doi:10.11887/j.cn.202301018 http://journal.nudt.edu.cn

针对无人机集群对抗的规则与智能耦合约束训练方法*

高显忠[1],项 磊[2],王宝来[2],贾高伟[1],侯中喜[1]
(1. 国防科技大学 空天科学学院,湖南 长沙 410073;2. 国防科技大学 计算机学院,湖南 长沙 410073)

摘 要:基于无人机集群智能攻防对抗构想,建立了无人机集群智能攻防对抗仿真环境。针对传统强化学习算法中难以通过奖励信号精准控制对抗过程中无人机的速度和攻击角度等问题,提出一种规则与智能耦合约束训练的多智能体深度确定性策略梯度(rule and intelligence coupling constrained multi-agent deep deterministic policy gradient, RIC-MADDPG)算法,该算法采用规则对强化学习中无人机的动作进行约束。实验结果显示,基于RIC-MADDPG方法训练的无人机集群对抗模型能使得红方无人机集群在对抗中的胜率从53%提高至79%,表明采用"智能体训练—发现问题—编写规则—再次智能体训练—再次发现问题—再次编写规则"的方式对优化智能体对抗策略是有效的。研究结果对建立无人机集群智能攻防策略训练体系、开展规则与智能相耦合的集群战法研究具有一定参考意义。

关键词:无人机集群;MADDPG算法;智能体决策;对抗模型;规则约束
中图分类号:V279 **文献标志码**:A **开放科学(资源服务)标识码(OSID)**:
文章编号:1001-2486(2023)01-157-10

Rule and intelligence coupling constraint training method for UAV swarm confrontation

GAO Xianzhong[1], *XIANG Lei*[2], *WANG Baolai*[2], *JIA Gaowei*[1], *HOU Zhongxi*[1]
(1. College of Aerospace Science and Engineering, National University of Defense Technology, Changsha 410073, China;
2. College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China)

Abstract: Based on the concept of the intelligent combat of UAV (unmanned aerial vehicle) swarms, the UAV swarms intelligent combat simulation environment was established. Aiming at the problem that it is difficult to accurately control the speed and attack angle of UAVs in the confrontation process through reward signals in traditional reinforcement learning algorithms, the RIC-MADDPG (rule and intelligence coupling constrained multi-agent deep deterministic policy gradient) algorithm was proposed. The algorithm uses rules to constrain the actions of UAVs in reinforcement learning. The simulation results show that the wining-rate of red UAV swarm, trained by the method based on the RIC-MADDPG, can be improved from 53% to 79%. This proves that the strategy of "agent training—problem finding—rule making—agent training again—problem finding again—rule making again" is effective for the optimization of agent combat strategy. The research results can be a reference for establishing the training system of the intelligent combat strategy of UAV swarms and conducting the research of swarm tactics coupling rule and intelligence.

Keywords: UAV swarms; MADDPG algorithm; agent decision making; countermeasure model; rule-constrained

近年来,随着无人机小型化、智能化、集群化技术快速发展,无人机智能集群作战已从理论走向战争实践,成为军事领域最活跃、最创新、最贴近实战的发展方向,已成为新型战斗力生成的重要创新发展途径。有矛就有盾,无人机智能集群技术的发展,持续推动着反无人机集群技术的发展。纵观人类武器发展史,当一种新质作战力量诞生后,应对该种作战武器最有效的方式往往是该武器本身。在无人机集群与反无人机集群武器的竞争式对抗发展过程中,也不可避免地将走向无人机集群与无人机集群对抗的作战样式[1]。为对抗无人机集群的攻击,最有效的方法就是利用无人机集群对入侵的无人机集群进行拦截,这将导致无人机集群之间的空中对抗,凸显出无人机集群对抗策略研究的重大意义[2]。

当前,在无人机集群对抗的方式、方法、策略

* **收稿日期**:2021-02-20
基金项目:国家自然科学基金资助项目(11602298)
作者简介:高显忠(1985—),男,重庆璧山人,副研究员,博士,E-mail:gaoxianzhong@nudt.edu.cn;
侯中喜(通信作者),男,陕西宝鸡人,教授,博士,博士生导师,E-mail:hzx@163.com

方面,还处在初步阶段,亟须开展深入研究。目前主流的无人机集群对抗算法主要包括三类:基于专家系统、基于博弈论和基于强化学习的算法。

在前期有人机对抗过程中,人类专家总结整理出了一些空战经验,通过这些经验可以建立专家知识库,可以应用在小规模的无人机集群对抗场景中。目前周欢等针对无人机集群控制系统方面存在的一些问题,提出了一种基于规则实现的无人机集群系统飞行与规避自主协同控制方法[3]。罗德林等在大规模无人机集群对抗决策系统中采用多 agent 理论方法,为每一个无人机单独设立行为规则集并给出决策方法,建立了无人机对抗模型[4],但是模型过度依赖专家指定的针对性规则,当环境发生变化时,规则必须重新制定。为了解决此问题,Xing 等研究了一种动态群与群无人机作战问题,提出了一种自组织攻防对抗决策(offense-defense confrontation decision-making, ODCDM)算法,该 ODCDM 算法采用分布式体系结构来考虑实时实现,其中每个无人机被视为智能体,并能够通过与邻居的信息交换来解决其局部决策问题[5],可以有效地解决大规模无人机集群对抗问题。基于专家系统的算法虽然可以有效地解决无人机集群对抗问题,但是当无人机集群规模较大时,集群系统过于复杂,导致专家知识库难以建立。

基于博弈论的方法可以在没有最优策略先验知识的情况下学习如何对抗。陈侠等利用传统有限策略静态博弈模型与纯策略纳什均衡的求解方法对多无人机协同打击任务开展研究,但是无法应用于集群规模较大的对抗中[6]。Duan 等基于捕食猎物粒子群优化(predator-prey particle swarm optimization, PP-PSO)的博弈论方法,将多个无人作战飞行器在军事行动中的动态任务分配问题分解为每个决策阶段的双人博弈问题,使得各阶段的最优分配方案均符合混合纳什均衡,之后,利用 PP-PSO 求解,对多无人机的空战模型问题进行了探索性研究[7]。Park 等基于博弈论方法设计了无人机的得分函数矩阵,建立无人机视距内对抗过程中的机动自动生成方法,在动态环境下寻找最优作战策略[8]。Alexopoulos 等采用多人动态博弈分解的方法来求解多无人机追讨问题[9]。基于博弈论的方法存在状态量过多、求解过于复杂等问题,同样无法应用于大规模无人机集群对抗环境中。

