Migrating Federated Learning to Centralized Learning with the Leverage of Unlabeled Data

Tianqing Zhu (✉ tianqing.zhu@ieee.org)
China University of Geosciences

Xiaoya Wang
China University of Geosciences

Wei Ren
China University of Geosciences

Dongmei Zhang
China University of Geosciences

Ping Xiong
Zhongnan University of Economics and Law

Research Article

Keywords: Federated Learning, Non-IID, Teacher-Student, Ensemble Learning

Posted Date: April 12th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1544372/v1

Migrating Federated Learning to Centralized Learning with the Leverage of Unlabeled Data

Xiaoya Wang[1], Tianqing Zhu[1*], Wei Ren[1], Dongmei Zhang[1] and Ping Xiong[2]

[1]School of Computer Science, China University of Geosciences, Wuhan, 430074, Hubei, China.
[2]School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan, 430073, Hubei, China.

*Corresponding author(s). E-mail(s): tianqing.zhu@ieee.org;
Contributing authors: wangxiaoya@cug.edu.cn;
weirencs@cug.edu.cn; cugzdm@foxmail.com;
pingxiong@zuel.edu.cn;

Abstract

Federated learning carries out cooperative training without local data sharing, the obtained global model performs generally better than independent local models. Benefiting from the free data sharing, federated learning preserves the privacy of local users. However, the performance of the global model might be degraded if diverse clients hold non-IID training data. This is because the different distributions of local data lead to weight divergence of local models. In this paper, we introduce a novel teacher-student framework to alleviate the negative impact of non-IID data. On the one hand, we maintain the advantage of the federated learning on the privacy-preserving, and on the other hand, we take the advantage of the centralized learning on the accuracy. We use unlabeled data and global models as teachers to generate a pseudo-labeled dataset, which can significantly improve the performance of the global model. At the same time, the global model as a teacher provides more accurate pseudo labels. In addition, we perform a model rollback to mitigate the impact of latent noise labels and data imbalance in the pseudo-labeled dataset. Extensive experiments have verified that our teacher ensemble performs a more robust training. The empirical study verifies that the reliance on the centralized pseudo-labeled data enables the global model almost immune to non-IID data.

Keywords: Federated Learning, Non-IID, Teacher-Student, Ensemble Learning

1 Introduction

Federated learning is a promising decentralized private learning, in which all local users cooperatively and decentrally train a global model without exposing local private data [1]. Each client trains the local model and sends parameters or gradients to the server for a global aggregation. The server returns a global model to clients for more training iterations. Compared with the centralized learning, federated learning can preserve clients' privacy with taking the cost of a modest accuracy loss [2]. However, data from clients may not always have been independent in practice. The non-IID (non-independent and identical distributed) data causes clients to get local models with weight divergence, and deteriorate the performance of the corresponding aggregated global model with a further accuracy loss on the learning model [3, 4]. The non-IID data issue is also called the problem of statistical heterogeneity.

The leverage of unlabeled data is one of the approaches to alleviate the adverse effects of non-IID data [5–13]. It has some advantages: The unlabeled data are easier to be collected than the labeled data; fewer privacy concerns might be raised on unlabeled data. Some literature [5–8] leverages unlabeled data on clients with local semi-supervised learning methods to improve federated learning's performance. Local semi-supervised learning relies on clients to make use of unlabeled data, which puts more storage and computation pressure on local clients. This might be unacceptable for resource-limited devices. Others [9–13] have an opposite hypothesis that the unlabeled data is on the server. The data is used for auxiliary centralized learning and alleviating the weight divergence of local models. However, most of them have limited leverage over the data when encountering non-IID data. Federated distillation [9–11] communicates logits between server and clients and aggregates the logits to generate pseudo labels for unlabeled data. The aggregation accuracy of the logits relies on the size of the unlabeled dataset and has a significant decline when non-IID data is encountered, resulting in a decrease in model performance. Ensemble learning on unlabeled data with the local models and the global model as the base learners [12, 13] has an unstable pseudo-labeling accuracy due to the existence of biased local models.

In this work, we introduce a novel *teacher-student* framework to alleviate the negative impact of non-IID data. On the one hand, we maintain the advantage of the federated learning on the privacy-preserving, and on the other hand, we take the advantage of the centralized learning on the accuracy. We focus on the utilization of unlabeled data on the server side to improve the performance of federated learning. We collect the aggregated global models by time series as teachers (see Figure 1) and give pseudo labels to the unlabeled data in an ensemble way.

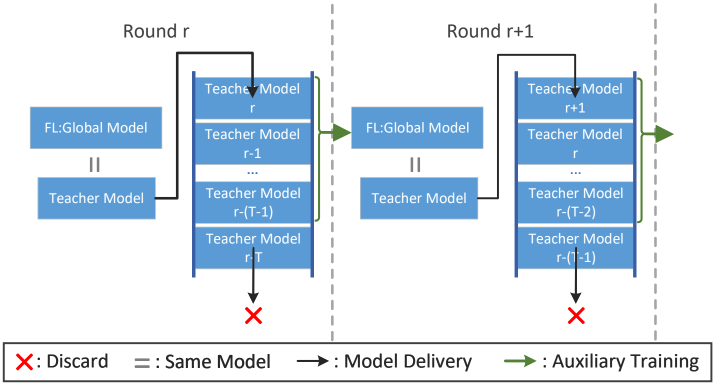


Fig. 1 We collect global models in different training rounds to form an ensemble with T teachers. At the same time, the teacher ensemble promotes the next round of global model training.

Different from previous work, we change the generation of the global model. As the result of training with the pseudo-labeled data, we obtain a better global model and a better teacher model to update a base learner in the ensemble. The pseudo-labeling of the ensemble migrates the knowledge contained in the distributed labeled data to the centrally collected unlabeled data, which has a similar feature space to local data, a more uniform distribution, and more accurate pseudo labels. We make the distributed federated learning has a closer performance to centralized learning.

Compared with pseudo-labeling by a single model, teacher ensemble can provide more accurate pseudo labels. The key to ensemble learning is the base learners with good performance and diversity. The diversity is the differences between the learners especially the differences in output [14, 15]. As multiple models join in the teacher ensemble, combining their different outputs through a specific voting method produces more accurate output. Due to the randomness of the model training process, at least before the global model converges, the collection of global models by time series still maintains diversity and gives play to the advantages of ensemble learning. In addition, the collection method ensures the consistency and stability of the global model. We show that the ensemble of aggregated global models can make the pseudo-labeling maintain high accuracy and confidence, even if on the non-IID data.

We evaluated our method under different local distributions on CIFAR-10/100 with a CNN and a ResNet-8. Experiments show that our method improved the utilization of unlabeled data from the perspective of the quantity of valid data, label accuracy, and distribution, and achieved higher test accuracy. With the knowledge learned from local labeled data transferring to the unlabeled data, we achieved comparable performance with centralized supervised learning with the same data size as the unlabeled data.

2 Preliminary

Federated learning is a distributed machine learning framework. It is proposed to use the computation capability of edge devices to collaboratively train a global model on a server, which performs better than any independent local model generally. At the same time, each local device does not send personal data to the server and keeps it locally to protect data privacy. The collaboration between the various local devices is embodied in the aggregation in federated learning. Federated Averaging (FedAvg) [1] is one of the most commonly used aggregation methods, it is proposed to use a weighted average of all local models' parameters as the aggregated global model. The weight is proportional to the amount of data on the local devices.

We consider there are N local clients and M of which participate in each training round. Each model is parameterized by w, a labeled dataset is denoted as $D = \cup(x, y)$ with distribution P, where $x \to \mathbb{R}^d$ is an input instance in d dimensional feature space, $y \in \{1, 2, \ldots C\}$, C is the number of categories for classification. Given a predictor $f : x \to y'$ and a loss function $l : y' \times y \to \mathbb{R}$, the risk of a model parameterized by w on a classification task on D is defined as $L(D;\ w) := \mathbb{E}_{(x,y)\in D}\left[l(f(x;\ w), y)\right]$.

Federated Learning. There are two components in federated learning: multiple clients with local models and a server with a global model. The server provides an initial model to the clients, while the clients apply their private data to update the model and submit the model parameters to the server. As the aggregated global model is used as a new initial model, federated learning starts a new round of training. This procedure will iterate for R rounds.

FedAvg [1] is a standard aggregation method in federated learning. With model parameters exchange, clients collaboratively train a global model without exposing their data. By averaging the parameter values of multiple models, FedAvg aggregates the multiple models into a single model to complete the aggregation process in federated learning.

