文章编号:1671-4598(2024)01-0297-07 DOI:10.16526/j.cnki.11-4762/tp.2024.01.042 中图分类号:TP393 文献标识码:A

基于强化学习的无人机网络资源分配研究

范文帝，王俊芳，党 甜，杜龙海，陈 丛
（中国电子科技集团公司 第54研究所，石家庄 050081）

摘要：以无人机网络的资源分配为研究对象，研究了基于强化学习的多无人机网络动态时隙分配方案，在无人机网络中，合理地分配时隙资源对改善无人机资源利用率具有重要意义；针对动态时隙分配问题，根据调度问题的限制条件，建立了多无人机网络时隙分配模型，提出了一种基于近端策略优化（PPO）强化学习算法的时隙分配方案，并进行强化学习算法的环境映射，建立马尔可夫决策过程（MDP）模型与强化学习算法接口相匹配；在gym仿真环境下进行模型训练，对提出的时隙分配方案进行验证，仿真结果验证了基于近端策略优化强化学习算法的时隙分配方案在多无人机网络环境下可以高效进行时隙分配，提高网络信道利用率，提出的方案可以根据实际需求适当缩短训练时间得到较优分配结果。

关键词：深度强化学习；多无人机网络；动态时隙分配；资源分配；近端策略优化

Research on Resource Allocation in UAV Networks Based on Reinforcement Learning

FAN Wendi, WANG Junfang, DANG Tian, DU Longhai, CHEN Cong
(The 54th Research Institute of CETC, Shijiazhuang 050081, China)

Abstract: The resource allocation of UAV networks is taken as a research object, a dynamic time slot allocation scheme in multi-UAV networks based on reinforcement learning is investigated. In UAV networks, it is important to reasonably allocate time slot resources to improve UAV resource utilization. Aiming at the dynamic time slot allocation problem, a time slot allocation model for multi-UAV networks is established according to the constraints of the scheduling problem. A time slot allocation scheme based on the proximal policy optimization (PPO) reinforcement learning algorithm is proposed to carry out the environment mapping of the reinforcement learning algorithm, and build a Markov decision process (MDP) model to match the interface of the reinforcement learning algorithm. The model training is performed in the Gym simulation environment to validate the proposed time slot allocation scheme. The simulation results show that based on the proximal policy optimization reinforcement learning algorithm, the time slot allocation scheme can efficiently perform the time slot allocation and improve the network channel utilization in a multi-UAV network environment. The proposed scheme appropriately reduces the training time according to actual demands, which obtains optimal allocation results.

Keywords: deep reinforcement learning; multi-UAV networks; dynamic time slot allocation; resource allocation; proximal policy optimization

0 引言

随着近几年无人机在关键技术上的突破，在军用和民用领域具有高机动性和低成本特性的无人机用途十分广泛。在人们的日常生活生产活动中，无人机越来越多地出现，在精准农业、应急救援、交通管制、货物递送等领域发挥着重要作用[1-3]。

然而随着无人化进程的加深，单无人机能力不足、资源有限、任务执行低效，如同自然界中动物通过成群结伴来弥补个体能力的有限[4]，无人机集群执行复杂任务成为无人机应用的重要模式，无人机集群信息实现实时传递的关键是无人机通信网络。

但无人机通信网络在实际操作中存在着众多问题，比如，电力资源、功率资源等等都存在着资源利用率不高的问题。资源分配问题是智能场景下研究的一个热点，一些启发式算法被提出以解决资源分配问题，文献[5]运用禁忌搜索算法解决火力资源分配；文献[6]在资源分配问题上应用模拟退火算法，模拟退火算法容易实现，鲁棒性强、质量高，但优化过程所需时间较长；文献[7]提出武器动态资源分配问题，用动态规划的方法来解决，但是递归的使用在编码过程中使得效率非常低；文献[8]完善标准灰狼算法，提出解决资源分配难题的离散灰狼算法。群体智能算法的速度虽然与传统算法相比有所提升，但仅仅针对

收稿日期:2023-11-02； **修回日期**:2023-11-20。
基金项目:国防基础科研计划资助(JCKY2020210B021)。
作者简介:范文帝(1998-),男,研究生。
通讯作者:王俊芳(1963-),男,博士,研究员,博导。
引用格式:范文帝,王俊芳,党 甜,等.基于强化学习的无人机网络资源分配研究[J].计算机测量与控制,2024,32(1):297-303,311.

