Viewing Federated Learning through the Lens of Small Dataism

Zhang Qian
School of Management, Lanzhou University, Lanzhou Gansu

Abstract: With the increasing awareness of privacy protection, large-scale data collection and sharing has become increasingly difficult, leading to a trend of data "islandization" . To address the problem of joint modeling under data privacy protection, federated learning technology trains a joint model by interacting only with model updates instead of the source data, opening up a new paradigm for data joint modeling. It has made in-depth explorations in aspects such as data heterogeneity, communication efficiency, and model privacy security. However, when viewed from a higher-level perspective of privacy-protected joint modeling, the "immaturity" of federated learning technology will be exposed. This article sorts out the needs and difficulties of joint modeling in real scenarios, analyzes the limitations of federated learning, and introduces small dataism as a guiding ideology to solve the problem of privacy-protected joint modeling. On this basis, it discusses future research directions to help researchers break out of the original framework of federated learning and explore a wider range of technical possibilities.

Keywords: Federated learning; Small dataism; Data privacy protection

I. Privacy-protected joint modeling

Privacy-protected joint modeling is a method of jointly analyzing and modeling data from multiple data holders while ensuring data privacy. Its purpose is to combine the data of multiple data holders without sharing the original data, in order to improve the performance of machine learning and data analysis models, while protecting the privacy of each party 's data. Compared with traditional centralized data mining, privacy-protected joint modeling requires that the original data cannot be shared and that data cannot be aggregated by a third party for centralized mining; compared with distributed data mining, privacy-protected joint modeling emphasizes data privacy, which requires that the original data cannot be shared, and it is necessary to avoid uploading data to leak local sensitive information as much as possible, while distributed data mining only designs the entire framework from the perspective of speeding up mining.[1]

In the context of the era of big data, the demand for joint modeling is extensive, covering various mining tasks such as classification, regression, clustering, and association analysis. At the same time, the relevant laws and regulations on data privacy are constantly being improved, which puts forward various restrictions and norms for the joint modeling process, making it necessary to balance between model usability and privacy security. This has also given birth to many new technologies (such as federated learning), which aim to improve the performance of the mo del under the condition of meeting some data usage specifications.

II. Federated learning

A. Technical connotation

With the advancement of computer computing power, machine learning, as a technology for analyzing and processing massive data, has been widely used in human society. However, the development of machine learning technology faces two major challenges: one is that data security is difficult to guarantee, and the problem of privacy data leakage urgently needs to be solved; the other is that there are data barriers between different industries and departments due to network security isolation and industry privacy, resulting in the formation

of “data islands” and the inability to safely share data. The performance of machine learning models trained independently by each department cannot reach global optimization.

To solve the above problems, Google proposed the federated learning (FL) technology in 2016. It effectively guarantees the privacy and security of users by transferring the data storage and model training phases of machine learning to local users and only interacting with the central server with model updates. Federated learning allows users to protect their privacy in the process of machine learning and can train machine learning models without aggregating source data into training data sharing. Federated learning is essentially a distributed machine learning technology, and its process is shown in Figure 1.

