Optimizing the Collaboration Structure in Cross-Silo Federated Learning

Wenxuan Bao [1] Haohan Wang [1] Jun Wu [1] Jingrui He [1]

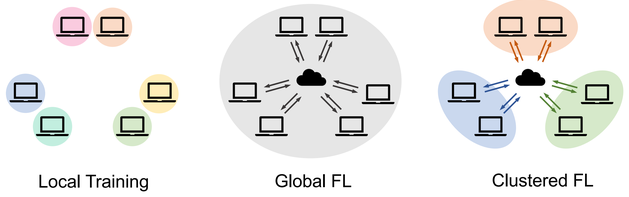
[1]University of Illinois Urbana-Champaign

AIFARMS Artificial Intelligence for Future Agricultural Resilience, Management, and Sustainability

Background

We consider an FL system with N clients, each client i has m_i samples $\hat{\mathcal{D}}_i = \{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k=1}^{m_i}$ i.i.d. drawn from its local distribution $\mathcal{D}_i$. h is the ML model and ℓ is the risk function. Denote the quantity distribution $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_N] = [\frac{m_1}{m}, \cdots, \frac{m_N}{m}]$ where $m = \sum_{i=1}^N m_i$.

- Client i's *local empirical risk*: $\hat{\epsilon}_i(h) = \frac{1}{m_i}\sum_{k=1}^{m_i} \ell(h(\boldsymbol{x}_k^{(i)}), \boldsymbol{y}_k^{(i)})$
- Client i's *local expected risk*: $\epsilon_i(h) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}_i}\ell(h(\boldsymbol{x}), \boldsymbol{y})$



Negative transfer When clients have non-IID data, the global model $\hat{h}_\beta$ significantly degrades, and even performs worse than the local model $\hat{h}_i$, i.e.,

$$\epsilon_i(\hat{h}_{\boldsymbol{\beta}}) > \epsilon_i(\hat{h}_i), \quad \text{for some clients } i, \quad \text{where } \hat{h}_{\boldsymbol{\beta}} = \arg\min_{h\in\mathcal{H}} \sum_{j=1}^N \beta_j \hat{\epsilon}_j(h), \hat{h}_i = \arg\min_{h\in\mathcal{H}} \hat{\epsilon}_i(h)$$

Reason: Distribution shift - the model is trained with $\sum_{i=1}^N \beta_i \hat{\mathcal{D}}_i$, but tested on $\mathcal{D}_i$.

Clustered FL alleviates negative transfer by grouping clients into coalitions. Each client only shares model with clients in the same coalition.

Question: What is the *optimal collaboration structure*, i.e., which clients should collaborate to train shared model?

Analysis

Theorem 3.3 (informal). Let $\hat{h}_{\boldsymbol{\alpha}_i} = \arg\min_{h\in\mathcal{H}} \sum_{j=1}^N \alpha_{ij}\hat{\epsilon}_j(h)$ where $\sum_{j=1}^N \alpha_{ij} = 1$. With probability at least $1-2\delta$,

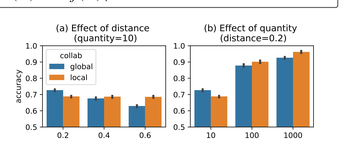
$$\epsilon_i(\hat{h}_{\boldsymbol{\alpha}_i}) \le \underbrace{\epsilon_i(h_i^*)}_{\text{irreducible error}} + 2\underbrace{\phi_{|\mathcal{H}|}(\boldsymbol{\alpha}_i, \boldsymbol{\beta}, m, \delta)}_{\text{quantity-aware function}} + 2\sum_{j\neq i}\alpha_{ij} \underbrace{D(\mathcal{D}_i, \mathcal{D}_j)}_{\text{distribution distance}}$$

where

- the quantity-aware function $\phi_{|\mathcal{H}|}(\boldsymbol{\alpha}_i, \boldsymbol{\beta}, m, \delta) = \sqrt{\left(\sum_{j=1}^N \frac{\alpha_{ij}^2}{\beta_j}\right)\left(\frac{2d\log(2m+2)+\log(4/\delta)}{m}\right)}$, and
- the distribution distance $D(\mathcal{D}_i, \mathcal{D}_j) = \max_{h\in\mathcal{H}} |\epsilon_i(h) - \epsilon_j(h)|$.

