计算机科学
COMPUTER SCIENCE

基于分层抽样优化的面向异构客户端的联邦学习

鲁晨阳, 邓苏, 马武彬, 吴亚辉, 周浩浩

引用本文

鲁晨阳, 邓苏, 马武彬, 吴亚辉, 周浩浩. 基于分层抽样优化的面向异构客户端的联邦学习[J]. 计算机科学, 2022, 49(9): 183-193.

LU Chen-yang, DENG Su, MA Wu-bin, WU Ya-hui, ZHOU Hao-hao. Federated Learning Based on Stratified Sampling Optimization for Heterogeneous Clients[J]. Computer Science, 2022, 49(9): 183-193.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220500263

基于分层抽样优化的面向异构客户端的联邦学习

鲁晨阳　邓　苏　马武彬　吴亚辉　周浩浩
国防科技大学信息系统工程重点实验室　长沙 410073
(cy_lu@nudt.edu.cn)

摘　要　联邦学习是一种新的面向隐私保护的分布式学习范式,相比传统分布式机器学习方法,其特点为各客户端通信、设备算力和存储能力存在较大差异(设备异构),各客户端数据分布和数量存在较大差异(数据异构)以及高通信消耗等。在客户端异构条件(包括设备异构和数据异构)下,客户端的数据分布区别较大,导致模型收敛速度显著降低,特别是在极端的数据异构情况下,传统的联邦学习算法无法收敛,并且训练曲线随着本地迭代轮次的增加出现大幅的波动。针对联邦学习中,客户端异构给模型训练带来的影响,提出了利用分层抽样优化的联邦学习算法——FedSSO。FedSSO使用了基于密度的聚类方法将总体客户端划入不同的聚类中,使得每个聚类中的客户端具有较高的相似度,再按样本权重从不同聚类中抽取可用客户端参与训练,因此所有种类的数据都会按样本权重参与每轮训练,使模型加速收敛到全局最优解;同时,设定了学习率递减和本地迭代轮次选择机制,以保证模型的收敛性。从理论和实验中证明了FedSSO的收敛性,并且在公开数据集MNIST,Cifar-10和Sentiment140上与其他联邦学习算法进行了对比,实验结果证明FedSSO的训练效果更优。

关键词:联邦学习;隐私保护;聚类;分层抽样;分布式优化;收敛性分析
中图法分类号　TP301

Federated Learning Based on Stratified Sampling Optimization for Heterogeneous Clients

LU Chen-yang,DENG Su,MA Wu-bin,WU Ya-hui and ZHOU Hao-hao
Science and Technology on Information Systems Engineering Laboratory,National University of Defence Technology,Changsha 410073,China

Abstract　Federated learning(FL) is a new distributed learning framework for privacy protection,which is different from traditional distributed machine learning:1)differences in communication,computing,and storage performance among devices(device heterogeneity),2)differences in data distribution and data volume(data heterogeneity),and 3)high communication consumption. Under heterogeneous conditions,the data distribution of clients varies greatly,which leads to the decrease of model convergence speed. Especially in the case of highly heterogeneous condition,the traditional FL algorithm cannot converge and the training loss curve will fluctuate greatly with the increase of local iterations. In this work,a FL algorithm based on stratified sampling optimization(FedSSO) is proposed. In FedSSO,a density-based clustering method is used to divide the overall client into different clusters. Then,some available clients are proportionally extracted from different clusters to participate in training. Therefore,various data are involved in each training round to ensure that FL can accelerate convergence to the optimal solution. The strategy of learning rate decay and the choice of local iterations is set to ensure the convergence. The convergence of FedSSO algorithm is proved theoretically and experimentally,and the superiority of FedSSO is demonstrated by comparing it with other FL algorithms on public MNIST,Cifar-10,and Sentiment140 datasets.

Keywords　Federated learning,Privacy protection,Clustering,Stratified sampling,Distributed optimization,Convergence analysis

联邦学习是一种新的分布式机器学习架构[1-2],它允许多个设备(在联邦学习中称为客户端)在不需要上传本地数据的情况下,共同训练一个全局模型[3-4]。与传统的分布式机器学习相比,它们的主要区别如下:1)联邦学习中客户端对本地设备和数据有独立的控制权;2)联邦学习中客户端通常不可靠(边缘节点通常因设备和通信问题而断开连接);3)联邦学习中的通信消耗高于计算消耗;4)联邦学习中的节点数据分布是非独立同分布的(non-IID)[5];5)联邦学习中各客户端的本地数据数量分布极不均衡。这些新的特性对设计和分析联邦学习的算法提出了新的挑战。

主要挑战之一是客户端的异构性[6],包括数据和设备异构性。客户端的异构性在现实条件下广泛存在,例如:1)异构的数据分布,每个客户端上的数据都是本地生成的,因此客户端之间的样本生成机制可能不同(如不同的国家或地区);

到稿日期:2022-05-30　返修日期:2022-07-04
基金项目:国家自然科学基金面上项目(61871388)
This work was supported by the National Natural Science Foundation of China(61871388).
通信作者:马武彬(wb_ma@nudt.edu.cn)

2)协变量偏移,例如,在手写识别中,不同的人书写同一个单词的方式不同;3)标签分布偏斜(先验概率漂移);4)数量倾斜或者不平衡;5)设备算力和通信差异。

考虑一个现实的情况,当我们应用联邦学习的方法去训练一个手机输入法模型时[7-8],不同手机的运算速度、内部数据、网络情况等都不相同,最新款手机比老款手机有更快的运行速度和传输速度,位于城镇等信号较好位置的手机比位于乡村等信号有干扰地区的手机通信传输更稳定。在模型训练过程中,老旧的手机训练速度慢,往往无法按时完成训练任务,网络环境差的手机在传输模型时也更容易出现信号丢失的情况,这就造成了参数服务器接受到的数据的分布和实际数据分布之间的差异。由于客户端异构的存在,某些类的数据更频繁地参与训练过程,给训练数据引入了误差。

为了降低客户端异构带来的影响,本文提出了 FedSSO 算法,该算法在不需要获取客户端原始数据的情况下,对所有客户端进行先期聚类,将客户端划入不同的簇中,然后按照样本权重,每轮从不同簇中抽取一定数量可用的客户端进行梯度聚合。通过设置适当的参数,证明了该算法能达到 $O(E^2/T)$ 的收敛速度;通过对算法的收敛性进行分析,证明了学习率递减对于收敛的重要性。同时,本文从理论和实验上证明,在高度异构的数据集上,传统联邦学习增加本地迭代次数会导致模型收敛速度变慢,而 FedSSO 算法可以减小这种影响,因此可以通过增加本地迭代次数的方式来实现模型训练加速。本文的主要贡献如下:

(1)证明了即使对于一个凸目标函数的优化问题,在计算精确的梯度(非随机梯度)时,传统的联邦学习算法无法在客户端异构条件下收敛到全局最优解。特别是在高度客户端异构条件下,传统联邦学习方法可能因为恒定的学习率和高本地迭代轮次而发散。

(2)从理论上证明了传统联邦学习在客户端条件下发散的原因,并且提出了收敛算法 FedSSO,该算法采用分层抽样的方式,每轮训练时从先期划分好的客户端集群中按照权重比例抽取一定数量的可用客户端参与训练。从理论上证明了 FedSSO 以 $O(E^2/T)$ 的速度收敛到全局最优解。

(3)采用标准数据集 MNIST,CIFAR-10 和 Sentiment140 来评估 FedSSO 算法,并且与 FedAvg[2] 和 FedProx[6] 进行了比较。评价结果证明了 FedSSO 算法在异构数据集上有更高的训练精度和更快的训练速度。

本文第 1 节介绍了联邦学习的研究背景和异构联邦学习的相关研究;第 2 节从理论上分析了客户端异构对模型收敛的影响,并提出了 FedSSO 算法的框架;第 3 节给出了 FedSSO 收敛性的理论证明,并对算法的收敛性展开了讨论,明确了递减的学习率对模型收敛性的重要性,同时分析了本地迭代轮次的选择策略;第 4 节在 Mnist,Cifar-10 和 Sentiment140 数据集上,应用不同的神经网络模型对算法进行了全面的评估。实验结果证明,相比 FedAvg 和 FedProx 算法,FedSSO 算法在异构数据集上有实质的改进。

1 相关工作

近年来,在大数据领域,机器设备存储能力和算力的提升,极大地推动了大规模的数据中心设置的分布式机器学习的发展[9-12]。传统的分布式机器学习需要将总体数据集中到一个节点或者数据中心上进行训练。然而,随着手机、智能穿戴设备、传感器等移动设备的本地算力提升,以及近年来对用户数据隐私保护的限制[13],与将数据传输到中心节点相比,在分布式的设备上对本地数据进行训练,再传输训练参数到参数服务器的方法更加有效。这个问题被称为联邦学习,它需要解决大规模训练数据、隐私保护、异构的数据和设备等问题[14-16]。

McMahan 等[2]于 2016 年提出了一种基于迭代模型平均的深度网络联合学习方法(FedAvg),由于该方法的学习任务是通过由中央服务器协调的客户端联合来运行,各客户端像是组成了一个松散的联邦,因此被命名为联邦学习。相比数据中心式的分布式机器学习,联邦学习的一个主要优点是将模型训练和直接访问原始数据的需求分离开来,对数据隐私有严格要求或者数据难以集中共享的场景有着重大的意义;同时,FedAvg 算法采用多轮本地迭代的方式加速了学习效率,这对通信消耗的降低有很大的帮助。

Peter 等[17]讨论了联邦学习的最新进展,总结了目前联邦学习面临的急切挑战:1)客户端数据分布异构;2)客户数据隐私保护;3)通信限制;4)面对恶意攻击的鲁棒性;5)公平性问题。他们还指出,联邦学习中客户端的数据异构和设备异构极大地影响了学习的效率,这也是联邦学习目前面临的急切挑战之一。

Zhao 等[18]提出,当客户端数据处于非独立同分布时,应用 FedAvg 算法会有较高的精度损失,因此他们改进了 FedAvg 算法,提出可以使用土方运算来计算权重散度,提高联邦学习在异构分布数据中的模型准确度,并提出了一种基于数据共享策略的联邦学习,在中央服务器上创建共享的公共数据集来改进训练效果。但这种方式人为加入了误差,并且共享数据的方式本质上违背了数据隐私保护的原则。

