A Scientific Research Recommendation System Based on Privacy-Preserving Training Dataset

Shaohua Liu[1,*], Lu Lv[2], Xiaoguang Su[1], Gang Shen[3]

[1]*Department of Management Engineering and Equipment Economics, Naval University of Engineering, Wuhan, China*
[2]*College of Life Sciences, South-Central Minzu University, Wuhan, China*
[3]*School of Computers, Hubei University of Technology, Wuhan, China*

Abstract

Scientific research recommendation system can provide the valuable reference for researchers to choose topics and determine research direction. However, traditional scientific research recommendation obtains the model by training the behaviour dataset of researchers stored in the centre, which may lead to the disclosure of researchers' sensitive information. In this paper, we propose a scientific research recommendation system based on privacy-preserving training dataset. Specifically, we use the federated learning mechanism and threshold homomorphic encryption technology to make the scientific research recommendation model available without uploading the raw dataset, which can protect the privacy of the researchers' behaviour dataset. Additionally, we also use a method to process the dataset of low-quality researchers to improve the accuracy of recommendation model. Through analysis, not only the researchers' privacy can be protected, but also the recommendation model accuracy can be optimized. The experimental results show that the proposed scheme can satisfy the functional requirements of scientific research recommendation system.

Keywords

Scientific research recommendation, Privacy-preserving, Federated learning

1. Introduction

With the massive growth of scientific research information data, scientific research recommendation system will become a right-hand man for researchers to choose their own scientific research interests [1], [2]. Scientific research recommendation system can realize active recommendation by analysing the interactive behaviour of researchers, and provide researchers with more accurate research directions and hot topics according to their research interests. Therefore, an excellent scientific research recommendation system needs to be trained through high-quality dataset. In general, the training dataset of recommendation model comes from a large number of the behaviour dataset of researchers who access the system. Moreover, these behaviour dataset often reflect the researchers' research interest and identity information. If this private information is leaked, it may have a negative impact on the lives of researchers.

The traditional recommendation system centralizes the dataset to a central server for training. In this way, the centralized storage of researcher's behaviour dataset on the server may lead to the risk of disclosure of private information. The reason is that the data information can be easily obtained by malicious third parties or untrusted cloud servers. In order to overcome these problems, many scholars have proposed to use federated learning mechanism to train the model [3-5]. This mechanism means that all data owners train the data locally and upload the trained gradient to the central server so that the raw data is not disclosed. However, there are still some obstacles to using federal learning methods to solve problems. On the one hand, adversary can obtain some researchers' sensitive

ICCEIC2022@3rd International Conference on Computer Engineering and Intelligent Control
EMAIL: *Corresponding author: lshhmail@163.com (Shaohua Liu)

information from the uploaded gradients. On the other hand, there are some unreliable researchers who have low-quality dataset [5], [6]. Since there may be great numerical difference between the gradient trained from low-quality dataset and the ideal gradient, the low-quality dataset affects the accuracy of recommendation model.

To combat that, we propose a scientific research recommendation system based on privacy-preserving training dataset. First of all, we test the local gradient and rule out the unreliable ones. Of course, we will ensure that a certain number of dataset are used to train the recommendation model. The contributions of this article can be summarized as follows:

- First, we propose a scientific research recommendation system based on privacy-preserving training dataset. This scheme uses federal learning mechanism and threshold homomorphic encryption technology to protect the researchers' privacy.
- Second, the proposed scheme can mitigate the negative impact of low-quality data caused by unreliable researchers.
- Finally, we also conduct a large number of experiments to verify that the proposed scheme has better performance in terms of security and efficiency.

The rest of this article is organized as follows. In Section 2, we describe the relevant primitives that this scheme needs to use. We introduce the system model and specific scheme in Section 3 and Section 4, respectively. We present the security and performance analysis in Section 5. Finally, we summarize the proposed scheme.