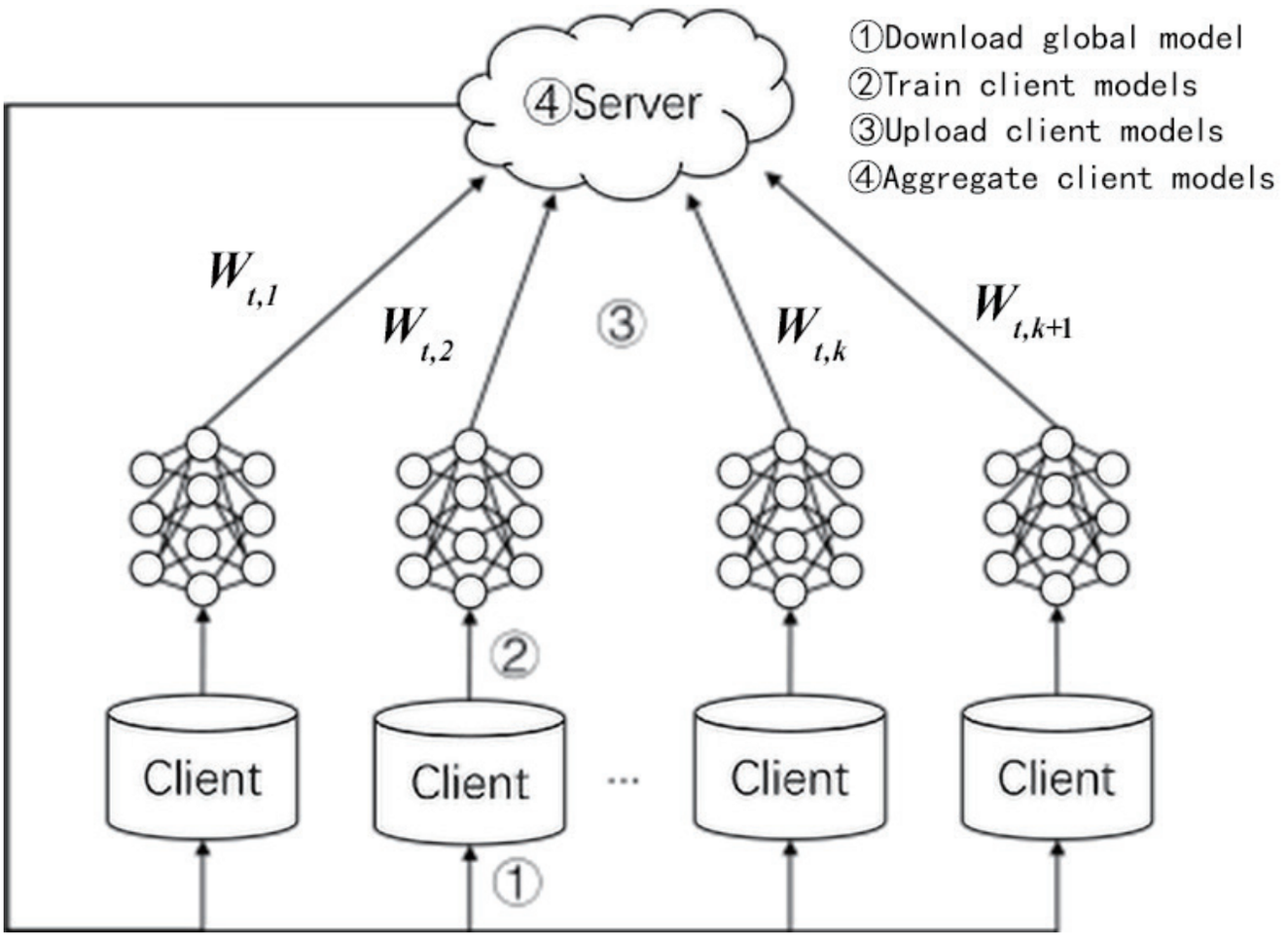

Fig1 Federated Learning Process

Clients (such as tablets, mobile phones, IoT devices) train local models on local data under the coordination of the central server (such as service providers). The central server is responsible for aggregating local models to obtain the global model. After multiple rounds of iterations, the final model obtained is similar to the result of centralized machine learning, effectively reducing many privacy risks caused by traditional machine learning source data aggregation.

The iterative process of federated learning is as follows.

1. The client downloads the global model from the server.

2. The client trains the local data to obtain the local model (the local model update of the ith client in the rth round of communication).

3. The various clients upload the local model update to the central server.

4. The central server receives the data from all parties and performs weighted aggregation operations to obtain the global model (the global model update in the rth round of communication).

In summary, federated learning technology has the following characteristics.

1. The original data participating in federated learning are all retained on the local client, and the interaction with the central server is only model update information.

