SCIENCE CHINA
Information Sciences

July 2022, Vol. 65 170307:1–170307:2
https://doi.org/10.1007/s11432-021-3415-1

• **LETTER** •

Special Focus on Cyber Security in the Era of Artificial Intelligence

ACCEL: an efficient and privacy-preserving federated logistic regression scheme over vertically partitioned data

Jiaqi ZHAO[1], Hui ZHU[1*], Fengwei WANG[1], Rongxing LU[2], Hui LI[1],
Zhongmin ZHOU[3] & Haitao WAN[3]

[1] *State Key Laboratory of Integrated Networks Services, Xidian University, Xi'an* 710071*, China;*
[2] *Faculty of Computer Science, University of New Brunswick, Fredericton* E3B 5A3*, Canada;*
[3] *China Mobile (Suzhou) Software Technology Co., Ltd., Suzhou* 215153*, China*

Received 29 August 2021/Revised 27 December 2021/Accepted 24 January 2022/Published online 22 June 2022

Citation Zhao J Q, Zhu H, Wang F W, et al. ACCEL: an efficient and privacy-preserving federated logistic regression scheme over vertically partitioned data. Sci China Inf Sci, 2022, 65(7): 170307, https://doi.org/10.1007/s11432-021-3415-1

Dear editor,

With the age of big data coming, massive data are being generated distributedly all the time and stored as the form of data islands; meanwhile, data privacy and security are strengthened with the introduction of some privacy laws, which thus bring huge challenges to traditional centralized machine learning. Consequently, federated learning (FL) [1], which can construct global machine learning models over multiple participants while keeping their data localized, gains widespread attention and shows its vast prospects in many fields [2]. At each training round of FL, the local updates are calculated locally with participants' training data, which are further aggregated by a server to update the global model until it converges.

Nevertheless, there are still some challenging issues in FL. On the one hand, in most scenarios, the data held by different participants share the common users but differ in features (i.e., vertically partition [3]), which causes difficulties in constructing global models. On the other hand, the uploaded local updates still contain data information and can be used to infer or even recover raw training data, which threatens the users' privacy considerably.

To tackle these challenges, massive FL schemes have been proposed, which are mainly based on homomorphic encryption (HE) or differential privacy (DP). Unfortunately, massive complex calculations of traditional HE will cause unacceptable computational cost, meanwhile, added noises in DP schemes will reduce the model accuracy inevitably. Moreover, it is noteworthy that only a few FL existing schemes [4,5] support vertical model training, and they only support two-party model training and are less efficient.

In this study, we propose an efficient and privacy-preserving federated logistic regression scheme over vertically partitioned data, namely ACCEL. With ACCEL, multiple participants, which have vertically partitioned data, can train a high-accuracy logistic regression model securely and efficiently. Specifically, by combining our proposed data aggregation matrix construction algorithm and a symmetric homomorphic encryption (SHE) [6] technique, local training data and global model can be protected well from inference attacks in the whole training process. Meanwhile, multi-round interactions between the cloud service provider and participants are not required in ACCEL, which reduces the training overhead significantly.

System model. ACCEL consists of two parties, namely the participant and cloud service provider (CSP). CSP is a cloud service provider with powerful storage and computing capability. Each $P_k \in \{P_1, P_2, \ldots, P_K\}$ is a participant with vertically partitioned data and can connect with CSP. Without loss of generality, we assume that P_K is the initiator who has the data labels and obtains the final global model.

Description of ACCEL. The proposed ACCEL mainly contains four phases described in the following. Moreover, some preliminaries are introduced in Appendix A.

- System initialization. In this phase, initiator P_K first selects the security parameters (k_0, k_1, k_2) and generates the public parameter PP and secret key SK of SHE. Then, for each P_k $(k \in [1, K-1])$, P_K generates a ciphertext pair $\{\mathrm{Enc}(0)_0, \mathrm{Enc}(0)_1\}$ by encrypting value 0 with SK, which is sent to corresponding participant P_k. Finally, P_K randomly initializes the global model $\theta^{(0)} = (\theta_0^{(0)}, \theta_1^{(0)}, \ldots, \theta_D^{(0)})$ and the training hyperparameters containing learning rate α, regularization parameter λ, and accuracy parameter κ.

