doi:10.3969/j.issn.1003-3106.2021.05.004
引用格式:轩书哲,柯良军.基于多智能体强化学习的无人机集群攻防对抗策略研究[J].无线电工程,2021,51(5):360-366.
[XUAN Shuzhe,KE Liangjun.Study on Attack-Defense Countermeasure of UAV Swarms Based on Multi-agent Reinforcement Learning[J].Radio Engineering,2021,51(5):360-366.]

基于多智能体强化学习的无人机集群攻防对抗策略研究

轩书哲[1,2],柯良军[1,2]
(1. 机械制造系统工程国家重点实验室,陕西 西安 710049;
2. 西安交通大学 自动化科学与工程学院,陕西 西安 710049)

摘　要: 针对大规模无人机集群攻防对抗问题,提出了一种基于近端策略优化(Proximal Policy Optimization,PPO)的改进多智能体(Multi-agent Proximal Policy Optimization,M-PPO)算法。该算法采用了Actor-Critic框架,但与PPO不同,为实现智能体之间的协作,算法使用了包含全局信息的Critic网络和局部信息的Actor网络。此外,算法采用了集中训练、分散执行的框架,训练得到的模型能够在不依赖通信的基础上实现协作。为了研究该算法的性能,设计了一个考虑无人机飞行约束和真实飞行环境的大型无人机集群对抗平台,并进行仿真实验。实验结果表明,M-PPO算法在攻防对抗问题中的效果显著优于PPO和深度确定性策略梯度(Deep Deterministic Policy Gradient,DDPG)等主流算法。

关键词: 无人机;攻防对抗;多智能体强化学习;三维环境
中图分类号: TP391.4　　**文献标志码:** A　　**开放科学(资源服务)标识码(OSID):**
文 章 编 号: 1003-3106(2021)05-0360-07

Study on Attack-Defense Countermeasure of UAV Swarms Based on Multi-agent Reinforcement Learning

XUAN Shuzhe[1,2], KE Liangjun[1,2]
(1. *State Key Laboratory for Manufacturing Systems Engineering, Xi'an* 710049, *China*;
2. *School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an* 710049, *China*)

Abstract: In order to solve the problem of attack-defense countermeasure of large-scale unmanned aerial vehicle (UAV) swarm, an improved Multi-agent algorithm (Multi-agent Proximal Policy Optimization, M-PPO) based on proximal policy optimization algorithm (PPO) is proposed. The algorithm uses Actor-Critic framework. Unlike PPO, M-PPO uses the Critic network with global information and the Actor network with local information to achieve the cooperation between agents. In addition, the algorithm adopts the framework of centralized training and decentralized execution. The trained model can achieve cooperation without communication. In order to study the performance of the algorithm, a large UAV swarm countermeasure platform considering UAV flight constraints and real flight environment is designed. The experimental results show that M-PPO algorithm is better than PPO algorithm and deep deterministic policy gradient (DDPG) algorithm.

Keywords: UAV; attack-defense countermeasure; multi-agent reinforcement learning; three-dimensional environment

0　引言

无人机集群是一个由多架无人机相互协作、执行共同任务的统一系统。近年来,无人机集群技术得到了极大的发展,在搜索与救援、巡逻与监视、消防与作战等领域得到了广泛应用。相较于单无人机系统,无人机集群拥有更加强大的环境适应能力和更高的控制冗余度,能够协同完成更复杂的任务。

作为一种具有代表性的无人机集群系统,无人机集群攻防对抗系统规模庞大、系统复杂、具有非常高的系统随机性和状态不确定性。无人机集群攻防对抗问题的本质是一种"疆土防御"问题[1]和"追捕逃脱"问题[2]。在该问题中入侵者需要尽可能地接近目标领地,而防御者试图拦截入侵者的入侵。环境中的无人机既要考虑自身个体的自治与发挥,又

收稿日期:2020-11-25
基金项目:国家自然科学基金资助项目(61973244,61573277)
Foundation Item: Project Supported by the National Natural Science Foundation (61973244,61573277)

