中国科学: 信息科学 2024 年 第 54 卷 第 7 期: 1775–1792
SCIENTIA SINICA Informationis
论文

引入反事实基线的无人机集群对抗博弈方法

王尔申[1,2], 陈纪浩[1], 宏晨[3,4*], 刘帆[1], 陈艾东[3,4], 景竑元[3,4]

1. 沈阳航空航天大学电子信息工程学院, 沈阳 110136
2. 沈阳航空航天大学民用航空学院, 沈阳 110136
3. 北京联合大学多智能体系统研究中心, 北京 100101
4. 北京联合大学机器人学院, 北京 100101
* 通信作者. E-mail: hchchina@sina.com, xxthongchen@buu.edu.cn

收稿日期: 2023–10–17; 修回日期: 2024–01–07; 接受日期: 2024–02–19; 网络出版日期: 2024–07–09

国家重点研发计划 (批准号: 2018AAA0100804)、国家自然科学基金 (批准号: 62173237)、北京联合大学科研 (批准号: ZK30202304, SK160202103, ZK50201911, ZK30202107)、卫星导航系统与装备技术国家重点实验室开放基金项目 (批准号: CEPNT2022A01)、辽宁省属本科高校基本科研业务费专项 (批准号: 20240177, 20240215) 和沈阳市科技计划项目 (批准号: 22-322-3-34) 资助

摘要 无人机在协同对抗博弈上的应用越来越广泛和深入, 尤其是无人机集群在协同探测、全域对抗、策略骗扰等对抗任务中, 发挥着越来越重要作用, 可靠高效的无人机集群博弈方法是当前的研究热点. 本文将反事实基线思想引入到无人机集群对抗博弈环境, 提出一种基于反事实多智能体策略梯度 (counterfactual multi-agent policy gradients, COMA) 的无人机集群对抗博弈方法; 在具有无限连续状态、动作的无人机对抗环境中, 基于无人机动力学模型, 设置符合实际环境的击敌条件和奖励函数, 构建基于多智能体深度强化学习的无人机集群对抗博弈模型. 红蓝双方无人机集群采取不同的对抗博弈方法, 利用多智能体粒子群环境 (multi-agent particle environment, MPE) 对红蓝双方无人机集群进行非对称性对抗实验, 实验结果表明平均累积奖励能够收敛到纳什均衡, 在解决 4 vs. 8 的对抗决策问题方面, COMA 方法的平均命中率较 DQN 和 MADDPG 分别提升 39% 和 17%, 在平均胜率方面比 DQN 和 MADDPG 分别提升 34% 和 17%. 最后, 通过对 COMA 方法的收敛性和稳定性的深入分析, 保证了 COMA 方法在无人机集群对抗博弈任务上的实用性和鲁棒性.

关键词 无人机集群, 对抗博弈, 多智能体, 深度强化学习, 纳什均衡

1 引言

无人机 (unmanned aerial vehicle, UAV) 集群是由若干配备多种任务载荷的低成本小型无人机组成的无人化对抗博弈系统, 通过自主学习共同完成特定复杂对抗任务. 作为典型的多智能体系统, 无人机集群以难防御、强进攻、低成本、自主学习, 使用灵活等优势深刻改变着集群对抗博弈模式[1~4]. 随

引用格式: 王尔申, 陈纪浩, 宏晨, 等. 引入反事实基线的无人机集群对抗博弈方法. 中国科学: 信息科学, 2024, 54: 1775–1792, doi: 10.1360/SSI-2023-0305
Wang E S, Chen J H, Hong C, et al. UAV swarm adversarial game method with a counterfactual baseline (in Chinese). Sci Sin Inform, 2024, 54: 1775–1792, doi: 10.1360/SSI-2023-0305

着无人机智能化水平的提高和集群控制技术的飞速发展, 无人机集群对抗自主决策方法将成为未来无人机对抗博弈的关键技术.

解决无人机集群对抗自主决策问题的一种思路是利用进化方法, 进化方法是一类受生物进化理论启发而形成的计算方法, 常用于解决优化、搜索和对抗等问题, 其核心思想是通过模拟生物进化的过程, 找到问题的最优或次优解. Kaneshige 等 [5] 使用人工免疫机制解决空中机动选择问题, 将敌机视为抗原, 通过相对位置速度表征, 将机动动作视为抗体, 利用遗传算法和进化算法模仿免疫系统对应抗原的自适应能力, 这种机制使得智能体具有较强的记忆能力, 能记录过往成功的经验以在相似场景下快速响应. Duan 等 [6] 提出了一种基于捕食者 – 猎物粒子群优化 (predator-prey particle swarm optimization, PP-PSO) 的博弈方法, 将多无人机对抗任务建模为双人博弈, 并通过 PP-PSO 方法来解决. Zhou 等 [7] 针对多无人机协同飞抵目标空域完成对抗任务的问题进行了建模, 利用蚁群算法和所提的多无人机控制算法进行仿真实验, 实验表明该算法能有效提升无人机集群对抗获胜率. Isler 等 [8] 将随机策略与狮子追捕策略结合, 研究了两个追踪者对一个高速运动逃跑者的协同追捕算法, 在简单连通多边形环境中验证了所提算法的有效性. Chen 等 [9] 利用模糊规则对多无人机对抗问题进行离散化, 并采用粒子群优化方法求解纳什均衡 (Nash equilibrium), 该方法解决了协同博弈问题, 模拟结果呈现了该方法的可行性和有效性. 然而, 用进化方法解决博弈问题需要固定一个策略并且和对手博弈多次, 或者与对手的仿真模型进行大量模拟博弈. 尽管获胜频率作为该策略获胜概率的无偏估计, 可用于指导下一轮策略选择, 然而每一次策略的调整都源于多次博弈. 仅有每一轮比赛的最终结果会被纳入考虑, 而博弈过程中的中间事件将被忽略. 如果对抗获胜, 就会认为这次对抗中所有的动作都有功劳, 而与每一步具体动作有多关键无关. 这些功劳甚至会被分配给那些从未出现的动作. 因此, 进化方法在面对多智能体长时间持续性对抗任务时能力略显不足.

解决无人机集群对抗自主决策问题的另一种思路是利用强化学习方法 [10]. 强化学习是一种对目标导向与决策问题进行理解并自动化处理的计算方法, 它常用马尔可夫 (Markov) 决策过程建立数学模型, 已在解决智能决策方面表现出不俗能力和良好发展态势, 特别是在复杂动态博弈环境中. 强化学习在智能体和环境交互的灵活性方面具备天然优势 [11]. 强化学习利用智能体与环境的直接交互来学习, 不需要可仿效的监督信号和对周围环境的完全建模, 在解决持续性复杂决策任务时有较大优势. 多智能体强化学习是强化学习的一个分支, 其研究多个智能体在共享环境中相互作用, 并通过智能体的学习来实现其目标. 无人机集群属于典型的多智能体系统, 与单智能体强化学习相比, 多智能体强化学习的复杂度更高、更难以训练: 一方面随着智能体数量的增加, 相应的策略空间呈指数级增加, 其难度远超围棋等棋类游戏; 另一方面随着异构智能体的加入, 多智能体间需要更高效和可靠的通信、协作和配合.

近年来, 随着 AlphaGo [12], AlphaGo Zero [13], AlphaZero [14], AlphaStar [15], AlphaFold [16] 等深度强化学习 (deep reinforcement learning, DRL) 方法的出现, 深度强化学习已成为一个热门的研究方向. Deepmind 提出了基于值方法的深度 Q 网络 (deep Q-networks, DQN) [17], 率先将深度神经网络与 Q-Learning 相结合, 为深度强化学习的发展奠定了坚实基础. 随后产生了许多基于 DQN 的变种, 如 Dueling DQN [18], Double DQN [19] 等, 并获得了更好性能.

针对无人机集群博弈的复杂性和强化学习自身特点, 一些学者应用强化学习对无人机集群博弈进行了研究. Gong 等 [20] 针对多无人机协同对抗问题, 建立了多无人机对抗环境. 提出了一种基于网络化分散的部分可观测马尔可夫决策过程 (networked decentralized partially observable Markov decision process, NDec-POMDP) 的对抗协同策略框架, 仿真结果验证了所提协同对抗决策框架的可行性和有效性. Chen 等 [21] 基于多智能体强化学习理论, 建立多无人机协同攻防演化模型, 提出一种多无人机

协同攻防自主决策方法, 提高了多无人机攻防对抗的效能. Li 等 [22] 基于强化学习的演员 – 评论家框架, 在无人机的演员网络中引入门循环单元, 使得无人机能根据历史决策信息做出合理决策, 采用注意力机制来设计集中式的评论家网络, 并在无人机集群对抗场景中对算法进行了验证. Zhang 等 [23] 提出了一种基于注意力机制的深度强化学习分布式方法, 该方法设计了可用于无人机协作短程对抗任务的奖励函数, 并采用 Unity3D 无人机仿真平台进行了训练.

