https://doi.org/10.1051/jnwpu/20224010047

基于 DDPG 的无人机追捕任务泛化策略设计

符小卫, 徐哲, 王辉

(西北工业大学 电子信息学院, 陕西 西安 710129)

摘 要:无人机追逃对抗问题是当今空战领域的研究热点,传统解决方案对此问题存在诸多限制,如模型难以适应复杂动态环境从而快速做出决策、对不同任务场景泛化性较差等问题。基于 DDPG (deep deterministic policy gradient)算法设计了无人机追逃对抗策略;在此基础上,设计多种逃逸无人机的对抗机动策略,利用课程学习思想,在 DDPG 的训练过程中逐步提高逃逸无人机的智能程度,从而递进式地训练追捕无人机的对抗策略。仿真结果表明,相较于直接进行训练,利用课程学习的方法所训练的追捕无人机的追捕策略能够更快收敛,并能更好地执行对敌机的追捕任务,且能够适用于具有多种对抗机动策略的敌机,有效地提升了无人机追逃对抗决策模型的泛化性。

关 键 词:无人机;追逃对抗;深度强化学习;DDPG;课程学习
中图分类号:V279 **文献标志码**:A **文章编号**:1000-2758(2022)01-0047-09

随着无人机性能的提升,其在战场上发挥的作用也将不单单是战场侦查与监视,更多的应该是执行对地攻击、防空火力压制、空战格斗任务,逐步完成从常规的侦察平台到作战平台的转换[1]。无人机间的攻防对抗是新型智能空战的主要作战样式,引起了各个国家军队和学者的关注。空战对操控的实时性要求很高,需要完成对无人机准确、实时的操控,提升无人机的智能化水平刻不容缓,也是未来军用无人机能否成为战场主力的关键[2]。

追逃对抗问题是无人机攻防对抗的核心问题[3],指的是具有利益冲突双方无人机之间的博弈。其中,追捕无人机通过追捕战术机动使逃逸无人机进入到自己的火力攻击范围内,逃逸无人机通过规避战术机动策略,逃离敌机的火力攻击区。

当前对无人机追逃对抗策略研究,主要是通过微分对策法[4]、专家系统法[5]、影响图决策法[6]等,但这些方法存在的缺点是获得解析解的难度比较大、灵活性不足、适用性有限。机器学习中的深度强化学习结合了深度学习强大的感知能力和强化学习的试探学习能力,可以让无人机的决策系统具备自学习的特点[7]。

目前,深度强化学习技术已经被用于无人机的任务决策和运动决策问题。张耀中等[8]基于 DDPG 算法研究了无人机集群对敌方来袭目标的协同追击问题,设计了针对具体追击任务的一种引导型回报函数,经训练后无人机集群能较好地完成协同追击任务。陈灿等[9]针对不同机动能力的无人机间攻防对抗问题,基于集中评判-分布执行的多智能体强化学习算法,研究了多无人机协同攻防自主决策的方法,为多无人机系统对抗提供新的研究思路。在对目标的追踪问题上,史豪斌等[10]尝试将视觉伺服控制与强化学习结合,利用 Sarsa 学习算法训练使得旋翼无人机自主调节视觉伺服增益,验证了该方法相对于 PID 控制与基于图像的视觉伺服控制方法具有更好的追踪效果。苏治宝等[11]在未知环境下利用 Q-learning 算法实现多移动机器人围捕移动目标,给出具体设计方案并验证影响围捕效果的因素。深度强化学习技术在无人机控制技术的落地应用证明了其应用于无人机的追捕对抗机动决策的可行性。对这些文献分析可知,所研究的任务较为单一,而对多种机动策略的敌机追捕任务不能有效迁移,因此需要研究一种对于不同逃逸策略的敌机

收稿日期:2021-06-11 **基金项目**:航空科学基金(2020Z023053001)资助
作者简介:符小卫(1976—),西北工业大学副教授,主要从事无人机控制、管理与决策与航空火力控制研究。
e-mail:fxw@nwpu.edu.cn

可泛化的有效适用方法。

本文以无人机追逃对抗问题为研究背景,基于DDPG 算法建立了无人机追逃对抗的数学模型和优化目标,训练无人机学习追逃对抗策略;在研究的基础上,设计多种逃逸无人机的对抗机动策略,基于课程学习思想的训练方式,逐步提高逃逸无人机的智能程度从而递进式地训练无人机的追捕对抗策略。所提出的训练算法能够成功训练出泛化性强的无人机追捕对抗策略,能够追捕不同对抗机动策略的敌机。