Li 等[6]从目标函数入手,通过给模型的目标函数加上一个限制项,使得每个客户端在使用本地数据更新时,新的模型不会偏离全局模型太多,以此来减小数据异构带来的影响。Ghosh 等[19]和 Sattler 等[20]认为,由于数据异构的存在,我们无法得到一个精度足够的全局模型,因此提出了一种将客户端划分至不同的聚类,然后在各聚类中训练一个单独的全局模型的方式,他们采用不同的聚类方法对客户端的局部经验损失函数或者节点梯度进行聚类。由于各簇内的客户端具有较高的相似度,因此训练出来的簇内全局模型有较高的准确率,然而这种方法会得到多个全局模型,每个模型在某个簇类客户端中的效果很好,但模型泛化性较差,违背了共同训练的宗旨,并且他们运用的聚类方式都需要提前指定聚类的个数,在实际应用中存在困难。Yan 等[21]考虑了客户端间歇性可用的情况,认为不同客户参与到训练中的次数不一样,导致训练的模型向参与训练多的客户端数据倾斜,因此,在每轮选取客户端参与训练时,应优先选取参与训练次数少的客户端进行训练,以尽量保证每个客户端参与训练的次数相同。

2 算法设计

联邦学习中的优化模型为:

$$\min_{w}\{F(\boldsymbol{w})=\sum_{k=1}^{N}\rho_k F_k(\boldsymbol{w})\} \tag{1}$$

其中,N 是客户端总数,ρ_k 是第 k 个客户端的权重。假设第 k 个客户端的本地数据分布为 D^k,$\xi_t^k\sim D_t^k$ 是从本地数据中独立选择的样本,$F_k(\boldsymbol{w})=\ell_k(\boldsymbol{w},\xi_k)$,为模型的损失函数。在标准 FedAvg 算法中(假设第 t 轮迭代),首先中央参数服务器向所有参与训练的客户端广播最新的全局参数 $\boldsymbol{w}_t$,然后各参与客户端进行 E 轮的本地迭代:

$$\boldsymbol{w}_{t+i+1}^k=\boldsymbol{w}_{t+i}^k-\eta_{t+i}\nabla F_k(\boldsymbol{w}_{t+i}^k),i=0,1,\cdots,E-1 \tag{2}$$

其中,η_t 为学习率,假设每轮选取 $K(1\leqslant K<N)$ 个客户端参与训练,由中央服务器对收集到的客户端梯度进行聚合:

$$\boldsymbol{w}_{t+E}\leftarrow\frac{N}{K}\sum_{k=1}^{K}\rho_k\boldsymbol{w}_{t+E}^k \tag{3}$$

总体数据的分布是所有本地数据分布的混合:$D=\sum_{k=1}^{N}\rho_k D_k$。当客户端数据属于独立同分布时,对于所有的 $k\in N$,$D_k=D$。然而,在现实生活中,不同客户端的数据分布往往并不相同,因此我们的理论分析是基于数据非独立同分布的假设。

2.1 数据异构的影响

例 1 我们考虑一个分布式的优化问题,假设目标函数为一个凸函数,目标是从 N 个客户端中学习一维数据的平均值。$\xi_i\sim D_i$,均值 $e_i=E[\xi_i]$。我们可以将该目标转化为一个最小化均方差的问题:

$$\begin{aligned}f(x)&=\sum_{i=1}^{N}\rho_i f_i(x)=\sum_{i=1}^{N}\rho_i E_{\xi_i\sim D_i}[(x-\xi_i)^2]\\&=\sum_{i=1}^{N}\rho_i\ (x-e_i)^2+\sum_{i=1}^{N}\rho_i E_{\xi_i\sim D_i}[(\xi_i-e_i)^2]\end{aligned} \tag{4}$$

为了计算的方便,我们假设每个客户端含有的数据量相同,可以得到最优解为:

$$x^*=1/N(\sum_{i=1}^{N}e_i) \tag{5}$$

假设 τ_i 为因为通信丢失、客户端设备差异等原因造成的第 i 个客户端的权重偏移量,则目标函数将收敛于:

$$1/N(\sum_{i=1}^{N}e_i)+\sum_{i=1}^{N}\tau_i e_i \tag{6}$$

证明:对目标函数求导可得:

$$\nabla f(x)=2\sum_{i=1}^{N}\rho_i(x-e_i) \tag{7}$$

令导数为 0,可得:

$$x=\sum_{i=1}^{N}\rho_i e_i \tag{8}$$

根据每个客户端含有数据量相同的假设,对任意 $\rho_i(i\in N)$,$\rho_i=1/N$,因此 $x^*=1/N(\sum_{i=1}^{N}e_i)$。在现实条件下计算目标函数收敛值时,$\rho_i=1/N+\tau_i$,($\tau_i$ 为客户端 i 的聚合权重偏移量),$x=1/N(\sum_{i=1}^{N}e_i)+\sum_{i=1}^{N}\tau_i e_i$。当且仅当 $e_1=e_2=\cdots=e_n$(数据分布为 IID)或者对所有 $i\in\{1,2,\cdots,N\}$,$\tau_i=0$ 时,$x=x^*$。因此,传统的联邦学习算法在面对数据异构情况时,会导致不好的结果。

2.2 FedSSO 算法架构

如 2.1 节所述,数据异构和设备异构严重降低了 FedAvg 算法的性能。在联邦学习中,总体的数据分布是各客户端本地数据分布按权重的混合,在 FedAvg 算法中,这个权重为样本权重。这个设定只考虑了各客户端的数据量差异,并没有考虑到客户端的硬件设备差异以及通信差异等问题,例如考虑了联邦学习的经典实例,在训练手机输入法模型时,最新款的手机比老款的手机有更快的运行速度和传输速度,位于城镇等信号较好位置的手机比位于乡村或信号干扰地区的手机通信传输更稳定,这就造成了参数服务器受到的数据分布和实际分布之间存在差异。由于客户端异构的存在,某些类的数据更频繁地参与训练过程,从而给训练数据引入了误差。为了缓解误差问题,我们考虑在每轮训练时,按照数据的样本分类,选取全部数据分布类型的数据进行训练,保证每个类型的数据参与训练的概率基本相同,使训练数据分布为各客户端样本分布的无偏混合,这样就消除了训练数据中的偏差,并且建立了收敛性的结果。

FedSSO 算法的详细过程如算法 1 所示,FedSSO 算法的客户端选择原则是从不同簇中选择可用的客户端(见算法 1 中的第 2-8 行)。参数服务器首先初始化全局模型,然后将模型 $\boldsymbol{w}_0$ 广播至所有客户端,客户端根据接收到的模型,对本地数据的一个样本进行训练,从而得到本地模型参数 $\boldsymbol{w}_1^k$,参数服务器收集每个客户端的本地模型参数信息,每个客户端按式(9)进行更新:

$$\boldsymbol{w}_{t+1}^k\leftarrow\boldsymbol{w}_t-\eta_t\cdot\nabla\ell(\boldsymbol{w}_t;\xi_k) \tag{9}$$

客户端在第 $t+1$ 轮训练时,接收到上一轮的全局模型 $\boldsymbol{w}_t$,利用本地的数据集样本 ξ_k,对模型执行梯度下降算法后返回新的模型参数 $\boldsymbol{w}_{t+1}^k$。每轮训练时,上一轮的全局模型 $\boldsymbol{w}_t$ 和学习率 η_t 都是相同的,那么新的模型参数是一个只与本地数据 ξ_k 相关的参数,即参数服务器收集到的模型参数包含了客户端的数据分布信息。

可以采用 OPTICS[22] 聚类方法(Ordering Points to Identify the Clustering Structure,OPTICS)对收集到的模型参数进行聚类,将客户端划分到不同的簇中。

OPTICS 是一种基于密度的聚类算法,它将簇定义为通过密度连接的最大点集,并将具有足够密度的区域划分为簇。与 K-means 和 BIRCH 相比,OPTICS 可以在有噪声的空间数据中发现任意形状的聚类,而 K-means 和 BIRCH 仅适用于凸样本集聚类。与 DBSCAN 方法相比,OPTICS 对输入参数不敏感,提高了聚类的稳定性。与其他聚类方法相比,OPTICS 有几个优点:1)不需要事先知道形成的簇类数量;2)可以发现任何形状的簇类;3)可以检测噪声点并消除某些恶意攻击节点的影响;4)对输入参数不敏感。

每轮训练时按比例从每个簇中抽取可用的客户端参与训练,这样就保证了所有种类的数据都参与了每轮训练,降低了客户端异构带来的影响。每轮参与训练的客户端从参数服务器接收到最新的全局模型参数后,使用本地数据计算当前参数下的梯度,迭代 E 次随机梯度下降后将最新参数发回参数服务器,由参数服务器对传回的参数进行加权平均(见算法 1 中的第 9-13 行)。

算法 1 FedSSO

输入:初始化模型参数 $\mathbf{w}_0$;客户端总数 N;本地迭代轮次 E;客户端训练比例 β(每轮存于训练的客户端集合 K=βN);总训练轮次 T;学习率 η;聚类算法参数 min_samples,xi

输出：$\mathbf{w}_T$

1. $\forall i\in\{1,2,\cdots,N\}, \mathbf{w}_1^i \leftarrow \mathbf{w}_0-\eta\nabla\ell_i(\mathbf{w}_0,\xi_0^i)$

2. $c_1,c_2,c_3,\cdots\in C\leftarrow$OPTICS{min_samples, x_i, $(\mathbf{w}_1^1,\mathbf{w}_1^2,\cdots)$}

3. $\mathbf{w}_1\leftarrow\sum_{i=1}^{N}\rho_k\mathbf{w}_1^i$

4. for t=2 to T do

5. . for each cluster $(c_1,c_2,\cdots)\in C$ in parallel do

6. randomly select available clients

7. $S_t^i\leftarrow\beta\cdot num(c_i)$

8. $S_t\leftarrow(S_t^1,S_t^2,\cdots)$

9. for each client $k\in S_t$ do

10. $\mathbf{w}_t^k\leftarrow\mathbf{w}_{t-1}$

11. loop E iteration

12. $\mathbf{w}_t^k\leftarrow\mathbf{w}_t^k-\eta\nabla\ell_k(\mathbf{w}_t^k,\xi_t^k)$

13. $\mathbf{w}_t\leftarrow\frac{N}{K}\sum_{k\in S_t}\rho_k\mathbf{w}_t^k$

3 收敛性分析

本节证明了对于强凸、光滑的函数和异构数据集，FedSSO算法以$O(E^2/T)$收敛到全局最优；还分析了算法的收敛性条件以及学习率递减和本地迭代轮次选择机制的必要性。

3.1 符号和假设

我们对函数$F_1,F_2,\cdots,F_N$做了如下假设。

假设1 $F_1,F_2,\cdots,F_N$是L-smooth平滑的：

$\forall\boldsymbol{v}$ and $\boldsymbol{w}$, $f(\boldsymbol{v})\leqslant f(\boldsymbol{w})+(\boldsymbol{v}-\boldsymbol{w})^{\mathrm{T}}\nabla f(\boldsymbol{w})+L/2\cdot\|\boldsymbol{v}-\boldsymbol{w}\|^2$