但是, 在多智能体强化学习环境中, 如果团队内部共同完成一个任务, 则智能体会共享一个奖励函数, 从而带来多智能体的信用分配问题, 即无法区分团队中某个智能体的策略对整个团队任务的贡献. 如果不考虑信用分配问题, 则可能导致智能体学到的策略是局部最优 [24]. 虽然可以为每个智能体设计单独的奖励函数, 但这些单独的奖励在合作环境中并不普遍存在, 也不能鼓励单个智能体为更大的团队利益牺牲, 这将在很大程度上阻碍多智能体在挑战性任务中的学习效率. Foerster 等 [25] 提出了反事实多智能体策略梯度 (counterfactual multi-agent policy gradient, COMA) 方法, 该方法利用反事实基线来减少估计方差, 并解决了多智能体信用分配问题.

事实上, 在无人机集群对抗博弈中, 无人机集群内部往往需要协调和配合, 以提高整体任务完成率. 如何最大化无人机之间的协同, 对信用进行合理分配, 以获得最优的无人机行为策略, 仍是当前需要面对的主要挑战. 本文将 COMA 方法引入到具有无限连续状态和动作的无人机对抗环境中, 基于无人机动力学和攻防态势, 设计符合实际环境的击敌条件和奖励函数, 构建基于多智能体深度强化学习的无人机集群对抗博弈模型. 红蓝双方无人机采取不同的对抗博弈方法, 利用多智能体粒子群环境进行非对称性对抗实验, 使用平均累积奖励、平均命中率和平均胜率作为评价指标. 结果表明平均累积奖励能够收敛到纳什均衡, COMA 方法比其他流行的深度强化学习方法更具优越性, 对 COMA 方法收敛性和稳定性的验证分析保证了其在无人机集群对抗任务上的实用性和鲁棒性.

2 问题描述与建模

2.1 部分可观测马尔可夫博弈

当涉及参与者超过二人时, 可表述为随机博弈 (stochastic game, SG) 模型 [10], 也称为马尔可夫博弈 (Markov game, MG) 模型. 马尔可夫博弈可以表示为一个六元组:

$$\left\langle N, S, A, P, \{R_i\}_{i\in N}, \gamma \right\rangle,$$

其中 N 为参与者数量, 当 N>2 时, 为多智能体部分可观测马尔可夫决策过程; S 是智能体与环境交互的状态集合; i 表示智能体的编号, A_i 是智能体 i 的动作空间, 多智能体联合动作空间为 A, $A = A_1 \times A_2 \times \cdots A_N$; $P : S \times A \times S' \to [0, 1]$ 是环境的状态转移函数, $P(s, a, s')$ 表示通过联合动作 $a \in A$ 从状态 $s \in S$ 到下一个状态 $s' \in S$ 的概率; R_i : 智能体 i 在状态 s 下执行动作 a_i 所获得的奖励; $\gamma \to [0, 1]$ 为折扣因子, 用来表征长期奖励与当前奖励的相对重要性.

部分可观测马尔可夫博弈过程可以描述如下: 在每个时间步 t, 环境都会存在一个状态 s_t, 智能体 i 通过观测概率 Z_i 获得自身的观测状态 $O_{i,t}$, 来执行动作 $a_{i,t}$, 在每个时间步智能体是同时决策的. 我们使用 $(\cdot_i, \cdot_{-i})$ 表示当前智能体 i 与其他 $N-1$ 个智能体. 智能体联合动作 $a_t = (a_{i,t}, a_{-i,t})$ 传入环境, 环境根据 $P(s'|s, a)$ 过渡到下一状态 s_{t+1}, 然后环境反馈给智能体 i 即时奖励 $R_i(s_t, a_t, s_{t+1})$. 智能体 i 的目标是找到一个最优行为策略 $\pi_i(a|o) \in \Pi_i$.

在对抗环境中, 多个智能体同时与环境进行交互, 在交互过程中学习各自的策略, 但是每个智能体不能获得完全信息, 因此将问题建模为部分可观测马尔可夫博弈 (partial observable Markov game,

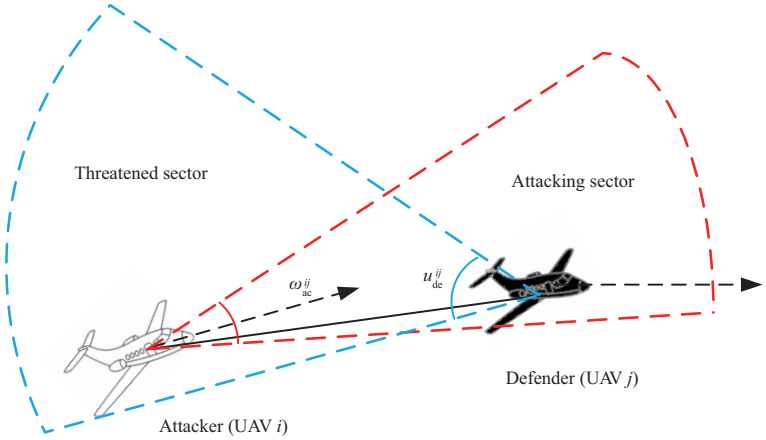

图 1　(网络版彩图) 无人机对抗态势

Figure 1　(Color online) UAV confrontation situation

POMG), 博弈的每个智能体的目标是最大化各自的累积折扣奖励.

2.2　无人机集群对抗环境

设定无人机的攻击方和防守方均采用三自由度模型, 故为了建立无人机在二维环境中的动力学模型, 需要给出无人机的滚转角、航向角、X 轴和 Y 轴的速度变化以及 X 轴和 Y 轴的坐标位置变化. 我们以文献 [26] 提出的无人机动力学模型为基础, 通过引入 $F = mv$ 动量定理公式, 分别对航向角速度、X 轴和 Y 轴的速度和坐标位置变化公式进行了转化, 本文的每架无人机的动力学模型如下式所示:

$$\begin{cases} \phi = \phi + r_\phi \mathrm{d}t, \quad -30^\circ < \phi < 30^\circ, \\ r_\varphi = 9.81 \cdot m/F \cdot \mathrm{d}t \cdot \tan\phi, \\ \varphi = \varphi + r_\varphi \mathrm{d}t, \quad -180^\circ < \varphi < 180^\circ, \\ v_x = \sin\varphi \cdot F/m \cdot \mathrm{d}t, \\ v_y = \cos\varphi \cdot F/m \cdot \mathrm{d}t, \\ x = x + v_x \mathrm{d}t, \\ y = y + v_y \mathrm{d}t, \end{cases} \tag{1}$$

其中, ϕ 表示滚转角, r_ϕ 表示滚转角速度, r_φ 表示航向角速度, m 表示无人机质量, F 表示驱动力, φ 表示航向角, v_x 和 v_y 分别表示无人机在 X 和 Y 轴的速度, (x, y) 分别表示无人机在 X 和 Y 轴的坐标位置, $\mathrm{d}t$ 表示时间 t 的微分变量. 为了有效地模拟无人机的运动状态, 无人机演员网络 θ^π 输出一个由驱动力和滚转角速度组成的向量 $[F, r_\phi]$, 无人机的滚转角 ϕ、航向角 φ、X 轴速度 v_x 和 Y 轴速度 v_y 随时间的推移而变化.

建立笛卡尔 (Descartes) 坐标系, 构建对抗的二维态势平面图. 如图 1 所示, 显示了两个对抗中无人机的场景. 在任何时刻 t, 一对无人机对抗的情况可以用二元组 $\langle \omega_{\mathrm{ac}}^{ij}, \mu_{\mathrm{de}}^{ij} \rangle$ 表征. $\omega_{\mathrm{ac}}^{ij}$ 是攻击者 i 的攻击角, μ_{de}^{ij} 是目标 j 的防卫角.