1 问题描述与建模

1.1 单无人机追逃对抗问题

针对追逃博弈问题,建立有控制约束的二局中人零和微分博弈模型。无人机追逃的几何模型如图1 所示。

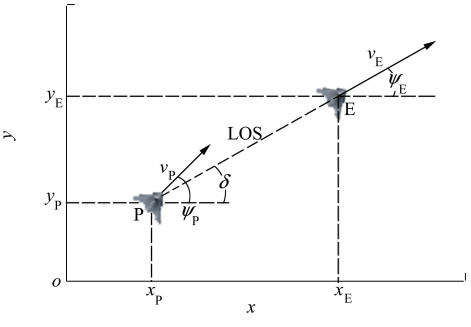


图 1 二维平面追逃博弈几何模型

图 1 中,P,E 分别代表追捕无人机和逃逸无人机,v_P,v_E 分别指追捕无人机和逃逸无人机的速度大小,ψ_P,ψ_E 分别指追捕无人机和逃逸无人机的航向角,规定航向角逆时针为正方向,δ 为目标视线(line of sight,LOS)的夹角视线角,目标视线指追捕无人机 P 指向逃逸无人机 E 的射线。追捕无人机的目标是以最短的时间捕获目标,逃逸无人机的目标是远离追捕无人机,避免在预设时间段被捕获或者最大化延迟被追捕无人机捕获的时间,追逃博弈标准微分博弈数学描述为公式(1)和(2)。

$$\min T_c = f(v_P, \psi_P, v_E, \psi_E, L) \tag{1}$$

$$\max T_c = h(v_P, \psi_P, v_E, \psi_E, L) \tag{2}$$

式中:L 为追逃双方的距离;T_c 是追捕无人机 P 捕获逃逸无人机 E 的时刻。公式(1)、(2)分别为追捕无人机和逃逸无人机的优化目标函数。

1.2 无人机运动学模型

无人机的运动学方程为公式(3)。

$$\begin{cases} \dot{x}_i = v_i \cos\psi_i \\ \dot{y}_i = v_i \sin\psi_i \\ \dot{v}_i = a_i \\ \dot{\psi}_i = \omega_i \end{cases} \quad (i = \mathrm{P,E}) \tag{3}$$

式中:ω_i 表示无人机的角速度;a_i 表示无人机的加速度大小。无人机的运动控制变量约束为公式(4)。

$$\begin{cases} v_{i\min} \leqslant v_i \leqslant v_{i\max} \\ -a_{i\max} \leqslant a_i \leqslant a_{i\max} \\ -\omega_{i\max} \leqslant \omega_i \leqslant \omega_{i\max} \end{cases} \quad (i = \mathrm{P,E}) \tag{4}$$

式中:v_{Pmin},v_{Emin} 是双方无人机的最小速度;v_{Pmax},v_{Emax} 是双方无人机的最大速度;a_{Pmax},a_{Emax} 是双方无人机的最大加速度;ω_{Pmax},ω_{Emax} 是双方无人机的最大角速度,规定角速度为正值表示航向角逆时针变化,即航向角增大。

无人机初始状态为公式(5)。

$$\begin{cases} x_i(t_0) = x_{i0} \\ y_i(t_0) = y_{i0} \\ v_i(t_0) = v_{i0} \\ \psi_i(t_0) = \psi_{i0} \end{cases} \quad (i = \mathrm{P,E}) \tag{5}$$

敌我距离在追捕无人机的捕获范围内,即为捕获成功,如(6)式所示。捕获范围 l_c 可以是无人机的载荷作用范围或者武器攻击范围。

$$d_{\mathrm{PE}} \leqslant l_c \tag{6}$$

式中,d_{PE} 是追捕无人机与逃逸无人机的距离。

2 基于 DDPG 的无人机追捕策略

在用 DDPG 算法解决无人机追逃对抗问题时,首先需定义无人机的状态空间、动作空间、奖励函数。本节先具体介绍模型设计过程,再介绍 DDPG 训练算法。

2.1 模型设计

2.1.1 状态空间

设定无人机携带机载 GPS 设备和陀螺仪,可以获得自身的位置信息和速度信息即 $\boldsymbol{\xi}_P = [x_P, y_P, v_P, \psi_P]$;携带机载脉冲多普勒火控雷达载荷设备,能获得探测目标的位置信息和速度信息 $\boldsymbol{\xi}_E = [x_E, y_E, v_E, \psi_E]$。

为了增加算法的适应性,减小神经网络的输入处理负担,聚焦于策略优化,本文使用相对位置关系来建立状态空间模型,其表达式如公式(7) 所示。

$$\boldsymbol{S} = [\alpha_{\mathrm{P}}, \alpha_{\mathrm{E}}, \alpha_{\mathrm{PE}}, d_{\mathrm{PE}}, v_{\mathrm{P}}, \Delta v_{\mathrm{PE}}] \tag{7}$$