假设2 $F_1,F_2,\cdots,F_N$是μ-strong强凸的：

$\forall\boldsymbol{v}$ and $\boldsymbol{w}$, $f(\boldsymbol{v})\geqslant f(\boldsymbol{w})+(\boldsymbol{v}-\boldsymbol{w})^{\mathrm{T}}\nabla f(\boldsymbol{w})+\mu/2\cdot\|\boldsymbol{v}-\boldsymbol{w}\|^2$

假设3 假设ξ_t^k为从第k个设备的本地数据中均匀地随机采样。每个客户端中，随机梯度的方差是有界的：

$E\|\nabla F_k(\mathbf{w}_t^k,\xi_t^k)-\nabla F_k(\mathbf{w}_t^k)\|^2\leqslant\sigma_k^2$, for $k=1,\cdots,N$

假设4 随机梯度的期望平方范数一致有界：

for all $k=1,\cdots,N$ and $t=0,\cdots,T-1$, $E\|\nabla F_k(\mathbf{w}_t^k,\xi_t^k)\|^2\leqslant G^2$

(1)数据异构的量化指标。假设F^*和F_k^*分别为目标函数F和F_k的最优解，我们可以得到$\Gamma\leftarrow x^*-\sum_{k=1}^{N}\rho_k x_k^*$来量化数据异构的程度。当客户端数据分布为IID时，$\Gamma=0$。数据异构程度越高，$|\Gamma|$值越高。

(2)设备异构的量化指标。假设τ_i为第i个客户端在模型聚合时其聚合权重相比其参数权重在期望上的差值(这个差值受设备算力、通信环境等的影响)。假设F_k^*为目标函数F_k的最优解，可计算得到$M=\sum_{k=1}^{N}\tau_k F_k^*$，为设备异构的量化指标。

3.2 例1的收敛性分析

首先证明FedSSO在例1中是收敛的，而FedAvg会产生一个带偏差的结果。根据2.2节中的设定，我们将不同客户端划分入不同簇中$\{c_1,c_2,\cdots,c_n\}$，每个簇c_i含有客户端数量为n_{c_i}。位于相同簇内的客户端具有相似的均值，每个簇的均值$E(\xi_{ci})=e_{ci}$。$\exists\delta_{ci}\geqslant 0$，对任意一个客户端$k\in\{c_i\}$，$|x_k^*-e_{ci}|\leqslant\delta_{ci}$。我们可以将目标函数改写为：

$$\begin{aligned}f(x)&=\sum_{i=1}^{n}\rho_{ci}f_{ci}(x)\\&=\sum_{i=1}^{n}\rho_{ci}(x-e_{ci})^2+\sum_{i=1}^{n}\rho_{ci}E_{\xi_{ci}\sim D_{ci}}[(\xi_{ci}-e_{ci})^2]\end{aligned}\tag{10}$$

因为每个簇类中抽取的都是可用客户端，不存在通信丢失等客户端不可用的情况，所以$\rho_{ci}=\frac{n_{ci}}{N}$，式(10)可改写为：

$$f(x)=\sum_{i=1}^{n}\frac{n_{ci}}{N}(x-e_{ci})^2+\sum_{i=1}^{n}\frac{n_{ci}}{N}E_{\xi_{ci}\sim D_{ci}}[(\xi_{ci}-e_{ci})^2]$$

当学习率$\eta\leqslant 2/L$，采取梯度下降方法计算，可知上式的解为：$x=\sum_{i=1}^{n}\frac{n_{ci}}{N}e_{ci}$。已知$x^*=\sum_{i=1}^{N}\frac{e_i}{N}$，可得：

$$|x-x^*|=\left|\sum_{i=1}^{n}\frac{n_{ci}}{N}e_{ci}-\sum_{i=1}^{N}\frac{e_i}{N}\right|\leqslant\frac{1}{N}\sum_{i=1}^{n}n_{ci}\delta_{ci}\tag{11}$$

当聚类足够精确时，可以认为每个簇中的数据是同分布的，$\delta_{ci}\to 0$目标函数收敛到最优解。

3.3 分层抽样带来的提升

本节将讨论对比随机抽样的方法，采用分层抽样带来的提升。首先证明分层抽样实现了对于总体数据的一个无偏抽样，即：

$$E_{S_t}(w_t)=E_{S_t}\Big(\sum_{k\in S_t}w_k(S_t)\cdot\boldsymbol{w}_t^k\Big)=\sum_{k=1}^{N}\rho_k\boldsymbol{w}_t^k\tag{12}$$

其中，$w_k(S_t)$代表对于客户端集合S_t，客户端k的聚合权重。

假设总体数据集中共有m种不同的分布模式，同样地，为了方便分析，我们假设每个客户端的本地数据数量相同。通过聚类，可以将全部的客户端划分成m个集合，根据假设，我们认为每个聚类集合中的客户端具有相同的数据分布模式，因此其训练出来的神经网络模型具有相同的参数信息，即：

$$\forall i\in[1,m],\forall j,k\in c_i,\boldsymbol{w}^j=\boldsymbol{w}^k\tag{13}$$

由式(13)可知，同一聚类中的客户端模型参数相同，因此用$\boldsymbol{w}_t^{c_i}$代表第t轮训练时，第i个聚类中的模型参数，可知：

$$\forall i\in[1,m],\forall k\in c_i,\sum_{k\in c_i}\boldsymbol{w}_t^k=n_{c_i}\cdot\boldsymbol{w}_t^{ci}\tag{14}$$

根据算法1，我们可以在每轮模型聚合时，从不同簇中聚合模型，假设聚类数量为m，可得：

$$E_{S_t}(\boldsymbol{w}_t)=E_{S_t}\Big(\sum_{k\in S_t}r_t^k\boldsymbol{w}_t^k\Big)=\sum_{i=1}^{m}\frac{n_{ci}}{N}\boldsymbol{w}_t^{ci}\tag{15}$$

由式(14)可以得到：

$$\sum_{i=1}^{m}\frac{n_{ci}}{N}\boldsymbol{w}_t^{ci}=\frac{1}{N}\sum_{i=1}^{m}n_{ci}\boldsymbol{w}_t^{ci}=\frac{1}{N}\sum_{k=1}^{N}\boldsymbol{w}_t^k=\sum_{k=1}^{N}\rho_k\boldsymbol{w}_t^k\tag{16}$$

由此证明了采用分层抽样的方式可以实现对全部数据的无偏抽样。

同时，我们将证明：使用分层抽样的方式可以降低客户端的聚合权重方差，使模型更新更加平稳。对于传统的随机抽样方法，各客户端聚合权重等同于其样本权重，$\rho_k\leftarrow n_k/M$，其中n_k为客户端k的样本数据，M为总体样本数据，根据本节关于客户端本地数据数量相同的假设，可简化为$\rho_k\leftarrow 1/N$。根据总体数据有m种不同分布的假设，我们选取每轮参与训练的客户端数量为m，在随机抽取的方式中，m个客户端根据伯努利分布$B(\rho_k)$被随机抽取，对于客户端k，其聚合权重方差为：

$$Var(w_k(S_t))=\frac{1}{m^2}mVar(B(\rho_k))$$
$$=\frac{1}{m}\rho_k(1-\rho_k)$$
$$=\frac{1}{mN}(1-\frac{1}{N}) \tag{17}$$

在分层抽样中,我们首先将总体客户端划分成 m 个聚类集合,然后在每个集合中抽取一个客户端参与训练,用 $Var_c(w_k(S_t))$ 表示客户端 k 在分层抽样方式下的聚合权重方差,可知:

$$Var_c(w_k(S_t))=\frac{1}{m^2}Var(B(r_t^k))$$
$$=\frac{1}{m^2}m\rho_k(1-m\rho_k)$$
$$=\frac{1}{mN}(1-\frac{m}{N}) \tag{18}$$

可知 $Var(w_k(S_t))\geqslant Var_c(w_k(S_t))$,仅在 $m=1$,即各客户端数据独立同分布时,二者相等。

3.4 收敛性分析结果

3.4.1 设备全部参与

本节讨论 FedSSO 算法在全客户端参与训练下的收敛性问题。事实上,由于 FedSSO 算法针对的是客户端选择策略的改变,当全体客户端都参与训练时,FedSSO 可以等同于 FedAvg 算法。FedAvg 算法的收敛性已经有广泛的证明,但之前的证明在对参数聚合权重的预设上,未考虑通信丢失和客户端设备异构等情况。事实上,上述情况给训练目标和最优解引入了偏差。在证明过程中,我们将引入变量来表示因客观原因导致的客户端聚合权重改变。

假设算法在 T 轮迭代后终止,返回 $\boldsymbol{w}_T$ 作为解。E 为客户端本地迭代轮次,我们要求 T 为 E 的整数倍,这样就可以像预期一样输出 $\boldsymbol{w}_T$。

假设 $\boldsymbol{w}_t^k$ 表示第 t 轮训练中第 k 个客户端的模型参数,E 为本地迭代次数,取集合 $\Phi_E=\{nE\mid n=1,2,3,\cdots\}$,代表客户端和参数服务器进行通信的步骤。如果 $t+1\in\Phi_E$,参数服务器将各本地模型集合得到全局模型,并将最新的全局模型发放给客户端。当 $t+1\notin\Phi_E$ 时,各客户端利用本地数据更新本地模型参数。由于客户端需要在本机进行多轮迭代,我们取一个中间变量 $\boldsymbol{v}_{t+1}^k$ 表示由 $\boldsymbol{w}_t^k$ 单步 SGD 后的结果。全部客户端参与的更新结果可表示为:

$$\boldsymbol{v}_{t+1}^k=\boldsymbol{w}_t^k-\eta_t\nabla F_k(\boldsymbol{w}_t^k,\xi_t^k) \tag{19}$$

$$\boldsymbol{w}_{t+1}^k=\begin{cases}\boldsymbol{v}_{t+1}^k, & \text{if } t+1\notin\Phi_E\\ \sum_{k=1}^N\rho_k\boldsymbol{v}_{t+1}^k, & \text{if } t+1\in\Phi_E\end{cases} \tag{20}$$

