IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 28 March 2024, accepted 23 April 2024, date of publication 2 May 2024, date of current version 10 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3396279

RESEARCH ARTICLE

Pretraining Client Selection Algorithm Based on a Data Distribution Evaluation Model in Federated Learning

CHANG XU[1], HONG LIU[1,2], KEXIN LI[1], WANGLEI FENG[3], AND WEI QI[1,2]

[1]School of Information and Electrical Engineering, Hangzhou City University, Hangzhou 310000, China
[2]Academy of Edge Intelligence, Hangzhou City University, Hangzhou 310000, China
[3]College of Information Engineering, Zhejiang University of Technology, Hangzhou 310012, China

Corresponding author: Hong Liu (liuhong@hzcu.edu.cn)

This work was supported in part by Hangzhou City University Scientific Research Fund under Grant X-202106, and in part by the Key Research and Development Program of Hangzhou under Grant 2023SZD0073.

ABSTRACT Federated Learning (FL) allows task initiators (servers) to utilize data from task participants (clients) to train machine learning models while protecting data privacy. However, in the FL system, when the client data are non-independently identically distributed (Non-IID), appropriate metrics are chosen to accurately evaluate the quality of the client data, accordingly to select a reasonable subset of clients, and thus ensure the accuracy of the FL aggregation model. In this paper, based on the experimental results, a data distribution evaluation model is proposed, which is based on two metrics: the volume of client data and its increment and the balance of global client data. This data distribution evaluation model enables more accurate evaluation of clients with Non-IID characteristics. Based on this evaluation model, this paper further proposes an FL client subset selection algorithm. This algorithm accurately evaluates the data value of each client, enabling the server to select the most valuable subset of clients before FL training, thus improving the accuracy of the federated learning aggregation model in scenarios with Non-IID client data. When training the FL aggregation model using the proposed method on datasets composed of CIFAR-10, Fashion-MNIST, and DEAP distributions, compared to the optimal baseline, the average precision scores increased by 5.99%, 4.79%, and 4.29% respectively. The improvement in accuracy is more pronounced in scenarios with Non-IID data, such as in the DEAP dataset distribution with the highest Non-IID degree, where the accuracy increased by 5.30% compared to the optimal baseline.

INDEX TERMS Client selection, data distribution, federated learning, machine learning.

I. INTRODUCTION

In recent years, the rise of the Internet of Things (IoT) and the surge in the number of smart devices have generated massive amounts of data, which provides new opportunities to further improve the accuracy of AI models. However, the ensuing data privacy issues pose new challenges for model developers [1]. As a result, scholars have proposed Federated Learning (FL) - a distributed computational training method designed to protect user privacy [2]. FL takes full advantage of the abundant data resources and computational power on each device, with the server as the coordination center. Each client uses its local data to train the model and uploads only model-related information to the server for aggregation. Eventually, the aggregated model is formed on the server side, and the model is applied to realize the recognition task [3].

However, FL also faces numerous new challenges. In particular, due to the general non-independently identically distributed (Non-IID) nature of the data from each client, the optimization direction of the model on local data may result in a significant difference from the optimization direction of the aggregated model on the aggregated global data [4], [5]. This difference may make FL lower than traditional centralized AI methods in key performance metrics such as training accuracy and convergence speed. Other researchers found that when dealing with a multi-classification task on

The associate editor coordinating the review of this manuscript and approving it for publication was Seifedine Kadry.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

the CIFAR-10 dataset, FL decreased its accuracy by 55% compared to the traditional approach due to the client data exhibiting Non-IID characteristics.

Therefore, selecting balanced, diverse, and highly representative data has become crucial for enhancing FL performance. Traditional FL usually randomly selects a portion of clients to form a subset of clients to participate in the training [6], [7], and this approach may lead to an imbalance of the data distribution, which in turn makes the model prefer data from some specific clients, and ultimately affects the accuracy of the model. Existing solutions usually evaluate the model parameters or gradients uploaded by the client during the server training process as a way to guide the client's selection [8], [9]. However, this approach still faces several problems: First, a large amount of client data needs to be purchased, leading to an increase in cost [10]. Second, the evaluation process consumes a large amount of computational resources and has a high time cost [11], [12], [13]; and finally, the client needs to upload the model parameters or gradients, which may lead to privacy leakage issues [14]. Therefore, FL requires effective value evaluation and screening of clients before training, aiming to mitigate the negative impact of client Non-IID characteristics on FL performance, while ensuring that client privacy is protected.

For the multi-classification task scenario, this paper proposes a client selection algorithm based on data quality evaluation metrics, which can accurately evaluate the value of each client before training to construct a reasonable subset of clients under the premise of protecting data privacy. By utilizing the data provided by this client subset for FL, the resulting aggregated model trained not only has high recognition accuracy but also effectively reduces the capital, communication, and computational costs in the FL process. We validate this on the image dataset CIFAR-10, Fashion-MNIST, and the emotion recognition dataset DEAP. The main contributions of this paper are as follows:

- Through experiments, we identified the characteristics of the functions required to assess data volume and its increment, as well as the balance of global data, and established evaluation metrics for evaluating the quality of client data based on these characteristics.
- We propose an algorithm based on client data quality evaluation metrics for selecting a subset of clients required before FL training. We apply this client subset selection algorithm to a standard federated learning task system example.
- We simulated and built the whole set of FL application flow and verified that this client selection algorithm has a high accuracy of the aggregation model obtained from federated model training in Non-IID scenarios for image classification and EEG-based emotion recognition scenarios.