In local training, each client i performs supervised learning with its labeled private data D_i after being initialized with a global model:

$$w_i = w_i - \eta \frac{\partial L\left(D_i;\ w_i\right)}{\partial w_i} \tag{1}$$

where w_i is the parameter of the local model on client i, η is the learning rate, $L(\cdot)$ is the loss function. We use Cross Entropy Loss as the loss function in this work. In FedAvg, the server averages the local models to obtain an updated global model:

$$w_g = \frac{1}{\sum_{i\in S}|D_i|}\sum_{i\in S}|D_i| \cdot w_i \tag{2}$$

where S is the collection of clients who participate in the training, w_g is the averaged global model. The proportion of the local data to the total training data is used as the weight of the local model parameters. In the standard process of federated learning, the global model w_g will then be sent back to clients for the next round of training, and so on for finite rounds to complete the whole learning process as shown in Figure 2(a).

Non-IID Data. In contrary to ideal identically and independently distributed data, non-IID data is a real and natural existence. Data generated by different users, from different geographic locations and in different time windows leads to the non-identicalness of data distributions [16]. With P_i and P_j denoting the data distributions of any two clients i and j, x is a sample with a label y, the non-IID data in federated learning typically refers to the differences between P_i and P_j. From the perspective of conditional distribution, i.e. $P(x, y) = P(y|x)P(x) = P(x|y)P(y)$, there are four main ways to show the differences: a. different $P(y|x)$ with same $P(x)$; b. different $P(x)$ with same $P(y|x)$; c. different $P(x|y)$ with same $P(y)$; d. different $P(y)$ with same $P(x|y)$. In addition, the unbalancedness in different clients' data also is a kind of non-IID setting. In fact, data distributions between clients may contain a more complex mixture of these effects.

Most of the existing work on simulating non-IID data focuses on making a different $P(y)$. Two methods are mainly used: i. distribute data directly in proportion to a given percentage according to established data preferences of different clients; ii. follow Dirichlet distribution randomly distribute data to clients. The former is much easier to operate.

When encountered with non-IID data, the distribution differences between local data will lead to weight divergence of local models, and the model obtained by FedAvg will deviate from the ideal model. Introducing additional data to training is a regular approach to alleviate the adverse effects of non-IID data. Especially, unlabeled data is much easier to be collected than labeled data, especially when the unlabeled data contains less privacy and no relevance to a certain user. Since both of them are produced by users, the distribution of the collected unlabeled data is somewhat consistent with the joint distribution of all local data.

3 Migrating decentralized federated learning to centralized learning

3.1 Overview of the teacher-student framework

In this section, we propose a novel teacher-student framework to migrate decentralized federated learning to centralized learning and alleviate the impact of non-IID data in federated learning. With a teacher ensemble, we migrate the knowledge of feature-to-label mapping from distributed labeled data to centralized unlabeled data. With auxiliary training, we aim to make the decentralized federated learning have a closer performance to centralized learning.

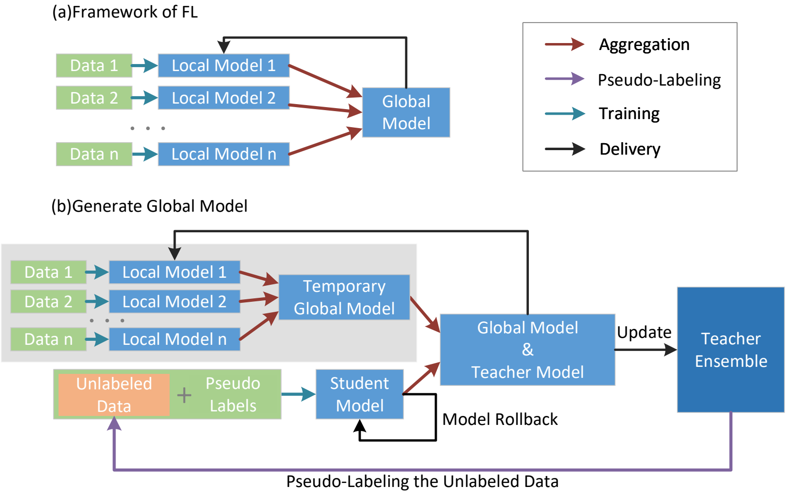


Fig. 2 Overview of the method. (a) shows the framework of federated learning, we use classical FedAvg as the aggregation method. (b) shows the way we generate the global model.

In federated learning, the statistical heterogeneity of local data caused the weight divergence of local models, thus deteriorating the performance of the global model. Rather than sending the aggregated model back directly, we change the way to obtain the global model. Figure 2 shows the overview of our method. We take the averaged model as a temporary global model and the student model learned from the pseudo-labeled data as an auxiliary to generate a teacher, while as a new global model. The teacher ensemble is consist of T adjacent global models as shown in Figure 1. The pseudo-labeling of the ensemble generates the pseudo labels, which are used as supervision to learn a student model. Generally, the prediction of multiple teacher models in an ensemble can achieve better accuracy than that of a single model. The generated teacher instead of the flimsy averaged model will be sent back to local clients for their next round of training. To mitigate the impact of possible noise and imbalance of the pseudo-labeled data, we reset the global model to varying degrees before the next global update.

We assume that the server can meet additional storage and computation requirements. The above process would be repeated finite times until it converged to an ideal student model. We summarize the whole training process as Algorithm 1.

3.2 Pseudo-Labeling the Unlabeled Data

We collect unlabeled data and obtain the pseudo labels by the pseudo-labeling of the teacher ensemble. For the sake of simplicity, we do not use the logits

Algorithm 1 Illustration of D2C-FL. We collect global models in federated learning to form a teacher ensemble. With pseudo-labeling an unlabeled dataset with the ensemble, we perform further training on the centralized pseudo-labeled data and re-update the global model.

```
Require: The initial model weights $w_{init}$; unlabeled data $D_u$; size of teacher ensemble $T$; number of participants in each round $M$; local labeled data $D_i$; learning rate $\eta$.
Ensure: global model weights $w_g^{'}$.
Initialization:
$w_g^{'} \leftarrow w_{init}$
Get Teacher Ensemble ready with $T * w_{init}$.

procedure SERVER
    for $r \leftarrow 1$ to $R$ do
        Sample $M$ clients as $S$ to participate in training.
        for each client $c_i \in S$ do
            $w_i \leftarrow w_g^{'}$;
            $w_i \leftarrow Update(w_i, D_i, \eta)$;                    ▷ Equation (1)
            Send $w_i$ to the server;
        end for
        $w_g \leftarrow$ Aggregate the local models;                   ▷ Equation (2)
        Get pseudo-labeled dataset $D_{pseudo}$;                        ▷ Equation (3)-(9)
        $w_{stu} \leftarrow Update(w_{stu}, D_{peudo}, \eta)$;             ▷ Equation (10)
        Get a new teacher $w_{teacher}$ and rollback student model $w_{stu}$;    ▷ Equation (11) and Equation (12)
        Replace the most corrupt model in the Teacher Ensemble with $w_{teacher}$;
        $w_g^{'} \leftarrow w_{teacher}$
    end for
end procedure
```

output of the ensemble as soft labels to learn a student model as what would be done in distillation. We directly use the assigned class value by logits as the hard labels, then the dataset composed of pseudo labels and unlabeled data has the same form as the dataset used in general supervised learning.

To get a single-value label for a sample, we assign the class value with the highest probability among all the teachers as the pseudo label. The decision relying on the highest probability indicates that the most confident teacher in the ensemble determines the label of the sample. It gives full play to the high-quality teacher in the ensemble. The pseudo-labeling process can be expressed as:

$$y_{max\_pseudos} = \{\arg\max\{f_i\left(x_u;\ w_i\right)\} \mid i \in \{1, 2, \cdots, T\}\} \tag{3}$$

$$y_{max\_probs} = \{\max\{f_i\left(x_u;\ w_i\right)\} \mid i \in \{1, 2, \cdots, T\}\} \tag{4}$$

$$y_{pseudo} = \begin{cases} y_{max\_pseudos}\left[\arg\max\{y_{max\_probs}\}\right] & \max\{y_{max\_probs}\} \geq \tau \\ \text{NULL} & \max\{y_{max\_probs}\} < \tau \end{cases} \tag{5}$$

where x_u is a sample of unlabeled data, $f_i(\cdot)$ is the output of teacher i in teacher ensemble, i.e. the logits, w_i is the parameters of the teacher model. $y_{max\_pseudos}$, y_{max_probs}, y_{pseudo} are the labels given by T teachers, the max prediction probabilities of the T teachers and the final given pseudo label for sample x_u respectively. The threshold τ is used to determine whether the data qualifies for subsequent training. If the highest probability reaches the threshold τ, the corresponding label will be assigned to the sample as its pseudo label and further learned by a student model. On the contrary, if the highest probability is lower than the threshold τ, the sample will be discarded from the subsequently generated dataset. The discard of the data with low prediction probability will mitigate the over-fitting of the student model to the noise labels.