式中:α_{P},α_{E} 分别是追捕无人机和逃逸无人机速度方向与目标视线 LOS(由追捕无人机位置指向逃逸无人机的向量) 的夹角;α_{PE} 是追捕无人机速度方向与逃逸无人机速度方向的夹角;Δv_{PE} 是指追捕无人机与逃逸无人机的速度大小差,其计算如公式(8) 所示。

$$\Delta v_{\mathrm{PE}} = v_{\mathrm{P}} - v_{\mathrm{E}} \tag{8}$$

2.1.2　动作空间

如公式(3)和(4)所示,无人机的控制输入为一个二维向量,即动作空间

$$\boldsymbol{A} = [a_i, \omega_i] (i = \mathrm{P}, \mathrm{E}) \tag{9}$$

2.1.3　状态转移方程

强化学习中智能体在当前状态 S_t 采取行为 A_t 转移到下一状态 S_{t+1} 存在一个概率问题,该概率称为状态转移概率 P_s^a。在本文的模型中,设定状态转移概率为 1,表示下一时刻的状态量是根据当前状态和动作完全决定的。双方无人机具体的运动状态转移方程如公式(10) 所示。

$$\begin{pmatrix} x_i \\ y_i \\ v_i \\ \psi_i \end{pmatrix}_{t+1} = \begin{pmatrix} x_i \\ y_i \\ v_i \\ \psi_i \end{pmatrix}_t + \begin{pmatrix} v_i \cos\psi_i \\ v_i \sin\psi_i \\ a_i \\ \omega_i \end{pmatrix}_t \Delta T (i = \mathrm{P}, \mathrm{E}) \tag{10}$$

2.1.4　回报函数

本文回报函数的设计采用稀疏回报和引导型回报函数相结合的方式,如公式(11)所示。

$$\begin{cases} r_{t1} = d_{t-1} - d_t \\ r_t = k r_{t1} + r_{t2} + r_{t3} \end{cases} \tag{11}$$

式中:r_t 代表无人机总奖励;r_{t1} 是设计的引导型奖励回报,为追逃双方的距离变化回报,即只有当双方距离变小时追捕无人机获取正奖励;d_t,d_{t-1} 分别代表 t 和 $t-1$ 时刻追捕无人机和逃逸无人机的距离,k 是比例系数;r_{t2} 表示追捕无人机离逃逸无人机过远的稀疏奖励;r_{t3} 表示追捕无人机完成任务的稀疏奖励,定义分别为公式(12) 和(13)。

$$r_{t2} = \begin{cases} -R_{\mathrm{far}}, & \text{if} \quad d_t > D_{\mathrm{far}} \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

式中,D_{far} 表示相对距离阈值,如果相对距离超过阈值,则认为追捕无人机离逃逸无人机过远,给予无人机负回报 R_{far}。

$$r_{t3} = \begin{cases} R_{\mathrm{finish}}, & \text{if} \quad d_t < l_{\mathrm{c}} \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

式中:l_{c} 的物理意义同公式(6),当追捕无人机完成对逃逸无人机的捕获任务,给与无人机正回报奖励 R_{finish}。

2.2　DDPG 算法

DDPG 算法是由 DeepMind 团队提出的,并在连续动作空间下经验证取得了很好的效果。它是基于 Actor-Critic(AC)算法框架,采用 DQN 中经验回放机制和双网络结构进行改进,是一种深度确定性策略梯度算法[12]。

DDPG 框架主要包括环境、存放样本的经验池、actor 网络模块及 critic 网络模块。强化学习中智能体通过与环境的不断交互产生样本数据,将这些样本存入到一个经验池中,在经验池中采用随机策略抽取一定数量的样本(mini batch samples)来训练网络参数。为了提高算法的学习效率,DDPG 算法采用双网络结构,无论是 actor 网络还是 critic 网络,都包含 eval 网络和 target 网络,2 个 target 网络与 eval 网络对应,是一对结构完全相同的神经网络,eval 网络会在每步训练更新网络参数 (θ^u,θ^Q),而 target 网络参数($\theta^{u'}$,$\theta^{Q'}$) 则定期通过软更新方式复制 eval 网络参数,更新公式如(14) 式所示。

$$\begin{cases} \omega' \leftarrow \tau\omega + (1-\tau)\omega' \\ \theta' \leftarrow \tau\theta + (1-\tau)\theta' \end{cases} \tag{14}$$

式中,τ 为软更新系数,一般取 0.01。同时为了提高智能体对环境的探索能力,需要在 actor 中的 eval 网络输出动作 $\mu(s_t)$,增加随机噪声 N。一般设定动作噪声服从正态分布,即公式(15)。

$$N \sim N(\mu, \sigma^2) \tag{15}$$

取 $\mu = 0$,σ 随着迭代步数的增加逐渐减小,降低对环境的探索性,同时无人机不同于其他模型,需要考虑无人机的机动性能约束,因此最终动作 a 是经过噪声和机动约束限定后的实际动作,如公式(16) 所示。

$$a = f(\pi_\theta(S) + N) \tag{16}$$

式中,f 为无人机动作约束函数,其定义参考公式(17)。

$$a = f(a') = \begin{cases} a_{\max}, a' \geqslant a_{\max} \\ a', -a_{\max} < a' < a_{\max} \\ -a_{\max}, a' \leqslant -a_{\max} \end{cases} \tag{17}$$

actor 网络与 critic 网络使用不同的损失函数进行训练。在使用批数据时,critic 网络的损失函数,如公式(18)所示。

$$L(\theta^Q)=\frac{1}{K}\sum_i\left(Q(s_i,a_i\mid\theta^Q)-y_i\right)^2 \quad (18)$$

