计算机科学
COMPUTER SCIENCE

基于Edge-TB的联邦学习中客户端选择策略和数据集划分研究

周天阳, 杨磊

引用本文

周天阳, 杨磊. 基于Edge-TB的联邦学习中客户端选择策略和数据集划分研究[J]. 计算机科学, 2024, 51(6A): 230800046-6.

ZHOU Tianyang, YANG Lei. Study on Client Selection Strategy and Dataset Partition in Federated Learning Basedon Edge TB [J]. Computer Science, 2024, 51(6A): 230800046-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230800046

基于 Edge-TB 的联邦学习中客户端选择策略和数据集划分研究

周天阳　杨　磊
华南理工大学软件学院　广州 510006
(zhoutianyang2002@163.com)

摘　要　联邦学习是分布式机器学习在现实中的应用之一。针对联邦学习中的异构性,基于 FedProx 算法,提出优先选择近端项较大的客户端选择策略,效果优于常见的选择局部损失值较大的客户端选择策略,可以有效提高 FedProx 算法在异构数据和系统下的收敛速度,提高有限聚合次数内的准确率。针对联邦学习数据异构的假设,设计了一套异构数据划分流程,得到了基于真实图像数据集的异构联邦数据集作为实验数据集。使用开源的分布式机器学习框架 Edge-TB 作为实验测试平台,以异构划分后的 Cifar10 作为数据集,实验表明,采用新的客户端选择策略的改进 FedProx 算法较原算法在有限的聚合轮数内准确率提升 14.96%,通信开销减小 6.3%;与 SCAFFOLD 算法相比,准确率提升 3.6%,通信开销减小 51.7%,训练时间减少 15.4%。

关键词:分布式机器学习;联邦学习;优化算法;正则化;近端项
中图分类号　TP181

Study on Client Selection Strategy and Dataset Partition in Federated Learning Based on Edge TB

ZHOU Tianyang and YANG Lei
Department of Software Engineering, South China University of Technology, Guangzhou 510006, China

Abstract　Federated learning is one of the applications of distributed machine learning in reality. In view of the heterogeneity in Federated learning, based on FedProx algorithm, this paper proposes a client selection strategy that preferentially selects the client with large near end items. The effect is better than the common client selection strategy that selects the client with large local loss value, which can effectively improve the Rate of convergence of FedProx algorithm under heterogeneous data and systems, and improve the accuracy within limited aggregation times. According to the hypothesis of heterogeneous data in federated learning, a set of heterogeneous data partition process is designed, and the heterogeneous federated dataset based on the real image dataset is obtained as the experimental dataset. Using the open-source distributed machine learning framework Edge-TB as the experimental testing platform and the heterogeneous partitioned Cifar10 as the dataset, the experiment proves that, using the new client selection strategy, the accuracy of the improved FedProx algorithm improves by 14.96%, and the communication overhead reduces by 6.3% compared to the original algorithm in a limited number of aggregation round. Compared with the SCAFFOLD algorithm, the accuracy is improved by 3.6%, communication overhead is reduced by 51.7%, and training time is reduced by 15.4%.

Keywords　Distributed machine learning, Federated learning, Optimization algorithm, Regularization, Proximal term

1　引言

联邦学习最早由谷歌于 2016 年提出,主要为了解决“数据孤岛”和隐私保护的问题[1],典型应用案例有谷歌输入法的智能提示功能[2]。与传统的分布式机器学习相比,联邦学习具有客户端拥有对本地数据样本的绝对控制权、通信代价高昂、数据异构和系统异构等特点。

异构系统和异构数据给联邦学习算法的设计带来了很大的挑战,需要一个兼顾通信开销、收敛速度和模型准确率的客户端选择策略。FedAvg 是经典的联邦学习算法[1],尽管它考虑了设备异构性和数据异构性,但是算法在实际异构条件下表现一般。许多研究在 FedAvg 算法的基础上进行改进,以期在异构数据上获得更好的表现,如 FedProx[3],FedNova[4],SCAFFOLD[5],FedAsync[6]等。文献[7]中提出的 FedCS 算法优先选择通信和训练快的设备,以加快模型的聚合速度。文献[8-9]采用设置局部模型更新阈值的方法,只有大于更新阈值的客户端才需要上传本地模型,使得每轮聚合上传的局部模型数量小于被选中的客户端数量,减少了通信量。文献[10]对异构客户端选择策略的收敛性进行了分析,量化了选择策略对收敛速度的影响,指出选择具有更高局部损失的客户端可以加快模型收敛。文献[11]在上述研究的基础上,给出了基于局部损失值和训练耗时的客户端效应计算公式,可以用来衡量客户端对全局模型的贡献。文献[12]按客户端返回的局部模型,用聚类算法把客户端分成 K 类,在选择客户端时从每一类中随机选择一个,这样可以尽可能让被选中的客户端所持有的样本接近全局样本的特征,可以提高全局模型的准确率。但是这种方法没有考虑到通信开销的影响。文献[13]使用 Double-DQN 选择客户端,把当前的全局模型经过 PCA 降维后作为状态,经过 Double-DQN 和 1 个 softmax 层输出每个客户端被选择的概率值,系统奖励通讯轮数少的

