A Survey of Federated Learning on Non-IID Data

HAN Xuming[1], GAO Minghan[2], WANG Limin[3], HE Zaobo[1], WANG Yanze[1]

(1. Jinan University, Guangzhou 510632, China；
2. Changchun University of Technology, Changchun 130012, China；
3. Guangdong University of Finance & Economics, Guangzhou 510320, China)

DOI: 10.12142/ZTECOM.202203003

https://kns.cnki.net/kcms/detail/34.1294.TN.20220818.0945.002.html, published online August 18, 2022

Manuscript received: 2022-06-10

Abstract: Federated learning (FL) is a machine learning paradigm for data silos and privacy protection, which aims to organize multiple clients for training global machine learning models without exposing data to all parties. However, when dealing with non-independently identically distributed (non-IID) client data, FL cannot obtain more satisfactory results than centrally trained machine learning and even fails to match the accuracy of the local model obtained by client training alone. To analyze and address the above issues, we survey the state-of-the-art methods in the literature related to FL on non-IID data. On this basis, a motivation-based taxonomy, which classifies these methods into two categories, including heterogeneity reducing strategies and adaptability enhancing strategies, is proposed. Moreover, the core ideas and main challenges of these methods are analyzed. Finally, we envision several promising research directions that have not been thoroughly studied, in hope of promoting research in related fields to a certain extent.

Keywords: data heterogeneity; federated learning; non-IID data

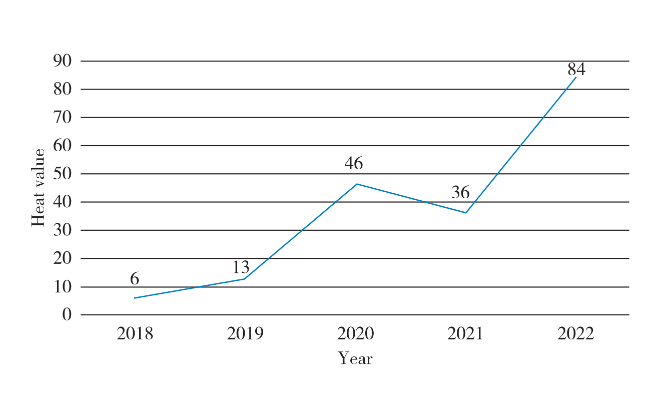
Citation (IEEE Format): X. M. Han, M. H. Gao, L. M. Wang, et al., "A survey of federated learning on non-IID data," *ZTE Communications*, vol. 20, no. 3, pp. 17–26, Sept. 2022. doi: 10.12142/ZTECOM.202203003.

1 Introduction

The potent ability of machine learning methods[1] comes from learning the representation and internal laws of a large number of sample data. The extensive use of edge devices makes data collection less expensive. However, the sample data collected by edge devices are often scattered and small-scale in the real world. Hence, it is not easy to train a practical machine learning model solely on edge devices' data. Centralizing edge data for training also becomes more challenging with the gradual improvement of laws and regulations related to data security, and therefore federated learning (FL) comes into being.

In FL, a central server can unite different clients (such as edge devices and an entire organization) to cooperatively train a global model that performs well on most clients while preserving privacy. FL has attracted much attention from researchers in recent years due to its excellent characteristics and is widely used in various fields, including mobile edge devices[2], the Internet of Things (IoT)[3], and medical collaboration[4]. Fig. 1 shows the heat value of FL in Google search in the past five years[5]. The higher the heat value, the more interested people are in federated learning in the current year.

Although FL solves the problem of cooperative learning with small data under privacy constraints, it still faces some challenges. Due to the geographic distribution and usage patterns of edge devices, the data in the edge devices tend to be skewed to varying degrees (including label skew, feature skew, volume skew, time skew, and hybrid skew), also known as data heterogeneity. The data heterogeneity challenges the data's independent and identically distributed (IID) assumption. The existence of data heterogeneity challenges the assumption of IID data, adding complexity to problem modeling,



▲Figure 1. Heat value of federated learning in Google search[5] theoretical analysis, and empirical evaluation of solutions. As a result, the global model becomes difficult to adapt to individual clients. On the non-independent and identically distributed (non-IID) data, the FL of a single global model, FedAvg[6], proved ineffective by experiments. Hence, a survey of improved methods is necessary for researchers to further analyze and solve FL problems on non-IID data.

Existing surveys on FL on non-IID data[7–8] focus on the action position (such as data, models, and architecture) of processing non-IID methods and cannot show the purpose and motivation of the improved methods. To remedy this regret, this survey investigates many improved methods that mitigate the impact of non-IID data on FL and provides a new perspective on FL methods for analyzing non-IID data. These methods are categorized into heterogeneity reducing strategies and adaptability enhancing strategies from the perspective of core motivations. On this basis, a detailed classification is carried out respectively. The main contributions of this survey are summarized as follows:

1) This survey provides a brief overview of FL concepts, methods, and challenges posed by the non-IID data setting.

2) This survey proposes a unique perspective based on core motivations. According to the core motivations of a method, existing state-of-the-art methods are classified into heterogeneity reducing strategies and adaptability enhancing strategies. On this basis, many FL methods for non-IID data are reviewed, and their basic ideas and main challenges are analyzed.

3) We look forward to future research trajectories in some related fields on non-IID data.

The rest of the article is organized as follows. Section 2 provides an overview of FL and its non-IID data setting. Section 3 presents our induction of a unique taxonomy based on core motivations. Section 4 analyzes the ideas and main challenges of heterogeneity reducing strategies. Section 5 analyzes the ideas and main challenges of adaptability enhancing strategies. In Section 6, we look forward to future research directions in FL on non-IID data. Finally, we summarize the work of this survey.

2 Preliminary Knowledge

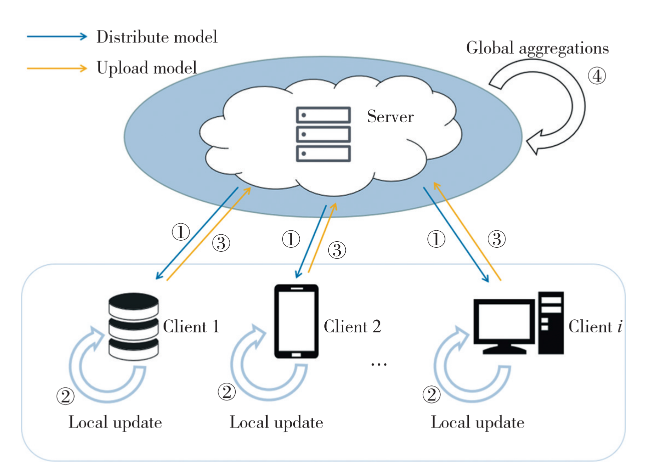
In this section, we provide an overview of FL and non-IID data settings of FL for understanding the problem of FL on non-IID data.