式中:K 为批数据的样本个数;$Q(s_i,a_i\mid\theta^Q)$ 为 eval 网络评价当前时刻状态与 actor 中的 eval 网络生成动作的价值;y_i 的定义如公式(19)所示。

$$y_i=r(r_i,a_i)+\gamma Q'(s_{i+1},\mu'(s_{i+1}\mid\theta^{u'})\mid\theta^{Q'}) \quad (19)$$

式中:$r(r_i,a_i)$ 为当前时刻状态执行动作 a_i 后的即时奖励;γ 为奖励衰减系数;$Q'(s_{i+1},\mu'(s_{i+1}\mid\theta^{u'})\mid\theta^{Q'})$ 为 target 网络评价下一时刻状态 s_{t+1} 和 actor 中的 target 网络对下一时刻状态所选动作 $\mu(s_{i+1})$ 的价值。

critic 网络的参数更新如公式(20)所示。

$$\begin{cases}\theta^Q=\theta^Q+\alpha_Q\nabla_{\theta^Q}L(\theta^Q)\\ \nabla_{\theta^Q}L(\theta^Q)=\dfrac{1}{K}\cdot\\ \displaystyle\sum_{i=1}^{K}\left(y_i-Q(s_i,a_i\mid\theta^Q)\right)\nabla_{\theta^Q}Q(s,a\mid\theta^Q)\mid_{s=s_i,a=a_i}\end{cases} \quad (20)$$

式中:α_Q 是 critic 网络的学习率;actor 网络的参数更新如公式(21)所示。

$$\begin{cases}\theta^u=\theta^u+\alpha_u\nabla_{\theta^u}J\\ \nabla_{\theta^u}J=\dfrac{1}{K}\cdot\\ \displaystyle\sum_i\left(\nabla_aQ(s,a\mid\theta^Q)\mid_{s=s_i,a=u(s_i)}\nabla_{\theta^u}Q(s\mid\theta^u)\mid_{s=s_i}\right)\end{cases} \quad (21)$$

式中,α_u 是 actor 网络的学习率。

3 基于课程学习的 DDPG 算法流程

3.1 基于课程学习的训练方法

深度强化学习存在的问题是无人机在学习复杂任务时训练算法收敛比较慢,原因是样本探索效率比较低。本文利用课程学习方式,提高样本探索效率,增快训练算法的收敛速度。课程学习是指机器学习任务中逐渐增加任务难度以增快学习速度的方法。课程学习核心思想是需要逐步调整学习的任务分布,智能体更容易在简单任务中获得奖赏回报,先选择相对简单的任务进行策略学习,然后将策略迁移到复杂任务上,能够有效降低复杂任务的探索难度[13]。

本文针对无人机追逃对抗决策问题,设计了基于课程学习核心思想的训练方法,训练方法总共分为 3 个步骤,具体如表 1 所示。3 个步骤中不同的是逃逸无人机的机动决策方式,逃逸无人机的智能程度呈现递增。在 DDPG 的训练过程中通过不断提高逃逸无人机的决策模型智能程度进而设置从简而难的学习任务,递进式地训练无人机的追捕对抗策略,能够有效提升模型的泛化性。

表 1 追逃对抗机动训练方法设计

步骤	训练控制策略设置
1	设定逃逸无人机做固定航向的匀速直线运动,训练追捕无人机的机动策略
2	设定逃逸无人机做一种随机运动,即每步的机动决策在合理的约束范围内随机产生,训练追捕无人机的机动策略
3	设定逃逸无人机做一种经典的逃脱策略机动[14],训练追捕无人机的机动策略

运用 DDPG 算法学习训练之前,需要无人机与环境进行交互,获取大量的实验数据作为训练样本。本文实验双方无人机的初始位置和速度在设定的范围内服从均匀分布,随机产生作为各自初始状态,追捕无人机根据 actor 策略网络输出并经过动作限制处理得到控制动作信号,而逃逸无人机依据表 1 的步骤选择自身逃逸策略,双方无人机根据当前状态及动作信息,利用公式(10)更新计算得到下一时刻的状态,然后追捕无人机获得环境反馈的立即奖励,这样一个追捕无人机与环境的交互经验[s_t,α_t,r_t,s_{t+1}]就产生,存入经验池中,等待无人机与环境交互一定的步数后,按照随机抽样方法抽取一定数量的样本,进而对 4 个网络的参数按公式(20)、(21)和(14)进行更新。在等待经验池填满时候开始训练。

3.2 DDPG 离线算法训练流程

本文实验中,控制周期设置为仿真步长。需要注意的是,状态 s_t 和 α_t 的下标 t 代表时间步而不是实际飞行时间,实际飞行时间为 $T=t\Delta T$。基于 DDPG 的无人机追捕机动策略训练算法流程如下所示:

1. 初始化存储量大小为 M 的经验池 D

2. 初始化 actor 网络和 critic 网络的策略网络 eval:$\mu(s;\theta^{u})$ 和 $Q(s,a\mid\theta^{Q})$;
   将 eval 网络的参数拷贝给对应 target 网络:$\theta^{u'}\leftarrow\theta^{u},\theta^{Q'}\leftarrow\theta^{Q}$
   actor 网络和 critic 网络的目标网络 target 初始化完成:$u(s;\theta^{u'})$ 和 $Q(s,a\mid\theta^{Q'})$
3. For episode = 1 to MaxEpisode do
4. 初始化 OU 过程 $N(t)$
5. 在设定范围内随机初始化追捕无人机、逃逸无人机初始状态,获得仿真环境初始状态 s_0
6. For t = 1 to MaxStep do
7. 通过状态 s_t 选择追捕无人机动作 $a_t=f_{\text{clip}}(u(s_t\mid\theta^{u})+N_t)$,$N_t$ 是 OU 随机噪声,f_{clip} 表示动作限制处理过程(超过约束范围采取边界值)
8. 根据训练方式为逃逸无人机选择机动策略,决定机动动作
9. 将控制信号输入到仿真环境中的无人机中,积分得到下一步时间无人机状态,计算得到环境状态 s_{t+1}
10. 同时得到环境的立即回报
11. 将经验样本$[s_t,a_t,r_t,s_{t+1}]$ 存储在 D 中
12. 从 D 中随机抽样得到大小为 BatchSize 的样本集 $\{[s_t,\alpha_t,r_t,s_{t+1}]\}$
13. 更新 critic 动作网络的策略网络 eval 参数 θ^{Q}
14. 更新 actor 动作网络的策略网络 eval 参数 θ^{u}
15. 更新 2 个网络中 target 网络参数 $\theta^{Q'},\theta^{u'}$
16. If 满足回合结束机制,Break
17. End For

4　仿真验证

4.1　实验验证

实验采用图 1 场景想定,同时采用常见的红蓝对抗作战法,设定红方为追捕无人机,蓝方为逃逸无人机。仿真环境全部基于 Python 语言编写,利用 Pycharm Community 2020.2 和 Anaconda3 平台,深度学习环境采用 Tensorflow 1.14.0,计算机配置为 CPU Inter i7-9700F@ 3.00GHz,内存 16 GB。

实验设定的训练超参数介绍如下:折扣因子 γ 取 0.9,经验池 M 大小取 30 000,抽取样本数取 64,actor 和 critic 网络学习率分别取 0.001 和 0.002,单次回合最大时间步取 300,总共训练 4 000 回合,仿真步长取 0.1。实验环境参数如表 2 所示。

表 2　无人机追逃对抗仿真实验参数

实验参数名称	参数数值
空间维数	2
追捕无人机初始位置 $x_{\text{P}},y_{\text{P}}$/m	[0,50]×[0,50]
追捕无人机初始速度 v_{P}/(m·s^{-1})	12.5
追捕无人机初始航向 ψ_{P}	[0,2π)
追捕无人机速度范围/(m·s^{-1})	[9,15]
追捕无人机加速度范围/(m·s^{-2})	[-2,2]
追捕无人机角速度范围/(rad·s^{-1})	[-0.6,0.6]
逃逸无人机初始位置 $x_{\text{E}},y_{\text{E}}$/m	[0,50]×[0,50]
逃逸无人机初始速度 v_{E}/(m·s^{-1})	10
逃逸无人机初始航向 ψ_{E}	[0,2π)
逃逸无人机速度范围/(m·s^{-1})	[9,15]
逃逸无人机加速度范围/(m·s^{-2})	[-2,2]
逃逸无人机角速度范围/(rad·s^{-1})	[-0.6,0.6]
过远距离阈值 D_{far}/m	600
捕获半径 l_{c}/m	20

actor 网络的 target 网络和 eval 网络采用单隐层的前馈神经网络,每层神经元个数为[6,128,2],隐藏层采用 relu(x)作为激活函数,输出层采用 tanh(x)作为激活函数,让智能体的输出动作限定在一定的范围内;critic 网络也采用单隐层的前馈神经网络,网络输入为无人机的状态与 actor 网络生成的动作,所以其 target 网络和 eval 网络每层神经元个数为[8,128,1],隐藏层采用 relu(x)作为激活函数,输出层采用 tanh(x)作为激活函数。训练使用 Adam Optimizer 作为网络参数的优化器。

4.2　方法有效性实验

本文定义 4 种不同的训练方式,4 种学习训练方式前 3 种分别只采用表 1 的 3 种学习训练方式,第 4 种则采用课程学习的方式。第 4 种训练方式训练时,步骤 1 先基于第 1 种训练方式训练追捕无人机的机动决策网络,然后将训练好机动决策网络的参数(网络权值、阈值)进行迁移,然后步骤 2 利用第 2 种训练方式再次训练追捕无人机的机动决策网络,同理步骤 3 以第 3 种训练方式再次训练追捕无人机的机动决策网络。

