Modeling and Simulation 建模与仿真, 2025, 14(4), 665-671
Published Online April 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.144319

Hans汉斯

基于梯度范数感知最小化的去中心化联邦学习算法

方诚意，胡建华，黄佳龙

上海理工大学理学院，上海

收稿日期：2025年3月21日；录用日期：2025年4月14日；发布日期：2025年4月21日

摘　要

去中心化联邦学习是通过一组设备执行隐私保护的分布式学习，它有效降低了中心化联邦学习的通信成本和信息泄露风险。然而，设备之间的非独立同分布数据会影响模型效果。为了解决这个问题，几乎所有的算法都利用经验风险最小化作为局部优化器，但这很容易造成客户端本地训练过拟合，造成算法全局模型的泛化能力下降。本文利用梯度范数感知最小化，提出基于梯度范数感知最小化的去中心化联邦学习算法，使全局模型损失函数的表面更加平滑，提升模型的泛化能力。

关键词

去中心化联邦学习，分布式算法，非独立同分布数据

A Decentralized Federated Learning Algorithm Based on Gradient Norm Aware Minimization

Chengyi Fang, Jianhua Hu, Jialong Huang

College of Science, University of Shanghai for Science and Technology, Shanghai

Received: Mar. 21st, 2025; accepted: Apr. 14th, 2025; published: Apr. 21st, 2025

Abstract

Decentralized Federated Learning performs privacy-preserving distributed learning across a group of devices, reducing the communication costs and information leakage risks associated with centralized federated learning. However, the non-independent and identically distributed (Non-IID)

文章引用: 方诚意, 胡建华, 黄佳龙. 基于梯度范数感知最小化的去中心化联邦学习算法[J]. 建模与仿真, 2025, 14(4): 665-671. DOI: 10.12677/mos.2025.144319

data among devices can negatively impact the model's performance. To address this issue, most algorithms adopt empirical risk minimization as the local optimizer, which often leads to overfitting during local client training and results in decreased generalization ability of the global model. This paper proposes a Decentralized Federated Learning Algorithm based on Gradient Norm-Aware Minimization, which smooths the loss surface of the global model and enhances its generalization performance.

Keywords

Decentralized Federated Learning, Distributed Algorithms, Non-IID Data

1. 引言

在数字时代，电子设备数量的激增产生了大量不同种类的数据。传统的机器学习和深度学习算法在处理这些数据时严重依赖集中式数据收集和处理，面临着严重的通信压力和隐私泄露风险。

联邦学习(FL) [1]为传统算法的局限性提供了一种解决方案，能够在多个设备上同时训练本地模型，每个设备根据其本地数据生成一个模型。原始数据既不发送也不可见，避免了数据传输带来的数据隐私泄露风险并且降低通讯压力。目前大多数 FL 算法都基于中心化联邦学习算法(CFL)，由一个中央服务器从其他设备接收每轮设备本地训练完成后的模型并执行聚合，再将聚合后的模型传输给设备以确保有效整合来自每个设备的训练结果并获得全局模型。但中央服务器的存在依旧使得对服务器的攻击能窃取全局模型导致隐私泄露，并且服务器的通信压力仍然存在。

去中心化联邦学习(DFL) [2]，去除了中央服务器，设备在没有任何服务器的情况下相互通信共享模型参数。DFL 由于没有了固定的中央服务器，设备之间的通信网络更加灵活和多样化，并且进一步降低了隐私泄露风险、节省通信成本。然而，由于设备之间的非独立同分布数据和网络结构引起的模型聚合局部性，使得 DFL 在设备局部模型之间存在严重的不一致性，这种不一致性可能会导致局部模型严重过拟合[3]，影响模型最终效果。

