Collaborative Learning by Detecting Collaboration Partners

Shu Ding, Wei Wang

{dings, wangw}@lamda.nju.edu.cn

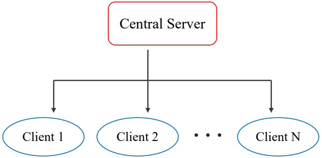
NANJING UNIVERSITY

NeurIPS 2022

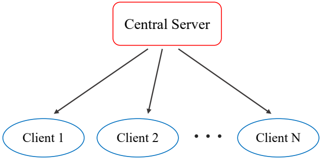
Collaborative Learning

- Massive amounts of data are naturally dispersed over numerous clients. Each client only has limited data.
- **Collaborative learning** is a promising paradigm that enables the clients to learn models through collaboration.

- Two settings



- *Centralized model*: return one single model for all clients



- *Personalized model*: return different models for different clients

Motivation

- Centralized model

  One single model may perform badly on clients whose distributions are different from the average distribution.

- Personalized model

  Learning personalized models is impractical when the number of clients N is very large since this costs unaffordable computational resources.

- Can we return K $(K \ll N)$ appropriate models for N heterogeneous clients and expect that the returned models have comparable performance to personalized models?

Setting

- Preliminaries
  - Clients $\{C_1, \ldots, C_N\}$ with distributions $\{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$
  - Each client C_i has access to m_i examples $S_i = \left\{\left(\boldsymbol{x}_1^i, y_1^i\right), \ldots, \left(\boldsymbol{x}_{m_i}^i, y_{m_i}^i\right)\right\}$ drawn from $\mathcal{D}_i$
  - Total number of examples $M = \sum_{i=1}^N m_i$
- Collaborative learning scenario
  - Train the model over the weighted union of all samples $S_{\boldsymbol{\alpha}} = \sum_{j=1}^N \alpha_j S_j$
  - The model for C_i can be learned by minimizing $\hat{\mathcal{L}}_{\boldsymbol{\alpha}_i}(h) = \sum_{j=1}^N \alpha_{ij} \hat{\mathcal{L}}_{S_j}(h)$ with collaboration vector $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iN}) \in \Delta^N$

Theoretical Analysis

Theorem (Generalization Bound)

Let $\mathcal{H}$ be the hypothesis space with VC-dimension d. Denote $h_i^\star = \arg\min_{h\in\mathcal{H}} \mathcal{L}_{\mathcal{D}_i}(h)$ *and* $\hat{h}_{\boldsymbol{\alpha}_i} = \arg\min_{h\in\mathcal{H}} \hat{\mathcal{L}}_{\boldsymbol{\alpha}_i}(h)$*. For any given* $\delta \in (0,1)$ *and* $\forall i \in \{1,\dots,N\}$*, with probability at least* $1-\delta$*:*

$$\mathcal{L}_{\mathcal{D}_i}(\hat{h}_{\boldsymbol{\alpha}_i}) - \mathcal{L}_{\mathcal{D}_i}(h_i^\star) \leqslant 2\sum_{j=1}^{N} \alpha_{ij} d_{\mathcal{H}}(\mathcal{D}_i,\mathcal{D}_j) + 2\mu\sqrt{\sum_{j=1}^{N}\frac{\alpha_{ij}^2}{m_j}}\sqrt{8(d\log(2M) + \log\frac{8}{\delta})}.$$

Here $d_{\mathcal{H}}(\mathcal{D}_i,\mathcal{D}_j) = \sup_{h\in\mathcal{H}} \left|\mathcal{L}_{\mathcal{D}_i}(h) - \mathcal{L}_{\mathcal{D}_j}(h)\right|$ *is the Integral Probability Metrics (IPM).*

Theoretical Analysis

Theorem (Optimal Collaboration Vector)

Let $\Xi_i^j = d_{\mathcal{H}}(\mathcal{D}_i, \mathcal{D}_j)$ *and* $\lambda = \mu\sqrt{8(d\log(2M) + \log\frac{8}{\delta})}$. *For client* C_i, *sort* $\{\Xi_i^1, \dots, \Xi_i^N\}$ *in ascending order to get* $\{\Xi_i^{\sigma(1)}, \dots, \Xi_i^{\sigma(N)}\}$. *The optimal* $\boldsymbol{\alpha}_i^\star$ *for client* C_i *is given by*

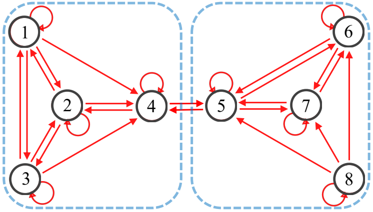
$$\alpha_{ij}^\star = \left[\frac{m_j(\zeta - \Xi_i^j)}{\sum_{q \leqslant q_i} m_{\sigma(q)}(\zeta - \Xi_i^{\sigma(q)})}\right]_+ .$$