通信作者:杨磊(sely@scut.edu.cn)

回合，惩罚通讯轮数多的回合。但总体来看，在客户端的选择策略上，现有研究很难兼顾通信开销、收敛速度和模型准确率，缺少一种简单有效的方式。

联邦学习采用的数据集一般有 3 个来源：真实场景下的数据集、人工合成的数据集、人工划分的已有的数据集。真实场景下的数据集获取困难、数量有限、使用上也不够灵活。以 EMNIST 数据集为例，LEAF 的作者把它划分成 3 000 份样本子集[14]，用户所能做的仅仅是从这 3 000 份样本中随机抽样作为每个客户端的本地数据集，如果实验的客户端数量不同，总的样本数就不同，这就很难比较不同研究的算法的优劣。LEAF 还提供了一种 Synthetic Dataset 的划分方式，支持用户指定客户端的数量、标签类型的数量和特征的维度等。但这个数据集是纯粹人工合成的，用高斯分布产生的随机数作为样本的特征向量和模型真实权重，把权重和特征向量相乘加上噪声，经过 softmax 取最大值生成标签，数据集缺乏真实性。人工划分数据集的方法很好地取得了两者之间的平衡，一方面，研究者可以灵活调整划分策略，在不损失总样本数的情况下分成任意份；另一方面，人工划分的数据集是基于真实的、通用的数据集，研究者很清楚每个数据集的难易程度，从而选择合适的神经网络模型。关于人工划分的数据集，文献[15-17]一共给出了 5 种方案，包括数量倾斜等，并分别验证了不同方案对联邦学习算法的影响。但是这 5 种方案是正交的，依然缺少一种能涵盖异构数据所有特点的划分方式。文献[17]关于特征倾斜的两种实现方式并不是很合理：合成数据集真实性差、对图片加噪音可能对模型精度产生影响。综合已有研究来看，需要研究一种兼顾标签倾斜、特征倾斜和数量倾斜的数据集划分方法。

Edge-TB 是一个通用的分布式机器学习的框架，作为一种混合的测试平台，它既有真实的物理设备保障计算的保真度，又可以在计算能力强的设备上使用容器技术搭载多个模拟节点，扩大系统规模，还可以灵活设置网络环境，具有实验成本低、保真度和灵活性高的特点[18]。本文以 Edge-TB 框架为基础，针对当前联邦学习研究中存在的不足，利用 Edge-TB 的真实异构网络通信环境，研究一种新的联邦学习数据集划分方法，可以涵盖异构数据的所有特征，包括数量分布倾斜、标签分布倾斜、特征分布倾斜；提出一种考虑近端项和局部损失值两种选择的客户端价值函数，实现兼顾通信开销、收敛速度和模型准确率的一种简单有效的客户端选择策略。

2 客户端选择策略

2.1 问题定义与形式化

假设一共有 m 个客户端，每个客户端有本地数据集 D_i，大小 $n_i=|D_i|$，总样本量 $n=\sum_{i=1}^{m}n_i$。联邦学习的目标是找到一个最优全局模型 w^*。$w^*=\operatorname{argmin}(F_N(w^*))$，其中 $F(w^*)=\sum_{i=1}^{m}d_iF_i(w^*)$，$d_i=\frac{n_i}{n}$，$F_i(w^*)=\frac{1}{n_i}\sum_{j=1}^{n_i}f(w^*,D_{i,j})$，$f(w,\xi)$代表损失函数。

记第 t 次聚合中被选中的客户端集合为 $c^{(t)}$，在 FedAvg 算法和 FedProx 算法中，全局模型 $w^{(t)}=\sum_{i\in c^{(t)}}\frac{n_i}{n^{(t)}}w_i^{(t)}$，其中 $n^{(t)}=\sum_{i\in c^{(t)}}n_i$。本文研究的问题是找到一种客户端选择策略 S，使得聚合后的全局模型 $w^{(t)}$ 尽可能接近 w^*，同时要求第 t 轮中最慢的节点花费的时间 $time^{(t)}=\max(time_i^{(t)})$ 尽可能少。

