小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2025-0113
2026 年 4 月 第 4 期
Vol.47 No.4 2026

混合动作空间下的无人机协助移动群智感知任务分配

杨桂松[1],张旭东[1],何杏宇[1,2]

[1](上海理工大学 光电信息与计算机工程学院,上海 200093)
[2](上海理工大学 出版学院,上海 200093)
E-mail:sherri_he@163.com

摘 要: 在无人机协助的移动群智感知(MCS,Mobile Crowd Sensing)系统中,由于任务分配决策是离散动作,任务执行者移动速度是连续动作,现有研究难以对二者进行同步联合优化,从而限制了感知收益的最大化.针对该问题,本文提出一种基于 E-HPPO 的无人机协助 MCS 任务分配算法.首先,针对任务分配决策的离散动作和任务执行者移动速度的连续动作,建立了具有离散动作与连续动作的两个神经策略模型的强化学习模型.其次,优化 H-PPO 算法,提出的 E-HPPO 通过分层结构对离散动作和连续动作进行处理,首先生成离散动作任务决策序列,然后根据离散动作策略网络结果生成连续动作,减少动作求解空间,提高二者之间的关联性.实验结果表明,所设计算法与现有算法和方案相比,能够有效处理离散与连续动作之间的关系,提高 MCS 平台的最终感知收益,并表现出良好的稳定性.

关 键 词: 移动群智感知;无人机辅助;混合动作空间;任务分配;深度强化学习
中图分类号: TP393 **文献标识码:** A **文 章 编 号:** 1000-1220(2026)04-0829-07

Task Allocation in UAV-assisted Mobile Crowd Sensing Based on Hybrid Action Spaces

YANG Guisong[1],ZHANG Xudong[1],HE Xingyu[1,2]

[1](School of Optical-Electrical & Computer Engineering,University of Shanghai for Science & Technology,Shanghai 200093,China)
[2](College of Publishing,University of Shanghai for Science & Technology,Shanghai 200093,China)

Abstract: In UAV-assisted Mobile Crowd Sensing (MCS,Mobile Crowd Sensing) systems,the challenge lies in the synchronisation and joint optimisation of the discrete task assignment decision and the continuous task performer's moving speed,which existing research has proven difficult to address. This paper proposes an E-HPPO-based task allocation algorithm for UAV-assisted MCS to address this problem. The proposed methodology involves the establishment of a reinforcement learning model comprising two neural policy models of discrete and continuous actions. These models are used to represent the discrete action of task allocation decision and the continuous action of the task performer's moving speed. Secondly,the H-PPO algorithm is optimised,and the proposed E-HPPO processes discrete and continuous actions through a hierarchical structure. This structure first generates discrete action task decision sequences,and then generates continuous actions based on the results of the discrete action policy network. This reduces the action solution space and improves the correlation between the two. The experimental results demonstrate that the designed algorithm can effectively handle the relationship between discrete and continuous actions,improve the final perceived benefit of the MCS platform,and exhibit good stability compared with existing algorithms and schemes.

Keywords: mobile crowd sensing;UAV-assisted;hybrid action space;task assignment;deep reinforcement learning

0 引 言

近年来,移动群智感知(MCS)作为一种新的数据收集与信息服务模式,逐渐成为研究热点.随着 MCS 应用的普及,任务数量增多且感知数据类型多样,但在一些特殊区域(如工业园区、灾区等),普通用户因条件限制无法完成感知任务,导致数据缺乏.为提高任务完成率,常需激励参与者,这增加了感知成本.无人机作为智能设备,具有更强的感知能力,能提供更多维度的数据(如高空视角),且可以标准化数据收集流程,因此被广泛应用于灾难救援、监视与环境监测中.无人机能够替代人类执行危险和重复性任务,因此需要引入无人机来协助传统的移动群智感知来更好的完成任务.

无人机协助 MCS 任务分配目前面临以下挑战,包括:1)能效问题:如何优化无人机的任务分配与路径规划以减少能耗;2)任务复杂性:如何应对大规模任务分配及任务的高风险性;3)大规模集群优化:如何在大规模无人机群体中实现高效任务分配.针对这些问题,学者们提出了多种方法和框架:文献[1]将任务分配建模为多臂老虎机问题以应对任务能力缺乏先验知识的情况;文献[2]通过动态激励机制与实时任务调度算法,显著提升了时间敏感型移动群智感知系统的任务完成率;文献[3]提出基于平台和参与者中心的拍卖机制以应对任务多样性;文献[4]针对时效性和位置依赖任

收稿日期:2025-03-14 收修改稿日期:2025-04-11 基金项目:国家自然科学基金项目(61602305,61802257)资助;上海市自然科学基金项目(18ZR1426000,19ZR1477600)资助. 作者简介:**杨桂松**,男,1982 年生,博士,副教授,CCF 会员,研究方向为物联网、卫星网络、边缘计算与普适计算等;**张旭东**,男,2000 年生,硕士研究生,研究方向为移动群智感知;**何杏宇**(通信作者),女,1984 年生,博士,副教授,研究方向为无人系统和群智计算.

务,考虑时间和地点信息分配任务;文献[5]基于参与者的偏好和可靠性进行任务匹配;文献[6]通过融合移动预测与差分隐私机制,提出隐私保护任务分配,在保障用户轨迹数据安全的同时优化任务分配效率文献[7]提出一种覆盖感知的无人机辅助任务分配方法,通过动态建模用户移动性与无人机覆盖能力的协同关系;文献[8]提出多对多匹配算法解决多任务分配;文献[9]提出动态更新多任务分配方法. 文献[10]等通过双阶段双边匹配优化了无人机任务分配与路径规划,文献[11]等通过图注意力网络与深度强化学习的协同优化实现任务分配的动态决策与资源协同调度,文献[12]通过多智能体深度强化学习解决数据感知与计算卸载问题,文献[13]则开发了分布式 DRL 框架"DRL-eFresh"以优化数据收集率和能耗. 此外,研究者[14]将多代理深度强化学习方法应用于无人机辅助救灾网络,解决了联合数据感知与计算卸载难题. 这些研究表明,无人机的引入显著提高了 MCS 任务分配的效率和灵活性,但仍需进一步探索如何在保证数据质量的前提下控制成本,同时优化多任务场景下的整体效用与单任务质量.