4.2.1 训练过程

本文采用平均奖励指标观察所设计训练算法的训练情况和收敛情况，定义每 100 回合的平均总奖励作为统计值，开始不足 100 回合时使用仅有所有回合总奖励的平均值。

如图 2a)～2c)所示，分别为第 1 种、第 2 种、第 3 种学习训练方式下 DDPG 算法的平均奖励曲线，横坐标表示训练回合数，纵坐标表示每个回合相应的平均奖励(每 100 回合的总奖励值平均)。从图中可以看出，当只采用单一训练方式时，DDPG 训练算法会在 500 回合内稳定收敛。

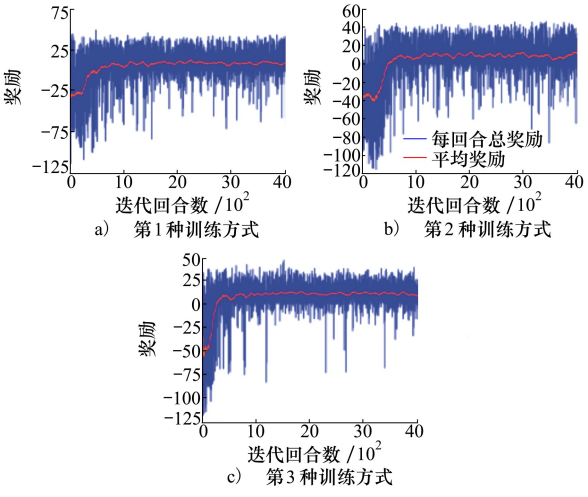


图 2 单一训练方式平均奖励曲线图

深度强化学习训练的策略网络一般只适用于具体特定的环境，而在现实环境下，逃逸无人机的对抗机动策略可能随时改变，而训练好的机动决策网络执行的效果可能不佳，适用性较差，这也是强化学习普遍存在的模型泛化性缺点。而本文针对此问题，首先在一个简单的环境中预先训练机动决策网络，而当环境逐渐复杂，在面对不同智能程度的逃逸无人机时，将适用于简单环境的策略迁移到该任务上来，进行一次重训练，而不用从随机初始化的网络参数开始，从而降低在复杂任务上的训练难度，这也是第 4 种训练方式的目的所在。它的本质是根据逐渐复杂的样本分别，逐步调整学习网络的参数，以适用于更为复杂的学习任务。

为了表现基于课程学习的训练方法的优势，将第 4 种训练方式中步骤 2 和步骤 3 的平均奖励曲线分别和第 2 种训练方式、第 3 种训练方式的平均奖励曲线图 2b)、图 2c)进行对比，如图 3a)～3b)所示。

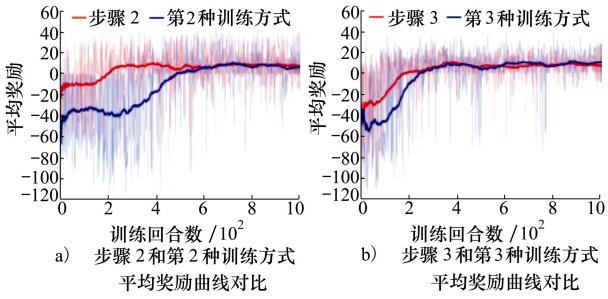


图 3 第 4 种训练方式平均奖励曲线图

在图 3a)和图 3b)都表明了在已经训练好的网络参数上再进行新环境(逃逸无人机的逃逸对抗策略不同)的重训练，相比于直接在新环境上从零开始的训练，都更快取得了算法收敛效果。

4.2.2 验证过程

为了进行充分的仿真测试，本文分别对 4 种训练方式得到的训练结果进行试验。追逃双方无人机的初始位姿、初始速度均随机产生，逃逸无人机采用经典的逃逸策略机动，共进行 10 000 次蒙特卡洛测试。测试的数据结果如表 3 所示，分别包含追捕任务的成功率、捕获时间、平均奖励。

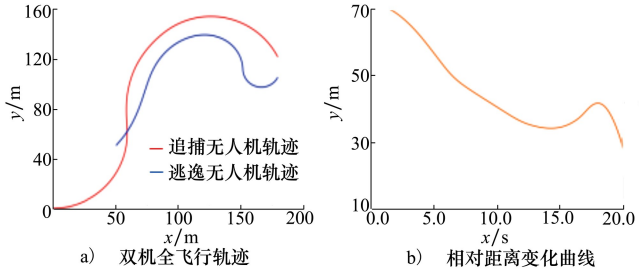
表 3 4 种训练方式测试结果

训练方式	奖赏均值(方差)[最小值,最大值]	成功率/%	捕获时间均值/s(方差)[最小值,最大值]
第 1 种	6.47 (264.39) [-42.58,32.73]	95.00	7.81 (62.13) [3.10,24.45]
第 2 种	6.84 (257.64) [-38.36,33.72]	98.00	6.80 (57.68) [3.38,21.87]
第 3 种	7.63 (178.56) [-32.79,36.45]	99.00	6.52 (33.26) [2.57,16.60]
第 4 种	**7.97** **(164.63)** **[-29.11,37.28]**	**99.00**	**6.30** **(31.69)** **[2.33,16.48]**