为了解决数据异构问题，文献[4]提出了一个概率驱动的八卦框架，以揭示非相邻客户端之间的相似关系，并指导相似客户端之间的聚合。AsyDFL [5]引入了邻居选择和梯度推送，要求每个边缘节点仅向邻居的子集传输梯度，以提高资源效率。这些工作主要集中在补充本地信息和识别网络拓扑中的最佳聚合关系上。DFedSAM [6]通过锐利度感知最小化(SAM)生成局部平坦模型，缓解设备局部模型过拟合问题。但 SAM 只考虑了零阶平坦度，无法区分在给定扰动范围内的低泛化误差最小值和高泛化误差最小值。为了解决这一问题，本文在 DFL 中引入梯度范数感知最小化(GAM) [7]，关注扰动半径内的最大梯度范数，避免损失函数发生剧烈变化，提升设备局部模型的泛化能力，从而优化模型总体性能。本文的主要贡献有以下两点：

1) 提出了基于梯度范数感知最小化的去中心化联邦学习算法 DFedGAM，通过将梯度范数感知最小化引入 DFL，有效改善了设备过拟合的问题，提升了本地模型的泛化能力，从而提升了模型整体性能。

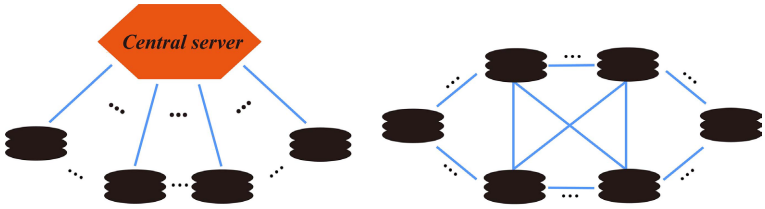
2) 通过实验证明了提出的算法与 7 个基线方法相比的优越性，包括 DFL 和 CFL 方法。

2. 去中心化联邦学习介绍

2.1. 去中心化联邦学习网络结构

目前大多数联邦学习(FL)模型都是基于中心化联邦学习(CFL)的。CFL 能够在多个设备或节点上同时训练模型，通过在每个设备上进行多次本地计算，由单个设备作为中央服务器从其他设备接受模型并执行聚合，如图 1(a)所示。但中央服务器带来了潜在的问题，包括单点故障导致的整体网络瘫痪，中央服务器的通信瓶颈和信息泄露风险。

去中心化联邦学习(DFL)的出现克服了这些局限性，DFL 通过让多个设备的模型聚合代替中央服务器聚合所有设备的模型来实现去中心化，大大减少了对单个中央服务器的依赖，如图 1(b)所示。设备之间的分散网络拓扑结构可以被建模为无向连通图 $G=(N,V,W)$，其中 $N=\{1,2,\cdots,m\}$ 表示设备集，$V\in N\times N$ 表示通讯通道集，每个通道连接两个不同的设备。



(a) 中心化联邦学习框架　　(b) 去中心化联邦学习框架

Figure 1. Illustrations of CFL (a) and DFL (b)
图 1. CFL (a)和 DFL (b)网络结构图

2.2. 去中心化联邦学习模型

在去中心化联邦学习中，目标函数是以下有限和随机非凸最小化问题：

$$\min_{x\in R^d} f(x)=\frac{1}{m}\sum_{i=1}^{m} f_i(x),\ f_i(x)=E_{\xi\sim D_i}F_i(x;\xi) \tag{1}$$

其中 D_i 表示第 i 个设备中的数据分布，m 是设备的数量，x 为参数，$F_i(x;\xi)$ 是与设备本地数据样本相关的局部目标函数。

在模型训练过程中，每个设备首先初始化本地模型，然后根据本地数据执行多步模型训练，第 t 轮设备 i 的第 k 次本地迭代可以表示为：

$$y^{t,k+1}(i)=y^{t,k}(i)-\eta g^{t,k}(i) \tag{2}$$

其中 $E\left[g^{t,k}(i)\right]=\nabla F_i\left(y^{t,k};\xi\right)$，$\eta$ 是学习率。经过 K 次本地迭代后，每个设备中的参数更新为 $z^t(i)=y^{t,K}(i)$，并将参数发送给其邻居。然后每个设备对本地和接收到的参数进行聚合，具体更新公式为：