为了提高环境适应性和协同效率,强化学习方法被引入无人机协助移动群智感知任务分配的问题研究[14]. 由于系统中存在离散的任务分配决策和连续的移动速度决策,存在混合动作空间问题,现有强化学习方法往往是通过转换获得单一形式的动作空间,也就是将连续动作离散化或反向转化. 例如,文献[15,16]通过将离散动作连续化和连续动作离散化训练,或文献[17]使用策略优化离散动作空间如 PPO 来解决该问题. 文献[18]提出的 HyAR 方法则通过将连续动作和离散动作转化成一个处于中间状态的动作的来优化混合动作空间.

目前,已有研究尝试通过将离散动作空间连续化或者连续动作空间离散化来应对混合动作空间问题. 然而,这些方法普遍存在适应性不足的问题. 例如将离散动作连续化处理将导致空间复杂度大大增加,而连续动作离散化将导致强化学习过程中丢失重要信息. 因此在无人机协助 MCS 任务分配混合动作空间处理问题上,急需一种将连续动作与离散动作二者进行同步联合优化的策略.

本文提出一种基于 E-HPPO 的无人机协助 MCS 任务分配策略. 首先建立具有离散动作策略网络与连续动作策略网络的强化学习模型. 其次通过 E-HPPO 算法,提出离散动作策略网络与连续动作策略网络联合优化,生成离散动作任务决策序列,根据离散动作任务决策序列生成连续动作任务执行者移动速度. 该策略考虑了不同动作之间的关联性,有效处理离散与连续动作之间的关系,提高 MCS 平台的最终感知收益,并表现出良好的稳定性.

1 MCS 任务分配模型

1.1 MCS 系统模型

移动群智感知(MCS)系统包括 3 层架构:任务执行层、MCS 平台层和任务请求层.

1.1.1 任务执行层

任务执行层由移动用户和无人机组成,负责根据云平台分配的任务前往指定地点完成感知操作并上传数据,其协同效率和数据传输可靠性对系统性能至关重要. 本文的任务参与者主要为无人机与工人,其模型可以用$\{X_I, Y_I, B_I, C_I, V_I\}$来描述,$I$ 取值为 W_n 和 U_m,I 为 W_n 代表工人模型,I 为 U_m 代表无人机模型.

a)参与者位置:X_I 和 Y_I 代表参与者的 X 坐标和 Y 坐标,用于确定参与者在任务场景中的位置.

b)参与者报酬 B_{W_n}:B_{W_n}表示工人执行任务所需要的报酬,其含义为该工人在执行任务时需要的报酬为 B_{W_n}每公里. B_{U_m}表示无人机执行任务所需能耗,其含义为该无人机在执行任务时需要的能耗为 B_{U_m}每公里.

c)参与者能力 C_I:C_I 代表工人或者无人机执行任务的获取感知收益的能力. 工人的感知能力受限于设备精度和人为因素,文献[5]通过实验表明,普通用户的感知数据质量通常为理想值的 70% ~90%,文献[7]指出,工人执行任务的可靠性受环境干扰影响,其有效数据采集率约为 0.75 ~0.85,因此本文工人取值在 0.7 ~0.9 之间. 无人机配备专业传感器,且飞行路径可控,数据质量更稳定. 文献[10]表明无人机在标准化数据收集中误差率低于 10%,故能力值接近 1.0. 文献[12]通过实际测试验证,无人机在灾害监测任务中的感知准确率达 92% ~98%,因此无人机取值在 0.9 ~1.0 之间.

d)参与者速度 V_I:V_I 表示工人或者无人机执行任务的移动速度. 参考行人移动速度研究,成年人的步行速度通常为 1.2 ~1.5 m/s,但任务执行者可能使用自行车或电动滑板车,速度提升至 4 ~5 m/s,因此本文工人通常取值在 4 ~5 之间. 商用多旋翼无人机[10]的巡航速度为 8 ~12 m/s,本文取中间值以平衡能耗与效率,无人机取值在 9 ~10 之间.

1.1.2 任务请求层

任务请求层是任务需求的发起者,向云平台提交任务目标、地点、时间约束及质量要求等信息,为云平台调度与执行提供依据,直接影响任务效果与成果质量. 任务 T 可以用集合$\{X_{T_K}, Y_{T_K}, P_{T_K}, D_{T_K}, R_{T_K}\}$来描述,具体定义如下:

a)任务的位置:X_{T_K}和 Y_{T_K}分别代表感知任务 T 的 x 坐标和 y 坐标,用于确定任务在场景中的位置.

b)任务优先级: P_{T_K}表示任务 T_K 优先级. 每个任务都被分配一个优先级参数,数值越高表示优先级越高. 任务的优先级与任务的感知收益直接相关,也影响移动用户和无人机在执行任务时的感知收益. 高优先级任务往往对感知质量有更大的影响. 如果任务在当前时隙内未完成,其优先级会逐渐下降.

c)任务的完成速度:D_{T_K}代表任务 T_K 的完成速度,衡量任务在给定时间内的执行效率. 时间敏感任务的感知收益衰减现象[2]广泛存在于应急响应和实时数据采集场景中. 本文 c 参考 Sigmoid 衰减因子,任务 T_K 设有截止时间 t_K,如果任务越接近截止时间才完成,则感知质量的收益会显著降低. 为了量化这种时间对感知质量的影响,本文引入了一个衰减因子,该因子通过以下公式定义,并与任务感知收益相乘:

$$D_{ij}=\begin{cases}\dfrac{1}{1+e^{-\epsilon(t_K-t_{iK})}}, \text{if } t_K \geqslant t_{iK}\\ 0, \text{if } t_K < t_{iK}\end{cases} \tag{1}$$

其中,t_K 表示任务的截止时间,t_{iK}表示实际完成任务的时间,ϵ 为调节任务完成时间对收益影响的参数.

d)任务的基础感知收益:R_{T_K}表示在理想情况下,感知任务 T 完成时的基础收益. 这一收益在没有时间衰减等因素影响下,反映了任务完成的基本回报.

