Maverick Matters

Client Contribution and Selection in Federated Learning

Huang, Jiyue; Hong, Chi; Liu, Yang; Chen, Lydia Y.; Roos, Stefanie

DOI
10.1007/978-3-031-33377-4_21

Publication date
2023

Document Version
Final published version

Published in
Advances in Knowledge Discovery and Data Mining - 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2023, Proceedings

Maverick Matters: Client Contribution and Selection in Federated Learning

Jiyue Huang[1], Chi Hong[1], Yang Liu[2], Lydia Y. Chen[1](✉), and Stefanie Roos[1](✉)

[1] Delft University of Technology, Delft, The Netherlands
{j.huang-4,c.hong,y.chen-10,s.roos}@tudelft.nl
[2] AI Industry Research (AIR), Tsinghua University, Beijing, China
liuy03@air.tsinghua.edu.cn

Abstract. Federated learning (FL) enables collaborative learning between parties, called clients, without sharing the original and potentially sensitive data. To ensure fast convergence in the presence of such heterogeneous clients, it is imperative to timely select clients who can effectively contribute to learning. A realistic but overlooked case of heterogeneous clients are Mavericks, who monopolize the possession of certain data types, e.g., children hospitals possess most of the data on pediatric cardiology. In this paper, we address the importance and tackle the challenges of Mavericks by exploring two types of client selection strategies. First, we show theoretically and through simulations that the common contribution-based approach, *Shapley Value*, underestimates the contribution of Mavericks and is hence not effective as a measure to select clients. Then, we propose FEDEMD, an adaptive strategy with competitive overhead based on the Wasserstein distance, supported by a proven convergence bound. As FEDEMD adapts the selection probability such that Mavericks are preferably selected when the model benefits from improvement on rare classes, it consistently ensures the fast convergence in the presence of different types of Mavericks. Compared to existing strategies, including *Shapley Value*-based ones, FEDEMD improves the convergence speed of neural network classifiers with FedAvg aggregation by 26.9% and its performance is consistent across various levels of heterogeneity.

Keywords: Federated learning · data heterogeneity · client selection · shapley value · wasserstein distance

1 Introduction

Federated Learning (FL) enables clients (either individuals or institutes who own data) to collaboratively train a global machine learning models by exchanging locally trained models instead of data [16,18]. Thus, Federated Learning allows the training of models when data cannot be transferred to a central server and is hence often a suitable alternative for medical research and other domains, such as finance, with high privacy requirements. The effectiveness of FL, in terms of accuracy and convergence, highly depends on how the local models are selected and aggregated.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-33377-4_21.

H. Kashima et al. (Eds.): PAKDD 2023, LNAI 13936, pp. 269–282, 2023.
https://doi.org/10.1007/978-3-031-33377-4_21

In FL, clients tend to own heterogeneous datasets [14] rather than identically and independent distributed (*i.i.d.*) ones. The prior art has recently addressed the challenge of heterogeneity from either the perspective of skewed distribution [28] or skewed quantity [23] among all clients. However, a common real-world scenario, where one or a small group of clients monopolize the possession of a certain class, is universally overlooked. For example, in the widely used image classification benchmark, Cifar-10 [12], most people can contribute images of cats and dogs. However, deer images are bound to be owned by comparably few clients. We call these types of clients *Mavericks*. Another relevant example, shown in Fig. 1, arises from learning predictive medicine from clinics who specialize in different conditions, e.g., AIDS and Amyotrophic Lateral Sclerosis, and own data of exclusive disease types. Without involving Mavericks into the training, it is impossible to achieve high accuracy on the classes for which they own the majority of all training data, e.g., rare diseases.

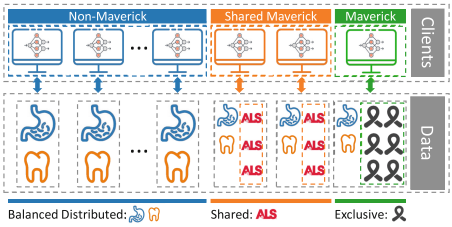


Fig. 1. Illustration of Mavericks.

Given its importance, it is not well understood when to best involve Mavericks in FL training, because the effectiveness of FL, in terms of accuracy and convergence, highly depends on how those local models are selected and aggregated. The existing client selection[1] considers either the contribution of local models [3] or difference of data distributions [19]. The contribution-based approaches select clients based on contribution scores preferring clients with higher scores [7], whereas the distance-based methods choose clients based on the pairwise feature distance. Both types of selection methodologies have their suitable application scenarios and it is hard to weigh the benefits of one over the other in general.

In this paper, we aim to effectively select Mavericks in FL so that users are able to collaboratively train an accurate model in a low number of communication rounds. We first explore *Shapley Value* as a contribution metric for client selection. Although *Shapley Value* is shown to be effective in measuring contribution for the *i.i.d.* case, it is unknown if it can assess the contribution of Mavericks and effectively involve them via the selection strategy. Moreover, we propose FEDEMD, which selects clients based on Wasserstein distance [2] of the global distribution and current distribution. As FEDEMD adapts the selection probability such that Mavericks are preferably selected when the model benefits from improvement on rare classes, it consistently ensures the fast convergence in the presence of different types of Mavericks.