Averaging the predictions of teachers as the output of ensemble is another way to get pseudo labels and it is a commonly used method in ensemble learning. The corresponding pseudo-labeling process can be expressed as:

$$y_{avg\_pseudo} = \arg\max\left\{\frac{1}{T}\sum_{i=1}^{T}\left\{f_i\left(x_u;\ w_i\right)\right\}\right\} \tag{6}$$

$$y_{avg\_prob} = \max\left\{\frac{1}{T}\sum_{i=1}^{T}\left\{f_i\left(x_u;\ w_i\right)\right\}\right\} \tag{7}$$

$$y_{pseudo} = \begin{cases} y_{avg\_pseudo} & y_{avg\_prob} \geq \tau \\ \text{NULL} & y_{avg\_prob} < \tau \end{cases} \tag{8}$$

where y_{avg_pseudo}, y_{avg_prob} are the labels and corresponding probabilities given by the average output of T teachers.

After all the pseudo labels are checked by the threshold τ and assigned to the unlabeled data, we get a new labeled dataset finally:

$$D_{pseudo} = \cup\left(x_u, y_{pseudo}\right) \tag{9}$$

To distinguish it from the local data D_i, we use D_{pseudo} to denote it. Then the student model will be updated through

$$w_{stu} = w_{stu} - \eta\frac{\partial L\left(D_{pseudo};\ w_{stu}\right)}{\partial w_{stu}} \tag{10}$$

where w_{stu} is the parameters of student model. The student model will serve as an auxiliary to generate a new global&teacher model and update the teacher ensemble.

3.3 Update the Teacher Ensemble

We update the teacher ensemble as shown in Figure 1. In the first T rounds, the student model remains the same as the averaged global model. Therefore, the first T collected teachers are exactly the first T averaged global models. As we do not want the randomly initialized teachers to give bad predictions on the unlabeled data before the target number of available models gathering in the ensemble, the collection of teachers requires T rounds of preparation.

Once the target number of teachers is reached, we are able to perform pseudo-labeling to get a new dataset. Therefore, we can train a new student model which is different from the averaged model. A new teacher will be generated by

$$w_{teacher} = w_g^{'} = \alpha \cdot w_g + (1-\alpha) \cdot w_{stu}, \text{ where } \alpha \in [0,1] \tag{11}$$

where $w_{teacher}$ exactly is the parameters of the generated teacher, $w_g^{'}$ means the teacher will be sent back to clients as a new global model for a new training round. The α is the weight of the averaged global model. When $\alpha = 0$, the student model will become the new teacher and be sent to clients directly which is consistent with [12]; when $\alpha = 1$, the averaged global model will be the new teacher, the unlabeled data and the student model lose their auxiliary values. Neither the student nor the averaged global model alone is the best choice for the teacher.

The most corrupt teacher in the ensemble will be replaced by the new teacher as the update of the ensemble. As long as the averaged model or the student model does not converge, our teacher ensemble can maintain its diversity which was important for ensemble learning.

3.4 Model Rollback

To alleviate the confirmation bias [17] of the global model on the latent noise labels and data imbalance in the pseudo-labeled data, we add randomness to each round of training with the model rollback.

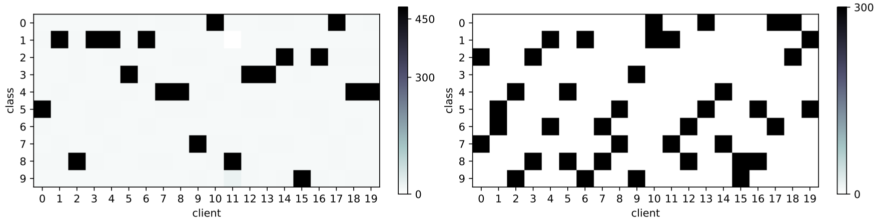
Rollback of student. If the teachers in the ensemble have a worse alignment between the prediction probabilities and test accuracy, the threshold τ cannot filter out the noisy data in pseudo-labeling. To this end, we perform a rollback on the student model for its iterative updating. Specifically, we take a weighted average of the student model and a new randomly initialized model and assign it to the student model to be further updated.

$$w_{stu} = \beta \cdot w_{stu} + (1-\beta) \cdot w_{\text{init}} \text{ , where } \beta \in [0,1] \tag{12}$$

where w_{init} is the new randomly initialized model. The β is the weight of the student model in the re-initialization. $\beta = 0$ means resetting student model to a completely random model at the beginning of each training round; $\beta = 1$ means withdrawing the rollback of student. When learning with pseudo-labeled data, resetting the model in each training round can prevent the model from over-fitting the wrong labels and get a more stable convergence.

Rollback of teachers. As detailed in Subsubsection 3.3, each updated student model must have a combination with the averaged temporary global model to generate a new teacher. We regard the combination as a rollback of the teacher. Compared with the student as a teacher directly, we regress the teacher to an intermediate value between the student model and the temporary global model. Even though the student model draws a much richer and even dataset, it is not appropriate to make it alone becoming a teacher directly. As using student model to re-pseudo-label the dataset it draws on, the wrong predictions will be learned and reinforced again, thus deteriorating the performance of the global model [18].

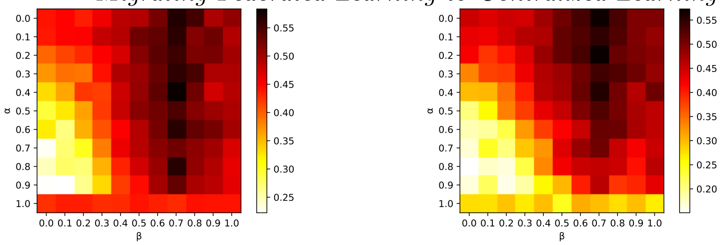
Analysis. We validated our method with different α and β on CIFAR-10 with a simple CNN (2 convolutional layers). We considered 100 clients, 10 of them participated in training with 20000 local labeled data and 10000 unlabeled data. Each client performed supervised learning with the same amount of 600 samples. We simulated two different non-IID settings, the data distributions of 20 randomly sampled clients can be seen in Figure 3. The test accuracy with different α and β can be seen in Figure 4. It is shown that neither training based on a randomly initialized model (i.e. $\beta = 1$) nor the student in the last round (i.e. $\beta = 0$) alone can achieve the best test accuracy. Also, making the student directly as a teacher (i.e. $\alpha = 0$) is not necessarily the optimal solution, which is also possible to have α be an intermediate value. But taking the model obtained by average aggregation as a teacher (i.e. $\alpha = 1$) is definitely not the optimal solution. Since when $\alpha = 1$, the global model has nothing to do with the student model. Furthermore, the global model will not be affected by the β at all as shown at the bottom of Figure 4. $\alpha = 1$ means the student model becomes a redundant model on the server side and local clients learned nothing from the unlabeled data. The rollback of the student and teachers realizes a better performance and of which on student model has more effect on the global model obviously.



(a) 0.8-bias

(b) 2-class

Fig. 3 Illustration of the two non-IID settings on CIFAR-10. 0.8-bias: 80% of data on each client belongs to its preference class and the rest of the data belongs to the other classes uniformly. 2-class: The data on each client evenly belongs to two classes.



(a) 0.8-bias

(b) 2-class

Fig. 4 Test accuracy with different α and β on CIFAR-10 with CNN.

3.5 Discussion on the method

Cost analysis. We exchange model parameters between server and clients in each communication round, which is consistent with general methods regardless of communication cost. As the fact that the leverage of unlabeled data makes a drastic increase of model's test accuracy and faster convergence than FedAvg, it reduces required communication rounds and indirectly saves communication cost. Additional computing and storage costs resulting from additional training of the student model are borne by the server and do not burden local clients.

Unlabeled data. The assumption of large amounts of unlabeled data that do not involve privacy may hinder our method applying in a wider range of scenarios, as the disputes in Subsubsection 5.3. Some generative methods have been used in federated learning to generate synthetic data to assist model learning. Ideally, the generative methods may only be used to assist global learning and add no additional computing and storage costs to clients.

Robustness to attacks. Since we have migrated the focus of model learning from decentralized labeled data to centralized unlabeled data, it's intuitive that our method will be more robust against malicious clients, it needs to be verified.

4 Experimental evaluation

4.1 Experimental setup

Datasets and models. In the experiment, we evaluate the classification on CIFAR-10/100 [19]. CIFAR-10/100 contains 50000 training images and 10000 test images with 10/100 classes. To highlight the guidance of unlabeled data to the whole training performance, we distribute more data to the server as unlabeled. We assign 20000 of training data to the clients, and the rest of the training data is assigned to the server along with the whole test set. 80% of the data on the server side is used as unlabeled data to train and 20% of the data is used to test the global model. We use a CNN Net with two convolutional and three fully-connected layers, which has the same structure as LeNet-5 [20]. We also compare the performances of ResNet-8 following [12] for a more complex model.

Federated learning settings. Referring to the setting in [21], we consider that there are 100 clients locally and 10 clients are selected in each round to participate in the training. Each client keeps 600/100 pieces of data which is taken from the 20000 data pool according to the client's special data distribution preference for CIFAR-10/100. When the data in the pool is deficient, it will be supplemented with the taken data. As a result, a certain amount of data duplication between clients.

