数据新鲜度驱动的协作式无人机联邦学习智能决策优化研究

范 文 韦 茜 周 知 于 帅 陈 旭*
(中山大学计算机学院 广州 510006)

摘 要：联邦学习是6G关键技术之一，其可以在保护数据隐私的前提下，利用跨设备的数据训练一个可用且安全的共享模型。然而，大部分终端设备由于处理能力有限，无法支持复杂的机器学习模型训练过程。在异构网络融合环境下移动边缘计算(MEC)框架中，多个无人机(UAVs)作为空中边缘服务器以协作的方式灵活地在目标区域内移动，并且及时收集新鲜数据进行联邦学习本地训练以确保数据学习的实时性。该文综合考虑数据新鲜程度、通信代价和模型质量等多个因素，对无人机飞行轨迹、与终端设备的通信决策以及无人机之间的协同工作方式进行综合优化。进一步，该文使用基于优先级的可分解多智能体深度强化学习算法解决多无人机联邦学习的连续在线决策问题，以实现高效的协作和控制。通过采用多个真实数据集进行仿真实验，仿真结果验证了所提出的算法在不同的数据分布以及快速变化的动态环境下都能取得优越的性能。

关键词：移动边缘计算；联邦学习；深度强化学习；无人机；信息年龄
中图分类号：TN915 **文献标识码**：A **文章编号**：1009-5896(2022)09-2994-10
DOI: 10.11999/JEIT211406

A Research on Collaborative UAVs Intelligent Decision Optimization for AoI-driven Federated Learning

FAN Wen WEI Qian ZHOU Zhi YU Shuai CHEN Xu
(*School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou* 510006, *China*)

Abstract: Federated learning is one of the key technologies of 6G, which can use cross-device data to train a usable and safe sharing model on the premise of protecting data privacy. However, most end devices have limited processing capabilities and can not support complex machine learning model training processes. In the framework of Mobile Edge Computing (MEC) in a heterogeneous network convergence environment, multiple Unmanned Aerial Vehicles (UAVs) are used as aerial edge servers to move flexibly within the target area in a collaborative manner, and collect fresh data in time for federated learning and local training to ensure real-time data learning. Multiple factors, such as data freshness, communication cost and model quality, are considered, and the flight trajectories of UAVs, the communication decisions with the user equipment, and the collaborative work between UAVs are comprehensively optimized. Moreover, a priority-based decomposable multi-agent deep reinforcement learning algorithm is used to solve the continuous online decision-making problem of multiple UAVs federated learning to achieve effective collaboration and control. By using multiple real data sets for simulation experiments, simulation results verify that the proposed algorithm can achieve superior performance under different data distributions and in rapidly changing complex dynamic environments.

Key words: Mobile Edge Computing (MEC); Federated learning; Deep reinforcement learning; Unmanned Aerial Vehicle (UAV); Age-of-Information (AoI)

1 引言

在传统计算范式中，用户设备通常将原始数据上传至集中云服务器进行处理，但是这不可避免地造成极大的传输开销和数据隐私泄露。针对该问题，联合利用移动边缘计算(Mobile Edge Computing, MEC) [1]和联邦学习[2]设计解决方案逐渐成为研究焦点。一方面，边缘服务器分担用户设备的联邦学习本地训练任务，既减轻用户设备的计算负载，又降低向云端传输数据造成的开销；另一方面，利用本地化模型训练结果聚合成全局共享模型，避

收稿日期：2021-11-30；改回日期：2022-05-13；网络出版：2022-06-07
*通信作者： 陈旭 chenxu35@mail.sysu.edu.cn
基金项目：国家自然科学基金(U20A20159, 61972432)
Foundation Items: The National Natural Science Foundation of China (U20A20159, 61972432)

免了隐私数据泄露的弊端，有利于实现快速、高效的训练过程。然而，边缘服务器通常是位置固定的且覆盖范围有限的，这将导致其无法灵活有效地处理复杂变化的强实时性任务[3]。

随着下一代网络系统如6G通信网络的快速发展，高性能无人机(Unmanned Aerial Vehicle, UAV)已被视为具备感知、计算和存储能力的空中边缘服务器[4]。与传统的安装在地面基站上的固定边缘服务器相比，无人机利用其高度敏捷性、灵活性和移动性实现按需部署，增强了系统的覆盖范围[5]。在许多强实时性应用场景(如交通管理、环境和灾难监测、战场监视等[6])中，多个无人机在不同区域中移动，及时接收众多分散的用户数据，以协作的方式完成复杂的移动边缘计算任务，训练具有高可用性和高实时性的机器学习模型(例如，图像分类模型)[7]。进一步地，在联邦学习模式下，多无人机完成训练后只需要将本地模型参数上传至云服务器进行全局模型聚合，实现训练模型的共享和隐私保护。

值得注意的是，无人机的感知半径有限，且有限的机载电池会约束无人机的移动范围，因此无法保证每个用户设备产生的数据都能及时地被无人机接收并处理。而在移动边缘计算场景中，数据的实时处理对其可用性和模型的实时更新非常重要。为此，文献[8]在模型中采用数据的信息年龄(Age-of-Information, AoI) 来刻画数据的新鲜程度，将其定义为数据最近一次成功传输后经过的时间[9]。但是，它们忽略了数据在区域中等待的时间，这对MEC 场景中无人机的模型训练和通信决策是至关重要的，特别是在多无人机协作训练的情况下。本文将数据的新鲜程度，即数据在端设备上等待的时间与被无人机接收并处理的时间之和定义为数据的信息年龄[10]，通过最小化信息年龄来优化无人机移动边缘计算决策，提升联邦学习性能，增强数据处理实时性。因此，如何规划无人机的路径和制定通信决策，以及如何在无人机之间展开协同工作，合理地分配计算资源，同时满足能耗和时延的限制，成为本文需要解决的关键问题。

针对上述挑战，本文提出了一种崭新的基于数据新鲜程度的协作式无人机联邦学习范式，通过多无人机协同地智能地进行移动、通信和计算卸载决策，高效地完成了边缘数据处理任务，显著地降低了无人机的能量消耗并保证了模型高准确率和低数据信息年龄。本文进一步提出一种多智能体深度强化学习(Deep Reinforcement Learning, DRL)算法，有效地处理复杂状态空间，实现多无人机的高效协作和智能决策优化。本文的主要贡献包括4个方面：

(1) 提出面向实时边缘数据处理的多无人机协作式联邦学习范式，能够充分发挥无人机辅助移动边缘计算和联邦学习的优势，避免了云中心集中式数据处理的用户隐私保护弱和任务处理时延大等不足；