不需要环境模型信息的强化学习算法,通过与环境不断交互,最大化接收到的奖励来优化自身策略。何金等基于强化学习算法,通过对空战中优势区域和暴露区域的定义,采用双深度 Q 网络(double deep Q network, DDQN)对连续状态空间无人机隐蔽接敌问题进行了研究[10]。Li 等基于深度确定性策略梯度(deep deterministic policy gradient, DDPG)建立了一个智能决策框架,该策略可以使得有人/无人机参与近距离的一对一空中对抗,并通过自学提高空中对抗中的智能决策水平[11]。张耀中等更进一步,基于 DDPG 算法,开展无人机集群通过相互协作追击敌方来袭目标的研究,结果表明,通过训练,无人机集群在追击任务中的成功率可达 95%,表明该算法在无人机集群方面具有广阔应用前景[12],但是该方法在攻击对抗方面的效果仍有待深入研究。陈灿等针对不同机动能力无人机群体间的攻防对抗问题,建立了多无人机协同攻防演化模型,基于多智能体强化学习理论,研究了多无人机协同攻防的自主决策方法,实现了多无人机的稳定自主学习[13]。Xu 等针对无人机区域侦察和空对空对抗的典型任务场景,采用深度强化学习方法,开发了无人机自主决策方法,构建任务决策模型,并对基于遗传算法的决策模型进行优化,仿真结果验证了该方法的有效性[14]。

本文在总结现有研究成果基础上,基于无人机集群攻防对抗构想,考虑无人机动力学约束,建立了多无人机集群对抗仿真环境。以无人机实际攻防中的具体战术问题为对象,基于多智能体深度确定性策略梯度(multi-agent deep deterministic policy gradient, MADDPG)算法,建立无人机对抗模型,对无人机集群与集群的对抗形式进行深入研究。针对传统强化学习算法中难以通过奖励信号精准控制对抗过程中无人机的速度和攻击角度等问题,提出了一种智能与规则耦合约束训练策略,有效提高无人机集群的对抗能力。

1 多智能体深度确定性策略梯度算法介绍

MADDPG 算法是对 DDPG[15] 在多智能体领域的拓展,是一种基于演员 - 评论家框架的算法[16]。MADDPG 算法中有两个神经网络模块:演员模块和评论家模块。演员模块获取环境中的当前状态、选择相应的动作,评论家模块依据当前的状态和动作信息计算一个 Q 值,作为对演员模块输出动作的评估反馈,演员模块则通过评论家模块的反馈来更新策略,做出在当前状态下的最优动作[17]。MADDPG 算法中的演员模块输出的是一个具体的动作,可以在连续动作空间中进行

学习。MADDPG 算法最核心的部分就是在训练的时候引入可以观察全局信息的评论家模块来指导演员模块的训练,而执行的时候只使用有局部观测的演员模块采取行动,进行中心化训练和非中心化执行[18]。MADDPG 的算法架构如图 1 所示。

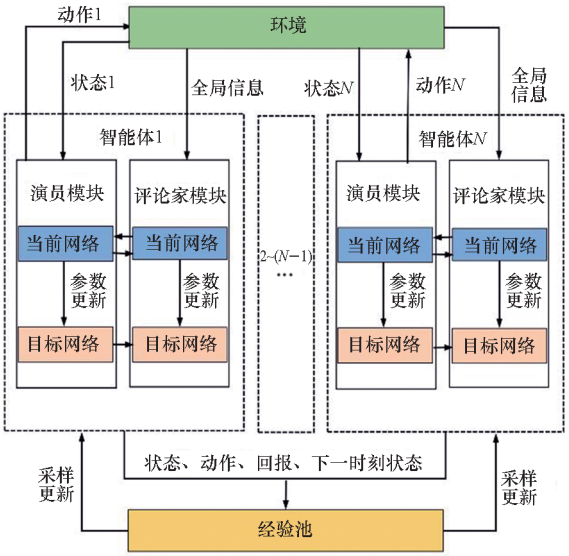


图 1　MADDPG 算法架构

Fig. 1　Algorithm architecture of MADDPG

在图 1 中 MADDPG 网络结构由环境、经验池、多个智能体组成,其中每个智能体均由演员 - 评论家网络模块构成。智能体获得环境输入的状态信息,通过演员模块输出动作与环境进行交互,并将交互过程中产生的样本存储在经验池中,评论家模块则获取全局信息来指导演员模块更新策略。智能体从经验池中抽取经验样本进行学习。演员模块和评论家模块中均存在两个结构相同但作用不同的网络,分别为估值网络和目标网络。在训练中只需要对估值网络的参数进行训练,而目标网络的参数则每隔一定的时间从估值网络中复制[19]。

设 n 个智能体的观测集合为 $x=\{o_1,\cdots,o_n\}$,智能体的随机策略集合为 $\pi=\{\pi_1,\cdots,\pi_n\}$,其参数分别表示为 $\theta_\pi=\{\theta_1,\cdots,\theta_n\}$,动作集合为 $a=\{a_1,\cdots,a_n\}$,则第 i 个智能体的累积期望奖励和策略梯度为:

$$J(\theta_i)=E_{s\sim p^\pi,a_i\sim\pi_i}\left[\sum_{t=0}^{\infty}\gamma^t r_{i,t}\right] \tag{1}$$

$$\nabla_{\theta_i}J(\theta_i)=E_{s\sim p^\pi,a_i\sim\pi_i}[\nabla\theta_i\log\pi_i(a_i\mid o_i)\cdot Q_i^\pi(x,a_1,\cdots,a_n)] \tag{2}$$

其中:γ^t 为折扣因子,表示当前动作对后续动作期望奖励的影响;$r_{i,t}$ 表示期望奖励;p^π 指执行策略 π 时,全局状态 s 的概率分布;$Q_i^\pi(x,a_1,\cdots,a_n)$ 是第 i 个智能体的状态 - 动作值函数,它将所有智能体的动作 a 及状态信息 x 作为输入,输出智能体 i 的 Q 值。将 MADDPG 算法扩展至确定性策略 μ,对于式(2)的梯度则写为:

$$\nabla_{\theta_i}J(\mu_i)=E_{x,a\sim D}[\nabla_{\theta_i}\mu_i(a_i\mid o_i)\cdot\nabla_{a_i}Q_i^\mu(x,a_1,\cdots,a_n)\mid_{a_i=\mu_i(o_i)}] \tag{3}$$