一个问题进行解决，很难扩展到其他问题的应用，在解决动态资源分配问题上效果不佳。

在无人机通信网络中，由于有限的频谱资源和动态的无人机用户需求，时隙分配最常被用来解决资源竞争和协调的问题。然而，时隙分配是一个经典的 NP 难题，困扰着众多研究者。其中一些传统的求解方式，如枚举法、分支定界法、动态规划法，这些都是很容易实现但又很缓慢的搜索方式；对于传统的智能算法，如遗传算法、差分进化算法，要想扩展有一定难度。巨大的状态空间和不断变化的环境使得传统的决策方法在这方面的效果并不理想。

随着各个研究领域深度强化学习技术的不断发展，强化学习被重新应用到此项问题。更多基于深度强化学习技术解决资源分配问题的方法被提出[9-12]，强化学习算法训练出的模型不仅决策速度快，而且扩展性强，能够在各种场景中广泛应用，能够有效应对动态变化的环境，非常适用于动态资源分配问题。强化学习是一种基于经验学习和探索的智能算法，可以通过对实时环境的反馈以及循环性训练，不断优化决策结果。

文献［13］对无人机网络辅助移动边缘计算执行任务卸载策略这个领域做了一系列的研究探索，说明了基于强化学习的无人机网络的资源分配和路径规划可以在当前相关研究中达到不错性能；文献［14］集中探讨了无人机群网络频谱分配方面的研究，研究了使用强化学习的动态信道分配和动态时隙分配算法；文献［15］主要探讨了无人机集群中的资源分配和虚拟网络映射问题，该文献对现有的资源分配算法进行了分析，总结了其优劣之处，同时提出了一种全新的资源分配算法；文献［16］是以深度强化学习算法为基础的资源分配问题研究，结合专家经验与深度强化学习算法，探索引入专家数据对资源分配算法性能的影响。

为了对资源分配问题进行更好的研究，本文将在深度强化学习技术的帮助下，研究无人机通信网络中如何更好地解决时隙分配问题。

本文基于近端策略优化算法[17-18]，设计了一个更高效的方法求解时隙分配问题。设计方法中的 PPO 算法使用 Kullback-Leibler（KL）散度来限制更新前后的策略变化，解决了传统梯度下降优化算法[19-20]只能用最小步长进行更新的限制，PPO 算法通常能优效地训练智能体执行复杂的控制任务。

1 无人机网络时隙资源分配方案

1.1 问题描述

无人机通信网络由 M 无人平台节点组成，由集合 $M=\{1,2,\cdots,M\}$ 表示，$m\in M$ 表示无人平台节点的编号。其中，各无人平台节点之间构成多跳场景，节点维护自身的一跳和两跳邻居表，分别用 $N_1(m)$ 和 $N_2(m)$ 表示第 m 无人节点的一跳和两跳邻居集合。为了便于表示，利用矩阵 $\boldsymbol{C}$ 表示所有节点的一跳邻接矩阵：

$$\boldsymbol{C}=\begin{Bmatrix} C_{1,1} & C_{1,2} & \cdots & C_{1,3} \\ C_{2,1} & C_{2,2} & \cdots & C_{2,3} \\ \vdots & \vdots & \ddots & \vdots \\ C_{M,1} & C_{M,2} & \cdots & C_{M,M} \end{Bmatrix} \tag{1}$$

其中：$c_{m,n}\in\{0,1\}$ 表示第 m 节点和第 n 节点间是否存在链路，$c_{m,n}=1$ 表示两个节点之间存在链路，第 m 节点和第 n 节点互为一跳邻居，即 $m\in N_1(n)$，否则，两个节点之间不存在链路。

无人机通信网络因其拓扑结构的动态性导致难以预先设定中心节点作为集中控制设备，进行全局的时隙资源动态分配。为了在无人机通信网络中优化调度时隙资源，需设计以节点为中心的时隙分配算法。在一个时隙调度周期中，假设共有 K 空闲数据时隙可分配给 M 个无人平台节点用于发送数据，k 表示时隙的编号。时隙的带宽被表示为 B。利用二元变量 $t_{m,k}\in\{0,1\}$ 表示第 k 时隙到第 m 无人节点的分配情况，其中，$t_{m,k}=1$ 表示第 k 时隙被分配到第 m 无人节点用于传输数据，否则 $t_{m,k}=0$。

此时，网络的信道利用率被定义为：

$$\rho=\frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K}t_{m,k} \tag{2}$$

代表的是网络分配的时隙数和所有能够分配的时隙数的比值。

时隙分配问题可以构建为以提高信道利用率为目标的优化问题。

另外，无人节点在请求时隙时，同时上传自身网络状态信息，包括节点吞吐量、节点负载、时延等。考虑到不同无人平台节点上通信性能的差异性，将同一时隙分配给不同节点所带来的网络性能增益不同，因此为分配到节点的时隙设置效用函数。效用函数由每个时隙调度周期内节点上报的吞吐量、负载、时延等构成。为了消除不同性能指标间的量纲差异，对其进行归一化，下面以负载为例：

负载：假设节点的负载状态为 Q_m，每个节点可允许的最大负载和出现的最小负载分别是 $Q_{m,max}$、$Q_{m,min}$，负载的归一化指标为：

$$Q'_m=\frac{Q_m-Q_{m,min}}{Q_{m,max}-Q_{m,min}} \tag{3}$$

每个节点的效用函数利用 U_m 表示，指该节点负载的归一化指标的加权和，其表达式为：

$$U_m=\omega Q'_m \tag{4}$$

其中：ω 的取值为 1。

因此，为节点分配时隙的时候需要同时考虑网络的信道利用率以及节点本身的网络状态。

在广播调度问题中，存在两个限制条件：一是非传输限制，即在一个调度周期内，网络中的每个节点至少有一个时隙被调度；二是非冲突限制，即调度过程中要避免两类冲突，第一类冲突是节点不能同时发送和接收数据，第二类冲突是一个节点不能同时接收超过一个传输[21]。