2. Preliminaries

2.1. (*t*, *n*) threshold Paillier cryptosystem

In the proposed scheme, we use the (*t*, *n*) threshold Paillier cryptosystem [7] to realize the encryption of sensitive information. The advantage of threshold Paillier cryptosystem is that it not only has additive homogeneity, but also has threshold, that is, only those who are equal to or more than a certain number (i.e., *t*) of shares can obtain the decryption key. The cryptosystem includes the following algorithms:

- **Key generation**: Choose two large primes *p*, *q* and calculate $n = pq$, and select a generator $g \in Z^*_{n^{s+1}}$. Then, the public key is $pk = (g, n^s)$, the private key is $s_i = f(i), 1 \le i \le n$.
- **Encryption**: Given a plaintext *m*, use a random $r \in Z^*_{n^{s+1}}$ to calculate the ciphertext $c = g^m r^{n^s} \bmod n^{s+1}$.
- **Share decryption**: Each private key share holder calculates its own share $c_i = c^{2\Delta \cdot s_i} \bmod n^{s+1}$, where $\Delta = n!$.
- **Share combining**: By using the Lagrange interpolation algorithm [7], the ciphertext *c* can be recovered by combining *t* shares of c_i.

The homomorphic property of the above algorithm is as follows:

$$\begin{aligned} c &= E_{pk}(m_i + m_j) = g^{(m_i + m_j)} (r_i r_j)^{n^s} \bmod n^{s+1} \\ &= E_{pk}(m_i) \cdot E_{pk}(m_j) \end{aligned} \tag{1}$$

2.2. Federated learning

Traditional machine learning is to train dataset together, so that it is possible to leak the raw data to adversaries. To combat this privacy issue, Google first proposed a framework for federated learning in 2016, which allows distributed users to train locally without exposing their raw data. Federated learning is technology that uses distributed optimization methods to protect data privacy in multi-party cooperation [8]. It allows multiple clients to cooperate with each other under the coordination of a central server, and a complete machine learning model can be obtained even if the data is scattered among the clients.

Typically, federal learning consists of the following four steps:
1) All clients train on local data independently;
2) The client encrypts the trained gradient and uploads it to central server;
3) The central server aggregates all uploaded gradients securely;
4) The central server sends the global model to each client.

3. System model, threat model and requirements

3.1. System model

As shown in **Figure 1**, the system model of the proposed scheme includes three entities, namely a trusted third party (TTP), a central server (CS) and researchers who provide dataset. Each researcher computes the local gradient by training his/her behaviour dataset locally, and then uploads the gradient to CS. After that, CS aggregates all uploaded gradients to train a global research recommendation model. At the same time, the global model is fed back to each local researcher, and they train the new gradient according to the global. The above iteration does not end until the accuracy of the global model meets certain requirements. The entities in the system model are described as follows:

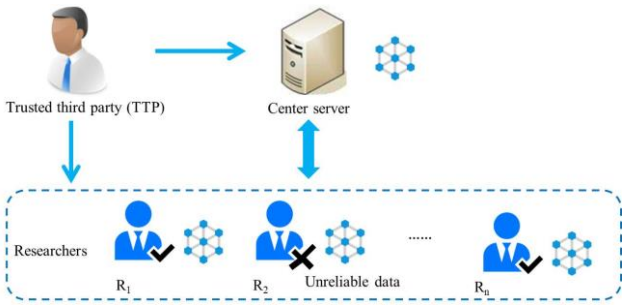

Figure 1 System model

We use cosine similarity to compare the correlation between local gradient and ideal gradient. The initial ideal gradient is a preset initial value.

3.2. Threat model and requirements

In the proposed scheme, the threat model comes from external adversaries and internal adversaries. Central server can become internal adversary if corrupted by adversaries. It is possible to use its convenience to obtain the researchers' behaviour information stored. Based on the given threat model, the requirement of the proposed scheme is to protect the privacy of gradient information provided by researchers, that is, the sensitive information will not be disclosed in the process of gradient transmission and storage. In addition, to improve the accuracy of recommendation system, unreliable participants should be screened before uploading the gradients.

4. The proposed scheme

In this section, we introduce the proposed scheme. First, our scheme considers the problem of unreliable researchers, that is, local gradient generated by *i*th iteration must be compared with the ideal gradient of this round in order to improve the accuracy of the recommendation model. Then, the security of gradient in transmission and storage procedures is also considered. The specific scheme includes the following four phases: system initialization, processing of low-quality dataset researcher, local gradient encryption and generation of recommendation model.

4.1. System initialization

TTA is responsible for initializing the system. Given a security parameter κ, TTA generates a public key pk for all entities and assigns a set of private keys $\{sk_1, sk_2, \cdots, sk_i, \cdots, sk_I\}$ to each researcher R_i. $G^* = \{G_0^*, G_1^*, \cdots, G_i^*, \cdots, G_{I-1}^*\}$ is a global ideal gradient, which is generated by pre-training the scientific research recommendation model. Here, G_i^* denotes the ideal gradient of the ith iteration.

