基于改进双延迟深度确定性策略梯度法的无人机反追击机动决策

郭万春[1]，解武杰[1]，尹　晖[2]，董文瀚[1]
(1. 空军工程大学航空工程学院，西安，710038；2. 空军工程大学教研保障中心，西安，710051)

摘要　针对近距空战下的自主机动反追击问题，建立了无人机反追击马尔科夫(Markov)决策过程模型；在此基础上，提出了一种采用深度强化学习的无人机反追击自主机动决策方法。新方法基于经验回放区重构，改进了双延迟深度确定性策略梯度(TD3)算法，通过拟合策略函数与状态动作值函数，生成最优策略网络。仿真实验表明，在随机初始位置/姿态条件下，与采用纯追踪法的无人机对抗，该方法训练的智能无人机胜率超过 93%；与传统的 TD3、深度确定性策略梯度(DDPG)算法相比，该方法收敛性更快、稳定性更高。

关键词　深度强化学习；近距空战；无人机；双延迟深度确定性策略梯度法
DOI　10.3969/j.issn.1009-3516.2021.04.003
中图分类号　V279　　**文献标志码**　A　　**文章编号**　1009-3516(2021)04-0015-07

Research on UAV Anti-Pursing Maneuvering Decision Based on Improved Twin Delayed Deep Deterministic Policy Gradient Method

GUO Wanchun[1], XIE Wujie[1], YIN Hui[2], DONG Wenhan[1]
(1. Aeronautical Engineering College, Air Force Engineering University, Xi'an 710038, China;
2. Teaching & Reseerch Support Center, Air Force Engcieeng Univerity)

Abstract　In view of the problem of autonomous maneuvering counter-pursuing in close air combat, a Markov decision-making process model for UAV counter-pursuing is established, and for the above-mentioned reasons, an autonomous maneuvering decision-making method for unmanned aerial vehicles (UAVs) based on deep reinforcement learning is proposed. The new method is based on the empirical replay area reconstruction, and improves the Twin Delayed Deep Deterministic policy gradient (TD3) algorithm, and generates the optimal strategy network by fitting the strategy function and the state action value function. The simulation experiments show that under condition of random initial position/attitude, being confronted with the drones adopted by the pure pursuit methods, the winning rate of intelligent drones trained by this method exceeds 93%. Compared with traditional TD3 and Deep Deterministic policy gradient (DDPG), this method is faster at convergence and higher in stability.

Key words　deep reinforcement learning; close air combat; UAV; twin delayed deep deterministic policy gradient method

收稿日期：2021-04-21
作者简介：郭万春(1996—)，男，安徽马鞍山人，硕士生，研究方向：飞行器导航、制导与飞行控制。E-mail:1223110879@qq.com
通信作者：董文瀚(1979—)，男，陕西西安人，教授，博士生导师，研究方向：飞行器导航、制导与飞行控制。E-mail:dongwenhan@sina.com

引用格式：郭万春，解武杰，尹晖，等.基于改进双延迟深度确定性策略梯度法的无人机反追击机动决策[J]. 空军工程大学学报(自然科学版)，2021，22(4)：15-21. GUO Wanchun, XIE Wujie, YIN Hui, et al. Research on UAV Anti-Pursing Maneuvering Decision Based on Improved Twin Delayed Deep Deterministic Policy Gradient Method[J]. Journal of Air Force Engineering University (Natural Science Edition), 2021, 22(4): 15-21.

近年来,各种控制理论和方法研究为自主空战决策提供了解决方案。文献[1]利用差分博弈论,将空战模型建模为一个确定的、完全信息的追逃博弈模型。文献[2]研究了一种实时自主一对一的近似动态规划空战方法。文献[3]采用了一种基于可达性的方法来解决追逃博弈问题。此外,还有多级影响图法[4]、滚动时域法[5]和基于统计学原理的方法[6]等。由于现实环境的不确定性以及真实测试昂贵、耗时和危险等原因,这些探索大多停留在理论研究阶段,难以付诸工程实践和实战。

深度强化学习(deep reinforcement learning, DRL)正成为利用 AI 解决决策问题的主流研究方向[7-10]。文献[11]采用了深度 Q 学习(deep Q-learning network,DQN)的方法控制无人机的速度和转角,然而 DQN 对次优动作高估的状态动作值超过最优动作的动作值时将无法找到最优动作,并且它只能处理离散的、低维的动作空间,这与大多实际情境不符。文献[12]使用异步的优势行动者评论家算法(asynchronous advantage actor-critic,A3C)训练无人机进行空战,利用多线程的方法,同时在多个线程里分别与环境进行交互学习,避免了 DQN 中出现的经验回放相关性过强的问题,但是训练出的无人机空战性能有待提高。文献[13]假定对战的两架飞机速度恒定,使用深度确定性策略梯度算法(deep deterministic policy gradient, DDPG)训练了无人机,虽然取得了良好的训练效果,但是训练出的追击策略较为单一,并且没有考虑飞机的火力打击区域,仅仅以两机间的距离在某一范围内作为成功打击目标的准则。