此时，时隙分配问题被构建为：

$$\max_{t_{m,k}}:\frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K}U_m t_{m,k} \tag{5}$$

$$s.t.\begin{cases}\mathrm{C1}:\sum\limits_{k=1}^{K}t_{m,k}\geqslant 1,\forall\, m\in M\\ \mathrm{C2}:c_{m,n}+t_{m,k}+t_{n,k}\leqslant 2,m\neq n\\ \mathrm{C3}:c_{m,q}t_{m,k}+c_{n,q}t_{n,k}\leqslant 1,m\neq n\neq q\end{cases} \tag{6}$$

其中：公式（5）为加权信道利用率，式（6）中 C1 保障了在一个时隙调度周期内所有无人平台节点至少存在 1 次时隙调度。假设两跳以上的节点可以利用空间复用的方式避免干扰，C2 约束同一个时隙不会分配给相邻 2 个节点，C3 约束分配到同一个时隙的节点间最多存在 1 条链路。

1）当存在中心控制节点时：

在无人机通信网络中，当存在一个无人平台节点是中心控制节点时，所有节点维护的邻居表和网络状态等信息可以被收集到集中控制节点处，并基于这些数据对上述优化问题进行全局求解。上述时隙分配问题是以二元时隙分配变量为优化对象的整数线性规划，可以利用混合整数线性规划（MILP）求解器或者深度强化学习算法进行求解。

MILP 求解器是一种计算机程序，可以解决包含整数变量的线性规划问题。这些求解器使用分支定界法等算法来解决整数规划问题，但它们通常比手动实现更高效。

2）当不存在中心控制节点时：

然而，由于无人机通信网络中难以设定集中控制节点，求解上述以优化全局网络性能为目标的时隙分配问题时难以获取所有节点的网络状态信息。为了进行时隙调度，假设节点要维护的邻居表包括一跳和两跳，同时还会维护一跳和两跳邻居的时隙分配表，可以将上述优化问题拆分为以节点 m 为中心的时隙分配问题：

$$\max_{t_{m,k}}:U_m\sum_{k=1}^{K}t_{m,k} \tag{7}$$

$$s.t.\begin{cases}\mathrm{C1}:\sum\limits_{k=1}^{K}t_{m,k}\geqslant 1,\forall\, m\in M\\ \mathrm{C2}:t_{m,k}\leqslant 2-t_{n,k}-c_{m,n},m\neq n\\ \mathrm{C3}:c_{m,q}t_{m,k}\leqslant 1-c_{n,q}t_{n,k},m\neq n\neq q\end{cases} \tag{8}$$

此时，优化对象是所有时隙到节点 m 的调度情况。对于节点 m 而言，其一跳邻居、两跳邻居、以及一跳邻居和两跳邻居的时隙占用情况都是已知的。上述以节点 m 为中心的时隙分配问题是整数线性规划问题，公式（8）中 C1 保障了在一个时隙调度周期内所有无人平台节点至少存在 1 次时隙调度，C2 和 C3 约束了变量 $t_{m,k}$ 的上界，C2 约束同一个时隙不会分配给相邻 2 个节点，C3 约束分配到同一个时隙的节点间最多存在 1 条链路。另外，分析该优化问题可知，对于无人平台节点 m，所分配到的时隙数量越多，其优化目标函数越大，其变化是随着其占用的时隙数量增加而单调递增的。此时，节点 m 的时隙分配结果可以直接以贪婪的思路最终确定。

上文将全网的时隙分配问题拆分为以节点 m 为中心的时隙分配子问题进行求解，不再需要将所有节点的网络状态信息收集到集中控制设备，但是这种求解方式难以获取全局最优的结果。并且，值得注意的是，时隙调度的最终目标是为所有节点分配数据时隙。在以节点 m 为中心的时隙分配问题中，将同一时隙分配给不同节点所带来的网络性能增益不同，这说明节点时隙分配的顺序对节点调度的公平性和稳定性具有重要影响，为此，需要提出基于优先级的节点时隙分配。

基于优先级的节点时隙分配需要按照节点优先级由高到低对节点进行时隙分配，节点都维护着一个优先级列表。节点的优先级用变量 P_m 表示，P_m 取值越大，节点 m 的优先级越高。节点优先级是指由于无人机通信网络中各通信节点的角色不同，一些通信节点承担重要的指令传输任务，因此必须保持稳定传输，减少网络的丢包。相比于其他节点，承担重要数据传输任务的节点优先级较高，这是由硬件设备和通信任务决定的优先级，将节点的初始优先级表示为 P_m^0。另外，为了避免固定优先级设置导致低优先级节点长时间无法获取时隙资源，优先级还会随着节点网络状态变化而动态变化，在此考虑的是无人节点流量负载 Q_m 对优先级的影响。

