Adaptive Clustering-Based Model Aggregation for Federated Learning with Imbalanced Data

Dong Wang, Naifu Zhang, and Meixia Tao
Dept. of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China
Emails: {wangdong1217, arthaslery, mxtao}@sjtu.edu.cn

Abstract

In this paper, we focus on one of the key factors which influence the federated learning (FL) performance - data imbalance. By using clustered federated learning (CFL) process to deal with imbalanced data, an adaptive FL clustering algorithm based on cosine similarity (CCS) is proposed to cluster edges devices with the similar data distribution. Then, we propose the convergence bound analysis which based on training data distribution for CFL. Based on the conclusion of convergence bound analysis, a model aggregation algorithm for FL called RFedAvg is proposed. Numerical results obtained using popular convolutional neural network (CNN) model and MNIST dataset show that the proposed framework can achieve faster model convergence rate and higher learning accuracy than the benchmark model aggregation frameworks.

I. Introduction

Federated learning (FL) is a promising edge learning framework that enables multiple edge devices to collaboratively train a common artificial intelligence (AI) model without exchanging raw data [1]. By leveraging the local storage and computing resources at edge devices, FL can preserve data privacy and reduce communication cost compared with traditional cloud-centric learning. In the meantime, it also faces several technical challenges due to the heterogeneity of network environment and user behavior of different edge devices. One unique challenge is that the well-behaved assumption of independent and identically distributed (IID) data samples does not hold in general. Another main challenge, which may occur in cloud-centric learning but becomes more severe in FL, is *data imbalance*, i.e., the amount of data samples per class is not equally distributed. While many previous works have investigated FL under non-IID setting, such as [2] [3], very few has addressed the data imbalance issue. The aim of this work is to introduce a new model aggregation approach for FL with imbalanced data.

Data imbalance is a critical factor affecting the performance of FL. Take the prediction of the next word in typewriting as an example, suppose that device A is used by a young man, and device B is used by an old man. Compared with the old man, the young man types more frequently, which will produce more

training data samples of the next word prediction of device A. Meanwhile, the young man's preferences are very different from the old man's. These factors will generate data imbalance in global dataset of FL and result in the learned global model based on the datasets in devices A and B shows good performance in youth groups. As a result, the overall accuracy of the global model varies significantly.

In [4], it points out that increasing the balance of training dataset can effectively improve the performance of the global model. To deal with data imbalance problem, the common method is to use sampling techniques. These methods operate on the data itself (rather than the model) to increase the balance level of the training dataset, one is oversampling [4] which is to increase the sampling frequency of the minority data samples. Due to the data privacy constraints in FL, this method is not feasible in FL, because it is impossible to deal with data imbalance directly from the data level. An effective way to deal with data imbalance without compromising data privacy in FL is clustering [5]. By clustering edge devices with the same data distribution, and then considering each distribution as a whole, the divergence of local upload gradients can be reduced. Meanwhile, clustering can further reduce the complexity of the optimization problem. Ghosh et al. [6] proposed a clustering approach which depends on ℓ_2-distance to determine the distribution similarity of the devices. A clustering algorithm based on cosine similarity was proposed in [7]. However, in [6], [7], their clustering algorithm cannot effectively adapt to the high data heterogeneity scenario. Moreover, their clustering algorithm do not have the characteristic of adaptive, and it will degrade the clustering performance with the number of iterations increases. The existing work on clustered federated learning (CFL) usually focuses on multi-task learning [6] [7]. By clustering edge devices with same data distribution, and providing customized models for devices in different clusters, so as to solve the problem of data imbalance in FL. However, these well-trained models are not generalizable.

Motivated by the above issue, we study the clustering local gradients and model aggregation in FL taking the data imbalance into account. Different from [6] [7], we focus on trained a well-performance single global model, the purpose is to balance the data contribution for global model by designing the weight of each cluster's updated gradient, then improve the generalization of the global model. First, we design an adaptive clustering algorithm to cluster the edge devices with the similar local gradient that adapts to the time varying of the divergence of local gradient. Specifically, we design an adaptive threshold for cosine similarity clustering algorithm to cluster these upload local gradients. Second, we propose a new CFL model aggregation algorithm by using wight each intra-cluster model to balance the contribution of each class of the training data. Moreover, the weight of each cluster is based on the analysis of the dependence of the convergence rate on the data balance feature of the training dataset. We also use simulation experiments to verify the progressiveness of the proposed clustering algorithm and model aggregation algorithm.

Fig. 1: FL system framework.

II. Adaptive Clustering based Federated Learning

A. Notation and CFL Framework

Consider an FL system as shown in Fig. 1, where K edge devices collaboratively train a common model under the coordination of an edge server through a wireless channel. Let $\mathcal{K} = \{1, 2, ..., K\}$ denote the set of all devices. Let $\mathcal{D}_k = \left\{(x_k^1, y_k^1), (x_k^2, y_k^2), ..., (x_k^{D_k}, y_k^{D_k})\right\}$ denote the local dataset of device k, where x_k^i is the i-th training data sample, y_k^i is the corresponding ground-truth label, and D_k is the size of the local training dataset. Define $\mathcal{D} = \bigcup_{k\in\mathcal{K}} \mathcal{D}_k$ as the global dataset with size $D = \sum_{k\in\mathcal{K}} D_k$.