Here $[\cdot]_+ = max(\cdot, 0)$, ζ *is the larger root of equation* $\sum_{q \leqslant q_i} m_{\sigma(q)}\left(\zeta - \Xi_i^{\sigma(q)}\right)^2 = \lambda^2$, *and*

$$q_i = \arg\max_t \left\{t \middle| \zeta \geqslant \Xi_i^{\sigma(t)} \wedge \left(\sum_{q \leqslant t} m_{\sigma(q)} \Xi_i^{\sigma(q)}\right)^2 \geqslant \left(\sum_{q \leqslant t} m_{\sigma(q)}\right)\left(\sum_{q \leqslant t} m_{\sigma(q)} (\Xi_i^{\sigma(q)})^2 - \lambda^2\right)\right\}.$$

- $\hat{h}_{\boldsymbol{\alpha}_i^\star}$ with respect to the optimal $\boldsymbol{\alpha}_i^\star$ is referred as the *personalized* model for client C_i

- In the directed graph A, $\alpha^\star_{ij} > 0$ means C_j is beneficial to C_i. Clients with similar incoming edges are called **collaboration partners** since they need similar contribution from other clients.
- Intuitively, collaboration partners should be in the same group. We could probably return the same model for C_1, C_2, C_3, C_4 while it is inappropriate to return the same model for C_4, C_5.



In graph $A = (V, E)$, $|V| = N$, node i denotes C_i and the weight of edge from j to i is $\alpha^\star_{ij}$.

For the General Case

- Collaboration with Modularity Maximization
  - Construct matrix $\mathbf{U}$ to evaluate the **incoming-edge similarity** among clients

$$\mathbf{U} = \mathbf{D}_{in}^{-\beta}\mathbf{A}\mathbf{A}^{T}\mathbf{D}_{in}^{-\beta}$$

  - Use **Modularity** as the objective function to evaluate the quality of group partitions

$$Q(\mathcal{G}) = \frac{1}{2W}\sum_{i,j}\left[w_{ij} - \frac{d_i d_j}{2W}\right]\delta(g_i, g_j)$$

  - Relax the modularity maximization problem as a **SemiDefinite Programming**

$$\begin{aligned} \max \quad & \sum_{\mathcal{M}^+}\mathcal{M}_{ij}\boldsymbol{\nu}_i\cdot\boldsymbol{\nu}_j + \sum_{\mathcal{M}^-}-\mathcal{M}_{ij}\left(1-\boldsymbol{\nu}_i\cdot\boldsymbol{\nu}_j\right) \\ s.t. \quad & \boldsymbol{\nu}_i\cdot\boldsymbol{\nu}_i = 1, \forall i \in \{1,\ldots,N\}; \quad \boldsymbol{\nu}_i\cdot\boldsymbol{\nu}_j \geqslant 0, \quad \forall i \neq j, \\ & \boldsymbol{\nu}_i \in \mathbb{R}^K, \forall i \in \{1,\ldots,N\}. \end{aligned}$$

For the General Case

- Collaboration with Modularity Maximization
  - **Find reasonable group partitions** by solving the SDP
    Given matrix $\mathbf{U}$, let $Q(\mathcal{G})$ be the modularity value of the group partition $\mathcal{G}$ obtained by solving the SDP using rounding techniques. Then $Q(\mathcal{G}) > \kappa \mathrm{OPT}_{Q(\mathcal{G})} - (1-\kappa)$ where $\kappa = 0.766$ is the approximation factor.
  - **Detect bad clients**
    Edge $e_{ij} \in \mathbf{U}$ is a *weak edge* if its weight $w_{ij} < \frac{1}{N}$. A group is divided into several disjoint parts after removing all weak edges within the group. Clients do not belong to the largest part are *bad clients*. Bad clients cannot be provided with good performance guarantee.

For the General Case

- Collaboration with Modularity Maximization

  - **Number of bad clients**

    Given the group partition $\mathcal{G} = \{G_1, \ldots, G_K\}$ returned by Algorithm ACLMM, assume $N_k \geqslant 2\sqrt{Z_{in}}, \forall k \in \{1, \ldots, K\}$. Let $N_{min} = \min_k N_k$, then $|\mathcal{B}| \leqslant \frac{N_{min} - \sqrt{N_{min}^2 - 4Z_{in}}}{2}$, where $Z_{in} \leqslant \frac{N}{2(N-1)} \left[\frac{N^2 - KN}{K} - 2W\left((\kappa + 1)\mathrm{OPT}_{Q(\mathcal{G})} - \frac{K-1}{K} \right) \right]$.

