Review on a Privacy-Preserving and Efficient k-Nearest Neighbor Query and Classification Scheme Based on k-Dimension Tree for Outsource Data

Abstract-- Cloud computing technology has attracted the attention of researchers and organizations due to its computing power, efficiency and flexibility. Using cloud computing technology to analyze outsourced data is become a new data utilization model. However, due to the severe security risks that appear in cloud computing, most organizations now encrypt data before outsourcing data. Therefore, in recent years, many works on the k-Nearest Neighbor (denoted by k-NN) algorithm for encrypted data has appeared. However, two main problems in existing current research are either the program is not secure enough or inefficient. In this paper, based on the existing problems, we have designed a non-interactive privacy-preserving k-NN query and classification scheme. Our proposed scheme uses two existing encryption schemes: Order Preserving Encryption and the Parlier cryptosystem, to preserve the privacy of encrypted outsourced data, data access patterns, and the query record, and utilizes the encrypted the k-dimensional tree (denoted by kd-tree) to optimize the traditional k-NN algorithm. Our proposed scheme aim to achieve high query efficiency while ensuring data security. Extensive experimental results prove that this scheme is almost close to the scheme using plaintext data and the existing non-interactive encrypted data query scheme in terms of classification accuracy. The query runtime of our scheme is higher than the existing non interactive k-NN query scheme.

Keywords-- Privacy preserving, k- nearest neighbour, k-dimensional tree, outsourced data.

1. INTRODUCTION

Nowadays, machine learning and cloud computing have been widely used. Machine learning can mine hidden knowledge or patterns from massive data and is one of the most attractive technologies. The K-Nearest Neighbor (denoted by K-NN) algorithm is one of classic machine learning algorithms, which can find the nearest k points from a large data set based on the test object. It has been used in many studies, such as pattern recognition, Location-Based Services, DNA sequencing, online recommendation systems, and data analysis, etc.

KNN firstly computes similarities between input query and each data in dataset (Compute Similarity), converts the similarities in bitwise shared representation (Bit-Decomposition), and selects K data with the highest similarities (PE-FTK). Among the sub protocols.

Update the feature extractor of a KNN model. Train a student model in the public Domain.

1. Update the feature extractor for private-KNN: We initialize the feature extractor with a public extractor --- Histogram of Oriented Gradient (HOG) features. We use the neural network of the last iteration student model (except for the last softmax layer) to update the feature extractor, in the next iteration. Note that this interactive scheme will iteratively refine the feature embedding used by KNN without using any exclusive information.

2. Train a student model: When the feature extractor of Private-KNN is updated, we train a student model by labeling a limited number of student queries (the public data) with pseudo-labels. For each student query, we first generate a random subset from the entire private domain, and then pick the k nearest neighbors among the subset. The pseudo-label is generated with private voting of k neighbors, and the detailed aggregation process can be found in the main paper.

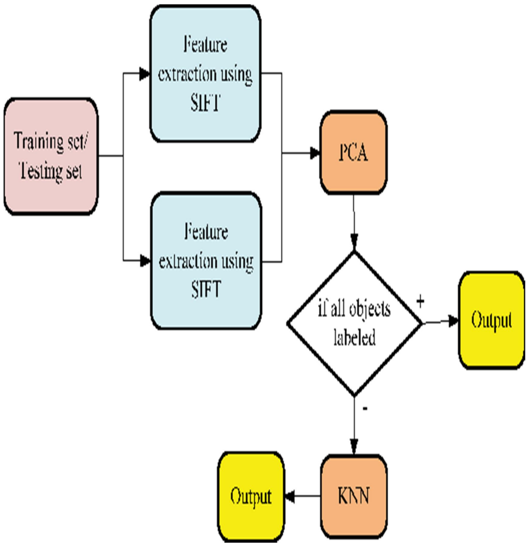
Training set/ Testing set
Feature extraction using SIFT
Feature extraction using SIFT
PCA
if all objects labeled
+
Output
-
KNN
Output

Fig 1: Architecture of NLP

2. OBJECTIVES

The Private-KNN is the data-efficient algorithm for differentially private (DP) deep learning under the knowledge transfer framework. It represents the practical solution that addresses this important problem scales to larger models while preserving theoretically meaning full DP guarantee.