式中,D 为经验池,包含 $(x,x',a_1,\cdots,a_n,r_1,\cdots,r_n)$,记录了所有智能体存储的经验样本,分别为当前时刻观察信息 x、下一时刻观察信息 x'、动作 a、奖励值 r。

演员模块通过最大化 $Q_i^\pi(x,a_1,\cdots,a_n)$ 来更新参数,更新规则如下:

$$\theta_i\leftarrow\theta_i+\alpha_{\mathrm{d}}\nabla_{\theta_i}J(\mu) \tag{4}$$

式中,α_{d} 为更新速率参数。

评论家模块使用的是全局信息,它通过最小化损失函数 $L(\theta_i)$ 来实现价值评估,如式(5)所示:

$$L(\theta_i)=E_{x,a,r,x'}[(Q_i^\mu(x,a_1,\cdots,a_n)-y)^2] \tag{5}$$

式中,

$$y=r_i+\gamma Q_i^{\mu'}(x',a_1',\cdots,a_n')\mid_{a_j'=\mu_j'(o_j)} \tag{6}$$

$\mu'=\{\mu_{\theta_1'},\cdots,\mu_{\theta_n'}\}$ 是具有延迟参数的目标策略集合。在式(5)、式(6)中智能体 i 通过环境信息和其他智能体的动作来计算 y 值,以此来函数逼近其他智能体的策略,使得评论家模块可以利用全局信息来指导演员模块。

评论家模块中的目标网络会根据输入的行为和状态得到 Q 值输出,并根据估值网络所产生的真实值来计算梯度损失用于训练网络,目标网络也会间隔一定时间步长后进行更新评论家模块。更新规则如下:

$$\theta_i\leftarrow\theta_i-\alpha_{\mathrm{d}}\nabla_{\theta_i}L(\theta) \tag{7}$$

在整个过程中,每个智能体独立采样,统一学习,并且每个智能体可以有独立的奖励机制。

2　无人机对抗仿真环境

2.1　任务场景

基于无人机集群对抗构想,本文设定的任务场景如下:蓝方以无人机集群的形式向红方基地攻击,试图从不同方向发起袭击。红方采用无人机集群对蓝方无人机集群进行防御拦截。红方无人机集群由多个无人机智能体组成,智能体基于智能对抗算法构建,以实现集群智能作战。

如图 2 所示,红方无人机部署在图中红色星所在的基地周围。本文任务场景中,双方均为 10

架无人机，并且双方无人机处于同一个二维平面内，不考虑高度因素。当蓝方目标突然出现，朝着红方基地移动并试图发起进攻时，红方出动无人机集群对蓝方无人机集群进行拦截，在保护基地的同时尽可能多地击落蓝方无人机。场景中蓝方的首要任务是成功靠近并攻击基地，同时也可对红方拦截的无人机进行攻击，若红方基地被任一蓝方无人机击中，则视为蓝方胜利。红方无人机集群的任务是实现对来袭的蓝方无人机集群进行防御拦截，若蓝方无人机集群全部被击毁则视为红方胜利。

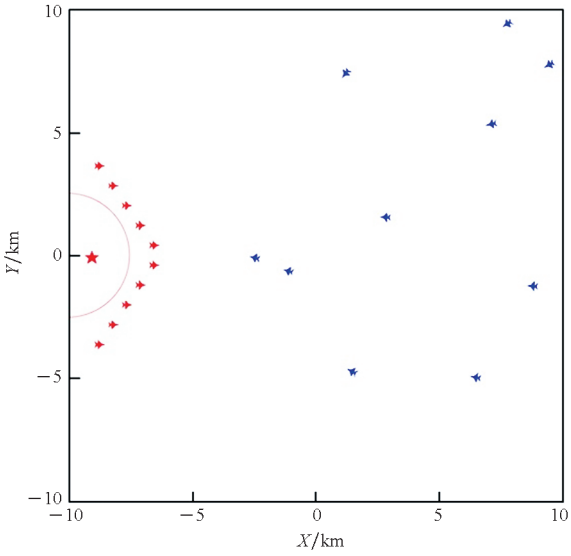


图 2 无人机对抗场景

Fig. 2 UAVs confrontation scenes

2.2 无人机集群对抗建模

任务场景中，无人机的动作是连续的，每架无人机具有三个属性：速度 V，航向 α，坐标位置(X，Y)。如图 3 所示，以战场中心为原点建立坐标系。无人机在对抗过程中要服从以下约束：

1)对抗环境边界约束：

$$-10\ \text{km} \leqslant X \leqslant 10\ \text{km} \tag{8}$$

$$-10\ \text{km} \leqslant Y \leqslant 10\ \text{km} \tag{9}$$

2)速度约束：

$$50\ \text{m/s} \leqslant V \leqslant 150\ \text{m/s} \tag{10}$$

3)最大偏航角约束：

$$-30° \leqslant \Delta\alpha \leqslant 30° \tag{11}$$

4)攻击范围约束：攻击范围是以攻击角 θ、攻击半径 r 形成的虚线扇形区域，它们分别为：

$$\theta = 45° \tag{12}$$

$$r = 200\ \text{m} \tag{13}$$

场景中，红蓝双方无人机同构。每架无人机击落对方目标具有一定的成功率，均设置为

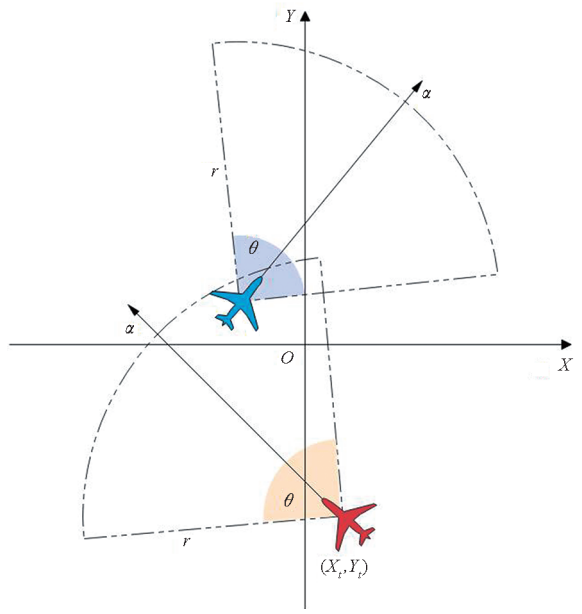


图 3 无人机攻击目标示意图

Fig. 3 Schematic diagram of UAV attack target

60%。假定基地雷达已经探测到蓝方所有无人机的位置和速度信息，红方基地与无人机以及各无人机之间具有通信能力，因此无人机在训练时能够对其他单位的位置、速度等信息完全感知，即无人机的状态空间包括所有无人机的位置、速度和航向信息以及基地的坐标位置。