由于节点负载过高会导致时延提高和丢包率增加，当节点监测到自身负载 Q_m 超过一定阈值 α_m 时，就需要请求更多数据时隙尽快传输缓存队列的流量，此时该节点将临时提高自身优先级并将更新后的优先级信息广播给邻居节点，这时节点的时隙分配顺序也会相应的提前，有利于节点获取占用更多时隙进行数据传输。当节点监测到自身负载 Q_m 低于一定阈值 β_m 时，可恢复为初始优先级，并将其更新消息广播给其他节点。为了避免节点优先级的频繁变更导致时隙分配无法收敛，设置负载阈值 α_m 和 β_m 的取值满足 $\alpha_m>\beta_m$。

当 $Q_m\geqslant\alpha_m$ 时，更新节点优先级的表达式可以定义为：

$$P_m=P_m^0+\frac{(Q_m-\alpha_m)^+}{Q_m} \tag{9}$$

当 $Q_m<\beta_m$ 时，节点优先级恢复为初始优先级 P_m^0。

当 $\beta_m<Q_m<\alpha_m$ 时，节点保持当前优先级不变。

基于优先级的节点时隙分配允许优先级较高的节点获得优先进行时隙划分的机会，通过调整时隙请求期间的参数，优先解决高优先级节点的时隙分配优化问题，使得重要紧急的消息可以高效传输。基于优先级的节点时隙分配流程如图 1 所示。

1.2　模型描述

为了提高多无人机网络的通信效率和整体系统性能协作，本文将运用深度强化学习的研究成果，设计一种用于解决时隙资源分配问题的方案。

在 2017 年，OpenAI 提出了 PPO 算法，这是一种深度强化学习算法，旨在通过使用 PPO 算法来让智能体解决问题。为此，需要先建立一个 MDP 模型，这个模型由状态空间（S）、动作空间（A）和奖励函数（R）组成[22]。在 t 时刻，环境处于状态 S_t，智能体根据当前状态选择策略 A_t，环境转移到下一状态 S_{t+1}，同时智能体获得相应的反馈奖励 R_t。

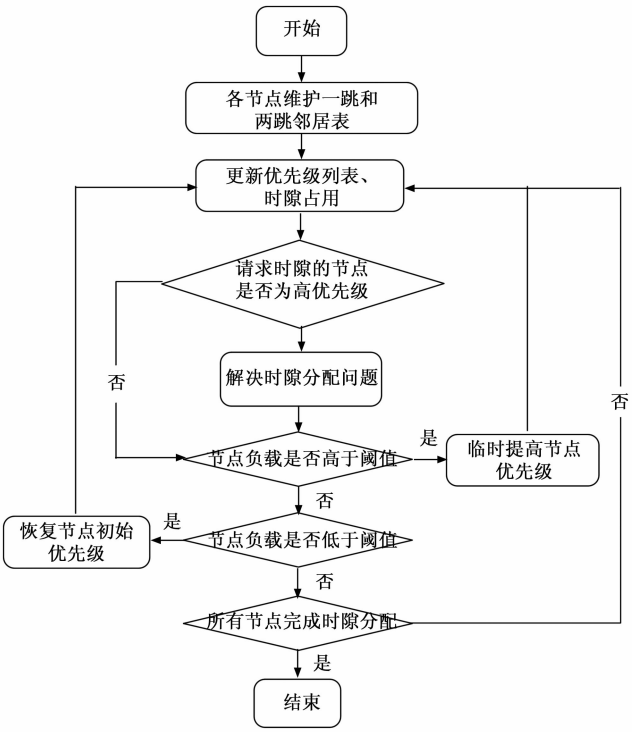

图 1 基于优先级的节点时隙分配流程

假设存在一个多无人机网络，其中包含一组环境状态 S、动作集合 A 以及回报奖赏 R。该网络中共有 N 架无人机，并且每个周期需要分配 T 个时隙。具体的动态时隙分配环境映射如下所述。

环境状态 S：为了评估无人机在执行动作后对环境状态的影响，环境状态主要用于描述这些影响的程度。这种程度可以包含积极的和负面的影响。根据环境状态的变化，可以更方便地设计后续的奖励函数[14]。环境状态可以被定义如下：

$$S = \{s_1, s_2, \cdots, s_N\}$$

在动态时隙分配的背景下，环境状态可由多个方面来描述，即无人机节点的负载和无人机节点的链接拓扑等。

动作集合 A：强化学习的另一个重要组成部分是动作。动作是无人机与环境交互的必备因素。通过不断尝试各种动作或策略，无人机可以获取环境反馈的信息和奖励值，以便快速了解采取何种动作或策略，以最大化环境收益[14]。动作集合可以定义如下：

$$A = \{a_1, a_2, \cdots, a_{N\times T}\}, a_i \in \{0,1\}$$

式中，a_i 表示第 $i \div T$ 时隙下的第 $i \bmod T$ 架无人机所采取的动作，也就是说该时隙下，无人机可以选择两种不同的动作再每个时隙中进行数据传输操作，$a_i = 0$ 表示无人机在该时隙中不进行数据传输；$a_i = 1$ 表示无人机在该时隙中试图进行数据传输。

奖赏函数 R：奖励函数的作用是教导智能体向对环境有好处的方向前进，并且限制其行动范围。一般来说，奖励函数会根据环境的特点进行制定。在动态分配的情况下，通过状态反馈来衡量环境的优劣[14]。