4.2. Processing of low-quality dataset researcher

Researchers with low-quality dataset should be screened before uploading local gradients. Otherwise, they will affect the accuracy of the scientific research recommendation system. Suppose $G^j = \{G_1^j, G_2^j, \cdots, G_i^j, \cdots, G_I^j\}$ is the *j*th researcher's local gradient, and *i* represents the number of iterations. Given the ideal gradient of *i*th iteration G_i^*, each participating researcher compares its own gradient with it, as follows:

$$sim(G_{i-1}^*, G_i^j) = \frac{G_{i-1}^* \cdot G_i^j}{\left\|G_{i-1}^*\right\| \cdot \left\|G_i^j\right\|} \tag{2}$$

where, $sim(\cdot)$ is the cosine similarity algorithm. According to equation (2), the higher the value of $sim(\cdot)$, the higher the reliability of local gradient. When the result of $sim(\cdot)$ is less than a certain value, it indicates that the researcher with the gradient is an unreliable participant. In the proposed scheme, it is assumed that *I* pariticipating researchers are needed to train the global recommendation system. Unreliable researchers are screened and new participants are reselected to ensure that a certain number of researchers come to train global model. The specific process is illustrated in **Algorithm 1**.

4.3. Local gradient encryption

Each reliable participant *j* encrypts his/her gradient as $c_j = Enc_{pk}(G^j)$ with a public key pk and then uploads it to the central server. After receiving all encrypted local gradients, the central server aggregates all encrypted gradients as follows:

$$\begin{aligned} c &= c_1 c_2 ... c_n = Enc_{pk}(G^1) Enc_{pk}(G^2) ... Enc_{pk}(G^n) \\ &= Enc_{pk}(G^1 + G^2 + ... + G^n) \end{aligned}$$

Algorithm 1: Processing of Low-quality Data Researcher

Input:

Global ideal gradient $G^* = \{G_0^*, G_1^*, \cdots, G_i^*, \cdots, G_{I-1}^*\}$,

local gradient $G^j = \{G_1^j, G_2^j, \cdots, G_i^j, \cdots, G_I^j\}$,

threshold *TH*

Output: reliable gradient

1: Initialize global ideal gradient G^*;

2: In *j*th iteration, given G^*;

3: for *i* to *I* **do**

4: **if** $Eqn.(2) \geq TH$ **then**

5: **Return** G_i^j;

6: **end if**

7: end for

4.4. Generation of recommendation model

The recommendation model is derived from the aggregate values of all upload gradients. Therefore, the central server needs t participants to use their private keys sk_j to calculate the secret share $c_j = c^{2\Delta \cdot sk_j} \bmod n^{s+1}$. Then, using t shares of c_j, the plaintext of aggregated gradients can be restored by Lagrange interpolation algorithm [7].

5. Security and performance analysis

In this section, we will discuss the security and performance of the proposed scheme. Additionally, the experiments are based on MNIST database and carried out on an operating system with intel (R) Core (TM) i7-9750H and 8G RAM.

5.1. Security analysis

The security of the proposed scheme focuses on how to protect the privacy of researchers who provide training dataset. Specifically, the participating researcher's gradient is protected.

***Proof*:** In the proposed scheme, the researchers involved in the training do not send their raw data to the central server, but only trained locally. Therefore, adversary will not be able to obtain the raw information of the researchers. In addition, each trained gradient is encrypted by threshold Paillier cryptosystem as $c_j = Enc_{pk}(G^j)$. From Section 4, the decryption key needs to be recovered by at least t participating researchers, so it is very difficult for adversary and central server to obtain the decryption key. Therefore, our scheme can protect researcher's dataset.

5.2. Performance analysis

Here, we mainly discuss the computation cost of encryption and decryption, and the efficiency of low-quality user verification in the proposed scheme.

In local gradient encryption phase, each researcher encrypts his/her gradient as $c_j = Enc_{pk}(G^j)$. And the recovery of global gradient is recovered in the model generation phase.