2.1 Federated Learning

FL[6] aims to get a single globally optimal model from data across thousands of clients to minimize the training loss for each client. The data and training processes of all clients must remain local to meet participants' needs for privacy. Fig. 2 shows the architecture of FL. The clients participating in the training are different types of devices with different hardware and software characteristics, and each client maintains a local model. In each training, the server distributes the initial model to the clients in the training. The clients update their model parameters by utilizing their local data, and upload their model parameters to the server for aggregation, thereby completing the updating of the current round of the global model. Finally, the server uses the updated global model as a new initial model to participate in the next training. The global model training objective function can be formulated as:

$$\Theta^{*} = \underset{\Theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} F_{i}(\theta_{i}, x_{ij}, y_{ij}), \tag{1}$$

where Θ^{*} is the global model parameter, θ_i is the local model parameter of the i-th client, n is the number of all clients, x is the data feature, y is the data label, and F_i is the empirical risk of the i-th client data.



① A server distributes the current global model to clients.
② The clients update their received model (e.g., stochastic gradient descent) with their own data.
③ The clients send model parameter updates (gradients) to the server.
④ The server aggregates the received gradients and updates the global model.

▲**Figure 2. Architecture of federated learning (FL) in each training**

2.2 Non-IID Data Setting

FL requires that the client data involved in the training satisfy the IID assumptions. However, the collection of data in the real world often depends on the usage of a specific device. There is a certain degree of heterogeneity in the distribution and quantity of data between clients, also known as skew [8–9]. Depending on the skew situation, we categorize it as follows in this survey.

•Label skew. The label distribution of data between clients is different. Since each client may rate the same feature differently, records with the same characteristics may have different labels. For example, many people love furry pets, but people allergic to animal hair may not think so.

•Feature skew. The features of data between clients do not overlap or partially overlap. Due to differences in viewing angle and modality, records with the same label may have very different characteristics or even be completely different. For example, when two cameras at different positions capture the same object, the description (front view and left view) of the object's features may be quite different.

•Quantity skew. The quantity of client data varies. Due to differences in computing and storage capabilities of client devices, the numbers of data that devices can use for federated training may be different. For example, there will be hundreds of fold differences in the frequency of temperature measurements and the number of temperature data stored between home and industrial electronic thermometers, which may lead to a preference for clients with larger data sets.

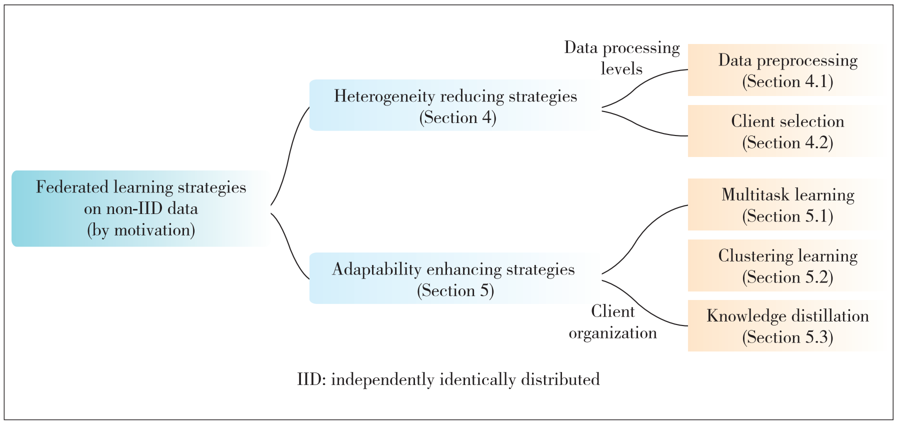
•Time skew. The distribution of client data is time-dependent. The data collected by the device may vary by day, night, or season. For example, the usage and driving characteristics of shared bicycles may be significantly different in the morning and the evening, and the transmission characteristics of COVID-19 may also be significantly different in summer and winter.

•Hybrid skew. Client data have two or more skews of the above.

Client data take on the characteristics of non-IID by the different types of skew mentioned above. Due to the difference in the distribution of clients, the convergence direction of a small number of clients may deviate from most of the other clients when FL is trained on non-IID data. This is known as client drift[10], which is an essential factor that impairs the effect of FL.

3 Federated Learning Strategies on Non-IID Data

This survey provides a comprehensive examination of the FL on non-IID data in recent years. On this basis, it classifies existing FL strategies on non-IID data from the perspective of motivation, mainly including heterogeneity reducing strategies (Section 4) and adaptability enhancing strategies (Section 5). Then, the specific methods of the two strategies are further subdivided according to the data processing level and client organization. Our proposed taxonomy is shown in Fig. 3, which is the basis for a comprehensive review and systematic analysis of existing methods. Fig. 4 shows the setup of two basic strategies, namely heterogeneity reducing strategies, which perform preprocessing before the client participates in federated training so that the data participating in federated training is close to the IID data, and adaptability enhancing strategies, which use various means to obtain a personalized model to enhance the model's adaptability to non-IID data when the client performs (or completes) federated training. This section will describe both strategies in detail.