从表 3 可以看出，在表格每一列中，用加粗字体标出了每列的最值(平均奖赏和成功率为最大值，捕获时间为最小值)，第 4 种训练方式基于课程学习的 DDPG 算法训练流程取得了最大的平均奖赏、成功率和最短的捕获时间，可见基于课程学习的

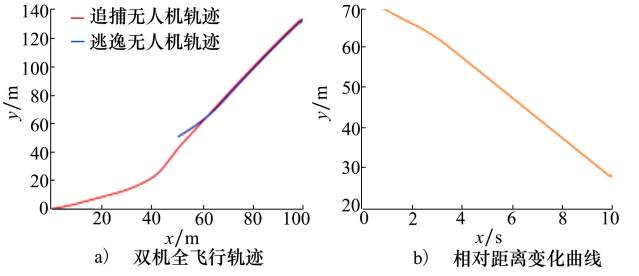
DDPG 算法取得了最好的训练效果和鲁棒性。

设置相同的实验设定,无人机追逃对抗过程的仿真如图 4~7 所示。从图 4~7 的 a)图可以看出,4 种训练方式都使得追捕无人机能够在一定程度上实现了追捕逃逸无人机的效果。而图 4~7 的 b)图可以看出,在设置相同的实验初始条件下,由于训练时设定敌机的逃逸策略分别为第 1 种(固定航向匀速直线的逃逸策略)和第 2 种(随机运动的逃逸策略)方式,与实验时敌机的逃逸策略(采用经典逃脱策略)不同,因此取得了较差的结果(第 1 种双机距离最终不稳定,追捕效果极差,第 2 种双机最终距离(28 m)也较远)。第 3 种训练和试验时的敌机的逃逸策略均相同,因此相较前 2 种,双机最终距离(26 m)有所减小。而本文设计的第 4 种采用课程学习的训练方式所训练的决策网络使得追捕无人机的追捕能力有了一定的提升,相较于前 3 种取得了较好的追捕效果,双机的最终距离(22 m)进一步减小。



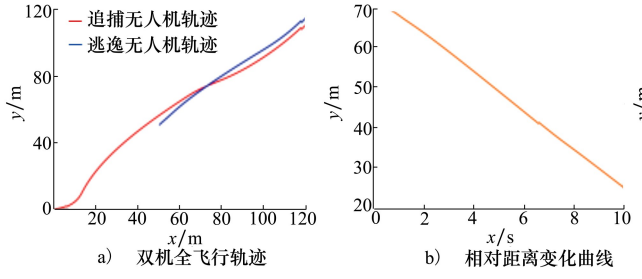
a)　双机全飞行轨迹　　b)　相对距离变化曲线

图 4　第 1 种训练方式对战飞行轨迹和测试结果



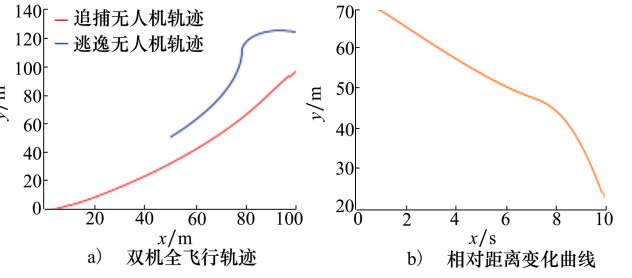
a)　双机全飞行轨迹　　b)　相对距离变化曲线

图 5　第 2 种训练方式对战飞行轨迹和测试结果



a)　双机全飞行轨迹　　b)　相对距离变化曲线

图 6　第 3 种训练方式对战飞行轨迹和测试结果



a)　双机全飞行轨迹　　b)　相对距离变化曲线

图 7　第 4 种训练方式对战飞行轨迹和测试结果

综上所述,第 4 种训练方式,基于课程学习的思想,通过逐渐提升逃逸无人机的逃逸对抗策略智能程度,设置任务从简到难,应用时取得了不错的战术效果,具有一定的鲁棒性。

5　结　论

本文针对空战中无人机追逃博弈问题,基于 DDPG 算法设计了无人机的追捕对抗策略。利用课程学习的训练方式,在 DDPG 的训练过程中逐渐提升逃逸无人机的智能程度,递进式地训练无人机的追捕对抗策略,研究的主要结论为:

1) 仿真表明了所设计的模型经过 DDPG 算法的学习能够达到稳定收敛,使得追捕无人机能够较好地追捕具有多种策略的逃逸无人机,成功率均达到 95%以上。

2) 针对逐步提升智能程度的逃逸无人机,基于课程学习的训练方法,相较于前 3 种对简单任务直接训练的方法,能够快速收敛;并能够适用于多种对抗机动策略的敌机,有效地提升了无人机追捕对抗决策模型的泛化性,为无人机执行较难任务的研究提供了新的思路。

下一步工作主要是研究如何简化基于课程学习的训练过程,以及将研究工作扩展到三维空间。

参考文献:

[1] 邵将, 徐扬, 罗德林. 无人机多机协同对抗决策研究[J]. 信息与控制, 2018, 47(3): 347-354
SHAO Jiang, XU Yang, LUO Delin. Cooperative combat decision-making research for multi UAVs[J]. Information and Control, 2018, 47(3): 347-354 (in Chinese)