Heterogeneity settings. In addition to the IID setting, we also consider two classical non-IID settings following [22]: α-bias, 2-class. **IID**: The data on each client is distributed evenly, with all the clients fetching the same amount of data from each class in the local data pool. α**-bias**: Each client has a data preference for a certain class to the degree of α. This means there will be an alpha ratio of data belonging to the preferred class and the rest of the data belongs to the other classes uniformly. In this experiment, we fixed α at 0.8 for corresponding training and evaluations. 2**-class**: The data on each client evenly belongs to two classes.

Training Details. We set both local epochs and global epochs to 5, the batch size is 128, the initial learning rate of local training is 0.001, and decay it by 0.95 in each epoch. Referring to the suggestion of [13], we directly set the size T of the teacher ensemble to 10 for experiments. The threshold τ for filtering unlabeled data is set to 0.75. The optimal (α, β) in model rollback varies according to different models and different datasets. $(0.2, 0.7)/(0.2, 0.4)$, $(0.8, 0.9)/(0.2, 0.4)$ are for CNN on CIFAR-10/100, ResNet-8 on CIFAR-10/100 respectively. Adam is used as the optimizer, which will be reset in each communication round to get a cyclical learning rate.

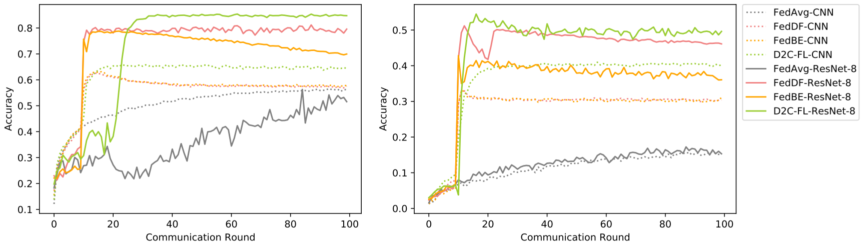
Baselines. FedAvg [1], as a typical aggregation method, is used as one of the baselines. We also compared our method with FedDF [12], FedBE [13]. The difference between them is the composition of the teacher ensemble. FedDF takes the local model as the teacher ensemble directly. FedBE implements the collection of teachers by using multiple linear combinations of local models to simulate sampling from local models distribution. The coefficients of linear combinations follow a Dirichlet distribution and the simulation achieves a comparable test accuracy with sampling models from Gaussian. The distillations with soft labels are eliminated to simplify the experiment, hard labels are used as a substitute to retrain a global student model. The One-Shot federated learning method [21] is not considered since it is similar to the one-round-version of FedDF, except it performs more local epochs training in each round. As the effect of our method largely depends on the leverage of centralized unlabeled data, we also compared our method with centralized learning on the same amount of labeled data.

4.2 Performance under different data settings

In this section, we compare the performance under different data settings. We evaluated the convergence of the global model from the aspect of test accuracy on CIFAR-10 and CIFAR-100 respectively.

Performance under the IID data setting. Figure 5 shows the test accuracies of all the baselines with CNN and ResNet-8 under the IID data setting. All methods with teacher ensemble begin pseudo-labeling at round 10, before which the accuracies are consistent with FedAvg. Benefits from the leverage of unlabeled data, our method achieves a significant improvement compared to FedAvg on both CIFAR-10 and CIFAR-100. Since the pseudo-labeling on the unlabeled data, our model gets a new richer dataset to learn and achieves a much better generalization effect. Due to less data on clients and the inherent complexity of the classification on CIFAR-100, FedAvg has a poor performance and the improvement of our method is more obvious.

By comparing with FedDF and FedBE, our method gets a significant improvement in the test accuracy with both CNN and ResNet-8. In addition to acquiring knowledge from the pseudo-labeled data which is consistent with FedDF and FedBE, we perform a rollback on teachers and the student to deal with the possible noise data and imbalance in the pseudo-labeled data. As a result, our model fits data features better and obtains more accurate mapping between the features and labels.



(a) CIFAR-10

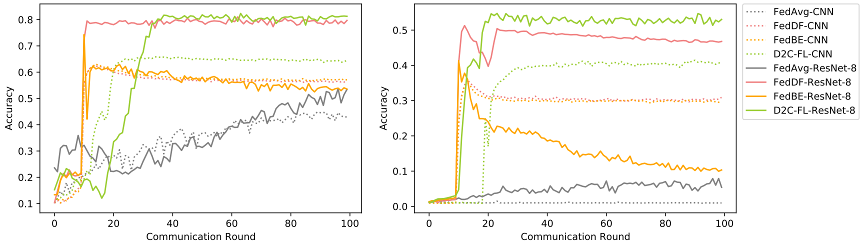
(b) CIFAR-100

Fig. 5 Comparison of the test accuracies under the IID data setting on CIFAR-10/100 with CNN and ResNet-8 in 100 communication rounds.

From Figure 5 we can see that ResNet-8 performs better than CNN generally. The generation and extensive use of pseudo-labeled data have resulted in a significant increase in test accuracy. Due to the poor leverage of unlabeled data (see Subsubsection 4.3), FedDF and FedBE with CNN suffer from a loss of test accuracy before the model converges. This also accords with the observation in [3], which showed that a pre-trained model would not learn from the FedAvg training on non-IID data and even had an accuracy drop on CIFAR-10. It is exactly the poor leverage of unlabeled data that makes it possible to generate another non-IID dataset and decrease the test accuracy. Those with ResNet-8 (except FedDF on CIFAR-10) may even keep declining and can not reach a convergence within 100 training rounds. A possible explanation for the result may be the stronger confirmation bias of ResNet which has been studied in [17].

In the case of training on fully labeled data, the residual block in ResNet solves the degradation problem of a deep neural network. When it comes to the training on pseudo-labeled data, the residual block would also be stubborn on the possible noise labels and cause the degradation again. Once confirming the knowledge in noise data, with the block of the threshold in filtering, it is hard for the poor ensemble to give pseudo labels. When the pseudo-labeled data is no longer generated, the global model gradually degenerates to FedAvg as local models fit the local labeled data, resulting in a decrease in test accuracy. Our method alleviates the problem by a different formation of teacher ensemble and a rollback on the teachers and student, making the model accuracy increase like monotonically on CIFAR-10. Due to the complexity of CIFAR-100, the test accuracy on CIFAR-100 has more obvious fluctuations, but it does not prevent our method achieves a higher test accuracy than others.

Performance unde the 0.8-bias data setting. The test accuracies with different models under the 0.8-bias data setting are compared in Figure 6. On the whole, our method still outperforms the baselines by a notable margin. With encountering the non-IID data, the test accuracy of FedAvg is likely to have a drop that matches the observation in [3]. Due to relying on pseudo-labeled data, FedDF, FedBE and our method with CNN do not show much influence in test accuracy and have a consistent performance with that in the IID data setting. While the test accuracy of FedBE with ResNet-8 has a steep decrease and degenerated to FedAvg, the test accuracy of our method keeps climbing with a slight fluctuation.

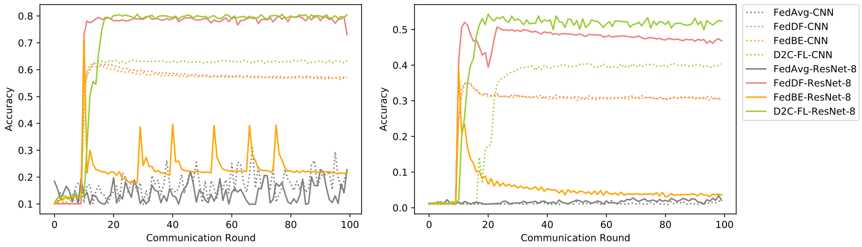


(a) CIFAR-10

(b) CIFAR-100

Fig. 6 Test accuracy of federated learning under the 0.8-bias data setting on CIFAR-10 with CNN and ResNet-8 in 100 communication rounds.

Performance under the 2-class data setting. Figure 7 shows the comparison between the baselines with different models under the 2-class data setting. With the greater degree of non-IID, it is even hard for FedAvg to have an increase of the test accuracy in 100 training rounds. The performance of FedBE with ResNet-8 on both CIFAR-10 and CIFAR-100 degrade to FedAvg like what is shown in the 0.8-bias data setting, but with a steeper degradation. While FedDF and FedBE with CNN and FedDF with ResNet-8 have a consistent performance with that in the IID data setting, we still perform better.