2.2 客户端选择策略设计

定义价值函数 $v_i^{(t)}$，用来衡量局部模型 $w_i^{(t)}$ 在第 t 轮聚合中对全局模型 $w^{(t)}$ 的贡献。为了加速全局模型的收敛，应该尽可能地让选择策略偏向对全局模型收敛贡献更大的局部模型，即每次选择较大的 $v_i^{(t)}$。

如果使用近端项定义 $v_i^{(t)}$，则 $v_i=\|w_i^{(t)}-w^{(t-1)}\|$，因为比较的是相对大小，所以不用乘上 FedProx 的系数 $\frac{\mu}{2}$。v_i 越大，说明局部模型越偏离全局模型，在聚合时对全局模型的影响越大。

借鉴文献[10-11]的做法，可以将局部损定义成 $v_i=f(w_i^{(t)},n_i)$或 $v_i=f(w^{(t-1)},n_i)$。前者是使用局部模型 $w_i^{(t)}$ 在第 t 轮训练后得到的局部模型计算局部损失值，后者使用中心服务器下发的上一轮全局模型 $w^{(t-1)}$ 计算局部损失值；前者代表客户端最新的局部模型在局部数据 n_i 上的损失，后者代表全局模型在局部数据 n_i 上的损失。

如果仅依靠价值函数作为客户端选择的唯一依据，就忽略了训练速度和通信开销。为了平衡训练速度，本文设置了一个时间阈值 T 作为超参数。记客户端的期望花费时间（训练时间加通信时间）为 t_i。将价值函数从大到小排序选择，如果 t_i 小于 T，就选择客户端 i 作为下一轮的参与者；当 t_i 大于 T 时，以 $1-\frac{t_i-T}{T}$的概率选择该客户端，因此 t_i 越大，被选择的概率越低。这样就兼顾了公平和效率。算法伪代码如算法 1 所示。

算法 1 优化后的 FedProx 算法（Server 端）

```
Input: aggregator, trainer_list, trainer 本地数据集 Di, 参与率 C, 聚合
       轮数 R, 本地迭代次数 E
Output: 全局模型 w
1. w←Initialization(seed)
2. For i ∈ trainer_list do
3.     ti←EstimateTime(i)
4.     vi←‖ w−EstimateW(i) ‖
5. end
6. For i =1 to R do
7.     selected_trainer←select(trainer_list, C)
8.     broadcast(selected_trainer, w)
9.     w←aggregate(wj), j∈ selected_trainer
10. end
11. def select(trainer_list, C):
12.     N←| trainer_list |
13.     n←⌈ N×K ⌉
14.     S←Ø
15.     vi ranked from big to small
16.     For i =1 to N do
17.     if ti<T then
18.         S ←S ∪ i
19.     else if ti<2 * Tthen
20.         S ←S ∪ i, with probability 1− (ti−T)/T
21.     end
```

```
22.  if |S| = n then
23.      break
24.  end
25. end
26. end
```

3 数据集划分

数据集划分问题可以形式化定义为：将数据集 D 划分给 m 个客户端，满足 $|D|=\sum_{i=1}^{m}|D_i|$，同时 $|D_i|$ 尽可能接近客户端数据量分布，同时每个客户端的特征分布 $p_i(x)$ 和标签分布 $p_i(y)$ 各不相同。

综合文献[15-17]，异构数据一般应该具有 3 种倾斜的特征。1）数量倾斜：即每个客户端的样本数量各不相同，有的特别多，有的特别少。2）标签倾斜：每个客户端拥有的标签种类、数量不一，如有的用户手机相册中照片全是动物，有的用户相册中鞋子、衣服和自拍很多，但是动物的照片没有或者很少。3）特征倾斜：客户端拥有样本的特征分布可能不一样，比如某个客户端的照片都是狸花猫，另一个客户端的照片都是橘猫，虽然标签都是猫，但是特征区别很大。

本文的数据集划分同时考虑数量倾斜、标签倾斜和特征倾斜，具体的异构数据集划分流程如图 1 所示（这里假设客户端数量为 N，数据集样本标签一共 K 种，样本总数为 M）。