1.1.3　MCS 平台层

云平台层作为系统核心调度与管理中心,负责处理任务需求、管理资源状态及位置信息,通过数据筛选提升质量,并根据预设策略高效分配任务,优化资源利用和任务完成效率. MCS 平台任务分配后感知收益由参与者执行任务感知收益与参与者执行任务的消耗之差获得,其计算如下:参与者与任务的距离:假设参与者 I 被指派执行任务 T_K,由于 Euclidea 距离往往不能准确反映两个位置坐标之间的真实距离,因此本文使用 Haversin 距离来表示两个位置之间的距离. 则参与者 I 与任务 T_K 的距离计算公式为:

$$D_{IT_K}=2\mathrm{r}\times \arcsin\sqrt{\sin^2\left(\frac{y_I-y_{T_K}}{2}\right)+\cos y_I\cos y_{T_K}\times\sin^2\left(\frac{x_I-x_{T_K}}{2}\right)} \quad (2)$$

总体感知收益:总体感知收益由工人的感知收益和无人机的感知收益组成,收益由参与者执行任务的距离 D_{IT_K},参与者与任务的距离越近,执行效率越高,参与者的感知能力 C_I,反映数据质量,与收益正相关,执行任务的基础感知收益 R_{T_K},任务固有价值,由请求者设定,以及执行任务的优先级 P_{T_K},高优先级任务需赋予更高权重这些因素获得,同时为其添加 $\mu_1\ \mu_2\ \mu_3\ \mu_4$ 这个 4 个权重系数,通过实验调优确保量纲一致性与实际场景适配性,其计算公式如下:

$$G_{total}=\mu_1 D_{IT_K}\times\mu_2 C_I\times\mu_3 R_{T_K}\times\mu_4 P_{T_K} \quad (3)$$

总体执行消耗:总体执行消耗 E_{total} 由工人 W_n 执行任务的报酬和无人机 U_m 执行任务的能耗两部分组成,其都由执行任务的距离 D_{IT_K} 乘于工人的报酬 B_{W_n} 或者无人机 B_{U_m} 能耗组成计算公式如下:

$$E_{total}=D_{IT_K}\times B_I \quad (4)$$

本文的无人机协作 MCS 任务分配模型由 MCS 平台,数个感知任务,无人机以及移动用户组成. MCS 平台从任务请求者那里获得感知任务信息后,对信息进行处理,结合当前平台记录的空闲移动用户以及无人机信息,进行优选,给出最优的任务分配策略,并将任务信息以及移动速度信息 传达给任务执行者. 任务请求者发布的感知任务分布在场景中的不同位置,无人机和移动用户能够前往相应位置执行感知任务. MCS 平台首先对任务信息、移动用户和无人机的信息进行初步处理和汇总,随后进行任务分配. 根据任务执行结果的反馈,平台再进行下一轮的任务分配. 在任意时隙 t 中,针对 MCS 场景内的感知任务,MCS 平台决定由哪位工人或无人机来执行任务.

1.2　通信模型

MCS 平台获得的总体收益由两部分组成:一部分是执行任务的移动用户和无人机获得的总感知收益 G_{total},另一部分是总体能耗 E_{total},由执行任务的移动用户所需的报酬以及无人机的能耗组成. 此外,本文引入收益系数 σ_1 以及报酬和能耗的权重 σ_2,使效用函数更加全面地描述 MCS 平台的总体收益. 基于上述,平台效用函数被定义为整体感知收益与能源消耗和补偿成本之间的差额. 平台效用函数可表示如下:

$$R=\sigma_1\times G_{total}-\sigma_2\times E_{total} \quad (5)$$

根据上述条件,本文将问题定义为多目标优化问题,问题模型及其约束为:

$$P1: arg^{max}_{W_n,U_m,v_{U_m},v_{W_n}}=R \quad (6)$$

$$s.t.\quad C1:\sum_{m=1}^{m}U_m+\sum_{n=1}^{n}W_n=K \quad (7)$$

$$C2: E_{total}\leqslant E_{max} \quad (8)$$

$$C3: E_{U_m}\leqslant UE_{max} \quad (9)$$

$$C4: D_{IT_K}\leqslant D_{max} \quad (10)$$

其中,约束条件(C1)确保平台将每项任务分配给一名移动用户或一架无人机,保证每个任务只能由一个移动用户或无人机执行. 约束条件(C2)确保分配给工人和无人机的任务的总报酬和能耗不超过平台的总预算. 约束条件(C3)确保每个执行任务的无人机的能耗不超过其自身的能耗限制. 最后,约束条件(C4)确保每个参与者执行任务的距离不超过其最大行程限制.

2　基于 E-HPPO 的 UAV-MCS 任务分配优化算法

本研究探讨了移动群智感知中的多目标人机协同任务分配问题,将非凸优化的多元变量问题转化为马尔可夫决策过程(MDP),通过强化学习方法优化状态和动作. 借助策略搜索与价值迭代,逐步逼近全局最优解. 如图 1 所示,本文将任务分配问题建模为最优 MDP 求解问题,提出基于 E-HPPO 的任务分配算法,并详述其设计与实现方案.