(2) 引入信息年龄以描述协作式无人机联邦学习的训练数据的新鲜程度，并据此对多无人机协同决策问题进行建模，以联合优化边缘数据处理的模型准确率、信息年龄以及总体能耗；

(3) 设计了一种新颖的具有全局和局部奖励的优先级多智能体深度强化学习算法，实现多无人机协同地移动、通信和任务卸载决策智能联合优化；

(4) 采用多个真实机器学习数据集进行仿真实验并设置了充分的对比实验，结果表明了本文提出的算法在不同数据分布下和在快速变化的复杂动态环境中都能实现优越的性能表现。

2　系统模型与问题形式化

2.1　区域模型

如图1所示，感知区域被划分为$\mathcal{M}=\{1,\ 2,\cdots,M\}$个子区域，每个子区域的中心位置设为用户设备，它感知并传输该子区域的实时数据至边缘服务器进行处理。在本系统中，由于安装在地面基站上的边缘服务器(后文简称为基站(Base Station, BS))的覆盖范围以及用户设备的射频功率有限，用户设备无法与基站直接通信。为了解决计算的局限性，系统部署多个无人机以接收和处理其覆盖范围内用户设备的实时数据。这些无人机配备了完成计算任务所必要的载荷，包括数据收发设备(如天线)、数据存储设备(如存储卡)和数据处理设备(如嵌入式CPU)，以及基本设备(如机体、电池、动力控制和飞行控制装置)及其相关传感器。无人机的载荷高度集成化使其数据存储、数据处理和移动的综合能力远在固定的边缘服务器之上。在本文中，无人机作为性能适中的边缘服务器，支持长、短距离无线通信，

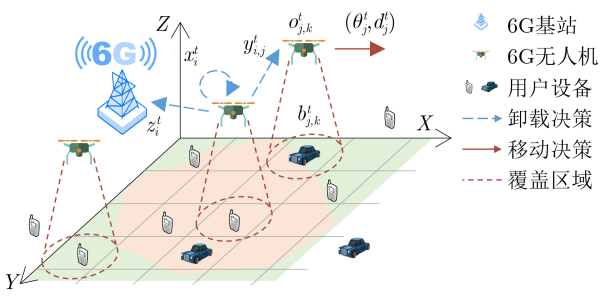

图 1 基于MEC的多无人机感知区域

能够为基站覆盖不了的区域提供计算服务。因此，无人机可以高效地充当边缘计算节点来完成本文的边缘计算任务。

在目标感知区域中，一组无人机$\mathcal{N}=\{1,2,\cdots,N\}$组成的智能体群以固定高度$H$飞行。在每个时隙$t\in\mathcal{T}=\{1,2,\cdots,T\}$结束时，无人机$i$以方向$\theta_i^t\in[0,2\pi)$和距离$d_i^t\in[0,l^{\max})$飞往下一个感知子区域，其中$l^{\max}$为无人机在单个时隙内的最大飞行距离。考虑采用二元变量$o_{i,k}^t\in0,1$表示无人机i在时隙t时的位置：当且仅当无人机$i\in\mathcal{N}$处于子区域$k\in\mathcal{M}$上空时，$o_{i,k}^t=1$；否则$o_{i,k}^t=0$。无人机的移动决策有约束条件

$$\sum_{k=1}^{M}o_{i,k}^t=1,\sum_{i=1}^{N}o_{i,k}^t\in\{0,1\} \tag{1}$$

式(1)表示每个时隙内无人机只能停留在一个子区域，并且多个无人机不能停留在同一个子区域。

假设无人机i的感知能力定义为其最大通信半径$R_i^{\max}$，任何在最大通信范围内的用户设备都被认为是可感知的并且其数据可收集的。无人机i从覆盖的子区域集合$\mathcal{M}_i\subset\mathcal{M}$中收集数据信息时，满足约束

$$b_{i,k}^tR_{i,k}^t\le R_i^{\max} \tag{2}$$

其中，$b_{i,k}^t\in0,1$是通信决策二元变量，它表示时隙t时无人机i是否与子区域k中的用户设备(以下统称用户设备k)通信：$b_{i,k}^t=1$表示通信，否则$b_{i,k}^t=0$。此外，无人机的通信还存在约束

$$\sum_{i=1}^{N}b_{i,k}^t\in\{0,1\} \tag{3}$$

其表示当用户设备k同时处于两个及以上无人机的覆盖范围内时，它将至多选择一个无人机进行通信。本文部分参数以及定义如表1所示。

2.2 联邦学习模型

系统的任务是多无人机作为本地训练节点，利用区域的实时数据来协同训练全局模型，为各类数据分析的智能应用提供服务。本文使用联邦学习框架，以多分类预测模型为例，进行多设备之间的协作学习。联邦学习是一个反复迭代直到全局模型收敛的过程，它的每一轮全局迭代包括以下步骤：

(1) 下载全局模型：无人机i从云服务器下载最新的全局模型$\boldsymbol{\omega}^{t-1}$，将其作为初始本地模型：$\boldsymbol{\omega}_i^{t,0}\leftarrow\boldsymbol{\omega}^{t-1}$；

(2) 本地模型训练：无人机i从覆盖区域内的用户设备接收训练数据，执行随机梯度下降方法进行本地模型的训练：$\boldsymbol{\omega}_i^{t,n}=\boldsymbol{\omega}_i^{t,n-1}-\eta\nabla L_i\left(\boldsymbol{\omega}_i^{t,n-1}\right),n\ge1$，是本地模型的学习率，$L_i\left(\boldsymbol{\omega}_i^{t,n}\right)$是损失函数。本地模型训练的停止条件是$\left\|\nabla L_i\left(\boldsymbol{\omega}_i^{t,n}\right)\right\|\le\epsilon\left\|\nabla L_i\left(\boldsymbol{\omega}_i^{t,n-1}\right)\right\|$，本地模型精度$\epsilon\in(0,1)$将影响本地训练的迭代次数[11]。训练停止时得到本地迭代次数$I_i^t(\epsilon)=n$，并将此时的本地模型表示为$\boldsymbol{\omega}_i^t\leftarrow\boldsymbol{\omega}_i^{t,n}$；

(3) 上传本地模型：每个无人机节点将训练后的本地模型$\boldsymbol{\omega}_i^t$上传至云服务器；

(4) 全局模型聚合：远程云服务器将接收到的本地模型通过联邦平均算法加权聚合，得到新的全局模型：$\boldsymbol{\omega}^t=\dfrac{\sum_{i=1}^{N}A_i^t\omega_i^t}{\sum_{i=1}^{N}A_i^t}$。其中，$A_i^t$表示无人机$i$从其覆盖区域的用户设备接收并用于本地训练的数据量。