要考虑无人机之间的交流与合作。在一个连续且动态变化的环境中，如何教会无人机协同工作具有十分重要的研究意义。

现有的无人机集群对抗方法有基于微分博弈的方法、基于专家系统的方法和基于引导率的方法等，这些方法在简单的、小规模静态环境中拥有较好的表现，但无法适用于规模较大的、复杂未知的场景。近年来，随着深度强化学习技术的提出和发展，强化学习被广泛应用于智能体对抗领域，如雅达利游戏和围棋等。与微分博弈、专家系统等传统的集群算法相比，强化学习不依赖于环境模型，主要通过学习奖励函数优化自身的策略，在大规模、复杂的环境中具有更大的优势。运用强化学习相关理论解决无人机集群对抗问题是一种具有广泛前途的方法。例如文献[3]首先提出并研究了具有2组对立任务的智能体决策问题，随后文献[4]提出随机博弈的决策问题可转化为多智能体强化学习问题。在微分博弈理论基础上，强化学习算法被用于解决“疆土防御”问题[5-7]。

为解决连续动作空间的大规模无人机集群攻防对抗问题，本文基于近端策略优化(Proximal Policy Optimization，PPO)算法的思想，提出了一种改进的多智能体强化学习算法M-PPO。算法使用了一个集中式的Critic框架和分布式的Actor框架。其中Critic网络包含了所有无人机的信息，保证智能体能够学习到无人机之间的协作；Actor网络仅依赖于自身观测值进行决策。同时集中训练、分散执行的方法使无人机仅在训练阶段进行通信，减少了通信开销。为了研究M-PPO算法，设计并开发了一个大型无人机集群攻防对抗平台，模拟无人机的飞行约束和真实的飞行环境。利用仿真环境，将M-PPO与其他几种常用的强化学习算法进行了比较，验证了本文所提算法的优越性。

1 相关工作

1.1 传统无人机集群攻防对抗算法

无人机集群攻防对抗问题的复杂性吸引了众多学者的关注。目前主流的无人机对抗算法包括引导率方法、微分博弈方法和专家系统法等，但这些方法都有一定的局限性。引导率方法是引导无人机到达指定目标点的算法，虽然其具有简单易实现的特点，但需要提前获知对方的控制策略或者要求对方具有相对固定的运动模式，因此难以适用于复杂的多智能体环境；微分博弈方法能够在智能体没有最优策略先验知识的前提下学习如何行动，但此种方法具有状态量多、微分方程复杂、方程解析式求解困难等问题；专家系统法是借助计算机模拟人类专家处理复杂问题的方法，其核心是根据相关领域的人类专家获取的先验知识建立其系统模型，并根据当前无人机的状态，通过模糊匹配等方法选择知识库中事先定义好的行动策略。专家系统法拥有悠久的研究历史和较为成熟的研究方案，应用较为广泛，但是其依赖于大量的人类专家制定的针对性规则，一旦环境发生调整，规则必须重新制定，可移植性较差。此外，专家系统法只能从事先定义好的规则库中选取规则，无法保证决策的正确性。

1.2 基于强化学习的集群攻防对抗算法

强化学习是一种不依赖模型和任何先验信息，通过不断“试错”和获得的奖励来优化自身行为的方法。在单智能体对抗问题中，一种具有代表性的强化学习方法是Q-Learning算法[8]，它通过表格的形式记录环境的所有状态-行为价值函数Q，并根据ε-贪婪策略选择动作。但该方法只适用于状态和动作空间离散的小规模问题。深度强化学习(Deep Q Network，DQN)算法[9]使用神经网络代替Q-Learning中的表格来拟合状态-行为价值函数，并使用经验回放池、双网络等技巧，成功将Q-Learning算法应用在连续状态空间中，在单智能体对抗问题上得到了广泛的应用，但仍然无法解决连续动作空间问题。

