计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2026,35(2):141−153 [doi: 10.15888/j.cnki.csa.010083] [CSTR: 32024.14.csa.010083]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于熵正则化近端策略优化的联邦客户端选择①

陈雨彤[1], 金子龙[2]

[1](南京信息工程大学 软件学院, 南京 210044)
[2](浙江理工大学 信息科学与工程学院 (网络空间安全学院), 杭州 310018)
通信作者: 金子龙, E-mail: zljin@outlook.com

摘　要: 近年来, 联邦学习 (federated learning, FL) 作为一种分布式机器学习范式, 因其能够在保护数据隐私的同时实现模型训练, 已在智能医疗、金融服务、物联网以及车联网等领域得到广泛应用. 在车联网 (IoV) 环境中, 由于节点高度动态和车辆资源的异构性, 并非所有客户端都适合参与联邦训练, 因此高效且鲁棒的客户端选择策略对于模型性能与系统效率至关重要. 然而, 传统 FL 方法大多依赖静态或启发式的客户端选择机制, 难以适应 IoV 场景中频繁变化的环境状态与客户端特性. 为此, 本文提出一种基于熵正则化近端策略优化 (entropy regularization proximal policy optimization, ERPPO) 的动态客户端选择方法, 并结合置信度加权聚合策略. 该方法通过在近端策略优化 (proximal policy optimization, PPO) 目标函数中引入策略熵正则项, 增强客户端选择策略的探索性, 以避免陷入局部最优. 同时, 置信度聚合机制基于客户端模型更新方差自适应调整聚合权重, 提升全局模型的收敛稳定性与鲁棒性. 实验结果表明, 所提方法在保障模型精度的前提下, 有效降低了通信开销, 并在动态环境下展现出优于传统方法的综合性能.

关键词: 联邦学习; 车联网; 客户端选择; ERPPO

引用格式: 陈雨彤,金子龙.基于熵正则化近端策略优化的联邦客户端选择.计算机系统应用,2026,35(2):141–153. http://www.c-s-a.org.cn/1003-3254/10083.html

Entropy Regularization Proximal Policy Optimization for Federated Client Selection

CHEN Yu-Tong[1], JIN Zi-Long[2]

[1](School of Software, Nanjing University of Information Science & Technology, Nanjing 210044, China)
[2](School of Information Science and Engineering (School of Cyber Science and Technology), Zhejiang Sci-tech University, Hangzhou 310018, China)

Abstract: In recent years, federated learning (FL) has emerged as a distributed machine learning paradigm that enables model training while preserving data privacy. It has been widely applied in domains such as smart healthcare, financial services, the Internet of Things (IoT), and the Internet of Vehicles (IoV). However, due to the highly dynamic nature of IoV environments and the heterogeneous computing resources among vehicles, not all clients are suitable for participation in federated training. Therefore, designing an efficient and robust client selection strategy is critical for ensuring model performance and system efficiency. Traditional FL methods often rely on static or heuristic client selection mechanisms, which fail to adapt to the frequently changing states and characteristics of clients in IoV scenarios. To address this issue, this study proposes a dynamic client selection approach based on entropy regularization proximal policy optimization (ERPPO), integrated with a confidence-weighted aggregation mechanism. By incorporating a policy entropy regularization term into the PPO objective function, the proposed method enhances the exploration capability of the client selection policy, thus mitigating the risk of local optima. Furthermore, the confidence-based aggregation strategy

① 基金项目: 国家自然科学基金面上项目 (62271264); 浙江省“尖兵领雁+X”重大科技计划 (2025C02033)
收稿时间: 2025-08-05; 修改时间: 2025-09-16; 采用时间: 2025-09-26; csa 在线出版时间: 2025-12-26
CNKI 网络首发时间: 2025-12-29

adaptively adjusts the aggregation weights based on the variance of local model updates, which enhances the convergence stability and robustness of the global model. Experimental results demonstrate that the proposed ERPPO framework not only reduces communication overhead but also achieves superior overall performance in dynamic environments while maintaining high model precision.

Key words: federated learning (FL); Internet of Vehicles (IoV); client selection; entropy regularization proximal policy optimization (ERPPO)

随着 5G-Advanced 与即将到来的 6G 技术的逐步普及, 全球数十亿物联网 (IoT) 设备正加速部署, 信息流通与数据交互进入爆发式增长阶段[1]. 在这一趋势下, IoT 设备不仅成为数据产生的重要源头, 还可能存储包含用户隐私的敏感信息. 传统的集中式云计算架构在面对如此庞大的数据体量时, 逐渐暴露出通信带宽瓶颈、存储压力及运营成本高昂等问题[2].

联邦学习 (federated learning, FL) 作为一种分布式机器学习框架, 因其在本地完成模型训练、仅共享模型参数而无需上传原始数据, 成为解决上述挑战的有效途径[3]. 该方法不仅显著降低了通信负载, 提升了数据隐私保护能力, 还为资源受限设备参与协同训练提供了可能, 极大促进了全局模型的泛化能力[4]. 目前, FL 已在智慧医疗[5]、入侵检测[6]、智能驾驶及推荐系统[7,8]等隐私敏感、数据分散的场景中展现出广阔的应用前景.

车联网 (Internet of Vehicles, IoV) 作为物联网 (IoT) 的典型应用场景之一, 通过实现车辆与基础设施、行人及云端的实时交互, 为智能交通系统的发展奠定了坚实基础[9]. 在车联网环境中, 车辆会持续产生海量高维的动态数据, 包括驾驶行为、道路状况和环境感知信息. 若将这些数据直接上传至云端进行集中处理, 将面临高昂的通信成本及潜在的数据泄露风险. FL 作为一种去中心化的分布式学习范式, 为车联网中的智能决策提供了隐私友好的解决方案[10]. 然而, 由于车辆节点具有高移动性, 且不同车辆在计算、存储和通信等资源上存在异构性, 让所有客户端参与联邦学习训练在实际应用中存在较大困难, 传统联邦学习方法面临适应性不足的挑战. 针对上述问题, 设计高效且灵活的客户端选择策略已成为提升车联网联邦学习性能的关键.

现有的客户端选择方法大多采用静态策略. Cho 等人[11]采用有偏的客户端选择策略, 偏向于选择局部损失较高的客户端, 提升全局模型在不同数据分布下的泛化性, 但可能导致训练过程偏向特定数据分布并增加训练延迟. 由于异构客户端的硬件能力、网络条件和计算资源的差异. 过多的训练失败将会降低整体训练效率, 而过度选择稳定性更高的客户端, 则可能陷入局部最优. 为了解决这一问题并保持训练过程的高效性与公平性, Huang 等人[12]在考虑客户端有效参与度和公平性的前提下, 提出了 E3CS. 此方法旨在平衡稳定性与多样性, 从而提升联邦学习系统的整体性能. 此外, Li 等人[13]针对客户端选择过程中选定参与者与未选定参与者之间的差异性, 提出了基于细粒度的客户端选择策略 PyramidFL. 该方法能够充分挖掘选定客户端内的数据分布特性与系统异构性, 从而优先选择那些在统计效用和系统效用上具有更高价值的客户端, 以提升训练效率和模型性能.