本文主要以分类模型的预测准确性来衡量联邦学习的效果。假设Acc^t表示t时隙的全局模型预测准确率，基于联邦学习的多分类预测模型的优化目标之一是最大化Acc^t。

2.3 任务卸载模型

假设每个无人机有3种任务卸载决策：本地处理、卸载至其它无人机处理(UAV-to-UAV, U2U)和卸载至基站处理(UAV-to-BS, U2B)。$x_i^t\in[0,1]$表示无人机i在本地处理的数据比例，$y_{i,j}^t\in[0,1]$表示无人机i通过U2U方式卸载至无人机j的数据比例，$z_i^t[0,1]$表示无人机i通过U2B方式卸载至基站的数据比例。为了刻画无人机之间的任务卸载情况，本文定义U2U有向连接图$\boldsymbol{G}^t=\{N,\boldsymbol{\varepsilon}^t\}$，其中$\boldsymbol{\varepsilon}^t=\{<i,j>:l_{i,j}^t=1,\forall i,j\in N\}$是边集，它表示在时隙$t$时无人机$i$与无人机$j$的链路连通情况。因此，有卸载决策约束

表 1 系统参数及其定义

参数	定义	参数	定义
$R_{i,k}^t$	无人机i与用户设备k的欧氏距离	c_i, c_j, c_B	无人机i、无人机j、基站的计算能力
W	通信带宽	σ, g_0	噪声功率和每米的信道功率增益
$p_{i,\mathrm{B}}^{\mathrm{tra}}$, $p_{i,\mathrm{U}}^{\mathrm{tra}}$	U2B, U2U数据传输功率	λ_{B}^t, λ_j^t	在基站、无人机j上的排队时延
$d_{i,\mathrm{B}}$, $d_{i,\mathrm{j}}$	无人机i与基站、无人机j的欧氏距离	d_{k_1,k_2}	无人机从用户设备k_1到k_2的移动距离
v_i^{mov}, p_i^{mov}	无人机i的移动速率、移动功率	p_i^{cmp}, p_i^{rev}	无人机i的数据计算功率、数据接收功率

$$x_i^t+\sum_{j=1,j\neq i}^{N}y_{i,j}^t l_{i,j}^t+z_i^t=1 \tag{4}$$

式(4)表示无人机i执行卸载决策的数据总量要与从覆盖区域内的用户设备接收的数据总量一致。在多无人机协作过程中，每个无人机进行数据通信时主要传输实时数据和模型。相对于任务数据量的大小，模型的大小一致且可以忽略。因此，本文主要考虑任务数据传输时所产生的通信时延开销和通信能耗开销。

2.4 信息年龄模型

为了更好满足实时应用需求，本文定义数据的AoI为由用户设备产生直到被无人机处理完成的时间长度。τ_k^t是用户设备k累计到时隙t的空闲时间长度，它表示从上一次与无人机通信到时隙t的时间间隔。因此，$\tau_k^t=\left(\tau_k^{t-1}+\Delta t\right)\left(1-\sum_{i=1}^{N}b_{i,k}^t\right)$，其中$\Delta t$表示时隙长度。用户设备$k$与无人机通信后，它的空闲时间为0，否则逐渐递增直到下一次与无人机通信。当用户设备k不与无人机通信时，它在该时隙内的AoI为其等待时间。

由于异构用户设备的数据感知能力存在差异，数据感知速度都不一样，如果已知用户设备k的数据感知速度φ_k，那么它在空闲时间内产生的数据量为$a_k^t=\tau_k^t\varphi_k$。当用户设备k与无人机i通信时，它将数据量a_k^t全部上传至无人机i。考虑到一个无人机可以同时与覆盖范围内的多个用户设备通信，因此，无人机i在时隙t时接收到的数据量为$A_i^t=\sum_{k=1}^{M}a_k^t b_{i,k}^t$，所产生的接收时延为$T_{i,\mathrm{rev}}^t=\dfrac{A_i^t}{P_i}$。其中，$P_i$为无人机$i$的数据接收能力。

无人机接收数据后，将根据自身的卸载决策进行数据处理，产生的时延包括以下3种：

(1) 本地处理时延$T_{i,i}^t=I_i^t(\epsilon)\dfrac{A_{i,i}^t}{c_i}$。其中，$A_{i,i}^t=A_i^t x_i^t$表示无人机$i$本地计算的数据量。$T_{i,i}^t$与联邦学习本地模型的迭代次数$I_i^t(\epsilon)$有关，迭代次数越多，所需要的计算时延越长。

(2) U2B卸载与计算时延$T_{i,\mathrm{B}}^t=\dfrac{A_{i,\mathrm{B}}^t}{v_{i,\mathrm{B}}^t}+I_i^t(\epsilon)\dfrac{A_{i,\mathrm{B}}^t}{c_{\mathrm{B}}}+\lambda_{\mathrm{B}}^t$。其中，$A_{i,\mathrm{B}}^t=A_i^t z_i^t$表示无人机$i$卸载至基站的数据量，$v_{i,\mathrm{B}}^t=W\log_2\left(1+\dfrac{\alpha p_{i,\mathrm{B}}^{\mathrm{tra}}}{d_{i,\mathrm{B}}^2}\right)$是无人机$i$到基站的数据传输速率，常数$=\dfrac{g_0G_0}{2}$[12]，$G_0\approx 2.284$。

(3) U2U卸载与计算时延$T_{i,j}^t=\dfrac{A_{i,j}^t}{v_{i,j}^t}+I_i^t(\epsilon)\dfrac{A_{i,j}^t}{c_j}+\lambda_j^t$。其中，$A_{i,j}^t=A_i^t y_{i,j}^t l_{i,j}^t$表示无人机$i$卸载至无人机$j$的数据量，$v_{i,j}^t=W\log_2\left(1+\dfrac{\alpha p_{i,U}^{\mathrm{tra}}}{d_{i,j}^2}\right)$表示无人机$i$到无人机$j$的数据传输速率。假设无人机之间的数据传输可以并行执行，那么无人机i通过U2U方式卸载并计算的时延为$T_{i,U}^t=\max\limits_{\forall j\in\mathcal{N},i\neq j}\left\{T_{i,j}^t\right\}$。

由于无人机可以同时通过U2U链路和U2B链路传输数据(即不同的传输模式)，并且可以同时执行数据传输和数据计算(即I/O和CPU可并行执行)，因此，数据处理时延取决于以上3种卸载方式所需时延的最大值。综上所述，用户设备k在时隙t内的数据的AoI为$T_i^t=\tau_k^t+\sum_{i=1}^{N}b_{i,k}^t\left(T_{i,\mathrm{rev}}^t+\max\left\{T_{i,i}^t,T_{i,\mathrm{B}}^t,T_{i,\mathrm{U}}^t\right\}\right)$。