2. The model trained jointly by the parties participating in federated learning will be shared by all parties.

3. The accuracy of the final model of federated learning is similar to that of centralized machine learning.

4. The higher the quality of the training data of the parties participating in federated learning, the higher the accuracy of the global model.

Federated learning protects users'sensitive data by interacting with model updates (such as gradient information) instead of the original data being stored locally, opening up a new paradigm for data security. As a classic solution for joint modeling, it has received in-depth research.[3] In terms of scenarios, horizontal federated learning and vertical federated learning are proposed for horizontal and vertical divisions of data. In terms of technology, extensive explorations and corresponding solutions have been proposed for aspects such as the heterogeneity of client data, communication efficiency between clients and servers, and privacy protection of model parameters.

B. Technical limitations

Although the framework of federated learning is becoming more and more mature, it is still possible to analyze many immature aspects of federated learning from the perspective of joint data modeling.

1. Specialization for classification problems

The framework of federated learning has already supported various mining models. The FATE system, as an industrial-level federated learning framework, has built-in secure joint modeling solutions for linear models, tree models, and neural networks. However, linear models, tree models, and neural networks are all technical approaches for classification/regression problems. The entire research field of federated learning is also concentrated on solving various difficulties in federated classification tasks, and there is little research on other mining tasks. In actual application scenarios, clustering is often the first step in data analysis. Especially in the case of unlabeled data, data analysts hope to discover the inherent cluster structure of data through clustering technology and assign the same pseudo-label to similar samples, and then assign specific meanings to pseudo-labels by experts to form labeled data, supporting subsequent classification tasks. When data is distributed across multiple terminals, each terminal has multiple data but does not meet the local cluster formation, and hopes to discover the cluster affiliation of local data in the global distribution in a privacy-protected manner. At this time, federated clustering becomes an important technical point, but in the existing research field, there is little attention to such problems, and the federated learning framework does not consider the needs of federated clustering.

2. Specialization for deep learning technology

The essence of the success of federated learning can be attributed to the federated average algorithm, which achieves the aggregation of local knowledge by weighting and averaging the neural network parameters trained locally, generating a global model. The aggregation effect of the federated average algorithm is theoretically consistent with the effect of centralized training, and the simple and efficient aggregation algorithm has become the foundation of the federated learning framework. The technical points of joint modeling problems are two, one is how to express the knowledge of the local model, and the other is how to aggregate the knowledge of different local models. The expression of local model knowledge is a field with high degrees of freedom. Model parameters, synthetic data, and summary structures can all represent a certain knowledge contained in the local data, and the key is whether this knowledge expression is convenient for aggregation. The second problem is to design a knowledge aggregation method based on the solution to the first problem, and aggregate the local knowledge into

global knowledge. Federated learning uses deep learning technology to express local knowledge and designs the federated average algorithm to make local knowledge easy to aggregate. However, the convenience of federated learning aggregation is based on the complexity of local knowledge expression. The parameters of neural networks range from millions to hundreds of millions, and the high-dimensional vectors interacted between local and servers bring additional communication overhead and privacy protection difficulties to the entire joint modeling process. The high dependence on deep learning technology is the reason for the current success of federated learning, but it is also an obstacle to the further promotion of federated learning.

3. Specialization for Model Parameter Averaging

Deep learning model parameters can be regarded as a high-dimensional vector containing rich local data knowledge. However, the server in the federated learning framework only utilizes the classification information in the model parameters, which is an underutilization of local information.

Even if deep learning is used to build a local classifier, model parameters are not the only expression of the classification information of the dataset. From the perspective of knowledge utilization rate, model parameters are not even an excellent carrier of classification knowledge. The core reason why the federated learning framework chooses model parameters as the carrier of classification knowledge is still the convenience of knowledge aggregation. In 2022, researchers were inspired by the dataset distillation technology and introduced it into federated learning. By building a small synthetic dataset based on the local dataset, the classification performance of the synthetic dataset and the local dataset trained on the same classification network is similar, and then the synthetic dataset is uploaded to the server to support the construction of the global classifier. This idea, overall, does not deviate from the federated learning framework, but changes the form of the carrier of local classification knowledge. It uses a small synthetic dataset to carry the classification information of the original dataset, and the synthetic data has the same format as the original data. The global model can be directly trained on the synthetic data uploaded by each terminal, capture the characteristics of global data, and build a global classifier, still achieving convenient local knowledge aggregation. This technical route once again points out the rigidity of the current federated learning field, where few researchers have tried to break through the original framework.