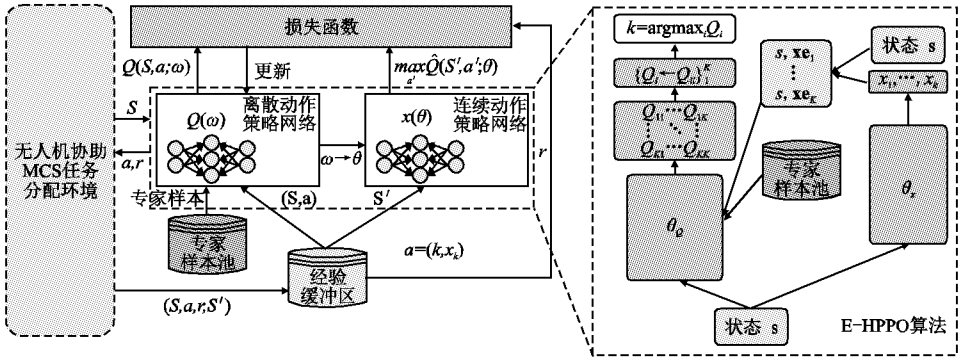

图 1　基于 E-HPPO 的强化学习模型

Fig. 1　Reinforcement learning model based on E-HPPO

2.1　基于马尔可夫决策过程的任务分配强化学习模型建模

本文研究的移动群智感知(MCS)场景需针对任务属性

及执行者(移动用户和无人机)设计优化的任务分配策略,以确保任务顺利完成. MCS 平台可视为强化学习智能体,通过设计最优分配策略满足任务需求. 尽管理论上可借助马尔可夫决策过程(MDP)四元组(S,A,R,P)精确解决任务卸载问题,但因任务和执行者状态动态变化,难以获得精确的状态转移概率 P. 为此,本文采用基于三元组(S,A,R)的无模型强化学习方法解决该问题. 整体强化模型如图 1 所示.

a)状态 state:强化学习的目标在于通过不断从历史数据中学习策略,逐步逼近全知视角. 因此,全面而精确地定义状态对于提升决策效率具有重要意义. 本文综合考虑了任务信息、移动用户和无人机的使用情况、任务与其对应的移动用户和无人机的位置信息,以及移动用户和无人机的个体属性,将时隙 t 的状态 s_t 定义如下:

$$s_t = \{P(t), C(t), O(t), I(t)\} \tag{11}$$

其中,$P(t) = \{P_1(t), P_2(t), P_3(t), \cdots, P_k(t)\}$ 表示 t 时隙内 k 项任务的基本属性,$C(t) = \{C_1(t), \cdots, C_n(t), \cdots, C_{n+m}(t)\}$ 表示 m 台无人机和 n 位工人在 t 时隙内是否处于空闲状态可以执行任务. $O(t) = \{O_1(t), \cdots, O_n(t), \cdots, O_{n+m}(t), \cdots, O_{n+m+k}(t)\}$ 表示 t 时隙内 m 台无人机、n 位工人和 k 项任务的位置信息. 最后,$I(t) = \{I_1(t), \cdots, I_n(t), \cdots, I_{n+m}(t)\}$ 表示 t 时隙内 n 位工人和 m 台无人机的基本属性.

b)动作 action:在时隙 t,MCS 平台依据环境状态 s_t 进行行动决策. 该决策包括离散变量的任务分配决策和连续变量的任务执行速度决策. 因此,时隙 t 的行动 a_t 定义如下:

$$a_t = \{\lambda(t), \theta(t)\} \tag{12}$$

其中,$\lambda(t) = \{\lambda_1(t), \lambda_2(t), \lambda_3(t), \cdots, \lambda_k(t)\}$ 表示 t 时隙内 k 项任务分别分配给了哪些工人和无人机的决策序列,$\theta(t) = \{\theta_1(t), \theta_2(t), \cdots, \theta_n(t), \theta_{n+1}(t), \cdots, \theta_{n+m}(t)\}$ 该决策表示 n 个移动用户和 m 台无人机的执行速度. 由于任务分配决策序列属于离散变量,而移动用户和无人机的执行速度为连续变量,故平台的决策空间由离散变量和连续变量共同构成.

c)奖励 reward:在时隙 t,用户设备在状态 s_t 下执行动作 a_t,并获得即时奖励 r_t,该奖励用于评估智能体卸载决策的优劣. 本文的优化目标是最大化系统效用,其具体形式由式(6)给出. 因此,奖励函数定义如下:

$$R(s_t, a_t) = \begin{cases} R & s.t.\ (7) \sim (10) \\ -\mu & otherwise \end{cases} \tag{13}$$

其中,μ 表示惩罚项. 当 MCS 平台的任务分配决策未满足公式(7)~公式(10)所列约束条件时,将受到相应的惩罚. 该机制旨在向智能体提供反馈信号,指示其所选择的行动具有次优性质,从而引导智能体优化其策略以满足约束条件并提升决策质量.

2.2 基于 E-HPPO 的无人机协助 MCS 任务分配算法

现有强化学习算法通常假设动作空间为连续或离散,对于处理连续动作空间的算法(如 A3C 和 DDPG),面临应对离散动作空间的挑战;而处理离散动作空间的算法(如 DQN 和 DDQN)则难以适应连续动作的复杂性. 常见解决方案是对动作空间进行量化,但精细离散化会增加维度,粗糙离散化则可能损失行为信息. 在人机协同任务分配中,动作空间的离散与连续混合特性带来了新的挑战. 为此,本文提出了一种基于改进 PPO 的 E-HPPO 算法,通过增强混合动作空间的优化能力,提升任务分配与决策效率. E-HPPO 基于 Actor-Critic 框架,Critic 部分为状态值网络输出状态价值,Actor 部分则针对混合动作空间设计优化方案,有效平衡了复杂性与效率.