人工智能 (AI) 的快速发展为 IoV 提供了更加智能化的服务能力, 使其能够在动态复杂的交通环境中实现高效的决策与资源管理. 近年来, 深度强化学习 (deep reinforcement learning, DRL) 方法逐渐应用于车联网场景, 主要集中在资源分配和任务卸载等领域. 例如, Hazarika 等人[14]激励车辆共享其空闲计算资源, 设计了一种基于软演员-评论家 (soft actor-critic, SAC) 的深度强化学习算法, 根据每个任务的优先级和计算需求对任务进行分类, 以实现功率的最优分配. Abishu 等人[15]则将数字孪生 (digital twin)、区块链和联邦多智能体深度强化学习 (FMADRL) 技术相结合, 提出了自适应资源分配与任务卸载方案. 具体而言, 该方法将多目标优化的资源分配与任务卸载问题建模为马尔可夫决策过程的多智能体扩展, 并采用基于多智能体深度确定性策略梯度 (FMADDPG) 算法进行求解. Quan 等人[16]提出一种基于混合多智能体深度强化学习算法 (HMADRL) 的自适应联合优化方案, 用于计算卸载和资源分配策略. 此外, 设计了一种集中式计算卸载和分布式资源分配框架以减少多个智能体之间的通信开销, 该方法提

高了系统在任务完成率、服务延迟和能耗方面的性能. Song 等人[17]为了最大化网络频谱能量效率 (SEE), 提出了一种名为联邦多智能体深度 Q 网络 (FMDQN) 的分布式侧链路资源分配方法. 该方法在满足给定任务的严格延迟限制的条件下具有良好的收敛性. 为了激励联邦学习中的客户参与模型训练, Fu 等人[18]提出了一种面向多联邦学习任务的需求均衡激励机制. 该机制考虑 IoV 场景下信道时变特性, 并设计了一种基于多智能体深度强化学习的方法来解决激励问题, 从而避免信息不对称的影响. 此方案不仅能够平衡各方需求, 还能增强用户参与度. Chen 等人[19]提出一种多无人机辅助车联网资源分配和协同卸载框架 RACOMU. 首先, 引入凸优化理论将原问题解耦, 然后通过求解 KKT (Karush-Kuhn-Tucker) 条件获得近似最优的传输功率和计算资源分配. 接下来, 设计了一种基于联邦深度强化学习 (FDRL) 的新型协作卸载策略, 该策略以分布式方式处理来自虚拟终端 (VT) 的卸载请求, 在任务处理延迟、决策时间和负载均衡度方面均取得更佳的性能.

然而, 尽管深度强化学习 (DRL) 在这些领域取得了一定进展, 关于客户端选择的问题却相对较少被关注. 客户端选择在联邦学习与车联网的协同优化中发挥着至关重要的作用, 但现有研究多聚焦于单一的任务分配或资源调度, 尚未充分探索如何利用 DRL 优化客户端选择以提升系统的稳定性与效率. 为应对车联网场景中环境动态变化及车辆资源异构性问题, 本文提出一种基于熵正则化近端策略优化的客户端选择算法 (entropy regularization proximal policy optimization, ERPPO). 该算法将客户端选择建模为马尔可夫决策过程, 综合考虑所选客户端比例、本地数据质量和系统能耗, 利用基于策略梯度的 DRL 方法做出决策. 具体贡献如下.

(1) 提出了一种动态客户端选择方法 (ERPPO) . 该方法在 PPO 策略优化过程中引入熵正则化项, 增强客户端选择策略的探索性, 使其能够根据车辆的实时状态灵活调整参与训练的客户端, 从而提高联邦学习在动态环境下的适应性和收敛效率.

(2) 为进一步提升模型聚合阶段的鲁棒性, 本文引入了置信度加权聚合机制. 该机制根据客户端上传模型更新的方差动态计算其置信度, 并自适应调整其聚合权重, 从而削弱不稳定客户端的影响, 提升全局模型更新的稳定性和训练效果.

(3) 实验结果表明, 与随机选择策略、贪婪算法以及双深度 Q 网络 (DDQN) 等方法相比, 本文所提出的动态客户端选择方法通过优化客户端的选择, 减少了不必要的通信和计算负担. 通过选择高效且稳定的客户端参与训练, 系统不仅降低了能耗, 还保证了模型的精度.

本文第 1 节总结 FL 客户端选择的相关研究工作. 第 2 节介绍车联网的 FL 框架以及系统消耗模型、数据质量评估. 第 3 节对所解决的问题进行描述. 第 4 节详细介绍基于 ERPPO 的车联网 FL 的客户端选择方案. 第 5 节对实验结果进行分析. 第 6 节对本文工作的贡献进行总结.

1 相关工作

近年来, 一些研究者致力于联邦学习中客户端选择方法的研究. 本节对 FL 在 IoV 中的应用以及客户端选择的静态及动态方法进行调查并做出概述.

1.1 IoV 场景下 FL 的应用

随着车联网技术的快速发展, 智能车辆和路侧基础设施产生了海量数据, 这些数据对于智能驾驶、交通优化和车辆协作具有重要价值. 作为一种去中心化的分布式学习范式, 联邦学习能够在保障数据隐私的同时实现跨车辆和基础设施的模型训练, 在车联网领域展现出巨大的应用潜力. 因此, 探索适用于车联网环境的 FL 方法, 对于提升智能交通系统的智能化水平和协作效率至关重要. 为了减轻传输负载并解决隐私问题, Lu 等人[20]提出了一种异步联邦学习方案, 并且开发了一种混合区块链架构. 该架构通过将学习到的模型集成到区块链中并执行两阶段验证, 保证了共享数据的可靠性. Xie 等人[21]为了解决拒绝服务 (DoS) 和欺骗等针对车联网的攻击方法对个人和社会保障构成的巨大威胁, 结合联邦学习的分布式计算资源和区块链的去中心化特性, 提出了一种名为 IoV-BCFL 的车联网入侵检测框架, 该框架能够实现分布式入侵检测和可靠日志记录, 并具有隐私保护功能. Zou 等人[22]为了解决不可信的中心化交易市场带来的安全挑战, 设计了一个声誉机制来衡量参与客户端的可靠性. 将最优定价机制建模为非合作博弈, 同时考虑到所有知识提供者之间的竞争, 有效提高了市场效用.

1.2 静态客户端选择策略

尽管联邦学习在车联网中展现出巨大的潜力, 但

其实际部署仍然面临诸多挑战. 其中, 客户端选择策略是影响 FL 性能的关键因素之一. 传统的 FL 方法通常采用静态客户端选择策略, 即在训练开始前预先确定一组固定的客户端, 并在整个训练过程中保持不变. 随机选择是客户端选择的常规方法. Amiri 等人[23]通过联合考虑信道条件和本地模型更新的重要性进行客户调度. Tan 等人[24]提出了一种基于声誉感知随机整数规划的联邦学习客户端选择方法 (SCS), 它可以以最佳方式选择和补偿具有不同声誉状况的客户端. 但是, 随机选择方法忽略了客户端之间的差异, 会使得模型聚合效率低下. 基于聚类的方法先根据客户端的资源、数据分布、位置等属性的相似性进行聚类, 然后再进行客户端的选择以提高模型的性能. Huang 等人[25]提出主动聚类联邦学习, 在每一轮模型训练中, 客户端首先确定自己所在集群, 然后每个集群都会根据不确定性采样、损失等指标筛选出一小部分对模型最有贡献的客户端. 李强等人[26]根据各客户端的计算能力将其分组, 在每轮训练中以每组客户端的平均准确率作为间接度量选择同组客户端, 在每组内根据各客户端的模型相似度对客户端进行聚类, 选择每组内不同聚类中的客户端, 实现了训练效率和全局模型准确率之间的良好平衡. Xiao 等人[27]提出了一种贪婪算法来动态选择图像质量更高的车辆, 将问题解耦成两个子问题, 分别使用拉格朗日对偶问题和启发式搜索进行求解, 实现了成本和公平的权衡.

上述方法主要是采用静态策略对客户端进行选择, 虽然在计算开销和实现复杂度上相对较低, 但在高动态、资源异构的车联网环境中, 固定的客户端选择可能导致训练过程中的资源利用率较低, 甚至会影响模型的收敛效果.