The goal is to learn a global optimal model $\mathbf{w}^*$ that minimizes its empirical loss function on the global dataset:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \quad \left(\frac{1}{K}\sum_{k\in\mathcal{K}} F_k(\mathbf{w})\right), \tag{1}$$

where $F_k(\mathbf{w}) = \frac{1}{D_k}\sum_{\mathbf{d}\in\mathcal{D}_k} f(\mathbf{w}, \mathbf{d})$ is the prediction error of device k on its training set $\mathcal{D}_k$, and $f(\mathbf{w}, \mathbf{d})$ is the empirical risk over each data sample $\mathbf{d}$.

We consider that the global training dataset $\mathcal{D}$ is imbalanced, i.e., the number of data samples in some classes are significantly larger or smaller than other classes. Let $M \geq 2$ denote the total number of classes in the global dataset $\mathcal{D}$. For each device k, denote D_k^m as the number of data samples in its local dataset $\mathcal{D}_k$ that belong to class m, for $m = 1, \cdots, M$. Then the total number of data samples in the global dataset that belong to class m can be expressed as $D^m = \sum_{k\in\mathcal{K}} D_k^m$. To measure the overall data imbalance level in the global dataset, we define the data balance feature vector as:

$$\mathbf{b} = \left(\frac{D^1}{D}, \frac{D^2}{D}, \cdots, \frac{D^M}{D}\right), \tag{2}$$

In the ideal case with fully balanced data, we have $\mathbf{b} = (\frac{1}{M}, \frac{1}{M}, \cdots, \frac{1}{M})$. Meanwhile, to measure the data imbalance degree of global dataset $\mathcal{D}$, from [8], Let

$$\beta(\mathbf{b}) = \frac{\max(b_m)}{\min(b_m)}, \tag{3}$$

where b_m denote the m-th value in $\mathbf{b}$. Here, larger β indicates that the global dataset $\mathcal{D}$ shows higher data imbalance degree, and the ideal case is fully balanced data, corresponding to $\beta = 1$.

In this paper, we use the clustering operation in CFL to cluster the edge devices with similar local gradient. Then, according to the estimated data imbalance level in each cluster, we propose a weighted method to acquire a reshaped data balance feature vector $\widetilde{\mathbf{b}}$, and the purpose is to reduce $\beta(\widetilde{\mathbf{b}})$. Specifically, let $\mathbf{g}_k^t = \nabla F_k(\mathbf{w}^t)$ denote the local gradient computed by device k at the t-th training round based on its local dataset $\mathcal{D}_k$. Upon receiving all $\mathbf{g}_k^t$, the edge server first adopts a certain clustering algorithm to classify them into a number of clusters according to their similarity as detailed in the next subsection. Let $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_N\}$ denote the resulting set of clusters, with each $\mathcal{C}_n$ consisting of the device indexes of cluster n. Here, N is the number of clusters and it can be different at different iterations. Then, the edge server performs per-cluster gradient aggregation as

$$\mathbf{g}_n^t = \frac{1}{\sum_{k\in\mathcal{C}_n} D_k} \sum_{k\in\mathcal{C}_n} D_k \mathbf{g}_k^t, \quad \forall n = 1, 2, \cdots, N. \tag{4}$$

Finally, the edge server updates the global model $\mathbf{w}^{t+1}$ as 5,

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta \sum_{n=1}^{N} p_n^t \mathbf{g}_n^t, \tag{5}$$

where $\eta > 0$ is the learning and $p_n^t \in [0, 1]$ denotes the weight of cluster n, satisfying $\sum_{n=1}^{N} p_n^t = 1$. We refer to the above FL procedure as weighted CFL. The design of the weighting factors $\{p_n^t\}$ shall be investigated in Section III.

B. Adaptive Clustering

As above analysis, if we can clustering the edge devices with the same data distribution, then, using the weight method to weight the gradient in each cluster can improve the global data balance feature. However, for the edge server, the local data distribution is unknown, and the method of estimating the local data distribution will destroy the data privacy of the edge device. Considering that each local gradient is obtained by learning from its local dataset. So, in this paper, we cluster the local gradients at each training iteration by their cosine similarity. The cosine similarity between any pair of local gradients $\mathbf{g}_k^t$ and $\mathbf{g}_j^t$ in round t is given by:

$$\alpha_{k,j}^t = \frac{< \mathbf{g}_k^t, \mathbf{g}_j^t >}{\|\mathbf{g}_k^t\| \|\mathbf{g}_j^t\|}, \tag{6}$$

where $< \cdot >$ is the inner product and $\| \cdot \|$ is the Euclidean norm. The larger value of $\alpha_{k,j}^t$ indicates that $\mathbf{g}_k^t$ and $\mathbf{g}_j^t$ are more similar, and vice versa. In the special case when $k = j$, one has $\alpha = 1$.

