无人机辅助MEC车辆任务卸载与功率控制近端策略优化算法

谭国平* 易文雄 周思源 胡鹤轩
(河海大学计算机与信息学院 南京 211100)

摘 要：无人机(UAVs)辅助移动边缘计算(MEC)架构是灵活处理车载计算密集、时延敏感型任务的有效模式。但是，如何在处理任务时延与能耗之间达到最佳均衡，一直是此类车联网应用中长期存在的挑战性问题。为了解决该问题，该文基于无人机辅助移动边缘计算架构，考虑无线信道时变特性及车辆高移动性等动态变化特征，构建出基于非正交多址(NOMA)的车载任务卸载与功率控制优化问题模型，然后将该问题建模成马尔可夫决策过程，并提出一种基于近端策略优化(PPO)的分布式深度强化学习算法，使得车辆只需根据自身获取局部信息，自主决策任务卸载量及相关发射功率，从而达到时延与能耗的最佳均衡性能。仿真结果表明，与现有方法相比较，本文所提任务卸载与功率控制近端策略优化方案不仅能够显著获得更优的时延与能耗性能，所提方案平均系统代价性能提升至少13%以上，而且提供一种性能均衡优化方法，能够通过调节用户偏好权重因子，达到系统时延与能耗水平之间的最佳均衡。

关键词：无人机辅助计算；移动边缘计算；近端策略优化；深度强化学习；功率控制和任务卸载
中图分类号：TN929.5 **文献标识码**：A **文章编号**：1009-5896(2024)00-0001-11
DOI: 10.11999/JEIT230770

Proximal Policy Optimization Algorithm for UAV-assisted MEC Vehicle Task Offloading and Power Control

TAN Guoping Yi Wenxiong ZHOU Siyuan HU Hexuan
(*College of Computer and Information*, *Hohai University*, *Nanjing* 211100, *China*)

Abstract: The architecture of Mobile Edge Computing (MEC), assisted by Unmanned Aerial Vehicles (UAVs), is an efficient model for flexible management of mobile computing-intensive and delay-sensitive tasks. Nevertheless, achieving an optimal balance between task latency and energy consumption during task processing has been a challenging issue in vehicular communication applications. To tackle this problem, this paper introduces a model for optimizing task offloading and power control in vehicle networks based on UAV-assisted mobile edge computing architecture, using a Non-Orthogonal Multiple Access (NOMA) approach. The proposed model takes into account dynamic factors like vehicle high mobility and wireless channel time-variations. The problem is modeled as a Markov decision process. A distributed deep reinforcement learning algorithm based on Proximal Policy Optimization (PPO) is proposed, enabling each vehicle to make autonomous decisions on task offloading and related transmission power based on its own perceptual local information. This achieves the optimal balance between task latency and energy consumption. Simulation results reveal that the proposed proximal policy optimization algorithm for task offloading and power control scheme improves not only the performance of task latency and energy consumption compared to existing methods, The average system cost performance improvement is at least 13% or more. but also offers a performance-balanced optimization method. This method achieves optimal balance between the system task latency and energy consumption level by adjusting user preference weight factors.

Key words: Unmanned Aerial Vehicles (UAVs) Assisted Computing; Mobile Edge Computing (MEC); Proximal Policy Optimization(PPO); Deep reinforcement learning; Power Control and Task Offloading

收稿日期：2023-07-28；改回日期：2024-01-05；网络出版：2024-01-28
*通信作者： 谭国平 gptan@hhu.edu.cn
基金项目：国家自然科学基金(61832005, U21B2016)
Foundation Items: The National Natural Science Foundation of China (61832005, U21B2016)

1 引言

车载应用大多是延迟敏感、计算密集型任务，仅靠车辆自身算力难以保证其低时延与低能耗要求。近年来，许多研究显示通过移动边缘计算(Mobile Edge Computing, MEC)实现车辆任务卸载，是有效保障车载应用低能耗和低时延通信的方法[1]。特别的，Beyond-5G 与 6G愿景已提出通过在无人机(Unmanned Aerial Vehicle, UAV)上部署MEC系统，提供空中边缘计算服务[2]，且组成的地-空网络既可以实现应急网络覆盖，也可以在用户密集区域实现低成本任务卸载，当出现地面基站临时故障或当网络覆盖范围内用户过于密集时，无人机可方便快捷的提供高质量的网络覆盖，以保障优质的用户体验。考虑车联网中计算任务特性及车辆移动性特点，该类任务需要高可靠的网络质量且某一区域计算任务量随时间变化较大，采用无人机辅助边缘计算是一种高效、低成本的方案。在现有研究中，已有较多针对车联网场景的任务卸载方案：文献[3]提出一种基于策略梯度的分布式深度强化学习方法，寻求在MEC辅助下获得最优卸载性能；文献[4]以系统能耗最小化为目标，研究了面向空天地异构网络的任务卸载问题；文献[5]考虑了通信和计算资源联合分配方法，以降低可靠性和时延总系统代价为目标，提出了一种多目标强化学习策略；文献[6]提出一种集中式多智能体算法，对多无人机辅助MEC系统下资源与功率分配进行了联合优化。然而，该类方法没有考虑车辆高移动性带来的高动态环境问题，如信道随机变化如何影响方案的长期性能。因此，这类方法不适合直接应用于车辆任务卸载方案。文献[7]则研究了一种新的多层服务架构下任务卸载服务场景，文献[8]全面讨论了车载系统的协同通信和计算问题。为提高移动通信资源利用率，研究证明非正交多址(Non-Orthogonal Multiple Access, NOMA)技术是有效方法之一[9]。相较于正交多址(Orthogonal Multiple Access, OMA)系统，NOMA在蜂窝边缘吞吐量、信道反馈松弛度以及传输时延等方面具有优越性。在NOMA系统中，支持的用户数量并未受到正交时频资源的严格限制，因此在资源不足的情况下，NOMA能够显著增加同时连接的用户数量。传统OMA系统中依赖于访问授权请求进行资源分配，然而，在部分NOMA的上行链路中，并不需要动态调度。这显著减少了传输的延迟和信令开销。因此，很多研究者提出了基于NOMA的无人机辅助资源分配与任务卸载方案：文献[10] 提出一种基于NOMA的新颖资源分配方案，该方案使用集中式控制方法进行资源分配，但使用分布式方法进行功率控制；文献[11]考虑平均数据吞吐量优化问题提出最优无人机边缘计算卸载方案。文献[12]则以最小化网络总能耗为目标，对无人机飞行轨迹优化与资源分配优化等问题进行了研究；文献[13]则基于相控阵天线接收信号提出一种能效优化方案；基于无人机辅助全双工技术，文献[14] 通过优化无人机布局及功率分配实现了最大化NOMA系统吞吐量；考虑能量收集时间、无人机位置等因素，文献[15]提出一种无人机辅助速率最大化资源分配算法；以服务质量感知为目标，文献[16]提出一种基于NOMA的资源分配方案，通过考虑计算任务资源需求、车辆与基站间距离等因素，降低了系统时延与能耗；文献[17,18]则提出一种基于NOMA的新颖调制编码技术，文献[19]则分析了其在车联网场景下的通信容量。