我们定义 $\bar{\boldsymbol{v}}_t=\sum_{k=1}^N\boldsymbol{q}_k\boldsymbol{v}_t^k$,$\bar{\boldsymbol{w}}_t=\sum_{k=1}^N\rho_t^k\boldsymbol{w}_t^k$,$q_k$ 为客户端样本权重,即在无通信丢失等误差影响下的权重,ρ_k 为真实权重,定义 τ_t^k 为第 t 轮第 k 个客户端的权重改变量,$\rho_t^k-q_k=\tau_t^k$。从期望的角度去计算误差,可知 $E(\tau_t^k\mid t\in\{1,2,\cdots,T\})=\tau_k$,表示第 k 个客户端在总体训练中的权重误差。定义最优解为 $\boldsymbol{w}^*$,事实上由于期望误差的存在,模型训练追求的最优解和实际最优解之间存在差距,定义 $\hat{\boldsymbol{w}}^*$ 为模型训练的目标解,可知:

$$E[F(\boldsymbol{w}^*)-F(\hat{\boldsymbol{w}}^*)]=E[\sum_{k=1}^N\rho_kF(\boldsymbol{w}_k^*)-\sum_{k=1}^Nq_kF(\boldsymbol{w}_k^*)]$$
$$=\sum_{k=1}^N\tau_kF(\boldsymbol{w}_k^*) \tag{21}$$

用 F^* 代表最优解,$\hat{F}^*$ 代表带误差的目标解,定义 $M=E(F_k^*)$。假设 $\bar{\boldsymbol{g}}_t=\sum_{k=1}^N\rho_k\nabla F_k(\boldsymbol{w}_t^k)$,$\boldsymbol{g}_t=\sum_{k=1}^N\rho_k\nabla F_k(\boldsymbol{w}_t^k,\xi_t^k)$,可知 $\bar{\boldsymbol{g}}_t=E\boldsymbol{g}_t$,$\bar{\boldsymbol{w}}_{t+1}=\bar{\boldsymbol{w}}_t-\eta_t\boldsymbol{g}_t$。

引理 1 如果函数 F 是 μ-strong 强凸的,我们可以得到:

$$\|\nabla F(\bar{\boldsymbol{w}}_t)\|^2\geqslant 2\mu[F(\bar{\boldsymbol{w}}_t)-F(\boldsymbol{w}^*)],\forall t\in\{1,2,\cdots,T\} \tag{22}$$

证明:由函数的强凸性可以得出:

$$F(y)-F(x)\geqslant\nabla F(x)^{\mathrm{T}}(y-x)+\frac{\mu}{2}\|y-x\|^2$$

对于上式大于等于符号右侧的项,对 y 求导可知,取 $\hat{y}=x-\frac{1}{u}\nabla F(x)$ 时为最小值,将 $\hat{y}$ 代入上式可得:

$$F(y)-F(x)\geqslant\nabla F(x)^{\mathrm{T}}(-\frac{1}{u}\nabla F(x))+\frac{\mu}{2}\|-\frac{1}{u}\nabla F(x)\|^2$$
$$=-\frac{1}{2u}\|\nabla F(x)\|^2$$

将 $\bar{\boldsymbol{w}}_t$ 和 $\boldsymbol{w}^*$ 代入上式得证。

引理 2 如果函数 F 是 L-smooth 的,同上述证明可以得到:

$$\|\nabla F(\bar{\boldsymbol{w}}_t)\|^2\leqslant 2L[F(\bar{\boldsymbol{w}}_t)-F(\boldsymbol{w}^*)],\forall t\in\{1,2,\cdots,T\} \tag{23}$$

引理 3 根据假设 4,假定学习率 η_t 是非增的,对所有 $t\geqslant 0$,$\eta_t\leqslant 2\eta_{t+E}$ 可以得到:

$$E[\sum_{k=1}^N\rho_k\|\bar{\boldsymbol{w}}_t-\boldsymbol{w}_k^t\|^2]\leqslant 4\eta_t^2(E-1)^2G^2 \tag{24}$$

证明:FedSSO 算法在经历 E 轮本地迭代后需要一次参数服务器的通信聚合,因此对于任何的 $t\geqslant 0$,取 $t_0\leqslant t$,$t-t_0\leqslant E-1$。根据定义,η_t 是非增的,$\eta_{t_0}\leqslant 2\eta_t$ 可以得到:

$$E[\sum_{k=1}^N\rho_k\|\bar{\boldsymbol{w}}_t-\boldsymbol{w}_k^t\|^2]$$
$$=E[\sum_{k=1}^N\rho_k\|(\boldsymbol{w}_t^k-\bar{\boldsymbol{w}}_{t_0})-(\bar{\boldsymbol{w}}_t-\bar{\boldsymbol{w}}_{t_0})\|^2]$$
$$\leqslant E\sum_{k=1}^N\rho_k\|\boldsymbol{w}_t^k-\bar{\boldsymbol{w}}_{t_0}\|^2$$
$$\leqslant\sum_{k=1}^N\rho_kE\sum_{t=t_0}^{t-1}(E-1)\eta_t^2\|\nabla F_k(\boldsymbol{w}_t^k,\xi_t^k)\|^2$$
$$\leqslant\sum_{k=1}^N\rho_kE\sum_{t=t_0}^{t-1}(E-1)\eta_t^2G^2$$
$$\leqslant\sum_{k=1}^N\rho_k\eta_{t_0}^2(E-1)^2G^2$$
$$\leqslant 4\eta_t^2(E-1)^2G^2$$

引理 4 假定学习率 η_t 是非增的,对所有 $t\geqslant 0$,$\eta_t\leqslant 2\eta_{t+E}$,可以得到:

$$E(\|\boldsymbol{g}_t\|^2)\leqslant\sum_{k=1}^N\rho_k\sigma_k^2+(4L^2(E-1)^2+\frac{L}{\mu})G^2+2L\Gamma \tag{25}$$

证明:由 $\|\cdot\|^2$ 的凸性和引理 2 可得:

$$E(\|\boldsymbol{g}_t\|^2)=D(\|\boldsymbol{g}_t\|)+\|\bar{\boldsymbol{g}}_t\|^2$$
$$\leqslant\sum_{k=1}^N\rho_k\sigma_k^2+\sum_{k=1}^N\rho_k\|\nabla F_k(\boldsymbol{w}_t^k)\|^2$$
$$\leqslant\sum_{k=1}^N\rho_k\sigma_k^2+2L\sum_{k=1}^N\rho_k(F_k(\boldsymbol{w}_t^k)-F_k^*)$$

我们定义了 $\Gamma=F^*-\sum_{k=1}^{N}\rho_k F_k^*$，上式可转化为：

$$\begin{aligned}E(\|\boldsymbol{g}_t\|^2)&\leqslant\sum_{k=1}^{N}\rho_k\sigma_k^2+2L[\sum_{k=1}^{N}\rho_k(F_k(\boldsymbol{w}_t^k)-F^*)+\sum_{k=1}^{N}\rho_k(F^*-F_k^*)]\\&=\sum_{k=1}^{N}\rho_k\sigma_k^2+2L[\sum_{k=1}^{N}\rho_k(F_k(\boldsymbol{w}_t^k)-F^*)+\Gamma]\end{aligned}$$

$$\begin{aligned}&\sum_{k=1}^{N}\rho_k(F_k(\boldsymbol{w}_t^k)-F^*)\\&=\sum_{k=1}^{N}\rho_k(F_k(\boldsymbol{w}_t^k)-F(\bar{\boldsymbol{w}}_t))+\sum_{k=1}^{N}\rho_k(F(\bar{\boldsymbol{w}}_t)-F^*)\\&=\sum_{k=1}^{N}\rho_k(F_k(\boldsymbol{w}_t^k)-F(\bar{\boldsymbol{w}}_t))+(F(\bar{\boldsymbol{w}}_t)-F^*)\\&\leqslant\sum_{k=1}^{N}\rho_k\cdot\frac{L}{2}\|\boldsymbol{w}_t^k-\bar{\boldsymbol{w}}_t\|^2+(F(\bar{\boldsymbol{w}}_t)-F^*)\\&\leqslant 2L\eta_t^2(E-1)^2G^2+(F(\bar{\boldsymbol{w}}_0)-F^*)\\&\leqslant 2L(E-1)^2G^2+\frac{1}{2\mu}G^2\\&=[L(E-1)^2+\frac{1}{2\mu}]G^2\end{aligned}$$

综上可得：

$$\begin{aligned}E(\|\boldsymbol{g}_t\|^2)&\leqslant\sum_{k=1}^{N}\rho_k\sigma_k^2+2L[\sum_{k=1}^{N}\rho_k(F_k(\boldsymbol{w}_t^k)-F^*)+\Gamma]\\&\leqslant\sum_{k=1}^{N}\rho_k\sigma_k^2+2L((2L(E-1)^2+\frac{1}{2\mu})G^2+\Gamma)\\&=\sum_{k=1}^{N}\rho_k\sigma_k^2+(4L^2(E-1)^2+\frac{L}{\mu})G^2+2L\Gamma\end{aligned}$$

定理 1 根据假设 1－假设 4，并且 L,μ,σ_k,G 如假设中所定义的，假定学习率 η_t 是递减的，$\gamma>0$，对所有 $t\geqslant 0$，$\eta_t\leqslant 2\eta_{t+E}$。全部客户端参与的 FedSSO 算法满足：

$$E[F(\bar{\boldsymbol{w}}_T)]-F^*\leqslant\frac{v}{T+\gamma}-M \tag{26}$$

其中：

$$B=\sum_{k=1}^{N}\rho_k\sigma_k^2+(4L^2(E-1)^2+\frac{L}{\mu})G^2+2L\Gamma$$

$$v=\max\left\{\frac{L\beta^2B}{4\beta\mu-2},(\gamma+1)(F(\bar{\boldsymbol{w}}_1)-\hat{F}^*)\right\},M=\sum_{k=1}^{N}\tau_k F_k^*。$$

证明：由假设 1 可知：

$$\begin{aligned}F(\bar{\boldsymbol{w}}_{t+1})-F(\bar{\boldsymbol{w}}_t)&\leqslant\nabla F(\bar{\boldsymbol{w}}_t)^{\mathrm{T}}(\bar{\boldsymbol{w}}_{t+1}-\bar{\boldsymbol{w}}_t)+\frac{L}{2}\|\bar{\boldsymbol{w}}_{t+1}-\bar{\boldsymbol{w}}_t\|^2\\&=-\eta_t\cdot\nabla F(\bar{\boldsymbol{w}}_t)^{\mathrm{T}}\boldsymbol{g}_t+\frac{L}{2}\eta_t^2\|\boldsymbol{g}_t\|^2\end{aligned}$$

由于模型更新采取的是随机梯度的方式，因此我们假设第 t 轮训练的样本为 ξ_t，对上式关于 ξ_t 取期望可得：

$$\begin{aligned}E_{\xi_t}[F(\bar{\boldsymbol{w}}_{t+1})]-F(\bar{\boldsymbol{w}}_t)&\leqslant-\eta_t\cdot\|\nabla F(\bar{\boldsymbol{w}}_t)\|^2+\\&\frac{L}{2}\eta_t^2E_{\xi_t}(\|\boldsymbol{g}_t\|^2)\end{aligned}$$