Note that the local gradient similarity varies over iterations and thus adaptive clustering is needed. Intuitively, at the beginning of the FL process, even with imbalanced data, all the local gradients are similar to each other. However, when the global model converges gradually, the local gradients become more divergent. This means that the similarity value $\alpha_{k,j}^t$ would decrease in general as the iteration t increases. Thus, we introduce an adaptive similarity threshold for clustering, denoted as ϵ_t, at each training round t, which is defined as,

$$\epsilon_t = \cos\left(\theta + \left(\frac{\pi}{2} - \theta\right)\frac{l_0 - l_t}{l_0 - l_*}\right). \tag{7}$$

Here, θ is an initial cosine angle close to 0, l_0, l_t and l_* are the initial value, current value and desired value, respectively, of the global loss function. By (7), the similarity threshold is adaptive to the current value of the global loss function. In specific, during the training process, if the current global loss function decreases, the cosine angle increases, resulting in a smaller similarity threshold, and vice versa. When the global model achieves the desired accuracy, that is, the value of the global loss function reaches the target l_*, the threshold becomes 0.

The detail of the proposed adaptive Clustering algorithm based on Cosine Similarity (CCS) is shown in Algorithm 1. In this Algorithm, we define $\mathcal{K}^0$ as the edge devices that have not yet been clustered, i.e., $\mathcal{K}^0 = \mathcal{K}$ (Line 1). Line $4 - 10$ is to find the local gradients that are similar to $\mathbf{g}_k^t$ (including itself), whose device indexed are denoted as $\widehat{\mathcal{K}}_k$. Line $11 - 13$ is to find the gradient $\mathbf{g}_k^t$ with the largest number of the similar local gradients and form these device index of local gradients and itself index as r-th cluster, denoted as $\mathcal{C}_r$. Line 14 is to exclude set $\mathcal{C}_r$ from set $\mathcal{K}^0$. This process is repeat until all the local gradients have been clustered (line 2).

III. Convergence Analysis and Global Model Update Algorithm

In this section, we first analyze the convergence bound of CFL and then propose a global model aggregation algorithm based on the convergence bound.

A. Convergence Bound

For the purpose of analysis, we make some assumptions on the loss functions F_k for each device k in $\mathcal{K}$.

Assumption 1 (L_1-smooth): $F_k(\cdot)$ is smooth with a positive modulus L_1, i.e., for any $\mathbf{u}$ and $\mathbf{v}$, it satisfies,

$$\|\nabla F_k(\mathbf{u}) - \nabla F_k(\mathbf{v})\| \leq L_1\|\mathbf{u} - \mathbf{v}\| \tag{8}$$

Algorithm 1 Adaptive CCS

Initialization: $\mathcal{C} \leftarrow \{\}$, $r = 1$, $\mathcal{K}^0 = \mathcal{K}$.
while $\mathcal{K}^0 \neq \emptyset$ **do**
  for each k in $\mathcal{K}^0$,
  $\hat{\mathcal{K}}_k = \emptyset$,
    for each j in $\mathcal{K}^0$,
      if $\alpha_{k,j}^t \geq \epsilon_t$,
        $\hat{\mathcal{K}}_k$ append j.
      end if
    end for
  end for
  Among $\{\hat{\mathcal{K}}_k, k \in \mathcal{K}^0\}$, find the set with the largest size, denoted as $\hat{\mathcal{K}}_k^*$.
  $\mathcal{C}_r = \hat{\mathcal{K}}_k^*$,
  $\mathcal{C}$ append $\mathcal{C}_r$,
  $\mathcal{K}^0$=$\mathcal{K}^0 \backslash \mathcal{C}_r$,
  $r = r + 1$.
end while
return $\mathcal{C}$.

Assumption 2 (Convex): $F_k(\cdot)$ is convex.

Assumption 3 (G-Bounded gradient): The expected ℓ_2-norm of $\nabla F_k(\mathbf{w})$ is bounded with a positive parameter G, i.e.,

$$\mathbb{E}\left[\|\nabla F_k(\mathbf{w})\|^2\right] \leq G^2. \tag{9}$$

To reveal the impact of data imbalance to the learning performance, we introduce an auxiliary model parameter $\widehat{\mathbf{w}}^{t+1}$, which is obtained by aggregating the model on the basis of data classes. Namely,

$$\widehat{\mathbf{w}}^{t+1} = \mathbf{w}^t - \eta \sum_{j=1}^{M} \frac{D^j}{D} \mathbb{E}\left[\nabla F^j(\mathbf{w}^t)\right], \tag{10}$$

where $\mathbb{E}\left[\nabla F^j(\mathbf{w}^t)\right]$ is the expectation of loss function of the j-th class in dataset $\mathcal{D}$ under model $\mathbf{w}^t$.

$$\mathbb{E}\left[\nabla F^j(\mathbf{w}^t)\right] = \frac{1}{D^j} \sum_{\mathbf{d} \in \mathcal{D}^j} \nabla F(\mathbf{w}^t, \mathbf{d}). \tag{11}$$

Now $\widehat{\mathbf{w}}^{t+1}$ is an explicit function of the balance feature vector $\mathbf{b}$ defined in (10). To facilitate the convergence bound analysis, we characterize the gap between the auxiliary model parameter $\widehat{\mathbf{w}}^{t+1}$ and the global optimal model parameter $\mathbf{w}^*$ as a function of $\mathbf{b}$:

$$R(\mathbf{b}) = \|\widehat{\mathbf{w}}^{t+1} - \mathbf{w}^*\|. \tag{12}$$

We make the following assumptions on $R(\mathbf{b})$ for any given $\mathbf{w}^t$,

Assumption 4 (Convex): $R(\mathbf{b})$ is convex.