2.3 无人机运动建模

设定当目标处于无人机攻击范围时，无人机将会自动对目标实施打击，因此无人机在环境中的运动状态只由航向和速度决定。无人机 i 的动作空间 $a_i = \{a_{i1}, a_{i2}, a_{i3}, a_{i4}\}$，$a^c = a_{i1} - a_{i2}$ 为无人机的速度改变值，$p = a_{i3} - a_{i4}$ 为无人机的航向改变值。无人机的运动方程如式(14)所示：

$$\begin{cases} \alpha_{t+1} = \alpha_t + p_t \cdot \alpha_{\max} \\ V_{t+1} = V_t + a_t \\ X_{t+1} = X_t + V_t \cdot \cos(\alpha_t) \cdot T \\ Y_{t+1} = Y_t + V_t \cdot \sin(\alpha_t) \cdot T \end{cases} \tag{14}$$

式中：α_t 为无人机在 t 时刻的航向；α_{t+1} 为无人机在 $t+1$ 时刻的航向；a_t 为速度在 t 时刻的改变值；V_t 为无人机在 t 时刻的速度；V_{t+1} 为无人机在 $t+1$ 时刻的速度；X_t、Y_t 为无人机在 t 时刻的位置；X_{t+1}、Y_{t+1} 为无人机在 $t+1$ 时刻的位置；p_t 为无人机在 t 时刻的转向值。

2.4 算法与环境交互关系

基于 MADDPG 算法的无人机智能体需要在合适的强化学习框架下进行对抗环境的训练和模拟。本文基于 OpenAI 的场景，构建适合用于无

人机强化学习对抗任务的环境和算法框架,形成无人机智能对抗平台。

本文将对抗任务分为训练和对抗两个阶段。如图 4 所示,在训练阶段,环境初始化并将环境信息和奖励值传递给智能体(由多智能体强化学习算法构建),环境信息中包含了智能体的速度、位置等状态信息,奖励函数包括智能体每一步获得的奖励值或惩罚值。智能体根据环境信息选择动作再输出给环境,对抗环境平台根据算法输入的动作生成新的环境信息和奖励值,再传给智能体,算法根据新的奖励值通过学习产生新的动作,形成循环。在对抗阶段,将训练完成后得到的智能体模型与对抗环境进行交互,检测模型的对抗能力和算法的性能。

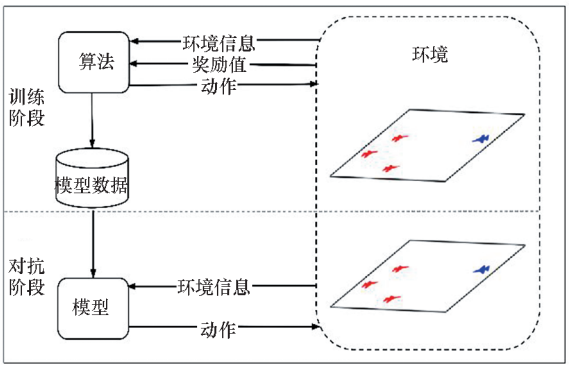


图 4　智能体与环境的交互关系

Fig. 4　Interactions between environment and agents

3　规则约束训练的无人机集群对抗

将 MADDPG 算法直接应用于红方无人机集群对抗中,可以实现对蓝方无人机集群的拦截对抗,但是胜率较低。通过对对战过程进行分析,基于 MADDPG 算法的集群对抗存在下面几个具体问题:①在靠近蓝方目标时,红方无人机因当前速度过大而导致目标逃离攻击范围,进而再去追击时需要经过更多的路程。如图 5 所示,红方无人机与一架蓝方无人机靠近,但蓝方未在其攻击范围内,因此需要重新靠近再次发起攻击。而红方因为当前速度过大,所以转了较大的弯才重新得以追击蓝方,但此时蓝方无人机已经进入红方基地范围并发起了攻击,导致红方无人机拦截失败。②红方无人机会朝着蓝方目标移动并且航向指向蓝方,但是当红方与蓝方都处在各自的攻击范围内时,双方可能同时被对方击毁。我们希望的是红方无人机以最佳态势接近蓝方目标,使得红方无人机较容易击中蓝方,而较难被蓝方无人机击中,形成如图 3 所示的有利态势。红方无人机存活率将得到显著提升,从而更好地完成拦截对抗任务。

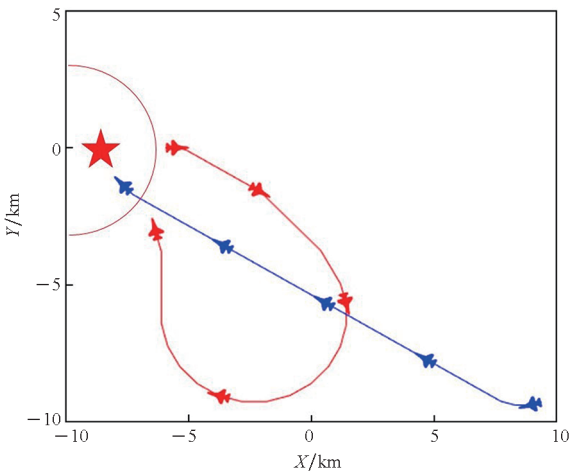


图 5　一对一对抗轨迹

Fig. 5　One-to-one confrontation trajectory

3.1　规则制定

综上情况,基于 MADDPG 的无人机没有很好地实现对蓝方目标快速追击和精准打击。出现这一现象的原因在于:智能体通过不断试错学习找到在模型约束下所遍历行为中的最优解,但这种最优解在实际执行时又远非理想的预期结果,这种现象普遍存在于强化学习的最优训练过程中。解决这一问题最终需要依据客观事实对奖励函数进行精细化设计,但是这对于大多数空战模型而言都是不现实的。这其实是强化学习通过奖励函数实现的“自驱式智能”决策与人类认识上的“客观式规则”决策之间的矛盾。