1.3 动态客户端选择策略

为了解决静态客户端选择策略导致的资源利用率低及无法适应车联网动态变化等问题, 研究人员引入强化学习 (RL) 方法, 以实现更加灵活、高效的动态客户端选择策略. Liao 等人[28]首先通过测量本地数据集的异质性和可靠性来定义数据质量度量. 然后联合考虑无线信道的动态和计算频率, 使用多臂老虎机算法来进行客户端选择. Ami 等人[29]提出了一种基于多臂老虎机的客户端选择方法 BSFL, 这种方法根据历史选择更新客户端对 FL 任务的贡献, 不仅减少了训练延迟, 还能够保证模型泛化能力. Zheng 等人[30]提出了一种基于 DRL 的 FL 框架 FedAEB. 它采用基于 SAC 网络的动态优化方法进行客户端选择和资源分配, 能够有效适应复杂和时变的系统, 提高了学习准确性. Zhang 等人[31]受到多智能体强化学习 MARL 在解决复杂控制问题方面取得成功的启发, 通过对智能体进行训练来执行高效的运行时客户端选择, 提高了模型准确性, 同时降低通信成本. Yu 等人[32]提出了一种基于多智能体近端策略优化 MAPPO 的动态客户端选择机制, 将系统建模为多智能体系统. 通过 RL 的自适应决策能力, 系统能够根据客户端的计算能力、网络状态动态调整客户端选择策略, 从而提升资源利用率、增强联邦学习在车联网环境中的适应性. 颜康等人[33]提出了一种全局感知独立近端策略优化 (GIPPO) 方法, 该方法可以根据客户端的状态自适应地选择客户端进行训练并调整迭代次数, 从而最小化整体能量消耗, 提升模型性能. Zhang 等人[34]将客户端选择机制建模为马尔可夫决策过程, 并采用 DDQN 算法解决该问题, 减少了系统开销.

现有强化学习方法在 IoV 动态环境中展现出一定优势, 但仍存在以下不足: 1) 部分方法依赖复杂的网络结构和大规模交互数据, 导致模型训练能耗较大; 2) 部分方法在训练早期收敛过快, 导致客户端选择缺乏多样性, 容易陷入局部最优; 3) 现有研究在全局聚合阶段普遍忽视了客户端上传更新的稳定性差异, 易导致全局模型收敛速度慢. 针对上述挑战, 本文提出了一种基于 PPO 的动态客户端选择方法, 并在其中引入熵正则化项与置信度加权聚合机制. 前者增强策略的探索性, 避免过早收敛; 后者利用客户端更新方差动态调整权重, 削弱不稳定更新的干扰, 从而提升全局收敛的稳定性与鲁棒性. 实验结果表明, 所提方法在保证模型精度的同时, 有效降低了能量消耗, 并提升了系统整体训练效率.

2 模型构建

在本节中, 首先给出了车联网中联邦学习的系统模型. 然后从能耗、时延、本地数据质量评估的角度阐述了基本的车联网系统模型.

2.1 IoV 中 FL 的系统模型

本节讨论了一个典型的车联网场景, 系统主要由 N 辆车辆、路边单元 (RSU) 以及一个基站组成. 车辆作为客户端持有本地数据集 D_i, 并利用自身计算能力

在本地进行模型训练. 路边单元与车辆进行交互, 接收车辆上传的本地模型参数或梯度, 并将全局模型下发至车辆. 基站负责协调各 RSU 与车辆之间的通信, 以及管理全局模型的聚合和分发, 确保系统的整体同步与收敛性. 如图 1 所示, FL 的过程包括以下 5 个步骤.

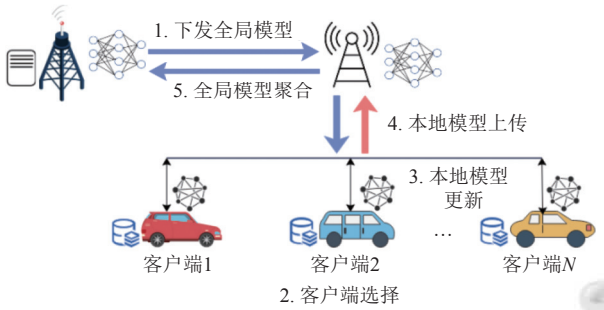

图 1 车联网中 FL 的系统模型

(1) 下发全局模型: BS 将当前全局模型w^t参数通过 RSU 广播给覆盖范围内的车辆集合.

(2) 客户端选择: 根据所提出的算法对参与模型训练的客户端进行选择.

(3) 本地模型更新: 被选中的客户端使用其本地数据D_i进行模型训练, 并更新本地模型w_i^t.

(4) 本地模型上传: 客户端将更新后的本地模型w_i^t上传至 RSU, 由 RSU 转发至 BS.

(5) 全局模型聚合: BS 在收集到参与车辆上传的模型后, 采用聚合策略更新全局模型.

总的来说, 联邦学习的优化目标为:

$$\min_{w} F(w,D) \triangleq \frac{1}{|D|}\sum_{i\in N} D_i F_i(w) \tag{1}$$

其中, w是全局模型的参数, $F(w,D)$是全局损失函数, $F_i(w)$是客户端 i 的本地损失函数, D_i是客户端 i 的本地数据集, D是全局数据集. 每个客户端 i 根据其本地数据进行模型训练, 目标是最小化本地的损失函数. 客户端 i 的局部模型可以通过梯度下降来更新, 可以表示为:

$$w_i^{t+1} = w_i^t - \eta \nabla F_i(w_i^t) \tag{2}$$

其中, η是学习率. 在本地更新后, 路边单元将模型聚合以生成新的全局模型w_i^{t+1}, 这里使用 FedAvg 的方法进行构建全局模型, 如式 (3) 所示:

$$w^{t+1} = \sum_{i=1}^{N} \frac{|D_n|}{|D|} w_i^{t+1} \tag{3}$$

2.2 延迟和能耗模型

在每一轮中, 客户端 i 的总延迟由模型更新和模型参数上传的时间组成. 由于下行链路带宽远大于上行链路带宽, 因此可以忽略服务器将模型发送到客户端的时间. 本地计算时间是指每个客户端在本地数据上执行训练过程所需的时间. 在联邦学习中, 客户端在本地进行模型训练, 并计算出本地的模型更新 (如权重、梯度等). 本地计算时间可以由式 (4) 表示:

$$T_i^{\text{loc}} = \frac{D_i q}{f_i} \tag{4}$$

其中, D_i为客户端 i 的数据量大小, q是 CPU 周期数, f_i是客户端 i 的 CPU 频率.

上传时间则是指每个客户端在本地训练完成后, 将其计算出的模型更新 (如梯度或权重) 上传至中央服务器 (或路边单元) 的时间. 上传时间由式 (5) 表示:

$$T_i^{\text{up}} = \frac{S_i}{r_i} \tag{5}$$

其中, S_i是客户端 i 的模型大小, r_i是客户端 i 的数据传输速率.

对于选定的客户端, 由于其存在本地模型训练以及模型更新上传的步骤, 因此会产生相应的能耗. 本地计算能耗是指客户端在本地训练模型时所消耗的能量, 通常取决于其计算能力和数据大小, 可以表示为:

$$E_i^{\text{loc}} = \mu D_i q f_i^2 \tag{6}$$

其中, μ为有效开关电容.

模型上传的能耗则是指客户端将本地计算结果 (如模型更新或梯度) 上传到中央服务器 (或路边单元) 时所消耗的能量, 可以表示为:

$$E_i^{\text{up}} = \frac{S_i}{r_i} p_i \tag{7}$$

其中, p_i为客户端 i 的传输功率.

2.3 本地数据质量

本地数据集的质量对联邦学习系统的性能有着深远的影响. 高质量的数据集有利于提高模型精度、加速模型收敛、减少计算开销. 为了确保联邦学习模型训练的高效性, 这里引入权重散度来量化本地数据集的质量, 定义为:

$$wd_i = \frac{\|w_i^0 - \bar{w}^0\|}{\|\bar{w}^0\|} \tag{8}$$

其中, w_i^0是客户端 i 上传的权重, $\bar{w}^0$是上一轮聚合后的

权重.