[2] 魏航. 基于强化学习的无人机空中格斗算法研究[D]. 哈尔滨: 哈尔滨工业大学, 2015
WEI Hang. Research of UCAV air combat based on reinforcement learning[D]. Harbin: Harbin Institute of Technology, 2015 (in Chinese)

[3] 孟秋楠. 有约束微分对策问题及其在空战对抗中的应用[D]. 沈阳: 沈阳航空航天大学, 2018
MENG Qiunan. Differential game with constrained its application in air combat[D]. Shenyang: Shenyang Aerospace University, 2018 (in Chinese)

[4] 谢剑. 基于微分博弈论的多无人机追逃协同机动技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2015
XIE Jian. Differential game theory for multi UAV pursuit maneuver technology based on collaborative research[D]. Harbin: Harbin Institute of Technology, 2015 (in Chinese)

[5] CHIN H H. Knowledge-based system of super maneuver selection for pilot aiding[J]. Journal of Aircraft, 1971, 26(12): 1111-1117

[6] MATHESON J E. Using influence diagrams to value information and control[M]. New York: John Wiley & Sons, 1988

[7] 李高垒, 马耀飞. 基于深度网络的空战态势特征提取[J]. 系统仿真学报, 2017, 29(增刊1): 98-105
LI Gaolei, MA YaoFei. Feature extraction algorithm of air combat situation based on deep neural networks[J]. Journal of System Simulation, 2017, 29(suppl 1): 98-105 (in Chinese)

[8] 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 Yaozhong, XU Jialin, YAO Kangjia, et al. Pursuit missions for UAV swarms based on DDPG algorithm[J]. Acta Aeronautica et Astronautica Sinica, 2020, 41(10): 314-326 (in Chinese)

[9] 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 Can, MO Li, ZHENG Duo, 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)

[10] 史豪斌, 徐梦. 基于强化学习的旋翼无人机智能追踪方法[J]. 电子科技大学学报, 2019, 48(4): 553-559
SHI Haobin, XU Meng. An intelligent tracking method of rotor UAV based on reinforcement learning[J]. Journal of Electronic Science and Technology University, 2019, 48(4): 553-559 (in Chinese)

[11] 苏治宝, 陆际联, 童亮. 一种多移动机器人协作围捕策略[J]. 北京理工大学学报, 2004, 24(5): 26-29
SU Zhibao, LU Jilian, TONG Liang. Strategy of Cooperative Hunting by Multiple Mobile Robots[J]. Journal of Beijing Institute of Technology, 2004, 24(5): 26-29 (in Chinese)

[12] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J]. Computer Science, 2015, 8(6): A187

[13] 杨瑞, 严江鹏, 李秀. 强化学习稀疏奖励算法研究—理论与实验[J]. 智能系统学报, 2020, 15(5): 888-899
YANG Rui, YAN Jiangpeng, LI Xiu. A survey on sparse reward algorithms in reinforcement learning-theory and experiment[J]. CAAI Transactions on Intelligent Systems, 2020, 15(5): 888-899 (in Chinese)

[14] SHEN H L, FURUKAWA T, DISSANAYAKE G, et al. A time-optimal control strategy for pursuit-evasion games problems[C] //Proceedings of International Conference on Robotics and Automation, New Orleans, LA, USA, 2004

Generalization strategy design of UAVs pursuit evasion game based on DDPG

FU Xiaowei, XU Zhe, WANG Hui

(School of Electronics and Information, Northwestern Polytechnical University, Xi′an 710129, China)

Abstract: UAVs pursuit evasion game is a research hotspot in the field of air combat. Traditional solutions have many limitations to this problem, such as the difficulty of the model to adapt to complex dynamic environments to quickly make decisions, and the poor generalization of different mission scenarios. Based on the DDPG(deep deterministic policy gradient) algorithm, a mathematical model of UAVs pursuit and evasion countermeasures is established in this paper. On this basis, this research designs a variety of countermaneuver strategies for escaping UAV, and uses the training method of course learning ideas. In the training process, the intelligence of the escaping UAV is gradually improved, so as to progressively train the confrontation strategy of the chasing UAV. The simulation results show that compared with direct training, the pursuit strategy of the chasing UAV trained by the research method of course learning can converge faster, and can better perform the hunting mission of enemy aircraft, and can be applied to a variety of enemy aircraft with a variety of maneuvering strategies, which effectively improved the generalization of the UAV′s pursuit and escape confrontation decision model.

Keywords: UAV; pursuit-evasion game; deep reinforcement learning; DDPG; curriculum learning

引用格式:符小卫, 徐哲, 王辉. 基于 DDPG 的无人机追捕任务泛化策略设计[J]. 西北工业大学学报,2022,40(1): 47-55

FU Xiaowei, XU Zhe, WANG Hui. Generalization strategy design of UAVs pursuit evasion game based on DDPG[J]. Journal of Northwestern Polytechnical University, 2022, 40(1): 47-55 (in Chinese)