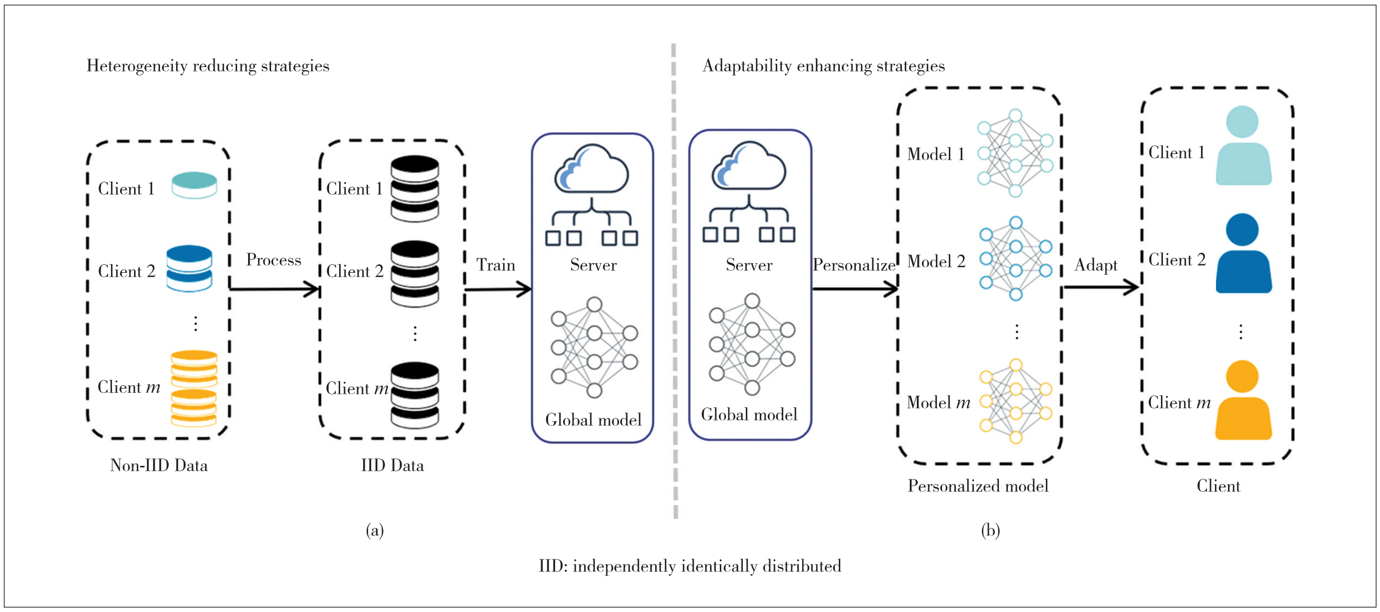
▲ **Figure 3. Taxonomy of federated learning (FL) strategies on non-IID data proposed in this survey**

3.1 Heterogeneity Reducing Strategies

The heterogeneity reducing strategies aim to keep the data involved in the federated training close to the IID data, as shown in Fig. 4(a). Since the data distribution between clients may differ, the server may experience client drift while organizing clients to train a single global model collaboratively. The client drift makes it difficult for the global model on the server to achieve the desired training effect. A natural idea for this challenge is that the server can perform federated training on data approaching the IID by reducing the heterogeneity between client data. Reduction in heterogeneity simplifies the difficulty of server model aggregation and reduces the risk that the model cannot converge. The primary means of heterogeneity reducing strategies summarized in this survey include data preprocessing and client selection. The methods of data preprocessing aim to directly or indirectly convert the non-IID data of the client participating in the training into IID data before the client uses the respective data to participate in the federated training. The methods of client selection aim to select the subset of clients with the slightest degree of data skew to participate in federated training.

3.2 Adaptability Enhancing Strategies

The adaptability enhancing strategies aim to learn personalized models for clients with different distributions, as shown in Fig. 4(b). The heterogeneity reducing strategies can effectively prevent the adverse effects of non-IID data in FL. However, even a high-quality global model may lose some of the client's private information. The loss of client information causes the model to degrade on specific clients. Therefore, some scholars have proposed an FL method with more robust client adaptability. Unlike the heterogeneity reducing strategies summarized in this survey, the server no longer trains a single global model in the adaptability enhancing strategies. Multiple refined models are adapted to clients with different data distri-



▲ **Figure 4. Heterogeneity reducing strategies and adaptability enhancing strategies: (a) Heterogeneity reducing strategies and (b) adaptability enhancing strategies**

butions by personalizing the global model aggregated by the server. The primary means of adaptability enhancing strategies summarized in this survey include federated multitask learning, federated clustering learning, and federated knowledge distillation. Federated multitask learning aims to find related subtasks in FL and use domain-specific knowledge to train similar models for them. Federated clustering learning aims to cluster clients with similar distributions into a class on client data with inherent partitions and train a cluster model to adapt to its inherent partitions. Federated knowledge distillation aims to transfer knowledge between the server and client models (or only between client models), improving its performance on unknown heterogeneous data.

4 Heterogeneity Reducing Strategies

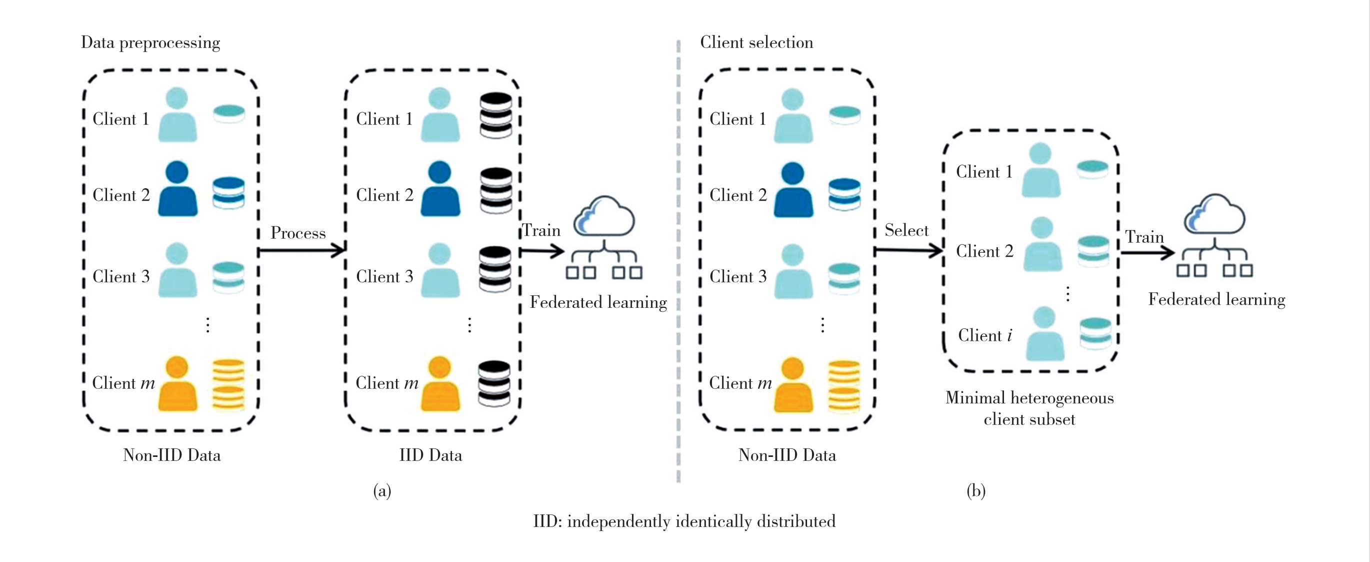
This section will introduce FL methods on non-IID motivated by reducing heterogeneity. The primary setting of these methods is shown in Fig. 5. This survey classifies them into data preprocessing methods and client selection methods according to the different levels of data processed by the server, where the data preprocessing method preprocesses the client's data before joining the training and converts the non-IID data into IID data, and client selection selects the client subset with the most negligible heterogeneity to join the training. Table 1 shows the advantages and disadvantages of these methods.