DDPG算法[10]是一种用于解决连续动作空间的强化学习算法。该算法基于策略梯度(Policy Gradient)框架，在动作连续的环境中取得了较为理想的效果，但当环境的动作空间较大时，算法往往难以收敛。另一方面，DDPG算法采用确定性策略，一个确定的状态s下只能采取一种动作，探索能力较差。

PPO算法[11]是OpenAI提出的另一种策略梯度算法。通过在损失函数中添加惩罚项来约束策略更新的幅度，PPO能够在复杂的场景中快速学习到正确的策略，被广泛应用于各种离散和连续的动作空间问题中[12]。此外，相较于DDPG，PPO使用了随机策略，基于动作的概率分布选取动作，能够实现更好的探索。

在处理无人机集群攻防对抗问题时，一种常用的方法是将多智能体问题直接建模成单智能体问题。这些方法通常假设一个统一的顶端智能体，该顶端智能体接收所有无人机的状态并输出动作值。但是随着智能体规模的扩大，问题的状态空间和动作空间维度指数增加，将造成维度灾难。同时无人机之间需要实时通信，会产生庞大的通信开销。

另一种处理多智能体问题的基本方法是将每个智能体视为一个独立的个体，如 Independent Q-Learning[13]。在该算法中，每个智能体只处理自己获取的信息，因此智能体之间是完全独立的。但是这些方法不仅没有考虑智能体之间的相互影响，也不能满足强化学习中的独立性要求，在复杂的场景中表现不佳。

2 问题描述与系统建模

本文考虑了一种无人机集群攻防对抗场景，场景模型如图 1 所示。

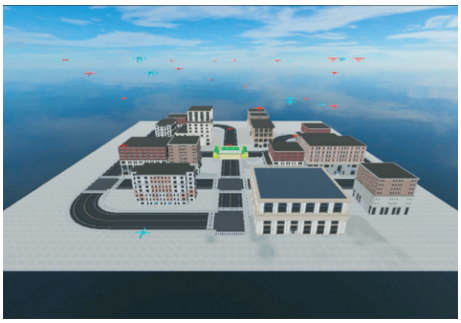

图 1 三维连续空间无人机环境模型

Fig.1 3D continuous space UAV environment model

红色无人机为进攻无人机，蓝色无人机为防御无人机，双方在一定范围内的建筑群中围绕目标区域（场景中心体育场）展开对抗。进攻无人机在保护自身的前提下试图入侵目标区域；防御无人机的目标则是阻止目标区域被侵入并尽可能地摧毁敌方。假定进攻无人机数量为 I，防御无人机数量为 J，所有无人机都是同构的，拥有相同的性能参数。无人机在飞行时要服从以下约束：

（1）初始坐标约束

场景中防御无人机在目标区域一定距离内随机产生，进攻无人机在目标区域一定距离外随机出现。对于进攻无人机 i 和防御无人机 j，其初始时刻到目标区域 g 的距离分别为 $d_{i,g}$ 和 $d_{j,g}$，$d_{i,g}$ 和 $d_{j,g}$ 应满足：

$$d_{i,g} \geqslant d_{\text{init}}\ , \tag{1}$$

$$d_{j,g} \leqslant d_{\text{init}}\ , \tag{2}$$

式中，d_{init} 为给定的初始安全距离。

（2）高度与边界约束

无人机飞行过程中受到高度限制，飞行高度过高或过低将受到惩罚。其飞行高度需满足如下约束：

$$h_{\min} \leqslant h \leqslant h_{\max}\ 。 \tag{3}$$

此外，建筑群四周是有界的，无人机不能超出其范围。

（3）速度与加速度约束

由于无人机避障的要求以及自身机能的限制，无人机的速度和加速度不可能无限大。在三维空间中，无人机的速度和加速度需满足最大值约束：

$$\left| v_{x,y,z} \right| \leqslant v_{\max_{x,y,z}}\ , \tag{4}$$

$$\left| a_{x,y,z} \right| \leqslant a_{\max_{x,y,z}}\ 。 \tag{5}$$

（4）最大偏航角约束

飞行过程中，无人机通过调整偏航角进行转向。发动机性能和机身气动结构的限制使无人机的偏航角无法达到 $\pm 90°$，而是存在一个最大范围。假设无人机航迹点 i 的坐标为 (x_i, y_i, z_i)，则从点 $i-1$ 到点 i 的航迹段的水平投影为 $\alpha_i = (x_i - x_{i-1}, y_i - y_{i-1})^{\mathrm{T}}$，那么最大偏航角 ϕ 约束为：