(a) CIFAR-10

(b) CIFAR-100

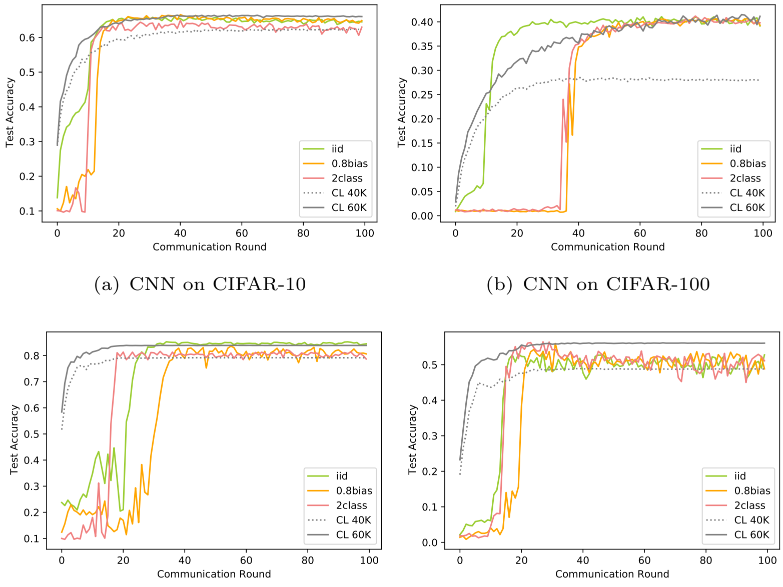
Fig. 7 Test accuracy of federated learning under the 2-class data setting on CIFAR-10 with CNN and ResNet-8 in 100 communication rounds.

Summarizing the performance under different data settings, our method consistently performs better than FedAvg and achieves generally higher test accuracy than FedDF and FedBE. When encountering non-IID data, FedAvg's performance degrades significantly, which is consistent with the observations in [3]. Other methods including ours leveraging centrally unlabeled data don't show much fluctuation in test accuracy. When the performance of our method with ResNet-8 on CIFAR-10 has a limited 0.5% improvement under the 2-class data setting, that under other data settings has a maximum 3% improvement on CIFAR-10 and 4% improvement on CIFAR-100, our method with CNN on CIFAR-10 and CIFAR-100 brings maximum 4% and 8% increases respectively.

Comparison with centralized learning. While our method performs model training on 20000 local labeled data and 40000 unlabeled data at the same time in a semi-supervised learning way, we compare it with supervised learning on 40000 and 60000 centralized labeled data respectively. Figure 8 shows the results. When we are almost immune to non-IID data, we achieve test accuracy comparable to supervised centralized learning. Benefits to the learning from labeled data on clients, we take full advantage of 40000 unlabeled data and achieve better performance than using only 40000 labeled data. Thus, we migrate the decentralized federated learning to centralized learning successfully.

4.3 Utilization of unlabeled data

Utilization of unlabeled data on CIFAR-10 with CNN. With 40000 × 80% unlabeled data on the server for global training, Figure 9 shows the evolution of pseudo-labeled data in the training process from the perspective of accuracy and quantity, as well as the distribution in the last training round. The preparation of teacher ensemble in our method results in the empty of the pseudo-labeled dataset in the first $T = 10$ rounds, and the model accuracy is consistent with FedAvg. To make the comparison more visualized, we suspend the pseudo-labeling in FedDF and FedBE at the same time, which also provides them a better initialization of teacher ensembles.



(a) CNN on CIFAR-10

(b) CNN on CIFAR-100

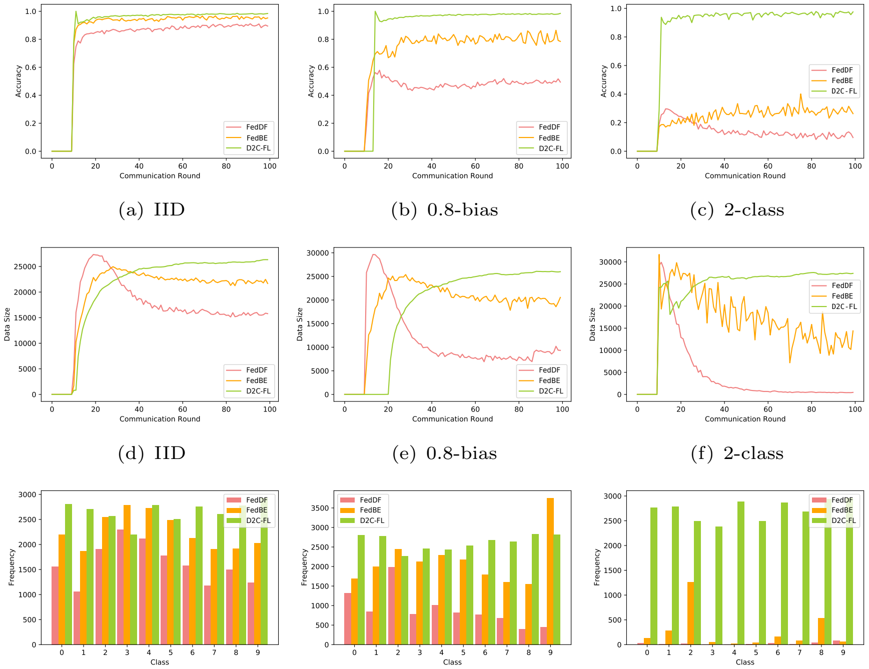
(c) ResNet-8 on CIFAR-10

(d) ResNet-8 on CIFAR-100

Fig. 8 Test accuracy of our method with 20000 labeled and 40000 unlabeled data, compared with the test accuracy of centralized learning with 40000/60000 labeled data. The comparison was performed on CIFAR-10 and CIFAR-100 with CNN and ResNet-8 respectively.

As Figure 9 shows, our method always has the highest pseudo-labeling accuracy in all the data settings. As we expected, it has almost become impervious to non-IID data while other methods have a decline in data accuracy with the deepening of non-IID. The pseudo-labeled data was selected by a prediction probability threshold τ, which indirectly reflects that our prediction results have much better alignment between confidence and accuracy (i.e. higher confidences, higher accuracy) in the confidence interval of $[\tau, 1.0]$. While other methods are less accurate than ours, the size of the dataset generated by them is also decreasing, much steeper in the case of non-IID. In our method, the accuracy and size are little affected by the non-IID data and rise steadily to better convergence. The bottom line of Figure 9 shows the data distribution in the last communication round. We can see that the generated data by our method has a more stable and even distribution, and it is an approximately IID dataset.

The high quality of the pseudo-labeled dataset generated by our method leads to incremental model accuracy. The evolution of the quality of the pseudo-labeled dataset is consistent with that of model accuracy shown in Subsubsection 4.2. As the generation of pseudo-labeled data leads to a sharp increase in model accuracy and its almost immunity to non-IID data, we successfully generate a centralized pseudo-labeled dataset for the global model to learn from.



(a) IID (b) 0.8-bias (c) 2-class

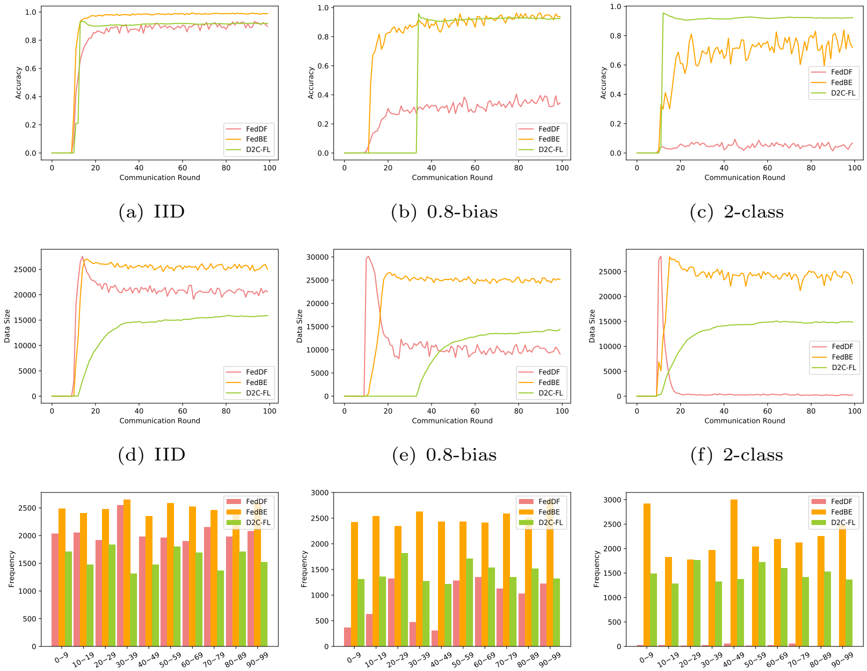
(d) IID (e) 0.8-bias (f) 2-class

(g) IID (h) 0.8-bias (i) 2-class

Fig. 9 Utilization of unlabeled data by all methods within 100 rounds using CNN on CIFAR-10.The top line shows the variation in the accuracy of the pseudo-labeled data, the middle line shows the variation in the quantity of the pseudo-labeled data and the bottom line shows the data distribution in the last communication round. It can be seen from the figure that the data generated by our method can always maintain considerable quality, especially on non-IID data.