II. RELATED WORK

The implementation of FL usually involves the participation of hundreds or even thousands of clients [15]. However, due to the constraints of budget and resource limitations, a small number of clients are generally selected to participate during actual federated learning training [16], and client selection has become a key factor in improving the performance of federated learning. Compared to the traditional method of randomly selecting clients [11], [17], much literature has proposed a variety of novel client selection methods. For example, literature [18] proposes an algorithm that prioritizes clients with higher local losses to participate in each round of training, thus accelerating the convergence of errors. In [19], the participating client models are evaluated by evaluating the model during each round of training to facilitate the speed of convergence. Literature [20] explores the problem of inactive clients and their incomplete updates for fast convergence of FL. In literature [21] and [22], Shapley values are introduced to evaluate the value of clients as a way to normalize the client selection process.

The above approach requires evaluating clients in each round of training, a process that consumes a large amount of computational resources and time and thus encounters challenges in practical applications. Given this, scholars have proposed a new algorithm, i.e., pre-evaluating the clients before the training is initiated, and letting the clients that are qualified by the evaluation results participate in each round of training for FL. For example, in [23], a method was designed in which the volume of data owned by a client is used as a client auction factor to attract clients with more data to join FL. Another method [24] measured the client value and selected clients by calculating the uniformity and diversity of the client dataset distribution. Reference [25] suggested that selecting clients with large and nearly uniformly distributed data improved model performance. The above methods face limitations in their applicability and accuracy of FL in the scenario of Non-IID data.

III. MAIN IDEA AND DESIGN METHOD

It has been found that the accuracy of federated learning is closely related to the quality of client data [15], [17], [24], [26], so the data quality of each client needs to be effectively evaluated. In this paper, a new client data quality evaluation metric is proposed, and an FL client selection algorithm is designed based on this evaluation model. Using this client selection algorithm, the clients are evaluated before the training, the clients that meet the conditions are filtered out, and the filtered clients participate in the training of each round of FL. The client selection algorithm is particularly well-suited for Federated Learning in Non-IID data scenarios, effectively enhancing FL accuracy.

Next, we will first introduce two core metrics for evaluating client data quality, then describe the client data quality evaluation metrics designed in this paper, and finally propose a client selection algorithm for selecting a reasonable subset of clients in FL.

A. DESIGN OF THE CLIENT SELECTION ALGORITHM

1) EVALUATION METRIC SELECTION

To design reasonable data quality evaluation metrics, this study explores the impact of data volume and data distribution on the performance of FL. Existing evaluation methods are mainly based on the total amount of client data and the variance of the client data distribution. In our preliminary experiments using the CIFAR-10 [27] dataset to train the ResNet-56 [28] model, we found that federated learning in non-IID data scenarios exhibits the following two characteristics: First, the volume of data in each class and its increment in the client dataset has an impact on the recognition accuracy of the aggregation model. Second, a balanced global data distribution ensures that the aggregation model maintains high recognition accuracy.

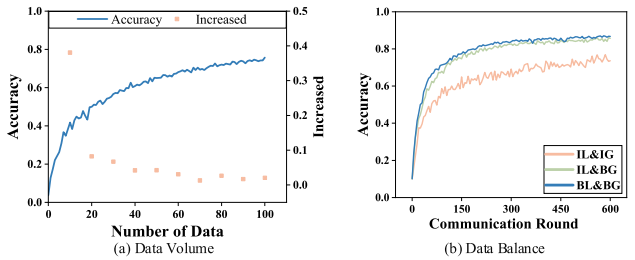


FIGURE 1. Impact of data volume and distribution on the FL training task.

a) The volume of client data and its increment. In the client data sets, the volume of data and its increment in each class have an impact on the recognition accuracy of the aggregated model. We use the ResNet-56 model to conduct multiple FL experiments on the CIFAR-10 dataset, setting up 10 clients for each experiment with 200 iteration cycles, and recording the recognition accuracy of each aggregated model. The experiments start from an extreme Non-IID data configuration, i.e., initially, each client has only 100 samples of data for a particular class in CIFAR-10. CIFAR-10 is a 10-categorization task, meaning that each client starts with only one class to train. Subsequently, 1 image from each of the other 9 categories was added to each client at a time during each FL retraining (9 images in total) to train and record the recognition accuracy. Ultimately, the process was repeated until the volume of data for each class reached 100 images. As a result, a total of 100 FL aggregation models were trained, and their recognition accuracies are shown in Figure 1a. Figure a illustrates the trend in the accuracy of these aggregation models as well as their variation(Δ Accuracy), where the horizontal axis indicates the volume of data added to categories other than the initial one. As can be seen from the figure, with the increase in the amount of data for each class at the client, the recognition accuracy of the aggregation model has been improved. However, once the data volume for each class on the client reaches a certain level, the recognition accuracy of the aggregation model does not improve significantly, resulting in a marginal diminishing effect.

b) The uniformity of the global data. A balanced global data distribution ensures that the aggregation model maintains high recognition accuracy. A more balanced global data distribution in federated learning can effectively reduce the impact of Non-IID on the recognition accuracy of FL aggregation models. As shown in Figure 1b, the accuracy of the aggregated model recognition is compared through three types of scenarios: The first, a balanced client local distribution and balanced global distribution (BL&BG); the second, a balanced global distribution and imbalanced client local distribution (IL&BG); and the third, an imbalanced global distribution and imbalanced client local distribution (IL&IG). In Figure 1b, it is shown that the aggregation model's recognition accuracy decreases from 87% to 74% when trained with FL using data from BL&BG, compared to FL using data from IL&IG. When trained with FL using data from IL&BG, the aggregated model experiences only a 1% decrease in recognition accuracy. In practice, the distribution of client data depends on the data it holds, so the server cannot directly interfere with the distribution of client data. However, the server can mitigate the negative impact of local data imbalance on the recognition accuracy of the integrated model by selecting a subset of clients participating in FL training to adjust the balance of the global data distribution.