从上述研究可知，基于NOMA的无人机辅助任务卸载技术在车联网系统具有很好的应用前景，然而，由于车辆高移动性、无线信道不确定性、任务到达随机性等多种因素导致任务卸载优化问题难以求解。已有研究主要采用使用确定性方法寻求次优解，且复杂度较高，难以应用于随机动态变化的车联网场景，针对该问题，本文主要工作为采用无人机辅助MEC架构，构建基于NOMA的车联网任务卸载与功率控制优化问题模型，并在模型中融入信道时变特性和车辆高移动性等动态变化特征，然后将该问题建模成马尔可夫决策过程，随后，基于近端策略优化的深度强化学习算法，研究一种分布式功率控制方法，使得车辆能够只需根据自身获取局部信息，自主决策任务卸载所需要发射功率及本地计算所需功率，从而达到时延与能耗的最佳均衡性能。

2 系统模型

本文研究构建的无人机辅助车辆MEC系统模型如图1所示，不失一般性，该模型由单基站、单无人机和多辆车组成，显然，通过该模型的组合易于推广到多基站和多无人机场景。

现假设无人机辅助MEC系统部署于城市道路场景，其中，基站位于道路两侧，且配置边缘服务器为道路覆盖范围内车辆提供计算服务；无人机位于道路上方，以固定高度、固定飞行速度沿道路方向飞行，且配置边缘服务器为道路覆盖范围内车辆提供计算服务。本文对车辆高速移动特征采用类似文献[20]的方法建模：道路上多辆车以一定速度行驶，且车辆间距遵循4秒时距原则，即车与车之间的时距为4 s。不失一般性，本文假设车辆与基站、无人机的通信方式采用NOMA技术。

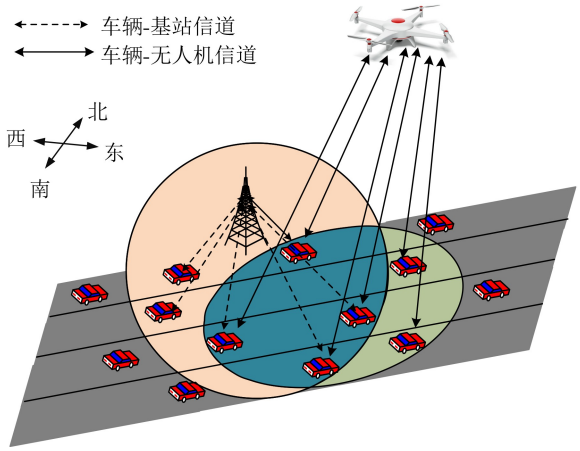


图 1 无人机辅助车辆MEC系统模型

假设车辆任务到达速率服从高斯分布。计算任务具有数据规格大且变化大的特点。当任务生成后先缓存至车辆缓存器，任务服务顺序服从先到先服务原则。因车内有大量计算密集型、时延敏感型任务需要及时处理，但车辆自身计算能力通常存在严重不足，必须将部分任务卸载至路侧基站或无人机上进行计算，然后将结果反馈给车辆。在传统集中式调度中，计算任务量需求随着车辆数量的增长而增长，而分布式架构将计算本地化，不会引起计算量膨胀化问题。集中式调度极度依赖于控制中心。存在控制中心宕机而方案失效情况，而分布式架构感知、计算、决策由用户实施，减小用户间、用户与控制中心间的耦合程度，提升方案的可用性。分布式架构不需要共享用户状态信息，可以有效保护隐私。不同于其他集中式调度方案，本文假设车辆无需全局状态信息，仅仅只需根据自身获取局部状态信息，便能够自主智能决策选择在本地、基站边缘服务器或无人机边缘服务器进行任务计算，且实现最佳功率分配，即能够以分布式的方式独立自主进行功率控制和任务卸载，实现时延与能耗的最佳均衡性能。本文主要符号定义如表1所示。

2.1 通信模型

本文采用NOMA信道模型，基站或无人机上收到的矢量信号$\boldsymbol{y}(t)$表示为

$$\boldsymbol{y}(t)=\sum_{i\in M}\boldsymbol{h}_i(t)\sqrt{p_i^o(t)}s_i(t)+\boldsymbol{n}(t) \tag{1}$$

其中$\boldsymbol{h}_i(t)\in\mathbb{C}^{N\times 1}$表示第$i$辆车与基站或无人机通信的信道状态信息，所有车辆信道在t时刻的信道状态信息表示为：$\boldsymbol{H}(t)=[\boldsymbol{h}_1(t),\boldsymbol{h}_2(t),\cdots,\boldsymbol{h}_M(t)]\in\mathbb{C}^{N\times M}$，$p_i^o(t)$为第$i$辆车在$t$时刻将任务卸载至无人机或基站所使用的传输功率。$\boldsymbol{n}(t)$为加性高斯白噪声，其均值为0，方差为$\sigma_R^2$，即$\boldsymbol{n}(t)\sim\mathcal{CN}(0,\sigma_R^2\boldsymbol{I}_N)$，$s_i(t)$为车辆发送至基站或无人机的信号。

表 1 缩略语表

符号	符号含义
σ_R^2	加性高斯白噪声方差
$\boldsymbol{n}(t)$	加性高斯白噪声矢量
$h_i^p(t)$	车辆i在t时刻的大尺度衰落
$\boldsymbol{h}_i^s(t)$	车辆i在t时刻的小尺度衰落矢量
h_r	单位距离信道功率增益
$\boldsymbol{p}_{\mathrm{BS}}$	基站位置
$\boldsymbol{p}_{\mathrm{UAV}}(t)$	t时刻无人机位置
$p_i(t)$	第i辆车在t时刻的位置
$d_i^o(t)$	第i辆车在t时刻与基站或无人机之间传输速率
$d_i^l(t)$	第i辆车在t时刻的本地计算量
$E_i^l(t)$	第i辆车在t时刻的本地能耗
$E_i^o(t)$	第i辆车在t时刻为任务卸载而使用的传输功率所消耗的能量
$q_n(\theta)$	重要性采样比率
λ	折扣系数
$E_i^o(t)$	网络损失函数
$\hat{A}_n^{\mathrm{GAE}(\gamma,\phi)}$	广义估计函数

在式(1)中，$\boldsymbol{h}_i(t)$由大尺度衰落$\boldsymbol{h}^{\mathrm{p}}(t)$和小尺度衰落矢量$\boldsymbol{h}_i^{\mathrm{s}}(t)$组成，表达式为

$$\boldsymbol{h}_i(t)=\boldsymbol{h}_i^s(t)\sqrt{h_i^p(t)} \tag{2}$$

在t时刻第i辆车至基站的信道大尺度衰落$h_{i,\mathrm{BS}}^p(t)$由路径损耗表示[21]，其表达形式为

$$h_{i,\mathrm{BS}}^p(t)=\frac{h_{\mathrm{r}}}{||\boldsymbol{P}_i(t)-\boldsymbol{P}_{\mathrm{BS}}||^{\eta}} \tag{3}$$

其中η为路径损耗系数，$\boldsymbol{P}_i(t)$为在t时刻第i辆车的位置，$\boldsymbol{P}_{\mathrm{BS}}$为基站位置，$h_{\mathrm{r}}$为单位距离信道功率增益。

第i辆车至无人机信道大尺度衰落$h_{i,\mathrm{UAV}}^p(t)$表达式为[22]

$$h_{i,\mathrm{UAV}}^p(t)=10^{-(30.9+(22.25-0.5\lg Y_{UAV})\lg d_{i,UAV}(t)+20\lg f_c)/10} \tag{4}$$

