Clustered Federated Learning for Heterogeneous Data (Student Abstract)

Xue Yu, Ziyi Liu, Yifan Sun, Wu Wang

Center for Applied Statistics, School of Statistics, Renmin University of China, China
{xueyu_2019, ziyiliu, sunyifan, wu.wang}@ruc.edu.cn

Abstract

Federated Learning (FL) aims to achieve a global model via aggregating models from all devices. However, it can diverge when the data on the users' devices are heterogeneous. To address this issue, we propose a novel clustered FL method (FPFC) based on a nonconvex pairwise fusion penalty. FPFC can automatically identify clusters without prior knowledge of the number of clusters and the set of devices in each cluster. Our method is implemented in parallel, updates only a subset of devices at each communication round, and allows each participating device to perform inexact computation. We also provide convergence guarantees of FPFC for general nonconvex losses. Experiment results demonstrate the advantages of FPFC over existing methods.

Introduction

Clustered FL was proposed to address the potential weakness of a global model by assuming that the devices can be partitioned into clusters and that the devices from the same cluster share the same model. CFL (Sattler, Müller, and Samek 2021) recursively bipartitions the devices top-down based on the similarity between devices' gradients. It requires the server to seek an optimal bipartition, which is computationally expensive. IFCA (Ghosh et al. 2020) can be viewed as the alternating minimization method. But it requires the specification of the number of clusters.

In this paper, we apply a nonconvex function to penalize the pairwise differences of models of devices and promote zero differences, and in turn, clustering. We propose FPFC to automatically determine the cluster-specific optimal models without any prior knowledge of the cluster structure.

Approach

Consider a FL setting with m devices and one server. Each device has parameter $\omega_i \in \mathbb{R}^d$, which incorporates possible heterogeneity in parameters flexibly. We assume that m devices form L disjoint clusters and devices in the same cluster have the same parameters. Let $G = \{G_1, \ldots, G_L\}$ be a mutually exclusive partition of $[m]$; then, $\omega_i = \alpha_l$ for all $i \in G_l$, where α_l is the common value for the l-th cluster. Our goal is to estimate L, identify the cluster membership of each device, and learn the underlying cluster-specific parameters $(\alpha_1, \ldots, \alpha_L)$. Our objective is:

$$\min_{\omega}\{F(\omega) = \sum_{i=1}^{m} f_i(\omega_i) + \frac{1}{2m}\sum_{i=1}^{m}\sum_{j=1}^{m} g(\|\omega_i - \omega_j\|, \lambda)\}.$$

Here $\omega = (\omega_1^\top, \ldots, \omega_m^\top)^\top$, $f_i(\omega_i) = \frac{1}{n_i}\sum_{s=1}^{n_i} \ell(\omega_i; z_i^s)$, where $\ell(\omega_i; z_i^s)$ represents a preselected loss function corresponding to the data point z_i^s and parameter ω_i. The second term is a pairwise fusion penalty with hyperparameter $\lambda > 0$, which promotes zero differences of parameters and partitions the devices into clusters. Convex functions could lead to biased estimates of parameters and can not correctly recover clusters as shown in our experiments, here we focus on the nonconvex SCAD penalty (Fan and Li 2001):

$$P_a(t, \lambda) = \begin{cases} \lambda|t|, & |t| \leq \lambda \\ \frac{a\lambda|t| - 0.5(t^2+\lambda^2)}{a-1}, & \lambda < |t| \leq a\lambda \\ \frac{\lambda^2(a+1)}{2}, & |t| > a\lambda. \end{cases}$$

Since the SCAD penalty is not differentiable everywhere, we approximate it by the differentiable surrogate:

$$\tilde{P}_a(t, \lambda) = \left(\frac{\lambda}{2\xi}t^2 + \frac{\xi\lambda}{2}\right) I(|t| \leq \xi) + P_a(t, \lambda) I(|t| > \xi),$$

where $\xi < \lambda$ and $I(\cdot)$ is the indicator function. Let $\tilde{g}(t, \lambda) = \tilde{P}_a(t, \lambda)$, we then apply a Douglas-Rachford splitting strategy by introducing a set of new parameters $\theta_{ij} = \omega_i - \omega_j$. The algorithm is presented in Algorithm 1. Different from ADMM, we separate the minimization over ω into m subproblems solved in parallel and allow the ω_i-minimization to be solved inexactly by running T_i steps of local GD/SGD updates. We also allow a small subset of devices to participate in training. Each θ_{ij}^{k+1}-minimization has a closed-form analytical solution and the server only needs to conduct assignment operations. Because no prior knowledge of the cluster is avaliable, we only perform clustering after convergence. We put devices i and j in the same cluster if $\|\theta_{ij}^K\| \leq \nu$, where ν is a threshold and can be chosen from $[\xi, 0.5]$. Finally, we have $\hat{L}$ estimated clusters $\hat{G}_1, \ldots, \hat{G}_{\hat{L}}$. The model parameters for the l-th cluster is $\hat{\alpha}_l = \frac{\sum_{i \in \hat{G}_l} n_i \omega_i^K}{\sum_{i \in \hat{G}_l} n_i}$.