Our main **contributions** for this work can be summarized as follows. *i)* We explore the effectiveness of both contribution-based and distance-based selection strategies for Mavericks. *ii)* Both our theoretical and empirical results show that the contribution of clients with skewed data or very large data quantity is measured below average by *Shapley Value*. *iii)* We propose FEDEMD, a novel adaptive client selection based on the Wasserstein distance, derive a convergence bound, and show that it significantly outperforms SOTA selection methods in terms of convergence speed across different scenarios of Mavericks.

[1] Note that here we only discuss selection on statistical challenges, the selections considering system resources, e.g., unreliable networks are left for other works.

2 Related Studies

Contribution Measurement. Although the self-reported contribution evaluation [7] is easy to implement, it is fragile too dishonest parties. Besides, existing work on contribution measurement can be categorized into two classes: *i)* local approach: clients exchange the local updates, i.e., model weights or gradients, and measure the contribution of each other, e.g., by creating a reputation system [11], and *ii)* global approach: all clients send all their model updates to the *federator* who in turn aggregates and computes the contribution via the marginal loss [1,25]. Prevailing examples of globally measuring contribution are Influence [1] and *Shapley Value* [22,25]. The prior art demonstrates that *Shapley Value* can effectively measure the client's contribution for the case when client data is *i.i.d.* or of biased quantity [22]. A work [24] has proposed federated *Shapley Value* to capture the effect of participation order on data value. The experimental results indicate that *Shapley Value* is less accurate in estimating the contribution of heterogeneous clients than for *i.i.d.* cases. However, there is no rigorous analysis on whether *Shapley Value* can effectively evaluate the contribution from heterogeneous users with skewed data distributions.

Client Selection. Selecting clients within a heterogeneous group of potential clients is key to enabling fast and accurate learning based on high data quality. The state-of-the-art client selection strategies focus on the resource heterogeneity [10,21] or data heterogeneity [3,4,14]. In case of data heterogeneity, which is the focus of our work, selection strategies [3,4,8] gain insights on the distribution of clients' data and then select them in specific manners. Goetz et. al [8] apply active sampling and Cho et. al [4] use Power-of-Choice to favor clients with higher local loss. TiFL [3] considers both resource and data heterogeneity to mitigate the impact of stragglers and skewed distributions. TiFL applies a contribution-based client selection by evaluating the accuracy of selected participants each round and chooses clients of higher accuracy. FedFast [19] chooses classes based on clustering and achieves fast convergence for recommendation systems. One recently work [17] focuses on reduce wall-clock time for convergence under high degrees of system and statistical heterogeneity. However, there is no selection strategy that addresses the Maverick scenario.

3 Federated Learning with Mavericks

In this section, we first formalize a Federated Learning framework with Mavericks. Then we rigorously analyze the contribution of clients based on *Shapley Value* and argue that the contribution of Mavericks is underestimated by the *Shapley Value*, which leads to a severe selection bias and a suboptimal integration of Mavericks into the learning process.

Suppose there are a total of N clients in a federated learning system. We denote the set of possible inputs as $\mathcal{X}$ and the set of L class labels as $\mathcal{Y} = \{1, 2, ..., L\}$. Let $f\colon \mathcal{X} \rightarrow \mathcal{P}$ be a prediction function and ω be the learnable weights of the machine learning tasks, the objective is then defined as: $\min \mathcal{L}(\boldsymbol{\omega}) = \min \sum_{l=1}^{L} p(y = l)\mathbb{E}_{\boldsymbol{x}|y=l}\left[\log f_l(\boldsymbol{x}, \boldsymbol{\omega})\right]$.

The training process of a FL system has the following steps[2]: *i)* INITIALIZATION. Initialize global model $\boldsymbol{\omega}_0$ and distribute it to the available clients, i.e., a set $\mathcal{C}$ of N clients. *ii)* CLIENT SELECTION. Enumerate the K clients $\mathcal{C}(\pi, \boldsymbol{\omega}_r)$, selected in round r with selection strategy π, by $C_1, \ldots, C_K$. *iii)* UPDATE AND UPLOAD. Each client C_k selected in round r computes local updates $\boldsymbol{\omega}_r^k$ and the *federator* aggregates the results. Concretely, with η being the learning rate, C_k updates their weights in the r-th global round by: $\boldsymbol{\omega}_r^k = \boldsymbol{\omega}_{r-1} - \eta \sum_{l=1}^{L} p^k(y=l) \nabla_{\boldsymbol{\omega}} \mathbb{E}_{\boldsymbol{x}|y=l} \left[\log f_l(\boldsymbol{x}, \boldsymbol{\omega}_{r-1})\right]$. *iv)* AGGREGATION. Client updates are aggregated to one global update. The most common aggregation method is quantity-aware FedAvg, defined as follows with n^k indicating the data quantity of C_k: $\boldsymbol{\omega}_r = \sum_{k=1}^{K} \frac{n^k}{\sum_{k=1}^{K} n^k} \boldsymbol{\omega}_r^k$. To facilitate our discussions, we also define the following:

Local Distribution: The array of all L class quantities $\mathcal{D}^i(y=l), l \in \{1, .., L\}$ owned by client C_i.