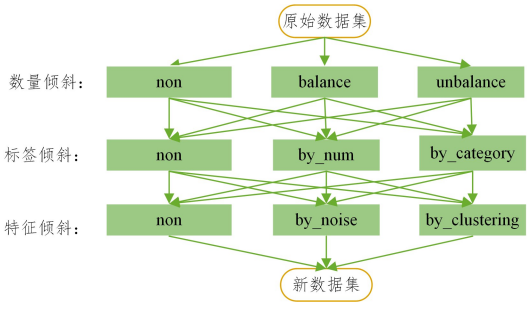


图 1 数据划分流程图

Fig. 1 Data partitioning flowchart

第一步 选择数量倾斜的方式（不倾斜、平衡和不平衡）。这一步主要确定每个客户端的样本数量。不倾斜是不固定样本的数量（根据标签倾斜的方式确定每个样本的数量），平衡是要求每个客户端样本数量相同，不平衡是要求每个客户端样本数量不平衡。文献[16-17]中指出，狄利克雷分布是一种先验分布，用来刻画样本数量的不平衡是比较合理的。因此令每个客户端的样本数 $n_i \sim Dir_N(\beta)$。其中 $N=\sum n_i$ 是总样本数，β 描述数量的倾斜程度，β 越小，数据倾斜程度越大。

第二步 选择标签倾斜的方式（不倾斜、按类别倾斜和按数量倾斜）。这一步主要确定每个客户端所持有的每种类型样本的数量 $n_{i,j}$。如果按类别倾斜，则要求用户输入客户端持有标签种类 k，把每种类型的样本分成 $\frac{Nk}{K}$ 份，每个客户端从这 $\frac{Nk}{K}$ 份中随机抽取标签类型不同的 k 份。如果是按数量倾斜，则令每种类型的标签数量按照狄利克雷分布分配给每个客户端，第 i 个客户端拥有第 j 个标签的数量 $n_{i,j} \sim Dir(\beta)$。当然，如果用户第一步选择了数量平衡和不平衡，那么用户选择标签倾斜后得到的第 i 个客户端样本总数 n_i' 和第一步得到的样本总数 n_i 可能不相等。在研究中选择从所有 $n_i' > n_i$ 的客户端中随机抽取 $n_i' > n_i$ 个样本，组成一个样本集合 Q，再让所有 $n_i' < n_i$ 的客户端分别从 Q 中随机抽取 $n_i - n_i'$ 个样本。

第三步 选择特征倾斜的方式（不倾斜、根据噪音倾斜和根据聚类倾斜）。这一步主要为每个客户端分配样本。如果选择不倾斜，就根据之前确定的客户端样本数量和每种标签类型的样本数量随机分配样本；如果选择按噪音倾斜，就在随机分配的基础上给每 i 个客户端加上噪声 $\frac{j}{N} \cdot Gauss(0,\sigma^2)$，$1 \leqslant j \leqslant N$（不同客户端加的噪音大小不一样，实现了特征倾斜）；如果选择按聚类倾斜，当分配到第 i 个客户端时，把第 j 种类型的标签采用 K-means 方法分成 $N-i+1$ 类，记作 q_k，$1 \leqslant k \leqslant N-i+1$，并将 q_k 从小到大排序，把恰好大于等于 $n_{i,j}$ 的 q_k 分配给第 i 个客户端。如果客户端需要的样本数 $n_{i,j}$ 大于任意一类，就将 q_k 从大到小排序，取 k 使得 $\sum_{k=1} q_k >= n_{i,j}$，从前 k 份中选择前 $n_{i,j}$ 个样本分配给用户。

根据上述数据集划分流程，就可以任意选择不同的倾斜方式，产生 24 种不同形式的划分方案。

4 结果与分析

4.1 实验环境

本文以 6 台腾讯云服务器作为实验设备，操作系统均为 Ubuntu 20.04.5 LTS。在 Edge-TB 框架中，需要一台设备作为 controller，用来配置系统网络拓扑、下达训练命令、收集训练信息等，其他设备作为 worker 搭载一个或多个节点（相当于真实世界的客户端和中心服务器）。Worker 本身可以作为系统中的物理节点，也可以搭载多个容器（虚拟节点）。图 2 是物理部署情况，表 1－表 3 分别列出了服务器的性能、实验网络拓扑和实验网络特点。可以看到，参与实验的 16 个节点的通信能力、计算能力是异构的，满足联邦学习系统异构的假设。

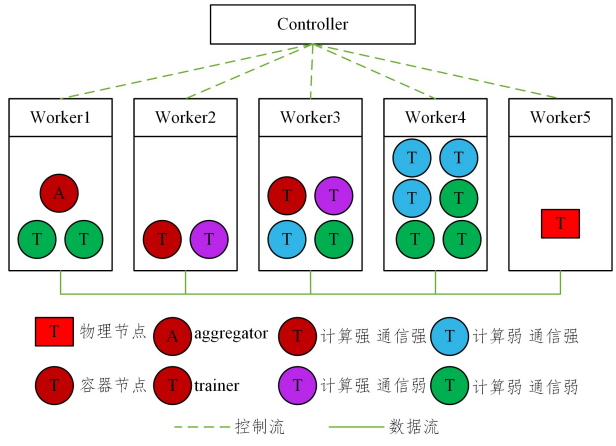


图 2 Edge-TB 平台的物理部署图

Fig. 2 Physical deployment diagram of Edge-TB platform

表 1 6 台服务器的基本指标

Table 1 Basic indicators of six servers

	CPU(cores)	内存/GB	带宽/Mbps
Controller	2	2	5
Worker1	8	32	5
Worker2	8	32	5
Worker3	8	32	5
Worker4	8	32	5
Worker5	4	8	5