最后，定义t时隙内目标感知区域的AoI为所有子区域中用户设备数据的AoI之和，即：$T^t=\sum_{k=1}^{M}T_k^t$。数据的信息年龄刻画了数据的新鲜程度，“年龄”越小意味着数据越新鲜。本文的一个重要目标是最小化整个目标感知区域的AoI，以保证每一个用户设备中的数据都能保持新鲜。

2.5 能耗模型

在每个时隙中，无人机需要完成两项任务：一是训练本地模型，即根据通信决策从覆盖的多个用户设备接收数据并执行卸载决策进行数据处理；二是根据移动决策飞往下一个目标子区域。在移动、数据接收、处理和传输的过程中，无人机会产生大量的能耗。本文对此过程产生的各种能耗进行定义：(1)无人机i从用户设备k_1飞行到k_2的移动能耗$E_{i,\mathrm{mov}}^t=\sum_{k_1=1}^{M}\sum_{k_2=1}^{M}p_i^{\mathrm{mov}}\dfrac{d_{k_1,k_2}}{v_i^{\mathrm{mov}}}o_{i,k_1}^{t-1}o_{i,k_2}^t$；(2)无人机$i$的数据接收能耗$E_{i,\mathrm{rev}}^t=p_i^{\mathrm{rev}}\dfrac{A_i^t+A_{i,\mathrm{rev}}^t}{P_i}$，其中$A_{i,\mathrm{rev}}^t=\sum_{j=1,j\neq i}^{N}A_j^t y_{j,i}^t l_{i,j}^t$表示无人机$i$通过U2U方式从其他无人机接收的数据量；(3)无人机i的计算能耗$E_{i,\mathrm{cmp}}^t=I_i^t(\epsilon)\,p_i^{\mathrm{cmp}}\dfrac{A_{i,i}^t+A_{i,\mathrm{rev}}^t}{c_i}$；(4)无人机$i$的数据传输能耗为$E_{i,\mathrm{tra}}^t=p_{i,U}^{\mathrm{tra}}\sum_{j=1}^{N}\dfrac{A_{i,j}^t}{v_{i,j}^t}+p_{i,B}^{\mathrm{tra}}\dfrac{A_{i,\mathrm{B}}^t}{v_{i,\mathrm{B}}^t}$。

最后，定义无人机i在时隙t时的总能耗为以上能耗之和，即：$E_i^t=E_{i,\mathrm{mov}}^t+E_{i,\mathrm{rev}}^t+E_{i,\mathrm{cmp}}^t+E_{i,\mathrm{tra}}^t$。本文另一个优化目标是最小化每个无人机的总能耗。

2.6 问题形式化

基于所构建的数学模型，本文希望在合理地规划无人机的飞行轨迹、智能地选择通信设备并分配联邦学习本地计算任务的前提下，找到一个可以长期最大限度地保持区域数据新鲜和模型的高预测准确率，同时最小化每个无人机能耗的解决方案。为此，将系统的优化目标表示为

$$\mathcal{P}1:\max\frac{1}{T}\sum_{t=1}^{T}\left[-T^t+\mu_1\mathrm{Acc}^t-\mu_2\sum_{i=1}^{N}E_i^t\right]$$
$$\text{s.t.式(1)—式(4)} \tag{5}$$

其中，权重因子μ_1和μ_2可以实现AoI、预测准确率和能耗的长期动态平衡。由式(5)可知，为了减少自身能耗，无人机偏向于在原地徘徊并做更少的通信决策；而为了保持区域数据长期新鲜，无人机会频繁移动以收集和处理用户设备的实时数据。但是，无人机频繁收集覆盖区域的用户数据将导致其通信时延和能耗的开销增大。此外，联邦学习模式基于收集到的任务数据进行多分类预测模型训练，以提升模型准确性为目标，却忽略了任务的实时性。但是在实际应用中，数据的实时性对于模型预测是十分重要的。如果基于过时的任务数据训练模型对新鲜的数据进行预测，那么其得到的预测性能将不理想。在本文中，多分类预测模型是通过联邦学习在多分类数据集上训练而得到的，模型的训练效果包括模型的准确性和模型的实时性。其中，模型的准确性是通过多分类预测任务的结果体现的，模型的实时性是由数据的新鲜程度决定的。

3 算法设计

本文所要解决的多无人机协作路径规划、通信决策和任务卸载决策问题属于复杂的离散变量和连续变量耦合的组合优化问题，采用传统的优化方法难以求解。因此，本文将该问题转化为马尔可夫决策问题，并设计基于深度强化学习的新型智能化优化算法来高效求解。

3.1 问题转化

本文采用马尔可夫决策过程来描述该协作式无人机智能决策问题，定义一个3元组$\mathcal{MDP}=<\mathcal{S},\mathcal{A},\mathcal{R}>$：

(1) $\boldsymbol{s}^t\in\mathcal{S}$表示时隙$t$时目标区域环境和每个无人机的状态，其中无人机$i$的状态为$\boldsymbol{s}_i=[k_i^t,\boldsymbol{\tau}^t,\boldsymbol{l}_i^t]$。$k_i^t$表示无人机$i$在时隙$t$时所在的位置，$\boldsymbol{\tau}^t=[\tau_1^t,\tau_2^t,\cdots,\tau_M^t]$表示用户设备$\mathcal{M}$在时隙$t$时的空闲时间，$\boldsymbol{l}_i^t=\left[l_{i,1}^t,l_{i,2}^t,\cdots,l_{i,N}^t\right]$表示无人机$i$在时隙$t$时是否可以向无人机$j$传输数据。

(2) $\boldsymbol{a}^t\in\mathcal{A}$表示时隙$t$时每个无人机执行的动作，其中无人机$i$的动作为$\boldsymbol{a}_i=[\theta_i^t,d_i^t,\boldsymbol{b}_i^t,x_i^t,\boldsymbol{y}_i^t,z_i^t]$。$\theta_i^t$，$d_i^t$分别表示无人机$i$在时隙$t$时的飞行方向、飞行距离，$\boldsymbol{b}_i^t=\left[b_{i,1}^t,b_{i,2}^t,\cdots,b_{i,M}^t\right]$表示无人机$i$在时隙$t$时是否与其覆盖区域内的用户设备$k$通信，$x_i^t$, $\boldsymbol{y}_i^t=\left[y_{i,1}^t,y_{i,2}^t,\cdots,y_{i,N}^t\right]$, z_i^t分别表示无人机i在时隙t时卸载决策为本地处理、U2U方式、U2B方式的数据比例。