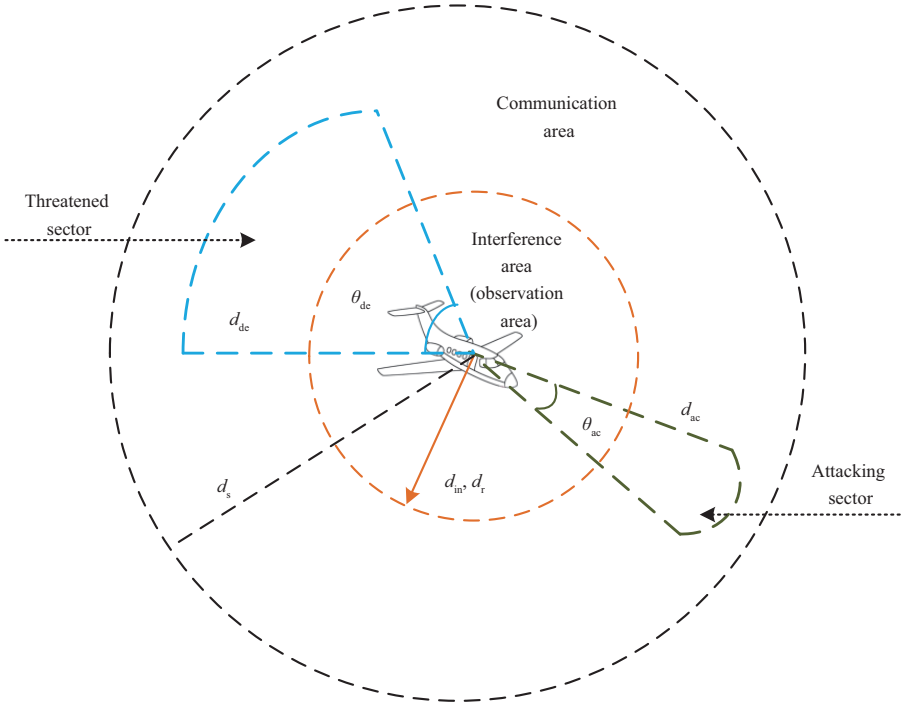

图 2 (网络版彩图) 无人机对抗参数

Figure 2 (Color online) UAV confrontation parameters

二元组的每个元素由下式计算:

$$\omega_{\mathrm{ac}}^{ij} = \arccos \frac{v_{x,i}(x_i - x_j) + v_{y,i}(y_i - y_j)}{d_{ij} \cdot ||v_i||_2}, \tag{2}$$

$$\mu_{\mathrm{de}}^{ij} = \arccos \frac{v_{x,j}(x_j - x_i) + v_{y,j}(y_j - y_i)}{d_{ji} \cdot ||v_j||_2}, \tag{3}$$

其中, $||v_{(\cdot)}||_2$ 是速度矢量的 2-范数, $d_{(\cdot)(\cdot)}$ 是攻击者 i 与目标 j 的欧氏距离, ac 和 de 分别代表攻击和防卫, $v_{x,i}$ 和 $v_{y,i}$ 是攻击者 i 分别在 X 轴和 Y 轴的速度分量, $v_{x,j}$ 和 $v_{y,j}$ 是目标 j 分别在 X 轴和 Y 轴的速度分量, x_i 和 y_i 是攻击者 i 分别在 X 轴和 Y 轴的坐标位置, x_j 和 y_j 是目标 j 分别在 X 轴和 Y 轴的坐标位置. 攻击角 $\omega_{\mathrm{ac}}^{ij}$ 和防卫角 μ_{de}^{ij} 根据攻击者 i 和目标 j 的速度、位置和距离动态变化, 攻击角 $\omega_{\mathrm{ac}}^{ij}$ 和防卫角 μ_{de}^{ij} 分别由式 (2) 和 (3) 计算得出.

全部无人机具有相同的对抗参数, 如图 2 所示. d_{ac} 和 θ_{ac} 表示攻击区的长度和角度, d_{de} 和 θ_{de} 表示受威胁区的长度和角度, d_{in} 表示干扰区域, d_s 表示通信区域, d_r 表示可观测区域.

在所考虑的简单场景中, 当以下 3 个条件成立时, 攻击者 i 可以杀死目标 j: (1) 攻击者与目标之间的距离 d_{ij} 小于攻击者的 d_{ac} 和目标的 d_{de}; (2) 目标在攻击者的攻击区域内; (3) 攻击者处于目标的受威胁区. 这些条件可以表述为下式:

$$d_{ij} \leqslant d_{\mathrm{ac}}^{i} \quad \text{and} \quad d_{ij} \leqslant d_{\mathrm{de}}^{j}, \tag{4}$$

$$\omega_{\mathrm{ac}}^{ij} \cdot 180/\pi \leqslant \theta_{\mathrm{ac}}^{i}/2, \tag{5}$$

$$\mu_{\mathrm{de}}^{ij} \cdot 180/\pi \leqslant 180 - \theta_{\mathrm{ac}}^{j}/2. \tag{6}$$

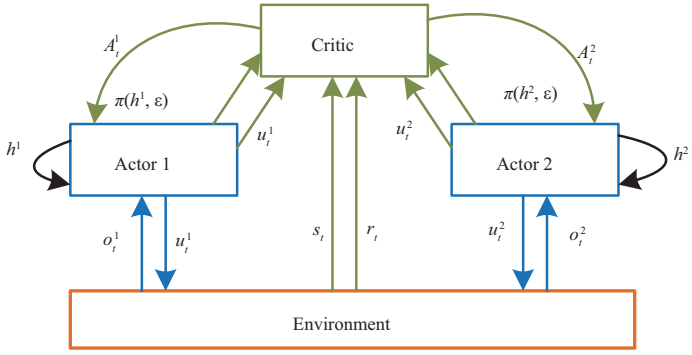

图 3　(网络版彩图) 训练框架

Figure 3　(Color online) Training framework

2.3　集中式训练分布式执行 (CTDE)

在真实世界的多智能体环境中, 智能体获得额外信息变得相当困难, 因此智能体必须采用分布式执行策略. 另一方面, 使用多智能体强化学习方法时, 需要在模拟器中进行训练, 然后再应用到实际设备上. 在训练过程中, 智能体可以在模拟器中获得额外的训练信息或自由通信, 以评估价值. 然而, 在实际执行策略时, 智能体不依赖这些额外信息, 而是仅根据本地观测生成策略. 这导致了第 3 种学习方法的出现, 称为集中式训练分布式执行 (centralized training with decentralized execution, CTDE) 方法.

CTDE 是多智能体强化学习最受欢迎的学习方式. 如图 3 所示, COMA 方法利用集中式学习 – 分布式执行模式, 在训练阶段将全局信息输入给集中式评论家网络, 集中式评论家网络指导演员网络进行动作决策. 演员网络 θ^{π} 获取可观测区域的信息 o_t, 演员网络 θ^{π} 根据策略 π 做出相应动作 u_t, h 表示隐藏状态, ε 是策略 π 的参数化, 智能体获得环境奖励 r_t, 评论家网络 θ^{c} 通过智能体的动作 u_t、环境状态信息 s_t 和策略 π 计算出优势函数 A_t, 然后利用策略梯度定理使奖励期望达到最大来更新演员网络 θ^{π}, 从而指导智能体的动作.

2.4　奖励函数

结合实际环境和奖励稀疏问题, 为无人机集群空中对抗博弈设计奖励函数, 该奖励函数适用于红方和蓝方无人机. 奖励函数有 5 部分组成, 分别为: 攻击奖励函数、受威胁奖励函数、干扰奖励函数、被干扰奖励函数、距离奖励函数. 无人机 i 在时刻 t 的攻击奖励函数设置为

$$r_{1,i,t}^{\mathrm{X}} = \lambda_{1,i}^{\mathrm{X}} \cdot \mathrm{sum}\left(\boldsymbol{w}_{i,t}^{\mathrm{X}}\right), \tag{7}$$

$$w_{i,t}^{\mathrm{X}} = \left[w_1^{\mathrm{X}^-},\ w_2^{\mathrm{X}^-},\ \ldots,\ w_{N^{\mathrm{X}-}}^{\mathrm{X}^-}\right]. \tag{8}$$

$\lambda_{1,i}^{\mathrm{X}}$ ($\lambda_{1,i}^{\mathrm{X}} \geqslant 0$) 表示攻击奖励函数的权重, 攻击奖励函数的权重应设置为正数, 鼓励红方和蓝方无人机的对抗博弈, X^- 表示敌机的总数量, $\boldsymbol{w}_{i,t}^{\mathrm{X}}$ 表示敌机身份的独热编码向量. 利用独热编码来确认被攻击的敌机信息. 同时无人机在对抗博弈过程中应学会避免被敌机威胁到, 即受威胁奖励函数应设置为负数, 无人机 i 在时刻 t 的受威胁奖励函数设置为

$$r_{2,i,t}^{\mathrm{X}} = -\lambda_{2,i}^{\mathrm{X}} \cdot \mathrm{sum}\left(\boldsymbol{u}_{i,t}^{\mathrm{X}}\right), \tag{9}$$

$$u_{i,t}^{\mathrm{X}} = \left[u_1^{\mathrm{X}^-},\ u_2^{\mathrm{X}^-},\ \ldots,\ u_{N^{\mathrm{X}-}}^{\mathrm{X}^-}\right]. \tag{10}$$