表 2 服务器搭载节点情况

Table 2 Server mounted nodes

服务器	节点名	CPU 核数	带宽(上行/下行)/(bMps)	内存/GB
Worker1	n1	4	—	22
Worker1	n2	1	1/2	4
Worker1	n3	1	1/2	4
Worker2	n4	3	3/5	15
Worker2	n5	3	1/2	15
Worker3	n6	2	3/5	10
Worker3	n7	2	1/2	10
Worker3	n8	1	3/5	5
Worker3	n9	1	1/2	5
Worker4	n10	1	3/5	5
Worker4	n11	1	2/4	5
Worker4	n12	1	2/4	5
Worker4	n13	1	1/2	5
Worker4	n14	1	1/2	5
Worker4	n15	1	1/2	5
Worker5	p1	4	3/5	5

表 3 系统客户端(节点)的特点

Table 3 Characteristics of system clients(nodes)

	计算能力强	计算能力弱	合计
通信能力强	3	4	7
通信能力弱	2	6	8
合计	5	10	15

Edge-TB 是一个通用的分布式机器学习框架，对联邦学习的支持不足。例如，Edge-TB 关于联邦学习的优化算法只支持 FedAvg。为了研究比较不同算法的优缺点，选择更好的算法作为客户端选择问题的基准算法，本文在 Edge-TB 框架内实现了 FedProx，FedNova，SCAFFLOD 和 FedAsync 4 种算法。此外还对 Edge-TB 中模型部署过程进行了优化，提高了模型的部署效率。

在神经网络结构方面，本文使用 keras 搭建了两个神经网络，一个用于 Fashion-MNIST，一个用于 Cifar-10。两个神经网络卷积核大小都是 3 * 3，都使用了 relu 激活函数和批量归一化。用于 Fashion-MNIST 的神经网络输入是 28 * 28 的特征值，网络结构为两个卷积层(32)、一个池化层、一个卷积层(64)、一个池化层、两个全连接层和一个 softmax 层，参数量为 229994。用于 Cifar-10 的神经网络输入是 32 * 32 * 3 的特征值，网络结构为两个卷积层(64)、一个池化层、两个卷积层(128)、一个池化层、两个卷积层(256)、一个池化层、两个全连接层和一个 softmax 层，参数量为 1674698。

4.2 数据集划分结果分析

综合考虑标签倾斜、特征倾斜和数量倾斜，按照第 3 节的 3 个步骤，异构数据集 Fashion-MNIST 划分后的可视化结果如图 3 所示。

取数量倾斜指数 $\beta=1$，标签倾斜方式选择按标签数量倾斜，倾斜指数 $\beta=0.5$，特征倾斜选择聚类的方式，得到的异构 Cifar-10 数据集如图 4 所示。