为了实现让智能体快速学习最优时隙分配策略的优化目标，本文设计了两部分奖励。第一部分是满足三个约束条件下的评估情况，第二部分是计算网络中通信的总负载量。

图 2 展示了模型的决策流程。

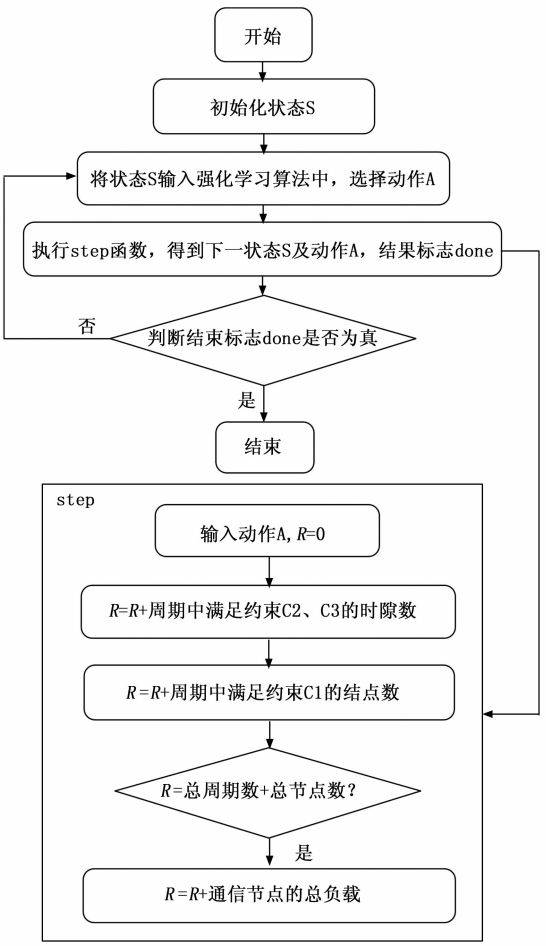

图 2 决策流程图

为了找到最佳的时隙分配策略，使得初始化状态 S 的预期奖励最大化，需要通过智能体与环境的交互来生成所有可能的时隙分配策略 π^*，以最大化系统整体的长期期望折扣奖励：

$$V_\pi(s_t, a_t) = E_\pi\left(\sum_{t=0}^{T} \gamma^t R_{t+1}(a_{t+1}, s_{t+1})\right) \tag{10}$$

最佳时隙分配策略可被表述为：

$$\pi^* = \operatorname{argmax}_\pi V_\pi = \operatorname{argmax}_\pi E_\pi\left(\sum_{t=0}^{T} \gamma R_{t+1}(a_{t+1}, s_{t+1})\right) \tag{11}$$

本文使用时隙分配系统时，不必考虑长远回报。为此，我们引入衰减因子作为衰弱机制 $\gamma \in [0,1]$。公式（11）中 s_{t+1} 描述无人机网络环境在 $t+1$ 时刻的状态；a_{t+1} 记录智能体在 $t+1$ 时刻的时隙分配行为；$R_{t+1}(a_{t+1}, s_{t+1})$ 记录 $t+1$ 时刻的即时奖励；$\gamma R_{t+1}(a_{t+1}, s_{t+1})$ 表示长时间的衰减奖励。

至此，我们已成功完成了强化学习中 MDP 建模的重要步骤。我们已经设计了模型中的状态、动作和奖励函数的三个关键要素。

2　基于近端策略优化算法的时隙分配算法框架

2.1　算法框架

如图 3 所示，本文所使用的 PPO 强化学习算法可以分为交互阶段和学习阶段两个关键阶段。

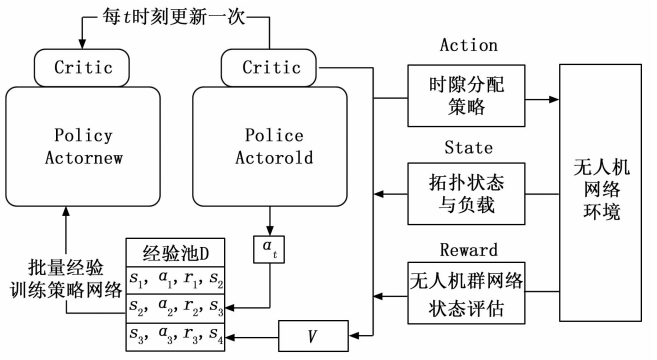

图 3　基于 PPO 算法的时隙分配系统体系结构

交互阶段：智能体与环境互动，利用环境传递的状态信息做出决策，环境再返回智能体下一时刻的状态和基于该策略的奖励。PPO 算法采用经验回放技术，运用经验数据池 D，帮助智能体更好地学习到最佳决策。该算法是一种在线学习策略，需要大量的样本数据用于重新训练模型。但是，数据集过大会导致训练时间过长，收敛性较弱；而数据集过小则会影响模型的训练效果[22]。PPO 算法利用重要性采样技术，通过创建一个固定大小的经验池 D，在解决上述问题时能够重复使用样本。