2) MATHEMATICAL MODELS FOR EVALUATION

The experiments in the previous section lead us to evaluate the quality of the data through the following two dimensions: First, the evaluation metric for data volume should emphasize the volume of data per class for each customer and the marginal benefits derived from its incremental growth;

TABLE 1. Notation and description.

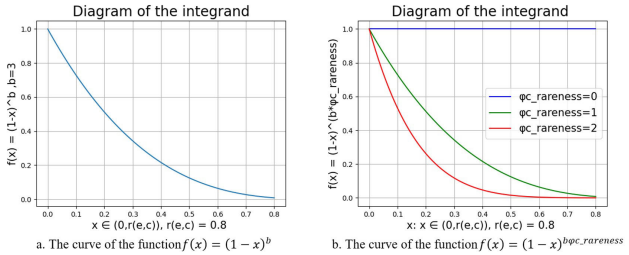
notation	description
φ_{e_volume}	The metric of data volume and its increment
φ_e^c	Coefficient of data volume and its increment for each class c of client e
c, C	Class, number of classes
n_e^c	Number of class c in client e
N_c	Total number of class c
E	Number of clients
$\varphi_{c\_rareness}$	Data balance metric
U	The average of the global data volume over the number of categories
C_p	Number of offsets for class
$\varphi(n_e^c)$	the total data evaluation metric
N	Total data volume
e	Client
$\varphi_{e\_evaluation}$	Data quality evaluation model

Second, when evaluating the client data distribution, more attention should be paid to the balance of the global data distribution.

Therefore, the evaluation metric for each client's data volume and its increment should fulfill the following two conditions: 1) The function should be monotonically increasing. 2) Its derivative should be monotonically decreasing and the growth of the function stabilizes after a certain threshold amount of data. Based on this we designed the data volume and its increment evaluation metric as shown in equation 1.

$$\begin{aligned} \varphi_{e\_volume} &= \sum_{c=1}^{C} (\varphi_e^c) \\ \varphi_e^c &= \int_0^{r(e,c)} (1-x)^b dx \\ r\,(e,c) &= \min\left\{ \frac{n_e^c}{N_c/E}, 1 \right\} \end{aligned} \tag{1}$$

In Equation 1, e denotes the client, and c denotes the data class. The metric φ_{e_volume} for the data volume and its increment for client e is calculated by cumulating the class data volume and its increment coefficient φ_e^c for each class c of the client e. In φ_e^c, $r(e,c)$ denotes the upper limit of data integration, i.e., the ratio of the number of categories c in customer n_e^c to the average number of categories c in all customers (i.e., the total number of categories c in all customers divided by the number of customers E) - $\frac{n_e^c}{Nc/E}$. If $\frac{n_e^c}{Nc/E}$ is greater than 1, set the value of $r(e,c)$ to 1. If $\frac{n_e^c}{Nc/E}$ is less than 1, it indicates that the quantity of class c in client n_e^c is lower than the average quantity of class c across all clients, indicating that class c is rarer. At this point the value of $r\,(e,c)$ increases with the number of categories c, ensuring that the data volume and its increment coefficient φ_e^c is large for categories with large data amounts. At the same time, the impact of adding additional data units to the overall evaluation metric exhibits diminishing marginal utility as the amount of data increases. Therefore, for the integrand function $f(x)$, the decreasing function chosen is $(1-x)^b$, as shown in Figure 2a. Here, $b\,(b>0)$ is the constant that adjusts the trend of the decreasing function. As the number of categories c in customer n_e^c increases, the increase in φ_e^c slows down until $\frac{n_e^c}{Nc/E}$ is greater than 1. The value of $r(e,c)$ no longer increases with the number of categories c, but is simply set to 1.



a. The curve of the function $f(x) = (1-x)^b$ b. The curve of the function $f(x) = (1-x)^{b\varphi c\_rareness}$

FIGURE 2. Schematic diagram of the product function.

The data volume and its increment metric proposed in this paper aim at evaluating the data characteristics of each class in the client in a more fine-grained way to ensure that each class in the subset of clients selected for FL has sufficient data, rather than just evaluating the overall data amount of the subset of clients.

A balanced global data distribution ensures that the FL aggregation model has higher recognition accuracy. We assign each class a data balance metric $\varphi_{c\_rareness}$ related to its rareness based on the distributional properties of the global data, which is defined as follows:

$$\varphi_{c\_rareness} = \left[1 - \left(\frac{U - Nc}{U + Nc}\right)\right] \tag{2}$$

In the above equation, U is the average of the global data volume over the number of categories, and N_c is the number of category c in the global data volume. Equation 2 measures the rareness of the amount of data in class c in the total amount of data, i.e., the rarer the data in a particular category is, the lower the data balance metric $\varphi_{c\_rareness}$ is. This reminds us that clients with rare data should be given due consideration in client selection algorithms to prevent rare data from being overlooked due to the inadequacy of statistical methods.