Assumption 1. All $f_i(\cdot)$ are continuously differentiable and there exists a Lipschitz constant $L_f > 0$ such that

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_f \|x - y\|, \forall x, y \in \mathbb{R}^d.$$

Algorithm 1: Fusion Penalized Federated Clustering (FPFC)

```
Initialize $\omega_1^0 = \ldots = \omega_m^0$, $\zeta_i^0 = \omega_i^0$, $\theta_{ij}^0 = \mathbf{0}$ and $v_{ij}^0 = \mathbf{0}$.
for $k = 0$ to $K-1$ do
    1: Server randomly selects a subset of devices $\mathcal{A}_k$.
    2: Server sends $\zeta_i^k$ to each device $i \in \mathcal{A}_k$.
    3: [Local update] For each device $i \in \mathcal{A}_k$: $\omega_i^{k,0} = \omega_i^k$
    for $t = 0$ to $T_i - 1$ do
        $\omega_i^{k,t+1} = \omega_i^{k,t} - \alpha[\nabla f_i(\omega_i^{k,t}) + \rho(\omega_i^{k,t} - \zeta_i^k)]$
    end for
    4: Each device $i \in \mathcal{A}_k$ sends $\omega_i^{k+1} = \omega_i^{k,T_i}$ back to the server.
    5: [Server update] For $i, j \in \mathcal{A}_k$ $(i < j)$, server computes $\delta_{ij}^{k+1} = \omega_i^{k+1} - \omega_j^{k+1} + \frac{v_{ij}^k}{\rho}$ and updates
    $\theta_{ij}^{k+1} = \arg\min_{\theta_{ij}} \tilde{g}(\|\theta_{ij}\|) + \frac{\rho}{2}\|\omega_i^{k+1} - \omega_j^{k+1} + \frac{v_{ij}^k}{\rho} - \theta_{ij}\|^2$
    $v_{ij}^{k+1} = v_{ij}^k + \rho(\omega_i^{k+1} - \omega_j^{k+1} - \theta_{ij}^{k+1})$
    $\theta_{ji}^{k+1} = -\theta_{ij}^{k+1}$, $v_{ji}^{k+1} = -v_{ij}^{k+1}$.
    For $i \notin \mathcal{A}_k$ and $j \notin \mathcal{A}_k$: $\theta_{ij}^{k+1} = \theta_{ij}^k$, $v_{ij}^{k+1} = v_{ij}^k$.
    For $i \in [m]$: $\zeta_i^{k+1} = \frac{1}{m}\sum_{j=1}^m (\omega_j^{k+1} + \theta_{ij}^{k+1} - \frac{v_{ij}^{k+1}}{\rho})$.
end for
```

Assumption 2. (Boundedness) $F^* = \inf_{\omega \in \mathbb{R}^d} F(\omega) > -\infty$.

Assumption 3. There exist $p_1, ..., p_m > 0$ such that $P(i \in \mathcal{A}_k) = p_i > 0$ for all $i \in [m]$.

This assumption implies that each device has a nonzero probability to participate in training. Define the mapping $\mathcal{G}(\omega, \theta, v) = \theta - \text{prox}_{\mathcal{L}_0}(\theta)$. Then, the condition $0 \in \partial_\theta \mathcal{L}_0(\omega^*, \theta^*, v^*)$ is equivalent to $\mathcal{G}(\omega^*, \theta^*, v^*) = 0$.