其中Y_{UAV}为无人机飞行高度，$d_{i,\mathrm{UAV}}(t)$为在t时刻第i辆车与无人机的距离，f_{c}为所使用的载波频率。

车辆至基站或无人机的信道小尺度衰落$\boldsymbol{h}_i^s(t)$由一阶自回归模型描述[23]，其表达式为

$$\boldsymbol{h}_i^s(t)=\rho_i\boldsymbol{h}_i^s(t-1)+\sqrt{1-\rho_i^2}\boldsymbol{e}(t) \tag{5}$$

其中ρ_i为第i辆车的连续时隙间归一化信道相关系数，$\boldsymbol{e}(t)$为服从复高斯分布的误差矢量。

现用$\boldsymbol{H}^{\dagger}(t)$表示$\boldsymbol{H}(t)$的伪逆矩，基站或无人机使用$\boldsymbol{H}^{\dagger}(t)$作为迫零解码器解码来自第$i$辆车的接收信号$\boldsymbol{y}(t)$，$\boldsymbol{H}^{\dagger}(t)$计算为

$$\boldsymbol{H}^{\dagger}(t)=\left(\boldsymbol{H}^{\mathrm{H}}(t)\boldsymbol{H}(t)\right)^{-1}\boldsymbol{H}^{\mathrm{H}}(t)\in\mathbb{C}^{M\times N} \tag{6}$$

其中$\boldsymbol{H}^{\mathrm{H}}(t)$为$\boldsymbol{H}(t)$的转置矩阵。

现用$\boldsymbol{g}_i^{\mathrm{H}}(t)$表示$\boldsymbol{H}^{\dagger}(t)$的第$i$行，则基站或无人机解码信号可以通过下式计算获得[24]

$$\boldsymbol{g}_i^{\mathrm{H}}(t)y(t)=\boldsymbol{g}_i^{\mathrm{H}}(t)\sum_{i\in M}\boldsymbol{h}_i(t)\sqrt{p_i^o(t)}s_i(t)+\boldsymbol{g}_i^{\mathrm{H}}(t)\boldsymbol{n}(t) \tag{7}$$

根据以上公式可得

$$\boldsymbol{g}_i^{\mathrm{H}}(t)\boldsymbol{h}_i(t)=\delta_{i,m}(t)=\begin{cases}1, & m=i\\ 0, & m\neq i\end{cases} \tag{8}$$

因此，式(7)又可表示为

$$\boldsymbol{g}_i^{\mathrm{H}}(t)\boldsymbol{y}(t)=\sqrt{p_i^o(t)}s_i(t)+\boldsymbol{g}_i^{\mathrm{H}}(t)\boldsymbol{n}(t) \tag{9}$$

根据式(9)可看出，当基站使用迫零解码器对接收信号进行解码后，$\sqrt{p_i^o(t)}s_i(t)$为有用接收信号，$\boldsymbol{g}_i^{\mathrm{H}}(t)\boldsymbol{n}(t)$为干扰噪声信号，因此，第$i$辆车的SINR可表示为

$$\gamma_i(t)=\frac{p_i^o(t)}{||\boldsymbol{g}_i^{\mathrm{H}}(t)||^2\sigma_R^2} \tag{10}$$

其中$\|\cdot\|$为L2范数。

在t时刻，第i辆车与基站或无人机之间单时隙传输数据量可利用香农公式计算为

$$d_i^o(t)=\tau_0 B\log_2(1+\gamma_i(t)) \tag{11}$$

其中，τ_0为单个时隙持续时间，B为信道带宽。

2.2 移动模型

本文采用笛卡尔坐标系对车辆瞬时位置进行建模，以场景左下角原点，车辆移动方向为X轴正方向，即由西往东，Y轴正方向为由南至北。因车辆变速行驶且随机变换车道的场景难以建模分析，为了简化问题，本文假设较短时间内车辆均以恒定速度在恒定车道上行驶，则第i辆车位置为

$$P_i(t)=(x_i(t),Y_i,0) \tag{12}$$

其中，Y_i为第i辆车在Y轴上的位置，注意本文假设车辆在恒定车道上行驶，因此Y_i为常量。$x_i(t)$为t时刻第i辆车在X轴上位置，其计算方式为

$$x_i(t)=x_i(t-1)+v_i\tau_0 \tag{13}$$

其中，v_i表示第i辆车行驶速度，注意本文假设车辆在道路上匀速行驶，因此v_i也为常量。

假设无人机在空中以恒定速度、恒定高度沿着道路方向飞行，此时无人机的位置表示为：

$$P_{\mathrm{UAV}}(t)=(x_{\mathrm{UAV}}(t),Y_{\mathrm{UAV}},Z_{\mathrm{UAV}}) \tag{14}$$

其中，Y_{UAV}为无人机在Y轴上的位置，Z_{UAV}为无人机在Z轴上的位置，由于假设无人机沿道路方向固定高度匀速飞行，则Y_{UAV}和Z_{UAV}也为常量。在t时刻无人机在X轴上的位置计算为

$$x_{\mathrm{UAV}}(t)=x_{\mathrm{UAV}}(t-1)+v_{\mathrm{UAV}}\tau_0 \tag{15}$$

其中，v_{UAV}为无人机飞行速度，由于假设车辆在空中匀速飞行，所以该值也为常量。

本文假设基站部署在路侧，其位置可表示为：

$$\boldsymbol{P}_{BS}=(X_{\mathrm{BS}},Y_{\mathrm{BS}},Z_{\mathrm{BS}}) \tag{16}$$

其中，X_{BS}，Y_{BS}，Z_{BS}分别为基站在X, Y, Z轴上的位置，由于基站位置固定，所以三者都为常量。

2.3 计算模型

本文假设车辆可以自主选择任务在本地或者边缘服务器上进行计算，当选择使用本地进行任务计算时，本地计算速率与车辆分配给车辆的CPU频率有关，则在t时刻第i辆车的本地计算量为

$$d_i^l(t)=\tau_0 f_i(t)/\varsigma \tag{17}$$

其中，$f_i(t)$为在t时刻第i辆车在本地进行任务计算所分配的CPU频率，ς为车辆计算1bit数据时所消耗的CPU周期数，因此，在t时刻第i辆车的本地能耗为[25]

$$E_i^l(t)=\left(f_i(t)\right)^3\mu\tau_0 \tag{18}$$

其中，μ为有效电容系数，该值取决于CPU硬件结构与质量等因素。

当车辆选择将任务卸载至基站或者无人机上进行计算时，其传输速率可用式(11)进行计算，本文假设车辆选择卸载任务量时考虑了当前边缘服务器剩余计算资源，以适当的选择卸载任务量以满足其时延预算。因回程流量通常较小，所以本文忽略计算时延与数据回传导致的传输时延。因此，在t时刻，第i辆车为任务卸载在一个时隙内消耗的能量计算为

$$E_i^o=p_i^o(t)\tau_0 \tag{19}$$

在t时刻，注意第i辆车中缓存器中剩余任务量与上一时刻剩余任务、当前时刻产生任务量、当前时刻本地计算和卸载至边缘服务器的任务量有关，可通过式(20)计算

$$T_i(t)=\left[T_i(t-1)+a_i(t)-\left(d_i^o(t)+d_i^l(t)\right)\right]^+ \tag{20}$$

其中，$a_i(t)$为在t时刻第i辆车产生的任务量，$[\cdot]^+=\max(0,\cdot)$。根据Little's定理[26]，注意任务时延与缓存器中任务量成正比，即缓存器中任务量越长，时延越大。