$$\cos\phi \leqslant \frac{\alpha_i^{\mathrm{T}}\alpha_{i+1}}{\|\alpha_i\| \cdot \|\alpha_{i+1}\|}, (i = 2,3,\cdots,n-1)\ 。 \tag{6}$$

（5）障碍物约束

环境中存在形状、大小各异的房屋障碍物。障碍物的坐标是随机的，只有当无人机与障碍物足够接近时，才能发现障碍物。飞行过程中无人机不能越过障碍物也不能与障碍物发生碰撞。一旦发生碰撞，则无人机被摧毁。无人机与障碍物的距离 l 应满足：

$$l \geqslant R_{\text{safe}} + l_{\min} + R_{\text{UAV}}\ , \tag{7}$$

式中，R_{safe} 为规定的安全距离；$l_{\min}$ 为障碍物在无人机方向上的长度；R_{UAV} 为无人机半径。

场景中无人机可通过雷达设备侦测到自身范围内一定数量的敌方和己方单位坐标，防御无人机可摧毁自身攻击范围内的敌方单位，当一架进攻无人机至少暴露在 K 架防御无人机的攻击范围内时，进攻无人机被摧毁。此外，考虑到能量消耗，无人机的飞行最大时长为 T。T 时刻内，任何一架进攻无人机足够接近目标区域，则进攻方获胜；所有进

攻无人机被摧毁或者无任何无人机入侵成功，则防御方获胜。

3 基于 M-PPO 的无人机集群对抗

3.1 马尔科夫决策过程

对于一个强化学习环境，可以使用马尔科夫决策过程（Markov Decision Process，MDP）描述智能体与环境的交互过程。一个 MDP 由一个五元组 $<S, A, P, R, \gamma>$ 组成，其中 S 和 A 表示状态空间和动作空间，$P: S \times A \rightarrow S$ 表示状态转移概率矩阵，$R: S \times A \times S \rightarrow [r_{\min}, r_{\max}]$ 表示即时奖励，$\gamma \in [0,1]$ 为回报折扣因子。在任意时刻 t，智能体 i 根据当前时刻状态 $s_t \in S$ 和策略 $\pi(a \mid s)$ 选择动作 $a_t \in A$，并根据状态转移矩阵 P 到达下一时刻状态 $s_{t+1} \in S$，同时得到对应的奖励 $r_t \in R$。智能体重复以上行为直到终止。

3.2 PPO 算法

PPO 是一类新型的策略梯度算法。策略梯度算法中智能体通过梯度上升的方式迭代更新策略 π 从而最大化期望回报 $\sum_t \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}[r(s_t,a_t)]$。传统的策略梯度算法，如 DDPG、AC、A3C 等，成功地在一系列控制与决策问题中取得甚至超过人类专家的水平，使得强化学习算法的适用性大大提高。但是此类方法对迭代步长十分敏感，解决实际问题时很难选择合适的步长。另一方面，这类方法采样效率较低，一个简单的问题往往需要大规模的采样。

为了解决以上问题，PPO 算法将前后策略输出动作的概率比值作为策略更新的依据，最大化目标函数：

$$L^{\mathrm{CPI}}(\theta) = \mathbb{E}_t\left[\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t \mid s_t)} A_t\right] = \mathbb{E}_t[r_t(\theta) A_t], \quad (8)$$