Global Distribution: The quantity of all clients' data by class as $\mathcal{D}_g = \sum_{i=1}^{N} \mathcal{D}^i(y = l), l \in \{1, .., L\}$.

Current Distribution at R: By summing up the class quantity of all clients' data reported, which have been chosen up to round R as: ${\mathcal{D}_c}^R = \sum_{t=1}^{R} \sum_{C_k \in \mathcal{K}^t} \mathcal{D}^{C_k}$.

Definition 1 (Maverick). *Let Y_{Mav} be the set of class labels that are primarily owned by Mavericks. An exclusive Maverick is one client that owns one or more classes exclusively. A shared Maverick is a small group of clients who jointly own one class exclusively. That is:*

$$D_i = \begin{cases} \{\{x_l, y_l\}^i_{l \in Y_{Mav}}, \{x_l, y_l\}^i_{l \notin Y_{Mav}}\}, \textit{if } C_i \textit{ is a Maverick} \\ \{x_l, y_l\}^i_{l \notin Y_{Mav}}, \textit{if } C_i \textit{ is not a Maverick}, \end{cases} \tag{1}$$

where D_i denotes the dataset for C_i, $\{x_l, y_l\}^i$ denotes the dataset in C_i with label l.

In the rest of the paper, we assume the global distribution organized by the server's preprocessing has high similarity with the real-world (test dataset) distribution, which is balanced, so that data $\{x_l, y_l\}_{l \notin Y_{Mav}}$ are evenly distributed across all parties, whereas $\{x_l, y_l\}_{l \in Y_{Mav}}$ either belong to one exclusive Maverick or are evenly distributed across all shared Maverick parties. We focus our analysis on exclusive Mavericks since shared Maverick are a straightforward extension. Based on the assumptions above, we obtain the following properties for Mavericks.

Property 1. Because the data distribution is balanced, Mavericks have a larger data quantity than non-Mavericks. Concretely, let n^n be the data quantity of a non-Maverick. Let n^m be the quantity for Mavericks, then $n^m = ((N/m - 1) \times Y_{Mav} + L) \times n^n$, where m is the number of Mavericks.

Property 2. Assume $N > 2$, the KL divergence of a Maverick's data to the normalized global distribution is expected to be larger than for a non-Maverick due to their

[2] Here we assume all the clients are honest. Since we focus on the statistical challenge, the impact of unreliable networking and insufficient computation resources is ignored.

specific distribution, i.e., $D_{KL}(\mathcal{P}_g||\mathcal{P}_m)$ ¿ $D_{KL}(\mathcal{P}_g||\mathcal{P}_n)$, where $\mathcal{P}_m$, $\mathcal{P}_n$ are the data distribution with class labels for Maverick and non-Maverick, where $\mathcal{P}_g$ denotes for global distribution.

3.1 *Shapley Value* for Mavericks

Definition 2 (*Shapley Value*). *Let $\mathcal{K} = \mathcal{C}(\pi, \boldsymbol{\omega}_r)$ denote the set of clients selected in a round including C_k, $\mathcal{K} \setminus \{C_k\}$ denote the set $\mathcal{K}$ without C_k.* Shapley Value *of C_k is:*

$$SV(C_k) = \sum_{S \subseteq \mathcal{K} \setminus \{C_k\}} \frac{|S|!(|\mathcal{K}| - |S| - 1)!}{|\mathcal{K}|!} \delta C_k(\mathcal{S}). \tag{2}$$

Here we let $\delta C_k(S)$ be the Influence [1]. Influence can be defined on loss, accuracy, etc., here we apply the most commonly used loss-based Influence written as $Inf_S(C_k)$ for set C_k.

Lemma 1. *Based on Shapley Value in Eq. 2, the difference of Maverick C_m's and non-Maverick C_n's Shapley Value is:*

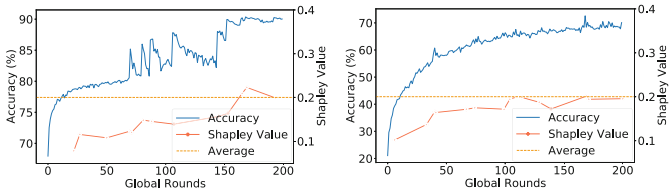
$$\begin{aligned} SV(C_m) - SV(C_n) = \frac{1}{|\mathcal{K}|!} \Bigg(& (|\mathcal{K}| - 1)!(\mathcal{L}(C_m) - \mathcal{L}(C_n)) \\ & + \sum_{S \subseteq S_-} |S|!(|\mathcal{K}| - |S| - 1)!(Inf_S(C_m) - Inf_S(C_n)) \\ & + \sum_{S \subseteq S_+} |S|!(|\mathcal{K}| - |S| - 1)!(Inf_S(C_m) - Inf_S(C_n)) \Bigg), \end{aligned} \tag{3}$$

with $S_- = \mathcal{K} \setminus \{C_n, C_m\}$, $S_+ = \mathcal{K} \setminus \{C_n, C_m\} \cup C_M$, $C_M \in \{C_n, C_m\}$. Note that we simplify $Inf_{S \cup C_i}(C_i)$ as $Inf_S(C_i)$ for readability.