根据引理 1 和引理 4，假设 $B=\sum_{k=1}^{N}\rho_k\sigma_k^2+(4L^2(E-1)^2+\frac{L}{\mu})G^2+2L\Gamma$，可得：

$$\begin{aligned}E_{\xi_t}[F(\bar{\boldsymbol{w}}_{t+1})]-F(\bar{\boldsymbol{w}}_t)&\leqslant-\eta_t\cdot\|\nabla F(\bar{\boldsymbol{w}}_t)\|^2+\frac{L}{2}\eta_t^2B\\&\leqslant-2\mu\eta_t[F(\bar{\boldsymbol{w}}_t)-\hat{F}^*]+\frac{L}{2}\eta_t^2B\end{aligned}$$

上式两边同时减去 $\hat{F}^*$，并求期望可得：

$$E[F(\bar{\boldsymbol{w}}_{t+1})-\hat{F}^*]\leqslant(1-2\mu\eta_t)E[F(\bar{\boldsymbol{w}}_t)-\hat{F}^*]+\frac{L}{2}\eta_t^2B$$

由于 η_t 是递减的，我们定义 $\eta_t=\frac{\beta}{t+\gamma}$，$\beta>\frac{1}{2\mu}$，$\gamma>0$，使得 $\eta_1\leqslant\frac{1}{2\mu}$，并且 $\eta_t\leqslant 2\eta_{t+E}$，可用归纳法证明：

$$E[F(\bar{\boldsymbol{w}}_t)-\hat{F}^*]\leqslant\frac{v}{t+\gamma},v=\max\left\{\frac{L\beta^2B}{4\beta\mu-2},(\gamma+1)(F(\bar{\boldsymbol{w}}_1)-\hat{F}^*)\right\}$$

首先根据 v 的定义可知，当 $t=1$ 时上式成立，假设对 t 上式成立，当 $t+1$ 时：

$$\begin{aligned}E[F(\bar{\boldsymbol{w}}_{t+1})-\hat{F}^*]&\leqslant(1-2\mu\eta_t)E[F(\bar{\boldsymbol{w}}_t)-\hat{F}^*]+\frac{L}{2}\eta_t^2B\\&\leqslant(1-2\frac{\mu\beta}{t+\gamma})\frac{v}{t+\gamma}+\frac{L}{2}\frac{\beta^2}{(t+\gamma)^2}B\\&\leqslant\frac{t+\gamma-1}{(t+\gamma)^2}\boldsymbol{v}-\frac{2\mu\beta-1}{(t+\gamma)^2}\boldsymbol{v}+\frac{L}{2}\frac{\beta^2}{(t+\gamma)^2}B\\&\leqslant\frac{v}{t+\gamma+1}\end{aligned}$$

由于 $F^*-\hat{F}^*=M$，$M=\sum_{k=1}^{N}\tau_k F_k^*$，我们可得：

$$E[F(\bar{\boldsymbol{w}}_{t+1})]-F^*\leqslant\frac{v}{t+\gamma+1}-M$$

其中，$v=\max\left\{\frac{L\beta^2B}{4\beta\mu-2},(\gamma+1)(F(\bar{\boldsymbol{w}}_1)-\hat{F}^*)\right\}$，$M=\sum_{k=1}^{N}\tau_k F_k^*$。

3.4.2 设备部分参与

本节讨论了 FedSSO 算法在部分客户端参与条件下的收敛性问题。由于全客户端参与的模式下，联邦学习会受到“straggler's effect”的严重影响，即所有的节点都要等一个最慢的节点，因此部分客户端参与的联邦学习有着更现实的应用。假设 $S_t\subseteq\{1,2,\cdots,N\}$，为 k-th 迭代时参与训练的客户端集合，S_t 由从各个簇中随机抽取的客户端组合而成，每轮抽取的客户端总数为 K。假设各客户端的数据量是平衡的，在每轮训练中选取的都是可用客户端，因此不会受到通信丢失等影响，故 $\rho_1=\rho_2=\cdots=\rho_N=1/N$，FedSSO 的聚合步骤可表示为：

$$\boldsymbol{w}_{t+E}\leftarrow\frac{N}{K}\sum_{k\in S_t}\rho_k\boldsymbol{w}_{t+E}^k \tag{27}$$

定义 $\rho_1=\rho_2=\cdots=\rho_N=1/N$ 似乎违背了联邦学习关于不平衡的假设，我们可以通过如下转化来解决这个问题。假设 $\widetilde{F}_k(\boldsymbol{w})=\rho_k N F_k(\boldsymbol{w})$，可以看作是对目标函数进行了缩放。则全局目标函数可转化为：

$$F(\boldsymbol{w})=\sum_{k=1}^{N}\rho_k F_k(\boldsymbol{w})=\frac{1}{N}\sum_{k=1}^{N}\widetilde{F}_k(\boldsymbol{w}) \tag{28}$$

以上假设受到了文献[23]的启发。

假设 $\boldsymbol{w}_t^k$ 表示第 t 轮训练中第 k 个客户端的模型参数，E 为本地迭代次数，取集合 $\Phi_E=\{nE|n=1,2,3,\cdots\}$，代表客户端和参数服务器进行通信的步骤。不同于全部设备参与中的情况，当 $t+1\in\Phi_E$ 时，参数服务器随机接收一部分客户端的信息进行中央聚合。相同地，我们取一个中间变量 $\boldsymbol{v}_{t+1}^k$ 表示由 $\boldsymbol{w}_t^k$ 单步 SGD 后的结果。部分客户端参与的更新结果可表示为：

$$\boldsymbol{v}_{t+1}^k=\boldsymbol{w}_t^k-\eta_t\nabla F_k(\boldsymbol{w}_t^k,\xi_t^k) \tag{29}$$

$$\boldsymbol{w}_{t+1}^k=\begin{cases}\boldsymbol{v}_{t+1}^k, & \text{if } t+1\notin\Phi_E\\\sum_{k\in S_{t+1}}\rho_k\boldsymbol{v}_{t+1}^k, & \text{if } t+1\in\Phi_E\end{cases} \tag{30}$$

定义 $\bar{\boldsymbol{v}}_t=\sum_{k=1}^{N}\rho_k\boldsymbol{v}_t^k$,$\bar{\boldsymbol{w}}_t=\sum_{k=1}^{N}\rho_k\boldsymbol{w}_t^k$。假设 $\bar{\boldsymbol{g}}_t=\sum_{k=1}^{N}\rho_k\nabla F_k(\boldsymbol{w}_t^k)$,$\boldsymbol{g}_t=\sum_{k=1}^{N}\rho_k F_k(\boldsymbol{w}_t^k,\xi_t^k)$,可知 $\bar{\boldsymbol{g}}_t=E\boldsymbol{g}_t$,$\bar{\boldsymbol{v}}_{t+1}=\bar{\boldsymbol{w}}_t-\eta_t\boldsymbol{g}_t$。

客户端选择策略如下:每轮训练开始前,参数服务器将最新的模型参数传输给所有客户端,客户端在接收到最新模型参数后,利用本地数据进行训练,然后将训练好的模型参数传输给参数服务器。参数服务器根据先期聚类信息,在每个簇中按权重比例接收到一定数目的客户端模型参数后,参数服务器便不再接收此簇中其他客户端的模型参数,每个簇在第 t 轮训练中选中的客户端集合表示为 S_t^{ci},第 t 轮训练客户端的集合为 S_t。

在 $t+1\notin\Phi_E$ 时,可知 $\bar{v}_{t+1}=\bar{w}_{t+1}$,但当 $t+1\in\Phi_E$ 时,二者并不相等。因此,我们需要在期望上建立二者的关系。经分析,主要有两个地方引入了随机性,一个是随机梯度和全梯度的误差,一个是部分客户端参与和全客户端参与的误差。我们在全客户端参与的算法收敛性分析中主要分析了随机梯度带来的误差,在这一部分主要解决部分客户端参与给算法带来的不确定性,用 E_{S_t} 表示在集合 S_t 上的期望。

引理 5 if $t+1\in\Phi_E$,假定 η_t 非增,对所有 $t\geqslant 0$,$\eta_t\leqslant 2\eta_{t+E}$,我们可以得到:

$$E_{S_t}[F(\bar{\boldsymbol{w}}_{t+1})-F(\bar{\boldsymbol{v}}_{t+1})]\leqslant\frac{LN}{2K(N-1)}\left(1-\frac{K}{N}\right)4\eta_t^2E^2G^2 \tag{31}$$

证明:根据假设 1 可知:

$$E_{S_t}[F(\bar{\boldsymbol{w}}_{t+1})-F(\bar{\boldsymbol{v}}_{t+1})]\leqslant E_{S_t}\left[\nabla F(\bar{\boldsymbol{v}}_{t+1})^{\mathrm{T}}(\bar{\boldsymbol{w}}_{t+1}-\bar{\boldsymbol{v}}_{t+1})+\frac{L}{2}\|\bar{\boldsymbol{w}}_{t+1}-\bar{\boldsymbol{v}}_{t+1}\|^2\right]$$

在客户端选择策略中,每轮随机选取的都是可用客户端,不存在通信丢失的情况,因此 $\rho_1=\rho_2=\cdots=\rho_N=\frac{1}{N}$,并且每轮训练的所有类型的客户端都会参与训练,可知 $E_{S_t}(\bar{\boldsymbol{w}}_{t+1})=\bar{\boldsymbol{v}}_{t+1}$,上式可简化为:

$$E_{S_t}[F(\bar{\boldsymbol{w}}_{t+1})-F(\bar{\boldsymbol{v}}_{t+1})]\leqslant\frac{L}{2}E_{S_t}[\|\bar{\boldsymbol{w}}_{t+1}-\bar{\boldsymbol{v}}_{t+1}\|^2]$$

根据假定 $\rho_1=\rho_2=\cdots=\rho_N=\frac{1}{N}$,同时已知 $\bar{\boldsymbol{w}}_{t+1}=\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{v}_{t+1}^k$,可得:

$$E_{S_t}\|\bar{\boldsymbol{w}}_{t+1}-\bar{\boldsymbol{v}}_{t+1}\|^2=E_{S_t}\left\|\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{v}_{t+1}^k-\bar{\boldsymbol{v}}_{t+1}\right\|^2$$