$\lambda^{\mathrm{X}}_{2,i}$ ($\lambda^{\mathrm{X}}_{2,i} \geqslant 0$) 表示受威胁奖励函数的权重, $\boldsymbol{u}^{\mathrm{X}}_{i,t}$ 表示无人机 i 受到敌机威胁时, 敌机身份的独热编码向量. 在无人机对抗博弈过程中, 向敌机施加干扰行为会干扰其正确决策, 从而有效提升己方的胜率, 无人机 i 在时刻 t 的干扰奖励函数设置为

$$r^{\mathrm{X}}_{3,i,t} = \lambda^{\mathrm{X}}_{3,i} \cdot \mathrm{sum}\left(\boldsymbol{b}^{\mathrm{X}}_{i,t}\right), \tag{11}$$

$$b^{X}_{i,t} = \left[b^{\mathrm{X}^-}_1,\ b^{\mathrm{X}^-}_2,\ \ldots,\ b^{\mathrm{X}^-}_{N^{\mathrm{X}^-}}\right]. \tag{12}$$

$\lambda^{\mathrm{X}}_{3,i}$ ($\lambda^{\mathrm{X}}_{3,i} \geqslant 0$) 表示干扰奖励函数的权重. $\boldsymbol{b}^{\mathrm{X}}_{i,t}$ 表示无人机 i 干扰到敌机时, 敌机身份的独热编码向量. 在无人机对抗博弈过程中, 不仅要干扰敌机, 同时也需要避免被敌机干扰, 受到敌机干扰视为惩罚, 被敌方干扰奖励值应设置为负数, 无人机 i 在时刻 t 的被干扰奖励函数设置为

$$r^{\mathrm{X}}_{4,i,t} = -\lambda^{\mathrm{X}}_{4,i} \cdot \mathrm{sum}\left(\boldsymbol{c}^{\mathrm{X}}_{i,t}\right), \tag{13}$$

$$c^{X}_{i,t} = \left[c^{\mathrm{X}^-}_1,\ c^{\mathrm{X}^-}_2,\ \ldots,\ c^{\mathrm{X}^-}_{N^{\mathrm{X}^-}}\right]. \tag{14}$$

$\lambda^{\mathrm{X}}_{4,i}$ ($\lambda^{\mathrm{X}}_{4,i} \geqslant 0$) 表示被干扰奖励函数的权重. $\boldsymbol{c}^{\mathrm{X}}_{i,t}$ 表示无人机 i 在遭受敌机干扰时, 敌机身份的独热编码向量. 为了解决奖励稀疏问题, 引入距离奖励并动态调整整体奖励函数, 无人机 i 在时刻 t 的距离奖励函数设置为

$$r^{\mathrm{X}}_{5,i,t} = -\lambda^{\mathrm{X}}_{5,i} \cdot \min\left(\sqrt{\left(x^{\mathrm{X}}_{i,t} - x^{\mathrm{X}^-}_{j,t}\right)^2 + \left(y^{\mathrm{X}}_{i,t} - y^{\mathrm{X}^-}_{j,t}\right)^2}, j \in N^{\mathrm{X}^-}\right). \tag{15}$$

$\lambda^{\mathrm{X}}_{5,i}$ ($\lambda^{\mathrm{X}}_{5,i} \geqslant 0$) 表示距离奖励函数的权重. 在上式中, 考虑了无人机 i 和全部敌方无人机之间的距离. 当无人机 i 与对手的距离越近, 无人机 i 得到惩罚越少. 距离奖励函数会使无人机 i 优先考虑距离最近的敌方无人机, 避免无人机 i 关注多个敌方无人机, 避免无人机 i 在不同敌方无人机之间重复寻找目标.

最后, 计算奖励函数 $r^{\mathrm{X}}_{i,t}$, 得出无人机 i 获得的即时奖励为

$$r^{\mathrm{X}}_{i,t} = r^{\mathrm{X}}_{1,i,t} + r^{\mathrm{X}}_{2,i,t} + r^{\mathrm{X}}_{3,i,t} + r^{\mathrm{X}}_{4,i,t} + r^{\mathrm{X}}_{5,i,t}. \tag{16}$$

3 深度强化学习方法

3.1 深度 Q 网络 (DQN)

深度 Q 网络 (DQN) [16] 是强化学习的一种方法, 它融合了神经网络和 Q-Learning 的理念, Q-Learning 作为一种经典的强化学习方法, 其核心是学习一个动作价值函数, 该函数为每个状态 – 动作对分配一个值, 以反映在某一状态下执行特定动作的价值. 在传统的 Q-Learning 中, 由于状态和动作的数量庞大, 为每个状态 – 动作对分配数值, 在实际中变得不切实际. 为了应对这一挑战, Mnih 等 [17] 利用神经网络强大的表征能力, 将高维输入数据作为强化学习中的状态, 并将其作为神经网络模型的输入. 随后, 神经网络输出每个动作对应的价值, 从而确定要执行的动作. 我们将在实验中探究 DQN 方法在无人机集群对抗博弈环境中的表现.

3.2 多智能体深度确定性策略梯度 (MADDPG)

多智能体深度确定性策略梯度 (multi-agent deep deterministic policygradient, MADDPG) [27] 方法是对深度确定性策略梯度 (DDPG) 方法的改进, 其中最主要的变化在于每个智能体的评论家部分能

够获取其他智能体的动作信息; 这一改进克服了智能体只能通过局部观测来评估和做出决策的限制, 同时解决了评论家网络评估的不准确性和演员网络决策能力的不足. 采用 MADDPG 方法的智能体在集中训练阶段, 评论家网络从智能体收集状态 – 动作对, 从而获得全局信息用于评估. 在分布执行阶段, 演员网络仅利用自身的局部观测来做出决策, 无需获取其他智能体的额外信息, 从而降低了复杂性. MADDPG 方法继承了 CTDE 和 DDPG 的特点, 因此成为了广受欢迎的多智能体强化学习方法之一. 我们将在实验中探究 MADDPG 方法在无人机集群对抗博弈环境中的表现.

3.3 引入反事实基线的无人机集群对抗博弈方法

DQN 方法在解决高维状态空间和动作空间时有较好的性能, 能较快地学到最优策略. MADDPG 方法结合了策略和价值函数的优点, 通过同时学习策略和价值函数来获得更好的学习性能, 并通过 CTDE 机制实现智能体间的协作. 但这两种方法存在多智能体间的信用分配问题, 导致智能体协同能力不高. 实际上, 在无人机集群对抗博弈环境中, 无人机在对抗过程中也需要合理的信用分配机制来进一步提升己方的合作效能, 而一种称为反事实多智能体策略梯度 (COMA) 的多智能体强化学习方法能有效应对智能体间的信用分配问题.

COMA 方法通过引入反事实基线, 来确定在这一次全局行为决策中单架无人机的动作贡献, 以解决信用分配问题 [25]. COMA 方法采用演员 – 评论家 [28] 的框架, 在这种方法中, 学习是集中式的, 因此使用集中的评论家网络, 评论家网络的条件是联合行动和联合状态信息; 每个无人机的策略条件是自身的观测信息, 演员网络是按照评论家网络估计的梯度来训练. 同时使用反事实基线, 将即时奖励与该无人机当前时刻行为用 “默认行为” 替代所获奖励进行比较, 通过比较这两个奖励值来判断无人机当前动作对任务的贡献.

COMA 的评论家网络如图 4 所示, 在时刻 t, 评论家网络的输入包括: 全局状态 s_t, 无人机 a 的观测 o_t^a, 无人机 a 的独热编码, 所有无人机上一时刻的联合动作 $\boldsymbol{u}_{t-1}$, 除无人机 a 之外的无人机的联合动作 $\boldsymbol{u}_t^{-a}$, 上述数据通过评论家网络的线性变换 (linear) 和激活函数 (ReLU) 后得到每架无人机所有动作的价值函数, 价值函数通过反事实基线得出每架无人机的优势函数 A_t^a, 然后利用策略梯度定理来更新演员网络. 计算评论家网络的梯度, 使用时间差分误差 (TD-error) 的方式更新评论家网络的权重, 并把损失函数的值降到最低.

