Information Sciences 451–452 (2018) 34–49

Contents lists available at ScienceDirect

Information Sciences

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

# Privacy-preserving ridge regression on distributed data

Yi-Ruei Chen*, Amir Rezapour, Wen-Guey Tzeng

*National Chiao Tung University, Hsinchu 30010, Taiwan*

ARTICLE INFO

*Article history:*
Received 15 August 2016
Revised 28 March 2018
Accepted 30 March 2018
Available online 30 March 2018

*Keywords:*
Privacy-preserving regression
Ridge regression
Data privacy
Recommendation system

ABSTRACT

Ridge regression is a statistical method for modeling a linear relationship between a dependent variable and some explanatory values. It is a building-block that plays a major role in many learning algorithms such as recommendation systems. However, in many applications such as e-health, explanatory values contains private information owned by different patients that are not willing to share them, unless data privacy is guaranteed. In this paper, we propose a protocol for conducting privacy-preserving ridge regression (PPRR) over high-dimensional data. In our protocol, each user submits its data in an encrypted form to an evaluator and the evaluator computes a linear model of all users' data without learning their contents. The core encryption method is equipped with homomorphic properties to enable the evaluator to perform ridge regression over encrypted data. We implement our protocol and demonstrate that it is suitable for dealing with high-dimensional data distributed among millions of users. We also compare our protocol with the state-of-the-art solutions in terms of both computation and communication costs. The results show that our protocol outperforms most existing approaches based on secure multi-party computation, garbled circuit, fully homomorphic encryption, secret-sharing, and hybrid methods.



## 1. Introduction

Ridge regression is a statistical method for modeling a linear relationship between a dependent variable and some explanatory values. It is a building-block that plays a major role in many learning algorithms such as recommendation systems. A recommendation system learns user profiles through feedback so that the recommended item falls into the interest without requiring the user to make an explicit query. Regression technique analyzes a set of collected data and summarizes them in a compact form for determining how relevant an item is for a user.

Techniques for constructing a ridge regression, like other traditional machine learning algorithms, require the data to be in plaintext form for building a model. It means that a user who engages in an online service has to share its data with a service provider for regression. However, the user may refuse to do so if the shared data contain its personal information. For instance, a nationwide electronic medical record system (EMR) that encourages sharing of medical knowledge, has the potential to improve care coordination, healthcare quality, and many other areas of healthcare. The medical records may contain sensitive individual information so that the patients' are not willing to share them unless data privacy is guaranteed. The problem can be effectively addressed with the existence of a fully trusted party that carries out the computations on the behalf of the users. However, it is hard to find such a party in practice. In order to collect data adequately, it is important to take privacy into consideration while designing data analysis algorithms.

* Corresponding author.
*E-mail addresses:* yrchen.cs98g@nctu.edu.tw (Y.-R. Chen), rezapour@cs.nctu.edu.tw (A. Rezapour), wgtzeng@cs.nctu.edu.tw (W.-G. Tzeng).

https://doi.org/10.1016/j.ins.2018.03.061
0020-0255/

The goal of this paper is to construct an efficient privacy-preserving ridge regression (PPRR) protocol over high-dimensional data. Each user $u$ can submits its data $\mathbf{x}_u \in \mathbb{R}^d$ and $y_u \in \mathbb{R}$ in an encrypted form to a party called *evaluator* who is responsible to collect and analysis data. The evaluator aggregates the received data and performs regression without knowing data contents. We introduce a Crypto Service Provider (CSP) to initialize the cryptographic parameters of the system and help the evaluator to complete the regression task. CSP is not allowed to learn users' data and the regression model in our system.

**Contributions.** We propose a novel privacy-preserving ridge regression protocol for a large number of distributed data over high-dimensional data. We utilize secure summation method [4,6] and homomorphic encryption schemes [3,10,19,23] for data encryption. We carefully designed an efficient encryption method, which is equipped with homomorphic properties to enable the evaluator to perform regression over the encrypted data. Our PPRR protocol benefits from a number of advantages as follows. Firstly, the encryption cost of a user is asymptotically optimal. It is linear in the data dimension $d$. It also saves bandwidth cost and decreases network latency when numerous users join the system for data submission. Secondly, users can be offline after data submission and they are not required to participate in the subsequent computations. Thirdly, the computation tasks of both the evaluator and the CSP for aggregating users' data are highly parallelable. Their tasks can be computed roughly $k$ times faster on $k$ processors. Fourthly, the evaluator and the CSP need to only engage in one round of communication for exchanging a $d \times d$ data matrix. Lastly, we remove the dependency in the terms of both the number of users and the size of data from the regression computation. This enables our protocol to equip with the ability to deal with massive datasets as in [22].

