International Journal of Network Dynamics and Intelligence Scilight

Article

Distillation-Based User Selection for Heterogeneous Federated Learning

Bowen Li and Wenling Li*

School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China
* Correspondence: lwlmath@buaa.edu.cn

Received: 7 September 2023
Accepted: 9 November 2023
Published: 26 June 2024

Abstract: Federated learning is a newly developing distributed machine learning technology, which makes it possible for users to train machine learning models with decentralized privacy data. Owing to huge communication overhead, the traditional federated learning algorithm samples user data randomly, which may reduce the performance of the model due to the statistical heterogeneity of users. In this paper, we propose a distillation-based user selection algorithm for federated learning in heterogeneous situations. Based on knowledge distillation, the soft targets of users are uploaded to the server as a basis for user selection. Our algorithm reduces the statistical heterogeneity of selected users, resulting in low additional communication and computation overhead. Experiments implemented on MNIST and fashion-MNIST show that the proposed algorithm obtains better model performance as compared to the federated averaging algorithm and several other user selection algorithms.

Keywords: federated learning; user selection; knowledge distillation

1. Introduction

In recent years, the usage of mobile devices has greatly increased. The widespread application of smartphones [1], internet of things devices [2], drones [3] and digital healthcare equipment [4] has generated a large amount of data, which is potentially valuable for machine learning. In the above scenarios, traditional centralized machine learning is no longer applicable, as significant communication overhead is required to gather a huge amount of data (held by users) on a server for centrally training models. In addition, the growing awareness of privacy makes people cautious about sharing their private data with others.

Federated learning [5], as a new machine learning paradigm, has been proposed to jointly train machine learning models among users without privacy leakage. The basic feature of federated learning is that only models, rather than raw data, are uploaded to the server from users. By performing local training on the user side and global aggregation on the server side, federated learning enables joint training of machine learning models while keeping raw data local.

Federated learning still faces many challenges in practical applications despite its advantages of processing large-scale distributed data [6]. Communication constraints are widely considered in the research of distributed systems [7], which are mainly caused by the up-link bandwidth limitation of wireless networks [8]. Frequent uploading of model parameters in federated learning causes huge overhead in the communication process, prohibiting the use of big models.

Another challenge of federated learning is user heterogeneity, which may lead to a significant decline in the model performance [9]. User heterogeneity can be divided into system heterogeneity and statistical heterogeneity. System heterogeneity refers to the differences (e.g. computing and communication capabilities) in devices among users. Statistical heterogeneity refers to the differences in data distributions among users. Non-independent and identically distributed (non-iid) user data is the main source of statistical heterogeneity in federated learning. For simplicity, we only take into account the statistical heterogeneity of users in this paper.

To address the two challenges mentioned above, various user selection algorithms have been proposed. The

user selection algorithm in federated learning has been first proposed in [5] to reduce communication overhead when aggregating model parameters. The impact of user heterogeneity would be mitigated if an appropriate user selection algorithm is applied. Taking advantage of the user model, a deep Q-learning-based mechanism has been proposed in [10] to select a subset of devices in each communication round. Note that, the application of the principal component analysis and reinforcement learning brings massive additional computing overhead on the server side.

A large number of user selection algorithms have been proposed, and the fundamental limitation of those algorithms is the requirement of large additional computational overhead on the server side. In order to address the above issues, we propose a distillation-based user selection algorithm for heterogeneous federated learning. In the proposed algorithm, a knowledge distillation algorithm is used to make each user compute soft targets (i.e. the posterior probability distribution of data samples belonging to each label) locally. The soft targets of users are collected by the server and used to obtain the user selection set through clustering algorithms. In addition, the knowledge distillation loss function is applied to improve model accuracy.

The main content and key contributions of this paper are listed below.

- We propose a distillation-based user selection algorithm for federated learning in heterogeneous situations. Our algorithm combines the knowledge distillation with the user selection to reduce the heterogeneity of selected users, without incurring significant computation and communication overhead.
- Based on the theory of the critical learning period of federated learning, experiments are designed to verify the impact of user selection rates on performance in early iterations.
- Experiments on MNIST and fashion-MNIST datasets are carried out to verify the performance of the proposed algorithm. Our algorithm outperforms multiple user selection algorithms in given scenarios.