最后，在t时刻，部署在基站或无人机上的边缘服务器剩余容量可通过式(21)计算

$$C(t)=\left[C(t-1)-\sum_{i=1}^{M}d_i^o(t)\right]^* \tag{21}$$

2.4 优化问题建模

本文假设车辆行驶过程中进行任务计算时，部署在车上的调度程序需要决定卸载任务的时机、卸载至边缘服务器计算的任务数量、以及分配到本地计算的任务数量。本文使用符号$T_i(t)$表示缓存器剩余任务量，即代表t时刻缓存器中需要计算的任务量。注意$T_i(t)$越大则表示时延越大。为了在时延与能耗之间取得最佳权衡的性能，本文优化目标系统代价函数定义为

$$\mathrm{cost}_i(t)=(1-\zeta)T_i(t)+\zeta\left(E_i^{\mathrm{BS}}(t)+E_i^{\mathrm{UAV}}(t)+E_i^{L}(t)\right) \tag{22}$$

其中$E_i^{\mathrm{BS}}(t)$, $E_i^{\mathrm{UAV}}(t)$, $E_i^{L}(t)$分别表示t时刻第i辆车将任务卸载至基站、无人机、本地计算所消耗的能量。参数ζ为权衡因子，其取值范围为$[0,1]$，其取值反映了能耗与时延的偏好选择。借助系统代价函数定义，本文优化目标为找到最佳计算任务卸载调度策略，最小化长期平均系统代价，即：

$$\min\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathrm{cost}_i(t) \tag{23}$$

因信道时变性、车辆运动随机性与任务生成随机性等严重不确定性因素，使得上述优化问题难以建立精确的数学模型进行求解，因此难以使用传统凸优化等确定性优化工具进行求解，为此，本文试图采用基于深度强化学习的方法，通过设计分布式的功率控制与任务卸载策略来实现上述优化目标。

2.4.1 状态空间设计

为了便于设计易于实现的分布式优化算法，本文假设在每个时隙，所有车辆仅根据自身获取到的局部信息做出决策，无须获取全局的状态信息。即，车辆获取到局部状态后，根据当前策略做出决策，车辆获得本次动作奖励，然后转移到下一状态。本文假设车辆能够连续感知信道状态，车辆将根据车辆中剩余任务量、当前时刻基站/无人机信道状态、边缘服务器中剩余计算容量实时做出决策，在t时刻第i辆车的状态空间表设计为

$$s_{i,t}=\left[T_i(t),g_{i,\mathrm{BS}}^{\mathrm{H}}(t),g_{i,\mathrm{UAV}}^{\mathrm{H}}(t),C_{\mathrm{BS}}(t),C_{\mathrm{UAV}}(t)\right] \tag{24}$$

其中，$T_i(t)$表示在t时刻第i辆车缓存器中剩余计算任务量，$\boldsymbol{g}_{i,\mathrm{BS}}^{\mathrm{H}}(t)$和$\boldsymbol{g}_{i,\mathrm{UAV}}^{\mathrm{H}}(t)$分别表示在$t$时刻第$i$辆车至基站和无人机的信道质量，该参数将使用公式(6)进行求解。$C_{\mathrm{BS}}(t)$和$C_{\mathrm{UAV}}(t)$分别表示在t时刻基站与无人机剩余计算容量。

2.4.2 动作空间设计

本文假设车辆通过当前策略使用连续动作进行功率控制，控制变量主要包括：本地计算所使用的CPU频率、任务卸载至基站进行计算所使用的传输功率、任务卸载至无人机进行计算所使用的传输功率，据此，第i辆车的动作空间设计为

$$a_{i,t}=[D_i(t),p_i^{\mathrm{BS}}(t),p_i^{\mathrm{UAV}}(t)] \tag{25}$$

其中，$D_i(t)$为第i辆车使用本地计算所分配的CPU总频率占比，$p_i^{\mathrm{BS}}(t)$和$p_i^{\mathrm{UAV}}(t)$分别表示第i辆车将任务卸载至基站和无人机计算时所使用的传输功率，本文假设三个控制变量都取连续值且受最大值约束。

2.4.3 奖励函数设计

本文优化目标为找到最优的计算卸载调度策略，最小化长期系统代价，即公式(23)所定义的目标函数，其中$\zeta\in[0,1]$为权衡因子，其不同取值反映了能耗与时延的均衡关系，具体取值则取决于用户偏好。为此，在t时刻第i辆车的奖励函数设计为

$$r_{i,t}=-(1-\zeta)T_i(t)-\zeta(E_i^{\mathrm{BS}}(t)+E_i^{\mathrm{UAV}}(t)+E_i^{L}(t)) \tag{26}$$

如上式所示，奖励值为系统代价值取反，模型目标为学习一种策略使得长期平均奖励最大。第i辆车的累积奖励为

$$J(\pi_i)=\sum_{t=1}^{N}\gamma^{t-1}r_{i,t} \tag{27}$$

其中，$\gamma\in(0,1]$为折扣系数，N为每一回合的时隙数。

3 算法设计

如上节所述，本文设计的状态空间和动作空间均为连续变化值，因此，拟采用适用于连续状态空间与状态空间的PPO算法对任务卸载优化问题进行求解，即本文拟设计基于PPO的的任务卸载与功率控制优化(PPO for Task Offloading and Power Control, PPO-TOPC)算法进行求解，其网络结构如下图所示。

如图2所示, 本文采用基于演员-评论家(Actor-Critic)架构的PPO算法进行优化问题求解，其中，“演员”角色由两个结构相同的网络组成：当前网络被用户用于根据当前策略选取出动作；目标网络则用于根据选取出的动作训练当前网络，待训练完成后再将当前网络的参数赋给本网络。“评论家”角色则由单个评价网络组成，主要用于对该策略做出的选择做出评价。本文在算法中使用神经网络以求解连续动作问题，网络并不输出某一个具体的策略，而是输出该策略的概率分布。例如，假设该策略服从高斯分布，则会在网络中输出均值和方差；然

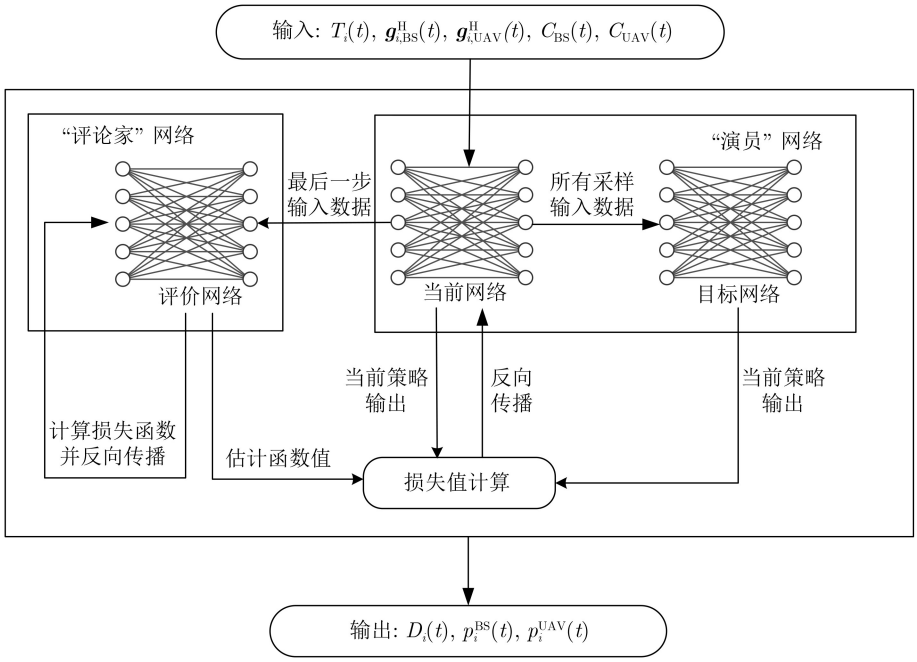


图 2 PPO-TOPC网络结构

后，再对该分布进行采样得到具体策略。本文使用的“演员”与“评论家”的当前网络和目标网络都是4层全连接层深度神经网络，每一层神经元个数设置为256，并且采用参数共享的深度神经网络架构，其中，本文采用广义优势估计作为估计函数$\hat{A}_n^{\mathrm{GAE}(\gamma,\phi)}$[27]，计算方式为