We build a comprehensive data quality evaluation model $\varphi_{e\_evaluation}$ by incorporating the data balance metric $\varphi_{c\_rareness}$ into the exponential term of the data volume and its increment evaluation metric φ_{e_volume}, as shown in Equation 3.

$$\begin{aligned} \varphi_{e\_evaluation} &= \sum_{c=1}^{C} \varphi(n_e^c) \\ \varphi(n_e^c) &= \int_0^{r(e,c)} (1-x)^{b\varphi c\_rareness} dx \end{aligned} \tag{3}$$

Figure 2b shows the graph of the integrand function containing the data balance metric $\varphi_{c\_rareness}$. The higher the data balance metric $\varphi_{c\_rareness}$, the lower the rate of decrease of the integrand function, while the data volume controls the upper limit of the integral. Therefore, the higher the data balance metric $\varphi_{c\_rareness}$, the larger the total data evaluation metric $\varphi(n_e^c)$ is for a certain upper limit of integration. This ensures that for a given volume of data, the rarer the data the higher the evaluation metrics.

The data quality evaluation model proposed in this paper has the following advantages:

a) Considered the impact of both the size of the data volume and the increment of data volume for each class within the client on the recognition accuracy of the FL aggregation model;
b) Variables N and N_c, based on global data design, prevent excessive emphasis on data balance distribution at the individual client level, while also avoiding the neglect of those data resources that, although rare, are of high quality overall.
c) A data balance metric $\varphi_{c\_rareness}$ is introduced into the index term to evaluate the combined effect of the size

of the data volume, the increment of the data volume and the global data balance on the data quality, thus enhancing the flexibility and adjustability of the data quality evaluation metrics.

B. CLIENT SELECTION ALGORITHM

Our client subset selection algorithm can be described by the following algorithm:

Algorithm 1 Client Selection and Payment Determination

Input: Reward R, Expected number of clients K, Client's pool $\mathbf{E}$;
Output: Selected clients set $\mathbf{S}$, Payment for all clients $\mathbf{P}$;
Server-side:
Collect global distribution $\boldsymbol{D} = \{N_c\}_{c \in C}$ from client's pool $\mathbf{E}$ and compute the total volume of data $N = \sum_{c \in C} N_c$;
Define data evaluation function φ based on $\boldsymbol{D}$ and N. Send φ to all clients;
Clients side:
Calculate data score s_e according to evaluation function φ, $s_e \leftarrow \varphi\left(n_e^c\right)$ and send it to the server;
Server-side:
Collect s_e from all clients and sort s_e by $s_{e_1} > s_{e_2} > \ldots > s_{e_3} > \ldots > s_{e_E}$; Select Top-K clients $\mathbf{S} \leftarrow [e_1, e_2, \ldots, e_K]$;
For each $e \in \mathbf{E}$ do
  If $e \in \mathbf{S}$ do
    $p_e = R/K$;
  Else
    $p_e = 0$;
  End if
End for
Return $\mathbf{S}$, $\mathbf{P} \leftarrow \{p_e\}_{e \in \mathbf{E}}$.

Note that the client selection algorithm is applied only once before FL starts the training task, as shown in Figure 3.

We designated the server as the FL task publisher and the client as the FL task applicant. They operate through the following four steps:

1) EVALUATION METRIC SELECTION

The server releases the task requirements (data, labeling, and hardware requirements) and the expected number of FL task clients K. Clients willing to participate in the training task submit a request to participate and the payment they want to receive, and only K clients can participate in the task.

2) DATA EVALUATION

The server sends a data collector, which only collects the volume and increment of data from the clients and the global distribution of the data. The server then defines and parameterizes a data evaluation function φ that is used to evaluate the value of each client's data. The function φ combines the collective statistical information from the clients to perform data evaluation for each client.

3) CLIENT SELECTION

After the scores of each client are determined, all clients are sorted according to the score, the top K clients with the highest score are selected, and the payment remuneration of the selected clients is determined.

4) FL TASK

The server starts the federated learning task based on the client selection result. Model training and updating are achieved by iteratively passing model parameters between the client and the server.

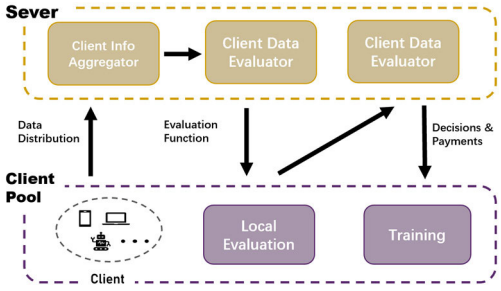


FIGURE 3. Federated learning client data selection process.

After completing the aforementioned client selection algorithm, FL will select suitable clients to form a client subset, and then proceed with the training and recognition of the model according to the original steps of FL. We name the client selection algorithm proposed in this paper as DDEM (Client Selection Algorithm Based on a Data Distribution Evaluation Model).

C. COMPLEXITY ANALYSIS OF ALGORITHMS

DDEM employs a more fine-grained strategy, allowing the server to evaluate the contribution of each category from each client to the training process before training. Based on this evaluation, a suitable subset of clients is selected to participate in the training. During the training process of federated learning, DDEM is no longer used to evaluate the client, so the complexity of the algorithm is lower. Specifically, we use $\varphi\left(\mathrm{n}_\mathrm{e}^\mathrm{c}\right)$ from Equation 3 to represent the calculation of the value of a single data class for a client data, where $\varphi(\mathrm{n}_\mathrm{e}^\mathrm{c})$ is the total data evaluation metric. Given that the server needs to compute these scores for all data categories (C) of all clients (E), the overall complexity of the DDEM algorithm can be represented as $E * C * \varphi(\mathrm{n}_\mathrm{e}^\mathrm{c})$. For example, if the server specifies a requirement for data from 10 categories ($\mathrm{C} = 10$), and there are 20 clients interested in joining the federated learning task, the server needs to calculate $\varphi\left(\mathrm{n}_\mathrm{e}^\mathrm{c}\right)$ a total of 200 times.