要调和这一矛盾,还需回到人类对客观事实的认识过程上来。人类空战过程中,也是先进行试错性尝试,然后通过对已有经验的总结、提炼,确定在某个状态下执行某种策略是最优的,进而形成特定情况下的战术条令、条例,也就是规则。然后再结合这些规则进行进一步的“智能体式”的试错与尝试,不断丰富和完善规则,从而由一名“新手”变成“老手”。受这一过程的启发,本文尝试通过“智能体训练—发现问题—编写规则—再次智能体训练—再次发现问题—再次编写规则”的方式,对智能体动作选择进行一定规则化约束来进一步优化智能体的策略。与纯粹的基于算法的智能体在环境中不断试错学习相比,无人机智能体使用一定的规则可以有更少的无效探索动作和更有效的攻击选择,同时也希望这样的规则可以指导智能体的训练。

因此,本文建立了一个基于规则实现的动作输出模块与算法进行融合。在动作输出选择方面,根据无人机当前在环境中的状态,在算法输出的动作和规则动作模块输出的动作两者之间进行判断决策以选择下一步无人机智能体的动作。为红方无人机的动作进行规则约束,主要从航向和速度两块编写规则,设计思想如下:

1)航向模块。为使得无人机智能体在靠近蓝方单位时,可以有一个更好的航向使得蓝方单位处于自身的攻击范围而自身不在蓝方单位的攻击范围内,需要让其提前在合适距离时转向。根据无人机的攻击距离和转向条件,设计当红方无人机位于蓝方攻击夹角之内时,若无人机与蓝方单位之间的距离满足 $2r < dis < 3r$,则计算并选择最快逃离蓝方攻击夹角的转向,反之则继续前行,其中 r 为无人机的攻击距离。无人机在提前逃离蓝方攻击角之后可再次正常调整航向重新朝向蓝方单位,之后蓝方将不能攻击到自身,从而降低被蓝方击中的概率。

2)速度模块。当蓝方单位距离较远时或智能体朝向蓝方单位的背面时,我们希望无人机智能体通过加速快速靠近蓝方。当无人机智能体距离蓝方单位较近时,若此时智能体朝向蓝方单位的正面,我们不希望在未击中蓝方后,由于速度过大而需要更多的距离再追击目标,这时就需要在合适的距离减速。因此无人机智能体的速度与距离蓝方单位的远近有关,于是在一对一的对抗环境中测试了无人机智能体在速度与距离不同比值时的拦截成功率,以此确定速度模块的规则。如图 6 所示,当无人机智能体的速度与距离的比值接近 1.1 时,拦截胜率最高。因此,当智能体的速度与距离比值小于 1.1 时选择加速动作,大于 1.1 时则选择减速动作。

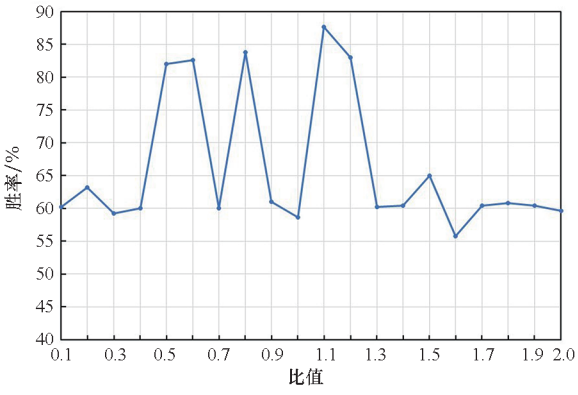


图 6 无人机速度与距离在不同比值时的胜率

Fig. 6 Winning rate of the UAV at different ratios of speed and distance

3.2 算法流程

基于 MADDPG 算法规则约束训练的无人机智能对抗架构如图 7 所示。将算法的动作输出和规则的动作模块进行整合,具体如下:

1)双方无人机在环境中对抗,环境将对抗过程产生的状态传给算法和规则模块。

2)算法根据当前状态生成智能体的下一步动作。而规则模块同时接收环境输入的状态和算法生成的动作,根据状态和动作来判断此时是否需要使用规则约束下一步的动作行为。

3)若不需要使用规则,则直接将算法生成的动作传给环境;若需要使用规则,则由规则模块生成规则动作并传给环境。

4)智能体使用规则约束的方法进行训练,直到本轮训练结束。

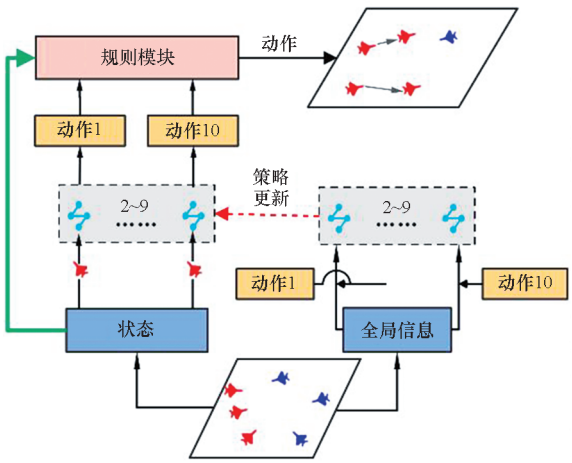


图 7 基于 MADDPG 规则约束训练的无人机对抗架构

Fig. 7 UAV countermeasure architecture based on MADDPG rule constraint training

4 实验验证

将基于 MADDPG 的规则与智能耦合约束训练方法命名为 RIC-MADDPG 方法。为简化研究红方无人机的速度控制与攻击角选择是否有了改善,首先在红蓝双方一对一的对抗环境中进行训练测试。进行多轮测试后选取其中典型的运动轨迹进行对比分析,然后分析对比在集群对抗环境中 RIC-MADDPG 相较于 MADDPG 算法的胜率。