$$\hat{A}_n^{\mathrm{GAE}(\gamma,\phi)}=\sum_{l=0}^{\infty}(\lambda\phi)^l\eta_{n+l}^v \tag{28}$$

其中，ϕ为方差和偏差的权衡值，λ为折扣系数；η_{n+l}^v可通过式(29)计算

$$\eta_{n+l}^v=r_{n+l}+\gamma v(s_{n+l+1};\omega)-v(s_{n+l};\omega) \tag{29}$$

其中，$v(\cdot;\omega)$为在策略ω下的价值函数，γ为折扣系数，r_{n+l}为第$n+l$步奖励值。s_{n+l}为$n+l$步状态。

当前网络的损失函数$L_n^{\mathrm{CLIP}}(\theta)$计算为

$$L_n^{\mathrm{CLIP}}(\theta)=\min(q_n(\theta)\hat{A}_n^{\mathrm{GAE}(\gamma,\phi)},\mathrm{clip}(q_n(\theta),1-\varepsilon,1+\varepsilon)\hat{A}_n^{\mathrm{GAE}(\gamma,\phi)}) \tag{30}$$

其中θ表示当前网络的策略参数，$r_n(\theta)$表示在当前网络策略参数下的第n步奖励值，在上式中引入clip函数是为了限制$r_n(\theta)$的幅值，以防止模型更新过快而导致的不稳定问题。该函数将其范围限制在$[1-\varepsilon,1+\varepsilon]$，其中$\varepsilon$为限制范围的超参数。

式(31)中的$q_n(\theta)$定义为

$$q_n(\theta)=\frac{\pi(a_n|s_n;\theta)}{\pi(a_n|s_n;\theta_{\mathrm{old}})} \tag{31}$$

其中，$\pi(\cdot)$表示所使用的策略，θ_{old}表示目标网络的策略参数，两者比值指示的是两个网络的差异性，即重要性权重。

借助上述定义，本文设计的PPO-TOPC算法流程如算法1所示：

如算法1所示，算法主要思想为首先，给当前网络与目标网络赋予相同的初始值($\theta_{\mathrm{old}}\leftarrow\theta$)；然后，使用当前网络与环境进行交互，输出当前策略

算法 1 PPO-TOPC算法流程

输入：车辆观测到的局部信息
输出：车辆为任务计算的所使用的CPU频率占比、输出功率

1.随机初始化神经网络模型参数；
2.**for** iteration = 1,2,…M **do**
3. 初始化仿真环境参数、训练模型网络参数；
4. **for** i = 1,2,…N **do**
5. 车辆观测局部信息s_i；
6. 将s_i输入到当前网络，得到决策动作a_i；
7. 将a_i作为参数输入至环境中得到下一状态s_{i+1}和当前动作奖励r_i；
8. 车辆在本地缓存$\{a_i,s_i,r_i\}$；
9. **end for**
10. 利用缓存数据和式(28)、式(33)计算估计函数和评价网络损失函数值并更新评价网络；
11. **for** step = 1,2,…K **do**
12. 利用缓存数据和式(31)计算重要性权重
13. 根据公式(30)更新当前网络；
14. **end for**
15. 使用当前网络权重来更新目标网络；
16.**end**

的概率分布；接着，根据概率分布采样获得当前动作，再将当前动作输入至环境，得到当前奖励和下一步状态，并将其存储至缓存器中；最后，将下一步状态作为当前网络的输入，循环执行以上步骤。

根据式(28)求解出估计函数值，将以上采样的最后一步所得到的状态输入至评价网络中，得到最后一步T的状态的价值$v_T = v_T(\theta_1)$，其中θ_1为评价网络参数，并计算其上一步折扣奖励。其中折扣奖励计算为

$$R_{T-1} = r_{T-1} + \kappa v_T \tag{32}$$

其中κ为折扣系数。

重复以上步骤，依次求解出每一步的折扣奖励$\boldsymbol{R} = [R_0, R_1, \cdots, R_T]$。随后将采样到的所有状态依次输入至评价网络中，此时可得到所有状态的价值$\boldsymbol{V} = [v_0, v_1, \cdots, v_T]$，评价网络的损失函数为

$$L(\theta_1) = \mathbb{E}\left[(\boldsymbol{R} - \boldsymbol{V})^2\right] \tag{33}$$

本算法迭代计算的主要思想是求解出损失函数后，通过反向传播更新评价网络。本文假设策略的分布服从高斯分布，将所有采样得到的状态输入到当前网络和目标网络，获取两种网络下的策略分布；将采样得到的动作输入到两种分布中，获得对应动作在策略分布下的取值，将这两种取值直接相除便得到重要性采样；随后，根据式(30)求解得到当前网络的损失函数值，再利用反向传播更新当前网络的参数；当利用所有数据更新完成当前网络参数后，再利用当前网络的权重更新目标网络的权重；至此完成一次迭代运算，通过循环以上步骤直至满足停止条件即可完成模型训练。本文使用Adam优化器进行训练，在模型中通过逐步减小学习率的方式来改进学习效果，即让模型在训练前期有更大的探索率以探索更优值，后期更易于达到稳定的收敛效果。

4　仿真结果分析

为检验本文所提PPO-TOPC方案的性能，本文通过仿真实验与以下3种算法进行了对比分析：(1) 随机控制方案：该算法对本地计算功率和为卸载任务计算所使用的功率进行随机选择，且服从均匀分布，即每一功率选择都是等概选取。(2)边缘计算方案：所有任务都卸载至边缘服务器进行计算，即当有任务生成时，以时延最小为优化目标，尽量将所有任务卸载至MEC进行计算。(3)基于深度Q网络(Deep Q-network, DQN)的功率分配方案：该方案优化目标与本文相同，唯一不同的是采用DQN算法进行优化决策运算。本文主要仿真参数如表2所示：

4.1　收敛性能

为了展示所提算法的收敛性能，图3展示了模型平均奖励值随训练回合数增长的收敛性能曲线。从图3可以看出，DQN算法在前几十回合即可达到收敛性能，而PPO-TOPC算法在600个回合左右才可以达到收敛性能，因此DQN算法的收敛速度要明显优于PPO-TOPC算法；但是在收敛后的奖励值可以看出，PPO-TOPC算法要显著优于DQN算法，因此能够最终获得更好的优化效果。