IV. EXPERIMENTATION AND RESULTS

Our experimental settings were as follows. In the experiments, a subset of clients was picked once using our method before each FL task started training. The selected clients were all added to the FL training task using the same hyperparameters, with the batch size set to 128, and in the simulated experiment with 100 clients, the batch size was set to 64; the learning rate was set to 0.003. All clients participated in local training for 1 epoch before joining the aggregation. The model aggregation was carried out on the server, and the aggregated model was tested for accuracy. The test accuracy and loss were recorded to serve as observations of the model's performance. The CIFAR10 experiment was halted after 600 rounds of model aggregation, while the DEAP experiment was concluded after 200 rounds of model aggregation.

A. SYSTEM SETTINGS

Our experiments were conducted on a Linux server with 40-core 4.0 GHz Intel Xeon CPUs and 8 NVIDIA V100 GPUs. On the server side, a document container with 8 CPU cores and 1 GPU was configured, and the clients shared the remaining computational resources equally. All functions are implemented via Python, where the FL function and the model training function are based on FedML and PyTorch, respectively.

B. APPLICATIONS

The pretraining client selection algorithm proposed in this paper is applied to a multi-classification task in an FL environment. To verify the generalizability and accuracy of the algorithms, two representative FL application scenarios are selected in this paper, and three multiclassification datasets are comprehensively evaluated under these scenarios.

1) APPLICATION #1

Image classification is a widely studied application domain. In this paper, we adopt the ResNet model to perform image classification tasks and conduct model training and evaluation on two datasets: CIFAR-10 and Fashion-MNIST (FMNIST). CIFAR-10 consists of 60,000 32×32 pixel color images that have been grouped into 10 different categories such as airplanes and automobiles. Among them, 50,000 images are used for training, and the remaining 10,000 are used for testing. FMNIST contains 70,000 grayscale images of 28×28 pixels classified into 10 categories such as pants and pullovers. Among them, 60,000 were used for training and 10,000 for testing. The training dataset is distributed to each client according to the Dirichlet distribution, which means that each client receives a different amount of training data and its distribution, while the test dataset has a balanced distribution.

2) APPLICATION #2

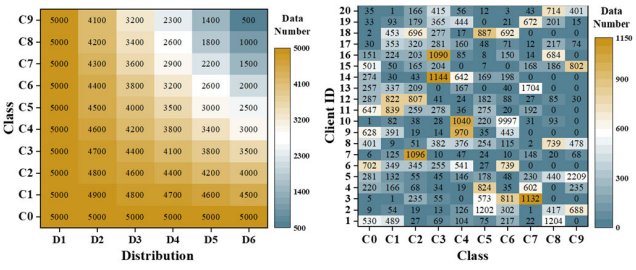
Applying deep learning techniques to the field of brain-machine interfaces, especially in emotion recognition, has shown tremendous potential [29], [30]. This study utilized the publicly available EEG dataset DEAP [31] to train the emotion recognition model. The dataset includes EEG and related physiological signals from 32 participants during prolonged audiovisual stimuli presentation, which can be used to analyze human emotional states. Likeness labels in DEAP are quantified as numerical values ranging from 1 to 9, constituting an imbalanced nine-category label for the Likeness prediction task. We sampled the DEAP dataset to obtain a sample set containing 90,000 labels with a balance distribution, and divided it into training and testing sets at an 8:2 ratio, to construct a Non-IID experiment.

C. CLIENT-SIDE DATA DISTRIBUTION

In reality, similar to emotion recognition, sample collection is based on the subjective thoughts of the subject or patient; therefore, the data distribution of the datasets used by the various organizations that own the data and participate in FL is often different. It is more likely that the data distributions of the different FL task participants are not independently identically distributed. Therefore, to simulate realistic scenarios, we constructed datasets with different distributions and distributed these data to clients.

D. DATA DISTRIBUTION GENERATION

We constructed datasets with different data distributions to test our method, and we treated the original dataset as distribution D1, as shown in Figure 4. For a comprehensive evaluation of our algorithm, we removed a certain number of samples from each class of the original data and downsampled the original dataset into sub-datasets with uneven sample sizes for each class. We generated five (D2-D6) subdatasets with different distributions by incremental deletion. Figure 8a illustrates the data generation process using the CIFAR10 dataset as an example, where Class denotes the class type and Distribution denotes that we created six different data distributions.



(a) Distribution list of CIFAR 10 (b) Allocating CIFAR 10-D1 to Clients

FIGURE 4. Generated data distributions of CIFAR10 and an example of its allocation.

E. DATA DISTRIBUTION TO CLIENTS

For each experiment, a generated dataset was distributed to clients. In this process, data samples from one of the dataset's classes were distributed to the client using a set of random numbers that followed a Dirichlet distribution. This process was repeated for all classes within the dataset. The size of the α parameter in the Dirichlet distribution influenced the variance of its random numbers. In the experiments described

in this paper, α is set to 0.5. Figure 4b illustrates an example of distributing CIFAR10-D1 to 20 clients.

Baseline: In FL, most existing client selection strategies are applied during the training phase, where these strategies select clients based on updated local models [32], [33], which is not applicable for pretraining data evaluation. Therefore, we compared our approach with four baselines in the literature that could be applied to select clients before the training phase.

1) DDS

This client selection algorithm evaluates the contribution of client distribution to FL using two evaluation metrics, statistical homogeneity, and content diversity and selects clients whose data tend to be uniformly distributed and diverse in content.