本文讨论自主空战中深度强化学习在无人机反追击的应用。考虑非对称性的追逃博弈问题,具有扇形火力打击区域的两架无人机进行追击/反追击的空中对抗,攻击(以下标注为 ATTACK)无人机采用纯追踪法(pure pursuit)打击目标,目的是训练速度不大于 ATTACK 无人机的智能(以下标注为 RL)无人机摆脱其追击并进行反制。

1 问题描述

无人机自主机动反追击使用参数化动作空间马尔科夫决策过程[14]的形式化框架,由一个五元组构成:$<S,A,P,r,\gamma>$。RL 无人机通过与环境交互学习状态到动作的映射关系以此获得最大累计期望回报。假设这是一个理想模型,环境的动态特性 $P(\cdot\,|(s,a))=1$ 是确定的,即不存在风等对无人机飞行有干扰的因素。时间步为 t 时观测到的无人机状态 $s_t \in S$。RL 无人机从可用的行动集合 A 中选用行动 $a_t \in A$,环境在 a_t 的作用下,转换至新状态 s_{t+1},在进行状态转移到下一个新状态的同时产生奖励 $r(s_t,a_t)$。RL 无人机根据新观测到的状态 s_{t+1},再做出新的决策,采取行为 a_{t+1},依次反复进行直至达到环境的终止状态。$\gamma \in [0,1]$ 为未来回报折扣因子,RL 无人机旨在寻找一个策略 π 使得其从任意初始状态 s_0 出发在达到终止状态时获得最大的累计奖励:

$$\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\Big|_{a_t=\pi(s_t)} \tag{1}$$

根据文献[15],无人机反追击模型可描述为:

$$\begin{cases}\dfrac{\mathrm{d}\theta}{\mathrm{d}t}=0\\ \dfrac{\mathrm{d}\varphi}{\mathrm{d}t}=\dfrac{N_a g\cos\gamma_c}{v}\\ \dfrac{\mathrm{d}x}{\mathrm{d}t}=v\cos\varphi\\ \dfrac{\mathrm{d}y}{\mathrm{d}t}=v\sin\varphi\\ \dfrac{\mathrm{d}z}{\mathrm{d}t}=0\end{cases} \tag{3}$$

式中:t 为仿真时间;(x,y,z)为无人机的坐标;v 为速度;$\theta=0$ 为航迹倾斜角初始值;与 x 轴方向夹角 φ 为偏航角;$N_a \in [-N_z^{\max},N_z^{\max}]$为无人机在水平面内的转向过载(其方向与机体纵轴垂直),转向过载可沿速度方向和垂直速度方向分解为切向过载 N_z 和法向过载 N_r,记速度方向与机体纵轴间的夹角为速度倾斜角 γ_c,$\boldsymbol{g}$ 为重力加速度,记 $V_{RL}^{\max}$ 为 RL 无人机速度最大值。

设 ATTACK 无人机和 RL 无人机的偏航角分别为 α 和 β,则其位置信息分别为 $X_{\mathrm{ATTACK}}=(x_1,y_1,\alpha)$,$X_{\mathrm{RL}}=(x_2,y_2,\beta)$。

根据文献[11],ATTACK 无人机对 RL 无人机进行火力打击的示意图如图 1 所示。

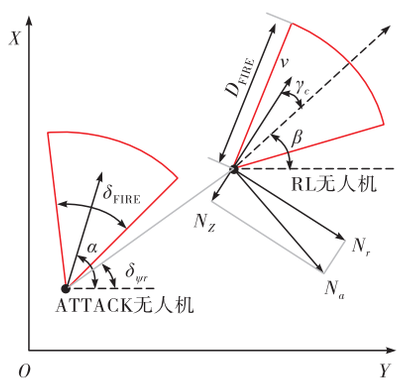

图 1 火力打击示意图

其连线方向与 x 轴方向夹角为 δ_{ψ_T},距离为 D_T。

2 架无人机具有相同的攻击范围，角度范围为$\frac{\pi}{b}$，打击半径 0.251 cm，以机体纵轴为对称轴，以 D_{FIRE} 作为打击半径，圆心角为 δ_{FIFE} 的一个扇形区域。