Utilization of unlabeled data in other cases. Figure 10 shows the evolution of pseudo-labeled data on CIFAR-100 with CNN. It is apparent that our method still gains an advantage in data accuracy, especially for non-IID data. Although our method does not have an advantage in terms of data quantity, benefits from high data accuracy and model rollback, our method can still achieve the best test accuracy. As the data size is not the only factor that determines model performance, the labeling accuracy and the prevention of over-fitting are both important.

Figure 11 and Figure 12 show the same comparisons on CIFAR-10 and CIFAR-100 with ResNet-8. We gain advantages in both data accuracy and quantity, along with the more stable and even distribution, we still get a pseudo-labeled dataset with high quality. Different from the performance with



(a) IID (b) 0.8-bias (c) 2-class

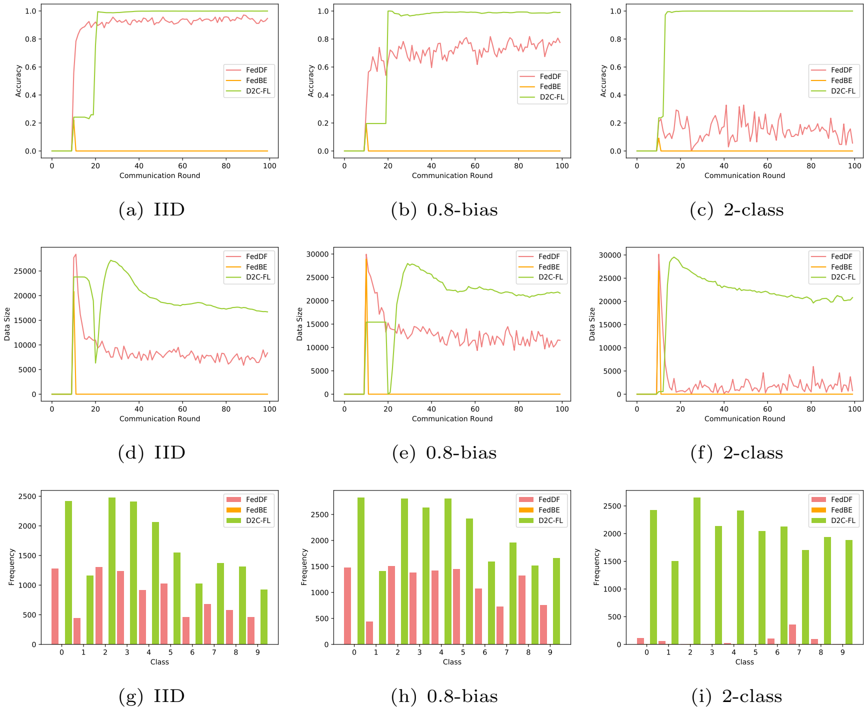
(d) IID (e) 0.8-bias (f) 2-class

(g) IID (h) 0.8-bias (i) 2-class

Fig. 10 Utilization of unlabeled data by all methods within 100 rounds using CNN on CIFAR-100.The top line shows the variation in the accuracy of the pseudo-labeled data, the middle line shows the variation in the quantity of the pseudo-labeled data and the bottom line shows the data distribution in the last communication round.

CNN, FedBE with ResNet-8 begins to degenerate to vanilla FedAvg after only a few rounds of training on the pseudo-labeled data. The sharp degradation of the pseudo-labeling causes the empty of the pseudo-labeled dataset, and its model accuracy continues to decline which has been shown in Subsubsection 4.2. We suspect that the aggregation of the local models and the residual block in ResNet-8 enhance the teacher's knowledge of noise data. A single local model as a teacher in FedDF would not have such a strong confirmation bias, the teachers in our method rely more on student model and β of which rollback to a randomly initialized model in each communication round, both of them keep the generation of pseudo-labeled data for global training.

With different sizes of unlabeled dataset. We investigate different sizes of unlabeled data on CIFAR-10 with CNN in Figure 13. $\{10000, 20000, 30000, 40000\}$ were used as the variable values in the experiment. The result shows that the model accuracy increases with the size of the unlabeled dataset and tends to increase. Although we can not pseudo-labeling and utilize all the unlabeled data due to the inevitable model error and the threshold τ, a larger unlabeled dataset always leads to a larger pseudo-labeled dataset and our method gains more than others regardless of the data settings.



(a) IID (b) 0.8-bias (c) 2-class

(d) IID (e) 0.8-bias (f) 2-class

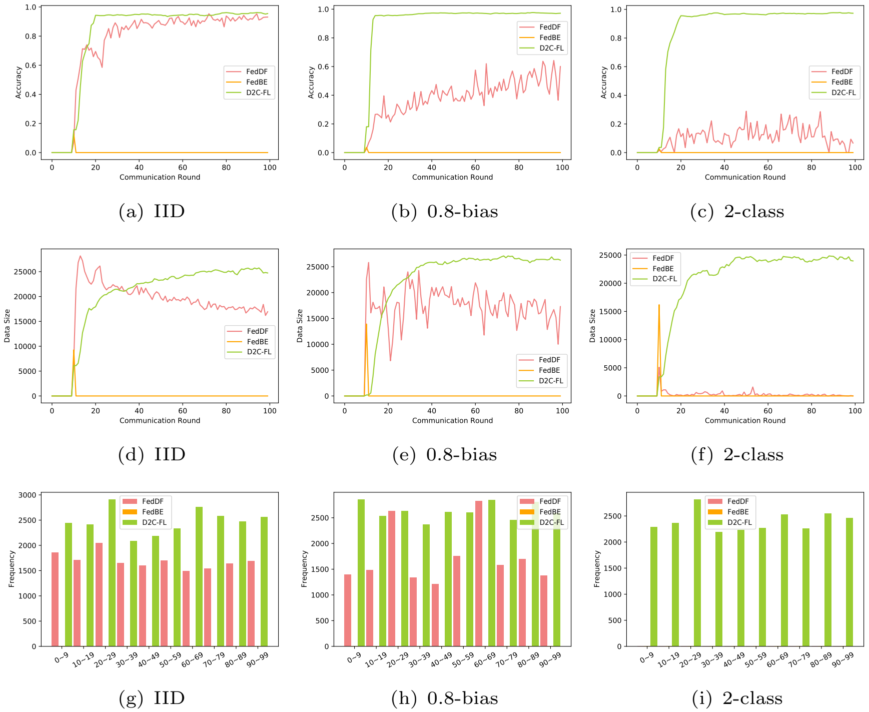
(g) IID (h) 0.8-bias (i) 2-class

Fig. 11 Utilization of unlabeled data by all methods within 100 rounds using ResNet-8 on CIFAR-10.The top line shows the variation in the accuracy of the pseudo-labeled data, the middle line shows the variation in the quantity of the pseudo-labeled data and the bottom line shows the data distribution in the last communication round.

4.4 Case studies

Different ensemble sizes. Keep 40000 data unlabeled on the server side, Figure 14 shows the utilization of the data and corresponding test accuracy with different teacher ensemble sizes. Due to the preparation of the teachers, the pseudo-labeling processes of different ensemble sizes begin in different communication rounds. We can see that a single teacher generates a pseudo-labeled dataset with the highest labeling accuracy but the smallest data size. With the increase of ensemble size, the labeling accuracy has a slight decrease while the data size has a noticeable increase. Since with the diversity of teachers, multiple teachers give samples multiple opportunities to pass the threshold τ and join in the pseudo-labeled dataset. Whenever a poor teacher generates a wrong label with high confidence, the sample will become noise data and be learned by the student model. As a result, the test accuracy has an irregular fluctuation in the $0 \sim 3\%$ range instead of a significant improvement. Nevertheless, multiple teachers are more likely to produce the best performance.

Different optimizers. We evaluate our method with *SGD* as the optimizer. The (α, β) is set to $(0.2, 0.8)$. As another learning rate strategy, we fix