2. Related Work

2.1. Federated Learning

The classic federated learning method adopts a centralized architecture where a number of users hold their own private data. The server collects and aggregates models trained by users. Federated averaging (FedAvg) [5] is the most widely used federated learning algorithm. In this algorithm, the federated learning system consists of a parameter server and several users. The server broadcasts global model parameters to all users. Each user uses its own dataset to calculate local parameters using stochastic gradient descent (SGD). After several iterations, the local parameters are uploaded to the parameter server. The server then performs the aggregation operation to get the global model, and the above process is cycled until convergence. Theoretically, FedAvg has the convergence guarantee on non-iid datasets in non-convex settings with partial user participation [11]. Although FedAvg performs well in simple image classification and natural language processing tasks, its performance drops significantly when the data distribution is non-iid.

2.2. User Selection in Federated Learning

The user selection strategy in federated learning specifies that part of the users participate in the update of the global model at each iteration. As the baseline algorithm of federated learning, FedAvg randomly selects a certain proportion of users at each global iteration. Although communication overhead is reduced, user heterogeneity is not taken into account while selecting users in FedAvg. An innovative study [12] has taken system heterogeneity and statistical heterogeneity into consideration together, and proposed an adaptive user selection algorithm. In order to estimate the upper bound of the user's gradient, additional communication and computation are required. Clustered sampling has been proposed in [13] to reduce the variability of user selection and increase the probability of users with a unique data distribution selected. A joint strategy for user selection and data offloading has been proposed in [14], enabling users with low data similarity to offload data from each other to improve efficiency. The above algorithms may face potential privacy leakage risks. The 'power-of-choice' selection strategy has been proposed in [15] to select users with larger loss functions. The above algorithm increases the accuracy and reduces the number of users participating in the update, thereby improving the rate of convergence at the cost of reducing robustness.

2.3. Knowledge Distillation

As proposed in [16], the knowledge distillation technology has made it possible to train a lightweight model whose performance is similar to a complex model on the same dataset. The core idea of knowledge distillation is to make the student model approximate the soft target of the teacher model by using the same sample for classification. The soft targets of the teacher model are believed to contain knowledge which can improve the performance of the student model. Although knowledge distillation has great potential in distributed machine learning, its application is

hindered by the privacy-sensitive characteristics of federated learning. In order to introduce knowledge distillation technology into federated learning, the basic concept of knowledge distillation is interpreted in various different aspects. Federated distillation has been proposed in [17] to reduce communication overhead, where the logit vector (instead of the model) has been shared between the server and users. An ensemble distillation method has been proposed in [18] where an unlabeled dataset from the server has been used to compute the logit output. Distillation-based semi-supervised federated learning has been introduced in [19] to share an unlabeled dataset among users to reduce the impact of statistical heterogeneity. Note that It is difficult to obtain a suitable public dataset in practical application scenarios due to the privacy and complexity of the data. In order to overcome this issue, our algorithm utilizes the knowledge distillation method without the need to share any datasets among users. FedKD [20] has been proposed based on adaptive mutual knowledge distillation and dynamic gradient compression techniques, which is both communication-efficient and effective. FedFTG has been proposed in [21] to mitigate the distribution discrepancy across users through direct model aggregation. EFDLS [22] has made it possible for users to solve different time series classification tasks using knowledge distillation.

3. Problem Formulation

In this section, we consider a distributed optimization problem under a basic federated learning setting, where all users share the same model structure. Each user is associated with a local loss function and the global loss function is a weighted sum of local loss functions

$$f(\mathbf{w}) = \sum_{i=1}^{N} p_i f_i(\mathbf{w}) \tag{1}$$

where $f_i(\cdot)$ is the loss function held by each user, and $f_i(\mathrm{w}) = \mathbb{E}_{\zeta_i \sim D_i} \mathcal{L}(\mathbf{w}; \zeta_\mathbf{i})$. N is the number of users. ζ_i is a single data sample held by the i-th user. For classification tasks, $\zeta_i = (x_i, y_i)$, where x_i and y_i are the feature and label of the data sample ζ_i, respectively. p_i is the weight of the i-th satisfying $\sum_{i=1}^{N} p_i = 1$ and $p_i \geq 0$. For simplicity, we assume that all users have the same data size as well as the same importance, therefore p_i is set to be $\frac{1}{N}$ for each user. $\mathbf{w}$ is the model parameter shared among users. D_i is the dataset held by the i-th user. In the iid setting, D_i has the same distribution for all users.