E-HPPO 的网络结构由两个独立且并行的 Actor 网络组成,分别为 离散 Actor 网络(Discrete Actor Network) 和 连续 Actor 网络(Continuous Actor Network),分别负责表示动作的离散部分和连续部分. 其中,左侧的 Discrete Actor Network 学习一个随机策略,用于选择离散动作. 右侧的 Continuous Actor Network 则学习一个随机策略,用于生成所有连续参数. 尽管在每次决策时会输出所有连续参数,实际执行的动作是基于离散动作和连续参数的联合选择结果. 此外,这两个 Actor 网络共享前端的状态编码网络(State Encoding Network)参数,以提取状态的共享表征,从而提高网络的特征学习能力和整体效率. E-HPPO 的动作空间定义如下:

$$A = \bigcup_{a \in A_d} \{(a, x) \mid x \in X_a\} \tag{14}$$

其中,A 表示整个动作空间,A_d 是一个有限集合,包含了所有可能的离散动作,即 $A_d = \{a_1, a_2, \cdots, a_k\}$. a 是从集合 A_d 中选择的一个特定的离散动作. X_a 是与离散动作 a 相关联的连续参数的集合,x 是从集合 X_a 中选择的连续参数,用于与动作 a 一起执行. E-HPPO 的动作选择采用 ε - *greedy* 策略,其中动作 a_t,连续动作 a_d 和离散动作 a_c,探索参数 ε,t 时刻环境状态信息 s_t,离散网络 Q(w)和连续网络 X(θ)具体流程图如图 2 所示.

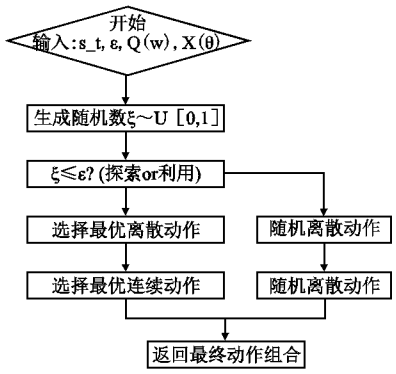

图 2 greedy 动作选择算法流程图

Fig. 2 Flowchart of greedy action selection algorithms

本文在 PPO 框架中引入优先级重放缓冲机制和专家样本以提升性能. 优先级重放通过动态调整样本权重,优先采样高贡献样本,降低低效样本采样概率,同时分层采样,解决低贡献样本的价值被忽视的问题;专家样本是由先验知识生成的优质状态-动作对,用于初始化经验回放池,其目标策略为贪心策略,离散动作选择最近且高能力的参与者,连续动作取最大速度. 专家样本利用贪心策略生成高质量数据,通过模拟器运行贪心策略,收集初始数据,加速训练初期收敛,缓解冷启动问题,增强泛化能力. 分层采样解决低贡献样本的价值被忽视的问题,贪心策略引入随机探索生成部分非最优动作的样本,以覆盖更多状态-动作对. E-HPPO 结合离散与连续动作空间,支持智能体在不同维度独立决策与调整,适应混合场景,显著提升复杂任务表现,相较于 DDQN 和 A3C 算法更具优势. 在 E-HPPO 中考虑一个拥有式(14)中定义的动作空间

的 MDP,对于 $a\in A$,有 $Q(s,a)=Q(s,a,x)$. 则有贝尔曼方程:

$$Q(s,a)=E_{r,s'}[r_t+\gamma max_{a'\in A}Q(s_{t+1},a')\mid s_t=s,a_t=a] \quad (15)$$

其中,γ 是折扣因子,表示未来奖励的当前价值.

目标值的定义基于状态价值函数 $V(s)$ 的估计,其公式如下:

$$A_t=-V(s_t)+r_t+\gamma r_{t+1}+\gamma^{T-t-1}r_T-1+\gamma^T V(s_T) \quad (16)$$

其中,t 是时间步索引,T 是小于一个 *episode* 长度的某个时间步.

H-PPO 使用 PPO 作为策略优化方法,其损失函数是一个裁剪的替代目标(CLIP),定义如下:

$$L_{CLIP}(\theta)=E_t[r_t(\theta)\ A_t,clip(r_t(\theta),1-\epsilon,1+\epsilon)\ A_t] \quad (17)$$

其中,$r_t(\theta)$是策略 π_θ 下动作 a_t 被选中的概率与旧策略 $\pi_{\theta old}$ 下动作 a_t 被选中的概率的比率,ε 是一个超参数.

E-HPPO 的网络参数更新公式,对于离散策略 $\pi_{\theta d}$和连续策略 $\pi_\theta c$,他们分别有自己的裁剪替代目标式(18)和式(19):

$$L^d_{CLIP}(\theta_d)=E_t[\min(r_{d,t}(\theta_d)A_t,clip(r_{d,t}(\theta_d),1-\epsilon,1+\epsilon)A_t)] \quad (18)$$

$$L^c_{CLIP}(\theta_c)=E_t[\min(r_{c,t}(\theta_c)A_t,clip(r_{c,t}(\theta_c),1-\epsilon,1+\epsilon)A_t)] \quad (19)$$

其中,$r_{d,t}(\theta_d)$和 $r_{c,t}(\theta_c)$分别是离散策略和连续策略的概率比率,它们仅考虑各自的策略更新,而不是联合分布. 通过最小化这些目标函数,可以分别更新离散演员网络和连续演员网络的参数.

通过上述改进,E-HPPO 的网络结构如图 1 右侧所示. 相比传统的混合动作强化学习算法,E-HPPO 在训练过程中表现出更高的稳定性与更快的收敛速度. 如算法 1 所示,其时间复杂度主要由训练周期 *episode* 和单个周期内的时间步数 T 决定,总体时间复杂度可表示为 $O(n_T n_{episode})$,其中 n_T 为每个周期包含的时间步数,$n_{episode}$为训练周期的总数. 这样的复杂度分析为算法的性能评估提供了理论依据,并进一步证明了其在复杂任务中的高效性.

