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

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Efficiently and securely harnessing cloud to solve linear regression and other matrix operations☆

Lu Zhou [a,b], Youwen Zhu [a,c,*], Kim-Kwang Raymond Choo [d]

[a] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
[b] Division of Computer Science, University of Aizu, Fukushima 965-8580, Japan
[c] Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210016, China
[d] Department of Information Systems and Cyber Security and Department of Electrical and Computer Engineering , The University of Texas at San Antonio, San Antonio, TX 78249-0631, USA

HIGHLIGHTS

- We analyze the security vulnerabilities of existing linear regression (LR) outsourcing scheme.
- We present a new encryption scheme to securely outsource large-scale LR to an untrusted cloud, in which the input dataset is perturbed by dense matrices with high efficiency.
- We observe an equivalent condition of LR answer verification, based on which we develop a novel verification algorithm for LR outsourcing.
- We provide detailed theoretical analysis and extensive simulation experiments to evaluate the new scheme.

ARTICLE INFO

Article history:
Received 29 April 2017
Received in revised form 30 August 2017
Accepted 10 September 2017
Available online xxxx

Keywords:
Cloud computing
Secure outsourcing
Linear regression
Verification
Data perturbation

ABSTRACT

In this paper, we study the problem of efficiently outsourcing large-scale linear regression of a customer to the public cloud while preserving the privacy of the customer's input and output results. To reduce the customer's computation costs, existing schemes generally use diagonal matrix multiplication to encrypt the input data. While such approaches are efficient, there are potential security limitations. For example, in this paper we reveal previously unknown limitations in the scheme of Chen et al. (2014). We then present a novel method to generate random dense matrices, and a new secure solution for outsourcing linear regression to cloud. In our proposed approach, we perturb the customer's input/output by adding random numbers and multiplying our constructed random dense matrices. A comparative summary demonstrates that the proposed approach has a stronger level of security, without incurring additional computation complexity. We also demonstrate that our constructed dense matrices can be utilized to efficiently enhance the security of outsourcing scheme for other large-scale matrix operations, including linear equation system and determinant computation.

1. Introduction

Outsourcing to the cloud allows a resource-constrained customer to leverage the computation resources of the cloud servers to complete a wide range of computational intensive tasks, typically based on a flexible pay-per-use manner [1,2]. There are several known challenges in such outsourcing activities. First, the input data and output results of the customer may contain sensitive information, and should not be available to even the cloud service provider processing the data. Thus, the customer needs to encrypt/perturb his/her input data prior to outsourcing, so that the cloud server only learns minimal information about the data and the computation outcomes. While existing conventional encryption schemes can hide such data, these schemes generally affect the functionality and performance of computation over the outsourced data.

Fully homomorphic encryption scheme is a viable solution, but at the time of this research there is no practical and efficient fully homomorphic encryption scheme [3,4]. In addition, the cloud server may return an inaccurate answer. For example, to maximize profits, the cloud server may seek to cut corners and return some invalid value if it cannot be checked. Software bugs and hardware errors may also lead to incorrect results. Thus, answer

☆ This paper is an extended version of the symposium poster (Zhu et al., 2016) with more than 70% new content.

* Corresponding author at: College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China.
E-mail address: zhuyw@nuaa.edu.cn (Y. Zhu).

https://doi.org/10.1016/j.future.2017.09.031
0167-739X/

verification is one of many challenges associated with computation outsourcing. In other words, the customer must have the ability to inexpensively verify the correctness of the returned value from the cloud. In other words, we need to ensure that the customer's efforts are minimized; otherwise, there is little incentive to use the cloud. In general, a practical outsourcing scheme should effectively address these three challenges.

Linear regression (LR) is a common statistical model in the scientific and engineering community, and has a wide range of applications [5]. For example, using a large-scale dataset generated from some real-world settings may require processing powers beyond that of a resource-limited device belonging to the customer [5,6]. Therefore, the customer can chooses to outsource the computations. Chen et al. [7] demonstrates in a recent work that if the customer perturbs his/her dataset by multiplying the diagonal matrix, the cloud server can perform LR over the perturbed data. However, there are limitations in this simple approach as we will show in Section 3.

In this paper, we are motivated to propose a secure and efficient solution for outsourcing large-scale LR to a public cloud that can achieve input privacy and result verification with high efficiency. In the proposed scheme, we encrypt the customer's dataset by adding random numbers and multiplying a random dense invertible matrix. This allows us to achieve stronger security. By including dense matrix multiplication, our scheme has the same time complexity as that of Chen et al. [7] (which uses diagonal matrix multiplication only). To guarantee that the cloud server correctly completes the outsourced computation task, we also introduce a new LR answer verification algorithm based on a series of equivalent transformations. In our verification algorithm, the customer can effectively find any deviation of the cloud server by checking a simple equation.