2) DICE

This algorithm uses the volume of client data and the variance of the client distribution as metrics for evaluating data quality, aiming to select clients with a large volume of data and a data distribution close to a uniform distribution.

3) QBS

The server sorts the clients based on their number of samples and selects the client with the highest number of samples to enter the joint learning task in order.

4) RS

The server randomly selects a specified number of clients from the full set of clients into the FL task.

We compared our algorithm (DDEM) with four different BASELINE algorithms to evaluate the ability of our method to select high-quality clients. Figures 5, 6, 7, and 8 show DDEM's ability to select 25% and 50% of customers, as well as our validation accuracy on two different datasets. Figures 5, 6, and 8 depict selection from 20 customers, while

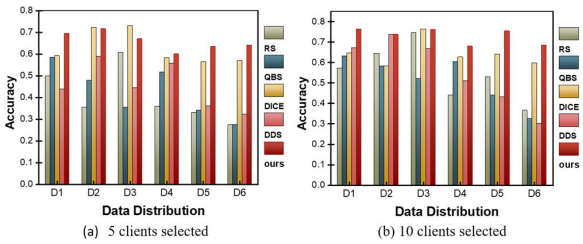


FIGURE 5. CIFAR 10: n clients selected from 20 clients.

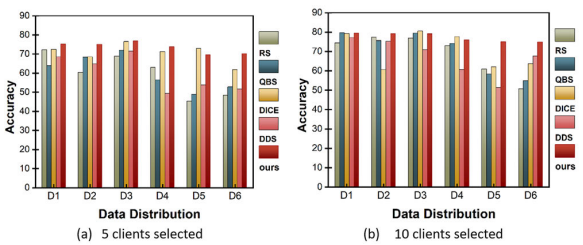


FIGURE 6. FMNIST: n clients selected from 20 clients.

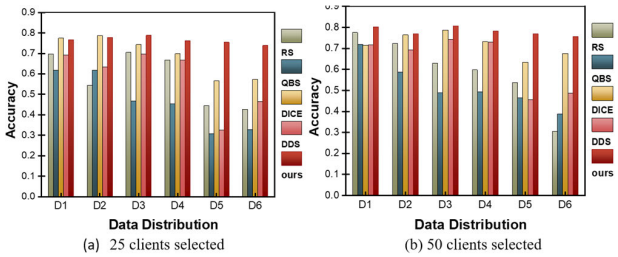


FIGURE 7. CIFAR 10: n clients selected from 100 clients.

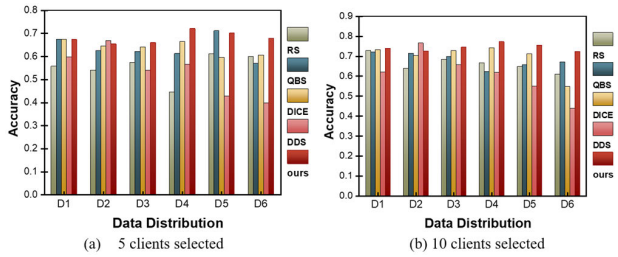


FIGURE 8. DEAP: n clients selected from 20 clients.

Figure 7 depicts the effect of selection from 100 customers on the CIFAR10 dataset.

As shown in Figures 5, 6, 7, and 8, our method obtains higher validation accuracies in most cases than other baselines (RS, QBS, DICE, DDS), especially when the global distribution is unbalanced (e.g., D4, D5, D6). As shown in the experiments in Figure 5, with 20 clients and on the CIFAR10 dataset, DDEM achieves an average model accuracy of 69.49% in a total of 12 experiments selecting subsets of 5 and 10 clients. Compared to the four baselines, DDEM showed an overall higher average validation accuracy on CIFAR10, with improvements of 21.83%, 22.36%, 5.99%, and 19.20%, respectively. As shown in the experiment in Figure 6, we did the same experiment in FMNIST to validate the effectiveness of the method, and the average model accuracy reached 75.36%. Compared to the four baselines, DDEM showed an overall higher average validation accuracy on FMNIST, with improvements of 11.10%, 9.95%, 4.79%, and 11.89%, respectively.

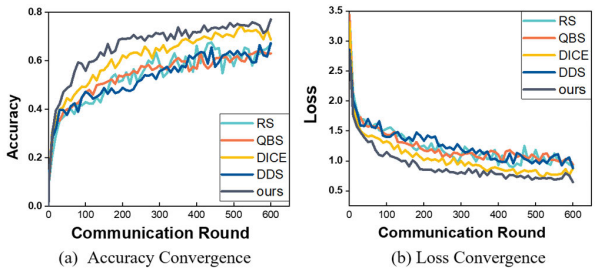
However, DDEM does not exhibit a significant advantage over the baseline with more balanced data distributions (e.g., D1, D2). This is because when the number of selected clients is small and the global data distribution is relatively balanced, selecting high-quality clients is not challenging, and other baseline strategies (e.g., DICE) can also achieve this goal.

The effectiveness of our data evaluation is also validated in the DEAP dataset, the results of which are presented in Figure 6. The average accuracies after FL, when 5 and 10 clients are selected through DDEM, are 68.19% and 74.32%, respectively. For the case of selecting 5 clients, the average accuracy improvements compared to other baselines (RS, QBS, DICE, DDS) are 12.60%, 4.48%, 4.29%, and 14.71%, respectively. Similarly, when selecting 10 clients, the average accuracy increases by 8.10%, 6.33%, 4.94%, and 13.50% compared to the respective baselines. Notably,

when selecting 10 clients from D6, which exhibits the most uneven distribution, the improvement effect is most significant, with enhancements of 11.44%, 5.30%, 17.48%, and 28.42% higher than the baseline approaches, respectively.