4.1 Data Preprocessing

Classical data preprocessing methods include oversampling[11] and undersampling[12]. Before performing machine learning training, increasing or decreasing the number of training times for specific types of samples in the dataset can effectively reduce data heterogeneity. This survey classifies the data preprocessing methods in FL into direct preprocessing and indirect preprocessing.

In the direct preprocessing methods, the server directly alters the data involved in training so that training is done on data close to the IID. TUOR et al.[13] proposed federated learning based on data correlation. They set up a small task-specific benchmark data set on the server and trained a benchmark model against it. The benchmark model was used to determine the relevance of local data on the clients and filter out data samples irrelevant to the learning task. Each client used only its selected subset of relevant data during FL. However, since the distribution of participating training clients in the federated environment is unknown, the setting of the benchmark dataset faces incredible difficulties. The difference between the benchmark dataset distribution and the real data set will be an essential factor in determining the model's performance. YOSHIDA et al.[14] proposed a hybrid FL. Clients allowed their data to be uploaded to the server to build an approximate IID dataset. The server then updated the global model with IID data and aggregated it with other locally updated models from clients. Such methods are relatively easy to implement and do not require a preset benchmark dataset. However, building IID datasets directly from the server raises data privacy concerns when a trusted central server is not guaranteed. YOON et al.[15] proposed mean augmented federated learning (MAFL), an average enhanced FL framework, which exchanged model parameters and additional data generated by the mix. Based on MAFL, FedMix is proposed to approximate the loss function of global mixing by Taylor expansion. This approximation involved only the average data from other clients, with some privacy to the server.

In the indirect preprocessing methods, the server designs the



▲Figure 5. Method settings for heterogeneity reducing strategies: (a) data preprocessing and (b) client selection

▼Table 1. Summary of specific methods based on heterogeneity reducing strategies

Methods	Ways	Advantages	Disadvantages
Data preprocessing	Direct	• Easy to implement	• May reveal privacy • Proxy dataset required
	Indirect	• Strong privacy	• Contextual information may be required • More complex to implement
Client selection	Context-based	• Faster model converges	• May reveal privacy
	Deep-learning-based	• No context required • Better effect	• Higher time and space costs

encoding method to obtain the encrypted data distribution indirectly and thus balance the data distribution. DUAN et al.[16] proposed a self-balancing FL framework named Astraea. Before training, the server classified the majority and minority classes based on the Z-score outlier detection algorithm and then performed data preprocessing to adapt to the classes. In training, the mediator of asynchronously receiving and applying client updates was proposed to average local imbalance. The mediator made the distribution of data collection close to unity by rearranging the clients to participate in the training. However, this algorithm requires the server to have more understanding of the context information of the client. WU et al. designed a generative convolutional autoencoder (GCAE) in Ref. [17]. By synthesizing minority class samples through GCAE, a class-balanced dataset was generated to retrain the client's local model. The process alleviated the non-IID of clients' data and achieved better-personalized prediction. Furthermore, because GCAE contains only a small number of model parameters, it can significantly reduce the communication overhead during model transfer.

4.2 Client Selection

Such work designs client-level data distribution balancing methods. Because the server does not need to preprocess the data directly, the methods avoid the privacy problems caused by directly processing the client data. This survey classifies them into context-based methods and deep-learning-based methods according to the difference in server selection of clients.

The context-based methods focus on utilizing available environmental information in FL. ZHAO et al.[18] proposed an enhanced FL method, Newt, which selects participating clients in heterogeneous FL. On the one hand, under the joint consideration of the client dataset and weight update size, the server selected the available clients in a specific FL task by setting selectors to explore the trade-off between accuracy performance and system progress for each round. On the other hand, the frequency of client selection was taken as an additional dimension to optimize the client selection algorithm. This allows the server to maintain fundamental fairness in its biased selection of clients. SHU et al.[19] proposed a computation and communication efficient federated learning via adaptive sampling of data and clients, called FLAS. The server captured data distribution among different clients and set adaptive thresholds during the learning process to improve local computing efficiency and accelerate client convergence. In addition, the server selected clients with the same convergence phase to reduce the communication cost between the client and the server. The context-based methods require the server to have a priori knowledge of the client data distribution, limiting the application of FL in environments with strict information constraints.

The deep-learning-based methods bring practical experience from deep learning to FL and use online learning to perform client selection. ZHANG et al.[20] designed an experience-driven FL method based on deep reinforcement learning. The server mitigated the negative impact of non-IID data by selecting a subset of participants and adaptively adjusting their batch size. This method can adaptively determine system parameters without knowing any prior information to control local model training and global aggregation and maximize the model accuracy of each round of communication. WANG et al.[21] proposed an experience-driven FL framework, FAVOR. An agent with dual deep Q-learning network (DDQN) training was designed to perform active client selection to obtain the optimal client terminal set. The agent offset bias was introduced by non-IID data and speeded up the FL process. One of the advantages of Q-learning is comparing the expected utility of available actions without prior environmental information. Therefore, this method can train and reuse data more effectively than the context-based methods in federated environments with strict information constraints. However, the deep-learning-based methods tend to have high costs in time and space, which imposes higher requirements on the performance of federated networks.

