional computation, storage and communication cost at negligible level than the traditional scheme. These additional costs do not affect the overall performance of spatial interpolation process. Due to the fact that spatial interpolation methods are not a time critical process as in real-time systems, researchers may prefer our solution when their privacy is important.

1) Computational Cost Analysis: Cloud computing is more efficient and profitable than the traditional client-server architecture. However, the privacy of the parties' data should be treated more carefully. Our solution which ensures the privacy with the help of homomorphic encryption and secure multi-party computation methods proposes that CSP has to carry out all necessary calculations instead of DO. Using privacy-preserving methods may increase computation time, but the resulting delay will be negligible for all parties.

DO constructs a kd-tree based on the spatial data in encrypted form and sends the tree to the CSP. CSP stores the received data as an ORAM server. Meanwhile, QO sends a query q to CSP where the q is encrypted with DO's public key. CSP finds the k number of nearest neighbours around the query. Then CSP calculates $\prod_{i=1}^{k} \xi(z_i) = \xi(z_1) \cdot \xi(z_2) \cdots \xi(z_k)$ and sends it to DO. Because, the stored data is encrypted with the DO's public key. So CSP cannot calculate the prediction value on its own. DO decrypts the cipher-texts coming from CSP and gets the prediction value. The final prediction value does not give any information about the query point of QO. After that, prediction value is encrypted by DO with QO's public key and sends it to CSP. CSP forward the encrypted value coming from DO to QO. Finally, QO decrypts the message and learns the prediction value. In the traditional method, DO receives a query location (x_p, y_p) from QO and computes the prediction value and sends the result back to the QO. But in our scenario there are three encryption, two decryption and two mathematical calculation in overall. These additional operations increase the computational cost.

2) Communication Cost Analysis: As expected, communication cost may increase when outsourcing data to cloud servers. DO sends its whole data in encrypted form to CSP. The CSP computes the nearest points and sends them to DO. DO has the required secret key to obtain the prediction value in plain text. DO encrypts the prediction value with the public key of the query owners and sends it back to CSP. QO receives the encrypted prediction value from CSP. The query owner has the necessary key to decrypt the cipher-text. Overall, there are five extra communication activities between DO, CSP and QO. However, in the traditional system there is only two communication between DO and QO.

3) Storage Cost Analysis: In a traditional system, DO is the only party that stores the database. However, in our proposed solution both DO and CSP have to store the database. As a result, the storage capacity requirement of the proposed solution is twice as much as the traditional system and negligible amount of space to store variables during communication and computation process.

B. Accuracy Analysis

Agrawal and Srikant [22] suggested a privacy-preserving technique called data perturbation. Their technique adds additional noise to ensure the privacy of the owner's data. But, this lead to the loss in the accuracy of the prediction model. However, we used a special type of encryption system called Paillier system [18]. In Paillier, the mathematical calculation can be done on the cipher-text, which does not affect the result of the operation. Therefore, our work produces the same prediction values as in traditional scheme. Beyond that it guarantees the confidentiality of the private data of all participants.

C. Privacy Analysis

In our work, data need to be protected in four different places from all three participants. CSP cannot learn the data which is stored in CSP in encrypted form with the key of DO. The second step, when QO sends his request in an encrypted form to hide its points from the CSP and DO. The third one is, CSP process the query to find the k-NN points for the requested coordinate. At this time CSP cannot predict the

value. Because all the data were encrypted using the DO's key. After determining the k nearest points, CSP sends the prediction value in encrypted form to DO. DO decrypts and gets the prediction value. In the fourth step, DO cannot get any information about the query coordinate. Therefore, DO cannot predict the location of the requested coordinate. Finally, DO encrypts the predicted value with the public key of QO and send it to CSP. Only QO can decrypt with its private key. CSP does not have any way to learn the predicted value. CSP forwards the cipher text to QO. QO decrypts the received cipher-text and get the predicted value. As a summary, DO cannot predict the query point from QO. CSP cannot predict and/or measure the stored data by the DO, query point from QO and predicted value by the DO. Finally, QO can only get the final predicted value from the DO.