- Data preprocessing and outsourcing. In this phase, local training data $\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(K)}$ are first preprocessed via entity resolution and data normalization, which can be represented as $\mathcal{D}^{(k)} = [X_1^{(k)}, \ldots, X_N^{(k)}]^{\mathrm{T}}$, where $X_n^{(k)} = [x_{n,1}^{(k)}, \ldots, x_{n,D_k}^{(k)}]$ and $X_n^{(K)} = [x_{n,1}^{(K)}, \ldots, x_{n,D_K}^{(K)}, y_n]$. After that,

[*] Corresponding author (email: zhuhui@xidian.edu.cn)

all data values of $\mathcal{D}_k$ are float numbers between 0 and 1, which should first be expanded to an integer through computing $x_{n,d}^{(k)} \leftarrow \lfloor \kappa \cdot x_{n,d}^{(k)} \rfloor$ and $y_n \leftarrow \kappa \cdot y_n$. Then, P_K calculates $\mathcal{V}^{(K)}$ and $\mathcal{M}^{(K)}$ as

$$v_d^{(K)} = \begin{cases} \kappa \cdot \sum_{n=1}^{N} x_{n,d}^{(K)}, \ d = 1, 2, \ldots, D_K, \\ \kappa \cdot \sum_{n=1}^{N} y_n, \ d = D_K + 1, \end{cases}$$

$$m_{i,j}^{(K)} = m_{j,i}^{(K)} = \begin{cases} \sum_{n=1}^{N} x_{n,i}^{(K)} x_{n,j}^{(K)}, \ 1 \leqslant i \leqslant j \leqslant D_K, \\ \sum_{n=1}^{N} x_{n,i}^{(K)} y_n, \ j = D_K + 1. \end{cases}$$

P_k $(k \in [1, K-1])$ calculates $\mathcal{V}^{(k)}$ and $\mathcal{M}^{(k)}$ as

$$\begin{cases} v_d^{(k)} = \kappa \cdot \sum_{n=1}^{N} x_{n,d}^{(k)}, \ d = 1, 2, \ldots, D_k, \\ m_{i,j}^{(k)} = m_{j,i}^{(k)} = \sum_{n=1}^{N} x_{n,i}^{(k)} x_{n,j}^{(k)}, \ 1 \leqslant i \leqslant j \leqslant D_K. \end{cases}$$

Finally, each element x in $\mathcal{V}^{(k)}$, $\mathcal{M}^{(k)}$, and $\mathcal{D}^{(k)}$ is encrypted as $[\![x]\!] = x \oplus (r_0 \odot \mathrm{Enc}(0)_0) \oplus (r_1 \odot \mathrm{Enc}(0)_1)$ and is sent to CSP, where r_0, r_1 are two random numbers, and $\oplus, \odot$ represent the homomorphic addition and multiplication.

- Data aggregation matrix construction. In this phase, CSP first calculates the data aggregation submatrix $[\![\mathcal{M}^{(u,v)}]\!]$ over ciphertexts. When $1 \leqslant u < v \leqslant K-1$, $[\![m_{i,j}^{(u,v)}]\!] \in [\![\mathcal{M}^{(u,v)}]\!]$ is calculated as

$$\bigoplus_{n=1}^{N} [\![x_{n,i}^{(u)}]\!] \odot [\![x_{n,j}^{(v)}]\!], \ i \in [1, D_u], j \in [1, D_v].$$

When $1 \leqslant u < v$ and $v = K$, $[\![m_{i,j}^{(u,K)}]\!] \in [\![\mathcal{M}^{(u,K)}]\!]$ is

$$\begin{cases} \bigoplus_{n=1}^{N} [\![x_{n,i}^{(u)}]\!] \odot [\![x_{n,j}^{(K)}]\!], \ i \in [1, D_u], j \in [1, D_K], \\ \bigoplus_{n=1}^{N} [\![x_{n,i}^{(u)}]\!] \odot [\![y_n]\!], \ i \in [1, D_u], j = D_K + 1. \end{cases}$$

Then, CSP constructs the data aggregation matrix $[\![\mathcal{M}]\!]$ as

$$\begin{bmatrix} NaN & [\![\mathcal{V}^{(1)}]\!] & [\![\mathcal{V}^{(2)}]\!] & \cdots & [\![\mathcal{V}^{(K)}]\!] \\ [\![\mathcal{V}^{(1)}]\!]^{\mathrm{T}} & [\![\mathcal{M}^{(1)}]\!] & [\![\mathcal{M}^{(1,2)}]\!] & \cdots & [\![\mathcal{M}^{(1,K)}]\!] \\ [\![\mathcal{V}^{(2)}]\!]^{\mathrm{T}} & [\![\mathcal{M}^{(1,2)}]\!]^{\mathrm{T}} & [\![\mathcal{M}^{(2)}]\!] & \cdots & [\![\mathcal{M}^{(2,K)}]\!] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ [\![\mathcal{V}^{(K)}]\!]^{\mathrm{T}} & [\![\mathcal{M}^{(1,K)}]\!]^{\mathrm{T}} & [\![\mathcal{M}^{(2,K)}]\!]^{\mathrm{T}} & \cdots & [\![\mathcal{M}^{(K)}]\!] \end{bmatrix},$$

and sends it to P_K for model training and estimating.