4.2　任务到达速率对优化性能的影响

为了分析任务到达速率对算法优化性能的影响，图4分别展示任务到达速率对平均系统代价、平均剩余任务量和系统能耗等三个性能指标的影响。首先，如图4(a)所示，所有方案的系统代价都随着平均任务到达速率的增大而增大，但是，值得注意的是，在任意平均任务到达速率水平，仿真结果显示本文所提PPO-TOPC方案均低于其他所有方案，综合性能表现最优，多个场景中，性能提升至少13%。此外，在平均任务到达速率较低时，随机选择方案的平均系统代价高于其他所有方案，但是，当平均任务到达速率增大到一定程度时，基于边缘服务器计算方案的系统代价将将高于其他所有方案，因此这两个方案在系统优化性能均衡方面表现欠佳。

表 2　主要仿真参数

参数	数值
γ	0.9
α	1e-4
$p^o_{\max}$ (w)	1
$p^l_{\max}$ (w)	1
v_{vh} (m/s)	15
v_{uav} (m/s)	10
Y_{UAV} (m)	50
C_{BS}(m)	300
h(m)	10
L(cycle/bit)	500

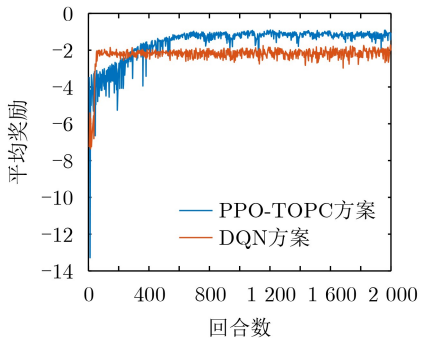


图 3 PPO-TOPC方案与基于DQN功率分配方案的收敛曲线

其次，如图4(b)所示，所有方案的平均剩余任务量都随着平均任务到达速率的增长而增长，即处理时延会增加。当平均任务到达速率小于55 Mbps时，所有方案的平均剩余任务量比较接近。当平均任务到达速率达到70 Mbps时，基于边缘服务器计算方案的平均剩余任务量急剧升高，主要原因是边缘服务器的计算容量不足以支撑所有数据及时计算，导致平均剩余任务量急剧增长。从图4(b)可以看出，基于DQN的方案和本文所提PPO-TOPC方案随着任务到达速率增长而增长的态势都比较缓慢，因此具有较好的鲁棒优化性能。特别值得注意的是，本文所提PPO-TOPC方案的平均剩余任务量一直低于基于DQN的方案，这意味着在时延方面具有更优的性能表现。

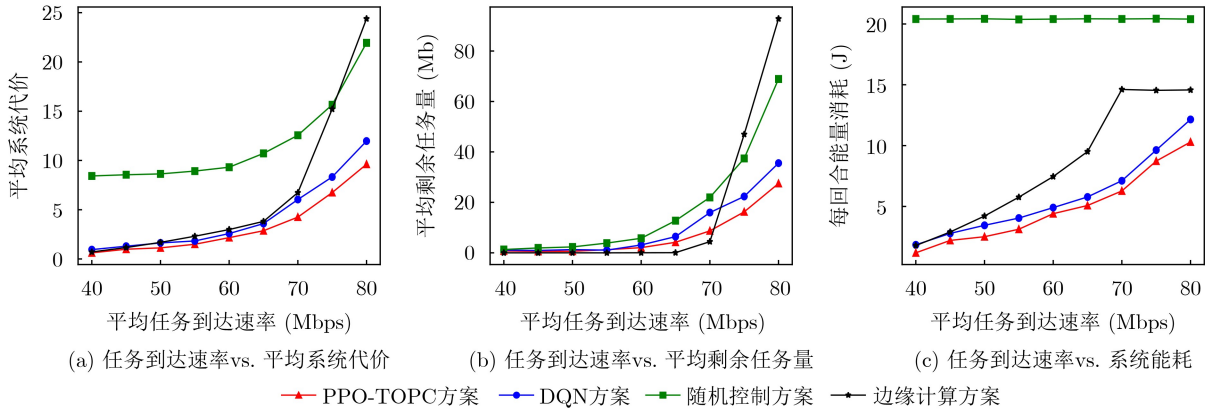
最后，如图4(c)所示，由于随机选择方案基于均匀分布随机选择本地计算与卸载任务的功率，其能耗性能表现不受平均任务到达速率变化的影响，但是其他所有方案的能耗水平都随着平均任务到达速率增长而增长。从该图可以看出，随着平均任务到达速率的增大，本文所提方案的能耗水平不仅增速缓慢，而且一直处于最低水平，表现出了最优的性能。随着平均任务到达速率的增大，基于边缘服务器计算方案的能耗水平急剧增大，当到达70 Mbps时，因服务器计算容量已达到饱和状态，导致不能卸载更多任务，因此能耗水平将不再增长，但是如图4(b)所示，其处理时延此刻会急剧增加，导致优化性能急剧下降。

特别地，相较于DQN而言，由于PPO可以直接处理连续动作空间而DQN需要离散化策略间接实现，所以本文所提方案更适用所提问题。且由于PPO采用“截断重要性采样比率”技术以限制每次更新时的策略变化范围，从而避免了过大的更新步长导致的训练不稳定问题。PPO使用“多步训练”技术，可在一次采样中同时更新多个时间步上的策略，从而提高了训练效率。

4.3 通信带宽对优化性能的影响

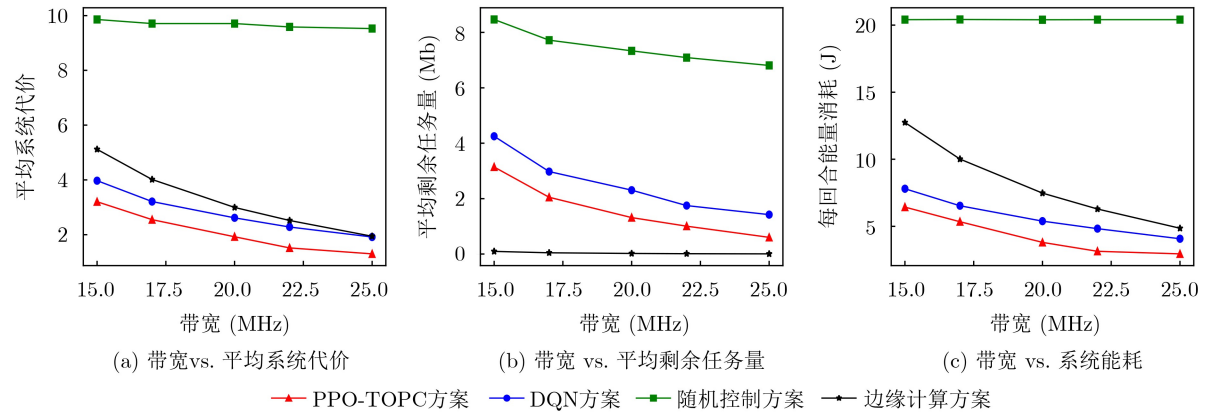
为了分析通信带宽对算法优化性能的影响，图5分别展示通信带宽对平均系统代价、平均剩余任务量和系统能耗等3个性能指标的影响。首先，如图5(a)所示，随着通信带宽的增大，所有方案的平均系统代价均会下降，而本文所提PPO-TOPC方案的平均系统代价一直处于最低水平，多个场景中，性能提升至少18%，因此取得了最优的系统性能。