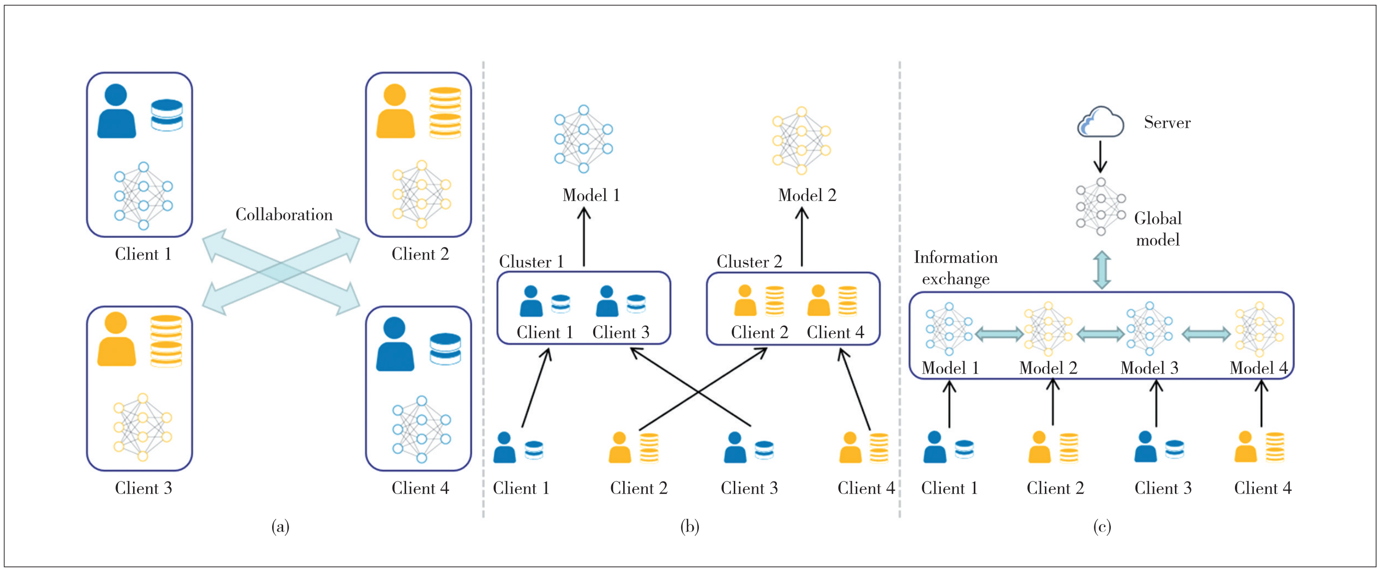
5 Adaptability Enhancing Strategies

This section will investigate FL methods on non-IID motivated by enhancing adaptability. The settings of the different methods are shown in Fig. 6. This survey classifies FL tasks into federated multitasking learning, federated clustering learning, and federated knowledge distillation based on how they are organized between the server and the client. Federated multitask learning finds relevant task clients for knowledge exchange and collaborative training in the training process. Federated clustering learning classifies clients with similar data distribution as clusters and learns the cluster model according to clusters. Federated knowledge distillation exchanges the knowledge in their models between the server and the client (or between the clients), which makes the model output close to each other. Table 2 shows the advantages and disadvantages of these methods.

5.1 Federated Multitask Learning

Multitask learning exploits the similarity between tasks while solving multiple potentially related tasks[22]. These tasks are somewhat related but not identical. By introducing multitask learning, each client learns knowledge from all relevant tasks, which facilitates the training of more adaptive client models. This survey divides federated multitask learning on non-IID into client-based methods, and subtask-division-based methods based on how multitasks are set up.

The client-based methods regard clients with different data distribution as different tasks. The server constructs an association matrix between the clients to organize the relevant client to participate in the collaboration. SMITH et al. [23] introduced multitask learning into FL and proposed a novel systems-aware optimization method, MOCHA. This algorithm extended primal-dual optimization into a federated multitasking setting and defined a data-local subproblem to separate computation across clients. MOCHA learned a personalized model for each client during joint optimization of multiple subproblems. However, this algorithm can only be applied to convex target problems, and all clients must be guaranteed to participate in each round of training. HUANG et al.[24] introduced the attention message passing mechanism into FL and proposed FedAMP. This algorithm implemented an attention mechanism by calculating the similarity between client model weights, iteratively encouraging more cooperation between similar clients. Based on this, a personalized cloud model was maintained for each client using a messaging mechanism. Cli-



▲Figure 6. Method settings for adaptability enhancing strategies: (a) federated multitask learning, (b) federated clustering learning, and (c) federated knowledge distillation

Table 2. Summary of specific methods based on adaptability enhancing strategies

Methods	Ways	Advantages	Disadvantages
Federated multitask learning	Client-based	• Easy to implement	• Possibility to isolate heterogeneous clients
	Subtask-division-based	• Part-time joins allowed	• Data quality sensitive
Federated clustering learning	Model-loss-based	• Easy to implement • Predictable effect	• Need to preset the number of clusters • Communication overhead is high
	Client-similarity-based	• No need to preset the number of clusters	• Lack of theoretical analysis
Federated knowledge distillation	One-way distillation	• Strong privacy	• Poor to heterogeneity model • Contextual information may be required
	Mutual distillation	• Robust to heterogeneous models • Suitable for a large number of clients	• Negative transfer possible • Lack of theoretical analysis

ents with similar models could achieve closer cooperation and improved cooperation efficiency through this positive feedback mechanism. However, such an algorithm may actively segregate clients with different distributions when the data are highly non-IID. These segregated clients will be more likely to converge to the local optimum. JAMALI-RAD et al. [25] proposed an FL algorithm with Taskonomy (FLT). Unlike FedAMP, this algorithm initialized an encoder with an on-server standard dataset and sent it to each client. The server received the latent representation obtained by the client compressed by the encoder and generated a task association matrix accordingly. On this basis, the personalization model was trained using the associations between clients. Since this algorithm only needs to pass the encoder to the client at one time, higher operating efficiency is obtained.

The subtask-division-based methods divide the FL task into multiple sub-tasks and perform multitask learning in a federated setting by organizing the subtasks. LI et al.[26] proposed Ditto, a federated multitask learning framework. Ditto added a regularization term to the original objective function of the client model. This algorithm used an objective function with added regularization terms for local training, while the original objective function was used for global training. This algorithm achieved a trade-off between robustness and fairness by adjusting a hyperparameter λ under the condition of personalization. Ditto achieves promising results on both convex and non-convex targets. MARFOQ et al.[27] designed a federated multitask model FedEM based on mixed data distribution. This algorithm assumed that the data distribution for each client was a mixture of M underlying distributions, with different distributions as subtasks. Clients used only the points sampled from the mixture distribution to construct an unbiased estimate of the true risk on each subtask and jointly learned shared component models and personalized hybrid weights through EM-like algorithms. Even if the two clients have completely different data distributions, both can benefit from knowing the same distribution drawn from all other clients' datasets. Notably, this method allows clients to join training at any time.