式中，A_t 为 t 时刻的优势函数估计值；π 是一个随机策略；θ_{old}，θ 分别表示策略 π 更新前后的参数。该目标函数可以解释为 PPO 尝试最大化相对于均值而言取得更大优势的动作的概率，并最小化取得更大劣势动作的概率。同时，为了防止策略更新幅度过大，算法引入约束项来限制策略更新。一种常用的限制方法是使用 KL 散度约束。约束后的目标函数为：

$$L^{\mathrm{KLPEN}}(\theta) = \mathbb{E}_t[r_t(\theta) A_t - \beta \mathrm{KL}[\pi_{\theta_{old}}(\cdot \mid s_t), \pi_\theta(\cdot \mid s_t)]], \quad (9)$$

式中，β 是自适应 KL 惩罚系数，在每次策略更新后被更新。

在实际应用中，研究人员发现，使用截断的方法进行约束能够取得更好的效果，此时策略更新的目标函数为：

$$L^{\mathrm{CLIP}}(\theta) = \mathbb{E}_t[\min(r_t(\theta) A_t, \mathrm{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) A_t)], \quad (10)$$

式中，ϵ 是一个超参数，用来限制策略更新的范围；$\mathrm{clip}(\cdots)$ 是截断函数，将 $r_t(\theta)$ 限制在 $[1-\epsilon, 1+\epsilon]$ 之间，$\min(\cdots)$ 函数用于选择二者中的最小值（即悲观边界）。

PPO 算法的核心思想是通过约束策略更新的幅度解决策略梯度算法中步长难以确定的问题，同时使用重要性采样提高样本利用效率，大大降低了算法的调试难度。

3.3 M-PPO 算法

传统的强化学习算法很难直接应用在多智能体环境中，一个重要的原因是训练过程中每个智能体都在不断变换更新。此时对于每个智能体而言，外部环境都是不稳定的，即对于任意的 $\pi_i \neq \pi'_i$，存在 $P(s' \mid s, a, \pi_1, \cdots, \pi_n) \neq P(s' \mid s, a, \pi'_1, \cdots, \pi'_n)$。在多智能体环境中，如果智能体 i 仅仅将其他智能体视为环境的一部分，将会忽略其他智能体的动态性，无法保证算法的收敛。

为解决这一问题，将 PPO 算法扩展到多智能体环境中，本文在 PPO 的基础上修改了 Critic 网络，在训练中引入可以观察全局的 Critic 来指导 Actor 的训练，从而将不可预测的环境转换成可预测的环境。同时，为进一步减小通信开销，M-PPO 采用了集中式 Critic 和分布式 Actor 策略，智能体之间拥有自己独立的 Actor 网络和一个共享的 Critic 网络。M-PPO 算法的结构如图 2 所示。

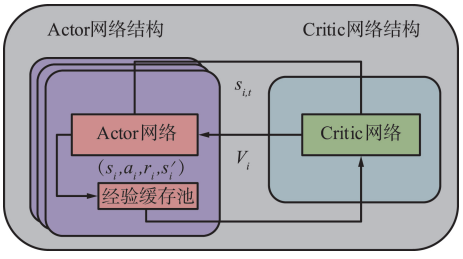

图 2 M-PPO 算法结构

Fig.2 Algorithm structure of M-PPO

对于含有 n 个无人机的强化学习环境，M-PPO

算法包含了 n 个 Actor 网络和 1 个 Critic 网络。对于智能体 i，t 时刻时其自身的局部状态值为 s_i，Actor 网络通过自身的局部观测值输出对应动作概率分布的均值 μ 和方差 σ，然后通过对 μ 和 σ 构建的正态分布函数采样得到最终的动作 a_i。环境执行动作 a_i 并将相关信息 (s_i, a_i, r_i, s_i') 存储在经验回放池中。

智能体与环境交互一定次数后，停止交互并随机从经验回放池中采样进行网络的训练。训练时 Critic 的输入为所有智能体的状态 $S(s_1, s_2, \cdots, s_n)$，输出为智能体对应的状态价值 $V(V_1, V_2, \cdots, V_n)$，其优化的损失函数为:

$$L^V = \sum_{i=0}^{n} \left(V_{i,\theta}(s_t) - V_{i,t}^{\text{targ}}\right)^2, \tag{11}$$