Assumption 5 (L_2-smooth): $R(\mathbf{b})$ is smooth with a positive modulus L_2, i.e., for any $\mathbf{b}$ and $\mathbf{b}'$, it satisfies,

$$R(\mathbf{b}) \leq R(\mathbf{b}') + \nabla R^T(\mathbf{b}')\|\mathbf{b} - \mathbf{b}'\| + \frac{L_2}{2}\|\mathbf{b} - \mathbf{b}'\|^2. \tag{13}$$

Let $\mathbf{b}^*$ denote the optimal balance feature vector that can minimize the gap function $R(\mathbf{b})$, and define $\Gamma_t = R(\mathbf{b}^*)$. After the model aggregation in Section II-A, $\widetilde{\mathbf{b}}$ is show as follows,

$$\widetilde{\mathbf{b}} = \left(\sum_{n=1}^{N} p_n^t \frac{D_{\mathcal{C}_n}^1}{D_{\mathcal{C}_n}}, \sum_{n=1}^{N} p_n^t \frac{D_{\mathcal{C}_n}^2}{D_{\mathcal{C}_n}}, \cdots, \sum_{n=1}^{N} p_n^t \frac{D_{\mathcal{C}_n}^M}{D_{\mathcal{C}_n}} \right). \tag{14}$$

In (14), $\mathcal{D}_{\mathcal{C}_n}$ is the dataset of n-th cluster, $D_{\mathcal{C}_n} = |\mathcal{D}_{\mathcal{C}_n}|$ is size of dataset $\mathcal{D}_{\mathcal{C}_n}$ and $D_{\mathcal{C}_n}^j$ is the number of j-th class data samples in dataset $\mathcal{D}_{\mathcal{C}_n}$.

Then, based on the proposed CFL model aggregation process in Section II-A, we obtain the following theorem.

Theorem 1 (Convergence upper bound): The convergence upper bound of the proposed weighted CFL after t rounds of iterations is given by

$$\begin{aligned} &\mathbb{E}\left[F(\mathbf{w}^{t+1}) - F(\mathbf{w}^*)\right] \\ &\leq \frac{L_1}{2}\left\|\mathbb{E}(\mathbf{w}^t) - \mathbf{w}^*\right\|^2 - \frac{L_1}{2}\eta^2 G^2 \\ &+ L_1 \eta G \left(\Gamma_t + \frac{L_2}{2}\left\|\widetilde{\mathbf{b}} - \mathbf{b}^*\right\|^2 \right). \end{aligned} \tag{15}$$

Proof: The main idea of this proof is to scale the update function of CFL global model by using Cauchy-Schwartz inequality. The detailed proof can be found at.

Remark: Theorem 1 indicates that the upper bound only depends on $\left\|\widetilde{\mathbf{b}} - \mathbf{b}^*\right\|^2$, from above analysis, the value of $\left\|\widetilde{\mathbf{b}} - \mathbf{b}^*\right\|^2$ depends on the data balance feature in each cluster which is fixed, and p_n^t. Therefore, in the next section, we minimize $\left\|\widetilde{\mathbf{b}} - \mathbf{b}^*\right\|^2$ by establishing the optimization problem of p_n^t.

B. Convergence Rate Optimization

From (15) and (14), because $\mathcal{D}_{\mathcal{C}_n}^j$ is fixed, therefore, only p_n^t can influence the convergence bound. Define $\mathcal{P} = (p_1^t, p_2^t, \cdots, p_N^t)$, from the result of Theorem 1, to tighten the upper bound, it is necessary to minimize $\left\|\widetilde{\mathbf{b}} - \mathbf{b}^*\right\|^2$. To solve this problem, the key issue is to obtain $\mathbf{b}^*$. Because the main goal of this paper is to propose a FL global model with strong generalization in data imbalance scenario. In [4], it points out that in data imbalance scenario, the model with strong generalization should fully balance each class of the data samples. That is, the data balance feature $\mathbf{b}$ should close to the ideal case with fully balanced data. So, in this paper, we define $\mathbf{b}^*$ as the fully data balance feature, which is $\mathbf{b}^* = \left(\frac{1}{M}, \frac{1}{M}, \cdots, \frac{1}{M}\right)$, and the problem can be formulated as:

$$\mathcal{Q}_1: \quad \min_{p_i^t} \quad \sum_{j=1}^{M}\left(\sum_{n=1}^{N} p_n^t \frac{D_{\mathcal{C}_n}^j}{D_{\mathcal{C}_n}} - \frac{1}{M}\right)^2 \tag{16a}$$

$$\text{s.t.} \quad 0 \le p_n^t \le 1, n = 1, 2, \ldots, N, \tag{16b}$$

$$\sum_{n=1}^{N} p_n^t = 1. \tag{16c}$$

Clearly, $\mathcal{Q}_1$ is convex and can be solved by a generic convex solver. Note that, to obtain the exact solution of $\mathcal{Q}_1$, we need to know the exact data balance feature $\left\{\frac{D_{\mathcal{C}_n}^j}{D_{\mathcal{C}_n}}\right\}$ for $\forall j$ and $\forall \mathcal{C}_n$, which, however, may not be available due to privacy concern. In this paper, we resort to the histogram method to acquire the difference in the number of $\mathcal{D}_k$ and $\mathcal{D}^*$ in each class of data sample. So we have the following definition,

Definition 1 (Data imbalance level): Let I_k denote the data imbalance level of dataset $\mathcal{D}_k$, and it is defined as,

$$I_k = \sum_{m=1}^{M}\left|\frac{D}{M} - D_k^m\right|. \tag{17}$$

Different from traditional FL, we need each device k participating in global aggregation not only to upload the local gradient but also upload its data imbalance level I_k. Because in any cluster n, the data balance feature between each device is similar, so, it can be considered that for any $k \in \mathcal{C}_n$, the trained data samples of the remaining devices in $\mathcal{C}_n$ can reduce data imbalance level of device k which is defined in (17). Then, for each cluster n, we use the following two steps to approximate the data imbalance level of cluster n.

