Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Privacy-preserving machine learning with multiple data providers

Ping Li [a], Tong Li [b], Heng Ye [c], Jin Li [a,*], Xiaofeng Chen [d], Yang Xiang [e]

[a] School of Computer Science, Guangzhou University, Guangzhou 510006, China
[b] College of Computer & Control Engineering, Nankai University, 300071, Tianjin, China
[c] Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing Jiaotong University, 3 Shangyuancun, Beijing 100044, China
[d] State Key Laboratory of Integrated Service Networks, Xidian University, 710126, Xi'an, China
[e] School of Information Technology, Deakin University, Melbourne Burwood, VIC 3125, Australia

HIGHLIGHTS

- To protect data privacy, multiple parties encrypt their data under their own public key of double decryption algorithm, before outsourcing it to cloud for storing and processing.
- To improve the efficiency and accuracy of the computation, cloud transforms the encrypted data into noised data, such that the machine learning algorithm can be executed on this noised data with ϵ-differential privacy.
- The proposed scheme is proven to be secure in the security model.

ARTICLE INFO

Article history:
Received 20 November 2017
Received in revised form 30 March 2018
Accepted 23 April 2018
Available online xxxx

Keywords:
Differential privacy
Homomorphic encryption
Outsourcing computation
Machine learning

ABSTRACT

With the fast development of cloud computing, more and more data storage and computation are moved from the local to the cloud, especially the applications of machine learning and data analytics. However, the cloud servers are run by a third party and cannot be fully trusted by users. As a result, how to perform privacy-preserving machine learning over cloud data from different data providers becomes a challenge. Therefore, in this paper, we propose a novel scheme that protects the data sets of different providers and the data sets of cloud. To protect the privacy requirement of different providers, we use public-key encryption with a double decryption algorithm (DD-PKE) to encrypt their data sets with different public keys. To protect the privacy of data sets on the cloud, we use ϵ-differential privacy. Furthermore, the noises for the ϵ-differential privacy are added by the cloud server, instead of data providers, for different data analytics. Our scheme is proven to be secure in the security model. The experiments also demonstrate the efficiency of our protocol with different classical machine learning algorithms.

1. Introduction

With the fast development of cloud computing, more and more data and applications are moved from the local to cloud servers, including machine learning and other data analytics. However, the cloud computing platform cannot be fully trusted because it is run by a third party. Cloud users lose the control of their data after outsourcing their data to the cloud. To protect the privacy, the data are usually encrypted before they are uploaded to the cloud storage. However, the encryption techniques render the data utilization difficult.

Though there are some traditional techniques such as homomorphic cryptographic techniques to provide solutions for the data utilization over encrypted data, they are inefficient in practice. To address this challenge, another important notion of differential privacy has been proposed. It can not only protect the privacy, but also provides efficient data operations.

However, most of the previous mainly focus on the data from a single user. It is common that the data always from different data providers for machine learning. Therefore, how to perform machine learning over cloud data from multiple users become a new challenge. Traditional differential privacy technique and encryption methods are not practical for this environment. On one hand, the data from different users are encrypted with different public keys or noises, which makes the computation be difficult. On the other hand, data have to be proceeded in different ways for different applications, which makes both the communication overhead and computation overhead be huge.

* Corresponding author.
E-mail addresses: liping26@mail2.sysu.edu.cn (P. Li), litongziyi@mail.nankai.edu.cn (T. Li), heng.ye@bjtu.edu.cn (H. Ye), jinli71@gmail.com (J. Li), xfchen@xidian.edu.cn (X. Chen), yang@deakin.edu.au (Y. Xiang).

https://doi.org/10.1016/j.future.2018.04.076
0167-739X/

Main idea. To tackle the above challenges, we propose a scheme named privacy-preserving machine learning under multiple keys (PMLM) to solve this problem. Since the secure multi-party computation (SMC) only supports the computation on the data encrypted under the *same public key* and the efficiency and accuracy of the computation need to be improved. Therefore, our PMLM scheme as an efficient solution is required that conducts the data encrypted under *different public keys* for different data providers and improves the efficiency and accuracy. Our novel technique based on a new public-key encryption with a double decryption algorithm (DD-PKE) and differential privacy. The DD-PKE is additively homomorphic scheme and holds two independent decryption algorithms which allows the outsourced data set to be transformed into randomized data. The differential privacy can be used to add statistical noises to the outsourced data set for data analyses and data computations.

Our PMLM scheme works as follows. First, we set up a public-key encryption with a double decryption algorithm (DD-PKE) to protect the data privacy of multiple data providers. During this phase, we do not take the differential privacy protection into consideration. We then use a cloud server to add different statistical noises to outsourced ciphertexts according to the different applications of the data analyst, and these noises are encrypted under a public key corresponding to the outsourced ciphertexts. Finally, the data analyst downloads this noise-added ciphertext data sets, decrypts it using his or her own master key and performs a machine learning task over this joint distribution with minimum error.