Comparison of *Shapley Value* and Influence: Rather than considering Influence for the complete set of K clients, Eq. 3 only considers Influence on a subset S. However, our derivations for Influence are independent from the number of selected clients and remain applicable for subsets S, meaning that indeed the second and the third term of Eq. 3 are negative. Similarly, the first term is negative as the loss for clients only owning one class is higher. However, *Shapley Value* obtains higher values for *i.i.d.* clients with large data sets than Influence since $\mathcal{L}(C_m) - \mathcal{L}(C_n)$ increases if the distance between C_m's distribution and the global distribution is small, in line with a previous work [9].

Property 3. *Shapley Value* and Influence share the same trend in contribution measurement for Mavericks.

Theorem 1. *Let C_m and C_n be a Maverick and a non-Maverick client, respectively, and denote by $SV_t(C_k)$ the Shapley value of C_k in round r. Then $SV_1(C_m) < SV_1(C_n)$ and $SV_t(C_m)$ converges towards $SV_t(C_n)$.*



(a) FMNIST-Maverick (b) Cifar-10-Maverick

Fig. 2. Relative *Shapley Value* during training under multiple exclusive and shared Mavericks.

We present the empirical evidences of how one or multiple Mavericks are measured by *Shapley Value*. We here focus on single exclusive Mavericks and leave multiple Mavericks, shared and exclusive, for our in-depth experimental evaluation in the supplementary material. We use Fashion-MNIST (Fig. 2a) and Cifar-10 (Fig. 2b) as learning scenarios and use random client selection with FedAvg.

Figure 2 shows the global accuracy and the relative *Shapley Value* during training, with the average relative *Shapley Value* of the 5 selected clients out of 50 indicated by the dotted line. The contribution is only evaluated when a Maverick is selected. Looking at Fig.(2a, b), The *Shapley Value* of the Maverick indeed increases over time but remains below average until round 160, providing concrete evidence of **Theorem** 1. Furthermore, the accuracy increases when a Maverick is selected, indicating that Mavericks contribute highly to improving the model. Thus, assigning Mavericks a lower contribution measure is unreasonable, especially in the early stage of the learning process. All of the empirical results are consistent with our theoretical analysis.

4 *FEDEMD*

In this section, we propose a novel adaptive client selection algorithm FEDEMD, which enables FL systems with Mavericks to achieve faster convergence compared with SOTA methods, including *Shapley Value*-based ones. The key idea is to assign a higher probability for selecting Maverick clients initially to accelerate convergence; later we reduce the selection probability to avoid skewing the distribution towards Maverick classes. To measure the differences in data distributions, we adopt Wasserstein distance (EMD) [2], which is used to characterize weight divergence in FL [27]. The Wasserstein distance (EMD) is defined as:

$$\text{EMD}\left(P_r, P_\theta\right) = \inf_{\gamma \in \Pi} \sum_{x,y} \|x - y\| \gamma(x, y) = \inf_{\gamma \in \Pi} \mathbb{E}_{(x,y)\sim\gamma} \|x - y\|, \tag{4}$$

where $\Pi(P_r, P_\theta)$ represents the set of all possible joint probability distributions of P_r, P_θ. $\gamma(x, y)$ represents the probability that x appears in P_r and y appears in P_θ.

Overview. The complete algorithm is shown in Algorithm 1, we here summarize the different components that make up the algorithm. ***i) Data Reporting and Initialization*** (Line 1–3): Clients report their data quantity so that the *federator* is able to compute the global data size array $\mathcal{D}_g$ and initialize the current size array $\mathcal{D}_c^1$.

ii) Dynamic Weights Calculation (Line 4–11): In this key step, we utilize a light-weight measure based on EMD to calculate dynamic selection probabilities over time, which achieve faster convergence, yet avoid overfitting, concretely we compute

$$Proba^r = softmax(\widetilde{emd}_g - t\beta\widetilde{emd}_c^r) \tag{5}$$

Algorithm 1: FEDEMD Clients Selection

Data: $\mathcal{D}^i$ for $i \in 1, 2, ..., N$.
Result: $\mathcal{K}$: selected participants.
1 **Set:** distance coefficient $\beta > 0$;
2 initialize probability $Proba^1$;
3 initialize current distribution $\mathcal{D}_c^1$;
4 $\mathcal{D}_g \leftarrow \sum_{i=1}^{N} \mathcal{D}^i$;
5 calculate $\widetilde{emd}_g$ by Eq. 6;
6 **for** *round r = 1, 2, ..., R* **do**
7 $\quad \mathcal{K}^r = rand(K, \mathcal{C}, Proba^r)$
8 $\quad \mathcal{D}_c^{r+1} \leftarrow \mathcal{D}_c^r + \sum_{C_k \in \mathcal{K}}^{r} \mathcal{D}^{C_k}$;
9 $\quad$ calculate $\widetilde{emd}_c^r$ by Eq. 7;
10 $\quad$ **for** *client i = 1, ..., N* **do**
11 $\quad\quad$ update $Proba^{r+1}$ by Eq. 5

where $Proba_i^r$ is the probability for selecting C_i in round r. β is a coefficient to weigh the global and current distance and shall be adapted for different initial distributions, i.e., different dataset and distribution rules. $\widetilde{emd}_g$ and $\widetilde{emd}_c^r$ are the normalized EMDs between the global/current and local distributions (Line 5, 9), namely