由于客户端选择策略采取的是不放回随机抽样的方式,因此每个客户端在每轮训练中只会被选择一次,可得:

$$\begin{aligned}&E_{S_t}\left\|\frac{1}{K}\sum_{k=1}^{K}\boldsymbol{v}_{t+1}^k-\bar{\boldsymbol{v}}_{t+1}\right\|^2\\&=\frac{1}{K^2}\Big[\sum_{k=1}^{N}\frac{K}{N}\|\boldsymbol{v}_{t+1}^k-\bar{\boldsymbol{v}}_{t+1}\|^2+\sum_{k,j\in S_{t+1},k\neq j}\frac{K(K-1)}{N(N-1)}\langle\boldsymbol{v}_{t+1}^k-\bar{\boldsymbol{v}}_{t+1},\boldsymbol{v}_{t+1}^j-\bar{\boldsymbol{v}}_{t+1}\rangle\Big]\\&=\frac{1}{K(N-1)}\left(1-\frac{K}{N}\right)\sum_{i=1}^{N}\|\boldsymbol{v}_{t+1}^k-\bar{\boldsymbol{v}}_{t+1}\|^2\end{aligned}$$

其中:

$$\sum_{k=1}^{N}\|\boldsymbol{v}_{t+1}^k-\bar{\boldsymbol{v}}_{t+1}\|^2+\sum_{k,j\in S_{t+1},k\neq j}\frac{K(K-1)}{N(N-1)}\langle\boldsymbol{v}_{t+1}^k-\bar{\boldsymbol{v}}_{t+1},\boldsymbol{v}_{t+1}^j-\bar{\boldsymbol{v}}_{t+1}\rangle=0$$

。因此:

$$\begin{aligned}&E_{S_t}[F(\bar{\boldsymbol{w}}_{t+1})-F(\bar{\boldsymbol{v}}_{t+1})]\\&\leqslant\frac{L}{2K(N-1)}\left(1-\frac{K}{N}\right)E[\sum_{i=1}^{N}\|\boldsymbol{v}_{t+1}^k-\bar{\boldsymbol{v}}_{t+1}\|^2]\\&\leqslant\frac{LN}{2K(N-1)}\left(1-\frac{K}{N}\right)E[\frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{v}_{t+1}^k-\bar{\boldsymbol{w}}_{t_0}\|^2]\\&\leqslant\frac{LN}{2K(N-1)}\left(1-\frac{K}{N}\right)4\eta_t^2E^2G^2\end{aligned}$$

定理 2 根据假设 1—假设 4,并且 L,μ,σ_k,G 如假设中所定义的,假定学习率 η_t 是递减的,$\gamma>0$,对所有 $t\geqslant 0$,$\eta_t\leqslant 2\eta_{t+E}$,$B$ 如定理 1 中的定义,$A=\frac{N}{K(N-1)}\left(1-\frac{K}{N}\right)4E^2G^2$,可得:

$$E[F(\bar{\boldsymbol{w}}_T)]-F^*\leqslant\frac{v}{\gamma+T} \tag{32}$$

其中,$v=\max\left\{\frac{L\beta^2(B+A)}{4\mu\beta-2},(\gamma+1)(F(\bar{\boldsymbol{v}}_1)-F^*)\right\}$。

证明:已知:

$$E_{S_t}[F(\bar{\boldsymbol{w}}_{t+1})-F^*]=E_{S_t}[F(\bar{\boldsymbol{w}}_{t+1})-F(\bar{\boldsymbol{v}}_{t+1})]+E_{S_t}[F(\bar{\boldsymbol{v}}_{t+1})-F^*]$$

当 $t+1\notin\Phi_E$ 时,$\bar{\boldsymbol{w}}_{t+1}=\bar{\boldsymbol{v}}_{t+1}$,可知:

$$E[F(\bar{\boldsymbol{w}}_{t+1})-\hat{F}^*]\leqslant(1-2\mu\eta_t)E[F(\bar{\boldsymbol{w}}_t)-\hat{F}^*]+\frac{L}{2}\eta_t^2B$$

其中:

$$B=\sum_{k=1}^{N}\rho_k\sigma_k^2+4L^2\eta_t^2(E-1)^2G^2+2L(F(\bar{\boldsymbol{w}}_0)-F^*)+2L\Gamma$$

当 $t+1\in\Phi_E$ 时,根据引理 5 可知:

$$E[F(\bar{\boldsymbol{w}}_{t+1})-\hat{F}^*]\leqslant(1-2\mu\eta_t)E[F(\bar{\boldsymbol{w}}_t)-\hat{F}^*]+\frac{L}{2}\eta_t^2(B+A)$$

其中,$A=\frac{N}{K(N-1)}\left(1-\frac{K}{N}\right)4E^2G^2$。$\eta_t$ 是递减的,$\eta_t=\frac{\beta}{t+\gamma}$,$\beta>\frac{1}{2\mu}$,$\mu>0$,并且 $\eta_t\leqslant 2\eta_{t+E}$,可以使用定理 1 中相同的方法去证明:

$$E[F(\bar{\boldsymbol{w}}_t)]-F^*\leqslant\frac{v}{\gamma+t}$$

其中,$v=\max\left\{\frac{L\beta^2(B+C)}{4\mu\beta-2},(\gamma+1)(F(\bar{\boldsymbol{w}}_1)-F^*)\right\}$。

3.5 学习率递减的必要性

本节证明了选择一个逐渐下降的学习率对于在客户端异构条件下的联邦学习收敛性十分必要。在之前的证明过程中,我们得到:

$$E_{\xi_t}[F(\bar{\boldsymbol{w}}_{t+1})]-F(\bar{w}_t)\leqslant-\eta_t\cdot\|\nabla F(\bar{\boldsymbol{w}}_t)\|^2+\frac{L}{2}\eta_t^2B \tag{33}$$

可见算法的更新过程类似 Markov 过程,也就是全局模型的下一步更新与历史无关,只与当前的参数有关。由于模型的更新由两个部分决定,我们可以看出对于式(33),第一项为负,第二项为正,因此学习率的选择对算法收敛性有重要的影响。

当模型参数接近最优解时,$\nabla F(\bar{\boldsymbol{w}}_t)\to 0$,若学习率 η_t 是一个常数,那么式(33)的第一项趋近于 0,第二项是一个正的常数,此时的模型更新将不会带来目标函数值的降低,只能得到

近似的最优解。因此，我们强调必须选择一个递减的学习率才能得到收敛至最优的结果。

3.6 本地迭代次数的选择机制

根据定理 2 中的结论，在适当的参数条件下，我们可以得到式(32)的主导项为：

$$O\left(\frac{\sum_{k=1}^{N}\rho_k\sigma_k^2+(1+\frac{N-K}{KN})E^2G^2+L\Gamma+\frac{L}{\mu}G^2}{(2\mu\beta-1)T}\right) \tag{34}$$

用 T_ε 表示算法为了达到 ε 精度所需要的迭代次数，则 $\frac{T_\varepsilon}{E}$ 为所需的通信次数，我们可以简化为：

$$\frac{T_\varepsilon}{E}\propto(1+\frac{N-K}{KN})EG^2+\frac{\sum_{k=1}^{N}\rho_k\sigma_k^2+L\Gamma+\frac{L}{\mu}G^2}{E} \tag{35}$$

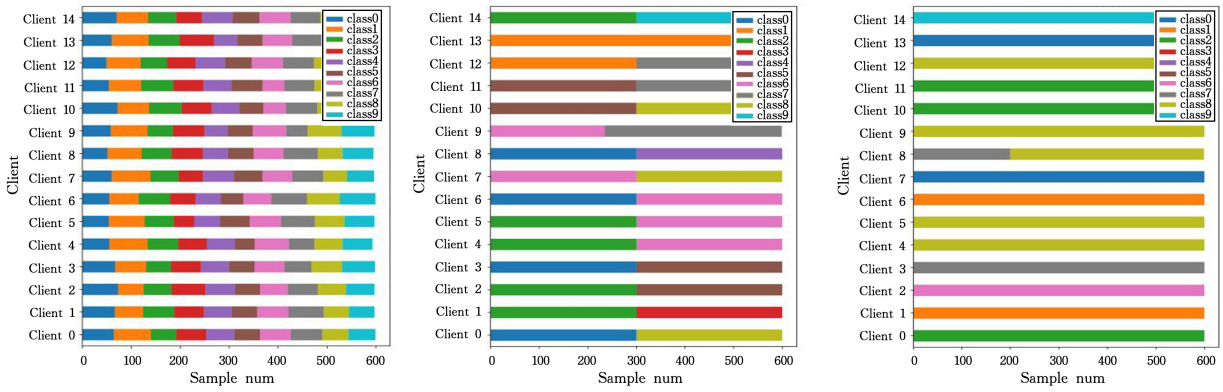
从上式可以看出，本地迭代轮次并不是越大越好，更高的本地迭代轮次可能导致通信次数的增加。事实上，对于不同的参数，存在一个最优本地迭代轮次 E，E 的计算与模型参数以及数据异构程度 Γ 有关，当数据异构程度较低，Γ 接近于 0，式子的后半部分为主导项，E 越大越好；当数据异构程度较大时，Γ 较大，式子前半部分为主导项，E 越小越好。

4 实验

本节将使用不同的数据集和模型来评估 FedSSO 算法，并将其与 FedAvg 算法以及 FedProx 算法进行对比。4.4 节将对客户端本地迭代轮次进行实验分析。

4.1 实验细节

我们使用了 3 个不同的标准数据集进行实验，这些数据集是根据之前联邦学习的相关工作总结出来的基准数据集。在凸问题上，我们使用多层感知机(Multilayer Perceptron，MLP)对比了不同算法在 MNIST 数据集[24]上的表现。为了模拟客户端的异构情况，将数据分布在 100 个客户端之间，每个客户端只含有样本数为 600 的数据；然后选择了一个更复杂的 Cifar-10 数据集，由于数据集中的图片来自于生活中常见的物品，如飞机、车辆等，相比手写字体数据集有更多的误差，因此我们同样将总体数据平均分布于 100 个客户端之中，每个客户端中只含有一个类别的数据。为了探究算法在非凸设置上的表现，使用一个 LSTM 分类器对 Sentiment140 数据集执行推文的文本情绪分析任务，其中每个账号对应一个设备，账号发布的推文为本地数据集。以 MNIST 数据集为例，在不同的异构设置下，取前 15 个客户端，其本地数据分布如图 1 所示。



(a)mnist_iid_data_distribution (b)mnist_noniid_data_distribution (c)mnist_noniid2_data_distribution

图 1 MNIST 数据集在不同异构设置下前 15 个客户端的数据分布图

Fig. 1 First 15 clients' data distribution in MNIST datasets under different data heterogeneity