4.1 实验设计

在仿真环境中,蓝方无人机集群使用规则进行控制,红方无人机集群使用智能对抗算法控制,对来袭的蓝方目标进行拦截打击。红方无人机的初始航向为 0,初始速度为 50 m/s。蓝方无人机的初始位置在环境地图中为随机位置,每架无人

机的初始航向为 180°,速度固定为 80 m/s。

MADDPG 和 RIC-MADDPG 算法的演员模块和评论家模块的隐藏层均具有四层隐藏层结构,每层隐藏层为拥有 128 个神经元的全连接层。算法的超参数设置见表 1。

表 1　超参数设置

Tab. 1　Hyper-parameter settings

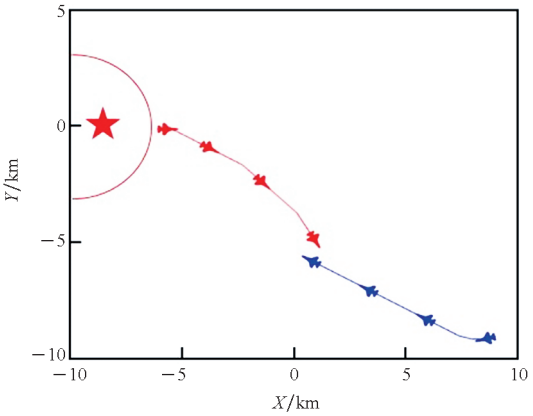
参数名称	参数值
奖励折扣率	0.95
网络隐藏层数	4
每层神经元个数	128
训练轮数	50 000
每轮训练最大步长	200
采样批次	1 024
经验池预设值	1 000 000

无人机在击毁一架敌方无人机时获得 +5 的奖励值,被敌机击毁则获得 -5 的惩罚。为了加快学习速度,引入了无人机与敌方目标之间的距离作为惩罚值,鼓励无人机去靠近敌方目标,将其设置为 $-\min(D_{\text{dis}})$,其中 D_{dis} 为无人机与所有敌方目标的距离的集合。

4.2　实验结果

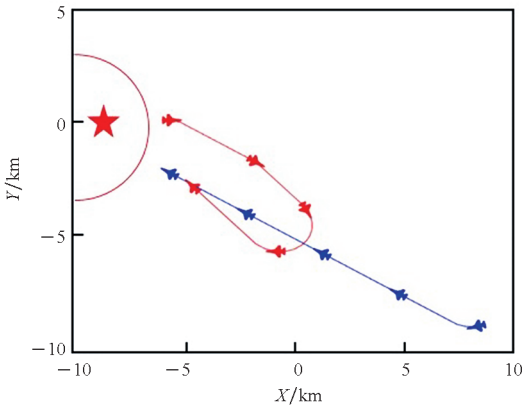
4.2.1　速度控制

经过训练后得到两次典型的使用规则约束方法后无人机的运动轨迹和速度变化,如图 8 和图 9所示。在使用规则约束训练策略后,从图 8(a)和图 9(a)中可以看出,红方无人机先是加速接近蓝方,在蓝方目标即将进入红方攻击范围内时红方进行了减速,这使得无人机有更多的空间调整好攻击角,最终红方在半途成功拦截了蓝方目标。从图 8(b)和图 9(b)中可以看到,



(a) 正面拦截

(a) Frontal interception



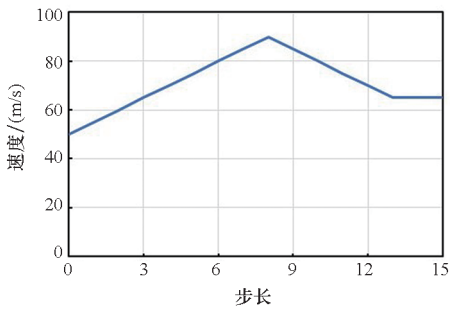
(b) 后方追击

(b) Rear pursuit

图 8　规则约束的对抗轨迹

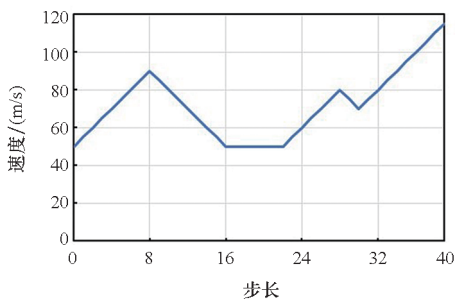
Fig. 8　Confrontation trajectories of rule constraints

红方第一次攻击未能命中蓝方目标,但与之前未使用规则约束训练策略有所不同,红方无人机在转弯前进行了减速,因此转弯半径较小,在转过弯后红方又通过加速快速接近蓝方,在转弯期间红方无人机还进行了小幅度的减速调整,使得航向能快速调整到朝向蓝方目标,最终红方在蓝方进入基地前成功将其击落,完成了拦截对抗任务。



(a) 正面拦截

(a) Frontal interception



(b) 后方追击

(b) Rear pursuit

图 9　红方无人机的速度变化曲线

Fig. 9　Speed curve of the red UAV

4.2.2 攻击角选择

同样,为了简化研究经过规则约束后无人机的攻击角是否有了合理的选择,依然在红蓝双方一对一的正面对抗环境中进行训练测试,从对抗测试结果中选取红蓝双方一次典型的运动轨迹进行分析。

在图 10 中可以看出,红方与蓝方的正面对抗中,当红方即将接近蓝方时选择了向左转向,之后再次调整航向使得蓝方处在自身的攻击范围内而蓝方由于攻击距离不够而无法攻击到红方,因此红方在正面的对抗中成功击毁了蓝方并且保证了自身的安全。

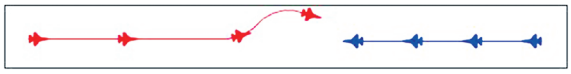


图 10 规则约束的正面对抗轨迹

Fig. 10 The positive confrontation trajectory of a rule constraint

4.3 无人机集群对抗

将红方基于 MADDPG 算法的无人机对抗模型和基于 RIC-MADDPG 方法的无人机对抗模型分别在相同的对抗环境中进行训练。两种算法的参数设置相同,经过 50 000 轮的训练后,获得了两种模型的奖励值变化曲线如图 11 所示。从图中可以发现,当两种模型训练均达到收敛后,RIC-MADDPG 相较于 MADDPG 的平均奖励值得到了提升,从 11 提升到 15,这说明使用规则约束训练策略后无人机对抗模型在训练中表现更为出色,能获得更多的奖励值。实验中同时也获取了红方无人机智能体平均每 1 000 轮的训练时间,其中 MADDPG 平均需要 515 s,而 RIC-MADDPG 平均仅需要 430 s,这大大节省了模型训练时间,提高了无人机智能体的学习效率。