STEP 1: Select the maximum value of data imbalance level of the device in n-th cluster, which is denote $I_{k^*}^n$.

STEP 2: Let $I_{\mathcal{C}_n} = I_{k^*}^n - \sum_{k \in \{\mathcal{C}_n \setminus k^*\}} D_k$ denote the data imbalance level of n-th cluster.

In [9], it points out, the bias of the weight norm which caused by data imbalance can be reduced by weighting each class leaning model with the reciprocal of the number of each classes data samples. By extending this result to this paper, we use the proportion of $\frac{1}{I_{\mathcal{C}_n}}$ in the sum of all $\frac{1}{I_{\mathcal{C}_n}}$ as the weight of $\widehat{\mathbf{w}}_n^t$ to acquire $\mathbf{w}^{t+1}$, specifically, $p_n^t = \frac{1}{I_{\mathcal{C}_n} \sum_{n=1}^{N} \frac{1}{I_{\mathcal{C}_n}}}$. Numerical results in Section IV show that the performance of the proposed model aggregation method can close to the optimal weight model aggregation method which acquired by the solution of $\mathcal{Q}_1$.

IV. Numerical results

In this Section, we validate the theoretical analysis and the performance of the proposed algorithm by simulation. The experimental platform is a personal computer equipped with a Intel(R) Core(TM)i5-4460 CPU @3.20GHz processor.

A. Experiment Settings

We adopt the open source framework Pytorch and uses real datasets to verify the effectiveness of the proposed RFedAvg algorithm compared to other three model aggregation policies. This experiment uses the MNIST dataset.

At the same time, CNN structure is used in this experiment. Total four layer network structure is used, including two convolution layers and two fully connection layers.

B. Benchmark Algorithms Setting

The scheduling policy used in this experiment is random scheduling policy, where the edge server randomly selects some devices as the current round scheduling device set.

We consider the following benchmarks:

FedAvg: In this scenario, no clustering operation is used. The model aggregation policy uses the traditional FedAvg update process.

Optimal: In this scenario, first we use CCS algorithm to cluster the local gradients, then we use optimizer toolkit to solve the problem $\mathcal{Q}_1$ to obtain the optimal solution $\mathcal{P}$. Note that this scheme is infeasible in practical since the local data distribution is unknown.

K-means: K-means clustering aims to partition the K observations into N ($\leq K$) sets $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_N\}$ so as to minimize the intra-cluster sum of squares (i.e. variance). Formally, the objective is to find:

$$\underset{\mathcal{S}}{\operatorname{argmin}} \sum_{n=1}^{N} \sum_{x \in \mathcal{S}_n} \|x - \delta_n\|^2 = \underset{\mathcal{S}}{\operatorname{argmin}} \sum_{n=1}^{K} |\mathcal{S}_n| Var(\mathcal{S}_n), \tag{18}$$

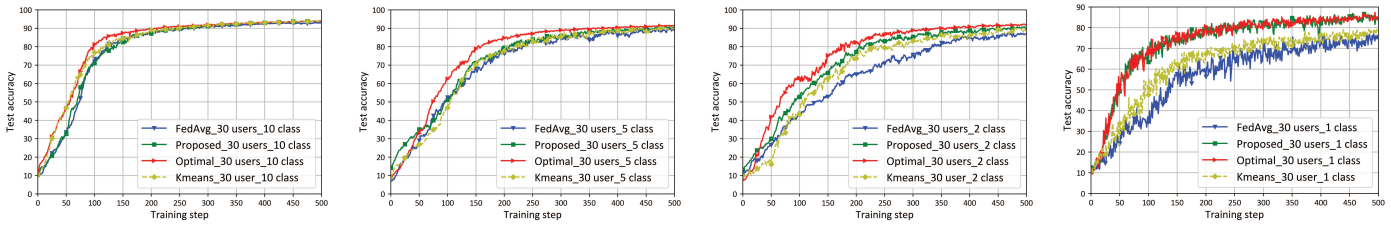
where δ_n is the mean of points in $\mathcal{S}_n$. In this scenario, because the main purpose is to verify the performance of the proposed clustering algorithm CCS, so the proposed model aggregation algorithm is used.

C. Data Distribution Setting for Each Device

In this experiment, we set 100 devices, corresponding to the index of 1-100. Then generate 100 discrete normal distribution numbers with a mean of 5 and variance of 10, meanwhile limit their values to positive integers from 1 to 10 as the index of the device containing the most class of samples. Then we generate $(1 \leq B \leq 10)$ integers that obey the uniform distribution $U \sim (0, 100)$ for each device, and the largest integer corresponds to the number of data samples that the most class of this device have, the remaining $B - 1$ numbers are randomly allocated to the remaining 9 class.