5.2 Federated Clustering Learning

Clustering is an unsupervised machine learning method that aims to generate multiple clusters of data with similar characteristics[28]. In FL, cluster models can be obtained by clustering clients and intra-cluster aggregation between global and local models. The cluster model has more robust adaptability to the clients within the clusters. This survey classifies the clustering into model-loss-based methods and client-similarity-based methods by their clustering bases.

The model-loss-based methods select a representative model as the cluster center, and clients can join the cluster of the model with the most negligible loss. GHOSH et al.[29] proposed the iterative federated clustering algorithm, IFCA. The server trained K models simultaneously and broadcasted the K models to all clients simultaneously. Each client joined a unique cluster by finding the model with the smallest loss. Cluster-based FL model aggregation was then performed on the server. However, since the server needs to broadcast K cluster models to all clients, its communication overhead is K times that of FedAvg. Based on Ref. [29], LI et al.[30] absorbed the idea of soft clustering and considered that different clusters have gradients and blurred boundaries. The authors divided clients into N-associated clusters and performed model fusion and local updates based on multiple cluster models. This method could utilize the information of boundary clients more effectively and realized information fusion between different clusters to a certain extent. Its communication cost is the same as in Ref. [28]. The loss-based method can ensure a specified number of cluster partitions. However, since several representative models need to be selected as cluster models (cluster centers), the clustering effect may be sensitive to the selected models and the number of them. Furthermore, the clients need to perform a cluster center model for each cluster to find the cluster with the smallest loss. This leads to extra computation for model loss-based methods.

The client-similarity-based methods take the similarity between model parameters or model parameter updates to represent client similarity. SATTLER et al. [31] proposed a recursive cluster FL method. After training the global model, this algorithm accorded with the cosine similarity between the last gradient updates of the client. Hierarchical clustering was used to iteratively bisect the client until the lower bound of the cosine distance within the cluster or the upper bound of the cosine distance between the clusters was satisfied. With the recursive method, the user does not need to pre-set the number of clusters. Even on non-convex optimization problems, solid mathematical guarantees for clustering quality can be provided. However, the model gradient-based methods have certain limitations. Because gradient descent methods may get stuck in local optima, those models' gradients pointing to the local optima cannot represent the similarity of these clients. Furthermore, in the gradient descent process, the model takes a mini-batch (a small subset of the data set) from the full data set each time to calculate the gradient, and then adjusts the parameters. The gradient directions given by these small mini-batches will vary. FRABONI et al.[32] proposed two aggregation sampling methods based on sample size and similarity. This algorithm pre-sets M different distributions and then puts clients into different distributions based on the number of samples or similarities. Experiments show that it can converge to a smaller value on non-IID data. ZHANG et al.[33] considered a measure of the same similarity between the client's computing power and its network conditions with the server and the skewed data distribution. Therefore, the client similarity was defined as the gradient direction and model update delay while solving the problems of data skew and system heterogeneity.

It is worth noting that clustering-based methods can achieve excellent results when applied to data distributions with transparent partitions. However, in the real world, such scenarios are minimal. More importantly, there is no good theoretical analysis to prove the validity of clustering basis, including methods based on model loss and model gradients.

5.3 Federated Knowledge Distillation

Knowledge distillation hopes to transfer the knowledge learned by machine learning models from specific tasks to related tasks[34]. The fundamental difference from traditional machine learning is that knowledge distillation relaxes the assumptions of IID data and allows for direct knowledge transfer between models. Therefore, in FL, clients can benefit from this process even if they have different data distributions. By reducing the importance of the data in the training process, a more adaptive client model can be obtained using federated knowledge extraction in a non-IID data setting. This survey classifies federated knowledge distillation into one-way distillation methods and mutual distillation methods based on the direction of knowledge transfer.

The one-way distillation methods can quickly transfer the knowledge contained in the dominant teacher model to the student model. In FL, both the server and the client can be set as teacher models. LIN et al. [35] proposed an ensemble distillation FedDF for federated model fusion. This algorithm built P groups of heterogeneous client models (which may vary in structure and numerical precision), evaluated on small batches of unlabeled data pre-stored by the server. The classification ensembles distill their logit output to train the student model on the server. This method improves the efficiency of client model training and has good robustness to data skew. LI et al.[36] proposed a federated learning method via model distillation (FedMD) by combining transfer learning and knowledge distillation. Each client used its own model prediction server to share the dataset to obtain class scores, and the server averaged the class scores as a global consensus. Each client learned this consensus through model distillation to obtain better client models. In this way, other clients' knowledge could be leveraged without the need to share its private data or model architecture explicitly. However, both the FedDF and FedMD have to pre-store a representative dataset on the client, which is a significant challenge. ZHU et al.[37] proposed a data-free knowledge distillation (FEDGEN) based on generative learning. The server used the client label prediction module (instead of the data) to learn a global generator that generated a feature representation matching the client-side labels. Each client model implemented knowledge distillation from the server to the client by sampling the generated feature representation. However, the one-way distillation may be challenging to achieve good results in the face of model heterogeneity.

The mutual distillation methods can be applied to diverse network architectures and are robust to heterogeneous models of different sizes. Better accuracy can also be achieved when training with a large number of clients. BISTRITZ et al. [38] proposed a distributed distillation algorithm that established a new topological relationship between clients, and each client could only connect and communicate with a few nearby devices. In each round of iterations, the clients accepted the soft network decisions of their neighbors in a chain, updated their soft network decisions through the consensus algorithm, and sent them to other neighbors. A more adaptive client model was obtained by limiting the loss of self-model features through knowledge distillation between adjacent clients. LI et al.[39] proposed FedH2L, which took the federated network as a collection of students, and all clients taught each other. To manage the global and client gradient conflict, they designed projected gradients to update the model to maximize intra-domain and cross-domain performance, performing well on non-IID data. Bidirectional distillation avoids the dependence on the powerful teacher model, and the student model can improve the learning efficiency and generalization ability of the network through online mutual learning. However, the methods of mutual distillation still lack theoretical analysis, and sometimes unavoidable negative knowledge transfer occurs. Participants may get caught up in groupthink, where the blind leads the blind.

6 Future Directions

Many methods have been proposed for the FL on non-IID data, but some problems are still not well solved. This section will discuss some of these challenges and examine their future research trajectories.

•Heterogeneity Analysis: A client heterogeneity analysis method is still missing, though the FL on non-IID data has received extensive attention and research. Specifically, existing methods for heterogeneity analysis based on model loss[29] or model parameter update (gradient)[31] have limitations such as being sensitive to manual settings and lack of theoretical proofs (details in Section 5.2) and cannot achieve the desired goal well. So how to design a client heterogeneity analysis method with good generalization is still an open problem.