我们选择 FedAvg 和 FedProx 算法作为 baseline，FedProx 中的参数 mu 按照原文选择 0.2。为了保证每次抽取的样本是总体样本的无偏估计，每轮从总体样本中(FedSSO 算法是从每个簇中抽取)随机不放回地抽取目标数量的样本，这样每个样本在每轮训练中只会出现一次。为了模拟不同程度的数据异构程度，我们采用了多样化的抽样策略。当需要模拟数据独立同分布情况时，对总体数据独立不放回地抽样划入每个客户端，这样每个客户端的数据分布都是总体样本的无偏估计，如图 1(a)所示；当需要模拟数据异构情况时，对总体数据按照标签大小排序，然后划分到不同的切片中，使得每个切片中只含有一类标签的数据，然后将切片数据随机划分到不同客户端中，为了模拟不同的数据异构程度，我们设计了两种异构数据集，即 non-IID 和 non-IID2，如图 1(b)和图 1(c)所示；non-IID 中每个客户端含有两类数据，non-IID2 中每个客户端只含有一类数据，模拟了极端数据异构的情况。

对于每个数据集，我们设置初始学习率为 0.01，在 FedSSO 算法实验中设置了 $\eta_t=0.01/(1+t)$ 的学习率递减机制。每轮选择的客户端数量占客户端总数的比例为 0.1，本地 batch size 为 10。聚类时，OPTICS 聚类方法的参数如下：密度为 2，半径为 0.25。

4.2 模型参数聚类的结果

根据算法 1，在不同数据异构条件下，对模型参数进行聚类后的结果如图 2 所示。

以 MNIST 数据集为例，如图 2(a)所示，在数据集独立同分布的设置下，所有客户端同属一个聚类，各客户端之间本地数据分布具有较高的相似度，可以认为客户端的本地数据分布与数据集总体数据分布相同；在不同的异构设置下(见图 2(b)和图 2(c))，客户端被划分入不同的集合，如图 2(b)所示，在第一类数据异构设置中，总体客户端被分成 29 个聚类集合，各集合中的模型参数相似度较高；图 2(c)为第二类数据异构设置的结果，可以看出，总体客户端被划分成 10 个聚类集合，相比第一类数据异构设置的结果，不同簇之间的模型参数差异度较大。

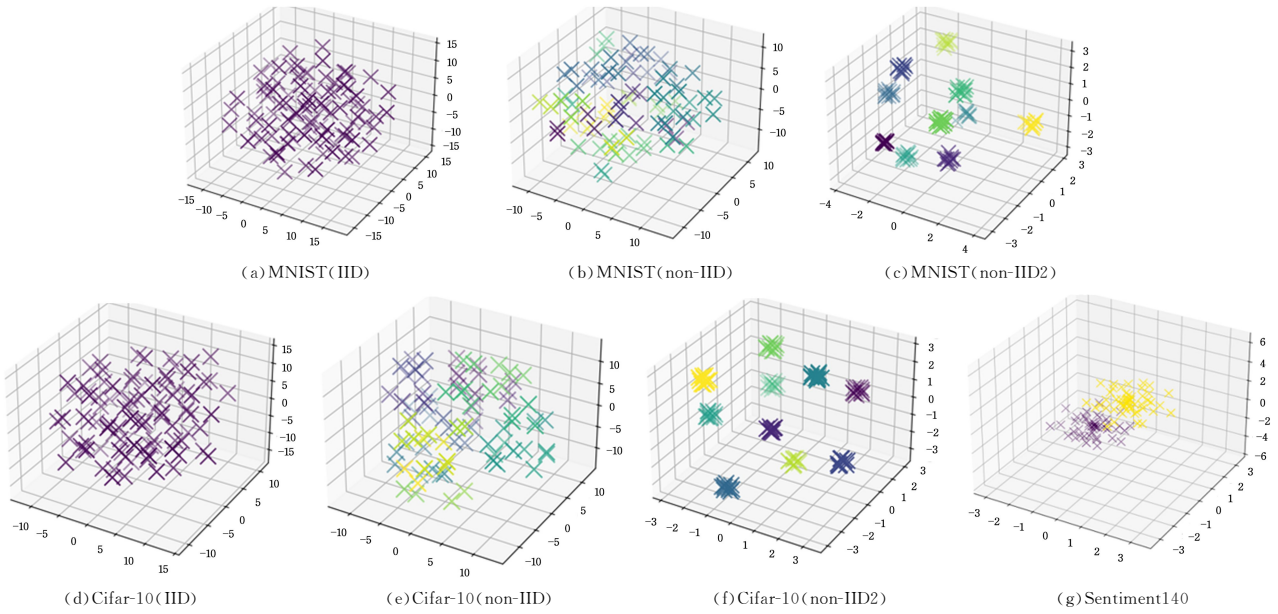


图 2 在 MNIST,Cifar-10 和 Sentiment140 数据集上不同异构设置下的模型参数聚类结果

Fig. 2 Clustering results of model parameters under different heterogeneous setting on MNIST,Cifar-10 and Sentiment140 datasets

4.3 实验结果

我们首先测试了在不同数据分布条件下的实验结果,如图 3 所示。对于 FedAvg 和 FedProx 算法,设置本地迭代轮次 $E=1$;对于 FedSSO 算法,测试了其在本地迭代轮次为 1 和 5 时的不同结果。对于凸问题的实验,我们在不同的数据异构条件下,在 MNIST 和 Cifar-10 数据集上进行了测试;对于非凸问题的实验,我们在 Sentiment140 数据集上进行了测试,由于数据集中每个推特用户为一个客户端,其所发推文为其本地数据,因此我们只考虑异构的单一分布。

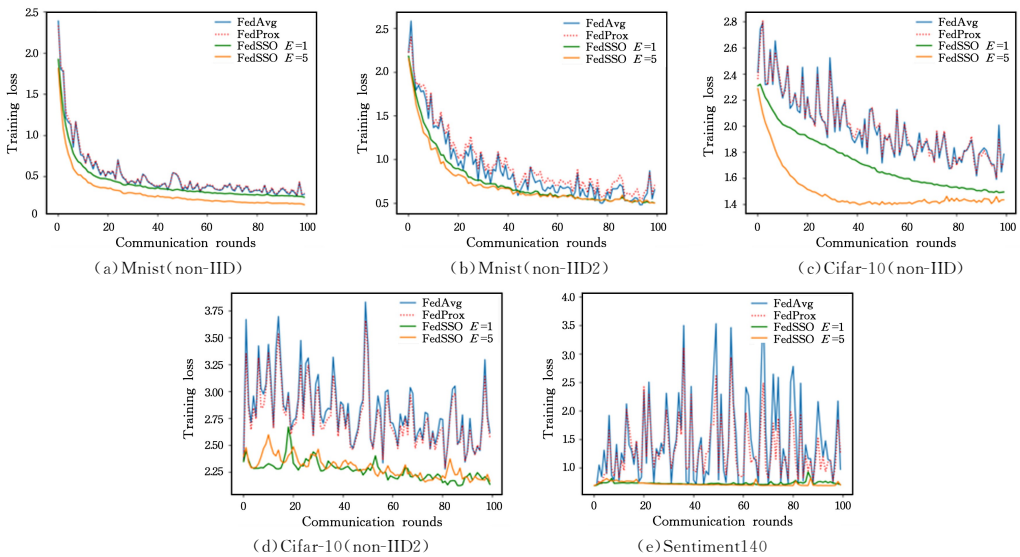


图 3 FedAvg,FedProx,FedSSO 在 MNIST,Cifar-10,Sentiment140 数据集上两种数据分布异构条件下的训练损失

Fig. 3 Training loss of FedAvg,FedProx,FedSSO on MNIST,Cifar-10,Sentiment140 datasets under two different data heterogeneous conditions

可以观察到,FedSSO 算法在 5 种实验设置中都能收敛;相反,FedAvg 和 FedProx 算法的训练曲线存在大幅的波动,特别是当数据分布极端异构时,模型无法收敛,这也验证了定理 2。特别地,不同数据集中数据之间的相似程度不同。在 MNIST 数据集中,由于其数据为数字的灰度图像不同类型数据之间的差异相比 Cifar-10(3 通道的彩色图像记录了生活中常见的物体)较小,因此在数据分布异构的条件下,训练模型的损失曲线波动较小。通过对 FedSSO 算法增加本地迭代轮次的实验,我们可以看出,当数据分布异构程度不高时,增加本地迭代次数可以加速模型的收敛(见图 3(a)和图 3(c))。然而,当数据分布极端不平衡时(见图 3(b)、图 3(d)、图 3(e)),增加本地迭代轮次的作用并不大,这个问题将在下一节详细讨论。

4.4 本地迭代轮次选择机制

本节首先展示了传统联邦学习方法(FedAvg)在不同数据异构设置下,增加本地迭代轮次时的收敛性。图 4 给出了实验结果。如 3.6 节所述,局部迭代轮的选择必须考虑数据异质性的影响。以 MNIST 数据集上的实验为例

(见图 4(a)—图 4(c)),当数据为 IID 时,可以通过增加本地迭代来加快收敛速度;当数据为 non-IID 时,增加局部迭代次数也可以加快收敛速度,但改善程度不如前一种情况;当数据为 non-IID2(数据分布极端异构情况)时,增加客户端本地迭代轮数将减缓训练损失的下降速度,当局部迭代轮数过高时模型无法收敛。同时,我们可以看到,随着数据异构程度的增加,模型的训练损失曲线波动逐渐增大,这验证了第 3.5 节所证明的内容:在异构条件下,设置学习率递减机制对于联邦学习的收敛是必要的。当不同客户端的数据分布不均匀时,增加局部迭代会加深模型参数之间的差异,导致训练损失曲线的大幅波动,此时过高的学习率会导致不同客户端之间的模型差异过大,从而导致无法收敛。因此,对于传统的联邦学习,当数据独立且分布相同时,我们可以通过增加局部迭代次数来加快模型的收敛速度。当数据分布异构时,增加局部迭代次数可能会导致模型的收敛速度减慢或无法收敛。图 5 给出了 FedSSO 在不同数据分布异构的情况下,设置不同本地迭代轮次时的收敛性情况。