TD-error 包括 TD(0), TD(λ) 两种更新方式, 本文采用 TD(λ) 方式进行更新, 其损失函数为

$$\text{Loss} = (y^{(\lambda)} - f(\cdot))^2, \tag{17}$$

$$y^{(\lambda)} = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{(n)}, \tag{18}$$

$$G_t^{(n)} = \sum_{l=1}^{n}\lambda^{l-1}r^{t+1} + \lambda^n f^{\theta^c}(\cdot), \tag{19}$$

其中, $y^{(\lambda)}$ 表示步数的加权和, λ 为折扣因子, r^{t+1} 为下一时刻的即时奖励; $G_t^{(n)}$ 为状态价值函数, 用来衡量无人机到达状态 s_t 的价值. 损失函数可表示为

$$\text{Loss} = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\left(\sum_{l=1}^{n}\lambda^{l-1}r^{t+1} + \lambda^n f^{\theta^c}(\cdot) - f(\cdot)\right)^2, \tag{20}$$

其中, $f(\cdot)$ 为评论家网络的函数值, $f^{\theta^c}(\cdot)$ 为评论家目标网络输出的预测函数值; 通过当前动作策略计

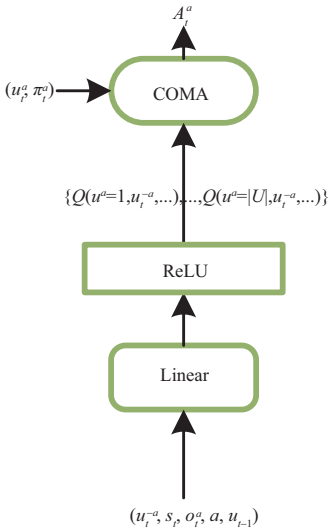

图 4 (网络版彩图) 评论家网络示意图

Figure 4 (Color online) Diagram of critic network

算每架无人机每个时间步的反事实基线, 其计算公式如下:

$$b(s, \boldsymbol{u}^{-a}) = \sum_{u'^a} \pi^a(u'^a|\tau^a) Q(s, (\boldsymbol{u}^{-a}, u'^a)), \tag{21}$$

其中, τ^a 表示无人机 a 的动作 – 观测历史, u'^a 是无人机 a 的遍历动作, $\boldsymbol{u}^{-a}$ 为除去无人机 a 的联合动作, $\pi^a(u'^a|\tau^a)$ 为无人机 a 的策略, $Q(s, (\boldsymbol{u}^{-a}, u'^a))$ 是无人机 a 每个动作的价值函数. 评论家目标网络输出该动作的反事实基线动作价值估计, 将无人机的联合动作价值与反事实基线进行比较, 无人机对应的评论家网络计算当前无人机在当前环境下的优势函数, 更新无人机的演员网络. 优势函数计算公式如下:

$$A^a(s, \boldsymbol{u}) = Q(s, \boldsymbol{u}) - \sum_{u'^a} \pi^a(u'^a|\tau^a) Q(s, (\boldsymbol{u}^{-a}, u'^a)), \tag{22}$$

其中, $A^a(s, \boldsymbol{u})$ 为优势函数, $\boldsymbol{u}$ 为所有无人机的联合动作, $Q(\cdot)$ 为动作价值函数. 采用反事实的多智能体策略梯度方法, 得到如下的演员网络的梯度:

$$g = \nabla_{\theta_\pi} \log \pi(\boldsymbol{u}|\tau_t^a) \left(Q(s, \boldsymbol{u}) - \sum_{u'^a} \pi^a(u'^a|\tau^a) Q(s, (\boldsymbol{u}^{-a}, u'^a)) \right). \tag{23}$$

利用策略梯度定理使奖励期望达到最大来更新演员网络, 将更新后的演员网络参数保存, 迭代更新, 直至满足对抗博弈的终止条件为止, COMA 算法伪代码如算法 1 所示. 反事实基线思想是 COMA 方法的核心部分, 本文引入反事实基线解决己方无人机间的信用分配问题. 在无人机集群对抗博弈环境中, 无人机的评论家网络在保持团队中其他无人机动作不变的情况下, 对无人机在每一时刻的所有动作进行评估来判定该时刻该无人机的动作价值. 评论家网络指导演员网络的更新, 使得演员网络输出的动作能够获得更多奖励. 本文在具有连续状态、连续动作的无人机对抗环境中, 基于无人机动力学模型和对抗态势, 设置符合实际环境的击敌条件, 综合考虑实际对抗环境和稀疏奖励问题, 分别设计攻击奖励函数、受威胁奖励函数、干扰奖励函数、被干扰奖励函数和距离奖励函数, 以适应无人机真

Algorithm 1 COMA algorithm

Input: Initialize actor network θ^{π}, critic network θ^{c}, replay buffer $\mathcal{D}$, the number of UAVs N, the number of episodes M, and the maximum episode length T.

for epiosde $= 1 : M$ **do**
  Set up UAVs' adversarial game environment;
  Initialize state s_0, at the step $t = 0$; initialize local observations for each UAV s_0^a; initialize actions for each UAV u_0^a=[r_ϕ,r_φ,F];
  for step $t = 1 : T$ **do**
    for UAV $a = 1 : N$ **do**
      Obtain local observations state s_t^a; obtain current strategy based on actor network $u_t^a = \text{Actor}(u_{t-1}^a, s_t^a; \theta_a)$;
    end for
    Execute all actions $\boldsymbol{u}$, receive rewards r_t, the next state s_{t+1} and completion marks d_t;
    Store interactive experience [s_t^a, u_t^a, r_t^a, s_{t+1}^a] to replay buffer $\mathcal{D}$;
    if all d_t is true **then**
      Reset environment;
    end if
    if fixed step update **then**
      for UAV $a = 1 : N$ **do**
        Compute loss function and update critic network θ^c;
        According to $A^a(s_t^a, u) = Q(s_t^a, u) - \sum_{u'^a} \pi^a(u'^a|\tau^a) Q(s_t^a, (u^{-a}, u'^a))$ compute the advantage function;
        According to $\delta\theta^{\pi} = \delta\theta^{\pi} + \nabla_{\theta^{\pi}} \log \pi(u'^a|\tau_t^a) A^a(s_t^a, u_t^a)$ compute the actor network gradient;
      end for
      According to $\theta_{i+1}^{\pi} = \theta_i^{\pi} + \alpha\delta\theta^{\pi}$ update the policy network;
    end if
  end for
end for

Output: The strategy distribution of all UAVs π_θ.

实对抗场景; 为了综合评估 COMA 方法在无人机对抗场景下的性能, 使用平均累积奖励、平均命中率和平均胜率作为评价指标.

4 仿真及实验

在多对多对抗场景中, 红蓝双方无人机相互交锋, 而无人机集群内部需要协同执行各自的任务, 使得这一博弈场景更加复杂. 图 5 为无人机集群对抗场景的一个示例, 对抗空域内共有 12 架无人机, 为 4 架红色无人机对抗 8 架蓝色无人机 (4 vs. 8). 红蓝双方在每个回合的初始位置相同, 它们在二维场景下搜索敌方, 学习攻击敌方和防止被敌方攻击到的策略. 图 6 为无人机对抗流程.

4.1 实验设置

在仿真对抗环境中, 包含了 4 架红色无人机和 8 架蓝色无人机 (4 vs. 8). 设定最大训练回合为 10000 次, 每回合最大步长为 200 步. 在仿真实验中, 每隔 10 步进行网络参数更新. 神经网络和强化学习环境的主要参数见表 1 和 2, 神经网络采用正交初始化, 优化器为同时考虑动量和自适应参数的 Adam. 图 7 为无人机对抗策略的设计流程.

为了验证 COMA 方法的训练效果, 将 COMA 方法与 DQN, MADDPG 方法对比, 这 3 种方法均在上述所述示例的无人机集群的对抗环境中应用, 以评估它们的训练效果. 为了清楚地展示模型优化