3 问题描述

在车联网场景下, 联邦学习中的客户端选择需要在保证模型精度和收敛速度的同时尽可能地减少能耗. 这涉及 3 个主要方面: 1) 客户端数量, 2) 能量消耗, 3) 权重散度. 当参与联邦学习模型训练的客户端比例较高时, 模型的收敛速度会加快, 但也会导致更高的能量消耗. 同时, 当模型具有较低的权重散度时, 全局模型的收敛会更稳定且泛化性能更好. 因此, 我们的优化目标是最大化所选客户端的比例、最小化能量消耗以及本地模型的权重散度.

优化目标可以概括如下:

$$\begin{cases} \max\limits_{\alpha} < \dfrac{N}{M}, \dfrac{1}{E}, \dfrac{1}{wd} > \\ \text{s.t.} \begin{cases} \text{C1}: \sum\limits_{i=1}^{M} \alpha_i = N,\ \alpha_i \in \{0,1\} \\ \text{C2}: E = \sum\limits_{i=1}^{M} \alpha_i E_i \\ \text{C3}: wd = \sum\limits_{i=1}^{M} \alpha_i wd_i \end{cases} \end{cases} \tag{9}$$

其中, N表示当前轮所选择的客户端数量, M表示客户端总数.

在约束 C1 中, α_i表示为客户端选择的二元变量, 用来表示客户端 i 是否被选中参与本轮训练. 当$\alpha_i = 1$时, 表示客户端 i 被选择, 当$\alpha_i = 0$时, 表示客户端 i 未被选择.

在 C2 中, E 表示所选择客户端消耗的能量总和, 它由客户端模型的本地计算能耗和模型上传能耗构成, 客户端 i 消耗的能量可以表示为:

$$E_i = E_i^{\text{loc}} + E_i^{\text{up}} \tag{10}$$

C3 中的wd则表示所有被选中的客户端的权重散度之和.

4 客户端选择优化方案

由于客户端的数据异质性和资源差异, 传统的随机选择策略往往会导致额外的通信开销和全局模型收敛缓慢, 难以满足车联网场景下对效率和稳定性的要求. 针对这一问题, 本文提出了一种动态客户端选择算法 ERPPO, 并结合置信度加权聚合机制, 在提升探索性的同时增强全局模型的收敛稳定性. 具体而言, 首先将客户端选择问题建模为马尔可夫决策过程 (Markov decision process, MDP). 在该框架下, 智能体通过与环境的交互不断优化选择策略: 智能体根据当前环境状态输出动作, 即选择参与训练的客户端集合; 随后环境反馈新的状态和奖励; 智能体再利用接收到的信息更新策略网络与价值函数网络. 不同于传统 PPO, 本研究在训练过程中引入策略熵正则化项, 以增强客户端选择的探索性, 避免策略过早收敛到局部最优. 同时, 在聚合阶段提出置信度加权机制, 依据客户端更新的方差动态调整其贡献度, 有效削弱不稳定更新的干扰. 系统框架如图 2 所示.

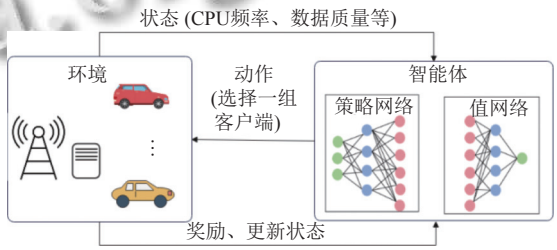

图 2 基于 ERPPO 的客户端选择框架

4.1 马尔可夫决策过程

我们首先将式 (9) 中客户端选择的优化问题建模为马尔可夫决策过程. MDP 由四元组(S, A, P, R)组成, 其中S、A、P、R分别表示为状态空间、动作空间、状态转移和奖励.

状态空间S: 状态s_i定义为客户端 i 的数据信息和计算能力, 表示为:

$$s_i = \{D_i, wd_i, f_i, r_i\} \tag{11}$$

其中, D_i为客户端 i 的数据大小, wd_i为客户端 i 的权重散度, f_i为客户端 i 的 CPU 频率, r_i为客户端 i 的数据传输速率.

动作空间A: 动作空间由路边单元对客户端的选择构成, 可以用笛卡尔积的形式表示:

$$A = \prod_{i=1}^{M} a_i \tag{12}$$

其中, $a_i = \{0,1\}$定义为客户端 i 的二进制候选动作集, 0 表示客户端 i 未被选择, 1 表示客户端 i 被选择.

状态转移P: 它表示从状态空间S到动作空间A的映射, 其中客户端在状态s_t中执行动作a_t, 并以一定的概率转换到下一个状态s_{t+1}.

奖励函数R: 对于第 3 节提出的联邦学习优化问

题 (式 (9)), 我们通过综合考虑客户端的选择比例、本地数据质量、能耗来设计奖励函数, 旨在最小化能量消耗的同时保证模型精度. 同时, 为了提高学习过程的稳定性, 我们还对奖励项进行归一化处理, 确保其具有相同的尺度. 因此, 奖励函数定义为:

$$R(s,a) = \omega_1 \frac{N}{M} - \omega_2 \overline{wd} - \omega_3 \overline{E} \tag{13}$$

其中, $\frac{N}{M}$表示所选客户端的比例, $\overline{wd}$表示归一化后的权重散度, $\overline{E}$表示归一化后的能量消耗, ω_1、ω_2、ω_3表示不同奖励项所对应的权重.

权重散度归一化表示为:

$$\overline{wd} = \frac{\sum_{i=1}^{M} a_i wd_i}{\sum_{i=1}^{M} wd_i} \tag{14}$$

其中, wd_i是客户端 i 的权重散度, M是候选客户端的数量.

能耗归一化表示为:

$$\overline{E}_i = \frac{\sum_{i=1}^{M} a_i E_i}{\sum_{i=1}^{M} E_i} \tag{15}$$

4.2 基于熵正则化 PPO 的客户端选择策略 ERPPO

为了解决 MDP 问题, 我们使用提出的 ERPPO 进行客户端选择. PPO 是 Schulman 等人[35]于 2017 年提出的策略梯度类强化学习算法, 旨在解决传统策略梯度方法训练不稳定和计算复杂的问题. 其核心思想是: 限制策略更新的幅度, 确保新策略与旧策略的差异在可控范围内, 从而避免策略更新导致的性能崩溃. 本文在 PPO 的基础上引入了熵正则化机制, 该机制通过自动调节探索力度, 实现了探索与利用之间的平衡. 因此, 这里为 ERPPO 建立了一个参数为θ的神经网络, 称为π_θ. 策略神经网络训练的目标函数定义为:

$$\begin{aligned}\underset{\theta}{\operatorname{argmax}} \frac{1}{|\Gamma_n| T_s} \sum_{\tau\in\Gamma_n} \sum_{t=0}^{T_s} \min\Bigg[& \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t, \\ & C_{\text{clip}}\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_t + \beta H(\pi_\theta(\cdot|s_t))\Bigg]\end{aligned} \tag{16}$$

其中, $\pi_\theta(a_t|s_t)$是在给定状态下选择动作的策略, C_{clip}是剪切操作符, 用于限制当前策略与旧策略之间概率比的值. T_s是轨迹收集的时间步长, Γ_n是不同客户端在时间步长内运行策略π_θ生成的轨迹集. τ_i是参与者 i 的轨迹, $\tau_i = \{s_t, a_t, r_t\}$, $t = 0,1,\cdots,T_s$, ε是剪切函数的系数, β是熵正则项的权重超参数, 控制策略探索程度的强弱. $H(\pi_\theta(\cdot|s_t))$是状态s_t下的策略熵, 可表示为:

$$H(\pi_\theta(\cdot|s_t)) = -\sum\nolimits_{at} \pi_\theta(a_t|s_t) \log \pi_\theta(a_t|s_t) \tag{17}$$