算法 1. 基于 E-HPPO 的 MCS 无人机协助任务分配算法

输入:任务,无人机和工人的数据状态信息

输出:任务分配决策序列和奖励,训练网络并更新网络

```
1. 使用专家样本初始化经验回放池 D.
2. 初始化评估网络权重 w 和 θ
3. 初始化探索参数 ε,小批量大小 B
4. for each episode do
5.     初始化状态 s_t
6.     for 1≤t≤T do
7.         根据算法 1 选择动作 a_t
8.         执行动作 a_t,获得奖励 r_t 和状态 s_{t+1}
9.         将[s_t,a_t,r_t,s_{t+1}]存储到 D 中
10.        从 D 中随机采用含有 B 条经验数据{s_b,a_b,r_b,s_{b+1}}_{b∈[B]}
           的小批量
11.        根据式(17)训练 B 个样本得到目标
12.        更新网络评估权重和目标网络权重
13.    end for
14. end for
```

3　实验结果与分析

在本节中,本文将评估所提出算法的性能. 在评估中,本文基于无人机协助 MCS 任务分配的场景,设计了一个仿真实验,并指定了具体的参数值. 然后,本文分析了算法的可行性. 最后,本文设计了几个对照组来验证算法的效率.

3.1　参数设置

为验证方法有效性并兼顾隐私与安全性,本文通过模拟仿真实验重构实际移动群智感知(MCS)场景,假设在洪涝灾害区域开展救援信息采集任务. 实验假设系统由多个工人、无人机及 MCS 平台组成,任务位置在 100×100 网格区域内随机分布,属性参照模型设计并在合理范围内随机生成. 在任务属性生成中截止时间 t_K 服从均匀分布 $U[1,T_{max}]$,其中 $T_{max}=10$ 时隙,模拟任务紧急的多样性,基础收益 R_{T_K}与优先级 P_{T_K}线性相关,高优先级任务的 ϵ 更高,体现严格时效性. 任务的完成奖励 μ_1 μ_2 μ_3 μ_4 4 个权重系数通过网格搜索确定,最终选择使平台收益最大化的组合为 $\mu_1=0.3$,$\mu_2=0.4$,$\mu_3=0.2$,$\mu_4=0.1$. 收益系数 σ_1 调节感知收益的贡献比例,参考能源效率权衡研究,默认取 0.6. 惩罚系数 σ_2 控制成本敏感性,通过固定 σ_1 参数调优根据平台预算约束实验调优取值为 0.4. 同时设置惩罚项 μ 为整体奖励 10%,在约束满足率与收敛速度间达到平衡. 工人和无人机初始位置,随机分布,能力值根据第 1.1.1 节描述随机生成,符合正态分布. 一共生成 10 组不同规模的数据集,每组包含 100 个训练 episode,验证算法可扩展性.

表 1　E-HPPO 超参数

Table 1　E-HPPO hyperparameters

名称	值
学习率 α	0.01
回放经验池大小 *pool*	10000
奖励折扣因子 γ	0.95
专家池大小 D	3000
初始探索率 ε	1.0
批量数据规模 B	128

实验平均运行 20 次以提高结果可信度,训练过程包含 10000 个,超参数配置如表 1 所示,包括经验回放池大小 10000、奖励折扣因子 0.95、网络参数更新频率 0.01、小批量数据规模 128、初始探索率 1.0 递减至 0.1. 所有实验都是在 Windows 11 操作系统上进行的,该系统运行在 Intel(R) Core (TM) i5-12490F CPU 和 8GB 内存上,使用的编程语言是 Python 3.8 版本,使用的实验平台是 Pycharm 2020.1.6. 所有模型基于 PyTorch 实现,采用两层全连接神经网络结构. 本实验最终输出结果为智能体执行任务感知收益与资源消耗之差,即算法实验中的输出奖励结果代表实际情况中平台最终感知收益. 上述设计旨在减少偶然误差,确保实验结果的可靠性与可重复性.

3.2　对比算法

由于现有的任务分配研究方法并不适合无人机协助 MCS,为了验证所提出的 E-HPPO 算法,本文设计了下列对比算法作为不同场景下的基线.

a) DDQN:DDQN 通过引入双 Q 网络改进了 DQN,以减少 Q 值高估偏差,其中一个网络用于动作选择,另一个用于目标计算. 但在连续控制任务中适用性较低.

b) A3C:A3C 是一种适用于连续动作空间的非策略算

法,基于带有深度神经网络的 Actor-Critic 框架,结合了价值函数和策略优化方法. 尽管通过经验回放和目标网络稳定训练,但其在离散控制任务中表现相对较弱.

c)P-DQN:P-DQN 是针对混合动作空间的强化学习算法,结合离散动作选择与连续参数优化. 通过深度 Q 网络选择离散动作,策略网络生成相应连续参数.

d)纯工人 MCS(FWS):与 E-HPPO 不同的是,E-HPPO 可以将任务分配给无人机执行,而在 FWS 中,MCS 平台只能将任务分配给工人执行.

e)贪婪算法(Greedy):与 E-HPPO 不同的是,任务分配中的贪婪策略在每次决策时都会优先考虑当前剩余的可用无人机或移动用户资源,以实现该任务的效益最大化.

f)随机任务分配算法(Random):与 E-HPPO 相比,随机任务分配策略是指在任务分配过程中随机选择合适的工人和无人机执行感知任务.

3.3 结果分析

3.3.1 E-HPPO 算法性能分析

图 3 展示了 E-HPPO 算法在不同学习率下的收敛表现. 实验环境中,工人数量设置为 5,无人机数量设置为 5,任务数量为 10,其坐标及相关属性均随机生成. 从图 2 可以看出,随着训练轮数的增加,算法逐渐趋于收敛. 然而,更高的学习率并不一定能够带来更快的收敛速度或更高的奖励收益. 在 E-HPPO 算法的实验中,学习率为 0.01 时,收敛效果最佳.