学习阶段：每次进行迭代时，智能体会利用新生成的 Actor 网络所提供的策略来选取动作，并与无人机网络环境进行互动，以获取样本数据。当一个完整的经验池 D 中的数据样本被收集完毕后，Actor 和 Critic 网络会对其中的数据进行学习。策略更新的流程如图 3 所示。

在每次迭代时，智能体使用新 Actor 网络生成的策略选取动作与无人机网络环境交互，得到一个 $(s_t, a_t, r_t, s_{t+1}, \pi_\theta)$ 样本数据，当收集完一个完整的经验池 D 的数据样本后，Actor、Critic 网络会对经验池 D 中的数据进行学习。

根据公式（10）可以计算出每个时刻的值函数：

$$\hat{A}_t = -V(s_t) + r_t + \gamma r_{t+1} + \cdots + \gamma^{T-t+1} r_{T-1} + \gamma^{T-t} V(s_T) \tag{12}$$

此时可以计算得到 Critic 网络的损失函数：

$$L_c = mean(square(\hat{A}_t)) \tag{13}$$

进行反向传播，以更新 Critic 网络。

对于经验池 D 中的每个状态，我们将其输入到两个 Actor 网络中，以获得相应的正态分布。同时，我们输入 D 中的每个动作以计算其在分布中的概率。通过将这些概率相除，我们得到了 PPO 算法的重要性采样权重：

$$r = \frac{p_\theta(a_t \mid s_t)}{p_{\theta_{old}}(a_t \mid s_t)} \tag{14}$$

可以计算 Actor 网络的误差函数：

$$L_a = \hat{E}_t[\min(r \times \hat{A}_t, clip(r, 1-\varepsilon, 1+\varepsilon) \times \hat{A}_t)] \tag{15}$$

其中：ε 是一个超参数，它用于限制策略更新幅度的 clip 函数。然后，通过反向传播更新 Actor 新网络，最终 PPO 的整体损失函数如下：

$$L_p = 0.5 \times L_c + L_a \tag{16}$$

2.2　算法流程

如算法 1 所示是一种改进的无人机网络时隙资源分配算法。

算法 1：基于 PPO 算法的时隙资源分配算法

输入：马尔可夫决策过程模型

输出：智能决策模型

1)设置模型超参数、训练总回合数、批大小、学习率、衰减率 γ，每次更新次数、最大步长、更新频率

2)初始化经验回放池 D

3)初始化无人机网络环境，生成网络拓扑

4)初始化算法模型，策略网络参数 θ

5)Actor 网络参数设置采样环境参数 $\theta^{old} \leftarrow \theta^{new}$

6)for $i = 1, 2, \cdots$ do

7)　重置环境

8)　for each step t do

9)　　观察无人机网络初始状态 s_t

10)　　将状态 s_t 输入到 Actor new 网络得到动作 a_t；

11)　　在环境中执行操作和下一个状态 s_{t+1} 后得到奖励 r_t，是否终止 d

12)　　将 $(s_t, a_t, r_t, s_{t+1}, \pi_\theta)$ 存入经验回访池 D 中

13)　　最终状态输入 Critic 网络中，根据公式(10)计算衰减后的奖励

14)　　if 更新步骤等于经验池 D 容量 then

15)　　将记录的状态组合输入到 Critic 网络根据公式(12)估计优势函数

16)　　使用梯度下降法优化目标函数并更新网络参数 θ^{new}

17)　　end

18)　end

19)　更新旧策略模型参数 $\theta^{old} \leftarrow \theta^{new}$

20)end

21)保存模型，绘制奖励曲线和网络损失曲线

3　仿真与分析

3.1　仿真原理及参数

仿真环境为 OpenAI Gym 框架，在 Anoconda 平台使用 Python 语言实现。

在无人机网络场景下，网络的拓扑结构多种多样，图 4 中为典型的网络拓扑，其中，链状拓扑结构下，节点时隙复用的方式更多，三种拓扑结构同样数量的节点下链状拓扑结构运算复杂度最高，因此选取链状拓扑结构进行仿真。

在仿真中设定节点数为 4，5，时隙数为 8，网络拓扑为线性拓扑。下面以 5 节点举例，网络的邻接矩阵为：

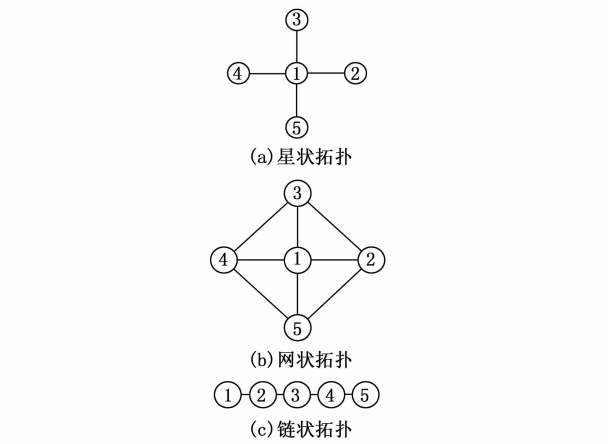

图 4 三种网络拓扑图

$$\boldsymbol{C}=\begin{pmatrix}0&1&0&0&0\\1&0&1&0&0\\0&1&0&1&0\\0&0&1&0&1\\0&0&0&1&0\end{pmatrix}$$