$$\widetilde{emd}_g = Norm([EMD(\mathcal{D}_g, \mathcal{D}^i)\big|_{i \in \{1,...,N\}}]), \tag{6}$$

which is constant through the learning process as long as the local distribution of clients stays the same. The larger $\widetilde{emd}_g$ is, the higher the probability $Proba_i^r$ that a client C_i is selected to increase model accuracy (Line 11), since C_i brings more distribution information to train $\boldsymbol{\omega}_r$. However, for convergence, a smaller $\widetilde{emd}_c$ is preferred in selection, so that $\widetilde{emd}_c$ depends on the round r:

$$\widetilde{emd}_c^r = Norm([EMD(\mathcal{D}_c^r, \mathcal{D}^i)\big|_{i \in \{1,...,N\}}]), \tag{7}$$

where $\mathcal{D}_c^r$ is the accumulated $\mathcal{D}^i$ of selected clients over rounds (Line 8). Let l denote one class randomly chosen by the *federator* except for the Maverick class from $\mathcal{D}$, here we apply normalization: $Norm(emd, \mathcal{D}) = \frac{emd}{\sum_{i=1}^{N} \mathcal{D}^i(y=l)/N}$.

iii) Weighted Random Client Selection (Line 7): At each round r, we select clients based on a probability distribution characterized by the dynamic weights [6] $Proba^r$:

$$\mathcal{K}^r = rand(K, \mathcal{C}, Proba^r). \tag{8}$$

Sampling K out of N clients based on $Proba^r$ has a complexity of $O(K \log(N/K))$, so comparably low. Thus, Mavericks with larger global distance and smaller current distance initially are preferred to be selected. The decrease of probability for selecting Mavericks elaborates based on the global and current distances changes over the learning procedure. As r increases, so does the impact of the current distance based on Eq. 5, reducing the probability to select a Maverick, as intended.

Convergence Analysis: To derive the convergence bound, we follow the setting of [15]. We let F_k be the local objective of client C_k and define $F(\boldsymbol{\omega}) \triangleq \sum_{k=1}^{N} p_k F_k(\boldsymbol{\omega})$, where p_k is the weight of client C_k when doing the aggregation. We have the FL optimization framework $\min_{\omega} F(x) = \min_{\omega} \sum_{k=1}^{N} p_k F_k(\boldsymbol{\omega})$. We make the L-*smooth* and μ-*strongly convex* assumptions on the functions $F_1, ..., F_N$ [15,20]. Let T be the total number of SGDs in a client, E be the number of local iterations of each client in each round. t is used to index the SGDs in each client. Thus, the relationship between E, t and global round r is $r = \lfloor t/E \rfloor$. F^* and F_k^* are the minimum values of F and F_k. $\Gamma = F^* - \sum_{k=1}^{N} p_k F_k^*$ is used to represent the degree of heterogeneity. We obtain:

Theorem 2. Let ξ_t^k be a sample chosen from the local data of each client. For $k \in [N]$, assume that:

$$\mathbb{E} \left\| \bigtriangledown F_k(\boldsymbol{\omega}_t^k, \xi_t^k) - F_k(\boldsymbol{\omega}_t^k) \right\|_2^2 \leq \sigma_k^2, \tag{9}$$

and

$$\mathbb{E} \left\| F_k(\boldsymbol{\omega}_t^k, \xi_t^k) \right\|_2^2 \leq G^2. \tag{10}$$

Then let $\epsilon = \frac{L}{\mu}$, $\gamma = \max\{8\epsilon, E\}$ and the learning rate $\eta_t = \frac{2}{\mu(\gamma+t)}$. We have the following convergence guarantee for Algorithm 1.

$$\mathbb{E}[F(\boldsymbol{\omega}_T)] - F^* \leq \frac{\epsilon}{\gamma + T - 1} \left(\frac{2(\Psi + \Phi)}{\mu} + \frac{\mu\gamma}{2} \mathbb{E} \left\| \boldsymbol{\omega}_1 - \boldsymbol{\omega}^* \right\|_2^2 \right),$$

where $\Psi = \sum_{k=1}^{N} (Proba_k^{\lfloor T/E \rfloor})^2 {\sigma_k}^2 + 6L\Gamma + 8(E-1)^2 G^2$ and $\Phi = \frac{4}{K} E^2 G^2$.

Since all the notations except T in Expression (2) are constants, we have $O(\frac{1}{T})$ convergence rate for the algorithm where $\lim_{T \to \infty} \mathbb{E}[F(\boldsymbol{\omega}_T)] - F^* = 0$.

5 Experimental Evaluation

In this section, we comprehensively evaluate the effectiveness and convergence of FEDEMD in comparison to *Shapley Value*-based selection and SOTA baselines. The evaluation considers both exclusive and shared Mavericks.

Datasets and Classifier Networks. We use public image datasets: *i)* Fashion-MNIST [26] for bi-level image classification; *ii)* MNIST [13] for simple and fast tasks that require a low amount of data; *iii)* Cifar-10 [12] for more complex task such as colored image classification; *iv)* STL-10 [5] for applications with small amounts of local data for all clients. We note that light-weight neural networks are more applicable for FL scenarios, where clients typically have limited computation and communication resources [19]. Thus, here we apply light-weight CNNs for all datasets.