The goal of federated learning is to find the optimal solution through solving the following joint optimization problem:

$$\mathbf{w}^* = \underset{\mathrm{w} \in \mathbb{R}^d}{\operatorname{argmin}} f(\mathbf{w}) \tag{2}$$

where d is the dimension of the model parameter.

In this article, we only study the classification problem, which satisfies the demand for knowledge distillation. The most commonly used training loss function for classification is the cross entropy. Based on [16], the knowledge distillation loss function can be written as

$$\mathcal{L}(\mathbf{w}; \zeta_\mathbf{i}) = \alpha T^2 * KLdiv\left(\mathbf{Q}_S^\tau, \mathbf{Q}_T^\tau\right) + (1-\alpha) * CrossEntropy(\mathbf{Q}_S, \mathbf{y}_{true}) \tag{3}$$

where $\mathbf{Q}_S^\tau$ and $\mathbf{Q}_T^\tau$ are the local soft targets and global soft targets using the same temperature $\tau (> 1)$. $\mathbf{Q}_S$ is the output label of the data ζ_i of the model $\mathbf{w}$, and $\mathbf{y}_{true}$ is the true label of data ζ_i. α is a trade-off parameter for soft loss and hard loss.

Based on basic federated learning settings, the training process is divided into two parts: local training and parameter aggregation. Each user downloads the current global model weights $\mathbf{w}^{t-1}$ from the server, and updates local parameters by using the following SGD:

$$\mathbf{w}_i^{t-1} \leftarrow \mathbf{w}_i^{t-1} - \eta \nabla L\left(\mathbf{w}_i^{t-1}; \zeta_i\right) \tag{4}$$

where η is the learning rate. The above process will be repeated B times, and B is the local batch size.

In the parameter aggregation step, the server collects and aggregates parameters from selected users by

$$\mathbf{w}^{\mathrm{t}} = \frac{\sum_{\mathrm{i} \in \mathrm{I}} \mathbf{w}_\mathrm{i}^{\mathrm{t}-1}}{|\mathrm{I}|} \tag{5}$$

where I is the user selection set. All users in I are selected to participate in the parameter aggregation step. |I| is the number of selected uses subject to |I| = C∗K, where C is the user selection rate and K is the number of users. The user selection algorithm is designed in the next section to obtain the user selection set.

4. Algorithm

In this section, a distillation-based user selection algorithm is proposed to overcome the negative impact of the non-iid user data.

The existing study of the user selection algorithm considers the rate of convergence of the training process theoretically. This is reasonable when the global optimal solution can be found using the gradient-based method (when the loss function is convex). In the case of non-convex loss functions, it is uncertain to which stationary point the parameter converges. How to make the model converge to a better stationary point is an important issue. Critical learning periods in federated learning proposed in [23] indicate that the initial learning phase plays a critical role in federated learning. The accuracy of the global model will permanently decrease if defective data (e.g. low resolution images) is sued in the initial learning phase, no matter how many additional training epochs are performed in federated learning. Inspired by the above research, the user selection algorithm in early epochs may be particularly important to the final model performance. The more evenly distributed the overall user data is, the higher the quality of the selected training samples is. Considering the communication constraints in practical application scenarios, it is not practical to use a large user selection rate in multiple epochs. To balance model performance and communication efficiency, the user selection rate is set to be 1 in early epochs, and is then declined to the following constant:

$$C_t = \begin{cases} 1 & (t \leqslant t_0) \\ C & (t > t_0) \end{cases} \tag{6}$$

In order to alleviate the statistical heterogeneity of users, we hope to obtain the distribution of user data. Due to the user privacy preserving characteristic of federated learning, the server cannot get access to the privacy data directly, which means that the users' data distribution cannot be obtained in an explicit expression. According to [9], there is an implicit connection between the distribution of the training samples on a device and the model weights trained based on those samples, and the weight divergence between users is strictly bounded by the users' data distribution. It has been proposed in [24] that when selecting users, users with similar computation ability can be clustered into a group to mitigate the straggler effect (i.e. additional time consumption caused by system heterogeneity). The above heuristic research can be extended to solve the problem of statistical heterogeneity. The idea of clustering users with similar data distributions can be used to accelerate model convergence when selecting users [13]. Taking the K-means algorithm as an example, the time complexity of clustering algorithms is positively correlated with the data dimension [25]. Directly using model parameters for clustering may incur significant computational overhead.