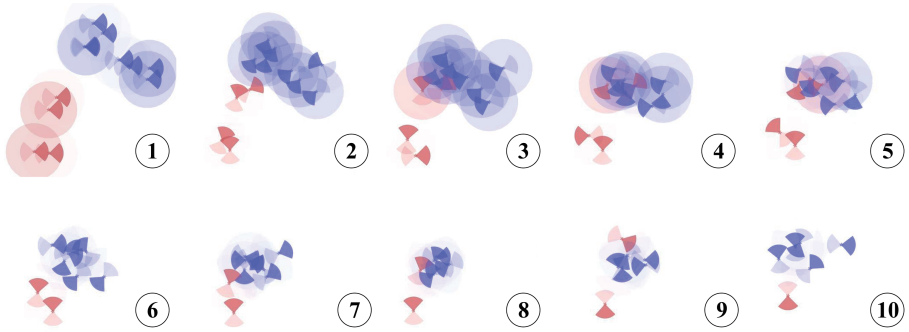

图 5 (网络版彩图) 无人机集群对抗场景示例

Figure 5 (Color online) Example of UAV swarm confrontation scene

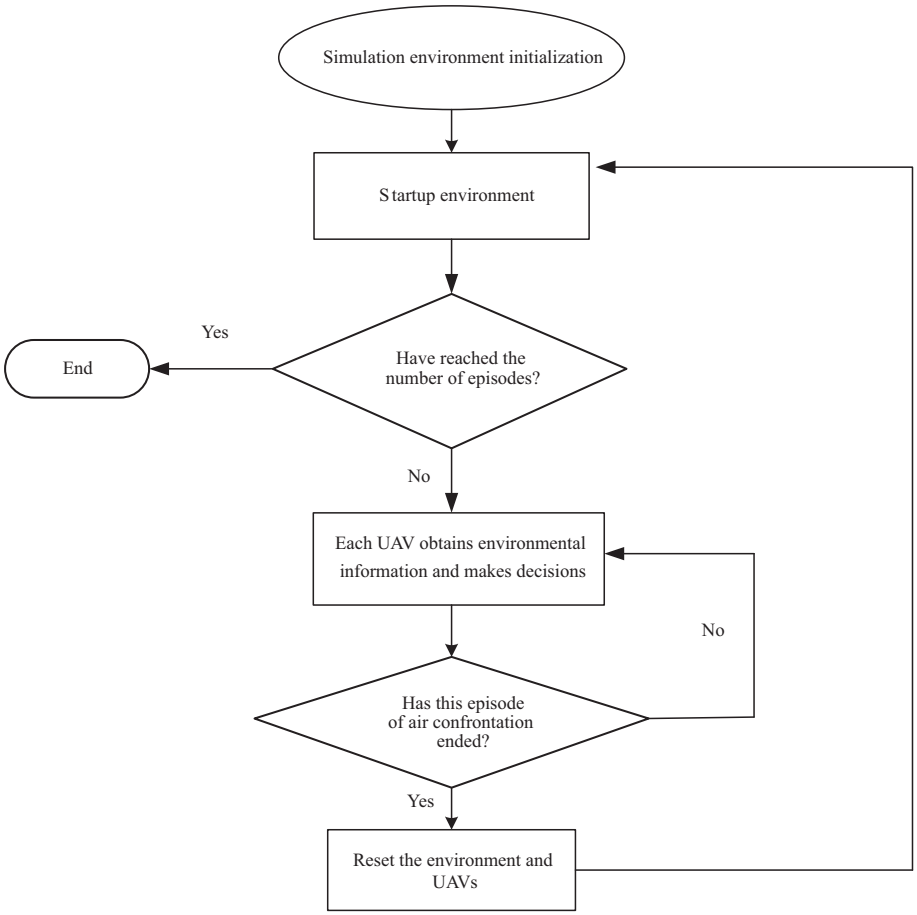

图 6 无人机集群空中对抗流程

Figure 6 Flow chart of UAV swarm air confrontation

的过程, 利用双方平均累积奖励 (η_{MAR}) 的总和作为评价指标, 具体描述如下:

$$\eta_{\mathrm{MAR}} = \frac{1}{|B|}\sum_{b=1}^{|B|}\sum_{i=1}^{N} R_i, \tag{24}$$

表 1　超参数设置

Table 1　Hyperparameter settings

Parameter	Value
Number of UAVs	12
Number of episodes	10000
Maximum length of episodes	200
Learning rate of the actor	1.5e−4
Learning rate of the critic	3e−4
Discount factor	0.99
Batch size	256
Network update frequency	10
Optimizer	Adam
Activation function	ReLU

表 2　无人机物理参数

Table 2　UAV physical parameters

Parameter	Value
Maximum attacking angle (°)	90
Maximum threatening angle (°)	90
Maximum attacking radius (km)	0.3
Maximum threatening radius (km)	0.3
Maximum interference radius (km)	0.7
Maximum detection radius (km)	0.6
Maximum acceleration (m/s^2)	2.0
Maximum speed (m/s)	2.0
Maximum roll angle (°)	30
Maximum course angle (°)	180

其中 B 代表存储的对抗回合 $|B|$ 的数量, N 代表无人机总数, 以及 R_i 代表每架无人机的累积奖励.

奖励的提升反映了无人机集群对抗训练的收敛情况, 然而, 无人机所学到的策略不一定是最有效的策略. 引入了平均命中率 η_{MHR} 作为回合内对抗效果的质量指标, 它用于衡量在对抗训练中击中对手无人机数量的归一化值. 在 “红 – 蓝” 对抗场景下, 双方都在努力提高平均命中率 η_{MHR}, 以最终实现完全消灭对手. η_{MHR} 值的提升意味着无人机学到的策略质量较高. 因此, 平均命中率描述如下:

$$\eta_{\mathrm{MHR}} = \frac{1}{|B|}\sum_{i=1}^{|B|}\frac{h_i}{N^{X^-}}, \tag{25}$$

其中 h_i 代表在存储的对抗回合 $|B|$ 的第 i 回合击中成功击中对手的次数, 而 N^{X^-} 代表对手的无人机数量. 实验采用平均胜率来描述 “红 – 蓝” 双方的对抗结果, 这一指标反映了完全摧毁对手无人机集群的能力, 而这正是红蓝两队追求的最直接目标. 同样地, 对于攻击者而言, 较高的平均胜率 η_{MWR} 代

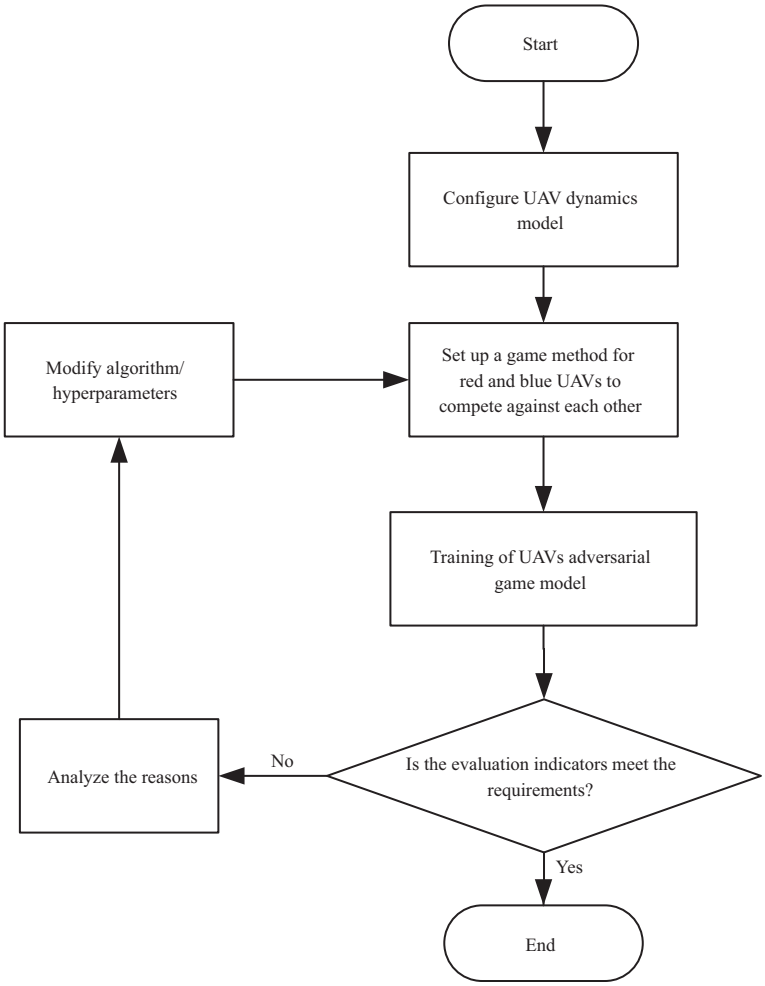

图 7 无人机对抗策略设计流程图

Figure 7 Design flow chart of UAV adversarial strategy

表无人机所学的策略有效性较高. 平均胜率定义如下:

$$\eta_{\mathrm{MWR}}=\frac{1}{|B|}\sum_{k=1}^{|B|}I(k),\quad I(k)=\begin{cases}1, & \text{win in the } k\text{-th episode},\\ 0, & \text{otherwise.}\end{cases}\tag{26}$$

4.2 实验结果

利用上述 3 个评估指标来分析无人机对抗过程, 其中红方无人机集群采取 COMA 方法, 蓝方无人机集群分别采用 MADDPG, COMA 和 DQN 方法. 首先, 对比不同对抗方法下无人机集群平均累积奖励曲线. 如图 8 所示, 平均累积奖励 η_{MAR} 的增加反映了无人机集群的训练过程逐渐收敛, 并且学习到了有效的策略, 基于 CTDE 框架的 COMA 方法包含了来自其他无人机的额外信息, 每架无人机可以在训练中根据评论家网络的全局估计值调整自己的策略. 因此所有的训练都能收敛. 特别地, 当红方采用 COMA 方法而蓝方采用 DQN 方法 (R:COMA vs. B:DQN) 时, 平均累积奖励能最快收敛. 由于 COMA 方法采用了 CTDE 框架, 在训练过程中无人机的评论家网络能共享其他无人机的信息, 而 DQN 方法不能共享其他无人机的信息; 同时, COMA 方法采用了反事实基线, 能更有效地进行信用分