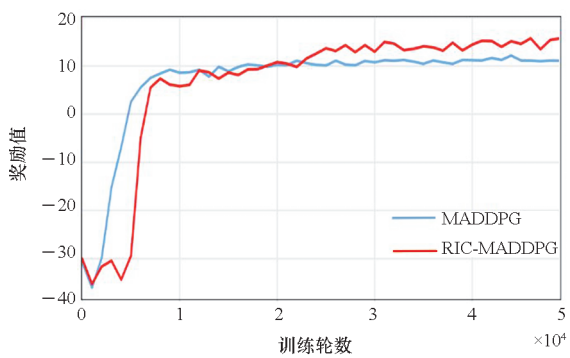


图 11 两种模型训练中的奖励值变化曲线

Fig. 11 Changes in the reward values of two models during training

将上一步训练完成的两种模型在对抗环境下各自进行 500 轮的对抗测试。经过测试后,获得如表 2 所示的红方的平均胜率和红方获胜局中击毁全部蓝方无人机所需的平均战斗步长。通过胜率对比可以看出,在使用规则与智能耦合约束方法后,红方的胜率从 53% 提升到了 79%,大大提高了无人机的胜率。对比红方获胜的每局中击毁对方所有无人机平均所需步长,RIC-MADDPG 方法比 MADDPG 减少了 9 步,红方无人机拦截击毁蓝方所有无人机所需的对战步长缩减,反映出基于 RIC-MADDPG 方法的无人机对抗模型整体对抗能力有了很大的提升。

表 2 对抗胜率和步长

Tab. 2 Against winning percentage and stride length

模型	胜率/%	步长
MADDPG	53	48
RIC-MADDPG	79	39

同时,也获得了如图 12 和图 13 所示的分别基于 MADDPG 和 RIC-MADDPG 方法的红方无人机获得胜利的典型运动轨迹。从中也可以看到,使用规则约束方法后无人机的对抗轨迹有了一些变化,红方无人机更多地从侧面去攻击蓝方。虽然与之前相比仍存在与蓝方无人机正面对抗并互相击中的情况,这是由于距离蓝方较近时,如果再通过调整航向从侧面攻击可能会使得蓝方很容易逃离攻击范围,因此无人机选择了正面对抗,降低蓝方攻击基地的成功率。但是这种情况明显减少,更多的是提前调整了方向进行侧面攻击。

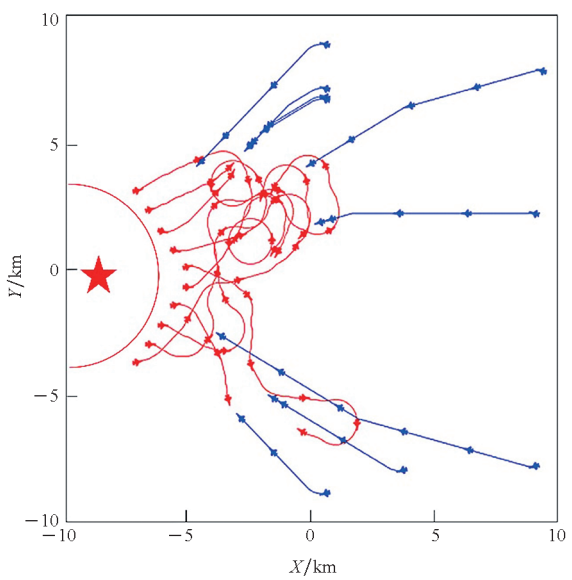


图 12 基于 MADDPG 算法的红蓝双方对抗轨迹

Fig. 12 Confrontation trajectories of red and blue UAVs based on MADDPG algorithm

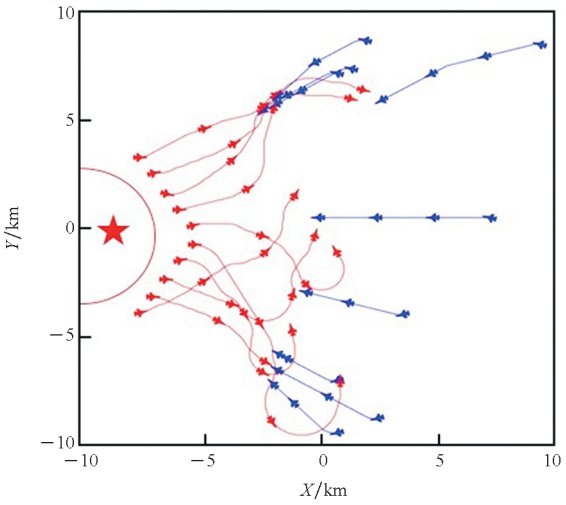


图 13　基于 RIC-MADDPG 算法的红蓝双方对抗轨迹

Fig. 13　Confrontation trajectories of red and blue UAVs based on RIC-MADDPG algorithm

5　结论

本文基于多智能体深度强化学习中的 MADDPG 算法对无人机集群对抗任务进行了研究,对蓝方无人机集群进攻红方基地,红方无人机集群进行防卫对抗的任务场景,在 OpenAI 的环境基础上构建了无人机集群对抗强化学习平台,并基于算法建立了无人机集群对抗模型。通过训练后,红方无人机集群对抗模型能有效对来袭的蓝方无人机集群进行拦截和追击。

针对模型在对抗测试中暴露出的两个问题,即对蓝方无人机拦截过程中,红方无人机速度无法精准控制导致不能快速精确地进行追击的问题,以及红方无人机在与蓝方无人机正面对抗中存在的攻击角选择问题,通过编写相应的速度控制和航向控制规则,在 MADDPG 算法的模型训练中使用规则约束训练策略来改善红方无人机在与蓝方无人机对抗中存在的速度和攻击角控制的问题,并提出了 RIC-MADDPG 方法。实验结果表明,使用规则与智能耦合约束的训练方法后,红方无人机集群在与蓝方无人机集群对抗中的胜率从 53% 大幅提高至 79%,获胜局中红方无人机击毁所有蓝方无人机所需的平均战斗步长从 48 步减少至 39 步,有效提升了无人机集群的作战能力和作战效率,也为后一步基地的防卫提供了更多的防御时间和空间。