$\hat{A}_t$是优势函数, 用于衡量某一状态-动作对相对于平均策略的优劣, 由广义函数估计:

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + (\gamma\lambda)^{T_{s-t+1}} \delta T_{s-1} \tag{18}$$

$$\delta_t = r_t + \gamma V_{\phi n}(s_{t+1}) - V_{\phi n}(s_t) \tag{19}$$

其中, δ_t是时间差分误差, γ是折扣因子, λ是平滑因子, r_t是当前轮获得的奖励, V_ϕ是值函数. 值函数的训练目标是:

$$\underset{\phi}{\operatorname{argmin}} \frac{1}{|\Gamma_n| T_s} \sum_{\tau\in\Gamma_n} \sum_{t=0}^{T_s} (V_\phi(s_t) - \hat{R}_t)^2 \tag{20}$$

$\hat{R}_t$是期望奖励:

$$\hat{R}_t = r_t + \gamma r_{t+1} + \cdots + \gamma^{T_{s-t+1}} r_{T_s-1} \tag{21}$$

在本文研究中, 我们应用 ERPPO 算法来优化车联网场景中的客户端选择策略: 通过在 PPO 中加入熵正则化项, 提升了策略的探索性, 防止模型陷入局部最优. 具体而言, 策略网络依据当前车辆的状态信息 (例如: 数据量、CPU 频率、数据传输速率等), 动态选择合适的客户端参与联邦学习的训练. 通过与环境的持续交互, 策略网络接收来自环境的反馈信号 (奖励), 并据此更新策略参数, 实现策略的自适应优化. 同时, 价值网络负责对给定状态-动作对进行价值估计, 即预测在特定状态下采取某一动作所能够获得的长期累积回报. 策略网络与价值网络通过协同优化, 有效平衡探索与利用, 使系统能够在动态多变的车联网环境中自适应调整客户端选择策略. 该协同机制不仅显著提升了客户端选择的灵活性与效率, 还能够加速全局模型的收敛过程, 进一步提升系统的整体训练性能与泛化能力.

算法 1 描述了基于 ERPPO 的客户端选择过程, 该算法将客户端的状态信息作为输入, 输出客户端选择策略. 在此算法中主要是对策略神经网络π_θ和值函数V_ϕ进行训练, 当目标函数收敛或迭代完成时, 算法终止.

算法 1. 基于 ERPPO 的客户端选择策略

输入: $f_i, w_i^0, r_i, i=1,2,\cdots,M$为候选客户端.

输出: 优化后的策略π_θ.

1. 根据w_i^0计算wd_i
2. 为策略π_θ初始化参数θ_0, 并为值函数V_ϕ初始化参数ϕ_0
3. for k=1,2,$\cdots$,T do

4. 服务器向客户端广播采集状态信息
5. 客户端返回状态信息
6. 依据当前策略π_θ, 采样选择动作a_t得到参与训练的客户端子集
7. 在T_s时间步长内收集不同客户端在策略网络下做出动作所产生的轨迹集Γ_n
8. 计算当前客户端选择策略下的预期奖励$\hat{R}_t$
9. 计算每个时间步下的优势函数$\hat{A}_t$
10. 更新策略网络$\theta_k \to \theta_{k+1}$
11. 更新价值函数$\phi_k \to \phi_{k+1}$
12. end for

4.3 置信度加权聚合

在车联网场景中, 客户端由于数据分布的异质性和设备性能差异, 上传的模型更新可能存在较大波动. 为了提高全局模型聚合过程的鲁棒性, 本文还提出一种基于置信度的加权聚合策略, 动态调整各客户端在全局模型更新中的权重. 客户端置信度与其模型更新的稳定性成正比. 为此, 我们采用本地模型更新的方差作为评估指标, 计算每个客户端更新的波动性:

$$Var_i = \frac{1}{|P|}\sum_{p=1}^{P}(\Delta W_{i,p} - \overline{\Delta W_i})^2 \tag{22}$$

其中, $|P|$是模型参数总数, $\Delta W_{i,p}$是客户端 i 在参数位置p的更新量, 可以表示为:

$$\Delta W_{i,p} = w_{i,p}^t - w_p^{t-1} \tag{23}$$

$\overline{\Delta W_i}$是指客户端在所有参数位置上更新量的平均值, 可表示为:

$$\overline{\Delta W_i} = \frac{1}{P}\sum_{p=1}^{P}\Delta W_{i,p} \tag{24}$$

客户端的置信度Z_i定义为更新方差的倒数:

$$Z_i = \frac{1}{Var_i + c} \tag{25}$$

其中, $c>0$ 是一个极小值, 用于避免除零错误.

为了保证聚合时各客户端权重的总和为 1, 将置信度进行归一化处理, 得到客户端 i 的权重ω_i^t:

$$\omega_i^t = \frac{Z_i^t}{\sum_{i=1}^{N} Z_i^t} \tag{26}$$

根据归一化后的权重对客户端模型更新进行聚合, 得到新的全局模型:

$$W^t = \sum_{i=1}^{N}\omega_i^t \cdot W_i^t \tag{27}$$

使用置信度进行加权聚合, 能够动态抑制更新不稳定的客户端 (如数据分布异常或计算误差导致的更新), 从而提高全局模型的鲁棒性和收敛速度.

5 实验结果与性能分析

在本节中, 主要是对提出的 ERPPO 方法进行实验验证.首先对实验设置进行介绍, 例如实验环境、IoV 模型参数以及所用数据集. 然后将本文的算法与对比算法进行对比分析, 证明该方法的有效性.

5.1 实验设置

(1) 实验环境设置

实验环境由 Windows 10 操作系统、Intel Core i7-12700 CPU、NVIDIA GeForce RTX 3060 GPU 组成. 软件框架基于 Python 3.8、PyTorch 1.9.0、Torchvision 0.10.0 和 CUDA 11.8.

(2) 模型参数设置

对于 IoV 中的基本参数设置如表 1 所示. 该实验对车联网场景进行模拟, 模拟的 FL 系统有 20 个客户端. 车辆的 CPU 频率、传输速率以及传输功率服从均匀分布.

表 1 实验参数设置

参数	值
CPU频率 (GHz)	[2.0, 2.8]
传输速率 (Mb/s)	[20, 40]
传输功率 (W)	[0.08, 0.12]
有效电容系数	10^{-28}
每比特CPU周期数 (cycles/bit)	1 000

(3) 数据集及网络模型设置

在实验中, 我们分别使用了 MNIST 和 CIFAR-10 两种数据集对所提出的方法进行验证.

MNIST 数据集: 该数据集包含 0–9, 共 10 类手写数字, 共有 70 000 张 28×28 的灰度图像, 其中包含 60 000 张用于训练的示例和 10 000 张用于测试的示例. 在该数据集上, 我们采用具有 2 个隐藏层的多层感知机制 (MLP), 即一个全连接神经网络, 通过 ReLU 函数激活.

CIFAR-10 数据集: 包含 10 个类别、60 000 张 3×32 的彩色图像. 这些类别分别是: 飞机、汽车、鸟、猫、鹿、狗、青蛙、马、船和卡车. 数据集分为 5 个训练批次和 1 个测试批次, 每个批次包含 10 000 张图像. 测试批次包含从每个类中随机选择的 1 000 张图像, 而训练批次则包含剩余的图像, 但某些训练批次可能包含

来自一个类的图像多于另一个类的图像. 由于 CIFAR-10 较之 MNIST 更加复杂, 因此使用卷积神经网络来训练. 在此数据集上, 使用的 CNN 模型由 3 个带有两层 3×3 的卷积层组成, 每个卷积层后面都连接一个 2×2 的最大池化层, 模型使用 ReLU 函数激活.