To address the above issues, we propose a distillation-based user selection algorithm. Instead of selecting users based on their data or models, the proposed algorithm uses their soft targets as an indicator for selecting users. Different from [18], our algorithm does not require an open dataset, but requires users' local data to generate soft labels. The clustering algorithm is implemented on the server side to obtain the user selection set. The distillation and aggregation processes of the proposed algorithm are shown in Figures 1 and 2.

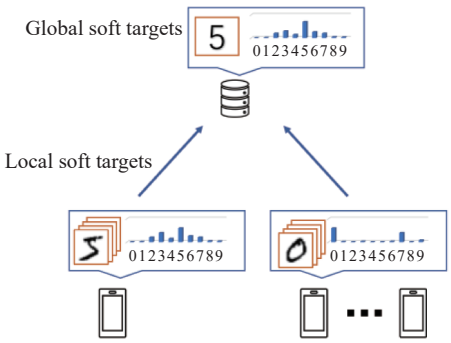

Figure 1. The knowledge distillation process of the proposed algorithm.

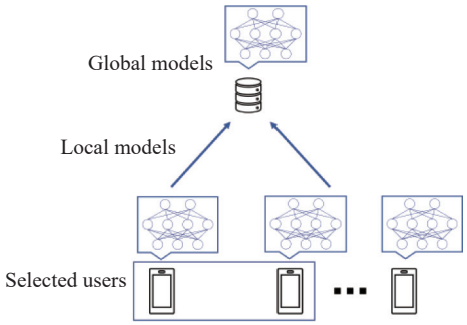

Figure 2. The federated aggregation process of the proposed algorithm.

As shown in Algorithm 1, on the server side, the global model and the global soft targets are broadcasted to all users at each global iteration. The server collects all soft targets from users and averages them to obtain the global soft targets for each label. For user selection, the server implements a clustering algorithm on the soft targets of each user. Sampling is conducted in each category to obtain a subset of users. After local training, the server collects local models uploaded by selected users and calculates the average to obtain new global model parameters. On the user side, each user receives the global model and global soft targets from the server, and uses a knowledge distillation loss function to update model parameters through RGD in (4). At the distillation step, the obtained local model is used to implement forward propagation based on local data samples to further obtain soft targets for each category.

$$q_{S,i,c}^{\tau} = \frac{exp\left(z_{i,c}/\tau\right)}{\sum_{j=1}^{C} exp\left(z_{i,j}/\tau\right)} \tag{7}$$

where C is the number of labels and $z_{i,j}$ is the output of the local model w_i^t on ζ_i. The soft targets are uploaded to the server for user selection, and the selected users upload their local models to the server.

Algorithm 1 Distillation-Based User Selection for Heterogeneous Federated Learning

Input: The K users are indexed by i, C_t is the user selection rate, τ is the knowledge distillation temperature, T is the maximum epoch

initialize $\mathbf{w}^0$,Q_T^{τ}

for each round $t = 0, 1, ..\mathrm{T}$ **do**

  for each user i **in parallel do**

    $\mathbf{w}_{\mathrm{i}}^{\mathrm{t}} \leftarrow LocalUpdate\left(i, \mathbf{w}^t, \mathbf{Q}_T^{\tau}\right)$

    $\mathbf{Q}_{S,i}^{\tau} \leftarrow Distillation\left(\mathbf{w}_i^t; \zeta_i, \tau\right)$

  end for

  $\mathbf{Q}_T^{\tau} \leftarrow \sum_{i=1}^{K} \frac{1}{K} \cdot \mathbf{Q}_{S,i}^{\tau}$

  $\mathrm{I} \leftarrow Clustering\left(\mathrm{Q}_{S,i}^{\tau}, max(C_t \cdot K, 1)\right)$