The error bound of client i is controlled by

- Collaboration structure $\boldsymbol{\alpha}_i = [\alpha_{i1}, \cdots, \alpha_{iN}]$
- Pairwise distribution distances $D(\mathcal{D}_i, \mathcal{D}_j)$
- Quantity distribution $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_N]$



Which collaboration structure minimizes the error bound? (Corollary 3.5)

- Clients prefer collaborators with smaller distribution distances.
- Clients with more data are pickier in the choice of collaborators.

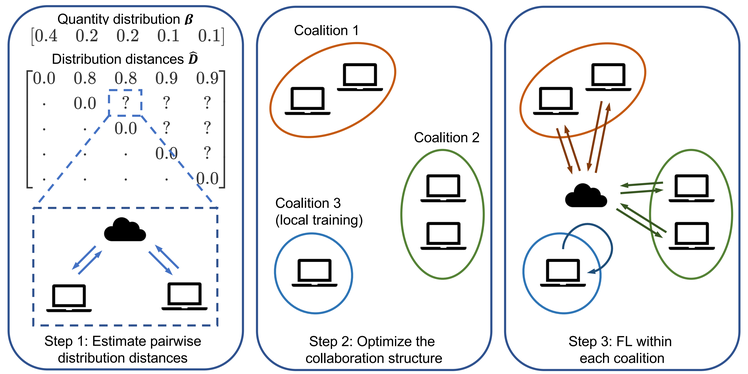
Paper Summary

FedCollab alleviates negative transfer in federated learning by clustering clients into non-overlapping coalitions based on their distribution distances and data quantities.

- **Theory**: We analyze how clustered FL performance is affected by two key factors: distribution distance and data quantity.
- **Algorithm**: We propose FedCollab to solve for the best collaboration structure.
- **Extensive experiments**: We test FedCollab under label shift, feature shift and concept shift with various models / datasets.

Proposed Method

FedCollab solves the collaboration structure by *minimizing an empirical estimation of the error bound in Theorem 3.3.*



Step 1: Estimate pairwise distribution distance Train a client discriminator $f : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ for each pair of clients with FL algorithm.

$$\hat{D}_{ij} = 2 \cdot \text{BalancedAccuracy}(f, \{\hat{\mathcal{D}}_i^{\text{valid}}, 1\} \cup \{\hat{\mathcal{D}}_j^{\text{valid}}, 0\}) - 1$$

- When $\mathcal{D}_i, \mathcal{D}_j$ are distinctly different, balanced accuracy $\approx 100\%$ and $\hat{D}_{ij} \approx 1$.
- When $\mathcal{D}_i, \mathcal{D}_j$ are similar, the classifier cannot outperform random guessing whose balanced accuracy $\approx 50\%$, thus $\hat{D}_{ij} \approx 0$.

Step 2: Optimize the collaboration structure Minimize the sum of empirical error upper bounds with greedy method.

$$\mathcal{L}(\boldsymbol{A}, \boldsymbol{\beta}, m, \hat{\boldsymbol{D}}) = \sum_{i=1}^N \left(\frac{C}{\sqrt{m}}\sqrt{\sum_{j=1}^N \frac{\alpha_{ij}^2}{\beta_j}} + \sum_{j=1}^N \alpha_{ij}\hat{D}_{ij}\right)$$