•Hyperparameter: Existing FL methods for non-IID data have achieved good performance. However, the vast number of hyperparameters presented in these methods adversely affects the debugging and use of FL networks. Furthermore, due to the significant differences in the number and usage of hyperparameters for different methods, it is not easy to evaluate the actual effectiveness of those proposed innovative methods fairly.

•Security Assurance: The FL applications tend to have high privacy and high-risk characteristics, such as in the business and medical field, because of FL's better privacy. Some recent studies have shown that the privacy guarantees of FL methods, such as FedAvg, can be easily broken by attackers using methods such as inversion[40] and inference[41]. So it is necessary to conduct more in-depth research on possible attacks and corresponding preventions and design FL methods with more security assurance.

•Interpretability: The interpretability of deep learning has always been the focus and difficulty of research. Since the federated setting has the characteristics of distributed training and data heterogeneity, how interpreting its training and decision-making process will be more complicated. There is little discussion on the interpretability of FL today, and more reliable explanations can enhance users' confidence in FL.

•Dedicated Datasets: In FL, researchers need to design their data partitioning algorithms for CIFAR100, Fashion MINIST, and other existing datasets. Since the algorithm's performance may be diverse under different data distributions, the algorithm's performance cannot be well proved using the self-divided data set. Dedicated homogeneous and heterogeneous datasets and data partition algorithms must be designed to align with real-world environments.

7 Conclusions

This survey provides an overview of FL on non-IID data. First, the background and settings of both FL and non-IID data are introduced. Then, according to the motivation of existing methods, a new taxonomy is proposed. Specifically, the existing methods are classified into two categories: heterogeneity reducing strategies and adaptability enhancing strategies. In addition, the core ideas, key technologies, and main challenges of the methods are emphasized. Finally, the future research trajectories for some existing challenges in this field are conceived. We hope this work will help researchers to further overcome the challenges of FL on non-IID data.

References

[1] LECUN Y, BENGIO Y, HINTON G. Deep learning [J]. Nature, 2015, 521 (7553): 436–444. DOI: 10.1038/nature14539

[2] LIM W Y B, LUONG N C, HOANG D T, et al. Federated learning in mobile edge networks: a comprehensive survey [J]. IEEE communications surveys & tutorials, 2020, 22(3): 2031–2063. DOI: 10.1109/COMST.2020.2986024

[3] NGUYEN D C, DING M, PATHIRANA P N, et al. Federated learning for Internet of Things: a comprehensive survey [J]. IEEE communications surveys & tutorials, 2021, 23(3): 1622–1658. DOI: 10.1109/COMST.2021.3075439

[4] PFITZNER B, STECKHAN N, ARNRICH B. Federated learning in a medical context: a systematic literature review [J]. ACM transactions on Internet technology, 2021, 21(2): 1–31. DOI: 10.1145/3412357

[5] Google. Google trends [EB/OL]. [2022-06-01]. https://trends.google.com/trends/explore?q=federated%20learning&geo=US

[6] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [C]//Artificial Intelligence and Statistics. PMLR, 2017: 1273–1282. DOI: 10.48550/arXiv.1602.05629

[7] TAN A Z, YU H, CUI L, et al. Towards personalized federated learning [J]. IEEE transactions on neural networks and learning systems, 2022. DOI: 10.1109/TNNLS.2022.3160699

[8] ZHU H Y, XU J J, LIU S Q, et al. Federated learning on non-IID data: a survey [J]. Neurocomputing, 2021, 465: 371–390. DOI: 10.1016/j.neucom.2021.07.098

[9] HSIEH K, PHANISHAYEE A, MUTLU O, et al. The non-IID data quagmire of decentralized machine learning [C]//International Conference on Machine Learning. PMLR, 2020: 4387–4398. DOI: 10.48550/arXiv.1910.00189

[10] KARIMIREDDY S P, KALE S, MOHRI M, et al. Scaffold: stochastic controlled averaging for federated learning [C]//International Conference on Machine Learning. PMLR, 2020: 5132–5143. DOI: 10.48550/arXiv.1910.06378

[11] KOVÁCS G. An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets [J]. Applied soft computing, 2019, 83: 105662. DOI: 10.1016/j.asoc.2019.105662

[12] GUZMÁN-PONCE A, SÁNCHEZ J S, VALDOVINOS R M, et al. DBIG-US: a two-stage under-sampling algorithm to face the class imbalance problem [J]. Expert systems with applications, 2021, 168: 114301. DOI: 10.1016/j.eswa.2020.114301

[13] TUOR T, WANG S Q, KO B J, et al. Overcoming noisy and irrelevant data in federated learning [C]//The 25th International Conference on Pattern Recognition (ICPR). IEEE, 2020: 5020–5027. DOI: 10.1109/ICPR48806.2021.9412599

[14] YOSHIDA N, NISHIO T, MORIKURA M, et al. Hybrid-FL for wireless networks: cooperative learning mechanism using non-IID data [C]//2020 IEEE International Conference on Communications. IEEE, 2020: 1–7. DOI: 10.1109/ICC40277.2020.9149323

[15] YOON T, SHIN S, HWANG S J, et al. FedMix: approximation of mixup under mean augmented federated learning [EB/OL]. [2021-06-01]. https://arxiv.org/abs/2107.00233v1

[16] DUAN M M, LIU D, CHEN X Z, et al. Self-balancing federated learning with global imbalanced data in mobile systems [J]. IEEE transactions on parallel and distributed systems, 2021, 32(1): 59–71. DOI: 10.1109/TPDS.2020.3009406

[17] WU Q, CHEN X, ZHOU Z, et al. FedHome: cloud-edge based personalized federated learning for in-home health monitoring [J]. IEEE transactions on mobile computing, 2022, 21(8): 2818–2832. DOI: 10.1109/TMC.2020.3045266

[18] ZHAO J X, CHANG X Y, FENG Y H, et al. Participant selection for federated learning with heterogeneous data in intelligent transport system [J]. IEEE transactions on intelligent transportation systems, 2022, 99: 1–10. DOI: 10.1109/TITS.2022.3149753

[19] SHU J G, ZHANG W Z, ZHOU Y, et al. FLAS: computation and communication efficient federated learning via adaptive sampling [J]. IEEE transactions on network science and engineering, 2022, 9(4): 2003–2014. DOI: 10.1109/TNSE.2021.3056655