5.2 性能比较

本节以基于随机选择策略 (Random)、基于贪婪算法 (Greedy) 以及基于双深度 Q 网络 (DDQN) 的客户端选择算法为基准, 评估了本文的动态客户端选择算法 ERPPO 的优越性. 为了模拟真实的车联网场景, 我们利用狄利克雷分布将数据集划分为非独立同分布 (Non-IID) 子集, 并将其分配给不同的客户端. 通过评估模型与基线模型在精度、损失和能耗方面的对比, 从而展现所提出方案在降低通信成本和提升模型准确性方面的优势.

(1) 奖励情况对比

如图 3 所示, 在前 200 轮训练阶段, 两种算法均表现出一定程度的波动, 这是由于客户端选择策略在探索阶段尚未稳定所致. 随着迭代的进行, 所提出的 ERPPO 算法能够更快收敛, 并在约 200 轮后趋于平稳; 相比之下, DDQN 的收敛速度明显较慢, 且在整个训练过程中始终存在较大的震荡. 该现象表明, ERPPO 在引入策略熵正则化后能够有效增强策略的探索性, 从而避免过早陷入局部最优, 并提升训练的稳定性. 就系统奖励水平而言, ERPPO 在收敛后整体高于 DDQN, 且波动幅度更小, 显示出更好的鲁棒性与泛化能力. 这一结果验证了在客户端选择过程中结合熵正则化与置信度加权聚合机制的有效性.

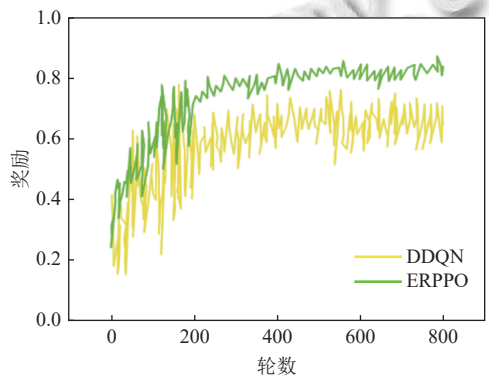

图 3 MNIST 下模型奖励对比

(2) 精度与损失对比

为了在 MNIST 和 CIFAR-10 数据集上比较所提方法的性能, 本节分别绘制了在独立同分布 (IID) 数据下各算法的精度情况, 分别在图 4 和图 5 中显示.

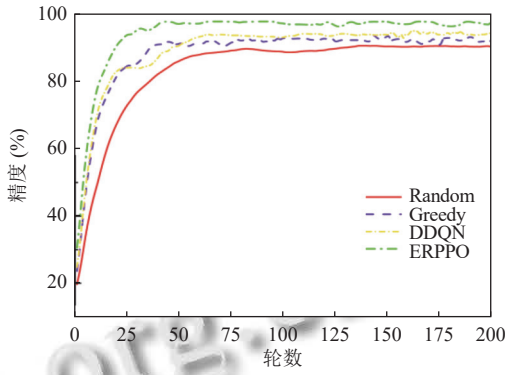

图 4 MNIST 下模型精度对比

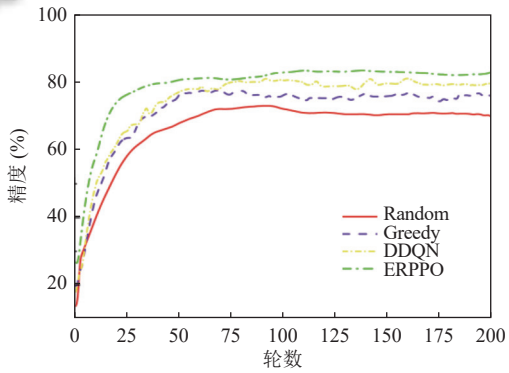

图 5 CIFAR-10 下模型精度对比

从图 4 中可以观察到, 随着训练轮数的增加, 所有方案最终都趋向于一个相对稳定的收敛阈值. 在 IID 数据分布下, MNIST 数据集上, 本文提出的 ERPPO 方法在收敛速度和最终精度上均优于其他基线算法. 与随机选择策略相比, 精度相对提升 7.96%; 与贪婪算法相比, 精度相对提升 6.14%; 与 DDQN 相比, 精度相对提升 3.43%. 从图 5 中可以观察到在 CIFAR-10 数据集上, ERPPO 方法同样表现出更快的收敛速度和更高的最终精度. 与随机选择策略相比, 精度相对提升 18.56%; 与贪婪算法相比, 精度相对提升 9.11%; 与 DDQN 相比, 精度相对提升 4.02%.

为证明模型的有效性, 我们进一步在 Non-IID 数据分布下进行实验验证, 图 6–图 9 展示了本模型和基线方法在 MNIST 数据集和 CIFAR-10 数据集中采用 Non-IID 数据分布下的精度和损失情况.

从图 6 可以看出, ERPPO 方法在所有轮次中均表现最佳, 最终精度稳定在 97.35% 左右; 而随机选择策略收敛速度最慢, 最终精度在 86.86% 左右. 贪婪算

法和 DDQN 两种策略的精度收敛趋势相似, 均优于随机选择策略, 但略低于 ERPPO.

图 7 展示了损失曲线的变化情况, 所有方法的损失均呈下降趋势, 其中 ERPPO 收敛速度最快, 最终达到最低损失值 0.109. 随机选择策略的损失下降最慢, 最终稳定在 0.251 附近, 这表明其在全局模型优化方面的效率较低.

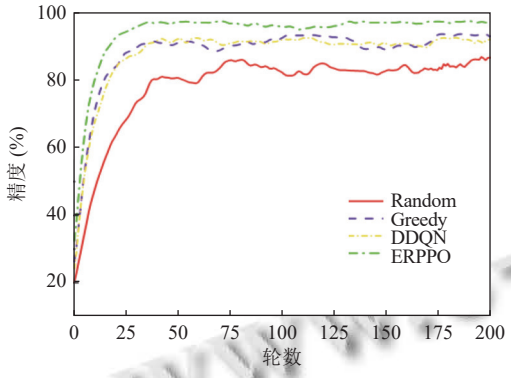

图 6 MNIST 下 Non-IID 精度对比

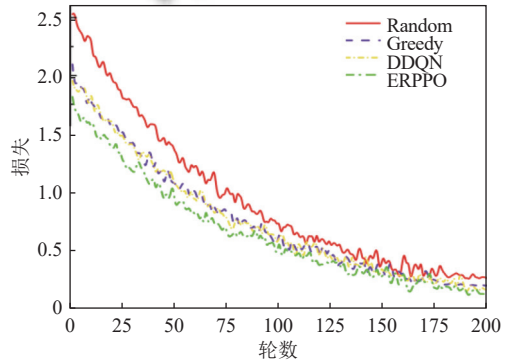

图 7 MNIST 下 Non-IID 损失对比

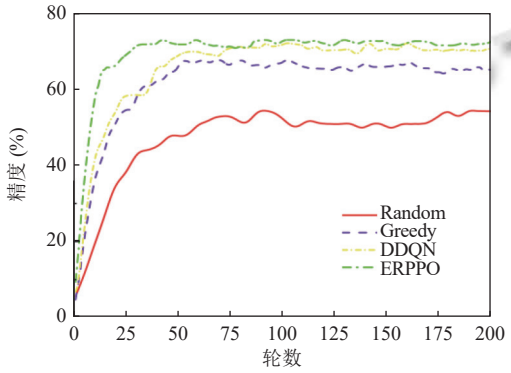

图 8 CIFAR-10 下 Non-IID 精度对比

对于 CIFAR-10 数据集, 从图 8 中可以看出, 当迭代次数达到 75 左右的时候, 各个算法精度开始收敛, 比 MNIST 数据集所需时间更多, 这是因为 CIFAR-10 数据集相较于 MNIST 数据集类别形态多变、图像更高维、背景更复杂, 需要更深的模型才能取得较高的准确率. 在训练轮数达到 200 轮时, 与随机选择策略相比, ERPPO 的精度从 54.229% 提升至 72.553%, 提升了 33.82%; 与贪婪算法相比, 精度提升 11.38%; 与 DDQN 相比, 精度提升 2.59%. 图 9 展示了各算法在 CIFAR-10 数据集下的损失变化趋势, 可以看出本文方法在收敛速度和稳定性上均优于对比算法. 与随机选择策略相比, ERPPO 的损失值从 0.513 降低到 0.416, 降低了 18.89%; 与贪婪算法相比, 降低了 13.68%; 与 DDQN 相比, 降低了 7.76%. 这一结果表明, ERPPO 能够在动态环境中选择更优客户端, 提升模型收敛性并有效减少全局模型的更新损失.