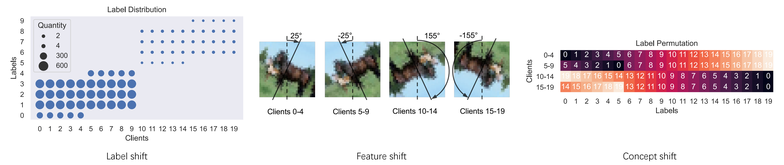
where C is a hyperparameter and the collaboration structure

$$\boldsymbol{A}_{ij} = \alpha_{ij} = \begin{cases} \frac{\beta_j}{\sum_{l\in\mathcal{C}_k}\beta_l}, & \text{if } i \in \mathcal{C}_k, j \in \mathcal{C}_k, \exists \text{ coalition } k \\ 0, & \text{otherwise} \end{cases}$$

Step 3: FL within each coalition Notice that since the collaboration structure and the FL model are optimized independently, FedCollab can be seamlessly integrated with any GFL or PFL algorithms in this stage.

Experiments

Setup 20 clients with unbalanced data quantities: $\beta_0 = \cdots = \beta_9 \gg \beta_{10} = \cdots = \beta_{19}$



FedCollab alleviates negative transfer for both global FL and personalized FL

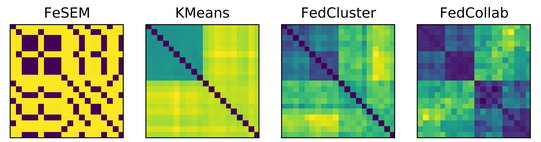
Method	Label Shift (FashionMNIST) Acc ↑	IPR ↑	RSD ↓	Feature Shift (CIFAR-10) Acc ↑	IPR ↑	RSD ↓	Concept Shift (CIFAR-100) Acc ↑	IPR ↑	RSD ↓
Local Train	86.05 ± 0.28	-	-	38.65 ± 0.44	-	-	29.82 ± 0.56	-	-
FedAvg	46.64 ± 0.12	46.00 ± 2.24	41.03 ± 0.24	44.31 ± 0.98	86.00 ± 4.18	4.62 ± 0.58	26.62 ± 0.12	50.00 ± 0.00	11.54 ± 0.45
+FedCollab	92.45 ± 0.07	100.00 ± 0.00	5.99 ± 0.41	52.61 ± 0.60	100.00 ± 0.00	3.30 ± 0.63	40.94 ± 0.22	100.00 ± 0.00	2.78 ± 0.30
FedProx	46.70 ± 0.08	45.00 ± 5.00	41.09 ± 0.29	44.45 ± 0.58	87.00 ± 4.47	4.74 ± 0.56	26.78 ± 0.14	50.00 ± 0.00	11.66 ± 0.36
+FedCollab	92.39 ± 0.15	100.00 ± 0.00	6.02 ± 0.37	52.73 ± 0.64	100.00 ± 0.00	3.16 ± 0.61	40.99 ± 0.17	100.00 ± 0.00	2.79 ± 0.34
FedNova	75.92 ± 1.14	45.00 ± 3.54	12.38 ± 1.25	46.98 ± 0.57	99.00 ± 2.24	3.42 ± 0.22	26.46 ± 0.13	50.00 ± 0.00	10.57 ± 0.32
+FedCollab	92.47 ± 0.13	100.00 ± 0.00	5.97 ± 0.39	52.72 ± 0.57	100.00 ± 0.00	3.18 ± 0.63	40.92 ± 0.36	100.00 ± 0.00	2.75 ± 0.43
Finetune	67.32 ± 3.17	48.00 ± 2.74	22.97 ± 2.82	44.17 ± 0.99	82.00 ± 2.74	5.14 ± 0.32	33.30 ± 4.79	50.00 ± 0.00	13.95 ± 0.57
+FedCollab	92.57 ± 0.15	99.00 ± 2.24	6.07 ± 0.30	51.53 ± 0.61	100.00 ± 0.00	2.92 ± 0.46	40.94 ± 2.36	100.00 ± 0.00	2.54 ± 0.30
Per-FedAvg	51.13 ± 4.10	49.00 ± 2.24	37.35 ± 4.15	43.78 ± 0.69	83.00 ± 9.08	4.74 ± 0.65	27.39 ± 0.24	50.00 ± 0.00	12.24 ± 0.46
+FedCollab	92.16 ± 0.25	97.00 ± 6.71	6.00 ± 0.25	52.64 ± 0.45	100.00 ± 0.00	3.03 ± 0.30	41.04 ± 0.26	100.00 ± 0.00	2.85 ± 0.49
pFedMe	55.31 ± 3.45	47.00 ± 4.47	33.71 ± 3.11	39.74 ± 0.85	60.00 ± 12.25	4.81 ± 0.74	27.04 ± 0.39	48.00 ± 2.74	10.39 ± 0.47
+FedCollab	92.18 ± 0.43	99.00 ± 2.24	6.40 ± 0.81	47.20 ± 1.29	97.00 ± 2.74	3.02 ± 0.30	37.47 ± 0.31	100.00 ± 0.00	3.04 ± 0.23
Ditto	68.73 ± 1.40	48.00 ± 2.74	20.29 ± 2.06	47.04 ± 0.30	97.00 ± 2.74	3.85 ± 0.35	32.50 ± 0.40	50.00 ± 0.00	12.22 ± 0.36
+FedCollab	92.55 ± 0.08	99.00 ± 2.24	6.11 ± 0.30	50.97 ± 0.75	99.00 ± 2.24	3.38 ± 1.55	40.33 ± 0.33	100.00 ± 0.00	2.16 ± 0.30