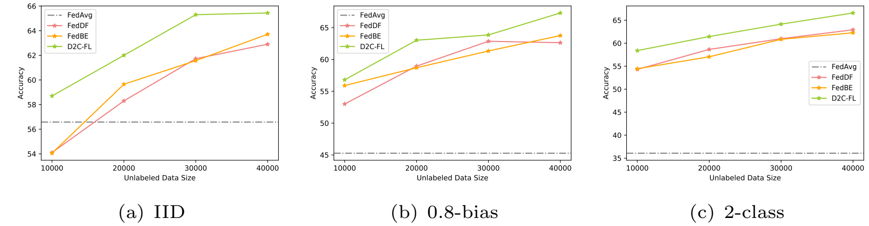


(a) IID (b) 0.8-bias (c) 2-class

(d) IID (e) 0.8-bias (f) 2-class

(g) IID (h) 0.8-bias (i) 2-class

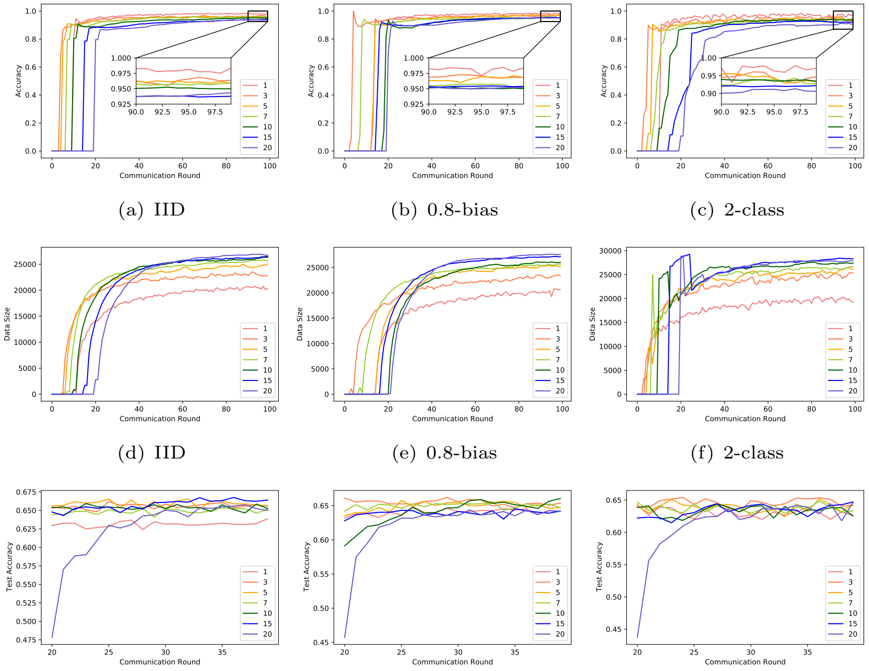
Fig. 12 Utilization of unlabeled data by all methods within 100 rounds using ResNet-8 on CIFAR-100.The top line shows the variation in the accuracy of the pseudo-labeled data, the middle line shows the variation in the quantity of the pseudo-labeled data and the bottom line shows the data distribution in the last communication round.



(a) IID (b) 0.8-bias (c) 2-class

Fig. 13 Performances with different sizes of the unlabeled dataset on CIFAR-10 with CNN. Our method always gains more from the unlabeled data in test accuracy.

the learning rate at 0.001 rather than a cyclical learning rate. The comparison shows in Figure 15. We can see that the learning with Adam has a faster convergence but an extremely slight over-fitting. Under the IID and the 0.8-bias data settings, *SGD* and *Adam* achieve comparable test accuracy. In the case of the 2-class data setting, with a deeper non-IID degree, *SGD* with a constant learning rate performs better. It indicates that keeping a constant rather than cyclical learning rate is better when encountering non-IID data.



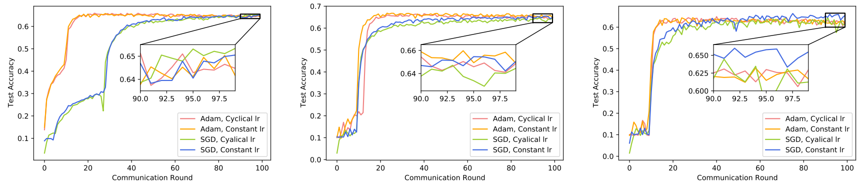
(a) IID (b) 0.8-bias (c) 2-class

(d) IID (e) 0.8-bias (f) 2-class

(g) IID (h) 0.8-bias (i) 2-class

Fig. 14 Effects of different teacher ensemble sizes on model accuracy on CIFAR-10 with CNN.

The advantage still maintained is that the test accuracy is little affected by the non-IID data.



(a) IID (b) 0.8-bias (c) 2-class

Fig. 15 Effects of different optimizers on model accuracy on CIFAR-10 with CNN.

5 Related Work

5.1 Federated Learning

Although federated learning has made good progress in both privacy protection and decentralized learning, there are some issues specific to federated learning that have been extensively studied. Like statistical heterogeneity due to different users characteristics, system heterogeneity as a result of different computing and storage capabilities of edge devices, communication cost in poor communication conditions, and security aiming at malicious clients or a malicious server [2, 23–26]. All of these would lead to poor convergence of the aggregation model. In this work, we mainly focus on the problem of statistical heterogeneity, related work of which will be discussed in the next subsection.

5.2 Challenges on Non-IID Data in Federated Learning

In conventional distributed learning, the model owner has rich data and distributes it to other devices to train a model collaboratively. Different from it, the server in federated learning does not have available data to learn, it uses not only the computing capability of local devices but also the data on them. As a result, the distributions of the data on different devices vary according to the users' personalities, become independent due to the possible connections between users. This non-IID data in the context of federated learning is also called statistical heterogeneity. Zhao et al.[3] had verified that the test accuracy of the global model trained by FedAvg decreased significantly under non-IID settings, the convergence rate also slowed down. How to eliminate the effect of non-IID data on the performance of the global model has been extensively studied.

Some existing researches deal with non-IID data by sharing part of the labeled data. Zhao et al.[3] shared 5% of labeled data to the server to improve the performance of the global model. Yoshida et al.[27] proposed to use 1% shared IID data to train a model with the same role as local models to participate in aggregation, so as to alleviate the impact of non-IID data. In this case, sharing labeled data has a great risk when the key information about data privacy lies exactly in the labels.

Active selection of participants is a good way to solve the problem of statistical heterogeneity. Lu et al.[28] proposed to select participants by clustering according to data distribution, to make local data used in training approximate to the global distribution. Wang et al.[22] proposed to use reinforcement learning to accomplish this participant selection process. The improvement brought by these methods is limited, and the model performance is still restricted by the non-IID data.

Sattler et al.[29] and Briggs et al.[30] didn't think that one model can fit the distribution of all clients' data, they proposed to cluster the local clients and aggregate different global models for different types of data distributions. When there are no explicit clusters of clients, Jamali-Rad et al.[31] proposed to learn the task correlation between clients with a contractive encoding of local data to perform more efficient federated aggregation of heterogeneous data.

Hu et al.[32] proposed to use GAN to train a feature extractor for local data to enhance the correlation between clients. The weight of each client who participates in aggregation is determined by the feature quality. Jeong et al.[33] proposed to use federated augmentation (FAug) to enhance local data, so as to make non-IID data between local clients become IID data. But clients executing FAug incurred additional computing and storage costs, making the conditions for participation more stringent. Li et al.[34] accelerated convergence by adding differences between local and global models as a regularization term, which was different from our purpose.

5.3 Data Disputes in Federated Semi-Supervised Learning

Federated Semi-Supervised Learning focuses on the problem of labels deficiency in federated learning [5]. It also refers to the leverage of additional unlabeled data to improve model performance. There are different scenarios based on different locations of labeled and unlabeled data. Part of the existing work studies that the server has labeled data, and the client has only unlabeled data because of the cost of labeling. Zhang et al.[6] used the consistent regularization loss [35] which was widely used in semi-supervised learning, and adopted the group-based model average method. Liu et al.[7] employed a minimax optimization-based client selection strategy to select the clients who hold high-quality models and used geometric median aggregation to robustly aggregate model updates.

At the same time, another part of the work believes that it is reasonable only for the client to have both labeled and unlabeled data. Directly, Albaseer et al.[36] performed the vanilla semi-supervised learning locally. Jeong et al.[5] proposed to decompose the parameters and perform disjoint learning on labeled data and unlabeled data respectively. Long et al.[37] sent different parts of model parameters to the server through a Teacher-Student framework, and the communication cost decreases with the convergence of the model.

There is also a scenario that the unlabeled data is only collected on the server side, while the client owns the labeled data. Jeong et al.[10] and Sattler et al.[9] aggregated the output of the supervised local model (i.e. logits) for each class to perform distillation. At the same time, Itahara et al.[11] shared all the labeled data, generated logits for each data for distillation. Although using logits could decrease the communication cost, the model performance was often poor. Lin et al.[12] still communicate model parameters between the clients and server. The local supervised models are used as an ensemble to give unlabeled data logits predictions. Chen et al.[13] fit the distribution of local models, sampling from the distribution to obtain an ensemble with higher quality. The unlabeled data were used to retrain the global model after being labeled with the logits.

In this work, we only study the last case. Following [12, 13], we use an ensemble to pseudo-labeling the unlabeled data.

5.4 Pseudo Labeling and Knowledge Transference

Generally, consistency regularization is a commonly used method in semi-supervised learning. It encourages the model to make the same prediction on the original sample and its perturbed samples. Differently, we use model predictions to generate pseudo labels to learn the knowledge in unlabeled data, which has been studied in [17]. In the context of federated learning, some existing researchers have used Ensemble Learning to make predictions. Ensemble learning constructs and integrates multiple base learners to make predictions, which could achieve better performances than predictions obtained from any base learner alone [38]. Guha et al.[21] and Lin et al.[12] drew on the idea of bagging in ensemble learning, they treated local clients as naturally formed bags to form the ensemble. The former carried out only once ensemble learning process while the latter iterated it finite times for incremental performance improvements. The basic models in bagging can be parallel trained, just as the training process on local clients can also be performed in a parallel manner. Mao et al.[39] and Chen et al.[13] both proposed to have a linear transformation on the base learners. Mao et al. aimed for pursuing the optimal projective direction of the linear transformation to have a better performance of the ensemble. Chen et al. simulated the sampling from the possible distribution of base learners through multiple linear transformations to catch better base learners. The sampled base learners consisted of a new ensemble and maintain its diversity.