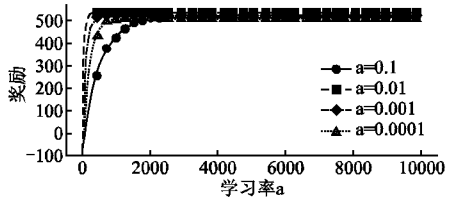

图 3 不同学习率下最终感知收益

Fig. 3 Final perceived benefits at different learning rates

为了评估专家经验池对 E-HPPO 算法性能的影响,实验设置了不同规模的专家经验池进行测试. 图 4 显示了专家经验池大小对模型奖励的具体影响. 结果表明,增大专家经验池的规模并不总能带来更高的奖励收益. 例如,在本实验中,专家经验池大小为 3000 时表现最佳,优于 4000 和 5000 的实验结果. 因此,考虑到较高的相关成本以及观察到的递减收益,专家经验池的规模应经过谨慎选择.

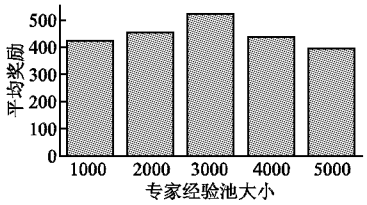

图 4 不同专家经验池下奖励表现

Fig. 4 Reward performance different expert experience pools

在进行 E-HPPO 与其他算法(DDQN、A3C、P-DQN)的学术性对比分析时,本文从 MCS 平台最终感知收益和收敛速度两个维度探讨.

在 MCS 平台最终感知收益的对比中,从图 5 可以看出,E-HPPO 算法相较于 DDQN、A3C 和 P-DQN 在不同的任务维度下展现出了显著的优势,总体提升在 15% ~25%. DDQN 算法通过采用两个 Q 网络来减少 Q 值高估,提高了学习的稳定性和可靠性. 然而,DDQN 在连续控制任务中的适用性不如 E-HPPO,后者作为一种改进的 PPO 算法,能够更好地处理混合动作空间,从而在总体奖励上超越 DDQN. A3C 算法虽然在连续行动空间中表现出色,但其基于价值和策略的方法可能导致在某些任务中不如 E-HPPO 那样高效. P-DQN 算法虽然能够处理混合动作空间,但在离散与连续动作相结合的问题上可能不如 E-HPPO 那样灵活和高效. E-HPPO 的优势源于分层结构处理混合动作空间,其对混合动作空间的适应性和更优的探索策略,这使得它能够在复杂的任务中获得更高的总体奖励.

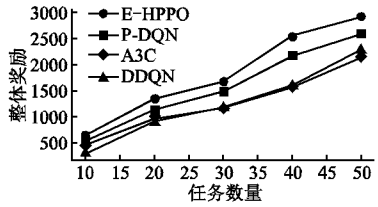

图 5 不同算法下 MCS 最终感知收益对比

Fig. 5 Comparison of final perceived benefits of MCS under different algorithms

从图 6 可见,E-HPPO 算法在收敛速度方面相较于 DDQN、A3C 和 P-DQN 展现了显著优势. 在收敛速度上,E-HPPO 在不同任务维度下的提升约为 7% ~13%. DDQN 尽管通过分离动作选择与 Q 值估计降低了估值偏差,但其收敛速度不及 E-HPPO. A3C 通过异步更新提高了训练稳定性,但在离散控制任务中的收敛速度仍逊于 E-HPPO. P-DQN 在混合动作空间中具备一定稳定性与收敛性能,但 E-HPPO 凭借基于信任域的策略优化和精细的动作选择机制,实现了更快的收敛. E-HPPO 的优势主要源于通过分离离散分配与连续速度决策,其对混合动作空间的高适应性、优化的探索策略以及精细的策略优化能力,使其在复杂强化学习任务中展现出更高效、更稳定的性能表现.

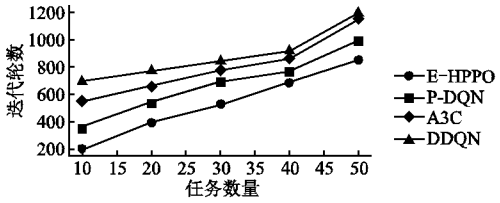

图 6 不同算法下迭代轮数对比

Fig. 6 Comparison of iteration rounds under different algorithms

3.3.2 任务分配方案性能分析

为了全面评估 E-HPPO 在无人机辅助移动人群感知任务分配中的有效性,本文对不同任务、无人机和工作人员配置下的常用 MCS 任务分配方法进行了系统性比较分析. 每种配置均进行了 10 次独立实验,以确保统计结果的稳健性,并通过平均处理减少随机性和可变性对结果的影响. 从图 7 可以看出实验结果表明,E-HPPO 在无人机辅助任务分配中显著优于其他方法. 在与 FWS 的对比中,E-HPPO 的平台总体感

知收益提升了 14% ~20%,凸显了无人机协作在感知效益上的显著优势. 相较于 Greedy,E-HPPO 提升了 10% ~15%,得益于其综合优化当前与未来回报的策略,而非仅关注单一任务的即时收益. 与 Random 方法相比,E-HPPO 的提升幅度达 20% ~25%,主要由于后者缺乏目标性和优化机制. 总体而言,E-HPPO 通过智能任务分配与资源调度,充分发挥了无人机与工作人员协同作业的潜力,在各类配置中均展现出显著的性能优势.