可见，针对联邦学习数据异构的假设，使用本文设计的一套异构数据划分流程能够得到同时具有数量异构、标签异构、特征异构的数据集，用做异构联邦数据集。

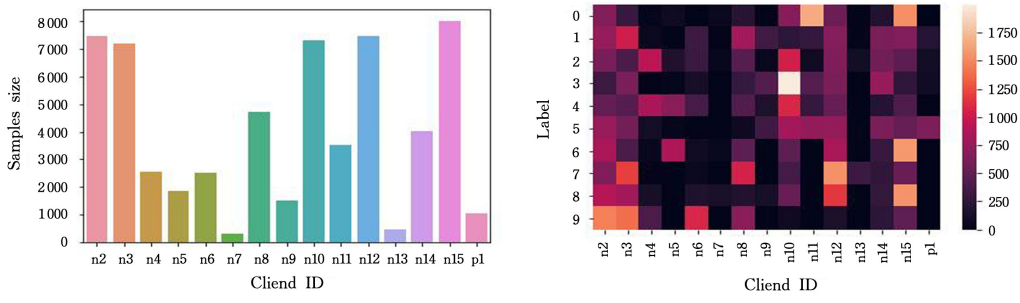


图 3 异构 Fashion-MNIST 数据集划分结果可视化

Fig. 3 Visualization of heterogeneous Fashion-MNIST dataset partition results

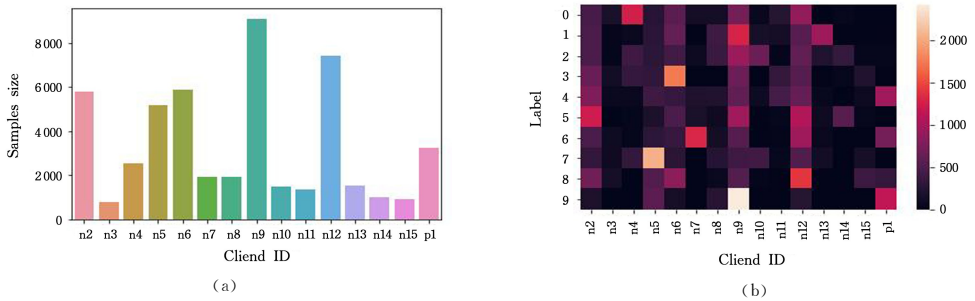


图 4 Cifar10 数据集划分后各客户端数据分布情况

Fig. 4 Data distribution of each client after Cifar10 dataset partitioning

4.3 FedAvg 算法在同构和异构数据集上的表现分析

使用基于 Fashion-MNIST 和 CIFAR-10 数据集生成的异构数据集和随机划分的同构数据集，在 15 个客户端(异构)的场景下对经典的联邦学习算法 FedAvg 进行测试，其中取聚合次数为 20 次，本地 epoch=1。实验结果如图 5 所示。

从图 5 可以看到异构数据对于 FedAvg 算法的准确率和训练速度都有着严重的影响。在有限的聚合次数内，面对较为复杂的数据集 CIFAR-10，FedAvg 算法在异构划分上的准确率远低于同构划分；在较为简单的 Fashion-MNIST 数据集上，即使最终全局模型的准确率差别不大，但是数据异构使得算法完成相同的聚合次数时需要的时间更多。由此可见，尽管 FedAvg 算法考虑了设备异构性和数据异构性，但是该算法在实际异构条件下表现一般。

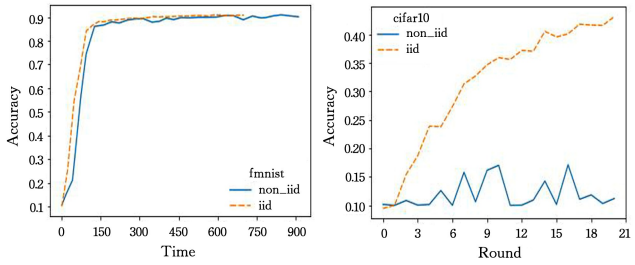


图 5　FedAvg 算法在同构和异构数据集上的表现

Fig. 5　Performance of FedAvg algorithm on isomorphic and heterogeneous datasets

4.4　不同算法在异构数据集上的表现分析

对于异步算法 FedAsync，为便于和同步算法比较，取每 5 次聚合为一轮。实验超参数设置为：本地迭代次数 E 为 1 次，聚合次数为 30 次，参与率为 0.3，FedPro 的 μ 在{0.01，0.05，0.1}中取最优值，FedAsync 取 α 为 0.8。

FedAvg，FedProx，FedNova，SCAFFOLD，FedAsync 这 5 种算法在异构数据集下的表现如图 6 所示。可以看到，对于较为简单的 Fashion-MNIST 图像分类任务，5 种方法均能达到 90%以上的准确率，其中 FedProx 和 SCAFFOLD 的准确率基本接近本地训练的准确率，两种算法的曲线几乎一致。面对复杂的图像分类任务 Cifar-10，5 种经典算法在有限的聚合次数内准确率都受到了很大的影响。准确率最佳的是 SCAFFOLD 方法，其次是 FedProx 和 FedAsync。以 FedAsync 为代表的异步算法虽然训练速度最快，但是对于中心服务器的通信能力要求高，通信开销大，容易产生丢包等问题。尽管 SCAFFOLD 的准确率在 5 种算法中最佳，但是算法需要客户端在每次聚合时将一个描述本地模型和全局模型差异的向量 $\boldsymbol{c}$ 发送给中心服务器，通信量和通信开销相比其他算法更大，几乎是 FedProx 等算法的两倍。综合通信开销和准确率，本文选择 FedProx 作为客户端选择研究的基准算法。