  $\mathbf{w}^{t+1} \leftarrow \sum_{i \in \mathbf{I}} \frac{1}{|T|} \cdot \mathbf{w}_i^t$

end for

Output: $\mathbf{w}^{\mathrm{T}}$

5. Evaluation

In this section, the performance of the proposed algorithm is tested through numerical simulation experiments. We train the convolutional neural network (CNN) model on the MNIST and fashion-MNIST dataset. The adopted CNN contains two convolutional layers and two fully connected layers. Followed by the max-pooling layer, two convolutional layers increase the number of channels to 20 by using a convolutional kernel size of 5 × 5. Then, two fully connected layers (with a dropout layer) reduce the output dimension to 10. The dropout rate is set to be 0.5. We use the following dataset partitioning method to process the non-iid data. The data samples in the training set are divided into several data shards and each of the shards contains data samples with the same label. Each user is randomly assigned with two data shards, which means that each user contains data from at most two different labels. In addition, we apply the K-means algorithm on the server side. The experimental results indicate that the K-means algorithm can complete clustering with low time consumption due to the low dimensionality of soft targets.

5.1. Full User Selection in Early Epochs

We first conduct an experiment on a toy example to demonstrate the potential relationship between the user selection rate and model accuracy. We perform simulations using the CNN model on MNIST dataset under the FedAvg framework where the user number is 100, the learning rate is $lr = 0.01$, the local batch size is $bs = 4$ and the local epoch is $ep = 3$. All users are selected in the first five epochs in the non-iid situation. Instead of being set to a constant, the user selection rate is set to be 1 in a few early epochs, and will be maintained at 0.1 after the early epochs. In addition, the standard FedAvg experiment is implemented to verify the effect of the full user selection in early epochs.

As shown in Figures 3 and 4, experimental results indicate that the full user selection in early epochs can improve the model accuracy by 5% and reduce the training loss in non-iid situations. Larger user selection rates can help to even out data distributions, and significantly improve training effectiveness in early epochs due to the impact of critical learning periods in federated learning. High user selection rates result in significant communication overhead, which further incurs additional time consumption. To address this issue, the user selection rate is set to a smaller constant in the following epochs.

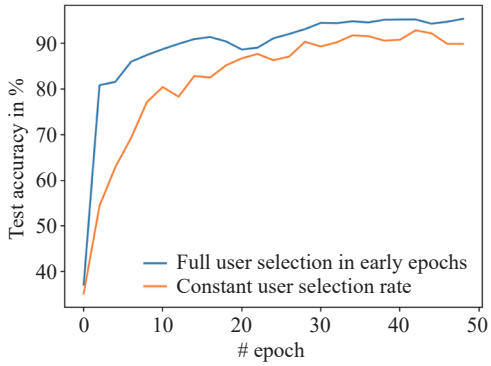

Figure 3. Influence of the user selection rate on test accuracy.

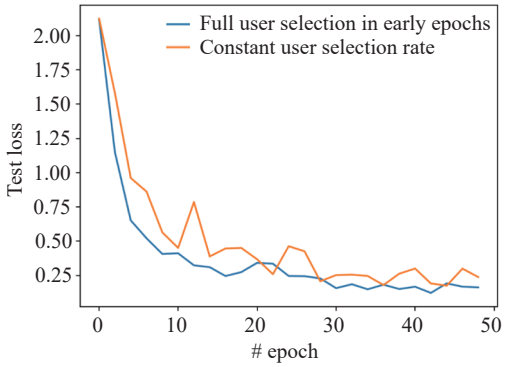

Figure 4. Influence of the user selection rate on test loss.

5.2. Distillation-Based User Selection Algorithm

The distillation-based user selection algorithm is conducted in the following experiment using the CNN model on MNIST dataset, where the user number is 100, the learning rate is $lr = 0.01$, the local batch size is $bs = 4$, the local epoch is $ep = 3$, the distillation weight is $\alpha = 0.1$ and the distillation temperature is $\tau = 4$. The K-means algorithm is used while clustering the local soft targets.

To determine the appropriate distillation weight α and distillation temperature τ, we apply our method to the MNIST dataset.

As shown in Table 1, choosing a large distillation weight may slow down the convergence speed, which might be due to the poor quality of the global soft target. We use the optimal value for the experiment.