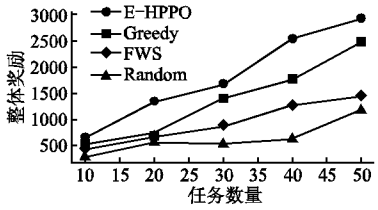

图 7　不同任务分配方案下最终感知收益对比

Fig. 7　Comparison of final perceived benefits

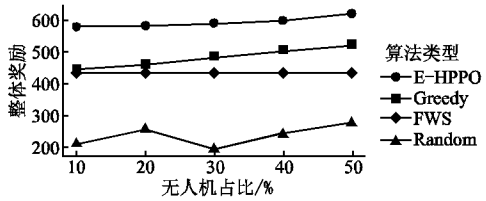

图 8　不同无人机占比下最终感知收益对比

Fig. 8　Comparison of final perceived benefits for UAV

图 8 结果表明,随着无人机占比的增加,平台总体感知收益显著提升,尤其在任务数量与复杂度较高的场景中,无人机的引入显著增强了任务执行效率和感知能力. 相较于高比例依赖工作人员的方案,尽管工作人员具有较高的灵活性和稳定性,其在处理复杂任务时的回报表现相对较弱,整体效率较低. 与传统完全依赖工作人员的 FWS 方法相比,E-HPPO 通过深度融合工作人员与无人机的协同效应,在任务分配回报率上实现了显著提升. 特别是,E-HPPO 通过综合考虑当前与未来的系统回报,优化任务分配策略,从而在长期感知收益上占据优势. 相反,Greedy 算法侧重于当前任务的即时回报,忽视了全局和长期效益,而 Random 方法由于缺乏目标性和优化机制,导致分配效率较低. E-HPPO 通过智能化的任务优化与资源调度,充分发挥了无人机与工作人员协同作业的潜力,显著提升了系统的整体收益. 综上,合理调配无人机与工作人员比例并结合智能调度算法,是实现资源最优配置与任务执行效能最大化的关键.

4　结　论

本文提出了一种无人机协助的移动群智感知(MCS)任务分配框架,并基于 E-HPPO 算法设计了人机协同的任务分配策略. 该框架通过优化成本与感知质量的平衡,充分利用了工人与无人机的优势,解决了传统 MCS 面临的任务规模需求和感知成本问题. 实验结果表明,E-HPPO 算法在降低成本和提高感知质量方面优于现有算法. 创新点在于将无人机引入 MCS 任务分配,采用多目标优化方法,解决了大规模数据收集中的能效问题和任务复杂性. 通过改进的强化学习算法,本文有效应对了混合动作空间的挑战,为未来 MCS 任务分配提供了新的解决方案. 未来,结合元学习和多智能体强化学习等先进算法,可以进一步提升任务分配策略的适应性与效率. 随着技术的进步,基于强化学习的任务分配策略将在智能交通、环境监测等领域展现广泛应用前景,并为提升无人机感知系统的性能提供更高效、灵活的解决方案.

References:

[1] Zhao H,Xiao M,Wu J,et al. Differentially private unknown worker recruitment for mobile crowdsensing using multi-armed bandits[J]. IEEE Transactions on Mobile Computing,2020,20(9):2779-2794.

[2] Xu Z,Sun H,Han W. Boosting task completion rate for time-sensitive MCS system[J]. Computer Networks,2024,251:110636,doi:10.1016/j.comnet.2024.110636.

[3] Cai Z,Duan Z,Li W. Exploiting multi-dimensional task diversity in distributed auctions for mobile crowdsensing[J]. IEEE Transactions on Mobile Computing,2020,20(8):2576-2591.

[4] Wang Z,Zhao J,Hu J,et al. Towards personalized task-oriented worker recruitment in mobile crowdsensing[J]. IEEE Transactions on Mobile Computing,2020,20(5):2080-2093.

[5] Wu F,Yang S,Zheng Z,et al. Fine-grained user profiling for personalized task matching in mobile crowdsensing[J]. IEEE Transactions on Mobile Computing,2020,20(10):2961-2976.

[6] Xie Z,Peng T,You W,et al. P2-TaskMP:privacy-preserving task allocation optimization based on mobility prediction[J]. Future Generation Computer Systems,2025:107720,doi:10.1016/j.future.2025.107720.

[7] Liu X,Wang Y,Gao H,et al. A coverage-aware task allocation method for UAV-assisted mobile crowd sensing[J]. IEEE Transactions on Vehicular Technology,2024,doi:10.1109/tvt.2024.3374719.

[8] Yang G,Wang B,He X,et al. Competition congestion-aware stable worker-task matching in mobile crowd sensing[J]. IEEE Transactions on Network and Service Management,2021,18(3):3719-3732.

[9] Dai C,Wang X,Liu K,et al. Stable task assignment for mobile crowdsensing with budget constraint[J]. IEEE Transactions on Mobile Computing,2020,20(12):3439-3452.

[10] Zhou Z,Feng J,Gu B,et al. When mobile crowd sensing meets uav:energy-efficient task assignment and route planning[J] IEEE Transactions on Communi cations,2018,66(11):5526-5538.

[11] Xu C,Song W. Intelligent task allocation for mobile crowdsensing with graph attention network and deep reinforcement learning[J]. IEEE Transactions on Network Science and Engineering,2023,10(2):1032-1048.

[12] Cai T,Yang Z,Chen Y,et al. Cooperative data sensing and computation offloading in uav assisted crowdsensing with multi-agent deep reinforcement learning[J]. IEEE Transactions on Network Science and Engineering,2021,9(5):3197-3211.

[13] Dai Z,Liu C H,Han R,et al. Delay sensitive energy-efficient uav crowdsensing by deep reinforcement learning[J]. IEEE Transactions on Mobile Computing,2021,22(4):2038-2052.

[14] Xiong J,Wang Q,Yang Z,et al. Parametrized deep q-networks learning:reinforcement learning with discrete-continuous hybrid action space[J]. arXiv preprint arXiv:1810.06394,2018.

[15] Delalleau O,Peter M,Alonso E,et al. Discrete and continuous action representation for practical rl in video games[J]. arXiv preprint arXiv:1912.11077,2019.

[16] Xiong J,Wang Q,Yang Z,et al. Parametrized deep q-networks learning:reinforcement learning with discrete-continuous hybrid action space[J]. arXiv preprint arXiv:1810.06394,2018.

[17] Fan Z,Su R,Zhang W,et al. Hybrid actor-critic reinforcement learning in parameterized action space[J]. arXiv preprint arXiv:1903.01344,2019.

[18] Li B,Tang H,Zheng Y,et al. Hyar:addressing discrete-continuous action reinforcement learning via hybrid action representation[J]. arXiv preprint arXiv:2109.05490,2021.