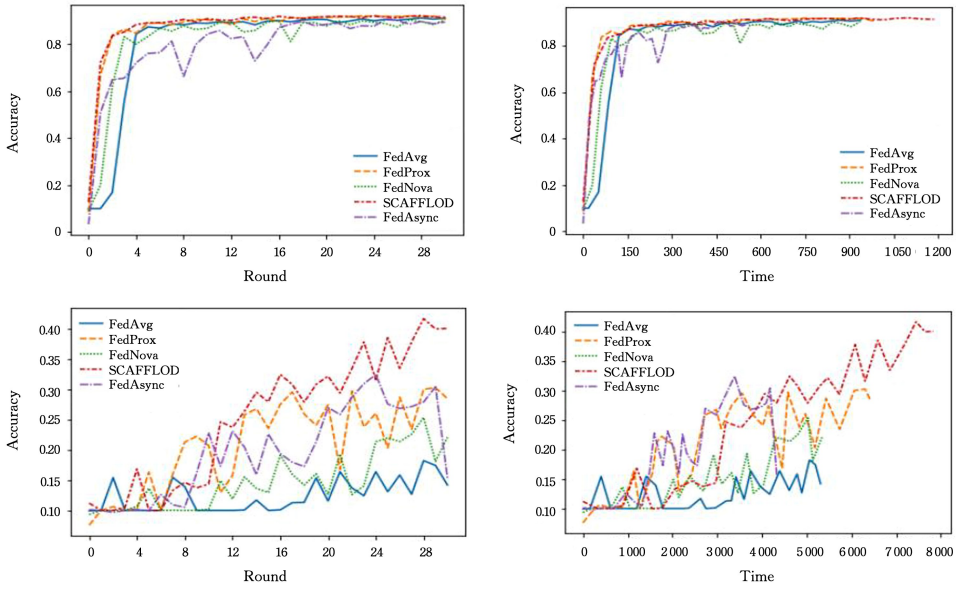


图 6　5 种经典的联邦学习算法在异构数据上的表现

Fig. 6　Performance of five classical federated learning algorithms on heterogeneous data

表 4　5 种算法在异构 Fashion-MNIST 上的统计数据

Table 4　Statistical data of five algorithms on heterogeneous Fashion-MNIST

算法	FedAvg	FedProx	FedNova	SCAFFOLD	FedAsync
最佳准确率	91.06%(28)	91.96%(23)	90.34%(30)	92.06%(24)	90.39%(25)
总时长/s	937	982	936	1186	559
每轮平均耗时/s	31.25	32.76	31.21	39.54	18.63
通信时长/s	1051	1036	1024	1988	5385

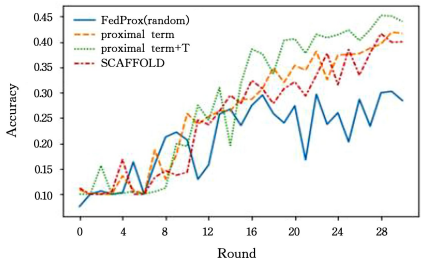
表 5　5 种算法在异构 Cifar-10 上的统计数据

Table 5　Statistical data of five algorithms on heterogeneous Cifar-10

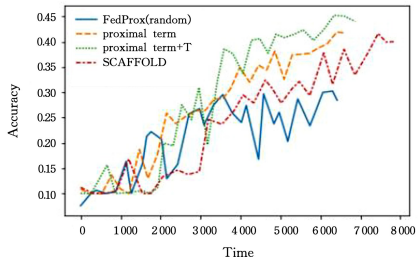
算法	FedAvg	FedProx	FedNova	SCAFFOLD	FedAsync
最佳准确率	18.21%(28)	30.24%(29)	25.32%(28)	41.62%(28)	32.40%(24)
总时长/s	5310	6417	5341	7834	4347
每轮平均耗时/s	177	213.9	178	261	144.9
通信时长/s	6886	6746	6815	13356	42598

4.5　不同客户端选择策略分析

以 FedProx 作为客户端选择研究的基准算法，分析对随机选择、根据近端项选择、根据阈值 T 和近端项选择等这 3 种客户端选择策略进行分析。考虑到 Fashion-MNIST 数据集难度偏低，只使用 Cifar-10 数据集作为实验数据集。分析结果如图 7 所示。



(a)



(b)

图 7 不同参与者选择策略在异构 Cifar-10 数据集上的表现

Fig. 7 Performance of different participant selection strategies on heterogeneous Cifar-10 dataset

表 6 FedProx 在不同客户端选择策略下的表现(异构 Cifar-10 数据集)

Table 6 FedProx performance under different client selection strategies(heterogeneous Cifar-10 dataset)

选择策略	FedProx(随机)	SCAFFOLD	近端项	近端项+时间阈值
最佳准确率	30.24%(29)	41.62%(28)	41.92%(29)	45.2%(28)
总时长/s	6417	7834	6869	6624
平均耗时/s	213.9	261	228.9	220.8
通信总时长/s	6746	13356	6443	6321