3. RELATED WORK

Y. Du, " Privacy-aware rnn query processing on location-based services in Mobile Data Management , Reverse K nearest neighbors query processing: experiments and analysis," VLDB Endowment, vol. Zhu et al. [2] improved the scheme of Wong et al. to provide privacy-preserving k-NN query. However, the data owner has to participate in the query process, however the scheme lacks a rigourous security proof. Hu et al.[3]proposed a k-NN query scheme based on privacy homomorphism encryption scheme. In their scheme, the k-NN query on the encrypted data is achieved by homomorphic properties. How one ensured an individual's privacy regarding his location and spatiotemporal behavioral patterns was proposed by Ho et al. (2016) through differential privacy mechanism which assumes that data trajectory is secure and users can only query knowledge derived from it. The proposed system demonstrated privacy preserving approach on frequent location pattern mining task. However, Yao et al. pointed out that both schemes can not resist chosen- plaintext attacks, and thus they proposed a new scheme based on partition-based secure Voroni diagram.

Elmehdwi et al. [5] designed a homomorphic encryption based k-NN query scheme over encrypted data. Although the data owner and the client achieve the privacy, the computation and communication over-heads are not very efficient. Recently, Xu et al. [6] also proposed a scheme with the sublinear computation complexity during the k-NN query process. [7] Dpsense: Differentially private crowd asources specirum sensing," in Processing of the 2015 ACM SIGSAC Conference on Computer and Communication Security. Private queries in location based services: anonymiuzers are not necessary in Processing on ACM SIGMOD. Evaluating k nearest neighbor query on road networks with no information leakage in WISE 2014.

4. TECHNIQUES

In the paper, use UDA to train the student model for SVHN and CIFAR-10 tasks, which allows us to save the privacy budget with a limited number of student queries. RKNN Query Answer Retrieval Algorithm used it.

1. The Encryption technique is Privacy Preserving:
In this scheme, the cloud can't obtain the value of any data (including query data records) because we use two encryption schemes, the OPE and Paillier cryptosystem, to encrypt the raw data and execute the *k*-NN algorithm. As mentioned earlier, the *honest-but-curious* cloud server cannot know any information about the encryption class, or even how many different classes in the scheme. And all intermediate results and the result of the query are encrypted data, and the cloud server cannot learn their plaintext data. Therefore, the encryption technique is privacy-preserving.

2. The Encrypted Data Comparison Protocol is Privacy-Preserving:

The data comparison protocol is a key component of the *k*-NN query algorithm. As men- tioned earlier, the comparison protocol is used in both the kd-tree technique and the comparison of distances. In our proposed scheme, we implemented the *k*-NN algorithm with the stateless OPE scheme presented in 2011. From this protocol, we can only obtain a comparison of two encrypted data *E*(*a*) and *E*(*b*) with- out leaking any plaintext values of *a* and *b*. Therefore, the cloud server cannot directly get the plaintext data content. Similarly, Nor can unauthorized users. Thus, this comparison protocol is privacy-preserving.

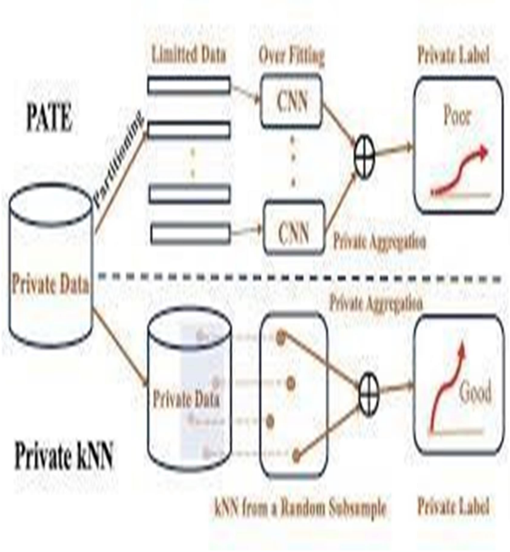

Fig 2: Proposed work models

.