- IPR - % of clients with $\epsilon_i(\hat{h}_{\boldsymbol{\alpha}_i}) < \epsilon_i(\hat{h}_i)$, i.e., FL model is better than local model
- RSD - standard deviation of $\{\epsilon_i(\hat{h}_i) - \epsilon_i(\hat{h}_{\boldsymbol{\alpha}_i})\}_{i=1}^N$, i.e., whether clients have uniform accuracy gains

FedCollab outperforms previous clustered FL algorithms

Method	Label Shift (FashionMNIST) Acc ↑	IPR ↑	RSD ↓	Feature Shift (CIFAR-10) Acc ↑	IPR ↑	RSD ↓	Concept Shift (CIFAR-100) Acc ↑	IPR ↑	RSD ↓
IFCA	91.49 ± 0.61	95.00 ± 5.00	5.62 ± 0.54	49.78 ± 1.01	100.00 ± 0.00	3.13 ± 0.52	30.74 ± 4.46	60.00 ± 22.36	11.28 ± 5.04
FedCluster	92.07 ± 0.47	95.00 ± 7.07	6.14 ± 0.49	44.86 ± 1.90	79.00 ± 17.10	5.64 ± 1.81	29.23 ± 2.18	62.00 ± 12.55	9.55 ± 0.69
FeSEM	56.79 ± 6.71	45.00 ± 11.18	36.12 ± 2.08	42.73 ± 0.37	82.00 ± 5.70	4.10 ± 0.62	31.92 ± 3.12	72.00 ± 12.55	9.81 ± 1.77
KMeans	69.30 ± 0.81	72.00 ± 2.74	35.87 ± 1.22	48.61 ± 1.15	96.00 ± 4.18	4.54 ± 0.74	34.24 ± 3.01	85.00 ± 13.69	6.47 ± 3.06
FedCollab	92.45 ± 0.07	100.00 ± 0.00	5.99 ± 0.41	52.61 ± 0.60	100.00 ± 0.00	3.30 ± 0.63	40.94 ± 0.22	100.00 ± 0.00	2.78 ± 0.30

- FedCollab explicitly uses the quantity distribution during collaboration optimization.
- FedCollab estimates high quality distribution distances.



Acknowledgement This work is supported by National Science Foundation under Award No. IIS-1947203, IIS-2117902, IIS-2137468, and Agriculture and Food Research Initiative (AFRI) grant no. 2020-67021-32799/project accession no.1024178 from the USDA National Institute of Food and Agriculture. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.