$$x^{t+1}(i)=\sum_{l\in N(i)} w_{i,l}z^t(i) \tag{3}$$

参数聚合完成后，每个设备使用聚合后的参数作为初始参数开始下一轮本地更新。

3. 基于梯度范数感知最小化的去中心化联邦学习算法

相比于 CFL，DFL 在聚合模型时没有中央服务器聚合全局参数，只能聚合邻居的局部参数，这使得每个设备在本地迭代完成后不能获得全局信息，只能获得局部信息。又由于 FL 设备数据之间的非独立同

分布特性，加剧了本地多次更新带来的设备模型漂移问题[8]。训练得到的设备模型之间产生的巨大差异又会导致设备本地训练过拟合，破坏本地模型的泛化能力，导致全局模型性能下降。

具体而言，我们要解决 DFL 的如下两个问题：

1) 非独立同分布数据和本地多次更新带来的设备模型漂移问题，这会导致本地模型偏离全局最优点，影响模型聚合结果。

2) 失去中央服务器后不能聚合全局模型的问题。模型局部聚合会加剧模型之间的差异，导致设备本地模型更新过拟合，降低泛化能力，损坏全局模型预测精度。

3.1. 梯度范数感知最小化

在锐利度感知最小化(SAM)中，通过对损失函数添加一个小的扰动计算优化点周围的零阶平坦度，测量最大损耗值和当前点之间的差值。

目标函数 $L(x)$ 在 x 处的零阶平坦度定义为：

$$R_{\rho}^{(0)}(x)=\max_{x'\in(x,\rho)}\left(L(x')-L(x)\right) \tag{4}$$

其中 ρ 是控制扰动范围大小的扰动半径。拥有零阶平坦度的 SAM 损失函数为：

$$L^{sam}(x)=L(x)+R_{\rho}^{(0)} \tag{5}$$

零阶平坦度通过平滑 x 附近的损失函数景观，提升了本地模型的泛化能力。但[7]发现仅仅通过比较扰动范围 ρ 内损失函数的数值大小选择更新方向并不是一直有效的。当 ρ 覆盖多个最小值时，SAM 无法测量损失函数波动频率，x 附近的损失函数可能波动变化很快，但数值差异较小，SAM 可能会把这样具有较差泛化能力的点作为更新方向。当 ρ 内只有一个最小值点时，SAM 也有可能因为观测半径是有限的而在最大损失与损失的上升趋势不一致时产生误判。为了解决这一问题，我们在 DFL 中引入了一阶平坦度，可以表示最小值附近的最大梯度范数。

目标函数 $L(x)$ 在 x 处的一阶梯度范数定义为：

引理 3.1 (ρ 一阶平坦度) [7]。对于任意的 $\rho>0$，损失函数 $L(x)$ 在 x 处的 ρ 一阶平坦度定义为：

$$R_{\rho}^{(1)}(x)=\rho\cdot\max_{x'\in(x,\rho)}\left\|\nabla L(x')\right\|_2$$

其中 ρ 是控制扰动范围大小的扰动半径，$\|\cdot\|_2$ 是 l_2 -范数。

一阶平坦度意味着损失函数 $L(x)$ 在 x 附近不会发生剧烈变化，从而平滑了 DFL 设备模型损失函数的景观，提升了泛化能力。

3.2. 基于梯度范数感知最小化的去中心化联邦学习算法模型

为了提升 DFL 设备本地模型的泛化能力，本文将 GAM 引入 DFL 模型中，提出基于梯度范数感知最小化的去中心化联邦学习算法。将损失函数定义为：

$$f_i(x)=E_{\xi\in D_i}\max_{\|\delta_i\|_2\le\rho}F_i\left(y^{t,k}(i)+\delta_i;\xi_i\right) \tag{6}$$