Our Contributions. In our PMLM scheme, we assume that the cloud server and data analyst are not collude with each other and that they are *semi-honest*. In all steps of PMLM scheme, the multiple users do not interact with each other. We show that our PMLM scheme is IND-CCA secure in the random oracle model.

In particular, the main contributions of this work are summarized as follows:

- In this work, the cloud server has the authority to add different statistical noises to the outsourced data set according to different queries of the data analyst rather than the data providers adding statistical noise by themselves with only one application.
- We use a DD-PKE cryptosystem to preserve the privacy of the data providers' data sets, which can be used to transform the encrypted data into a randomized data set without information leakage.
- In our PMLM scheme, the machine learning task is performed on a randomized data set with ϵ-differential privacy rather than on the encrypted data set. This process improves the computational efficiency and data analysis accuracy.

Organization of the Paper. The remainder of this paper is organized as follows. Section 2 provides a literature review over privacy-preserving machine learning based on differential privacy protection. Section 3 presents some notations and definitions on cryptographic primitives and differential privacy. In Section 4, we present the system model, the problem statement and the adversary model. In Section 5, we provide the PMLM scheme. Then, we present our simulation results in Section 6 and the security analysis in Section 7. Finally, the conclusions and directions for future work are presented in Section 8.

2. Related work

Machine learning is the process of programming computers to optimize a performance criterion using example data or prior experience. Because of its powerful ability to process large amounts of data, machine learning has been applied in various fields in recent years, including speaker recognition [1], image recognition [2,3] and signal processing [4]. To protect the data privacy in the machine learning model, two well-known lines of research should be considered in our work.

2.1. Homomorphic encryption in machine learning

There are many works considering the problem of privacy preserving for outsourced computation. Homomorphic encryption is one of the basic techniques, which can be also applied in machine learning. To protect the privacy of users' sensitive data, users only provide the encrypted data for data storing and data processing. For instance, Chen et al. [5] presented a privacy-preserving two-party distributed algorithm of back-propagation neural networks (BPNN) which allows a neural network to be trained without revealing the information about each of party. To preserve the privacy of input data and output result, they used a homomorphic scheme to keep the security. In their work, the BPNN conducts the vertically partitioned data, i.e., each party has a subset of feature vector. Due to their scheme only process vertically partitioned data, in the subsequent work, Bansal et al. [6] proposed a similar scheme for privacy-preserving BPNN over arbitrarily partitioned data between two parties. However, all works [5,6] cannot be applied to the multi-party scenario because directly extending them to the multi-party scenario will lead to the communication overhead.

Hence, Samet et al. [7] presented new privacy-preserving protocols for both the BPNN and extreme learning machine (ELM) algorithms with horizontally and vertically partitioned data among multiple parties. Graepel et al. [8] proposed secure machine learning scheme over encrypted data, they only trained two simple classifiers, linear means (LM) and fisher's linear discriminate (FLD). Dowlin et al. [9] proposed a scheme, called CryptoNets, which used an fully homomorphic encryption scheme of Bos et al. [10] to evaluate deep convolutional neural networks (CNN) with two convolutional layers and two fully connected layers. Hesamifard et al. [11] proposed a CryptoDL scheme, which is a solution to run deep NN algorithms on encrypted data and allow the parties to provide/ receive the service without having to reveal their sensitive data to the other parties. The main work of CryptoDL is combine the CNN with leveled homomorphic encryption (LHE). Gao et al. [12] considered a situation that a user requests a naive Bayes classifier server, both the user and the server do not want to reveal their private data to each other. Their key technique involves the use of a "double-blinding" technique, and they shown how to combine it with additively homomorphic encryptions and oblivious transfer to hide both parties' privacy. There are also many other solutions by using other outsourcing computation techniques, such as [13–20].

2.2. Differential privacy in machine learning

Differential privacy [21,22] is a popular approach to privacy protection for machine learning algorithms on data sets, including Bayesian inference, empirical risk minimization (ERM), stochastic gradient descent (SGD), and so on. The main idea of differential privacy in machine learning is to learn a simple rule automatically from the distributional information of the data set at hand without revealing too much about any single individual in the data set. In fact, we often want to perform privacy-preserving machine learning as accurately as possible, just like we perform non-privacy-preserving machine learning on the same number of examples.

Dwork [23] first considered the original definition of ϵ-differential privacy protection, where the parameter ϵ (> 0) is a real number and controls how much information is disclosed

Please cite this article in press as: P. Li, et al., Privacy-preserving machine learning with multiple data providers, Future Generation Computer Systems (2018), https://doi.org/10.1016/j.future.2018.04.076.