Theorem 1. *Suppose Assumptions 1-3 hold. Assume there exists $L_- > 0$ such that $\nabla^2 f_i \succeq -L_-\mathbf{I}$ with $\mu = \rho - L_- > 0$. If T_i, α, ξ, λ, a, and ρ are chosen such that $T_i > -\frac{2\log 2}{\log c}$, $0 < \alpha \leq \frac{1}{L_f + 2\rho - L_-}$, $\rho > \max\{\frac{L_f}{1-2c^{\frac{T}{2}}}, \frac{2\lambda}{\xi}, \frac{2}{a-1}, L_-\}$ where $c = 1 - \alpha\frac{2\mu(L_f+\rho)}{L_f+\rho+\mu}$ and $T = \min_{i\in[m]} T_i$, then we have the following statements:*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla_\omega \mathcal{L}_0(\omega^{k+1}, \theta^{k+1}, v^{k+1})\|^2] \leq \frac{C_1[F(\omega^0) - F^* + m\xi\lambda/4]}{K}$$

$$\mathbb{E}[\|\mathcal{G}(\omega^{k+1}, \theta^{k+1}, v^{k+1})\|^2] \leq \frac{m^2\lambda^2\xi^2}{(\lambda+\xi)^2}$$

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla_v \mathcal{L}_0[\omega^{k+1}, \theta^{k+1}, v^{k+1}]\|^2] \leq \frac{C_2[F(\omega^0) - F^* + m\xi\lambda/4]}{K},$$

where $C_1 = \frac{1}{\hat{p}}[\frac{6[(c^{-\frac{T}{2}}-1)^{-2}L_h^2 + \rho^2}{\rho - L_f - 2(c^{-\frac{T}{2}}-1)^{-1}L_h} + \frac{12\rho^3}{\rho^2 - L_{\tilde{g}}\rho - 2L_{\tilde{g}}^2}]$, $C_2 = \frac{L_{\tilde{g}}^2}{m\rho(\rho^2 - L_{\tilde{g}}\rho - 2L_{\tilde{g}}^2)}$, $\hat{p} = \min_{i\in[m]}\{p_i\}$.

Preliminary Results

Implementation. We evaluate the performance of FPFC on synthetic and two public datasets. We consider $m = 100$ devices which form $L = 4$ clusters and we follow the settings in (Li et al. 2020) to generate synthetic data. For MNIST/FMNIST, we follow the procedures in CFL (Sattler, Müller, and Samek 2021) to create non-IID data settings. We compare FPFC with several baselines and FPFC-ℓ_1 is a variant of FPFC with a convex ℓ_1 penalty. We select testing accuracy (Acc), number of identified clusters (Num) and adjusted Rand index (ARI) as evaluation metrics.

	Acc	Num	ARI
LOCAL	84.97%± 0.06	×	×
FedAvg	30.35%± 0.05	×	×
Per-FedAvg	56.28%± 0.06	×	×
IFCA	60.42%± 0.17	2.33±0.94	0.47± 0.33
CFL	86.60%± 0.05	42.0±4.55	0.28± 0.11
FPFC-ℓ_1	83.85%± 0.06	5.67± 2.87	0.63± 0.21
FPFC	**89.46%± 0.04**	**4.00±0.00**	**1.00± 0.00**

Table 1: Experimental results on synthetic dataset.

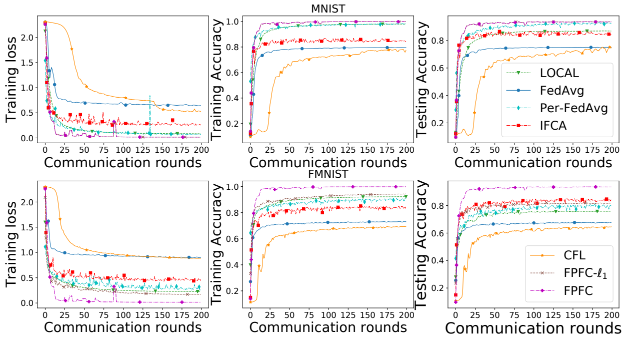

Figure 1: Experimental results on MNIST and FMNIST.

Results and Conclusions. Since LOCAL, FedAvg and Per-FedAvg can not cluster, we do not report their Num and ARI results. From Table 1 and Figure 1, FPFC outperforms all six baselines in prediction and device clustering. These results reveal that FPFC can automatically determine the number and structure of clusters and optimal model within each cluster. Future works may consider conducting extensive experiments and incorporating privacy-preserving techniques.

References

Fan, J.; and Li, R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456): 1348–1360.

Ghosh, A.; Chung, J.; Yin, D.; and Ramchandran, K. 2020. An Efficient Framework for Clustered Federated Learning. In *NeurIPS*.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated Optimization in Heterogeneous Networks. In *MLSys*.

Sattler, F.; Müller, K.-R.; and Samek, W. 2021. Clustered Federated Learning: Model-Agnostic Distributed Multitask Optimization Under Privacy Constraints. *IEEE Transactions on Neural Networks and Learning Systems*.