(3) $\boldsymbol{r}^t\in\mathcal{R}$表示时隙$t$时无人机执行动作后获得的奖励，它包含全局奖励$r_g^t=-T^t+{}_1\mathrm{Acc}^t$，以及局部奖励$\boldsymbol{r}_l^t=[r_1^t,r_2^t,\cdots,r_N^t]$，其中$r_i^t=-E_i^t$表示无人机$i$在时隙$t$时能耗的负数形式。

全局奖励越大，表示t时隙内目标区域的数据越新鲜，且预测模型准确率更高；而局部奖励越大，表示t时隙内无人机移动、通信、计算和传输消耗的能量越少。因此，优化目标转化为

$$\mathcal{P}2:\max\frac{1}{T}\sum_{t=1}^{T}\left[r_g^t+\mu_2\sum_{i=1}^{N}r_i^t\right]$$
$$\text{s.t.(1)—(4)} \tag{6}$$

3.2 算法设计

传统的多智能DRL算法，如多智能体深度确定性策略梯度(Multi-Agent Deep Deterministic Policy Gradient, MADDPG)算法，通常优化单一的整体奖励。但是这可能会使学习过程在优化全局对象和局部对象之间来回波动，从而导致收敛不稳定、收敛速度缓慢等问题。因此，本文将奖励函数分解为全局奖励和局部奖励，其中局部奖励是每个智能体的本地优化目标，即减少无人机的能耗；而全局奖励是智能体群组的共同优化全局目标，即提高目标区域数据的AoI和模型预测准确率。为了实现全局优化目标和局部优化目标之间的动态平衡，本文引入可分解的多智能体深度确定性策略梯度(DEcomposed Multi-Agent Deep Deterministic Policy Gradient, DE-MADDPG)方法[13]。

DE-MADDPG是一种采用双critic网络的多智能体DRL算法，其目标是同时朝着使全局奖励和局部奖励最大化的方向优化策略。在给定当前状态s_i时，每个智能体i中的分布式actor网络可以生成动作$\boldsymbol{a}_i$。Actor网络使用确定性策略梯度方法进行参数更新，其梯度可以表示为

$$\begin{aligned}\nabla J(\vartheta_i)=&\underbrace{\mathbb{E}[\nabla_{\vartheta_i}\pi_i(a_i|s_i)\nabla_{a_i}Q_\psi^g(s,a_1,a_2,\cdots,a_N)]}_{\text{Global}}\\&+\underbrace{\mathbb{E}[\nabla_{\vartheta_i}\pi_i(a_i|s_i)\nabla_{a_i}Q_{\phi_i}^l(s_i,a_i)]}_{\text{Local}}\end{aligned} \tag{7}$$

DE-MADDPG引入局部critic网络$Q_{\phi i}^l$来扩展MADDPG算法，其中$Q_{\phi_i}^l$是无人机i的分布式Q网络，旨在最大化本地奖励，从而最小化损失函数以更新权重。全局critic网络Q_ψ^g的作用是指导全局优化，其参数通过最小化损失函数进行训练。此外，虽然actor网络是分布式地部署在无人机，它根据自身的状态决定动作，但是actor网络的优化过程利用了全局的状态信息和动作信息，实现了全局信息共享。

在本文的多无人机动态决策场景中，状态空间和动作空间规模随着无人机数量和目标区域规模的

增加而迅速增加。为了学习有价值的样本进而优化策略以加速DRL的收敛过程，本文进一步结合优先级经验回放机制[14]与DE-MADDPG方法，设计了基于优先级的可分解多智能体深度确定性策略梯度算法(Prioritized Decompose Multi-Agent Policy Gradient, PD-MADDPG)。缓存中的每个样本都有一个优先级，为其样本的TD误差。TD误差越大的样本，其估计值与目标值差距越大，网络使用此样本进行训练时可以更快提升性能。

3.3 算法实现

本文将训练一个共享预测模型的联邦学习作为系统的主要任务(表2)，在训练过程中调用PD-MADDPG 算法提供阶段性通信和卸载决策(表3)，并将预测模型训练结果反馈给PD-MADDPG算法进行优化。PD-MADDPG算法在每一轮联邦学习的本地迭代中在线为无人机提供执行联邦学习的相关决策，并且在每轮全局迭代后，都进行离线网络训练。无人机在探索时依据当前状态执行动作，计算全局奖励r_g^t和局部奖励r_l^t。以上离线训练过程结束后，将训练得到的多个actor网络模型部署到对应的无人机上再执行。

3.4 算法复杂性分析

在表3中，无人机的数量N将直接决定网络中的分支数量。因此，函数INIT( )、EXPLORATION( )和EXPLOITATION($\boldsymbol{I}^t(\epsilon)$, Acc^t)的计算复杂度分别为$O(N)$，$O(N)$和$O(N\cdot(n_l+n_a)+n_g)$。n_l，n_a和n_g分别表示局部critic, actor和全局critic网络的复杂度，与网络神经元个数和层数有关。在算法1中，系统执行$T^{\max}$次全局迭代过程，每个无人机进行$N^{\max}$次局部模型参数更新。因此，算法1的计算复杂度为$O(N+T^{\max}\cdot(N+N\cdot N^{\max}\cdot n_\omega+N\cdot(n_l+n_a)+n_g))$。$n_\omega$是预测模型的复杂度，与模型神经元个数和层数有关。

4 实验结果与分析

4.1 仿真实验设置

在仿真实验中，目标区域被划分为10 m×10 m的网格。无人机的数量为N=3，其覆盖半径为$R_i^{\max}$=1 m，其飞行高度为H=0.1 m，其最大飞行距离为 $l^{\max}$=10 m。通信带宽为W=100 MHz。另外，基站固定于目标区域外[−1, −1]的位置。在联邦学习的多轮迭代中，本文设置最大的全局迭代回合数为$T^{\max}$=400，最大的本地模型训练回合数为$N^{\max}$= 500。每一轮全局迭代回合包含100次全局模型的更新过程。在每个本地模型迭代时，本地模型的学习率为η=0.01，目标精度为ϵ=1.0。折扣因子=0.9，更新速率ξ=0.01。基于优先级的经验回放缓冲区PER的大小为64。在优化目标公式中，参数μ_1=1000，μ_2=0.1。