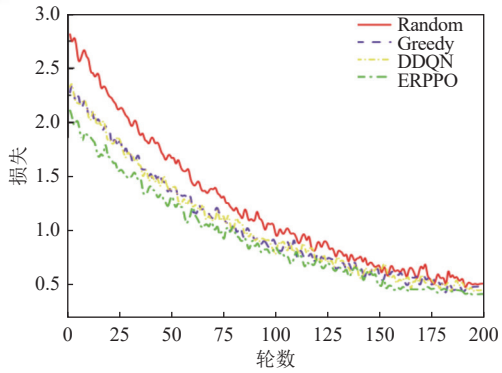

图 9 CIFAR-10 下 Non-IID 损失对比

(3) 能量消耗

图 10 和图 11 展示 ERPPO 与 3 个基线在 MNIST 和 CIFAR-10 数据集上达到同样精度所需的通信能量消耗对比. 可以观察到, ERPPO 在 MNIST 数据集和 CIFAR-10 数据集中均保持最小能耗成本.

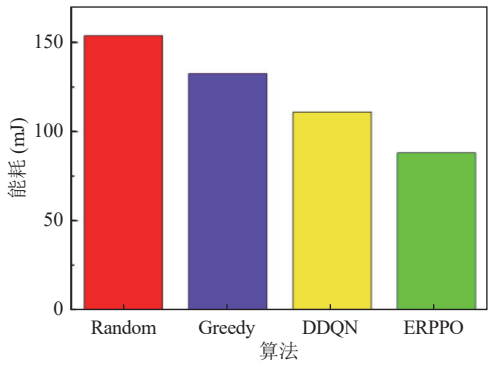

图 10 MNIST 数据集下能耗对比

图 10 展示了 MNIST 数据集下各算法能量消耗对比. 从图中可以看出, 基于随机选择策略的客户端选择

算法在能量消耗方面显著高于其他算法. 这是因为随机选择策略在每一轮训练中随机选择参与客户端, 无法有效感知各节点的计算能力、通信状态和数据分布特性, 从而引发较高的通信开销, 整体能耗显著增加. 而 DDQN 相较其他对比算法表现出更低的能量消耗, 这得益于其动态客户端选择策略能够自适应车联网环境的高动态性, 合理控制参与训练的客户端规模, 有效降低系统能耗.

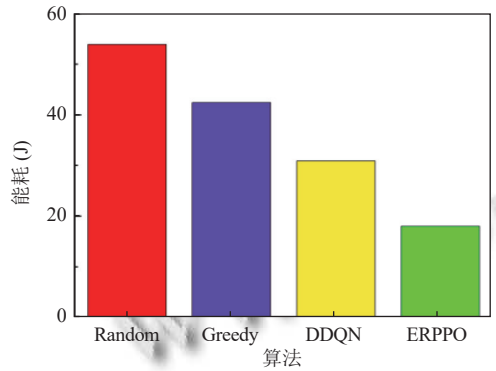

图 11 CIFAR-10 数据集下能耗对比

图 11 进一步给出了 CIFAR-10 数据集下各算法的能耗对比结果. Random、Greedy、DDQN 以及 ERPPO 选择机制在模型收敛后的能耗分别为 53.87 J、42.41 J、30.89 J 以及 17.99 J. 与随机选择策略相比, ERPPO 的能耗降低了 66.60%; 与贪婪算法相比, 能耗降低了 57.61%; 相较于 DDQN, ERPPO 也实现了 41.77% 的能耗节省. 在两种数据集下, 本文提出的算法都表现出了更低的能耗, 这是因为我们的方法同时考虑了能耗和精度, 实现了两者之间的有效权衡.

(4) 时延

图 12 展示了 MNIST 数据集下各算法的时延对比, 可以看出, 随机选择策略的时延最高, 这是由于其在客户端选择过程中不加区分, 导致经常选到计算与通信资源较差的节点, 从而增加整体训练时间. 贪婪算法在一定程度上降低了时延, 但仍未能有效避免低性能客户端的参与. 相比之下, DDQN 利用强化学习机制能够自适应环境变化, 减少了低效节点的选择, 因此整体时延更低. 而本文提出的 ERPPO 算法在收敛后表现出最低的时延, 较 Random、Greedy 和 DDQN 均有显著改善, 说明其在保证模型训练效果的同时, 能够有效缩短全局迭代的时间.

图 13 进一步展示了 CIFAR-10 数据集下的时延对比结果. 与 MNIST 的结论一致, ERPPO 在该数据集上的时延依然最低, 显著优于其他 3 种方法. 这主要得益于 ERPPO 在客户端选择中同时引入策略熵正则化和置信度加权机制, 使得所选客户端在计算能力与通信条件上更具均衡性, 从而避免了系统瓶颈节点对整体时延的影响.

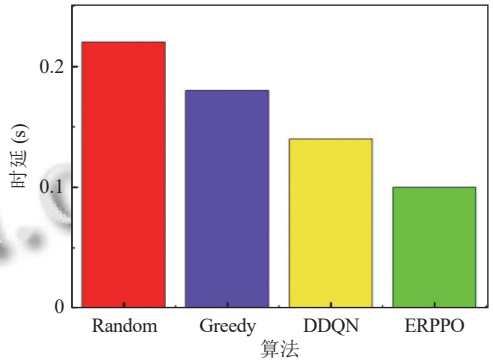

图 12 MNIST 数据集下时延对比

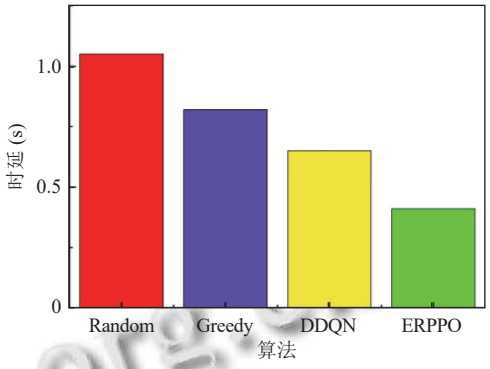

图 13 CIFAR-10 数据集下时延对比

6 结论

为应对 IoV 环境下 FL 面临的客户端异构性与高动态性挑战, 本文提出了一种名为 ERPPO 的动态客户端选择算法. 该方法将客户端选择建模为马尔可夫决策过程, 并在 PPO 框架中引入熵正则化项, 鼓励策略在训练初期保持探索多样性, 避免过早陷入局部最优. 为进一步提升全局模型的鲁棒性与泛化性能, 我们还设计了基于客户端模型更新方差的置信度加权聚合方法, 该方法通过对上传模型进行动态加权整合, 削弱不稳定客户端对全局模型的负面影响. 实验结果表明, ERPPO 在保证模型精度的同时能够有效降低能耗, 并在多项性能指标上优于随机选择、贪婪算法和 DDQN 等基线方法, 从而验证了所提方法的有效性.

参考文献

1 张爽, 张晨, 彭淑敏, 等. 面向 6G 的智能物联网通信关键技术综述. 无线电工程, 2025, 55(4): 699–713.

2 Wen J, Zhang ZX, Lan Y, *et al*. A survey on federated learning: Challenges and applications. International Journal of Machine Learning and Cybernetics, 2023, 14(2): 513–535. [doi: 10.1007/s13042-022-01647-y]

3 李少波, 杨磊, 李传江, 等. 联邦学习概述: 技术、应用及未来. 计算机集成制造系统, 2022, 28(7): 2119–2138.