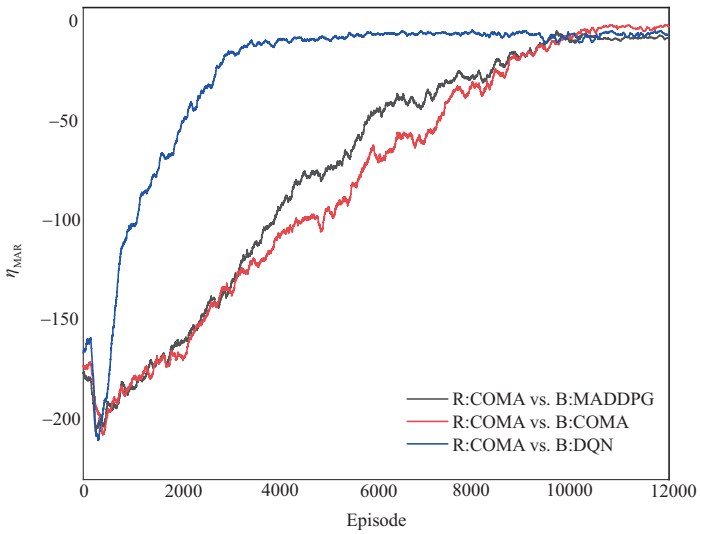

图 8　(网络版彩图) 无人机的平均累积奖励 η_{MAR}

Figure 8　(Color online) Mean accumulated rewards η_{MAR} of all UAVs

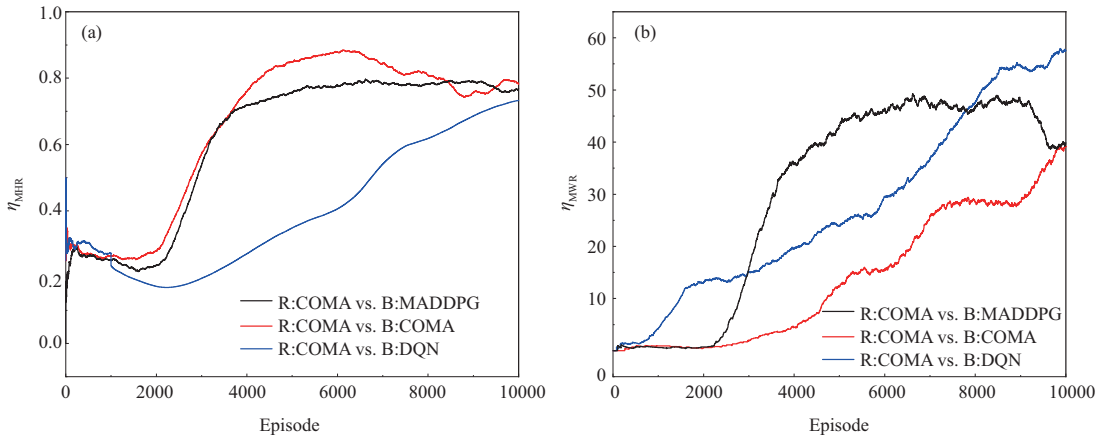

图 9　(网络版彩图) (a) 红方无人机的平均命中率 η_{MHR}; (b) 红方无人机的平均胜率 η_{MWR}

Figure 9　(Color online) The average hit rate η_{MHR} (a) and the average win rate η_{MWR} (b) of red UAVs

配, 使得红方无人机能更快地学到有效策略.

平均命中率 η_{MHR} 是表征对抗效果的详细指标. 图 9(a) 显示红方无人机集群命中率的情况, 可以看出随着回合数的增加, 所有方法下的红方命中率均能收敛, 这表明红方无人机集群学习到了有效策略. 特别地, R:COMA vs. B:COMA 时获得了最高的命中率. 这说明 COMA 的信用分配机制能提高双方集群内部的合作效率, 使得对抗双方收敛到了更高的纳什均衡, 从而使得红方无人机能够获得较高的命中率.

平均胜率 η_{MWR} 是反映对抗效果的终极指标. 图 9(b) 展示了红方的胜率曲线, 可以看出随着训练回合数的增加, 红方胜率具有增加的趋势, 表明红方无人机集群可以学习到有效的策略. 尤其是 R:COMA vs. B:DQN 获得了最高的 η_{MWR}. 对比图 8 和 9, 可以看出高的平均累积奖励 η_{MAR} 能获得高的平均胜率 η_{MWR}, 但是高的平均命中率 η_{MHR} 不一定能获得高的平均胜率 η_{MWR}.

表 3 不同对抗场景下的评价指标最大值 [a)]

Table 3 Maximum values of indicators under different scenes[a)]

	R:COMA vs. B:COMA	R:COMA vs. B:MADDPG	R:COMA vs. B:DQN
η_{MAR}	**−2.31**	−8.57	−12.47
η_{MHR}	**0.89**	0.76	0.64
η_{MWR}	0.41	0.48	**0.55**

a) Bold values represent the maximum values for different metrics.

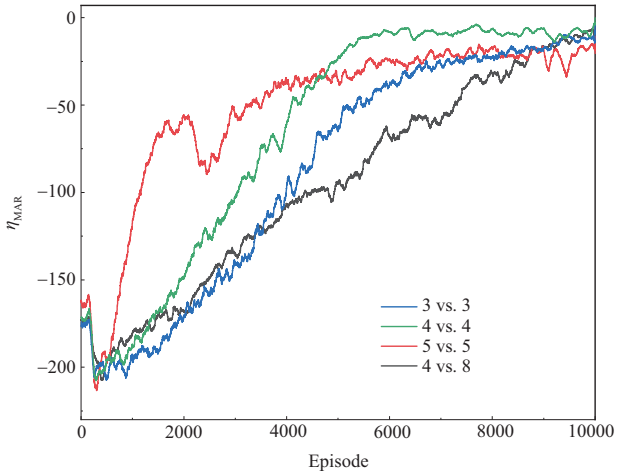

图 10 (网络版彩图) 不同集群规模下 R:COMA vs. B:COMA 时的平均累积奖励

Figure 10 (Color online) The average cumulative rewards of R:COMA vs. B:COMA under different swarm sizes

为了更进一步展示 COMA 方法相对于其他方法的优势, 将评价指标在不同对抗场景下的最大值绘制成表格, 如表 3 所示.

从表 3 中可以看出, 命中率的最大值 (0.89) 与图 9(a) 中的平均命中率 η_{MHR} 结果一致, 平均胜率的最大值 (0.55) 也与图 9(b) 中的平均命中率 η_{MWR} 结果一致. 但是, 累积奖励的最大值 (−2.31) 与图 5 中的平均累积奖励 η_{MAR} 结果并不一致, 这表明累积奖励的方差比较大.

为了进一步验证 COMA 方法的收敛性和稳定性, 进行了不同集群规模 (3 vs. 3, 4 vs. 4, 5 vs. 5 和 4 vs. 8) 下的仿真实验, 实验结果如图 10 所示, 当 R:COMA vs. B:COMA 时, 在不同的集群规模下, 随着训练回合数的增加, 所有对抗场景下的平均累积奖励值均能有效地收敛到纳什均衡, 表明针对不同的集群规模时 COMA 方法具有良好的泛化性, 从而保证了 COMA 方法的收敛性和稳定性.

最后, 为了深入验证 COMA 方法的收敛性, 绘制了 4 vs. 8 时每架无人机的平均累积奖励, 如图 11 所示, 可以看出随着回合数的增加, 红蓝双方的每架无人机的平均累积奖励均能收敛, 表明 COMA 方法也能让每架无人机学习到有效的策略, 从而保证了 COMA 方法在无人机集群对抗博弈任务上的实用性和鲁棒性.

5 结论

本文将反事实基线思想引入到无人机集群对抗博弈环境, 提出一种基于 COMA 的无人机集群对抗博弈方法, 建立基于多智能体深度强化学习的无人机集群对抗博弈模型, 利用多智能体粒子群环境

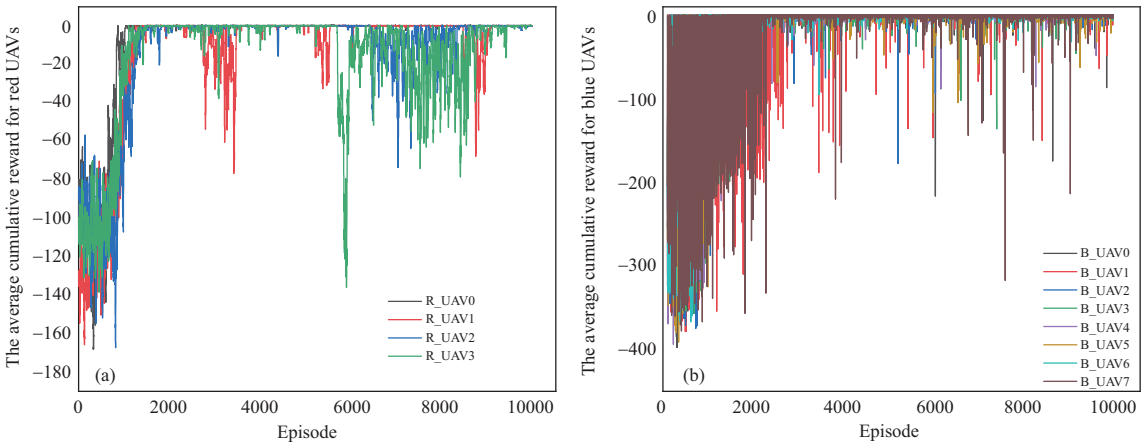

图 11 (网络版彩图) 4 vs. 8 时单架无人机的平均累积奖励. (a) 4 架红方无人机; (b) 8 架蓝方无人机

Figure 11 (Color online) The average cumulative reward of single UAV in the case of 4 vs. 8. (a) 4 red UAVs; (b) 8 blue UAVs