4.2 数据集与对比算法

本文采用3个真实的10分类数据集来进行仿真测试：(1) MNIST，由250个不同的人手写数字0, 1,…, 9构成；(2) Fashion-MNIST，由10个不同类别的28像素 × 28像素的灰度图像组成；(3) CIFAR-10，由10个物品类别的32×32的3通道彩色RGB图片组成。每个数据集中70%的数据用于训练分类预测模型，30%的数据用于测试其预测准确率。将训练集数据平均分配给每个用户设备，并设

表 2　联邦学习算法(算法1)

初始化最大全局模型训练回合数$T^{\max}$、最大本地模型训练回合数$N^{\max}$、学习率η、目标准确率ϵ和全局模型参数ω^0；

调用执行算法2函数INIT()；

for全局回合t=1,2,…,$T^{\max}$

　调用执行算法2的函数EXPLORATION()，获取无人机的决策；

　for无人机i=1,2,…,N

　　执行无人机与用户设备的通信，获取训练数据；

　　下载全局模型$\omega_i^{t,0}\leftarrow\omega^{t-1}$；

　　for局部回合n=1,2,…,$N^{\max}$

　　更新局部模型参数$\omega_i^{t,n}=\omega_i^{t,n-1}-\eta\nabla L_i\left(\omega_i^{t,n-1}\right)$；

　　if $\left\|\nabla L_i\left(\omega_i^{t,n}\right)\right\|\le\epsilon\left\|\nabla L_i\left(\omega_i^{t,n-1}\right)\right\|$ then

　　break；

　$I_i^t(\epsilon)=n$；

　无人机i上传局部模型$\omega_i^t(\omega_i^t\leftarrow\omega_i^{t,n})$；

进行全局模型聚合，调用执行算法2的函数EXPLOITATION($I^t(\epsilon)$, Acc^t)。

表 3 PD-MADDPG算法(算法2)

函数INIT( ):

for无人机i=1,2,$\cdots$,N

初始化局部critic、actor网络的权值为ϕ_i和ϑ_i、局部目标critic、actor网络的权值为$\phi_i^{'} \leftarrow \phi_i$和$\vartheta_i^{'} \leftarrow \vartheta_i$;

初始化全局critic网络的权值为ψ、全局目标critic 网络的权值为$\psi' \leftarrow \psi$、基于优先级的经验回放缓存PER。

函数EXPLORATION( ):

获得当前环境状态$\boldsymbol{s}^t=[\boldsymbol{s}_1^t,\boldsymbol{s}_2^t,\cdots,\boldsymbol{s}_N^t]$，当$t$= 0时随机初始化状态$\boldsymbol{s}^0$;

for无人机i=1,2,$\cdots$,N

while True

根据当前策略选择动作$\boldsymbol{a}_i^t=\pi_i\left(s_i^t\right)+\rho\mathcal{O}$，其中$\mathcal{O}$是高斯随机噪声，$\rho$随着$t$衰减;

if无人机i没有飞越边界，或与其它他人机的位置重合then

break;

return $a^t=[a_1^t,a_2^t,\cdots,a_N^t]$。

函数EXPLOITATION($I^t(\epsilon)$, Acc^t):

执行动作$\boldsymbol{a}^t=[\boldsymbol{a}_1^t,\boldsymbol{a}_2^t,\cdots,\boldsymbol{a}_N^t]$，获取新状态$s^{t+1}$，计算全局奖励$r_g^t$和局部奖励$r_l^t$;

将$[\boldsymbol{s}^t,\boldsymbol{a}^t,r_g^t,r_l^t,\boldsymbol{s}^{t+1}]$ 保存到PER;

if PER满then

从PER抽取一批样本$[\boldsymbol{s}^t,\boldsymbol{a}^t,r_g^t,\boldsymbol{s}^{t+1}]$;

更新全局critic 网络，根据$\psi' \leftarrow \xi\psi+(1-\xi)\psi'$更新全局目标critic网络，是更新速率;

for无人机i=1,2,$\cdots$,N

从PER抽取一批样本$[\boldsymbol{s}^t,\boldsymbol{a}^t,r_l^t,\boldsymbol{s}^{t+1}]$;

更新局部critic, actor网络，根据$\phi^{'} \leftarrow \xi\phi+(1-\xi)\phi^{'}$，$\vartheta' \leftarrow \xi\vartheta+(1-\xi)\vartheta'$更新目标局部网络。

置非独立同分布程度D来刻画每个用户设备数据的不同用户特性或者地理区域特性。$D=0$表示每个子区域的训练样本均匀地包含所有分类标签，$D\in(0,1)$表示所有数据均匀地属于D个标签，$D=1$表示每个子区域设备上的所有数据只属于一个标签。

本文使用4种优化整体奖励的算法进行对比实验：(1) P-MADDPG，将优先级经验回放缓存技术引入 MADDPG算法，所有无人机共用一个优先级缓存；(2) P-DDPG，将优先级经验回放缓存技术引入 DDPG算法，所有无人机分布式地训练各自的actor网络和critic网络，它们之间不共享信息，并且每个无人机上都设置分布式缓存；(3) GREEDY，列出每个时隙每个无人机所有可能的动作，在其中选择执行使整体奖励最优的动作(其搜索空间庞大和实现复杂度高，难以在实际应用中部署)；(4) RANDOM，每个无人机在每个时隙随机地产生动作，包括飞行方向、飞行距离、通信决策和卸载决策。

4.3 实验结果

4.3.1 基于联邦学习的预测模型效果分析

图2展示了本文提出的基于联邦学习的PD-MADDPG算法在不同的数据集和$D=[0, 0.5, 1, 2]$的预测准确率的表现。随着用户设备数据的非独立同分布程度的增加(从0到1)，预测准确率变差，并且收敛速度变慢。这是因为非独立同分布程度的增加导致每个用户设备中的数据标签种类变少。虽然某一种标签的样本数量会相对增加，但是多样性的降低会使得本地模型更加偏向于预测某几种标签的样本。对于全局模型而言，非独立同分布程度越大，本地模型就越发散。聚合发散的本地数据集会使模型性能变差，并使收敛回合数增加。只有当收集了足够多的样本标签后，全局模型的预测准确率才会逐步提高直到收敛。

表4展示了PD-MADDPG与4种对比算法在准确率性能上的差异。对不同数据集而言，所有算法的整体性能都会随着数据集的复杂度变大而变差，并且非独立同分布程度的增加会使模型预测准确率下降。其中，PD-MADDPG算法表现最优，预测准确率平均提升了16.3%，这是因为它将奖励分为全局奖励和局部奖励。P-MADDPG, P-DDPG和GREEDY算法整体优化预测准确率、数据AoI、无人机能耗，因此可能会导致优化目标失衡，即为了确保能耗而牺牲准确率。RANDOM算法的动作没有任何策略，其性能是最差的。