式中，$V_{i,\theta}$ 表示智能体 i 的 Critic 网络输出值; $V_{i,t}^{\text{targ}}$ 是通过贝尔曼方程计算得到的目标价值。Actor 网络的损失函数如式(10)所示。

通过以上方法，无人机在训练阶段进行通信，学习智能体之间的合作; 在执行阶段仅依赖自身局部感知做出动作，从而实现了不依赖于通信的协作策略。此外，为了减小网络训练的开销，智能体之间共享相同的 Actor 网络参数。

3.4 算法元素表示

对于集群中的无人机 i，给出其状态空间、动作空间和奖励函数。

3.4.1 状态空间

对于无人机 i，状态空间 $s_i = \{X_i, V_i, D_i, D_{ij_1}, D_{ij_2}, \cdots, D_{ij_k}\}$，其中 X_i, V_i, D_i 表示自身的坐标、速度和到目标点的距离，D_{ij_k} 表示无人机 i 与敌方无人机 j_k 的距离。状态空间中所有值都是连续有界的。

3.4.2 动作空间

对于无人机 i，动作空间为加速度 $A_i(a_{i,x}, a_{i,y}, a_{i,z})$，加速度的取值连续有界。由于无人机的飞行约束和障碍物的限制，不同时刻的可选择动作是不同的，无人机只能从当前可选择动作空间中选取动作。

3.4.3 奖励函数

由于无人机攻防对抗问题的独特性，无人机只有在回合结束时才可获得一个明确的奖励值，这种延迟奖励将极大拖慢智能体学习的速度。为加快学习速度，引入智能体到目标点的距离作为回合过程中的奖励值。

回合过程中，对于坐标点为 (x_i, y_i, z_i) 的进攻无人机 i，其目标区域用 (x', y', z') 为中心，r 为半径的球表示，则奖励 $R_{1,i}$ 可以定义为:

$$R_{1,i} = -\left(\sqrt{(x_i - x')^2 + (y_i - y')^2 + (z_i - z')^2} - r\right)。 \tag{12}$$

类似的，对于防御无人机 j，其奖励 $R_{2,j}$ 定义为:

$$R_{2,j} = -\frac{1}{K}\sum_{i=1}^{K} R_{1,i}, \tag{13}$$

式中，$R_{1,i}$ 表示防御无人机周围第 i 个进攻无人机的奖励值。

回合结束时，胜利方将获得一较大的正奖励，失败方将获得一绝对值较大的负奖励，此时的奖励 R 表示为:

$$\begin{cases} R_{\text{suc}} = \text{M} \\ R_{\text{fail}} = -\text{N} \end{cases}, \tag{14}$$

式中，M 和 N 为正整数。

4 实验验证

在实验中本文采用了含有 30 架进攻无人机和 20 架防御无人机的对抗场景，并使用第 3 节中介绍的仿真场景进行测试。每架无人机最多能感知周围 3 架其他无人机的坐标。无人机在三维空间中移动，最大移动速度为 10 m/s，场景中长、宽、高分别为 1 000，1 000，500 m。实验中 M-PPO 算法分别使用 2 个相互独立的多层神经网络表示 Actor 和 Critic，并使用 Adam 优化器进行梯度的更新。其中 Actor 网络隐藏层神经元个数分别为 64 和 64，Critic 网络隐藏层神经元个数分别为 128 和 64。所有网络的隐藏层均使用了 ReLU 激活函数。算法的一些超参数设置如表 1 所示。

表 1 M-PPO 算法超参数

Tab.1 Hyper-parameters of M-PPO

参数	值
学习率(Learning Rate)	0.005
截断常数 ϵ	0.2
批次大小(Batch Size)	512
回合步长	400
回报折扣因子 γ	0.95

将 M-PPO 与 2 种主流的强化学习算法进行了比较，对比算法的网络结构与超参数与 M-PPO 保持一致。2 种对比算法分别为: 智能体之间相互独立的 PPO 算法(记作 I-PPO)和智能体之间相互独立