其次，如图5(b)所示，随着通信带宽的增大，



(a) 任务到达速率vs. 平均系统代价 (b) 任务到达速率vs. 平均剩余任务量 (c) 任务到达速率vs. 系统能耗

图 4 任务到达速率对优化性能的影响



(a) 带宽vs. 平均系统代价 (b) 带宽 vs. 平均剩余任务量 (c) 带宽 vs. 系统能耗

图 5 通信带宽对优化性能的影响

边缘计算方案的平均剩余任务量一直处于最低水平，这是由于在该方案中本地不进行任何计算，因此不存在任何剩余任务量。此外，其他所有方案的剩余任务量都会随着通信带宽的增大而减小，从而具有更优的时延性能表现。从图5(b)可以看出，本文所提方案一直要优于其他两种方案，且当通信带宽达到一定程度时，每时隙平均剩余任务量会降低到1 Mb以下。

最后，如图5(c)所示，随着通信带宽的增大，本文所提PPO-TOPC方案的能耗水平会显著下降，且与其他方案相比较，一直处于最低水平，因此获得了最佳能耗优化性能。从该图还可以看出，即使在通信带宽较小的情况，本文所提方案的能耗水平依然处在较低水平，因此具备了良好的优化鲁棒性能。

4.4 用户偏好对优化性能的影响

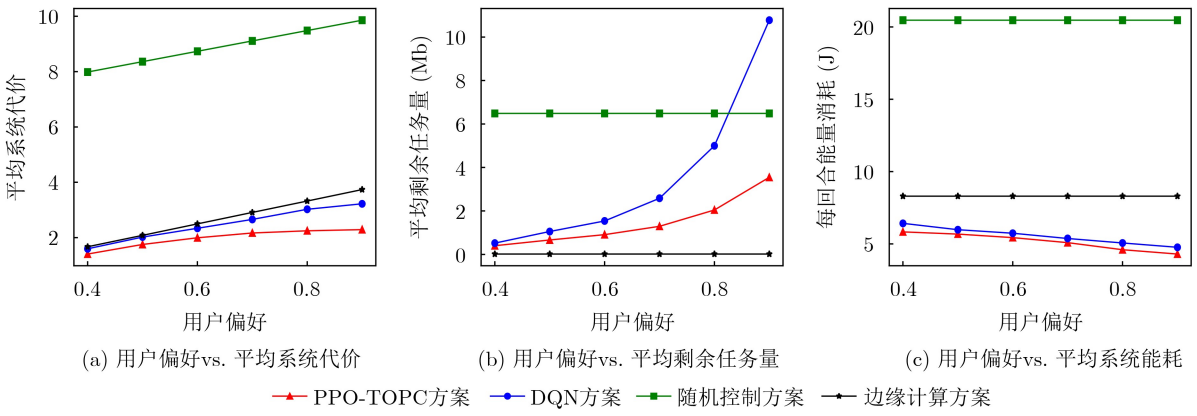
为了分析用户偏好对算法优化性能的影响，图6分别展示用户偏好对平均系统代价、平均剩余任务量和系统能耗等3个性能指标的影响。本文用户偏好使用系统奖励权重进行量化表示，权重值越高，意味着用户越偏好获得更低的能耗水平。首先，如图6(a)所示，随着奖励权重值的增大，所有方案的平均系统代价都会增高，但是，与其他方案相比较，本文所提方案的代价水平一直显著低于其他方案，多个场景中，性能提升至少15%，因此，能够取得最优的系统性能。

其次，如图6(b)所示，随着奖励权重值的增大，DQN方案与本文所提PPO-TOPC方案的平均剩余任务量都会增加，这是由于随着权重的增加，能耗占比将会增加，而平均剩余任务量占比则会降低，因此算法模型将倾向通过学习获得能耗更小、平均剩余任务量更大的优化方案参数。从该图可以看出，本文所提方案一直要优于DQN方案和随机控制方案。基于边缘服务器计算的方案由于使用完全卸载的策略，因此平均剩余任务量为空且一直保持不变，在系统资源保障充足的情况下，该方案能够获得较好性能且不受奖励权重变化的影响，但是，从图6(c)可以看出，因边缘服务器计算方案缺乏灵活的参数调节机制，其能耗性能也将一直维持不变，因此该方案无法提供在时延与能耗性能之间进行均衡的措施。

最后，如图6(c)所示， 随着奖励权重值的增大，DQN方案和本文所提方案的能耗水平都会降低，这是因为随着奖励权重增大，强化学习模型更倾向于学习使得系统能耗更小的性能优化参数，从而降低能耗水平；此外，从该图还可以看出，与其他方案相比，本文所提PPO-TOPC方案的能耗水平一直处于最低水平，因此，能够取得最佳的能耗性能。从该图还可以看出，随机选择方案与边缘服务器计算的能耗水平都与奖励权重值变化无关，因此，这两个方案都无法提供在时延与能耗性能之间进行优化均衡的措施。

5 结束语

基于无人机辅助MEC架构，本文构建出基于NOMA的车联网任务卸载与功率控制优化问题模型，并在模型中融入信道时变特性和车辆高移动性等动态变化特征，然后将该问题建模成马尔可夫决策过程，并提出一种基于近端策略优化的分布式深度强化学习算法，使得车辆能够只需根据自身获取局部信息，自主决策任务卸载所需要发射功率及本地计算所需功率，从而达到时延与能耗的最佳均衡性能。仿真结果表明，与现有的随机控制策略、边缘服务器计算策略和DQN强化学习策略相比较，本文所提PPO-TOPC方案能够显著获得更优的时延与能耗性能表现，且通过调节用户偏好权重因子，该方案还提供了一种在系统时延与能耗之间进行性能均衡优化的方法。基于本研究，未来将扩展到多基站、多无人机等场景进行后续研究。



(a) 用户偏好vs. 平均系统代价　(b) 用户偏好vs. 平均剩余任务量　(c) 用户偏好vs. 平均系统能耗

图 6 用户偏好对优化性能的影响

参 考 文 献

[1] ALAM M Z and JAMALIPOUR A. Multi-agent DRL-based Hungarian algorithm (MADRLHA) for task offloading in multi-access edge computing internet of vehicles (IoVS)[J]. *IEEE Transactions on Wireless Communications*, 2022, 21(9): 7641–7652. doi: 10.1109/TWC.2022.3160099.

[2] DU Shougang, CHEN Xin, JIAO Libo, *et al*. Energy efficient task offloading for UAV-assisted mobile edge computing[C]. Proceedings of 2021 China Automation Congress (CAC), Kunming, China, 2021: 6567–6571. doi: 10.1109/CAC53003.2021.9728502.

[3] LIU Haoqiang, ZHAO Hongbo, GENG Liwei, *et al*. A distributed dependency-aware offloading scheme for vehicular edge computing based on policy gradient[C]. Proceedings of the 8th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2021 7th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), Washington, USA, 2021: 176–181. doi: 10.1109/CSCloud-EdgeCom52276.2021.00040.

[4] 李斌, 刘文帅, 费泽松. 面向空天地异构网络的边缘计算部分任务卸载策略[J]. 电子与信息学报, 2022, 44(9): 3091–3098. doi: 10.11999/JEIT220272.

LI bin, LIU Wenshuai, and FEI Zesong. Partial computation offloading for mobile edge computing in space-air-ground integrated network[J]. *Journal of Electronics & Information Technology*, 2022, 44(9): 3091–3098. doi: 10.11999/JEIT220272.