In contrast to the parallel formation of the ensemble, we collect the base learners of the ensemble in tandem. The global model obtained in each round of federated learning will be collected as a base learner of our ensemble. Generally, the classification results obtained by multiple base learners, i.e. the ensemble, are better than those of a single classifier. The more accurate the pseudo-labeling technique, the less likely the target model is to over-fit the noise data. Since the strong temporal correlation of our base learners, the collection method is similar to boosting in ensemble learning.

6 Conclusion

In this work, we leveraged a large amount of unlabeled data to improve the performance of federated learning, especially on non-IID data. We assume the unlabeled data is from users but detached from the users and does not pose a privacy threat. We collected the global models obtained from each training round as teachers to make predictions on the unlabeled data. By using the predictions on the data as the pseudo labels, we migrated the focus of federated learning on decentralized labeled data to centralized pseudo-labeled data successfully. In addition, we used a model rollback to alleviate the impact of possible data imbalance and noise data in the pseudo-labeled data. Simulation shows that our method has a great improvement to federated learning, achieving similar and even higher accuracy compared to others. In addition, we achieved comparable performance with centralized supervised learning with the same data size as the unlabeled data. To explain the performance of the proposed method, we analyzed the utilization of pseudo-labeled data from the perspective of accuracy, quantity and distribution. The result is that our method can achieve greater utility of the unlabeled data and almost be immune to non-IID data.

References

[1] McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics, pp. 1273–1282 (2017)

[2] Lim, W.Y.B., Luong, N.C., Hoang, D.T., Jiao, Y., Liang, Y.-C., Yang, Q., Niyato, D., Miao, C.: Federated learning in mobile edge networks: A comprehensive survey. IEEE Communications Surveys & Tutorials **22**, 2031–2063 (2020)

[3] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V.: Federated learning with non-iid data. arXiv preprint arXiv:1806.00582 (2018)

[4] Li, X., Huang, K., Yang, W., Wang, S., Zhang, Z.: On the convergence of fedavg on non-iid data. In: International Conference on Learning Representations (2020)

[5] Jeong, W., Yoon, J., Yang, E., Hwang, S.J.: Federated semi-supervised learning with inter-client consistency & disjoint learning. In: International Conference on Learning Representations (2021)

[6] Zhang, Z., Yao, Z., Yang, Y., Yan, Y., Gonzalez, J.E., Mahoney, M.W.: Benchmarking semi-supervised federated learning. arXiv preprint arXiv:2008.11364 **17** (2020)

[7] Liu, Y., Yuan, X., Zhao, R., Zheng, Y., Zheng, Y.: Rc-ssfl: Towards robust and communication-efficient semi-supervised federated learning system. arXiv preprint arXiv:2012.04432 (2020)

[8] Diao, E., Ding, J., Tarokh, V.: Semifl: Communication efficient semi-supervised federated learning with unlabeled clients. arXiv preprint arXiv:2106.01432 (2021)

[9] Sattler, F., Marban, A., Rischke, R., Samek, W.: Communication-efficient federated distillation. arXiv preprint arXiv:2012.00632 (2020)

[10] Jeong, E., Oh, S., Kim, H., Park, J., Bennis, M., Kim, S.-L.: Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. arXiv preprint arXiv:1811.11479 (2018)

[11] Itahara, S., Nishio, T., Koda, Y., Morikura, M., Yamamoto, K.: Distillation-based semi-supervised federated learning for communication-efficient collaborative training with non-iid private data. IEEE Transactions on Mobile Computing, 1–1 (2021)

[12] Lin, T., Kong, L., Stich, S.U., Jaggi, M.: Ensemble distillation for robust model fusion in federated learning. In: NeurIPS (2020)

[13] Chen, H.-Y., Chao, W.-L.: Fed{be}: Making bayesian model ensemble applicable to federated learning. In: International Conference on Learning Representations (2021)

[14] Mao, S., Chen, J., Jiao, L., Gou, S., Wang, R.: Maximizing diversity by transformed ensemble learning. Appl. Soft Comput. **82** (2019)

[15] Zhou, Z.-H.: Ensemble Learning, pp. 181–210. Springer, ??? (2021)

[16] Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K.A., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R.G.L., Eichner, H., Rouayheb, S.E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P.B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S.U., Sun, Z., Suresh, A.T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F.X., Yu, H., Zhao, S.: Advances and open problems in federated learning. Found. Trends Mach. Learn. **14**, 1–210 (2021)

[17] Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K.: Pseudo-labeling and confirmation bias in deep semi-supervised learning. In: 2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020, pp. 1–8 (2020)

[18] Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings, vol. 30 (2017)

[19] Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features

from tiny images (2009)

[20] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**, 2278–2324 (1998)

[21] Guha, N., Talwalkar, A., Smith, V.: One-shot federated learning. arXiv preprint arXiv:1902.11175 (2019)

[22] Wang, H., Kaplan, Z., Niu, D., Li, B.: Optimizing federated learning on non-iid data with reinforcement learning. In: IEEE INFOCOM 2020-IEEE Conference on Computer Communications, pp. 1698–1707 (2020)

[23] Pang, J., Huang, Y., Xie, Z., Han, Q., Cai, Z.: Realizing the heterogeneity: A self-organized federated learning framework for iot. IEEE Internet of Things Journal **8**, 3088–3098 (2020)

[24] Laguel, Y., Pillutla, K., Malick, J., Harchaoui, Z.: Device heterogeneity in federated learning: A superquantile approach. arXiv preprint arXiv:2002.11223 (2020)

[25] Shen, S., Zhu, T., Wu, D., Wang, W., Zhou, W.: From distributed machine learning to federated learning: In the view of data privacy and security. Concurrency and Computation: Practice and Experience (2020)

[26] Zhu, T., Ye, D., Wang, W., Zhou, W., Yu, P.: More than privacy: Applying differential privacy in key areas of artificial intelligence. IEEE Transactions on Knowledge and Data Engineering, 1–1 (2020)

[27] Yoshida, N., Nishio, T., Morikura, M., Yamamoto, K., Yonetani, R.: Hybrid-fl for wireless networks: Cooperative learning mechanism using non-iid data. In: ICC 2020-2020 IEEE International Conference on Communications (ICC), pp. 1–7 (2020)

[28] Lu, R., Zhang, W., Li, Q., Zhong, X., Vasilakos, A.V.: Auction based clustered federated learning in mobile edge computing system. arXiv preprint arXiv:2103.07150 (2021)

[29] Sattler, F., Müller, K., Samek, W.: Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. IEEE Trans. Neural Networks Learn. Syst. **32**, 3710–3722 (2021)

[30] Briggs, C., Fan, Z., Andras, P.: Federated learning with hierarchical clustering of local updates to improve training on non-iid data. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–9 (2020)

[31] Jamali-Rad, H., Abdizadeh, M., Szabo, A.: Federated learning with taskonomy for non-iid data. arXiv preprint arXiv:2103.15947 (2021)

[32] Hu, K., Wu, J., Weng, L., Zhang, Y., Zheng, F., Pang, Z., Xia, M.: A novel federated learning approach based on the confidence of federated kalman filters. Int. J. Mach. Learn. Cybern. **12**, 3607–3627 (2021)

[33] Jeong, E., Oh, S., Kim, H., Park, J., Bennis, M., Kim, S.-L.: Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. arXiv preprint arXiv:1811.11479 (2018)

[34] Li, X., Liu, N., Chen, C., Zheng, Z., Li, H., Yan, Q.: Communication-efficient collaborative learning of geo-distributed jointcloud from heterogeneous datasets. In: 2020 IEEE International Conference on Joint Cloud Computing, pp. 22–29 (2020)

[35] Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C., Cubuk, E.D., Kurakin, A., Li, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence (2020)

[36] Albaseer, A., Ciftler, B.S., Abdallah, M., Al-Fuqaha, A.: Exploiting unlabeled data in smart cities using federated edge learning. In: 2020 International Wireless Communications and Mobile Computing (IWCMC), pp. 1666–1671 (2020)

[37] Long, Z., Che, L., Wang, Y., Ye, M., Luo, J., Wu, J., Xiao, H., Ma, F.: Fedsiam: Towards adaptive federated semi-supervised learning. arXiv preprint arXiv:2012.03292 (2020)

[38] Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q.: A survey on ensemble learning. Frontiers Comput. Sci. **14**, 241–258 (2020)

[39] Mao, S., Chen, J., Jiao, L., Gou, S., Wang, R.: Maximizing diversity by transformed ensemble learning. Appl. Soft Comput. **82** (2019)