Table 1 Test accuracy at the 30th epoch with different distillation weights and temperatures

	$\tau = 1$	$\tau = 4$	$\tau = 8$
$\alpha = 0.0$	88%	88%	88%
$\alpha = 0.1$	90%	93%	87%
$\alpha = 0.4$	82%	86%	81%
$\alpha = 0.7$	69%	67%	63%

We evaluate the performance of the proposed algorithm with FedAvg, power of choice selection strategies [15], active federated learning [26], FedCor [27], multinomial distribution sampling [28] and clustered sampling [13] under three different similarity measures in the same settings. FedCor is a federated learning framework built on a correlation-based user selection strategy, which utilizes the covariance stationarity to reduce the communication cost and boost the convergence rate of federated learning. The 'power-of-choice' framework is a communication-efficient and computation-efficient user selection framework that can flexibly span the trade-off between the convergence speed and solution bias. Active federated learning selects users based on the current model and the data from each user to maximize efficiency. We use FedAvg (iid) as an ideal scenario for comparison. For FedCor, 'power-of-choice' and active federated learning, we use the same hyperparameters in our method, including the learning rate, local batch size, and local iteration. Other hyperparameters are set to be the recommended values in the corresponding literature.

Better global soft targets can be obtained by using the full user selection in early epochs, and this accelerates the local training process. As shown in Figures 5 and 6, the proposed algorithm has more stable performance in terms of testing accuracy and testing loss. Compared to other algorithms, the proposed algorithm has smaller fluctuations. Due to the use of higher user selection rates in early epochs, the early performance of the proposed algorithm has been significantly improved. Compared to 'power-of-choice', active federated learning, FedCor, multinomial distribution sampling and clustered sampling, the proposed algorithm has a lower computational overhead on the server side, making it easier to deploy in environments with scarce server computing resources. Compared to FedAvg, the proposed algorithm saves 20% of communication rounds when achieving an accuracy of 90%. Although our method is surpassed by some other methods after 25 epochs, the model accuracy can still maintain a slow increase.

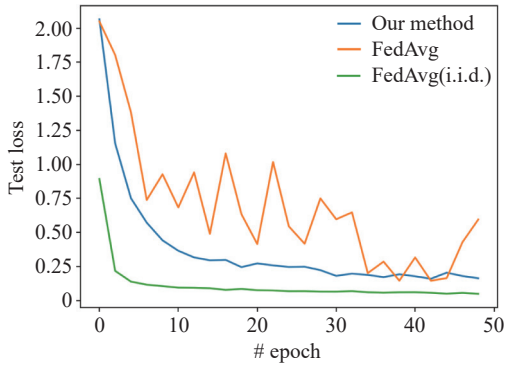

Figure 5. Influence of the user selection algorithm on test loss.

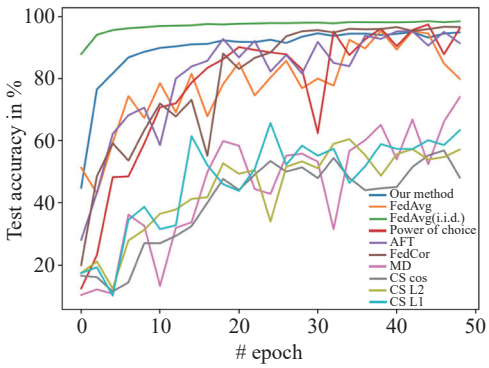

Figure 6. Comparison of test accuracy using different algorithms on MNIST.

In addition, we conduct experiments on the fashion-MNIST dataset with the same settings, and the experimental results are shown in Figure 7.

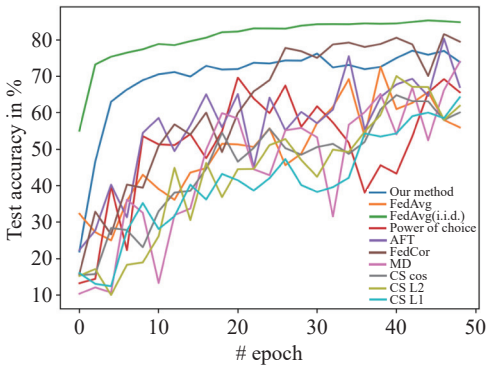

Figure 7. Comparison of test accuracy using different algorithms on Fashion-MNIST

As shown in Figure 7, despite being inferior to FedCor in model performance, our method outperforms other methods. Compared to MNIST, the features of the fashion-MNIST dataset are more complex and difficult to extract through the shallow network. Due to the use of the same model for the two tasks mentioned above, the accuracy of the model decreases on the fashion-MNIST dataset. The model accuracy of FedAvg in the iid situation decreases from 98% to 82%. Our method significantly surpasses the FedAvg algorithm in the same situation. In addition, benefiting from the clustering-based user selection strategy, the model accuracy in our method increases more smoothly compared to other methods. Our method significantly outperforms other methods in early epochs, as a better global model is obtained in the first epoch. The test accuracy may increase if a deeper CNN is used.