To investigate whether DDEM could be effective in scenarios with more clients, we increased the number of clients from 20 to 100 and conducted an experiment to select 25 and 50 clients from the pool of 100 clients. In this experiment, the total volume of data remained the same as the distribution depicted in Figure 5; only the number of clients was increased. As depicted in Figure 7, DDEM exhibits higher test accuracy than the other baselines when selecting 25 clients, showcasing average accuracy improvements of 18.35%, 29.87%, 7.40%, and 18.40%, along with a similar accuracy improvement when selecting 50 clients.

Figure 9 shows the accuracy and loss curves of DDEM and all baselines. As shown in the figure, as DDEM selects clients that join FL, the verification accuracy increases and loss decreases faster than those of other baselines. For example, it takes 90 rounds for DDEM to reach 60% accuracy, but 170, 390, 360, and 270 rounds for DICE, DDS, QBS, and RS, respectively. As shown in Figure 8b, DDEM has a much smaller training loss in a very large fraction of rounds.



(a) Accuracy Convergence (b) Loss Convergence

FIGURE 9. Performance comparison with the existing strategies.

In this section, we further analyze the effectiveness of the client selection algorithm by comparing the data distribution of the selected clients. We compared the selection outcomes of our approach with the best-performing baseline DICE in the selection of 5 clients out of 20 from CIFAR10-D6. Please note that D6 is an extreme case with a highly imbalanced global data distribution (refer to Figure 8a).

As shown in Table 2, both strategies select 3 identical clients (2, 6, 18). A comparison of the other 2 clients selected by the two strategies shows that our selection of clients 15 and 16 contains data samples from all categories, while the clients selected by DICE (i.e., 0 and 1) omit these rare categories (i.e., 7, 8, and 9). Since DDEM outperforms DICE in terms of validation accuracy, we can conclude that the DDEM selection algorithm is more effective in the case of data balance.

According to Table 2, DDEM selects fewer clients with less data and lower variance across different categories than DICE. During the selection process, DICE considers only local data variance, whereas DDEM values a globally balanced distribution and tends to select clients with rare data samples. With this algorithm, we can select more valuable samples at a smaller cost, reducing computational and communication costs during the training process.

TABLE 2. Selection results of DDEM and DICE CIFAR10-D6.

Client \ Class		C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	Distribution
Common	Client 2	5	120	112	311	472	28	0	319	360	0	
	Client 6	6	125	66	14	207	75	6	97	121	36	
	Client 18	33	212	23	99	29	64	141	109	157	25	
DICE	Client 0	530	38	59	349	33	346	209	0	0	0	
	Client 1	9	203	19	224	224	318	362	40	0	0	
ours	Client 15	151	378	29	4	85	44	125	12	69	174	
	Client 16	30	42	171	201	11	1	322	111	53	70	
ours	Data volume		2083		Variance			17544.01		Accuracy		65.91%
DICE			2963					71073.79				62.67%

V. CONCLUSION

Selecting a reasonable subset of clients to obtain appropriate data information for participation in FL can effectively ensure the recognition accuracy of the FL aggregation model. Based on this, we propose a client selection algorithm based on data distribution evaluation, which enhances the performance of the FL model, including the convergence speed and accuracy, without collecting private client data. The method is applied before training, which helps to reduce the computational and communication costs of the server. Additionally, through a specially designed evaluation model, the client data is comprehensively evaluated by considering both the data volume and its increment, as well as the global balance of data, and dynamic adjustments are made between them. Our algorithm performs well under a strong Non-IID distribution and can be applied to a wide range of application scenarios. In this paper, we compare four different benchmark methods, and the clients selected by our algorithm in different scenarios generally yield higher and more stable test accuracies, with average accuracy improvements of 5.99% and 4.29% on the CIFAR-10 and DEAP datasets, respectively. The improvement in accuracy is more pronounced in scenarios with Non-IID data, such as in the DEAP dataset distribution with the highest Non-IID degree, where the accuracy increased by 5.30% compared to the optimal baseline.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[2] A. Li, L. Zhang, J. Wang, J. Tan, F. Han, Y. Qin, N. M. Freris, and X.-Y. Li, "Efficient federated-learning model debugging," in *Proc. IEEE 37th Int. Conf. Data Eng. (ICDE)*, Apr. 2021, pp. 372–383.

[3] P. Kairouz, P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, and K. Bonawitz, "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, nos. 1–2, pp. 1–210, 2021.

[4] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-IID data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3400–3413, Sep. 2019.