5. DATA DESCRIPTION

In this paper, three schemes are used to implement the *k*-NN algorithm: the classical plaintext *k*-NN classier, the scheme in and our proposed scheme. In the scheme we proposed and the scheme proposed in, a stateless OPE scheme and the Parlier homomorphic encryption scheme are used to encrypt data, including outsourced data and query records of authorized users. To be specific, the two schemes adopt the OPE scheme to encrypt the property values of the data and use the Parlier encryption scheme to encrypt the class labels of the data. Moreover, in the process of data query and classification, the comparison protocol and the computation protocol are executed between encrypted data, so the privacy-preservation of all data is achieved. Therefore our proposed scheme and the scheme proposed in have the same security.

Table 1: Dataset Description For Dataset Used In KNN

Data sets	No of classes	No of instance	No of features
MNIST	1000	286	30
SVHN	100	45	10
CelebA	600	163	45

Table 2: Proposed Plan of Work

Duration Of Work	Action To Be Taken
Sep 2020	Literature review
Sep – Oct 2020	Publishing Review Paper & Implementing 1^{st} module i,e generating group of features
Oct – Nov 2020	Implementing of another modules group feature selections
Nov – Dec 2020	Paper Publication on Problem Definition
Dec 20 – Jan 2021	Classification of features
Jan – Feb 2021	Testing of Project
Feb – March 2021	Paper Publication on Result Analysis
March – April 2021	Report writing

6. CONCLUSSION

In this paper , proposed two novel solutions RKNN- HG nad RKNN-HRT to answer private RKNN queries without disclosing any information about the location of query point. Our solutions utilize Private Information Retrieval (PIR) mechanism to request data from an untrusted database server without the server learning about retrieved data or the query source. We evaluated our methods extensively using real datasets studying the effect of servral parameters on computational cost and data overhead size. Our results show the efficiency and effectiveness of our solutions. Our future work includes extending our solutions for Bichromatic RKNN queries and moving object queries for spatial crowdsourcing applications while protecting the location privacy of the participants**.**

REFERENCES

[1] Y. Du, “Privacy-aware rnn query processing location-based services,” in Mobile Data Management. IEEE, 2016, pp253-257.

[2] B. Balle, G. Barthe, and M. Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In Preprint, 2018.

[3] R. Bassily, A. Smith, and A. Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In Foundations of Computer Science (FOCS-14), pages 464–473. IEEE, 2014

[4] M. Bun and T. Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In Theory of Cryptography Conference, pages 635–658. Springer, 2016.

[5] N. Carlini, C. Liu, U. Erlingsson, J. Kos, and D. Song. The ´ secret sharer: Evaluating and testing unintended memorization in neural networks. In 28th USENIX Security Symposium (USENIX Security 19), pages 267–284, Santa Clara, CA, Aug. 2019. USENIX Association.

[6] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. The Journal of Machine Learning Research, 12:1069–1109, 2011.

[7] T. Cover and P. Hart. Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1):21–27, 1967.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. 2005.

[9] C. Dimitrakakis, B. Nelson, A. Mitrokotsa, and B. I. Rubinstein. Robust and private bayesian inference. In Algorithmic Learning Theory, pages 291–305. Springer, 2014. 1.

[10] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In Theory of cryptography, pages 265–284. Springer, 2006.

[11] C. Dwork, G. N. Rothblum, and S. Vadhan. Boosting and differential privacy. In Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on, pages 51–60. IEEE, 2010.

[12] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. What can we learn privately? SIAM Journal on Computing, 40(3):793–826, 2011. 1.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradientbased learning applied to document recognition. In Proceedings of the IEEE, 1998.

[14] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In ICCV, 2015.

[15] Mironov. Renyi differential privacy. In ´ 2017 IEEE 30th Computer Security Foundations Symposium (CSF), pages 263–275. IEEE, 2017.

[16] Mironov. Renyi differential privacy. In ´ Computer Security Foundations Symposium (CSF), 2017 IEEE 30th, pages 263– 275. IEEE, 2017.

[17] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In NIPS Workshop, 2011.

[18] Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In ACM symposium on Theory of computing (STOC-07), pages 75–84. ACM, 2007.

[19]M. Park, J. Foulds, K. Chaudhuri, and M. Welling. Variational bayes in private settings (vips). arXiv preprint arXiv:1611.00340, 2016