D. Simulation Results

In this experiment, we randomly select 30 devices in each round to participate in global aggregation and use MNIST dataset to verify the proposed algorithm. The simulation results show that the proposed model aggregation algorithm will obtain higher performance gain with the increase of the imbalance degree of training data (In this experiment, higher level degree of imbalance means that each device contains fewer classes of data).



(a) 10 classes. (b) 5 classes. (c) 2 classes. (d) 1 class.

Fig. 2: The test accuracy changes with the increase of iterations in MNIST dataset, 30 devices are randomly selected. (a) Each device has 10 classes data samples. (b) Each device has 5 classes data samples. (c) Each device has 2 classes data samples. (d) Each device has 1 class data samples.

Fig.2(a) demonstrates that the final test accuracy of the proposed algorithm is similar to the three benchmark algorithm, but the convergence rate of the proposed algorithm is lower than K means and optimal. This is because when the training data in each device has the total classes of data, then, each local dataset is almost close to balanced, the direction of local model update in each device is roughly the same. In this scenario, using distance-based clustering will more accurate, but as a result, performance improvement is not obvious. At the same time, no matter what kind of global gradient update policy is used, the model can converge in the end.

Similarly, when each device contains only 5 classes of data data samples, the global data imbalance degree is increase and the instability of gradient update direction also increases. The proposed algorithm can effectively adapt to this change. Fig.2(b) shows that the performance of proposed algorithm is about 0.5% higher than K-means, 1.5% higher than FedAvg and and about 0.5% lower than Optimal. If we further increase the imbalance degree of the training dataset, as shown in Fig.2(c)(d), the convergence rate and final model accuracy will be further improved, about 1% higher than K-means, 3% higher than FedAvg in 2 classes and about 8% higher than K-means, 10% higher than FedAvg, nearly to the Optimal in 1 class. The simulation results also indicate that the proposed clustering algorithm accuracy is higher than K-means.

V. Conclusion

In this paper, we propose a new adaptive clustering algorithm called CCS in FL. For the first time, we derive the convergence bound of CFL from the perspective of data distribution. We also propose a new framework for model aggregation in CFL. This new framework can effectively accelerate the convergence rate of FL system and improve the generalization performance of the global model in data imbalance scenario. Comprehensive experiments using real datasets substantiate the higher performance gain of the proposed model aggregation framework as compared with the benchmark policies.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics. PMLR*, 2017.

[2] X. Li, K. X. Huang, W. H. Yang, S. S. Wang, and Z. H. Zhang, "On the convergence of fedavg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2020.

[3] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "Fedbn: Federated learning on non-iid features via local batch normalization," 2021.

[4] C. X. Ling and C. Li, "Data mining for direct marketing: Problems and solutions," *Kdd*, vol. 98, no. 5, pp. 73–79, 1998.

[5] P Li, J L Huang, K H Zhang, and T T Bi, "Imbalance data classification method based on cluster boundary sampling RF-bagging," in *International Conference on Software Intelligence Technologies and Applications and International Conference on Frontiers of Internet of Things IET, 2015*, pp. 305–311.

[6] A. Ghosh, J. Hong, D. Yin, and K. Ramchandran, "Robust federated learning in a heterogeneous environment," *arXiv preprint arXiv:1906.06629*, 2019.

[7] F. Sattler, K.-R. Mller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multi-task optimization under privacy constraints," *arXiv preprint arXiv:1910.01991*, 2019.

[8] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[9] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Feature transfer learning for face recognition with under-represented data," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5697–5706.

APPENDIX A
PROOF OF THEOREM1

Based on assumption 1, loss function satisfies L_1 smooth, so we have,

$$F(\mathbf{u}) \leq F(\mathbf{v}) + \nabla^T F(\mathbf{v})\|\mathbf{u}-\mathbf{v}\| + \frac{L_1}{2}\|\mathbf{u}-\mathbf{v}\|^2. \tag{19}$$

Let $\mathbf{u} = \mathbf{w}^{t+1}$ and $\mathbf{v} = \mathbf{w}^*$, then we have

$$F(\mathbf{w}^{t+1}) \leq F(\mathbf{w}^*) + \nabla^T F(\mathbf{w}^*)\|\mathbf{w}^{t+1}-\mathbf{w}^*\| + \frac{L_1}{2}\|\mathbf{w}^{t+1}-\mathbf{w}^*\|^2. \tag{20}$$

Because $\mathbf{w}^*$ is the global optimal model, and $F(\cdot)$ is convex, so the first derivative of the loss function with respect to $\mathbf{w}^*$ is equal to $\mathbf{0}$, where $\mathbf{0}$ is all zero vector.

$$\nabla^T F(\mathbf{w}^*) = \mathbf{0}. \tag{21}$$

Substitute (21) into (20), then the inequality can be written as,

$$F(\mathbf{w}^{t+1}) - F(\mathbf{w}^*) \leq \frac{L_1}{2}\left\|\mathbf{w}^{t+1}-\mathbf{w}^*\right\|^2. \tag{22}$$

Because we use CFL to update the global model $\mathbf{w}^{t+1}$, from (4) and (5), then,

$$\begin{aligned} \widehat{\mathbf{w}}_n^t &= \mathbf{w}^t - \frac{\eta}{D_{\mathcal{C}_n}}\sum_{k\in\mathcal{C}_n} D_k \mathbf{g}_k^t \\ &= \mathbf{w}^t - \eta\sum_{k\in\mathcal{C}_n} e_k^t \nabla F(\mathbf{w}^t, \mathcal{D}_k), \end{aligned} \tag{23}$$

where $e_k^t = \frac{D_k}{D_{\mathcal{C}_n}}$.