Federated Learning System. The system considered has 50 participants with homogeneous computation and communication resources and 1 *federator*. At each round, the *federator* selects 10% of clients using different client selection algorithms. The *federator* uses average or quantity-aware aggregation to aggregate local models from selected

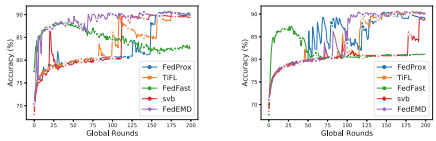
clients. We set one local epoch for both aggregations to enable a fair comparison of the two aggregation approaches. Two types of Mavericks are considered: exclusive and shared Mavericks with up to 3 Mavericks. We demonstrate the case of single Maverick owning an entire class of data in most of our experiments.

Evaluation Metrics. *i)* Global test accuracy for all classes; *ii)* Source recall for classes owned by Mavericks exclusively; *iii)* $R@99$: the number of communication rounds required to reach 99% of test accuracy of random selection results; *iv)* Normalized *Shapley Value* ranging between $[0, 1]$ to measure the contribution of Mavericks.

Baselines. We consider four selection strategies: Random [18], *Shapley Value*-based, FedFast [19], and TiFL [3][3] under both average and quantity-aware aggregation methods. Further, in order to compare with state-of-the-art solutions for heterogeneous FL that focus on the optimizer, we evaluate FedProx [14] as one of the baselines.

5.1 *FEDEMD* Is Effective for Client Selection

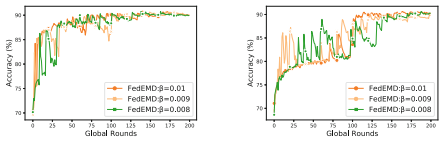
Figure (3a, b) show global accuracy over rounds. First we focus on the comparison between the contribution-based SVB and our proposed distance-based FEDEMD. FEDEMD achieves an accuracy close to the maximum almost immediately for FedAvg while SVB requires about 100 rounds (72 and 104 rounds for $R@99$ for SVB and FEDEMD). For average aggregation, both client selection methods have a slower convergence but FEDEMD still only requires about half the number of rounds to achieve the same high accuracy as SVB. Indeed, SVB fails in reaching $R@99$ within 200 rounds. The reason is that SVB rarely selects the Maverick in the early phase, as the Maverick has a below-average *Shapley Value*. We can also see the superiority of FEDEMD among results presented for the baselines in the figures. The detailed analysis will be discussed together with Table 1 below.



(a) FMNIST-quantity (b) FMNIST-average

Fig. 3. Comparison on FEDEMD with baselines.

We evaluate the effects of the hyper-parameter β in Fig. (4a, b). The server can apply a preliminary client selection simulation before training based on the self-reported data size array. FEDEMD works best when the average probability of selecting Maverick is within $[1/N - \epsilon, 1/N + \epsilon]$ based on our observation experiments, where $\epsilon > 0$ is a task-aware



(a) FMNIST-quantity (b) FMNIST-average

Fig. 4. Comparison on FEDEMD over different β.

[3] We focus on their client selection and leave out other features, e.g., communication acceleration in TiFL. We apply distribution mean clustering for FedFast following the setting in their paper.

small value. In our example with Fashion-MNIST, we choose β equal to 0.008, 0.009 and 0.01, with the results displayed in Fig. 4. These three values all satisfy the average probability above with $\epsilon \geq 0.002$. The results shows that all of the 3 numbers work for Fashion-MNIST, verifying the effectiveness of FEDEMD for various values of the hyper-parameter. However, there are also values of β that are not suitable, e.g., $\beta = 0.1$ for which the Maverick is selected too rarely.

Comparison with Baselines. We summarize the comparison with the state-of-the-art methodologies in Table 1. The reported $R@99$ is averaged over three replications. Note that we run each simulation for 200 rounds, which is mostly enough to see the convergence statistics for these lightweight networks. The rare exceptions when 99% maximal accuracy is not achieved for random selection are indicated by > 200.

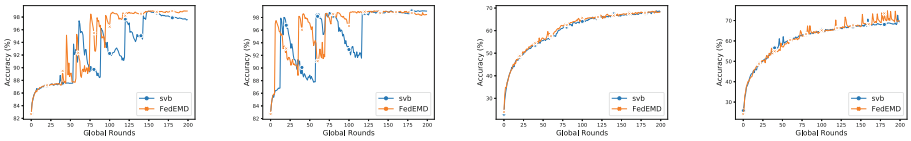
Due to its distance-based weights, FEDEMD almost consistently achieves faster convergence than all other algorithms. The reason for this result is that FEDEMD enhances the participation of the Maverick during the early training period, speeding up learning of the global distribution. For most settings, the difference in convergence rounds is considerable and clearly visible.

Table 1. Convergence rounds of selection strategies in $R@99$ Accuracy, under average and quantity-aware aggregation (Every result is averaged over three runs and is marked with standard deviation among all of the replication results).