6. Conclusion

In this paper, we have discussed the effect of user selection strategies on federated learning under statistical heterogeneity. Based on the characteristics of the heavy communication overhead of federated learning, we have proposed a distillation-based user selection strategy for heterogeneous federated learning. The knowledge distillation loss function has been applied to local training. Local soft targets have been calculated by all users and uploaded for user selection. Based on local soft targets, the server has obtained a user selection set through clustering. The proposed algorithm can reduce the statistical heterogeneity of users, accompanied by low additional communication overhead. Numerical experiments have been designed to evaluate the performance of the proposed algorithm. Experimental results have validated that the proposed algorithm outperforms FedAvg and several algorithms in a given scenario, achieving higher accuracy with the same number of communication rounds.

Author Contributions: Bowen Li: experiment designing, paper writing; **Wenling Li:** supervision, writing-review, instruction. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by NSFC (62376015, 61976013, U22B2038).

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, L.; Fan, Y.X.; Tse, M.; *et al*. A review of applications in federated learning. Comput. Ind. Eng., **2020**, *149*: 106854. doi: 10.1016/j.cie.2020.106854
2. Yao, F.; Ding, Y.L.; Hong, S.G.; *et al*. A survey on evolved LoRa-based communication technologies for emerging internet of things applications. Int. J. Network Dyn. Intell., **2022**, *1*: 4−19. doi: 10.53941/ijndi0101002
3. Zhang, Z.; Ma, S.Y.; Yang, Z.H.; *et al*. Robust semisupervised federated learning for images automatic recognition in internet of drones. IEEE Internet Things J., **2023**, *10*: 5733−5746. doi: 10.1109/JIOT.2022.3151945
4. Rieke, N.; Hancox, J.; Li, W.Q.; *et al*. The future of digital health with federated learning. npj Digital Med., **2020**, *3*: 119. doi: 10.1038/s41746-020-00323-1
5. McMahan, B.; Moore, E.; Ramage, D.; *et al*. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017*; PMLR, 2017; pp. 1273−1282.
6. Abdulrahman, S.; Tout, H.; Ould-Slimane, H.; *et al*. A survey on federated learning: The journey from centralized to distributed on-site learning and beyond. IEEE Internet Things J., **2021**, *8*: 5476−5497. doi: 10.1109/JIOT.2020.3030072
7. Wang, Y.A.; Shen, B.; Zou, L.; *et al*. A survey on recent advances in distributed filtering over sensor networks subject to communication constraints. Int. J. Network Dyn. Intell., **2023**, *2*: 100007. doi: 10.53941/ijndi0201007