III. Small Dataism and the Future Joint Modeling Architecture

The rapid development of artificial intelligence is rooted in the exploitation of a large amount of ready-made human knowledge through the internet. However, the “accessibility of big data” is not the norm in human society. On the contrary, a visible trend is that the large-scale collection of data is becoming increasingly difficult under the constraints of laws and regulations. The data gifts brought by the internet are fragile, and the social and economic benefits brought by them also have fundamental fragility. In this context, considering other development paths is not a bad compensation scheme.

Small dataism originates from the “principle of frugality” proposed by Gerd Gigerenzer, which can be intuitively understood as a kind of lazy wisdom that tries to infer answers from the least but most relevant information. Big data technology attempts to become an all-knowing and omnipotent god, mastering all information flows and summarizing the causes and effects of things to achieve prediction and judgment of things; while small dataism originates from the observation and simulation of human intelligence. Humans do not have the powerful computing power and memory ability of computers, but still have higher intelligence. The human brain has a series of mechanisms to help reduce the consumption of mental activities. It does not pursue all-knowing and omnipotent, but simple and efficient practicality, hoping to grasp the most knowledge at the lowest cost.

The current artificial intelligence algorithms represented by deep learning are like a “Godzilla” monster that needs to swallow a large amount of data to be satisfied. Its greed for data has a kind of technical spontaneity, and it has no respect for the commonly accepted value standards of “privacy” “civil rights” and “human free choice” . This makes the privacy data of citizens in modern society exposed to danger at all times. Based on the above criticism of deep learning technology, this article proposes a joint modeling architecture based on the idea of “small dataism” . In this architecture, the local terminal no longer uploads model parameters, but a highly refined specific knowledge expression, and then the central server implements knowledge aggregation and mining. The expression of local knowledge is free, and it is as concise as possible while carrying the specific information for the task; the strategy of knowledge aggregation is free, and simple knowledge can be aggregated in one round, while complex optimization processes can be aggregated in multiple rounds; the task is free, and classification, regression, clustering, and association analysis can all be included. The joint modeling under small dataism has a natural policy affinity with privacy protection laws and regulations. The reduction of uploaded data dimensions will greatly reduce the additional costs of privacy protection; it has greater inclusiveness for the mining needs of joint modeling, and classification models, clustering models, and association analysis models can all be included in this architecture; it has broader coverage of technical approaches, not limited to neural networks, and corresponding knowledge aggregation strategies can be developed for the knowledge expression methods of various mining algorithms, providing more flexible design space for joint modeling itself (Figure 2).