的 DDPG 算法(记作 I-DDPG)。分别使用 3 种算法训练环境中的防御无人机,并使用事先用 I-DDPG 训练好的网络模型控制进攻无人机。场景进行 2 000回合的训练,每回合步长为 400 步,共 80 万步。训练曲线如图 3 所示。

由图 3 可以看出,在经过 80 万步的训练后,3 种算法均成功收敛,其中 M-PPO 算法在训练 10 万步后得到收敛,而 I-PPO 和 I-DDPG 算法则收敛在 20 万步之后,这表明 I-PPO 和 I-DDPG 算法需要更长时间的探索与尝试。此外,相较于 I-DDPG 算法,I-PPO 算法和 M-PPO 算法在实验中取得了更高的平均奖励值。

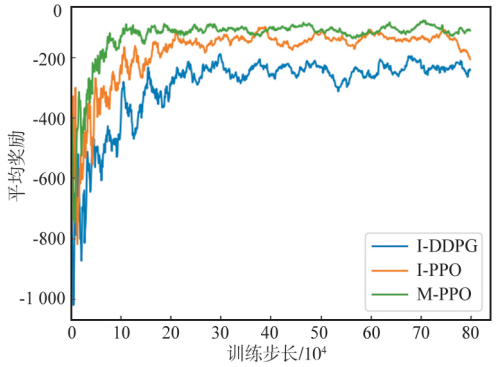

图 3 不同算法下防御无人机集群平均奖励

Fig.3 Average reward for defense UAV swarm under different algorithms

训练过程中每隔 4 000 步对不同算法进行 50 个回合的测试并记录防御无人机的防御成功率,结果如图 4 所示。由图 4 可知,收敛后的 M-PPO 算法比 I-PPO、I-DDPG 算法拥有更高、更稳定的成功率。其中 M-PPO 算法在 10 万步后成功率保持稳定,I-PPO 和 I-DDPG 算法在 20 万步后成功率稳定,与图 3 中奖励曲线相一致。

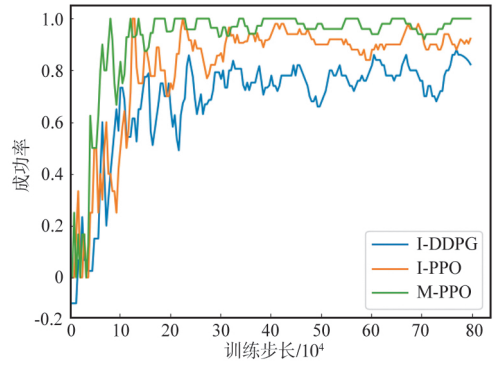

图 4 不同算法下防御无人机测试成功率

Fig.4 Success rate of defense UAV under different algorithms

5 结束语

本文研究了无人机集群攻防对抗场景,设计了一个基于真实环境约束的大规模无人机集群仿真平台,并提出了一种基于近端策略优化的改进多智能体强化学习算法 M-PPO。该算法使用了全局感知的 Critic 网络和局部感知的 Actor 网络,并使用了集中训练、分散执行的框架,训练后的网络能够在无通信的条件下学会合作。实验结果表明,相较于智能体相互独立的 PPO 和 DDPG 算法,M-PPO 算法训练所需时间更短,训练之后的成功率更高。

参考文献

[1] ISAACS R. Differential Games: a Mathematical Theory with Applications to Warfare and Pursuit,Control and Optimization [M] .New York: Wiley,1965.

[2] TABACHNIKOV S.Chases and Escapes: The Mathematics of Pursuit and Evasion by Paul J.Nahin [J] .The Mathematical Intelligencer,2009,31(2):78-79.

[3] LITTMAN M L .Markov Games as a Framework for Multi-agent Reinforcement Learning [M] .San Franciso: Morgan Kauffman Publishers,Inc.1994.

[4] BOWLING M,VELOSO M. Multiagent Learning Using a Variable Learning Rate [J] .Artificial Intelligence,2002,136(2):215-250.