其中 $y^{t,k}(i)+\delta_i$ 是扰动后的模型参数。

DFL 在引入 GAM 后可以更准确地找到位于平坦区域的最小值。在 ρ 覆盖多个最小值时，如果范围内局部最小值数量变大，一阶平坦度的最大梯度范数会增加，这表明在覆盖多个最小值时一阶平坦度可以表示锐度。当 ρ 只覆盖一个最小值点时，零阶平坦度在观测半径内不足以表示最大损失趋势，但一阶平坦度可以帮助了解损失趋势的信息。因此在许多零阶平坦度无法表示损失大小的情况下，一阶平坦度

仍然具有辨别性。下面给出模型的更新公式。

第 t 轮设备 i 的第 k 次本地迭代可以表示为：

$$y^{t,k+1}(i)=y^{t,k}(i)-\eta\tilde{g}^{t,k}(i) \tag{7}$$

其中 $\tilde{g}^{t,k}(i)=\nabla F_i\left(y^{t,k}+\delta\left(y^{t,k}\right);\xi\right)$。

扰动项 $\delta\left(y^{t,k}\right)$ 可以表示为：

$$\delta\left(y^{t,k}\right)=\rho\cdot\frac{\nabla\left\|g^{t,k}(i)\right\|_2}{\left\|\nabla\left\|g^{t,k}(i)\right\|_2\right\|_2} \tag{8}$$

其中 $g^{t,k}(i)=\nabla F_i\left(y^{t,k};\xi_i\right)$。

本地更新 K 次后和邻居聚合。我们将提出的基于梯度范数感知最小化的去中心化联邦学习算法称为 DFedGAM，具体模型如算法 1 所示。

算法 1. 基于梯度范数感知最小化的去中心化联邦学习算法

输入：总通信轮数 T，本地更新次数 K，客户端总数 m，学习率 η

输出：所有客户端通讯后的共识模型 x^T

初始化：随机初始化每个设备模型 $x^0(i)$

for $t=1,\cdots,T$ **do**

  for $i=1,\cdots,m$ **do**

    for $k=1,\cdots,K$ **do**

      令 $y^{t,0}(i)=x^t(i)$，$y^{t,-1}(i)=y^{t,0}(i)$

      对本地数据采样，计算梯度 $g^{t,k}(i)=\nabla F_i\left(y^{t,k};\xi_i\right)$

      $\delta\left(y^{t,k}\right)=\rho\cdot\frac{\nabla\left\|g^{t,k}(i)\right\|_2}{\left\|\nabla\left\|g^{t,k}(i)\right\|_2\right\|_2}$

      $\tilde{g}^{t,k}(i)=\nabla F_i\left(y^{t,k}+\delta\left(y^{t,k}\right);\xi\right)$

      $y^{t,k+1}(i)=y^{t,k}(i)-\eta\tilde{g}^{t,k}(i)$

    end

    $z^t(i)=y^{t,K}(i)$

    每个设备接收邻居的模型并进行聚合

    $x^{t+1}(i)=\sum_{l\in N(i)}w_{i,l}z^t(i)$

  end

end

4. 实验

本节将新算法与基于 CFL 和 DFL 的七个基线进行比较来评估其优越性，其中 FedAvg [1]、FedSAM [9]和 SCAFFOLD [8]是基于 CFL 的算法，D-PSGD [10]、DFedAvg [3]、DFedAvgM [3]和 DFedSAM [6]是基于 DFL 的算法。实验在数据集 CIFAR-10 上进行，分独立同分布(IID)和非独立同分布(non-IID)两种情况，非独立同分布(non-IID)的情况下采用狄利克雷(Dirichlet)数据 Dir-0.3 和 Dir-0.6 两种分布。