其中：C_{ij} 代表节点 i、j 的联通情况，计算得到允许同时开通的矩阵：

$$\boldsymbol{B}=\begin{pmatrix}0&0&0&1&1\\0&0&0&0&1\\0&0&0&0&0\\1&0&0&0&0\\1&1&0&0&0\end{pmatrix}$$

B_{ij} 代表节点 i、j 是否能够同时分配时隙。

随机生成节点当前负载为 $\boldsymbol{Q}=[10,12,4,4]$。

对载荷进行归一化后，$U=[0.416\ \ 67,0.833\ \ 33,1,0.333\ \ 33,0.333\ \ 33]$。

算法中奖励的衰减率为 0.99，PPO 算法中 clip 参数为 0.2。

3.2 仿真结果与分析

实验的损失函数如图 5 所示。

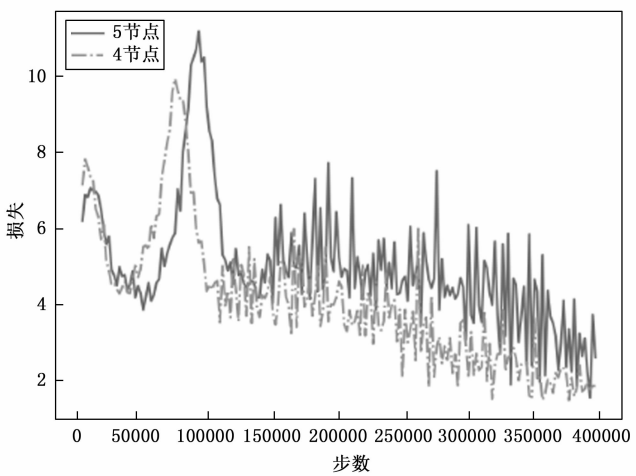

图 5 损失函数

可以看出模型的损失函数整体趋势随着训练步数的增加而降低，说明模型朝预计的目标方向学习，算法收敛。

图 6 为 PPO 算法的平均奖励曲线，纵轴代表每轮中所有步数的奖励之和，奖励随着训练步数的增加而增大，在前 150 000 步的训练中，提升幅度较大，快速得到局部最优值，后 250 000 步的训练奖励值提升幅度较小，花销大量时间寻找最优奖励。与 5 节点相比，4 节点奖励曲线增幅大致相同，比 5 节点快 100 000 步达到收敛得到最优解。

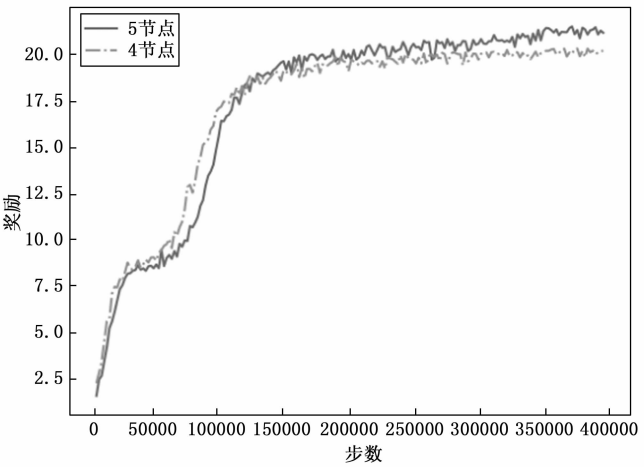

图 6 奖励函数

最终算法输出如下时隙分配矩阵：

$$\boldsymbol{A}_1=\begin{pmatrix}0&1&0&0&1\\0&1&0&0&1\\0&1&0&0&1\\1&0&0&1&0\\0&0&1&0&0\\0&1&0&0&1\\0&1&0&0&1\\0&1&0&0&1\end{pmatrix}$$

将负载代入公式（2）中计算加权信道利用率可得结果为 21.875%。

其中 200 000 步时的局部最优解的时隙分配矩阵为：

$$\boldsymbol{A}_2=\begin{pmatrix}0&1&0&0&1\\0&1&0&0&1\\0&1&0&0&1\\1&0&0&1&0\\0&0&1&0&0\\0&1&0&0&1\\0&1&0&0&1\\0&0&1&0&0\end{pmatrix}$$

计算加权信道利用率为 21.458%。使用贪心算法进行对比，优先遍历分配时隙，满足一个周期至少开通一次的约束，再选择载荷最大的节点分配时隙，得出如下时隙分配矩阵：

$$\boldsymbol{A}_3 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

可知加权信道利用率为 14.791%。

图 7 为三种时隙分配矩阵的数据对比，其中 PPO 局部最优为 200 000 步时的时隙分配情况，由此可见 PPO 算法在前 200 000 步即可得到局部最优解，但寻找到最优解需要花费翻倍的时间，以 5 节点网络为例，最优解比局部最优解提高了 1.94%，训练步数多出 200 000 步，因此 PPO 算法最优解与局部最优解存在折中关系，可以根据实际需求适当缩短训练时间得到加权信道利用率略微低于最优解的时隙分配方案。