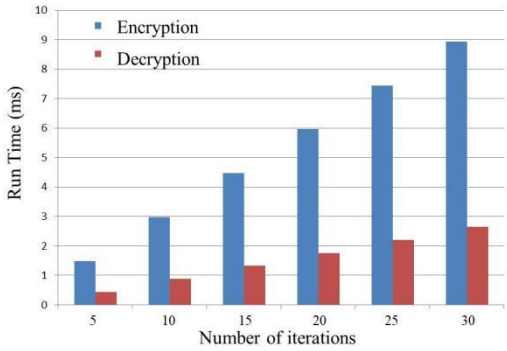

Figure 2 Computation cost of encryption and decryption

Figure 2 shows the computation cost of gradient encryption and decryption with the number of iterations. Next, we discuss the accuracy of the proposed scheme after considering the processing of low-quality dataset researchers. In order to better describe the experiments, we compare the proposed scheme with normal federated learning mechanism (NFM) scheme and filtered but not reselected federal learning mechanism (FBNRF) scheme in terms of accuracy. NFM refers to a federated learning scheme that does not deal with low-quality dataset researchers, and FBNRF denotes that low-quality dataset researchers have been screened but have not been reselected.

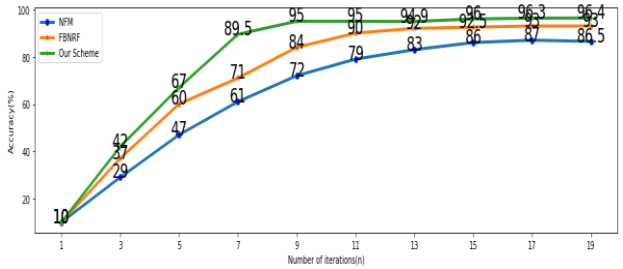

Figure 3 Accuracy

As shown in **Figure 3**, since the ideal global gradient is not obtained, the accuracy in the three scenarios in the initial iteration is about 10%. However, after three iterations, we can see that the accuracy of FBNRF and our scheme is better than that of NFM. In the 9th iterations, the accuracy of the three scenarios is 72%, 84%, 95%, respectively. And after completing the number of iterations, the accuracy of our scheme is higher than that of the other two schemes.

6. Conclusion

In this article, we propose a scientific research recommendation system based on privacy-preserving training dataset, we test the local gradient and rule out the unreliable ones and use the federated learning mechanism and threshold homomorphic encryption technology to make the scientific research recommendation model available without uploading the raw dataset. Security and performance analysis shows that the proposed scheme can meet the security and efficiency of dataset of scientific research recommendation system. In the future, we will study the privacy protection of query users and model parameters in the scientific research recommendation.

7. References

[1] Nishioka C., Hauke J., and Scherp A.: Influence of tweets and diversification on serendipitous research paper recommender systems, Peerj Comput. Sci., vol. 6, pp. e273, 2020.

[2] Zhou X., Liang W., Wang I. K., and Yang L. T.: Deep mining based on hierarchical hybrid networks for heterogeneous big data recommendations, IEEE Trans. on Comput. Soc. Syst., vol. 8, no. 1, pp. 171-178, 2021.

[3] Duan M., Liu D., Chen X., Liu R., Tan Y., and Liang L.: Self-balancing federated learning with global imbalanced data in mobile systems, IEEE Trans. Parallel Distrib. Syst., vol. 32, no. 1, pp. 59-71, 2021.

[4] Fang C., Guo Y., Hu Y., Ma B., Feng L., and Yin A.: Privacy-preserving and communication-efficient federated learning in internet of things, Comput. Secur., vol. 103, pp. 102199, 2021.

[5] Li Y., Li H., Xu G., Huang X., and Lu R.: Efficient privacy-preserving federated learning with unreliable users, IEEE Internet of Things J., vol. 9, no. 13, pp. 11590-11603, 2022.

[6] Hsieh K. et al.: Gaia: Geo-distributed machine learning approaching LAN speeds, in Proc. 14th USENIX Symp. Netw. Syst. Design Implement. (NSDI), 2017, pp. 629-647.

[7] Damgard I., and Jurik M.: A generalization, a simplification and some applications of paillier's probabilistic public-key system, in Proc. Int. Workshop Pract. Theory Public Key Cryptogr., 2001, pp. 119-136.

[8] Lu Y., Huang X., Zhang K., Maharjan S., and Zhang Y.: Blockchain empowered asynchronous federated learning for secure data sharing in internet of vehicles, IEEE Trans, Veh. Technol., vol. 69, no. 4, pp. 4298-4311, 2020.