4.1. 实验设置

客户端总数设置为 100，其中 10%的客户端参与通信。具体来说，所有客户端都对去中心化方法执

行本地迭代步骤，只有参与的客户端才能对集中式方法执行本地更新。对于所有实验，学习率η初始化为0.1，每轮通信的衰减率为0.995。最优扰动范围大小ρ在范围$\{0.05,0.1,0.2,0.5,1.0,2.0\}$内搜索。每个设备采用LeNet作为模型。在通信设置方面，为了确保与FL的公平比较，我们使用了随机拓扑，并确保每个客户端的邻居数量不超过CFL中服务器的连接数量。

4.2. 实验结果分析

从表1中可以清楚地看出，在CIFAR-10数据集的三种数据分布情况下，我们提出的DFedGAM均表现出优于基线算法的测试精度。具体而言，在通信轮次相同的条件下，本算法在关键性能指标上，显著超过了其他中心化和去中心化联邦学习框架。这一结果表明，本算法在充分利用去中心化网络拓扑的情况下，能够更高效地实现模型参数的聚合与优化，减小了由于非独立同分布数据和失去中心节点导致的性能下降。

在实验中，与DFedSAM算法相比，我们提出的基于梯度范数感知最小化的去中心化联邦学习算法在IID数据下精度提升1.30%，在Dir-0.3和Dir-0.6数据下分别提升1.19%和1.46%。这一提升可以归因于我们提出的算法更准确地计算了损失函数变化的趋势，提升了本地模型的泛化能力。

Table 1. Dirichlet data algorithm accuracy
表1. Dirichlet数据算法精度

算法	CIFAR-10		
	Dir-0.3	Dir-0.6	IID
FedAvg	78.01	78.92	80.14
FedSAM	80.22	81.35	82.79
SCAFFOLD	77.91	79.83	81.60
D-PSGD	59.56	60.21	63.05
DFedAvg	76.82	77.98	80.31
DFedAvgM	79.27	80.59	82.32
DFedSAM	79.65	80.17	81.40
DFedGAM	**80.60**	**81.34**	**82.46**

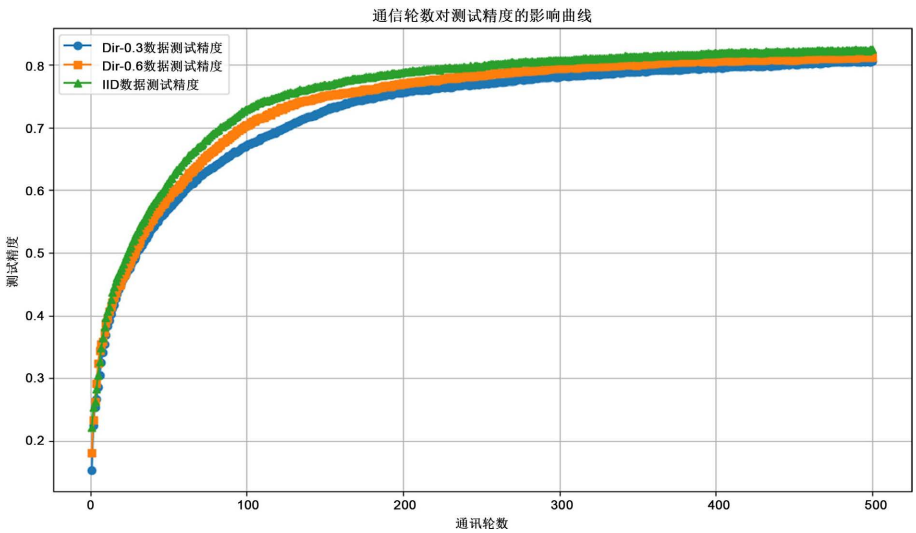


Figure 2. The curve of the impact of communication rounds on testing accuracy
图2. 通信轮数对测试精度影响曲线图

由图 2 中三种数据下通信轮数和测试精度的关系曲线可以看出，我们提出的算法测试精度随着通信轮数的增加而提高，并且有着较快的收敛速度。在本地数据被设置为不同异质性水平的情况下，我们的算法在每种数据情况下都具有鲁棒性。