Dataset	Average Aggregation					
	Random	**FedProx**	**TiFL**	**FedFast**	**SVB**	**FEDEMD**
MNIST	133 ± 44.47	118 ± 8.50	111 ± 21.66	>200 ± NA	147 ± 52.50	**99** ± 24.70
Fashion-MNIST	144 ± 51.47	135 ± 20.59	140 ± 8.62	>200 ± NA	**103** ± 56.00	131 ± 37.29
Cifar-10	141 ± 6.11	164 ± 15.00	147 ± 10.97	>200 ± NA	184 ± 9.24	**140** ± 15.13
STL-10	122 ± 49.94	186 ± 4.36	125 ± 57.50	171 ± 16.74	190 ± 3.06	**96** ± 4.93
Dataset	**Quantity-aware Aggregation**					
	Random	**FedProx**	**TiFL**	**FedFast**	**SVB**	**FEDEMD**
MNIST	72 ± 29.26	51 ± 8.19	84 ± 37.99	>200 ± NA	49 ± 2.52	**40** ± 5.57
Fashion-MNIST	111 ± 37.75	92 ± 12.12	146 ± 38.18	>200 ± NA	80 ± 40.13	**80** ± 10.79
Cifar-10	143 ± 26.29	144 ± 39.46	120 ± 9.45	174 ± 9.50	132 ± 26.50	**107** ± 10.58
STL-10	180 ± 0.58	179 ± 6.24	>200 ± NA	153 ± 34.88	181 ± 10.97	**95** ± 2.65

The only exception are easy tasks with simple averaging rather than weighted, e.g., Fashion-MNIST with average aggregation, which indicates our distribution-based selection method is especially useful for data size-aware aggregation and more complex tasks. Quantity-aware aggregation nearly always outperforms plain average aggregation as its weighted averaging assigns more impact to the Maverick. While such an increased weight caused by larger data size can lead to a decrease in accuracy in the latter phase of training, Mavericks are rarely selected in the latter phase by FEDEMD, which successfully mitigates the effect and achieves a faster convergence.

In order to demonstrate the comparison of FEDEMD and SVB across multiple datasets, here we also provide the experimental results with MNIST and Cifar-10, which is inline with our conclusion of Fashion-MNIST in Fig. 4 for better convergence performance of FEDEMD.



(a) MNIST-average (b) MNIST-quantity (c) Cifar-10-average (d) Cifar-10-quantity

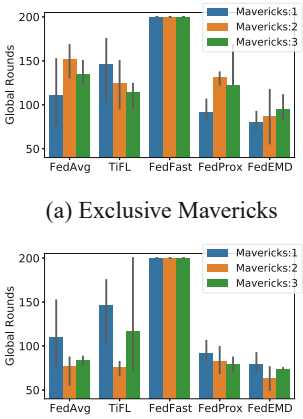
Fig. 5. Comparison on FEDEMD with SVB.

5.2 *FEDEMD* Works for Multiple Mavericks

We explore the effectiveness of FEDEMD on both types of Mavericks: exclusive and shared Mavericks.

We vary the number of Mavericks between one and three and use the Fashion-MNIST dataset. The Maverick classes are 'T-shirt', 'Trouser', and 'Pullover'. Results are shown with respect to $R@99$.

Figure (6a) illustrates the case of multiple exclusive Mavericks. For exclusive Mavericks, the data distribution becomes more skewed as more classes are exclusively owned by Mavericks. FEDEMD always achieves the fastest convergence, though its convergence rounds increase slightly as the number of Mavericks increases, reflecting the increased difficulty of learning in the presence of skewed data distribution. FedFast's K-mean clustering typically results in a cluster of Mavericks and then always includes at least one Maverick. In some initial experiments, we found that constantly including a Maverick hinders convergence, which is also reflected in FedFast's results. TiFL outperforms FedAvg with random selection for multiple Mavericks. However, TiFL's results differ drastically over runs due to the random factor in its local computations. Thus, TiFL is not a reliable choice for Mavericks. Comparably, FedProx tends to achieve the best performance among the SOTA algorithms but still exhibits slower convergence than FEDEMD as higher weight divergence entails higher penalty on the loss function.



(a) Exclusive Mavericks

(b) Shared Mavericks

Fig. 6. Convergence rounds $R@99$ for multiple Mavericks.

For shared Mavericks, a higher number of Mavericks indicates a more balanced distribution. Similar to the exclusive case, FEDEMD has the fastest convergence and FedFast again trails the others. The improvement of FEDEMD over the other methods is less visible due to the limited advantage of FEDEMD on balanced data. A higher number of Mavericks resembles the case of *i.i.d.*. Random performs the most similar to FEDEMD for shared Mavericks, as random selection is best for *i.i.d.* scenarios. Note that the standard deviation of FEDEMD is smaller, implying a better stability.

6 Conclusion

Client selection is key to successful FL as it enables maximizing the usefulness of different diverse datasets. In this paper, we highlighted that existing schemes fail when clients have heterogeneous data, in particular if one class is exclusively owned by one or multiple Mavericks. We first explore *Shapley Value*-based selection, theoretically showing its limitations in addressing Mavericks. We then propose FEDEMD that encourages the selection of diverse clients at the opportune moment of the training process, with guaranteed convergence. Evaluation results on multiple datasets across different scenarios of Mavericks show that FEDEMD reduces the communication rounds needed for convergence by 26.9% compared to the state-of-the-art client selection methods.