论文研究成果对建立无人机集群智能攻防策略训练体系、开展规则与智能相耦合的集群战法研究具有一定参考意义。

参考文献(References)

[1] 高显忠, 王克亮, 彭新, 等. 无人机粉碎机: 硬杀伤式反无人机蜂群关键技术解析[J]. 国防科技, 2020, 41(2): 33 – 38.
GAO X Z, WANG K L, PENG X, et al. Drone-smasher: the key technology analysis on the manner of hard kill to counter UAV swarm [J]. National Defense Technology, 2020, 41(2): 33 – 38. (in Chinese)

[2] 宋怡然, 申超, 李东兵. 美国分布式低成本无人机集群研究进展[J]. 飞航导弹, 2016(8): 17 – 22.
SONG Y R, SHEN C, LI D B. Research progress of distributed low cost UAV cluster in USA[J]. Aerodynamic Missile Journal, 2016(8): 17 – 22. (in Chinese)

[3] 周欢, 赵辉, 韩统, 等. 基于规则的无人机集群飞行与规避协同控制[J]. 系统工程与电子技术, 2016, 38(6): 1374 – 1382.
ZHOU H, ZHAO H, HAN T, et al. Cooperative flight and evasion control of UAV swarm based on rules[J]. Systems Engineering and Electronics, 2016, 38(6): 1374 – 1382. (in Chinese)

[4] 罗德林, 张海洋, 谢荣增, 等. 基于多 agent 系统的大规模无人机集群对抗[J]. 控制理论与应用, 2015, 32(11): 1498 – 1504.
LUO D L, ZHANG H Y, XIE R Z, et al. Unmanned aerial vehicles swarm conflict based on multi-agent system [J]. Control Theory & Applications, 2015, 32(11): 1498 – 1504. (in Chinese)

[5] XING D J, ZHEN Z Y, GONG H J. Offense-defense confrontation decision making for dynamic UAV swarm versus UAV swarm[J]. Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering, 2019, 233(15): 5689 – 5702.

[6] 陈侠, 李光耀, 赵谅. 多无人机协同打击任务的攻防博弈策略研究[J]. 火力与指挥控制, 2018, 43(11): 17 – 23.
CHEN X, LI G Y, ZHAO L. Research on UCAV game strategy of cooperative air combat task [J]. Fire Control & Command Control, 2018, 43(11): 17 – 23. (in Chinese)

[7] DUAN H B, LI P, YU Y X. A predator-prey particle swarm optimization approach to multiple UCAV air combat modeled by dynamic game theory [J]. IEEE/CAA Journal of Automatica Sinica, 2015, 2(1): 11 – 18.

[8] PARK H, LEE B Y, TAHK M J, et al. Differential game based air combat maneuver generation using scoring function matrix[J]. International Journal of Aeronautical and Space Sciences, 2016, 17(2): 204 – 213.

[9] ALEXOPOULOS A, BADREDDIN E. Decomposition of multi-player games on the example of pursuit-evasion games with unmanned aerial vehicles[C]//Proceedings of American Control Conference, 2016: 3789 – 3795.

[10] 何金, 丁勇, 高振龙. 基于 Double Deep Q Network 的无人机隐蔽接敌策略[J]. 电光与控制, 2020, 27(7): 52 – 57.
HE J, DING Y, GAO Z L. A stealthy engagement

maneuvering strategy of UAV based on Double Deep Q Network[J]. Electronics Optics & Control, 2020, 27(7): 52 – 57. (in Chinese)

[11] LI Y, HAN W, WANG Y Q. Deep reinforcement learning with application to air confrontation intelligent decision-making of manned/unmanned aerial vehicle cooperative system[J]. IEEE Access, 8: 67887 – 67898.

[12] 张耀中, 许佳林, 姚康佳, 等. 基于DDPG算法的无人机集群追击任务[J]. 航空学报, 2020, 41(10): 314 – 326.
ZHANG Y Z, XU J L, YAO K J, et al. Pursuit missions for UAV swarms based on DDPG algorithm[J]. Acta Aeronautica et Astronautica Sinica, 2020, 41(10): 314 – 326. (in Chinese)

[13] 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[J]. 航空学报, 2020, 41(12): 342 – 354.
CHEN C, MO L, ZHENG D, et al. Cooperative attack-defense game of multiple UAVs with asymmetric maneuverability[J]. Acta Aeronautica et Astronautica Sinica, 2020, 41(12): 342 – 354. (in Chinese)

[14] XU J, GUO Q, XIAO L, et al. Autonomous decision-making method for combat mission of UAV based on deep reinforcement learning [C]//Proceedings of IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference, 2019: 538 – 544.

[15] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[EB/OL]. (2015 – 09 – 09)[2021 – 02 – 10]. https://arxiv.org/abs/1509.02971.

[16] LOWE R, WU Y, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments [EB/OL]. (2017 – 06 – 07)[2021 – 02 – 10]. https://arxiv.org/abs/1706.02275.

[17] CHEN L, GUO T, LIU Y T, et al. Survey of multi-agent strategy based on reinforcement learning[C]//Proceedings of Chinese Control and Decision Conference, 2020: 604 – 609.

[18] 梁星星, 冯旸赫, 马扬, 等. 多Agent深度强化学习综述[J]. 自动化学报, 2020, 46(12): 2537 – 2557.
LIANG X X, FENG Y H, MA Y, et al. Deep multi-Agent reinforcement learning: a survey [J]. Acta Automatica Sinica, 2020, 46(12): 2537 – 2557. (in Chinese)

[19] 孙长银, 穆朝絮. 多智能体深度强化学习的若干关键科学问题[J]. 自动化学报, 2020, 46(7): 1301 – 1312.
SUN C Y, MU C X. Important scientific problems of multi-agent deep reinforcement learning [J]. Acta Automatica Sinica, 2020, 46(7): 1301 – 1312. (in Chinese)