由上述图表可见,使用近端项作为客户端选择策略有效提高了模型的收敛速度和准确率,并且通信开销相比 SCAFFLOD 方法减少了近一倍。使用时间阈值的方法在不影响精度的情况下,可以有效减少训练时间。

结束语 Edge-TB 作为通用的分布式机器学习框架,具有实验成本低、保真度和灵活性高的特点,是一个良好的支撑联邦学习相关研究的平台。本文提出的能得到同时具有数量异构、标签异构、特征异构数据集的异构数据集划分方法具有可性性。以偏向较大的近端项为客户端选择策略,取代原算法的随机选择策略,在不增加通信开销的前提下,极大地提高了有限聚合次数内的模型准确率,是一种简单有效的客户端选择策略。

下一步,将继续采用更多的不同异构数据集来检验本文所提的异构数据集划分方法,以及客户端选择策略,完善 Edge-TB 支持的联邦学习算法种类,为联邦学习相关研究构建一个更加具有鲁棒性的平台。

参 考 文 献

[1] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data[C]// Artificial Intelligence and Statistics. PMLR, 2017: 1273-1282.

[2] LI T, SAHU A K, TALWALKAR A, et al. Federated learning: Challenges, methods, and future directions[J]. IEEE Signal Processing Magazine, 2020, 37(3): 50-60.

[3] LI T, SAHU A K, ZAHEER M, et al. Federated optimization in heterogeneous networks[J]. Proceedings of Machine Learning and Systems, 2020, 2: 429-450.

[4] WANG J, LIU Q, LIANG H, et al. Tackling the objective inconsistency problem in heterogeneous federated optimization[J]. Advances in Neural Information Processing Systems, 2020, 33: 7611-7623.

[5] KARIMIREDDY S P, KALE S, MOHRI M, et al. Scaffold: Stochastic controlled averaging for federated learning[C]// International Conference on Machine Learning. PMLR, 2020: 5132-5143.

[6] XIE C, KOYEJO S, GUPTA I. Asynchronous federated optimization[J]. arXiv: 1903.03934, 2019.

[7] NISHIO T, YONETANI R. Client selection for federated learning with heterogeneous resources in mobile edge[C]// 2019 IEEE International Conference on Communications(ICC 2019). IEEE, 2019: 1-7.

[8] RIBERO M, VIKALO H. Communication-efficient federated learning via optimal client sampling[J]. arXiv: 2007.15197, 2020.

[9] CHEN W, HORVATH S, RICHTARIK P. Optimal client sampling for federated learning[J]. arXiv: 2010.13723, 2020.

[10] CHO Y J, WANG J, JOSHI G. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies[J]. arXiv: 2010.01243, 2020.

[11] LAI F, ZHU X, MADHYASTHA H V, et al. Oort: Efficient Federated Learning via Guided Participant Selection[C]// OSDI. 2021: 19-35.

[12] FRABONI Y, VIDAL R, KAMENI L, et al. Clustered sampling: low-variance and improved represent ativity for clients selection in federated learning[C]// International Conference on Machine Learning. New York: PMLR, 2021: 3407-3416.

[13] WANG H, KAPLAN Z, NIU D, et al. Optimizing federated learning on non-iid data with reinforcement learning[C]// IEEE Conference on Computer Communications(INFOCOM 2020). IEEE, 2020: 1698-1707.

[14] CALDAS S, DUDDU S M K, WU P, et al. Leaf: A benchmark for federated settings[J]. arXiv: 1812.01097, 2018.

[15] ZHAO Y, LI M, LAI L, et al. Federated learning with non-iid data[J]. arXiv: 1806.00582, 2018.

[16] ZHU H, XU J, LIU S, et al. Federated learning on non-IID data: A survey[J]. Neurocomputing, 2021, 465: 371-390.

[17] LI Q, DIAO Y, CHEN Q, et al. Federated learning on non-iid data silos: An experimental study[C]// 2022 IEEE 38th International Conference on Data Engineering(ICDE). IEEE, 2022: 965-978.

[18] YANG L, WEN F, CAO J, et al. Edgetb: A hybrid testbed for distributed machine learning at the edge with high fidelity[J]. IEEE Transactions on Parallel and Distributed Systems, 2022, 33(10): 2540-2553.

ZHOU Tianyang, born in 2002, postgraduate, is a student member of CCF (No. D6441G). His main research interest is federated learning.

YANG Lei, born in 1986, Ph.D, professor, is a member of CCF(No. 60282M). His main research interests include cloud and edge computing, distributed machinelearning and federated learning.