4 肖雄, 唐卓, 肖斌, 等. 联邦学习的隐私保护与安全防御研究综述. 计算机学报, 2023, 46(5): 1019–1044.

5 杨宇宁. 面向智慧医疗场景的联邦学习协同优化研究 [博士学位论文]. 北京: 北京邮电大学, 2025.

6 Friha O, Ferrag MA, Shu L, *et al*. FELIDS: Federated learning-based intrusion detection system for agricultural Internet of Things. Journal of Parallel and Distributed Computing, 2022, 165: 17–31. [doi: 10.1016/j.jpdc.2022.03.003]

7 唐伦, 文明艳, 单贞贞, 等. 移动边缘计算辅助智能驾驶中基于高效联邦学习的碰撞预警算法. 电子与信息学报, 2023, 45(7): 2406–2414.

8 Neumann D, Lutz A, Müller K, *et al*. A privacy preserving system for movie recommendations using federated learning. ACM Transactions on Recommender Systems, 2024, 3(2): Article No. 14. [doi: 10.1145/3634686]

9 Wang JH, Zhu K, Hossain E. Green Internet of Vehicles (IoV) in the 6G era: Toward sustainable vehicular communications and networking. IEEE Transactions on Green Communications and Networking, 2022, 6(1): 391–423. [doi: 10.1109/TGCN.2021.3127923]

10 Chellapandi VP, Yuan LQ, Brinton CG, *et al*. Federated learning for connected and automated vehicles: A survey of existing approaches and challenges. IEEE Transactions on Intelligent Vehicles, 2024, 9(1): 119–137. [doi: 10.1109/TIV.2023.3332675]

11 Cho YJ, Wang JY, Joshi G. Towards understanding biased client selection in federated learning. Proceedings of the 25th International Conference on Artificial Intelligence and Statistics. Valencia: PMLR, 2022. 10351–10375.

12 Huang TS, Lin WW, Shen L, *et al*. Stochastic client selection for federated learning with volatile clients. IEEE Internet of Things Journal, 2022, 9(20): 20055–20070.

13 Li CN, Zeng X, Zhang M, *et al*. PyramidFL: A fine-grained client selection framework for efficient federated learning. Proceedings of the 28th Annual International Conference on Mobile Computing and Networking. Sydney: ACM, 2022. 158–171.

14 Hazarika B, Singh K, Biswas S, *et al*. DRL-based resource allocation for computation offloading in IoV networks. IEEE Transactions on Industrial Informatics, 2022, 18(11): 8027–8038. [doi: 10.1109/TII.2022.3168292]

15 Abishu HN, Seid AM, Jhaveri RH, *et al*. Blockchain-empowered resource allocation in HAPS-assisted IoV digital twin networks: A federated DRL approach. IEEE Transactions on Intelligent Vehicles, 2025. 10(10): 4710–4726. [doi: 10.1109/TIV.2024.3492015]

16 Quan HY, Zhang QM, Zhao JH. Federated learning assisted intelligent IoV mobile edge computing. IEEE Transactions on Green Communications and Networking, 2025, 9(1): 228–241. [doi: 10.1109/TGCN.2024.3421357]

17 Song XQ, Hua YQ, Yang Y, *et al*. Distributed resource allocation with federated learning for delay-sensitive IoV services. IEEE Transactions on Vehicular Technology, 2024, 73(3): 4326–4336. [doi: 10.1109/TVT.2023.3328988]

18 Fu YC, Dong MY, Zhou LL, *et al*. A distributed incentive mechanism to balance demand and communication overhead for multiple federated learning tasks in IoV. IEEE Internet of Things Journal, 2025, 12(8): 10479–10492. [doi: 10.1109/JIOT.2024.3510561]

19 Chen ZY, Huang ZQ, Zhang JJ, *et al*. Resource allocation and collaborative offloading in multi-UAV-assisted IoV with federated deep reinforcement learning. IEEE Internet of Things Journal, 2025, 12(5): 4629–4640. [doi: 10.1109/JIOT.2024.3516838]

20 Lu YL, Huang XH, Zhang K, *et al*. Blockchain empowered asynchronous federated learning for secure data sharing in Internet of Vehicles. IEEE Transactions on Vehicular Technology, 2020, 69(4): 4298–4311. [doi: 10.1109/TVT.2020.2973651]

21 Xie NN, Zhang CX, Yuan QZ, *et al*. IoV-BCFL: An intrusion detection method for IoV based on blockchain and federated learning. Ad Hoc Networks, 2024, 163: 103590. [doi: 10.1016/j.adhoc.2024.103590]

22 Zou Y, Shen F, Yan F, *et al*. Reputation-based regional federated learning for knowledge trading in blockchain-enhanced IoV. Proceedings of the 2021 IEEE Wireless Communications and Networking Conference (WCNC). Nanjing: IEEE, 2021. 1–6.

23 Amiri MM, Gündüz D, Kulkarni SR, *et al*. Convergence of update aware device scheduling for federated learning at the wireless edge. IEEE Transactions on Wireless Communications, 2021, 20(6): 3643–3658. [doi: 10.1109/

TWC.2021.3052681]

24 Tan X, Ng WC, Lim WYB, *et al*. Reputation-aware federated learning client selection based on stochastic integer programming. IEEE Transactions on Big Data, 2024, 10(6): 953–964. [doi: 10.1109/TBDATA.2022.3191332]

25 Huang HL, Shi W, Feng YH, *et al*. Active client selection for clustered federated learning. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(11): 16424–16438. [doi: 10.1109/TNNLS.2023.3294295]

26 李强, 张凌羽, 孟祥宇. 资源高效的聚类协同联邦学习客户端选择方法. 吉林大学学报 (工学版), 2025, 55(10): 3337–3345. [doi: 10.13229/j.cnki.jdxbgxb.20231369]

27 Xiao HZ, Zhao J, Pei QQ, *et al*. Vehicle selection and resource optimization for federated learning in vehicular edge computing. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(8): 11073–11087. [doi: 10.1109/TITS.2021.3099597]

28 Liao YY, Feng J, Zhou ZJ, *et al*. Quality-aware client selection and resource optimization for federated learning in computing networks. Proceedings of the 2024 IEEE International Conference on Communications. Denver: IEEE, 2024. 2628–2633.

29 Ami DB, Cohen K, Zhao Q. Client selection for generalization in accelerated federated learning: A bandit approach. Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island: IEEE, 2023. 1–5.

30 Zheng F, Sun YZ, Ni B. FedAEB: Deep reinforcement learning based joint client selection and resource allocation strategy for heterogeneous federated learning. IEEE Transactions on Vehicular Technology, 2024, 73(6): 8835–8846. [doi: 10.1109/TVT.2024.3359860]

31 Zhang SQ, Lin JY, Zhang Q. A multi-agent reinforcement learning approach for efficient client selection in federated learning. Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022. 9091–9099.

32 Yu TQ, Wang XB, Hu JL, *et al*. Multi-agent proximal policy optimization-based dynamic client selection for federated AI in 6G-oriented Internet of Vehicles. IEEE Transactions on Vehicular Technology, 2024, 73(9): 13611–13624. [doi: 10.1109/TVT.2024.3383860]

33 颜康, 束妮娜, 吴韬, 等. 基于深度强化学习的高效联邦学习客户端选择方法. 信息对抗技术, 2025, 4(3): 84–96.

34 Zhang HJ, Xie ZJ, Zarei R, *et al*. Adaptive client selection in resource constrained federated learning systems: A deep reinforcement learning approach. IEEE Access, 2021, 9: 98423–98432. [doi: 10.1109/ACCESS.2021.3095915]

35 Schulman J, Wolski F, Dhariwal P, *et al*. Proximal policy optimization algorithms. arXiv:1707.06347, 2017.

(校对责编: 张重毅)