We implement our protocol to evaluate the computation overhead with realistic settings. In addition, we compare it with the state-of-the-art solutions in terms of both computation and communication costs. The experimental results show the efficiency improvement of our protocol in almost all factors against the existing protocols. For instance, the regression computation time for the evaluator in Nikolaenko et al.'s method [22] is about 1.3 min with $d = 20$, whereas, a similar computation in our protocol takes only 8.8 s. The improvement becomes more substantial as $d$ increases since the computational complexity of regression is $O(d^3)$. The importance of such improvement is immediately apparent in applications dealing with high-dimensional data.

We show that our protocol preserves data privacy assuming the evaluator and CSP are honest-but-curious and non-collusive. Also, we discuss the collusive cases for a subset of users, the evaluator and users, and the CSP and users. They can compromise data privacy of non-collusive users with a negligible probability. Moreover, our protocol can be extended to deal with a malicious evaluator or CSP, who attempts to misbehave for learning users' sensitive information.

**Related works.** The research of privacy-preserving regression has received considerable attention in recent years. Most of the proposed protocols focused on the data which are partitioned either horizontally or vertically across distributed servers [1,7–9,13,16–18,20,25,26,28]. A majority of these protocols employed secure multi-party computation (SMPC) framework to build a linear model over the joint dataset cooperatively. Du et al. [9] defined the S2-MLR (secure 2-party multivariate linear regression) and S2-MC (secure 2-party multivariate classification) problems. They developed a set of basic protocols for matrix computations (e.g., multiplication, inverse, etc.) as building blocks to solve the S2-MLR and S2-MC problems. Sanil et al. [25] further proposed a solution that enables multiple parties to compute the global regression coefficients from the local ones. They combined Powell's method and secure summation technique to enable each user to iteratively update the coefficients using its data.

Some of the approaches are based on secret sharing techniques [17,18]. Karr et al. [17] presented two secure linear regression methods based on secure data integration and secure multi-party computation, while considering different level of data confidentiality. The first method integrates datasets, while ensuring that no party can learn others data. The second one allows each party to share the local statistical information for computing the global regression coefficients using the secure summation protocol. In 2009, Karr et al. [18] used secure matrix product technique to allow multiple parties to estimate the coefficients and standard errors in linear regression. Parties can also verify the regression model without disclosing their private data. Recently, Cock et al. [7] proposed an information-theoretically secure linear regression protocol in the commodity-based model [2], which allows a trusted initializer pre-distributes some random strings. It later correlate the data to parties in the setup phase. Cock et al. [7] proposed a secure matrix multiplication and inversion protocol for the parties to compute the regression coefficients cooperatively. One shortcoming of the aforementioned SMPC based approaches is that they expect the data servers (or parties) to be on-line and participate in the computation throughout the entire process.

Another line of research made use of partially homomorphic encryption (PHE) schemes [16,20,24]. The learning algorithm is performed on ciphertext domain. Some solutions utilize fully homomorphic encryption (FHE) scheme [12] as a building block for encrypting user data and tackle the requirement of two non-colluding parties. However, data privacy is only assured when there is a trusted third party who can be responsible for key generation. Moreover, the approaches based on PHE schemes are usually complicated due to the limitations of the homomorphic property. On the other hand, the current FHE schemes are not efficient enough for large scale applications [21].

Recently, Nikolaenko et al. [22] proposed a hybrid approach for a large distributed dataset among million of users. Each user submits its encrypted data under an additive homomorphic encryption scheme. The evaluator aggregates them and executes a garbled regression circuit $\mathfrak{C}_{\mathrm{CSP}}$ to compute the regression coefficient $\mathbf{w}$. The circuit $\mathfrak{C}_{\mathrm{CSP}}$ is generated by CSP for implementing a linear system solver based on the Cholesky decomposition. Nevertheless, the garbled circuit method [27] is more efficient than the FHE schemes in regression phase. In particular, the size of $\mathfrak{C}_{\mathrm{CSP}}$ does not depend on the number of