In general, our contributions in this paper can be summarized as follows:

1. We present a new encryption scheme to securely outsource large-scale LR to a public cloud that is resilience to an untrusted cloud server. In other words, the privacy of the customer's data and its computation is preserved in the sense that the cloud server cannot learn any useful information about the customer's private data.
2. We achieve the same computation complexity as the simple scheme of Chen et al. [7], without suffering from the same limitation that we demonstrate in their scheme.
3. We develop a novel verification algorithm that allows a customer in our scheme to rapidly and easily identify any deviation of the cloud server.

We then evaluate the new scheme and demonstrate its security and efficiency. We also show that our proposal can be utilized to efficiently enhance the security of outsourcing scheme for other large-scale matrix operations, such as linear equation system and determinant computation.

We remark that this paper significantly extends the two-page poster paper [8]. Specifically, the earlier work [8] only provides a simple description about using a random dense invertible matrix to perturb the input dataset. In this paper, we describe the system model, design goals and preliminaries in Section 2. We reveal the limitations in the scheme of Chen et al. [7] (see Section 3), prior to presenting our proposal (see Section 4). We then demonstrate the security and the performance of the proposal in Sections 5 and 6, respectively. In Section 7, we discuss how our scheme can be deployed to securely outsource other matrix operations. Related work is presented in Section 8, prior to the conclusion of this paper in the last section.

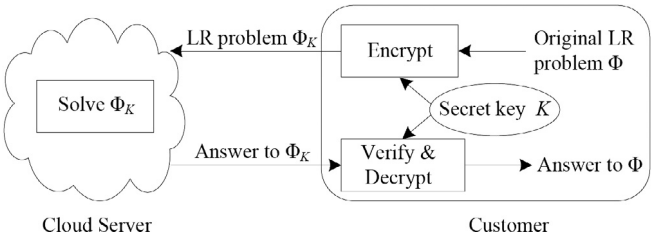

Fig. 1. Secure outsourcing LR in cloud computing architecture.

2. Problem statement and preliminaries

2.1. System model

In this paper, we consider the outsourcing system involving a customer and a cloud server, as illustrated in Fig. 1. Here, the customer has a large-scale data set $\{y_i, x_{i1}, x_{i2}, \ldots, x_{in}\}_{i=1}^{m}$ of m statistical units, each of which consists of n explanatory variables $\{x_{i1}, x_{i2}, \ldots, x_{in}\}$ and one response variable y_i. Generally, m is larger than n. All explanatory/response variables are numerical. For modeling of the relationship between response variable and explanatory variables, the customer wishes to solve an LR problem Φ taking m statistical units as input, but his/her computation power is insufficient to efficiently carry out the computation. Thus, the customer plans to outsource LR to the cloud server with the required resources (e.g. storage and computation).

The cloud server is, however, not fully trusted by the customer, and the latter does not want the cloud server to learn any meaningful information about his/her data (i.e. input) and computation outcomes (i.e. output). Therefore, the customer transforms Φ (original input dataset) into Φ_K (encrypted dataset) using his/her secret key K, and merely releases the perturbed LR problem Φ_K to the cloud server, rather than the original problem Φ. Next, the cloud server solves Φ_K, and returns the answer together with a proof (for proving the correctness of answer) to the customer. It should be pointed out that, in our scheme, the proof is the answer itself, and the cloud server does not need to prepare another special proof. Finally, the customer verifies that the cloud server has correctly completed the computation, and recovers the answer to his/her original LR problem Φ.

2.2. Design goals

We assume that the cloud server is the adversary (e.g. employee of the cloud service provider or an external attacker), who could be malicious. A malicious cloud server may be lazy, curious, and dishonest [2,7,9]. Concretely speaking, unless caught, a lazy and dishonest cloud server may not faithfully implement the solving algorithm. For example, this cloud server sends the customer a discretionary (invalid) answer in order to reduce operating cost. As being curious, the cloud server seeks to infer as much useful information about the customer's input/output as possible during the outsourcing computation. In additionally, the cloud server is assumed to know the mean value (and the average absolute value) of some columns of original dataset beyond the perturbed LR problem Φ_K. For example, the attacker may learn the average value of qth explanatory variable or its absolute value, $\frac{1}{m}\sum_{i=1}^{m} x_{qi}$ or $\frac{1}{m}\sum_{i=1}^{m} |x_{qi}|$, based on prior experience or publicly available reports (e.g. if the column denotes the salary or profit of a publicly listed organization, then one can refer to the organization's annual report).

Under the malicious threat model, our outsourcing scheme seeks to reduce the customer's computation overheads, preserve

Please cite this article in press as: L. Zhou, et al., Efficiently and securely harnessing cloud to solve linear regression and other matrix operations, Future Generation Computer Systems (2017), https://doi.org/10.1016/j.future.2017.09.031.