- Global model training and estimation. In this phase, P_K decrypts received $[\![\mathcal{M}]\!]$ as $\mathrm{Dec}([\![\mathcal{M}_{i,j}]\!])/\kappa^2$. Then, P_K deletes the last row of $\mathcal{M}$ and fills the first element with N. After that, $\mathcal{M}$ is rewritten as $\mathcal{M} = [A|B]$ and the last row of M is B. Finally, P_K trains the global model iteratively by executing $\theta^{(r+1)} = (1 - 2\lambda\alpha)\theta^{(r)} - \frac{\alpha}{N}(\frac{1}{4}\theta^{(r)}A - \frac{1}{2}B^{\mathrm{T}})$. After a certain number of training rounds, P_K computes the loss as $L(\theta) = \log 2 - \frac{1}{2N}L_B + \frac{1}{8N}L_A + \lambda\,\|\theta\|^2$, where $L_A = \mathrm{SUM}(\theta^{\mathrm{T}}\theta \circ A)$, $L_B = \theta B$, $\circ$ represents the Hadamard product, and SUM is the matrix summation. When $L(\theta)$ is judged to convergence, P_K obtains the final model.

Security analysis. In our threat model, we assume that CSP and all participants are honest-but-curious; i.e., they are obliged to execute the protocol process honestly, but try to deduce the model or data information as much as possible through observing intermediate parameters alone or even collusively. Based on the CPA-secure of SHE [7] and the theory of the linear equations, we demonstrate that our ACCEL can well protect the data and model privacy under the above threat assumption.

Experimental evaluation. We conduct experiments on a workstation with an Intel(R) Xeon(R) Gold 6226R CPU and 256.0 GB RAM. The security parameters are set as $k_0 = 1024$, $k_1 = 20$, and $k_2 = 200$. We first test the model accuracy on three real-world classification datasets and make a comparison with centralized training. The results show that, the prediction accuracy and loss value of ACCEL and centralized training are very close, although the convergence speed of ACCEL is slightly slower. Then, we evaluate the computational cost and communication overhead with generated synthetic datasets, and make a comparison with [4, 5]. The results demonstrate that, ACCEL has a 270× speedup of computational cost (Figure 1) and an up to 23× improvement of communication overhead.

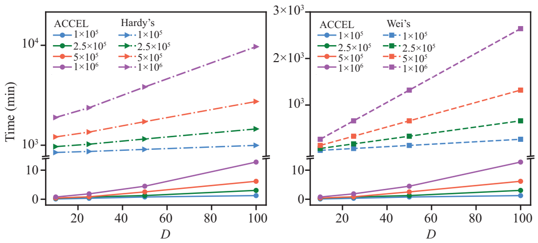

Figure 1 (Color online) Computational cost comparison.

Conclusion. In this study, we have proposed ACCEL, an efficient and privacy-preserving federated logistic regression scheme over vertically partitioned data. Security analysis shows that ACCEL can resist various inference attacks. In addition, the extensive experiments demonstrate that ACCEL has high accuracy for model training and low overhead for both computation and communication.

Acknowledgements This work was supported by National Key R&D Program of China (Grant No. 2021YFB3101300), National Natural Science Foundation of China (Grant Nos. 61972304, 61932015), and Science Foundation of the Ministry of Education (Grant No. MCM20200101).

Supporting information Appendix A. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of Machine Learning Research, Sydney, 2017. 1273–1282

2 Xing J, Tian J D, Jiang Z X. Jupiter: a modern federated learning platform for regional medical care. Sci China Inf Sci, 2021, 64: 202101

3 Yang Q, Liu Y, Chen T J. Federated machine learning: concept and applications. ACM Trans Intell Syst Technol, 2019, 10: 1–19

4 Hardy S, Henecka W, Ivey-Law H, et al. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. 2017. ArXiv:1711.10677

5 Wei Q, Li Q, Zhou Z. Privacy-preserving two-parties logistic regression on vertically partitioned data using asynchronous gradient sharing. Peer-to-Peer Netw Appl, 2021, 14: 1379–1387

6 Mahdikhani H, Lu R X, Zheng Y D. Achieving $O(\log^3 n)$ communication-efficient privacy-preserving range query in fog-based IoT. IEEE Internet Things J, 2020, 7: 5220–5232

7 Zheng Y D, Lu R X, Guan Y G. Efficient and privacy-preserving similarity range query over encrypted time series data. IEEE Trans Depend Secure Comput, 2021. doi: 10.1109/TDSC.2021.3061611