4.3.2　算法收敛性和可分解奖励分析

图3是在MNIST数据集中$D=2$时的算法奖励变化。在图3(a)中，除GREEDY算法之外，PD-MADDPG算法比其余算法的平均总奖励高48.4%，比基于DRL的算法高38.7%。它分别优化全局奖励和局部奖励，因此两者的性能都是最优的，即它能找到最合适的移动、通信和卸载决策，使得奖励在优化无人机能耗、数据AoI及预测准确率之间找到较好的平衡。当设置较小的局部奖励权重μ_2时，总奖励的收敛性主要受全局奖励的影响，因此两者的收敛性非常相似，如图3(b)所示。在该设置中，GREEDY算法偏向选择使全局奖励更大的动作。RANDOM算法中无人机会任意地移动并通信，因此数据新鲜程度普遍较高，全局奖励较高。PD-MADDPG算法通过全局critic网络来优化全局奖励，使得无人机执行有利于维持数据新鲜程度和预测模型准确率的动作，它比基于DRL的算法的平均全局奖励高37.1%。在图3(c)中，PD-MADDPG算法是最优的，因为无人机分布式actor网络的优化同时受全局critic网络和局部critic网络的影响，并且无人机之间是通过相互协作来进行决策的。它比所有算法的平均局部奖励高66.2%，比基于DRL的算法高48.3%。

4.3.3　基于实时联邦学习的协作式无人机计算系统的规模可扩展性分析

当无人机数量和通信范围不变时，以MNIST数据集中$D=0.5$为例，本文绘制了目标区域边长为5, 10, 15和20时，各算法在收敛后100个回合内

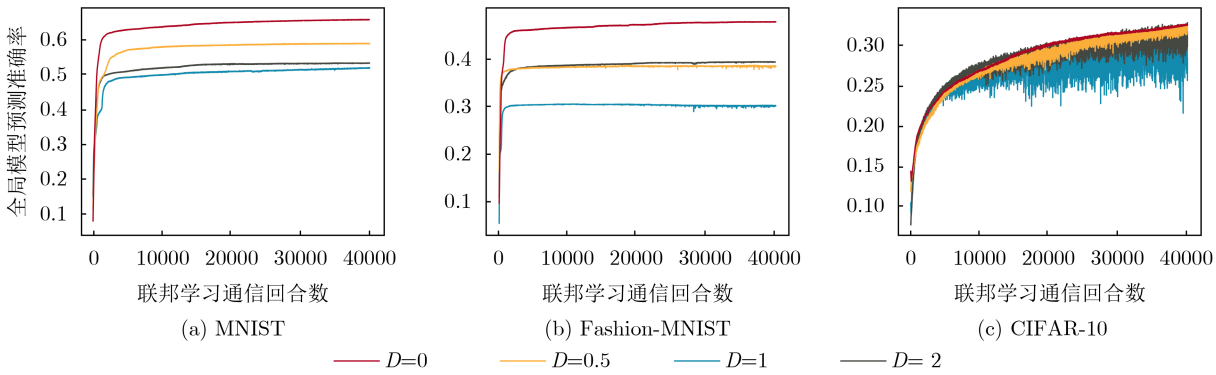

图 2 PD-MADDPG 算法的全局模型预测准确率

表 4　各算法在不同数据集的不同非独立同分布程度时的全局模型预测准确率

算法	MNIST				Fashion-MNIST				CIFAR-10			
	$D=0$	$D=0.5$	$D=1$	$D=2$	$D=0$	$D=0.5$	$D=1$	$D=2$	$D=0$	$D=0.5$	$D=1$	$D=2$
PD-MADDPG	0.661	0.590	0.519	0.533	0.451	0.371	0.288	0.379	0.343	0.338	0.336	0.328
P-MADDPG	0.619	0.384	0.483	0.442	0.361	0.370	0.190	0.367	0.324	0.305	0.302	0.312
P-DDPG	0.558	0.450	0.453	0.521	0.435	0.336	0.261	0.359	0.323	0.320	0.306	0.311
GREEDY	0.561	0.544	0.487	0.515	0.400	0.344	0.282	0.358	0.320	0.309	0.325	0.320
RANDOM	0.479	0.278	0.463	0.407	0.291	0.292	0.194	0.289	0.322	0.330	0.326	0.319

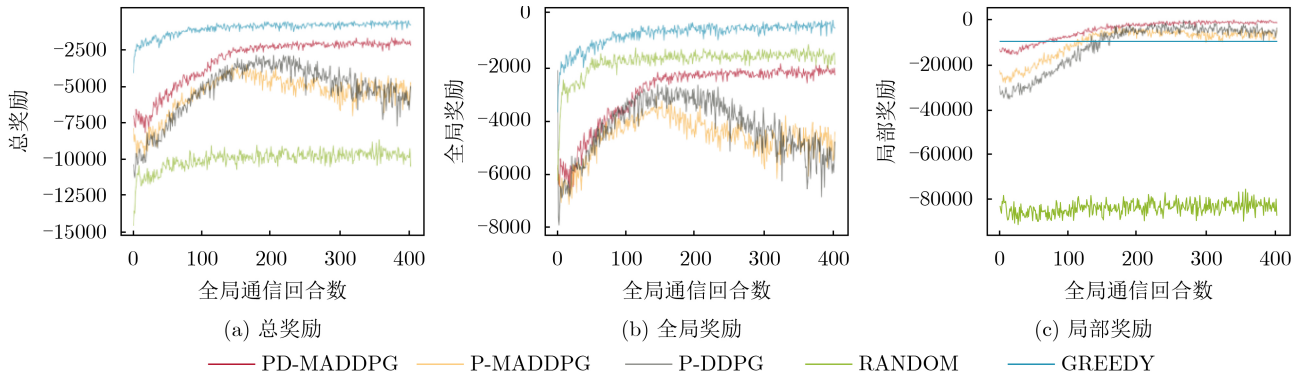

图 3 在 MNIST 数据集中$D=2$时各算法奖励的变化

平均总奖励的变化，如图4所示。当区域规模增大时，无人机服务的用户设备增多，在保证预测准确率和区域数据新鲜度的前提下，它的移动能耗增加，因此所有算法的平均总奖励都减少。其中，PD-MADDPG算法的平均总奖励的下降速率最慢，比所有算法的下降速率慢38.6%，比基于DRL的算法的下降速率慢23.5%。这说明本文提出的算法受区域变化的影响程度最小，无人机能找到更合适的协作方式，在保证其他优化目标的前提下减少无人机的移动能耗，因此其可扩展性是最好的。