8. Zhang, T.H.; Lam, K.Y.; Zhao, J.; *et al*. Joint device scheduling and bandwidth allocation for federated learning over wireless networks. *IEEE Trans. Wireless Commun*. **2023**. doi:10.1109/TWC.2023.3291701
9. Cho, Y.J.; Wang, J.Y.; Chirvolu, T.; *et al*. Communication-efficient and model-heterogeneous personalized federated learning via clustered knowledge transfer. IEEE J. Sel. Top. Signal Process., **2023**, *17*: 234−247. doi: 10.1109/JSTSP.2022.3231527
10. Wang, H.; Kaplan, Z.; Niu, D.; *et al*. Optimizing federated learning on Non-IID data with reinforcement learning. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications, Toronto, ON, Canada, 6–9 July 2020*; IEEE: New York, 2020; pp. 1698–1707. doi:10.1109/INFOCOM41043.2020.9155494
11. Yang, H.B.; Fang, M.H.; Liu, J. Achieving linear speedup with partial worker participation in Non-IID federated learning. In *Proceedings of the 9th International Conference on Learning Representations, 3–7 May 2021*; ICLR, 2021.
12. Luo, B.; Xiao, W.L.; Wang, S.Q.; *et al*. Tackling system and statistical heterogeneity for federated learning with adaptive client sampling. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications, London, UK, 2–5 May 2022*; IEEE: New York, 2022; pp. 1739–1748. doi:10.1109/INFOCOM48880.2022.9796935
13. Fraboni, Y.; Vidal, R.; Kameni, L.; *et al*. Clustered sampling: Low-variance and improved representativity for clients selection in federated learning. In *Proceedings of the 38th International Conference on Machine Learning, 18–24 July 2021*; PMLR, 2021; pp. 3407–3416.
14. Wang, S.; Lee, M.; Hosseinalipour, S.; *et al*. Device sampling for heterogeneous federated learning: Theory, algorithms, and implementation. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications, Vancouver, BC, Canada, 10–13 May 2021*; IEEE: New York, 2021; pp. 1–10. doi:10.1109/INFOCOM42981.2021.9488906
15. Cho, Y.J.; Wang, J.Y.; Joshi, G. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. arXiv: 2010.01243, 2020. doi:10.48550/arXiv.2010.01243
16. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. arXiv: 1503.02531, 2015. doi:10.48550/arXiv.1503.02531
17. Jeong, E.; Oh, S.; Kim, H.; *et al*. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-IID private data. arXiv: 1811.11479, 2018. doi:10.48550/arXiv.1811.11479
18. Lin, T.; Kong, L.J.; Stich, S.U.; *et al*. Ensemble distillation for robust model fusion in federated learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020*; ACM: Red Hook, 2020; p. 198. doi:10.5555/3495724.3495922
19. Itahara, S.; Nishio, T.; Koda, Y.; *et al*. Distillation-based semi-supervised federated learning for communication-efficient collaborative training with non-IID private data. IEEE Trans. Mobile Comput., **2023**, *22*: 191−205. doi: 10.1109/TMC.2021.3070013
20. Wu, C.H.; Wu, F.Z.; Lyu, L.J.; *et al*. Communication-efficient federated learning via knowledge distillation. Nat. Commun., **2022**, *13*: 2032. doi: 10.1038/s41467-022-29763-x
21. Zhang, L.; Shen, L.; Ding, L.; *et al*. Fine-tuning global model via data-free knowledge distillation for non-IID federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022*; IEEE: New York, 2022; pp. 10164–10173. doi:10.1109/CVPR52688.2022.00993
22. Xing, H.L.; Xiao, Z.W.; Qu, R.; *et al*. An efficient federated distillation learning system for multitask time series classification. IEEE Trans. Instrum. Meas., **2022**, *71*: 2517012. doi: 10.1109/TIM.2022.3201203
23. Yan, G.; Wang, H.; Li, J. Seizing critical learning periods in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence, 22 February–1 March 2022*; AAAI: Palo Alto, 2022; pp. 8788–8796. doi:10.1609/aaai.v36i8.20859
24. Cui, Y.G.; Cao, K.; Cao, G.T.; *et al*. Client scheduling and resource management for efficient training in heterogeneous IoT-edge federated learning. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst., **2022**, *41*: 2407−2420. doi: 10.1109/TCAD.2021.3110743
25. Ikotun, A.M.; Ezugwu, A.E.; Abualigah, L.; *et al*. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Inf. Sci., **2023**, *622*: 178−210. doi: 10.1016/j.ins.2022.11.139
26. Goetz, J.; Malik, K.; Bui, D.; *et al*. Active federated learning. arXiv: 1909.12641, 2019. doi:10.48550/arXiv.1909.12641
27. Tang, M.X.; Ning, X.F.; Wang, Y.T.; *et al*. FedCor: Correlation-based active client selection strategy for heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022*; IEEE: New York, 2022; pp. 10092–10101. doi:10.1109/CVPR52688.2022.00986
28. Li, T.; Sahu, A.K.; Zaheer, M.; *et al*. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems 2020, Austin, TX, USA, 2–4 March 2020*; MLSys, 2020; pp. 429–450.

Citation: Li, B.; Li, W. Distillation-Based User Selection for Heterogeneous Federated Learning. *International Journal of Network Dynamics and Intelligence*. 2024, 3(2), 100007. doi: 10.53941/ijndi.2024.100007