[20] ZHANG J, GUO S, QU Z H, et al. Adaptive federated learning on non-IID data with resource constraint [J]. IEEE transactions on computers, 2022, 71(7): 1655–1667. DOI: 10.1109/TC.2021.3099723

[21] WANG H, KAPLAN Z, NIU D, et al. Optimizing federated learning on non-

IID data with reinforcement learning [C]//IEEE Conference on Computer Communications. IEEE, 2020: 1698 – 1707. DOI: 10.1109/INFOCOM41043.2020.9155494

[22] STANDLEY T, ZAMIR A, CHEN D, et al. Which tasks should be learned together in multitask learning? [C]//International Conference on Machine Learning. PMLR, 2020: 9120 – 9132. DOI: 10.48550/arXiv.1905.07553

[23] SMITH V, CHIANG C K, SANJABI M, et al. Federated multitask learning [J]. Advances in neural information processing systems, 2017, 30. DOI:10.48550/arXiv.1705.10467

[24] HUANG Y T, CHU L Y, ZHOU Z R, et al. Personalized cross-silo federated learning on non-IID data [C]//Proceedings of the AAAI Conference on Artificial Intelligence. AAAI, 2021: 7865 – 7873. DOI: 10.48550/arXiv.2007.03797

[25] JAMALI-RAD H, ABDIZADEH M, SINGH A. Federated learning with taskonomy for non-IID data [J]. IEEE transactions on neural networks and learning systems, 2022. DOI: 10.1109/TNNLS.2022.3152581

[26] LI T, HU S, BEIRAMI A, et al. Ditto: fair and robust federated learning through personalization [C]//International Conference on Machine Learning. PMLR, 2021: 6357 – 6368. DOI: 10.48550/arXiv.2012.04221

[27] MARFOQ O, NEGLIA G, BELLET A, et al. Federated multi-task learning under a mixture of distributions [EB/OL]. [2022-06-01]. https://arxiv. org/abs/2108.10252

[28] LI Y, HU P, LIU Z, et al. Contrastive clustering [C]//2021 AAAI Conference on Artificial Intelligence. AAAI, 2021. DOI: 10.48550/arXiv.2009.09687

[29] GHOSH A, CHUNG J, YIN D, et al. An efficient framework for clustered federated learning [J]. Advances in neural information processing systems, 2020, 33: 19586 – 19597. DOI: 10.48550/arXiv.2006.04088

[30] LI C X, LI G, VARSHNEY P K. Federated learning with soft clustering [J]. IEEE Internet of Things journal, 2022, 9(10): 7773 – 7782. DOI: 10.1109/JIOT.2021.3113927

[31] SATTLER F, MÜLLER K R, SAMEK W. Clustered federated learning: model-agnostic distributed multitask optimization under privacy constraints [J]. IEEE transactions on neural networks and learning systems, 2021, 32(8): 3710 – 3722. DOI: 10.1109/TNNLS.2020.3015958

[32] FRABONI Y, VIDAL R, KAMENI L, et al. Clustered sampling: low-variance and improved representativity for clients selection in federated learning [C]//International Conference on Machine Learning. PMLR, 2021: 3407 – 3416. DOI: 10.48550/arXiv.2105.05883

[33] ZHANG Y, DUAN M, LIU D, et al. CSAFL: a clustered semi-asynchronous federated learning framework [C]//Proceedings of 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021: 1 – 10. DOI: 10.1109/IJCNN52387.2021.9533794

[34] PARK D Y, CHA M H, KIM D, et al. Learning student-friendly teacher networks for knowledge distillation [J]. Advances in neural information processing systems, 2021, 34. DOI: 10.48550/arXiv.2102.07650

[35] LIN T, KONG L, STICH S U, et al. Ensemble distillation for robust model fusion in federated learning [J]. advances in neural information processing systems, 2020, 33: 2351 – 2363. DOI: 10.48550/arXiv.2006.07242

[36] LI D L, WANG J P. FedMD: heterogenous federated learning via model distillation [EB/OL]. (2021-01-27) [2022-06-01]. https://arxiv.org/abs/2101.11296

[37] ZHU Z D, HONG J Y, ZHOU J Y. Data-free knowledge distillation for heterogeneous federated learning [J]. Proceedings of machine learning research, 2021, 139: 12878 – 12889. DOI:10.48550/arXiv.2105.10056

[38] BISTRITZ I, MANN A, BAMBOS N. Distributed distillation for on-device learning [J]. Advances in neural information processing systems, 2020, 33: 22593 – 22604

[39] LI Y Y, ZHOU W, WANG H M, et al. FedH2L: federated learning with model and statistical heterogeneity [EB/OL]. (2021-01-27) [2022-06-01]. https://ui.adsabs.harvard.edu/abs/2021arXiv210111296L/abstract

[40] JIN X, CHEN P Y, HSU C Y, et al. CAFE: catastrophic data leakage in vertical federated learning [EB/OL]. [2022-06-01]. https://arxiv. org/abs/2110.15122

[41] ZHANG J W, ZHANG J L, CHEN J J, et al. GAN enhanced membership inference: a passive local attack in federated learning [C]//2020 IEEE International Conference on Communications. IEEE, 2020, 1 – 6. DOI: 10.1109/ICC40277.2020.9148790

Biographies

HAN Xuming received his PhD degree from Jilin University, China. Now he is a professor and PhD supervisor at Jinan University, China. He is in charge of about 10 important scientific research projects and 80 journal papers and conference papers, and has publish four academic monographs. His research interests include artificial intelligence, federated Learning, and machine learning.

GAO Minghan is currently a graduate student in Changchun University of Technology, China. His research interests include federated learning, multitasking optimization, and clustering.

WANG Limin (20211016@gdufe.edu.cn) received her master's and PhD degrees in computer science and technology from Jilin University, China in 2004 and 2007, respectively. Now she is a professor with the Guangdong University of Finance & Economics, China. Her current research interests include big data analysis, evolutionary algorithm, and intelligent decision optimization. She is a member of China Computer Federation. She has published more than 90 research papers in international and domestic journals or international conferences.

HE Zaobo received his PhD degree from Georgia State University, USA, MS degree from Shaanxi Normal University, China, and BS degree from Yan'an University, China, all in the Department of Computer Science. Dr. HE is currently a professor in the Department of Computer Science at Jinan University, China. His research areas focus on data privacy and Internet of Things.

WANG Yanze is currently a graduate student in Jinan University, China. His research interests include federated learning, pattern recognition, and computer vision.