[5] HARMON ME,BAIRD L C,KLOPF A H.Reinforcement Learning Applied to a Differential Game [J] .Adaptive Behavior,1995,4(1):3-28.

[6] SMITH A E.Swarm Intelligence: from Natural to Artificial Systems [Book Reviews] [J] .Connection Science,2002,14(2):163-164.

[7] DESOUKY S F,SCHWARTZ H M. Self-learning Fuzzy Logic Controllers for Pursuit-evasion Differential Games [J] .Robotics and Autonomous Systems,2010,59(1):22-33.

[8] WATKINSC J C H,DAYAN P. Technical Note: Q-learning [J] .Machine Learning,1992,8(3-4):279-292.

[9] MNIH V,KAVUKCUOGLU K,SILVER D,et al.Human-level Control through Deep Reinforcement Learning [J] .Nature,2015,518(7540):529-533.

[10] LILLICRAP T P,HUNT J J,PRITZEL A,et al.Continuous Control with Deep Reinforcement Learning [J] .arXiv preprint arXiv: 1509.02971,2015: 1-10.

[11] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal Policy Optimization Algorithms [J]. arXiv preprint arXiv: 1707.06347, 2017: 1-10.
[12] BERNER C, BROCKMAN G, CHAN B, et al. Dota 2 with Large Scale Deep Reinforcement Learning [J]. arXiv preprint arXiv: 1912.06680, 2019: 1-7.
[13] TAN M. Multi-agent Reinforcement Learning: Independent vs. Cooperative Agents [C] // Proceedings of the Tenth International Conference on Machine Learning. Amherst, ICML, 1993: 330-337.

作者简介

轩书哲　男,(1995—),就读于西安交通大学控制工程专业,硕士研究生。主要研究方向:强化学习。

柯良军　男,(1976—),博士,教授。主要研究方向:智能计算、强化学习、无人系统智能感知与协同。

《无线电工程》专题征稿启事
——卫星导航安全与对抗

卫星导航系统是现代国防和国民经济建设的重大基础设施,由于当前电磁环境复杂化和应用场景多样化趋势,卫星导航应用安全面临严峻挑战。近年来,卫星导航安全与对抗技术发展迅速,在多源信息融合定位、卫星导航干扰与抗干扰方法、干扰信号检测与识别技术、导航对抗决策分析等研究中均取得了一定的成果。为集中展现卫星导航安全与对抗领域最新理论成果及技术进展,《无线电工程》期刊计划推出“卫星导航安全与对抗”专题,特向国内外广大专家、学者征集“卫星导航安全与对抗”方面的原创性研究论文。

一、征文范围包括(但不仅限于此)

- 卫星导航安全领域发展现状及最新进展
- 卫星导航抗干扰技术研究进展及方法验证
- 导航欺骗干扰信号仿真、检测与识别技术
- 复杂条件下多源信息融合定位方法
- 导航对抗决策分析
- 导航对抗等电子对抗手段联合攻防设计与效能评估
- 导航对抗相关技术与系统工程、人工智能、大数据等学科的多元融合探索与实践

二、特邀编委

丛佃伟,航天工程大学副教授、硕士生导师。主要研究方向为卫星导航系统性能测试评估、多源信息融合定位及测姿技术。近五年,先后主持国家自然科学基金、军委科技委首批国防科技项目基金等 8 项科研课题,获省部级科技进步奖励 3 项、全国高等学校测绘类专业青年教师讲课竞赛特等奖、国家发明专利及计算机软件著作权各 4 项,出版《GNSS 高动态定位性能检定理论与关键技术研究》专著及《漫话北斗导航》科普读物,荣立三等功 1 次。

三、重要日期

征文截止日期: 2021 年 06 月 20 日

计划出版日期: 2021 年 10 月

四、投稿指南

投稿邮箱(来稿请注明“卫星导航安全与对抗”专题):

1. **丛佃伟副教授:** congdianwei@ sina.com
2. **编辑部:** gch4954@ 163.com

编辑部联系电话: 0311-86924962