[5] CUI Yaping, DU Lijuan, WANG Honggang, *et al*. Reinforcement learning for joint optimization of communication and computation in vehicular networks[J]. *IEEE Transactions on Vehicular Technology*, 2021, 70(12): 13062–13072. doi: 10.1109/TVT.2021.3125109.

[6] NIE Yiwen, ZHAO Junhui, GAO Feifei, *et al*. Semi-distributed resource management in UAV-aided MEC systems: A multi-agent federated reinforcement learning approach[J]. *IEEE Transactions on Vehicular Technology*, 2021, 70(12): 13162–13173. doi: 10.1109/TVT.2021.3118446.

[7] LIU Zongkai, DAI Penglin, XING Huanlai, *et al*. A distributed algorithm for task offloading in vehicular networks with hybrid fog/cloud computing[J]. *IEEE Transactions on Systems*, Man, and Cybernetics: Systems, 2022, 52(7): 4388–4401. doi: 10.1109/TSMC.2021.3097005.

[8] HAN Xu, TIAN Daxin, SHENG Zhengguo, *et al*. Reliability-aware joint optimization for cooperative vehicular communication and computing[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2021, 22(8): 5437–5446. doi: 10.1109/TITS.2020.3038558.

[9] DONG Peiran, NING Zhaolong, MA Rong, *et al*. NOMA-based energy-efficient task scheduling in vehicular edge computing networks: A self-imitation learning-based approach[J]. *China Communications*, 2020, 17(11): 1–11. doi: 10.23919/JCC.2020.11.001.

[10] ZHANG Fenghui, WANG M M, BAO Xuecai, *et al*. Centralized resource allocation and distributed power control for NOMA-integrated NR V2X[J]. *IEEE Internet of Things Journal*, 2021, 8(22): 16522–16534. doi: 10.1109/JIOT.2021.3075250.

[11] 李斌, 刘文帅, 谢万城, 等. 智能反射面赋能无人机边缘网络计算卸载方案[J]. 通信学报, 2022, 43(10): 223–233. doi: 10.11959/j.issn.1000-436x.2022196.

LI Bin, LIU Wenshuai, XIE Wancheng, *et al*. Computation offloading scheme for RIS-empowered UAV edge network[J]. *Journal on Communications*, 2022, 43(10): 223–233. doi: 10.11959/j.issn.1000-436x.2022196.

[12] BUDHIRAJA I, KUMAR N, TYAGI S, *et al*. Energy consumption minimization scheme for NOMA-based mobile edge computation networks underlaying UAV[J]. *IEEE Systems Journal*, 2021, 15(4): 5724–5733. doi: 10.1109/JSYST.2021.3076782.

[13] WANG Ningyuan, LI Feng, CHEN Dong, *et al*. NOMA-based energy-efficiency optimization for UAV enabled space-air-ground integrated relay networks[J]. *IEEE Transactions on Vehicular Technology*, 2022, 71(4): 4129–4141. doi: 10.1109/TVT.2022.3151369.

[14] KATWE M, SINGH K, SHARMA P K, *et al*. Dynamic user clustering and optimal power allocation in UAV-assisted full-duplex hybrid NOMA system[J]. *IEEE Transactions on Wireless Communications*, 2022, 21(4): 2573–2590. doi: 10.1109/TWC.2021.3113640.

[15] GAN Xueqing, JIANG Yuke, WANG Yufan, *et al*. Sum rate maximization for UAV assisted NOMA backscatter communication system[C]. Proceedings of the 6th World Conference on Computing and Communication Technologies (WCCCT), Chengdu, China, 2023: 19–23. doi: 10.1109/WCCCT56755.2023.10052259.

[16] ASHRAF M I, LIU Chenfeng, BENNIS M, *et al*. Dynamic resource allocation for optimized latency and reliability in vehicular networks[J]. *IEEE Access*, 2018, 6: 63843–63858. doi: 10.1109/ACCESS.2018.2876548.

[17] KHOUEIRY B W and SOLEYMANI M R. An efficient NOMA V2X communication scheme in the Internet of vehicles[C]. Proceedings of the IEEE 85th Vehicular Technology Conference (VTC Spring), Sydney, Australia, 2017: 1–7. doi: 10.1109/VTCSpring.2017.8108427.

[18] CHEN Yingyang, WANG Li, AI Yutong, *et al*. Performance analysis of NOMA-SM in vehicle-to-vehicle massive MIMO channels[J]. *IEEE Journal on Selected Areas in Communications*, 2017, 35(12): 2653–2666. doi: 10.1109/

JSAC.2017.2726006.

[19] ZHANG Di, LIU Yuanwei, DAI Linglong, *et al*. Performance analysis of FD-NOMA-based decentralized V2X systems[J]. *IEEE Transactions on Communications*, 2019, 67(7): 5024–5036. doi: 10.1109/TCOMM.2019.2904499.

[20] TANG S J W, NG K Y, KHOO B H, *et al*. Real-time lane detection and rear-end collision warning system on a mobile computing platform[C]. Proceedings of the IEEE 39th Annual Computer Software and Applications Conference, Taichung, China, 2015: 563–568. doi: 10.1109/COMPSAC.2015.171.

[21] ZHAN Wenhan, LUO Chunbo, WANG Jin, *et al*. Deep-reinforcement-learning-based offloading scheduling for vehicular edge computing[J]. *IEEE Internet of Things Journal*, 2020, 7(6): 5449–5465. doi: 10.1109/JIOT.2020.2978830.

[22] ZHONG Ruikang, LIU Xiao, LIU Yuanwei, *et al*. Multi-agent reinforcement learning in NOMA-aided UAV networks for cellular offloading[J]. *IEEE Transactions on Wireless Communications*, 2022, 21(3): 1498–1512. doi: 10.1109/TWC.2021.3104633.

[23] NGO H Q, LARSSON E G, and MARZETTA T L. Energy and spectral efficiency of very large multiuser MIMO systems[J]. *IEEE Transactions on Communications*, 2013, 61(4): 1436–1449. doi: 10.1109/TCOMM.2013.020413.110848.

[24] ZHU Hongbiao, WU Qiong, WU Xiaojun, *et al*. Decentralized power allocation for MIMO-NOMA vehicular edge computing based on deep reinforcement learning[J]. *IEEE Internet of Things Journal*, 2022, 9(14): 12770–12782. doi: 10.1109/JIOT.2021.3138434.

[25] LIU Yuan, XIONG Ke, NI Qiang, *et al*. UAV-assisted wireless powered cooperative mobile edge computing: Joint offloading, CPU control, and trajectory optimization[J]. *IEEE Internet of Things Journal*, 2020, 7(4): 2777–2790. doi: 10.1109/JIOT.2019.2958975.

[26] TSE D and VISWANATH P. Fundamentals of Wireless Communication[M]. Cambridge: Cambridge University Press, 2005.

[27] SCHULMAN J, MORITZ P, LEVINE S, *et al*. High-dimensional continuous control using generalized advantage estimation[J]. arXiv: 1506.02438, 2015. doi: 10.48550/arXiv.1506.02438. **(查阅网上资料,不确定文献类型,请确认)**.

谭国平：男，教授/博导，研究方向为无线分布式机器学习、移动边缘计算、车联网.

易文雄：男，硕士生，研究方向为移动边缘计算、5G通信.

周思源：男，教授，研究方向为无线网络、Beyond　5G通信、物联网.

胡鹤轩：男，教授/博导，研究方向为深度学习与大数据、智能机器人、云计算.

责任编辑：马秀强