VI. Conclusion

Spatial interpolation is one of the essential operations of geographic information systems. There is a variety of interpolation methods used by researchers. k-NN interpolation method is preferred due to its simplicity and efficiency in computation. Recently, companies tend to move their data and computation needs to cloud servers. However, their privacy may be at risk. Therefore, we propose a k-NN spatial interpolation method which protects privacy of all participants during the process. We are aware of that privacy-preserving methods lead to additional storage, communication and computation cost. We analysed our solution in terms of these costs and presented that they are not that critical from the perspective of all participants.

References

[1] T. Tassa, "Secure mining of association rules in horizontally distributed databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 4, pp. 970–983, 2014.

[2] L. Sun, W.-S. Mu, B. Qi, and Z.-J. Zhou, "A new privacy-preserving proximal support vector machine for classification of vertically partitioned data," *International Journal of Machine Learning and Cybernetics*, vol. 6, no. 1, pp. 109–118, 2015.

[3] S. Kumar and R. Goudar, "Cloud computing-research issues, challenges, architecture, platforms and applications: A survey," *International Journal of Future Computer and Communication*, vol. 1, no. 4, p. 356, 2012.

[4] J. Li and A. D. Heap, "Spatial interpolation methods applied in the environmental sciences: A review," *Environmental Modelling & Software*, vol. 53, pp. 173–189, 2014.

[5] R. L. Krutz and R. D. Vines, *Cloud Security: A Comprehensive Guide to Secure Cloud Computing*. Wiley Publishing, 2010.

[6] S. Pearson and A. Benameur, "Privacy, security and trust issues arising from cloud computing," in *Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on*. IEEE, 2010, pp. 693–702.

[7] A. C. Yao, "Protocols for secure computations," in *23rd Annual Symposium on Foundations of Computer Science*. IEEE, 1982, pp. 160–164.

[8] A. Shamir, "How to share a secret," *Communications of the ACM*, vol. 22, no. 11, pp. 612–613, 1979.

[9] M. Ito, A. Saito, and T. Nishizeki, "Secret sharing scheme realizing general access structure," *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, vol. 72, no. 9, pp. 56–64, 1989.

[10] P. Bogetoft, D. L. Christensen, I. Damgård, M. Geisler, T. Jakobsen, M. Krøigaard, J. D. Nielsen, J. B. Nielsen, K. Nielsen, J. Pagter *et al.*, "Secure multiparty computation goes live," in *International Conference on Financial Cryptography and Data Security*. Springer, 2009, pp. 325–343.

[11] M. Andrychowicz, S. Dziembowski, D. Malinowski, and L. Mazurek, "Secure multiparty computations on bitcoin," in *2014 IEEE Symposium on Security and Privacy*, 2014, pp. 443–458.

[12] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: consistency properties," *International Statistical Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.

[13] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning*. Springer, 1998, pp. 137–142.

[14] X. Xu, H. Yu, G. Zheng, H. Zhang, and L. Xi, "The soil nutrient spatial interpolation algorithm based on knn and idw," in *Computer and Computing Technologies in Agriculture IX*. Springer International Publishing, 2016, pp. 412–424.

[15] S. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *International Journal of Mathematical Models and Methods in Applied Sciences*, vol. 1, no. 4, pp. 300–307, 2007.

[16] O. Goldreich and R. Ostrovsky, "Software protection and simulation on oblivious rams," *Journal of the ACM (JACM)*, vol. 43, no. 3, pp. 431–473, 1996.

[17] E. Shi, T.-H. H. Chan, E. Stefanov, and M. Li, "Oblivious ram with o ((logn) 3) worst-case cost," in *International Conference on The Theory and Application of Cryptology and Information Security*. Springer, 2011, pp. 197–214.

[18] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 1999, pp. 223–238.

[19] S. Goldwasser and S. Micali, "Probabilistic encryption," *Journal of computer and system sciences*, vol. 28, no. 2, pp. 270–299, 1984.

[20] T. ElGamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," *IEEE transactions on information theory*, vol. 31, no. 4, pp. 469–472, 1985.

[21] D. Boneh, E.-J. Goh, and K. Nissim, "Evaluating 2-dnf formulas on ciphertexts," in *Theory of Cryptography Conference*. Springer, 2005, pp. 325–341.

[22] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *ACM Sigmod Record*, vol. 29, no. 2. ACM, 2000, pp. 439–450.