  - **Theoretical guarantee**

    Let $\mathcal{G} = \{G_1, \ldots, G_K\}$ be the group partition returned by solving the SDP. $\hat{h}_{\boldsymbol{\alpha}_{G_k}}$ is the model returned by Algorithm ACLMM for client C_i in group G_k. $\mathrm{upp}(\hat{h}_{\boldsymbol{\alpha}_{G_k}})$ is the upper bound of the expected risk of $\hat{h}_{\boldsymbol{\alpha}_{G_k}}$ and $\mathrm{upp}(\hat{h}_{\boldsymbol{\alpha}_i^\star})$ is the upper bound of the expected risk of the personalized model $\hat{h}_{\boldsymbol{\alpha}_i^\star}$. The following result holds except for the bad clients in $\mathcal{B}$:

$$\mathrm{upp}(\hat{h}_{\boldsymbol{\alpha}_{G_k}}) - \mathrm{upp}(\hat{h}_{\boldsymbol{\alpha}_i^\star}) \leqslant O\left(\eta(1-\tau)\sqrt{\frac{N}{N-1}} \right).$$

For the Special Case

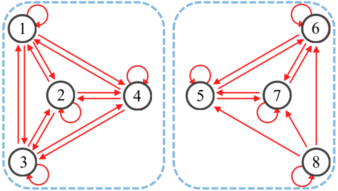
- Collaboration with Clustering

- **Potential structures**

  There exists a potential partition $\mathcal{P}^\star = \{P_1^\star, \dots, P_K^\star\}$ s.t. $\Phi(\mathcal{P}) = \sum_{k=1}^{K} \sum_{C_i \in P_k} d\left(\boldsymbol{\alpha}_i^\star, \bar{\boldsymbol{\alpha}}_k\right)$ is small. Assume that $\{\boldsymbol{\alpha}_1^\star, \dots, \boldsymbol{\alpha}_N^\star\}$ satisfy $(1+\gamma,\ \epsilon)$-*approximation-stability* property.

- **Detect bad clients**

  $\bar{d} = \frac{1}{N}\mathrm{OPT}_{\Phi(\mathcal{P})}$ is the average distance. $d^\star = \frac{\gamma \bar{d}}{\epsilon t}$ is the critical distance. C_i is the *bad client* if $d_1(\boldsymbol{\alpha}_i^\star) \geqslant d^\star$ or $d_2(\boldsymbol{\alpha}_i^\star) - d_1(\boldsymbol{\alpha}_i^\star) \leqslant \frac{t}{2} d^\star$.



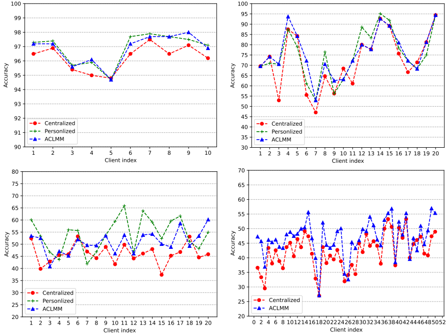
The example here has better structures than the aforementioned example.

For the Special Case

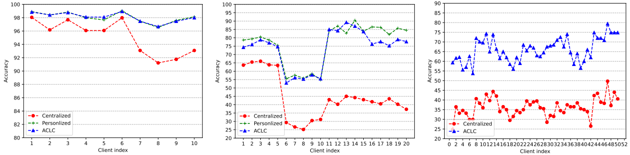
- Collaboration with Clustering
  - **Number of bad clients**
    Let $\mathcal{P} = \{P_1, \ldots, P_K\}$ be the group partition produced by Algorithm ACLC. Then $|\mathcal{B}| < (6 + \frac{t}{\gamma})\beta\epsilon N$ where $t > 2$ and $\beta > 1$ are given constants.
  - **Theoretical guarantee**
    Let $\mathcal{P} = \{P_1, \ldots, P_K\}$ be the group partition produced by Algorithm ACLC. $\hat{h}_{\boldsymbol{\alpha}_{P_k}}$ is the model returned by Algorithm ACLC for client C_i in group P_k. $\mathrm{upp}(\hat{h}_{\boldsymbol{\alpha}_{P_k}})$ is the upper bound of the expected risk of $\hat{h}_{\boldsymbol{\alpha}_{P_k}}$ and $\mathrm{upp}(\hat{h}_{\boldsymbol{\alpha}_i^\star})$ is the upper bound of the expected risk of the personalized model $\hat{h}_{\boldsymbol{\alpha}_i^\star}$. The following result holds except for the bad clients in $\mathcal{B}$:

$$upp(\hat{h}_{\boldsymbol{\alpha}_{P_k}}) - upp(\hat{h}_{\boldsymbol{\alpha}_i^\star}) \leqslant O\left(\frac{\gamma \mathrm{OPT}_{\Phi(\mathcal{P})}}{\epsilon t N}\right).$$

Experimental Results



- The model learned with ACLMM performs better than the centralized model and is comparable to the personalized model.



- The model learned with ACLC performs much better than the centralized model and is comparable to the personalized model.
- The gap between the model return by ACLC and the personalized model is small.

Thank you!