GREEDY算法在每次迭代中遍历所有可能的动作从而执行使整体奖励最优的决策。但是这将产生额外的运行能耗，而该能耗是算法运行代价。本文在能耗建模时更关注多无人机执行决策时产生的通信开销，因此没有在优化目标中考虑算法运行能耗，而是最小化无人机能耗。具体地，GREEDY算法的计算复杂度为$O\left(a^{N\cdot M}\right)$，其中$a$是问题的动作空间。一次迭代中，GREEDY 算法耗时202.04 s, PD-MADDPG算法耗时20.13 s，这说明GREEDY算法的时间复杂度比PD-MADDPG算法的高约10倍。由图4可知，仅当目标区域规模增加时，GREEDY算法的平均总奖励下降得比PD-MADDPG算法快，目标值之间的差距逐渐加大。随着动作变量和空间规模变大，GREEDY算法的复杂度呈指数级增加，因此其可扩展性是最差的。

5 结束语

本文主要研究了在实时边缘数据处理场景中，以无人机作为边缘服务器，通过智能地进行轨迹规划、通信决策和卸载决策来实现模型预测高准确率、高数据新鲜程度和低无人机能耗的优化问题。考虑到用户设备数据的实时性、隐私性和规模有限性，本文引入联邦学习在无人机上执行本地训练，然后聚合为全局模型，通过多轮迭代获得共享的预测模型。为了解决该多目标优化问题，本文设计了一种全局奖励和局部奖励融合的多智能体深度强化学习的算法，动态地进行多无人机的轨迹规划以及任务卸载和通信决策。最后，大量的仿真实验结果表明本文的PD-MADDPG算法的优越性，验证了所设计的系统和算法的合理性、有效性和可拓展性。

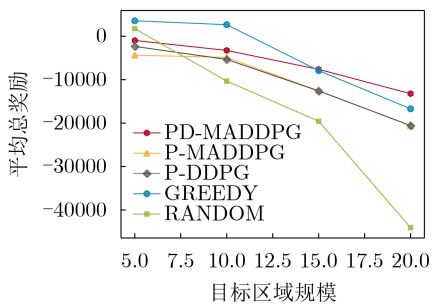

图 4 平均总奖励随目标区域规模的变化

参 考 文 献

[1] PHAM Q V, FANG Fang, HA V N, *et al*. A survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art[J]. *IEEE Access*, 2020, 8: 116974–117017. doi: 10.1109/ACCESS.2020.3001277.

[2] LIM W Y B, LUONG N C, HOANG D T, *et al*. Federated learning in mobile edge networks: A comprehensive survey[J]. *IEEE Communications Surveys & Tutorials*, 2020, 22(3): 2031–2063. doi: 10.1109/COMST.2020.2986024.

[3] WANG Jingrong, LIU Kaiyang, and PAN Jianping. Online UAV-mounted edge server dispatching for mobile-to-mobile edge computing[J]. *IEEE Internet of Things Journal*, 2020, 7(2): 1375–1386. doi: 10.1109/JIOT.2019.2954798.

[4] BRIK B, KSENTINI A, and BOUAZIZ M. Federated learning for UAVs-enabled wireless networks: Use cases, challenges, and open problems[J]. *IEEE Access*, 2020, 8: 53841–53849. doi: 10.1109/ACCESS.2020.2981430.

[5] JEONG S, SIMEONE O, and KANG J. Mobile edge computing via a UAV-mounted cloudlet: Optimization of bit allocation and path planning[J]. *IEEE Transactions on Vehicular Technology*, 2018, 67(3): 2049–2063. doi: 10.1109/TVT.2017.2706308.

[6] MAO Yuyi, YOU Changsheng, ZHANG Jun, *et al*. A survey on mobile edge computing: The communication perspective[J]. *IEEE Communications Surveys & Tutorials*, 2017, 19(4): 2322–2358. doi: 10.1109/COMST.2017.2745201.

[7] POUYANFAR S, SADIQ S, YAN Yilin, *et al*. A survey on deep learning: Algorithms, techniques, and applications[J]. *ACM Computing Surveys*, 2019, 51(5): 92. doi: 10.1145/3234150.

[8] SUN Yin, UYSAL-BIYIKOGLU E, YATES R D, *et al*. Update or wait: How to keep your data fresh[J]. *IEEE Transactions on Information Theory*, 2017, 63(11): 7492–7508. doi: 10.1109/TIT.2017.2735804.

[9] DAI Zipeng, LIU C H, HAN Rui, *et al*. Delay-sensitive energy-efficient UAV crowdsensing by deep reinforcement learning[J]. *IEEE Transactions on Mobile Computing*, To be published. doi: 10.1109/TMC.2021.3113052..

[10] KAUL S, YATES R, and GRUTESER M. Real-time status: How often should one update?[C]. 2012 Proceedings IEEE INFOCOM, Orlando, USA, 2012: 2731–273. doi: 10.1109/

INFCOM.2012.6195689.

[11] LUO Siqi, CHEN Xu, WU Qiong, *et al*. HFEL: Joint edge association and resource allocation for cost-efficient hierarchical federated edge learning[J]. *IEEE Transactions on Wireless Communications*, 2020, 19(10): 6535–6548. doi: 10.1109/TWC.2020.3003744.

[12] WANG Liang, WANG Kezhi, PAN Cunhua, *et al*. Deep reinforcement learning based dynamic trajectory control for UAV-assisted mobile edge computing[J]. *IEEE Transactions on Mobile Computing*, To be published. doi: 10.1109/TMC.2021.3059691.

[13] SHEIKH H U and BŐLŐNI L. Multi-agent reinforcement learning for problems with combined individual and team reward[C]. 2020 International Joint Conference on Neural Networks, Glasgow, UK, 2020: 1–8. doi: 10.1109/IJCNN48605.2020.9206879.

[14] CAO Xi, WAN Huaiyu, LIN Youfang, *et al*. High-value prioritized experience replay for off-policy reinforcement learning[C]. The 31st International Conference on Tools with Artificial Intelligence, Portland, USA, 2019: 1510–1514. doi: 10.1109/ICTAI.2019.00215.

范　文：女，硕士，研究方向为移动边缘计算、联邦学习、无人机等.

韦　茜：女，硕士生，研究方向为移动边缘计算、无人机等.

周　知：男，副教授，研究方向为联邦学习系统、分布式机器学习系统、边缘智能等.

于　帅：男，副教授，研究方向为边缘计算、联邦学习、深度强化学习等.

陈　旭：男，教授，研究方向为边缘计算与云计算、分布式人工智能、联邦学习等.

责任编辑：马秀强