[5] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," 2018, *arXiv:1806.00582*.
[6] W. Zhang, X. Wang, P. Zhou, W. Wu, and X. Zhang, "Client selection for federated learning with non-IID data in mobile edge computing," *IEEE Access*, vol. 9, pp. 24462–24474, 2021.
[7] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–7.
[8] Y. J. Cho, J. Wang, and G. Joshi, "Towards understanding biased client selection in federated learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2022, pp. 10351–10375.
[9] G. Wang, C. X. Dang, and Z. Zhou, "Measure contribution of participants in federated learning," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 2597–2604.
[10] R. Hu and Y. Gong, "Trading data for learning: Incentive mechanism for on-device federated learning," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.
[11] F. Shi, C. Hu, W. Lin, L. Fan, T. Huang, and W. Wu, "VFedCS: Optimizing client selection for volatile federated learning," *IEEE Internet Things J.*, vol. 9, no. 24, pp. 24995–25010, Dec. 2022.
[12] S. Ji, W. Jiang, A. Walid, and X. Li, "Dynamic sampling and selective masking for communication-efficient federated learning," *IEEE Intell. Syst.*, vol. 37, no. 2, pp. 27–34, Mar. 2022.
[13] A. M. Abdelmoniem, A. N. Sahu, M. Canini, and S. A. Fahmy, "REFL: Resource-efficient federated learning," in *Proc. 18th Eur. Conf. Comput. Syst.*, vol. 2023, pp. 215–232.
[14] Z. Lian, W. Wang, and C. Su, "COFEL: Communication-efficient and optimized federated learning with local differential privacy," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2021, pp. 1–6.
[15] C. Li, X. Zeng, M. Zhang, and Z. Cao, "PyramidFL: A fine-grained client selection framework for efficient federated learning," in *Proc. 28th Annu. Int. Conf. Mobile Comput. Netw.*, 2022, pp. 158–171.
[16] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konecný, S. Mazzocchi, B. McMahan, and T. Van Overveldt, "Towards federated learning at scale: System design," in *Proc. Mach. Learn. Syst.*, vol. 1, 2019, pp. 374–388.
[17] L. Fu, H. Zhang, G. Gao, M. Zhang, and X. Liu, "Client selection in federated learning: Principles, challenges, and opportunities," *IEEE Internet Things J.*, vol. 10, no. 24, pp. 21811–21819, Dec. 2023.
[18] Y. Jee Cho, J. Wang, and G. Joshi, "Client selection in federated learning: Convergence analysis and power-of-choice selection strategies," 2020, *arXiv:2010.01243*.
[19] M. Ribero and H. Vikalo, "Communication-efficient federated learning via optimal client sampling," 2020, *arXiv:2007.15197*.
[20] Y. Ruan, X. Zhang, S.-C. Liang, and C. Joe-Wong, "Towards flexible device participation in federated learning," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 3403–3411.
[21] Y. Liu, Z. Ai, S. Sun, S. Zhang, Z. Liu, and H. Yu, *FedCoin: A Peerto-Peer Payment System for Federated Learning*. Cham, Switzerland: Springer, 2020.
[22] T. Song, Y. Tong, and S. Wei, "Profit allocation for federated learning," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 2577–2586.
[23] R. Zeng, S. Zhang, J. Wang, and X. Chu, "FMore: An incentive scheme of multi-dimensional auction for federated learning in MEC," in *Proc. IEEE 40th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Nov. 2020, pp. 278–288.
[24] R. Saha, S. Misra, A. Chakraborty, C. Chatterjee, and P. K. Deb, "Data-centric client selection for federated learning over distributed edge networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 34, no. 2, pp. 675–686, Feb. 2023.
[25] A. Li, L. Zhang, J. Tan, Y. Qin, J. Wang, and X.-Y. Li, "Sample-level data selection for federated learning," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, May 2021, pp. 1–10.
[26] Y. Deng, F. Lyu, J. Ren, H. Wu, Y. Zhou, Y. Zhang, and X. Shen, "AUCTION: Automated and quality-aware client selection framework for efficient federated learning," *IEEE Trans. Parallel Distrib. Syst.*, vol. 33, no. 8, pp. 1996–2009, Aug. 2022.
[27] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009. [Online]. Available: https:// www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[29] Y. Cimtay and E. Ekmekcioglu, "Investigating the use of pretrained convolutional neural network on cross-subject and cross-dataset EEG emotion recognition," *Sensors*, vol. 20, no. 7, p. 2034, Apr. 2020.
[30] B. Chakravarthi, S.-C. Ng, M. R. Ezilarasan, and M.-F. Leung, "EEG-based emotion recognition using hybrid CNN and LSTM classification," *Frontiers Comput. Neurosci.*, vol. 16, Oct. 2022, Art. no. 1019776.
[31] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; Using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012.
[32] F. Lai, X. Zhu, H. V. Madhyastha, and M. Chowdhury, "Oort: Efficient federated learning via guided participant selection," in *Proc. 15th USENIX Symp. Operating Syst. Design Implement. (OSDI)*, 2021, pp. 19–35.
[33] L. Nagalapatti and R. Narayanam, "Game of gradients: Mitigating irrelevant clients in federated learning," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 9046–9054.

CHANG XU received the B.S. degree in automation from LiRen College, Yanshan University, Qinhuangdao, China, in 2020. He is currently pursuing the master's degree in control science and engineering with the Zhejiang University of Technology. He is conducting research with Hangzhou City University. His research interests include joint learning and EEG signal processing.

HONG LIU received the Ph.D. degree from the College of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2006. In 2009, he had a postdoctoral experience in electronic science with Zhejiang University. In 2014, he was a Visiting Researcher with The Hong Kong University of Science and Technology, Hong Kong, China. He is currently an Associate Professor with the School of Information and Electrical Engineering, Hangzhou City University, Hangzhou. His major research interests include system modeling, optimization, and control, and federated learning.

KEXIN LI is currently pursuing the master's degree in electronic information engineering with Hangzhou City University. Her research interest includes artificial intelligence.

WANGLEI FENG received the B.S. degree in electrical engineering and automation from Wenzhou University, Wenzhou, China, in 2022. He is currently pursuing the M.S. degree in control science and engineering with Zhejiang University of Technology. His research interests include federated learning and distributed machine learning.

WEI QI received the Ph.D. degree from the Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China, in 2009. He is currently an Associate Professor with the School of Information and Electrical Engineering, Hangzhou City University, Hangzhou. His major research interests include embedded systems, the Internet of Things, and artificial intelligence.