References

1. Adam, R., Aris, F.R., Boi, F.: Rewarding high-quality data via influence functions (2019)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on Machine Learning (ICML), pp. 214–223. PMLR (2017)
3. Chai, Z., et al.: TiFL: a tier-based federated learning system. In: Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing (HPDC), pp. 125–136 (2020)
4. Cho, Y.J., Wang, J., Joshi, G.: Client selection in federated learning: convergence analysis and power-of-choice selection strategies. arXiv preprint arXiv:2010.01243 (2020)
5. Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence Aad Statistics (AISTATS). JMLR Workshop and Conference Proceedings (2011)
6. Efraimidis, P.S., Spirakis, P.G.: Weighted random sampling with a reservoir. Inf. Process. Lett. **97**(5), 181–185 (2006)
7. Feng, S., Niyato, D., Wang, P., Kim, D.I., Liang, Y.: Joint service pricing and cooperative relay communication for federated learning. In: 2019 IEEE iThings and GreenCom and IEEE Cyber and CPSCom and SmartData, iThings/GreenCom/CPSCom/SmartData 2019, pp. 815–820. Atlanta, GA, USA, 14–17 July 2019. IEEE (2019)
8. Goetz, J., Malik, K., Bui, D., Moon, S., Liu, H., Kumar, A.: Active federated learning. arXiv preprint arXiv:1909.12641 (2019)
9. Huang, J., Talbi, R., Zhao, Z., Boucchenak, S., Chen, L.Y., Roos, S.: An exploratory analysis on users' contributions in federated learning. arXiv preprint arXiv:2011.06830 (2020)
10. Huang, T., Lin, W., Wu, W., He, L., Li, K., Zomaya, A.: An efficiency-boosting client selection scheme for federated learning with fairness guarantee. IEEE Transactions on Parallel and Distributed Systems (2020)
11. Kang, J., Xiong, Z., Niyato, D., Xie, S., Zhang, J.: Incentive mechanism for reliable federated learning: a joint optimization approach to combining reputation and contract theory. IEEE Internet Things J. **6**(6), 10700–10714 (2019)
12. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
13. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
14. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: Proceedings of Machine Learning and Systems (MLsys) (2020)

15. Li, X., Huang, K., Yang, W., Wang, S., Zhang, Z.: On the convergence of FedAvg on non-IID data. ICLR (2020)
16. Liu, S., Feng, X., Zheng, H.: Overcoming forgetting in local adaptation of federated learning model. In: Gama, J., Li, T., Yu, Y., Chen, E., Zheng, Y., Teng, F. (eds.) Advances in Knowledge Discovery and Data Mining. PAKDD 2022. LNCS, vol. 13280, pp. 613–625. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-05933-9_48
17. Luo, B., Xiao, W., Wang, S., Huang, J., Tassiulas, L.: Tackling system and statistical heterogeneity for federated learning with adaptive client sampling. In: IEEE INFOCOM 2022 - IEEE Conference on Computer Communications, London, United Kingdom, 2–5 May 2022, pp. 1739–1748. IEEE (2022)
18. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Proceedings of Artificial Intelligence and Statistics (AISTATS), pp. 1273–1282 (2017)
19. Muhammad, K., et al.: FedFast: going beyond average for faster training of federated recommender systems. In: Proceedings of the 26th ACM International Conference on Knowledge Discovery & Data Mining, pp. 1234–1242 (2020)
20. Nguyen, H.T., Sehwag, V., Hosseinalipour, S., Brinton, C.G., Chiang, M., Poor, H.V.: Fast-convergent federated learning. IEEE J. Sel. Areas Commun. **39**(1), 201–218 (2020)
21. Nishio, T., Yonetani, R.: Client selection for federated learning with heterogeneous resources in mobile edge. In: IEEE International Conference on Communications (ICC), pp. 1–7 (2019)
22. Sim, R.H.L., Zhang, Y., Chan, M.C., Low, B.K.H.: Collaborative machine learning with incentive-aware model rewards. In: International Conference on Machine Learning (ICML), pp. 8927–8936. PMLR (2020)
23. Wang, L., Xu, S., Wang, X., Zhu, Q.: Addressing class imbalance in federated learning. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI, IAAI, EAAI, pp. 10165–10173. AAAI Press (2021)
24. Wang, T., Rausch, J., Zhang, C., Jia, R., Song, D.: A principled approach to data valuation for federated learning. In: Yang, Q., Fan, L., Yu, H. (eds.) Federated Learning. LNCS (LNAI), vol. 12500, pp. 153–167. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-63076-8_11
25. Wei, S., Tong, Y., Zhou, Z., Song, T.: Efficient and fair data valuation for horizontal federated learning. In: Federated Learning, pp. 139–152 (2020)
26. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms (2017)
27. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V.: Federated learning with non-IID data. arXiv:1806.00582 (2018)
28. Zhu, Z., Hong, J., Zhou, J.: Data-free knowledge distillation for heterogeneous federated learning. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18–24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 12878–12889. PMLR (2021)