Substitute (24) into (5), then we have,

$$\begin{aligned} \|\mathbf{w}^{t+1}-\mathbf{w}^*\|^2 &= \left\|\sum_{n=1}^N p_n^t \widehat{\mathbf{w}}_n^t - \mathbf{w}^*\right\|^2 \\ &= \left\|\sum_{n=1}^N p_n^t (\mathbf{w}^t - \eta\sum_{k\in\mathcal{C}_n} e_k^t \nabla F(\mathbf{w}^t, \mathcal{D}_k)) - \mathbf{w}^*\right\|^2 \\ &= \left\|\mathbf{w}^t - \eta\sum_{n=1}^N p_n^t \sum_{k\in\mathcal{C}_n} e_k^t \nabla F(\mathbf{w}^t, \mathcal{D}_k) - \mathbf{w}^*\right\|^2. \end{aligned} \tag{24}$$

Rewrite (4) as follows,

$$\widehat{\mathbf{g}}_n^t = \sum_{k\in\mathcal{C}_n} e_k^t \mathbf{g}_k^t. \tag{25}$$

The global updated gradient $\mathbf{g}^t$ in t-th round equals,

$$\mathbf{g}^t = \sum_{n=1}^N p_n^t \widehat{\mathbf{g}}_n^t. \tag{26}$$

Substitution (25)(26) into (24),

$$\begin{aligned}\|\mathbf{w}^{t+1}-\mathbf{w}^*\|^2 &= \|\mathbf{w}^t-\eta\mathbf{g}^t-\mathbf{w}^*\|^2\\ &= \left\|\mathbf{w}^t-\eta\sum_{n=1}^{N}p_n^t\widehat{\mathbf{g}}_n^t-\mathbf{w}^*\right\|^2\\ &= \|\mathbf{w}^t-\mathbf{w}^*\|^2-2\eta\underbrace{\left\langle\mathbf{w}^t-\mathbf{w}^*,\sum_{n=1}^{N}p_n^t\widehat{\mathbf{g}}_n^t\right\rangle}_{B_1}+\eta^2\underbrace{\left\|\sum_{n=1}^{N}p_n^t\widehat{\mathbf{g}}_n^t\right\|^2}_{B_2}.\end{aligned} \tag{27}$$

Rewrite B_2 as follows,

$$\begin{aligned}B_2 &= \left\|\sum_{n=1}^{N}p_n^t\widehat{\mathbf{g}}_n^t\right\|^2\\ &\le \sum_{n=1}^{N}p_n^t\|\widehat{\mathbf{g}}_n^t\|^2.\end{aligned} \tag{28}$$

Expand B_1 we can get,

$$\begin{aligned}B_1 &= \left\langle\mathbf{w}^t-\mathbf{w}^*,\sum_{n=1}^{N}p_n^t\widehat{\mathbf{g}}_n^t\right\rangle\\ &= \sum_{n=1}^{N}p_n^t\left\langle\mathbf{w}^t-\mathbf{w}^*,\widehat{\mathbf{g}}_n^t\right\rangle\\ &= \sum_{n=1}^{N}p_n^t\left\langle\mathbf{w}^t-\widehat{\mathbf{w}}_n^t+\widehat{\mathbf{w}}_n^t-\mathbf{w}^*,\widehat{\mathbf{g}}_n^t\right\rangle\\ &= \left\langle\sum_{n=1}^{N}p_n^t\left(\mathbf{w}^t-\widehat{\mathbf{w}}_n^t\right),\widehat{\mathbf{g}}_n^t\right\rangle+\left\langle\sum_{n=1}^{N}p_n^t\left(\widehat{\mathbf{w}}_n^t-\mathbf{w}^*\right),\widehat{\mathbf{g}}_n^t\right\rangle\\ &= \eta\sum_{n=1}^{N}p_n^t\|\widehat{\mathbf{g}}_n^t\|^2+\underbrace{\left\langle\sum_{n=1}^{N}p_n^t\widehat{\mathbf{w}}_n^t-\mathbf{w}^*,\widehat{\mathbf{g}}_n^t\right\rangle}_{C_1}.\end{aligned} \tag{29}$$

By Cauchy-Schwarz inequality,

$$\begin{aligned}-C_1 &= -\left\langle\sum_{n=1}^{N}p_n^t\widehat{\mathbf{w}}_n^t-\mathbf{w}^*,\widehat{\mathbf{g}}_n^t\right\rangle\\ &\le \left\|\sum_{n=1}^{N}p_n^t\widehat{\mathbf{w}}_n^t-\mathbf{w}^*\right\|\left\|\widehat{\mathbf{g}}_n^t\right\|.\end{aligned} \tag{30}$$