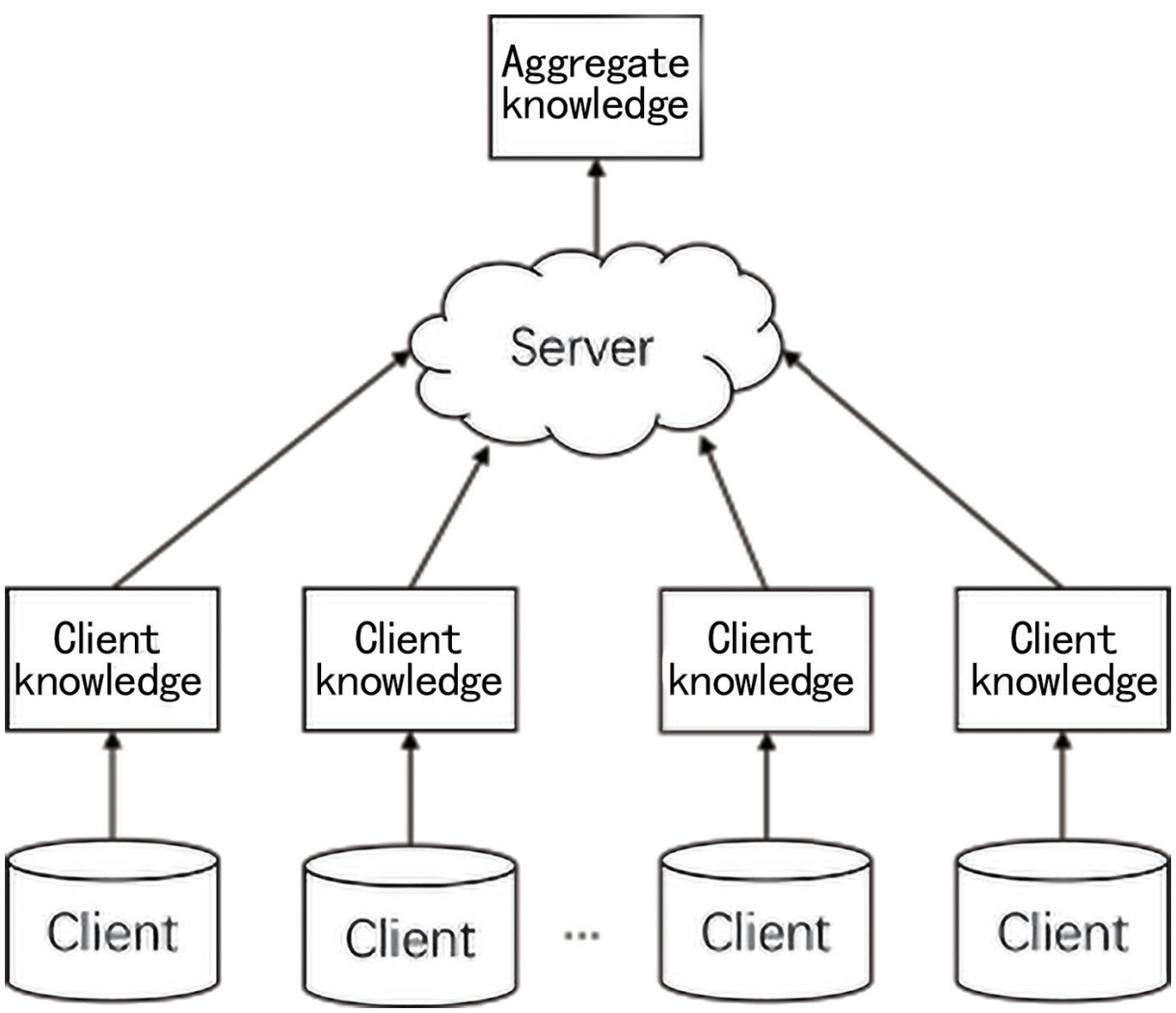

Fig.2: Joint Modeling under Small Dataism

IV. Conclusion

From a higher perspective, we are able to identify the limitations of the federated learning framework and, based on the concept of small dataism, briefly analyze a potential future architecture of the federated learning framework. This provides a general technical path that can be explored for the task of privacy-protected joint modeling.

References

[1] Zhou Chuanxin, Sun Yi, Wang Degang, et al. A Survey on Federated Learning[J]. Journal of Network and Information Security, 2021, 7(5): 77-92.

[2] Zhou Chuanxin. Research on Key Technologies of Data Security Sharing for Federated Learning[D]. Zhengzhou: Information Engineering University of Strategic Support Force, 2023.

[3] Chen Lei, Feng Xiaoqing. A Federated Learning Framework Based on Consortium Chain for Big Data Credit[J]. Credit, 2023, 41(7): 36-42.

[4] Xu Yingjin. Fifteen Lectures on the Philosophy of Artificial Intelligence[M]. Beijing: Peking University Press, 2021.