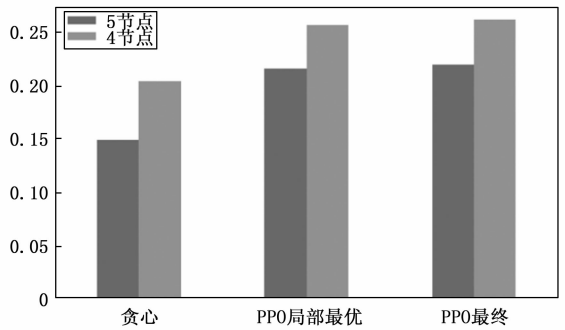

图 7　贪心算法、PPO 算法局部最优、最优解加权信道利用率对比

与贪心算法对比可以看出 PPO 算法的局部最优解和最优解都显著优于贪心算法，与贪心算法相比，PPO 算法局部最优解加权信道利用率提高了 45.07%，最优解集权信道利用率提高了 47.89%。

图 8 为节点数为 4、5，需分配的时隙数为 5～30 的 PPO 算法与贪心算法的加权信道利用率对比。当节点数固定时，PPO 算法得到的加权信道利用率随时隙的增加而增大；相比于 5 节点，4 节点网络下 PPO 算法得到的加权信道利用率更高；PPO 算法比贪心算法加权信道利用率提高了 45%左右。

通过仿真结果对 PPO 算法在时隙分配优化方面的有效性进行分析，该算法能够收敛到最优，最优解与局部最优解间存在折中关系，可以根据实际需求来选取，对比贪心算法，PPO 算法有巨大优势。

4　结束语

近些年来随着多无人机网络的研究与应用的不断深入，多无人机网络资源分配问题具有很重要的研究意义。本文根据多无人机网络差异化节点负载的特点，为了适配网络业务流量，针对多无人机网络动态时隙分配问题，将网络拓扑结构纳入考虑，提出了基于 PPO 算法的时隙分配方案。仿真结果表明，相比于贪心算法，提出的深度强化学习方案，有效提高了信道利用率。

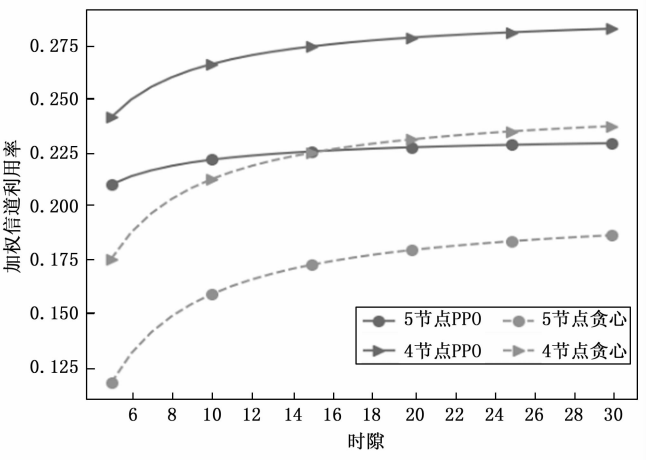

图 8　贪心算法、PPO 算法加权信道利用率对比

参考文献：

[1] LIU DX，XU YH，WANG JL，et al. Opportunistic UAV utilization in wireless networks：motivations，applications，and challenges [J]. IEEE Communications Magazine，2020，58 (5)：62－68.

[2] YU XY，WU DC，LIU DX，et al. The Heterogeneous Demands Satisfaction in IoT Network：Air-Ground Collaborative Deployment [J]. IEEE Transactions on Vehicular Technology，2021，70 (12)：12713－12724.

[3] 刘　育，孙见忠，李　航. 民用无人机的监管与规范探讨 [J]. 南京航空航天大学学报，2017 (201)：152－157.

[4] 梁晓龙，侯岳奇，胡利平，等. 无人集群试验评估研究现状分析及理论方法 [J]. 南京航空航天大学学报，2020，52 (6)：846－854.

[5] CULLENBINE C A. A tabu search approach to the weapons assignment model [D]. Air Force Institute of Technology，2000.

[6] 于　江，贺赛飞，张凤霞，等. 模拟退火算法在战场频率资源分配中的应用 [J]. 中国无线电，2018 (1)：34－38.

[7] SIKANEN T. Solving weapon target assignment problem with dynamic programming [J]. Independent research projects in applied mathematics，2008，32.

[8] 向子权，杨家其，李慧琳，等. 基于离散灰狼算法的资源分配问题求解 [J]. 华中科技大学学报（自然科学版），2021，49：81－85.

[9] 余　涛，王宇名，叶文加，刘前进. 基于改进分层强化学习的 CPS 指令多目标动态优化分配算法 [J]. 中国电机工程学报，2011 (19)：90－96.

[10] 阎　栋，苏　航，朱　军. 基于 DQN 的反舰导弹火力分配方法研究 [J]. 导航定位与授时，2019，6 (5)：18－24.

[11] 丁振林，刘冠龙，谢　艺，等. 基于强化学习与神经网络的动态目标分配算法 [J]. 电子设计工程，2020，28 (13)：54－60.

（下转第 311 页）