对红蓝双方无人机集群对抗进行仿真实验, 实验结果显示 COMA 方法优于流行的深度强化学习方法, 且平均累积奖励、平均命中率和平均胜率均能随着训练回合数的增加而收敛, 表明无人机集群对抗博弈能够收敛到纳什均衡. 另外, 分析了 COMA 方法对于不同无人机集群规模的泛化性, 验证了每架无人机的平均累积奖励的收敛性, 从而保证了 COMA 方法在无人机集群对抗博弈任务上的实用性和鲁棒性.

无人机集群中往往存在类型多样、功能各异的异质无人机, 如何高效解决大规模、异质无人机间的信用分配问题将是下一步的研究方向.

参考文献

1 Sun Z X, Yang S Q, Piao H Y, et al. A survey of air combat artificial intelligence. Acta Aeronaut Astronaut Sin, 2021, 42: 525799 [孙智孝, 杨晟琦, 朴海音, 等. 未来智能空战发展综述. 航空学报, 2021, 42: 525799]

2 Li S Y, Chen M, Wang Y H, et al. Human-computer gaming decision-making method in air combat under an incomplete strategy set. Sci Sin Inform, 2022, 52: 2239–2253 [李守义, 陈谋, 王玉惠, 等. 非完备策略集下人机对抗空战决策方法. 中国科学: 信息科学, 2022, 52: 2239–2253]

3 Yan F, Zhu X P, Zhou Z, et al. Real-time task allocation for a heterogeneous multi-UAV simultaneous attack. Sci Sin Inform, 2019, 49: 555–569 [严飞, 祝小平, 周洲, 等. 考虑同时攻击约束的多异构无人机实时任务分配. 中国科学: 信息科学, 2019, 49: 555–569]

4 Wang E S, Guo J, Hong C, et al. Cooperative confrontation model of UAV swarm with random spatial networks. J Beijing Univ Aeronaut Astronaut, 2023, 49: 10–16 [王尔申, 郭靖, 宏晨, 等. 基于随机空间网络的无人机集群协同对抗模型. 北京航空航天大学学报, 2023, 49: 10–16]

5 Kaneshige J, Krishnakumar K. Artificial immune system approach for air combat maneuvering. In: Proceedings of SPIE – The International Society for Optical Engineering, 2007. 6560: 68–79

6 Duan H, Li P, Yu Y. A predator-prey particle swarm optimization approach to multiple UCAV air combat modeled by dynamic game theory. IEEE CAA J Autom Sin, 2015, 2: 11–18

7 Zhou W Q, Zhu J H, Kuang M C. An unmanned air combat system based on swarm intelligence. Sci Sin Inform, 2020, 50: 363–374 [周文卿, 朱纪洪, 匡敏驰. 一种基于群体智能的无人空战系统. 中国科学: 信息科学, 2020, 50: 363–374]

8 Isler V, Kannan S, Khanna S. Randomized pursuit-evasion in a polygonal environment. IEEE Trans Robot, 2005, 21: 875–884

9 Chen X, Wang Y F. Study on multi-UAV air combat game based on fuzzy strategy. Appl Mech Materials, 2014,

494-495: 1102–1105
10 Wang E S, Liu F, Hong C, et al. MASAC-based confrontation game method of UAV clusters. Sci Sin Inform, 2022, 52: 2254–2269 [王尔申, 刘帆, 宏晨, 等. 基于 MASAC 的无人机集群对抗博弈方法. 中国科学: 信息科学, 2022, 52: 2254–2269]
11 Lazaridou A, Peysakhovich A, Baroni M. Multi-agent cooperation and the emergence of (natural) language. 2016. ArXiv:1612.07182
12 Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529: 484–489
13 Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. Nature, 2017, 550: 354–359
14 Silver D, Hubert T, Schrittwieser J, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. 2017. ArXiv:1712.01815
15 Vinyals O, Babuschkin I, Czarnecki W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature, 2019, 575: 350–354
16 Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nat Rev Mol Cell Biol, 2019, 20: 681–697
17 Mnih V, Kavukcuoglu K, Silver D, et al. Playing Atari with deep reinforcement learning. 2013. ArXiv:1312.5602
18 Wang Z, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning. In: Proceedings of International Conference on Machine Learning, 2016. 1995–2003
19 Hasselt H V, Guez A, Silver D. Deep reinforcement learning with double Q-Learning. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'16), 2016. 2094–2100
20 Gong Z, Xu Y, Luo D. UAV cooperative air combat maneuvering confrontation based on multi-agent reinforcement learning. Un Sys, 2023, 11: 273–286
21 Chen C, Mo L, Zheng D, et al. Cooperative attack-defense game of multiple UAVs with asymmetricmaneuverability. Acta Aeronaut Astronaut Sin, 2020, 41: 324152 [陈灿, 莫雳, 郑多, 等. 非对称机动能力多无人机智能协同攻防对抗. 航空学报, 2020, 41: 324152]
22 Li S, Jia Y, Yang F, et al. Collaborative decision-making method for multi-UAV based on multiagent reinforcement learning. IEEE Access, 2022, 10: 91385–91396
23 Zhang T, Qiu T, Liu Z, et al. Multi-UAV cooperative short-range combat via attention-based reinforcement learning using individual reward shaping. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2022. 13737–13744
24 Chang Y H, Ho T, Kaelbling L. All learning is local: multi-agent learning in global reward games. In: Proceedings of Advances in Neural Information Processing Systems, 2003
25 Foerster J, Farquhar G, Afouras T, et al. Counterfactual multi-agent policy gradients. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2018
26 McGrew J S, How J P, Williams B, et al. Air-combat strategy using approximate dynamic programming. J Guid Control Dyn, 2010, 33: 1641–1654
27 Lowe R, Wu Y I, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments. In: Proceedings of Advances in Neural Information Processing Systems, 2017
28 Konda V, Tsitsiklis J. Actor-critic algorithms. In: Proceedings of Advances in Neural Information Processing Systems, 1999

UAV swarm adversarial game method with a counterfactual baseline

Ershen WANG[1,2], Jihao CHEN[1], Chen HONG[3,4*], Fan LIU[1], Aidong CHEN[3,4] & Hongyuan JING[3,4]

1. *School of Electronic and Information Engineering, Shenyang Aerospace University, Shenyang* 110136*, China*;
2. *School of Civil and Aviation, Shenyang Aerospace University, Shenyang* 110136*, China*;
3. *Multi-Agent Systems Research Centre, Beijing Union University, Beijing* 100101*, China*;
4. *College of Robotics, Beijing Union University, Beijing* 100101*, China*
* Corresponding author. E-mail: hchchina@sina.com, xxthongchen@buu.edu.cn

Abstract The collaborative adversarial game of unmanned aerial vehicles (UAVs) is becoming increasingly widespread and profound, especially in collaborative detection, global confrontation, strategic deception and other confrontation tasks. Reliable and efficient UAV swarm game methods are currently a hot research topic. This paper introduces the counterfactual baseline concept into the UAV swarm adversarial environment and proposes a UAV swarm adversarial game method based on counterfactual multi-agent policy gradients (COMA). In the UAV confrontation environment with infinite continuous states and actions, merging the UAV dynamics, we set up realistic attack conditions and reward functions, and construct a UAV swarm adversarial game model based on multi-agent deep reinforcement learning. The red and blue UAVs adopt different adversarial game methods, and asymmetric adversarial experiments are conducted in the multi-agent particle environment (MPE). The experimental results show that the average cumulative rewards can converge to Nash equilibrium. For 4 vs. 8 adversarial decision-making scene, the average hit rate of COMA is 39% and 17% higher than that of DQN and MADDPG, while the average win rate is 34% and 17% higher than that of DQN and MADDPG, respectively. Finally, the practicality and robustness for UAV swarm adversarial game tasks are ensured through an in-depth analysis of the convergence and stability of COMA.

Keywords UAV swarm, confrontation game, multi-agent, deep reinforcement learning, Nash equilibrium