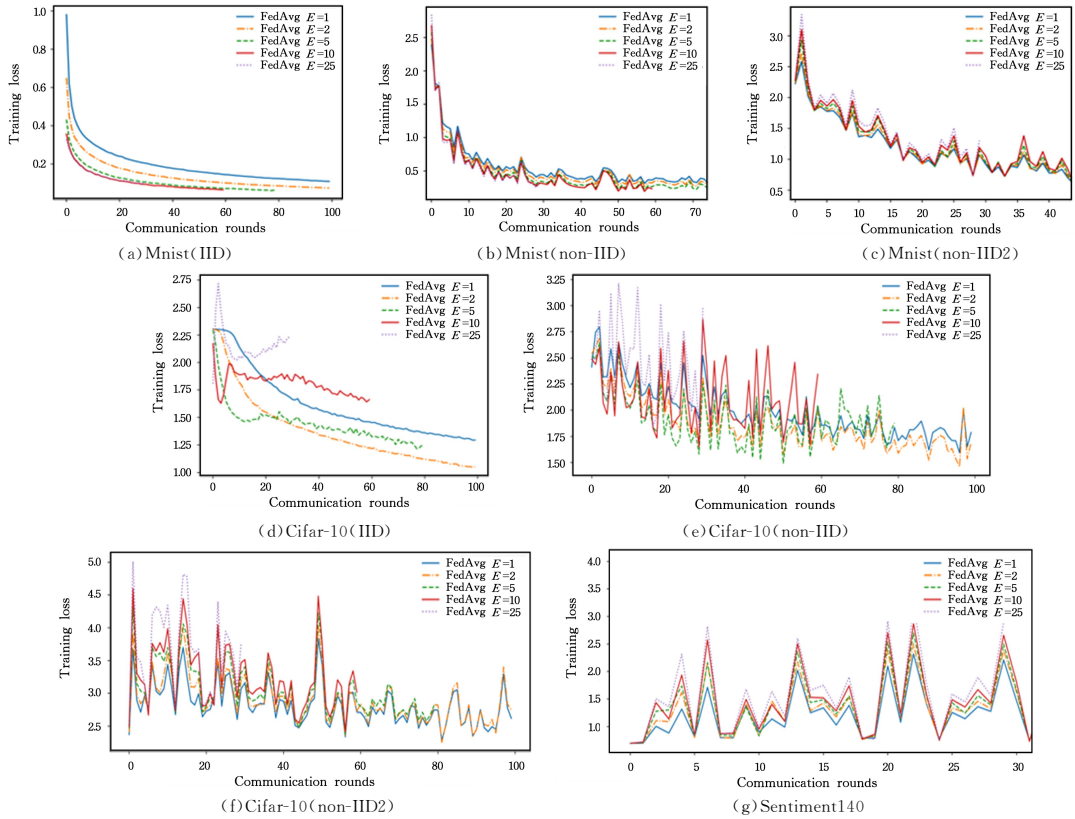


图 4 FedAvg 算法在 MNIST,Cifar-10,Sentiment140 数据集上在不同数据分布异构程度下的训练损失曲线

Fig. 4 Training loss of FedAvg on MNIST, Cifar-10 and Sentiment140 datasets under different heterogeneity of data distribution

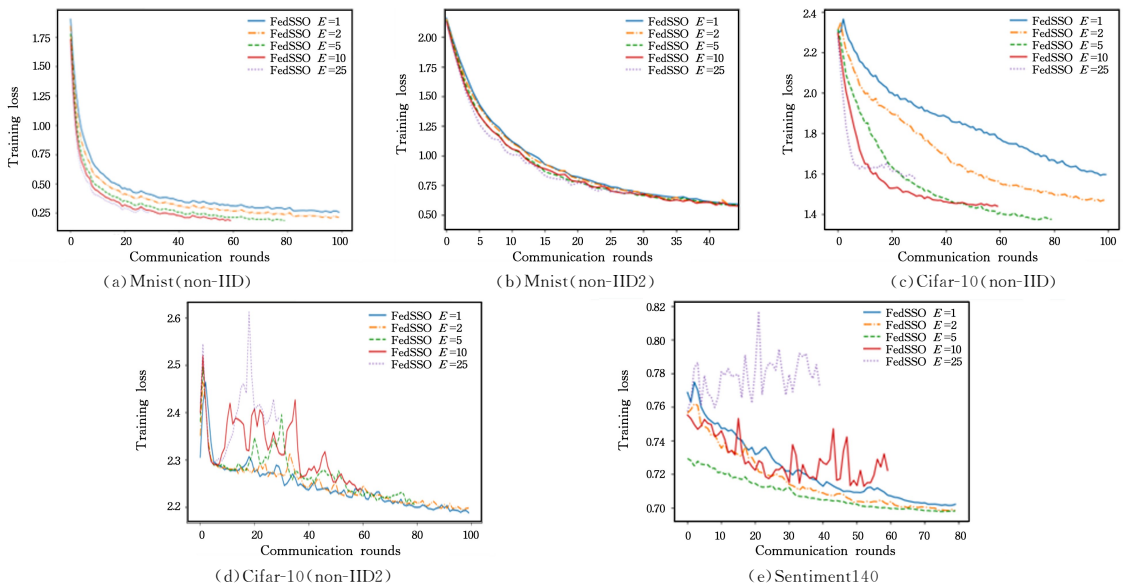


图 5 FedSSO 算法在极端异构条件下设置不同本地迭代轮次时的训练损失曲线

Fig. 5 Training loss of FedSSO for different local iteration rounds under extreme heterogeneous conditions

当数据异构程度较低时(见图 5(a)—图 5(c)),增加本地迭代轮次可以加速模型的收敛速度。然而,随着数据分布异构程度的增加,增加本地迭代轮次对模型收敛速度的加速作用逐渐降低,甚至出现了模型无法收敛的情况(见图 5(d)和图 5(e))。因此,FedSSO 算法可以在客户端异构的情况下,通过增加本地迭代轮次来加速收敛。但在数据极端异构的情况下,FedSSO 算法依然无法在高本地迭代轮次下保证收敛性。因此,在极端异构条件下,选择较低的本地迭代轮次是较好的选择。

结束语 本文提出了一种利用分层抽样优化的联邦学习算法——FedSSO 算法。FedSSO 采用基于密度的聚类方法将异构的客户端划分成不同簇的集合,使每个簇内的客户端具有较高的相似度,在每轮训练时,从所有簇中按比例抽取指定数量的客户端参与训练,保证了所有类型的数据都参与了每轮训练。在标准的联邦学习假设下,提供了 FedSSO 算法的收敛性证明,并且通过标准数据集上的实验验证了所提理论。同时,通过实验证明了 FedSSO 算法在异构数据集上相比 FedAvg 和 FedProx 算法有很大的提升。最后分析了算法的收敛性条件,证明了学习率递减对于模型收敛至关重要,同时,我们发现在异构条件下,提高本地迭代次数会降低模型的收敛速度,为此分析了此现象发生的原因并提出了解决方案,使得在异构条件下也可以通过提高本地迭代的方式加速模型收敛。

参 考 文 献

[1] 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.

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

[3] MCMAHAN H B,RAMAGE D,TALWAR K,et al. Learning Differentially Private Recurrent Language Models [J]. arXiv:1710.06963,2017.

[4] YANG Q,LIU Y,CHEN T,et al. Federated Machine Learning: Concept and Applications[J]. ACM Transactions on Intelligent Systems and Technology,2019,10(2):1-19.

[5] HSIEH K,PHANISHAYEE A,MUTLU O,et al. The Non-IID Data Quagmire of Decentralized Machine Learning [J]. arXiv:1910.00189,2020.

[6] LI T,SAHU A K,ZAHEER M,et al. Federated Optimization in Heterogeneous Networks[J]. arXiv. 1812.06127,2018.

[7] HARD A,RAO K,MATHEWS R,et al. Federated Learning for Mobile Keyboard Prediction [J]. arXiv. 1811.03604,2018.

[8] YANG T,ANDREW G,EICHNER H,et al. Applied Federated Learning:Improving Google Keyboard Query Suggestions [J]. arXiv. 1812.02903,2018.

[9] BOYD S,PARIKH N,CHU E,et al. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers[J]. Foundations & Trends in Machine Learning,2010,3(1):1-122.

[10] DEKEL O,GILAD-BACHRACH R,SHAMIR O,et al. Optimal Distributed Online Prediction Using Mini-Batches [J]. Journal of Machine Learning Research,2012,13(1):165-202.

[11] RICHTÁRIK P,TAKÁČ M. Distributed Coordinate Descent Method for Learning with Big Data [J]. Journal of Machine Learning Research,2016,17(75):1-25.

[12] ZHANG S,CHOROMANSKA A,LECUN Y. Deep learning with Elastic Averaging SGD [J]. arXiv. 1412.6651,2014.

[13] BONAWITZ K,IVANOV V,KREUTER B,et al. Practical Secure Aggregation for Privacy-Preserving Machine Learning [C]// The 2017 ACM SIGSAC Conference. ACM,2017:1175-1191.

[14] BONAWITZ K,EICHNER H,GRIESKAMP W,et al. Towards Federated Learning at Scale: System Design [J]. arXiv. 1902.01046,2019.

[15] MOHRI M,SIVEK G,SURESH A T. Agnostic Federated Learning [C]// International Conference on Machine Learning. PMLR,2019.

[16] HU H,WANG D,WU C. Distributed Machine Learning through Heterogeneous Edge Systems[C]// AAAI Conference on Artificial Intelligence. 2020:7179-7186.

[17] PETER K,BRENDAN H,MCMAHAN H B,et al. Advances and Open Problems in Federated Learning [J]. arXiv. 1912.04977,2019.

[18] ZHAO Y,LI M,LAI L,et al. Federated Learning with Non-IID Data [J]. arXiv. 1806.00582,2018.

[19] GHOSH A,CHUNG J,DONG Y,et al. An Efficient Framework for Clustered Federated Learning [J]. arXiv:2006.04088,2020.

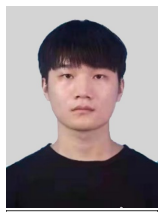
[20] SATTLER F,KR MÜLLER,SAMEK W. Clustered Federated Learning: Model-Agnostic Distributed Multitask Optimization Under Privacy Constraints [J]. IEEE Trans Neural Netw Learn Syst,2021,32(8):3710-3722.

[21] YAN Y,NIU C,DING Y,et al. Distributed Non-Convex Optimization with Sublinear Speedup under Intermittent Client Availability [J]. arXiv. 2002.07399,2020.

[22] ANKERST M,BREUNIG M M,KRIEGEL H P,et al. OPTICS:ordering points to identify the clustering structure [J]. SIGMOD Record:Special Interest Group on Management Data,1999,28(2):49-60.

[23] LI X,HUANG K,YANG W,et al. On the Convergence of FedAvg on Non-IID Data [J]. arXiv:1907.02189,2020.

[24] LECUN Y,BOTTOU L. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE,1998,86(11):2278-2324.



LU Chen-yang, born in 1997, postgraduate. His main research interests include federated learning and machine learning.

MA Wu-bin, born in 1986, Ph.D, associate research fellow. His main research interests include data engineering and cyber-physical systems.

(责任编辑:何杨)