Combine (27)(28)(29)(30), then we can get,

$$\begin{aligned}\|\mathbf{w}^{t+1}-\mathbf{w}^*\|^2 &\leq \|\mathbf{w}^t-\mathbf{w}^*\|^2-2\eta^2\sum_{n=1}^{N}p_n^t\|\widehat{\mathbf{g}}_n^t\|^2\\ &+2\eta\left\|\sum_{n=1}^{N}p_n^t\widehat{\mathbf{w}}_n^t-\mathbf{w}^*\right\|\|\widehat{\mathbf{g}}_n^t\|+\eta^2\sum_{n=1}^{N}p_n^t\|\widehat{\mathbf{g}}_n^t\|^2 \\ &=\|\mathbf{w}^t-\mathbf{w}^*\|^2-2\eta^2\sum_{n=1}^{N}p_n^t\|\widehat{\mathbf{g}}_n^t\|^2+\eta\left\|\sum_{n=1}^{N}p_n^t\widehat{\mathbf{w}}_n^t-\mathbf{w}^*\right\|\|\widehat{\mathbf{g}}_n^t\|.\end{aligned} \tag{31}$$

Then (22) can be rewritten as follows,

$$F(\mathbf{w}^{t+1})-F(\mathbf{w}^*)\leq\frac{L_1}{2}\|\mathbf{w}^t-\mathbf{w}^*\|^2-L_1\eta^2\sum_{n=1}^{N}p_n^t\|\widehat{\mathbf{g}}_n^t\|^2+\frac{L_1\eta}{2}\left\|\sum_{n=1}^{N}p_n^t\widehat{\mathbf{w}}_n^t-\mathbf{w}^*\right\|\|\widehat{\mathbf{g}}_n^t\|. \tag{32}$$

Based on Assumption 4 and Assumption 5, and combine (11) we have,

$$\begin{aligned}\mathbb{E}\left(\left\|\sum_{n=1}^{N}p_n^t\widehat{\mathbf{w}}_n^t-\mathbf{w}^*\right\|\right)&=\left\|\sum_{n=1}^{N}p_n^t\mathbb{E}(\widehat{\mathbf{w}}_n^t)-\mathbf{w}^*\right\| \\ &=\left\|\sum_{n=1}^{N}p_n^t\left(\mathbb{E}(\mathbf{w}^t)-\eta\frac{1}{D_{\mathcal{C}_n}}\sum_{\mathbf{d}\in\mathcal{D}_{\mathcal{C}_n}}\mathbb{E}\left(\nabla F(\mathbf{w}^t,\mathbf{d})\right)\right)-\mathbf{w}^*\right\| \\ &=\left\|\mathbb{E}(\mathbf{w}^t)-\mathbf{w}^*-\eta\sum_{m=1}^{M}\sum_{n=1}^{N}p_n^t\left(\frac{D_{\mathcal{C}_n}^m}{D_{\mathcal{C}_n}}\mathbb{E}\left(\nabla F(\mathbf{w}^t,\mathbf{d}_n^m)\right)\right)\right\|,\end{aligned} \tag{33}$$

where $D_{\mathcal{C}_n}^m$ denote the total number of m-th class data samples in m-th cluster's dataset $\mathcal{D}_{\mathcal{C}_n}$, and $\mathbf{d}_n^m$ denote the training data samples of m-th class in n-th cluster's dataset $\mathcal{D}_{\mathcal{C}_n}$.

In (14), we denote the reshaped data balance feature vector $\widetilde{\mathbf{b}}$ as follows,

$$\widetilde{\mathbf{b}}=\left(\sum_{n=1}^{N}p_n^t\frac{D_{\mathcal{C}_n}^1}{D_{\mathcal{C}_n}},\sum_{n=1}^{N}p_n^t\frac{D_{\mathcal{C}_n}^2}{D_{\mathcal{C}_n}},\cdots,\sum_{n=1}^{N}p_n^t\frac{D_{\mathcal{C}_n}^M}{D_{\mathcal{C}_n}}\right). \tag{34}$$

Then, combine (33) and (34), and from (2) and (12), (33) can be written as,

$$\mathbb{E}\left(\left\|\sum_{n=1}^{N}p_n^t\widehat{\mathbf{w}}_n^t-\mathbf{w}^*\right\|\right)=R\left(\widetilde{\mathbf{b}}\right), \tag{35}$$

Based on Assumption 5, the following inequality is obtained,

$$\begin{aligned}R\left(\widetilde{\mathbf{b}}\right)&\leq R(\mathbf{b}^*)+\nabla R(\mathbf{b}^*)\|\widetilde{\mathbf{b}}-\mathbf{b}^*\| \\ &+\frac{L_2}{2}\|\widetilde{\mathbf{b}}-\mathbf{b}^*\|^2 \\ &=R(\mathbf{b}^*)+\frac{L_2}{2}\|\widetilde{\mathbf{b}}-\mathbf{b}^*\|^2.\end{aligned} \tag{36}$$

Combine (36) and Assumption 3,

$$\begin{aligned}\mathbb{E}(F(\mathbf{w}^{t+1}) - F(\mathbf{w}^*)) \leq \frac{L_1}{2}\|\mathbb{E}(\mathbf{w}^t) - \mathbf{w}^*\|^2 - \frac{L_1}{2}\eta^2 G^2 \\ + L_1 \eta G \left(\Gamma_t + \frac{L_2}{2}\|\widetilde{\mathbf{b}} - \mathbf{b}^*\|^2\right).\end{aligned} \tag{37}$$

Theorem 1 is proved.