5. 总结

本文提出了一种基于梯度范数感知最小化的去中心化联邦学习算法，通过在去中心化联邦学习中引入梯度范数感知最小化技术，缓解非独立同分布数据造成的本地模型过拟合问题，提高了去中心化联邦学习模型性能。在数据集三种程度的异质设置下，模型均表现出优于其他模型的性能，体现出对非独立同分布数据较强的鲁棒性和有效性。但对于未来的研究，理论上解释更强的平坦度是否更适合提升泛化能力至关重要，本文缺少对一阶平坦度有效性的理论分析。非独立同分布数据以及本地多步更新带来的设备间模型差异是去中心化联邦学习亟待解决的问题，本文提出了一个较为有效的方法。未来还应继续研究如何提高去中心化联邦学习的性能。

基金项目

本文由国家自然科学基金项目(62073223, 12371308)、上海自然科学基金项目(22ZR1443400)资助。

参考文献

[1] McMahan, B., Moore, E., Ramage, D., Hampson, S. and Arcas, B.A. (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of the* 20*th International Conference on Artificial Intelligence and Statistics* (*AISTATS*) 2017, Fort Lauderdale, 20-22 April 2017, 1273-1282.

[2] Lalitha, A., Shekhar, S., Javidi, T. And Koushanfar, F. (2018) Fully Decentralized Federated Learning. *Third Workshop on Bayesian Deep Learning* (*NeurIPS*), Montréal. https://bayesiandeeplearning.org/2018/papers/140.pdf

[3] Sun, T., Li, D. and Wang, B. (2023) Decentralized Federated Averaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 4289-4301. https://doi.org/10.1109/tpami.2022.3196503

[4] Cai, X., Yu, N., Zhao, M., Cao, M., Zhang, T. and Lu, J. (2024) Decentralized Federated Learning in Partially Connected Networks with Non-IID Data. 2024 *Design, Automation & Test in Europe Conference & Exhibition* (*DATE*), Valencia, 25-27 March 2024, 1-6. https://doi.org/10.23919/date58400.2024.10546508

[5] Liao, Y., Xu, Y., Xu, H., Chen, M., Wang, L. and Qiao, C. (2024) Asynchronous Decentralized Federated Learning for Heterogeneous Devices. *IEEE/ACM Transactions on Networking*, **32**, 4535-4550. https://doi.org/10.1109/tnet.2024.3424444

[6] Shi, Y., Shen, L., Wei, K., Sun, Y., Yuan, B., Wang, X. and Tao, D. (2023) Improving the Model Consistency of Decentralized Federated Learning. *International Conference on Machine Learning*, Honolulu, 23-29 July 2023, 31269-31291.

[7] Zhang, X., Xu, R., Yu, H., Zou, H. and Cui, P. (2023) Gradient Norm Aware Minimization Seeks First-Order Flatness and Improves Generalization. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 20247-20257. https://doi.org/10.1109/cvpr52729.2023.01939

[8] Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S. and Suresh, A.T. (2020) Scaffold: Stochastic Controlled Averaging for Federated Learning. 2020 *International Conference on Machine Learning*, 13-18 July 2020, 5132-5143.

[9] Qu, Z., Li, X., Duan, R., Liu, Y., Tang, B. and Lu, Z. (2022) Generalized Federated Learning via Sharpness Aware Minimization. 2022 *International Conference on Machine Learning*, Baltimore, 17-23 July 2022, 18250-18280.

[10] Lian, X., Zhang, C., Zhang, H., Hsieh, C. J., Zhang, W. and Liu, J. (2017) Can Decentralized Algorithms Outperform Centralized Algorithms? A Case Study for Decentralized Parallel Stochastic Gradient Descent. *Advances in Neural Information Processing Systems*, 2017, Long Beach, 4-9 December 2017, 5330-5340.