ATTACK 无人机进行火力打击的规则为纯追踪法：其速度方向将时刻指向 RL 无人机，试图将 RL 无人机追击在自己的打击范围内，该策略在文献[16]中被证明是十分有效的追击手段。根据最优追逃策略[17]，设 ATTACK 无人机相邻两次时间步长内的偏航角变化量 φ_{ATT} 满足：

$$\varphi_{\mathrm{ATT}}=\begin{cases}-\varphi_{\max},\varphi_u\leqslant-\varphi_{\max}\\ \varphi_u,-\varphi_{\max}<\varphi_u<\varphi_{\max}\\ \varphi_{\max},\varphi_u\geqslant-\varphi_{\max}\end{cases}\tag{3}$$

式中：$\varphi_u=\alpha-\delta_{\psi_T}$，最大变化量为 $\varphi_{\max}=\frac{\pi}{6}$。RL 无人机旨在更新策略躲避 ATTACK 无人机的攻击并将其暴露在自己的打击范围内。

2　无人机反追击 Markov 决策过程建模

2.1　无人机飞行状态空间

由于两架无人机在同一高度上进行追击与反追击的空中对抗，记位置信息为 $D=(x_1,y_1,\alpha,x_2,y_2,\beta)$，在每个时间步长的开始，以 ATTACK 无人机的位置和偏航角为基准，将原有的坐标系逆时针旋转 α 角，使新坐标系的原点位于 ATTACK 无人机处，并且 x 轴方向与 ATTACK 无人机的偏航角重合。在新坐标系下，得出 RL 无人机的位置满足以下关系：

$$x'_2=(x_2-x_1)\cos\alpha+(y_2-y_1)\sin\alpha\tag{4}$$

$$y'_2=(x_2-x_1)\sin\alpha-(y_2-y_1)\cos\alpha\tag{5}$$

$$\beta'=\beta-\alpha\tag{6}$$

新坐标系下无人机的相对位置信息为：

$$D'=(0,0,0,(x_2-x_1)\cos\alpha+(y_2-y_1)\sin\alpha,(x_2-x_1)\sin\alpha-(y_2-y_1)\cos\alpha,\beta-\alpha)\tag{7}$$

值得注意的是，这个新坐标系是随着 ATTACK 无人机的位置和偏航角实时变化的，由于 ATTACK 无人机也在做机动，所以每一时间步的原点和坐标的横纵轴方向，在真实物理空间上是不一样的，引入这个坐标系只是为了描述它们的相对位置。相对位置信息的 6 维向量有 3 维始终为 0，因此通过相对坐标系可以进一步将无人机的飞行状态空间压缩一倍。构造新的观测状态为：

$$s=((x_2-x_1)\cos\alpha+(y_2-y_1)\sin\alpha,(x_2-x_1)\sin\alpha-(y_2-y_1)\cos\alpha,\beta-\alpha,N_z)\tag{8}$$

2.2　无人机飞行动作空间与终止奖励函数

在每个时间步的开始，无人机从其动作空间允许的速度和转向过载向环境提供一个动作，给定的动作会立即更新当前的速度和偏航角，在剩余的时间步长中保持不变。其中 ATTACK 无人机采用纯追踪法的策略，保持一个恒定的速度，可以选择从一个连续范围的转弯角度，使用纯追踪法可以让 ATTACK 无人机稳步拉近与对手的距离并接近对手使其置于火力打击范围。RL 无人机使用强化学习算法，它的动作空间包含速度和转向过载值，定义为：

$$A=\{v,N_a\}\tag{9}$$

式中：v 为速度；$N_a\in[-N_z^{\max},N_z^{\max}]$ 为水平面内无人机控制量，$N_z^{\max}$ 为无人机最大法向过载值。

两种无人机机动能力数据见表 1。

表 1　机动能力数据

参数	ATTACK	RL
速度/(km/s)	0.1	[0.05,0.1]
偏航角变化量	$\left[-\frac{\pi}{6},\frac{\pi}{6}\right]$	$\left[-\frac{\pi}{6},\frac{\pi}{6}\right]$
转向过载值	[−5,5]	[−5,5]

定义反追击成功时的回报，即：

$$r(s,a)=\begin{cases}1,D_T\leqslant D_{\mathrm{FIRE}}\cap|\beta-\delta_{\psi_T}|\leqslant\frac{\delta_{\mathrm{FIRE}}}{2}\\ 0,\mathrm{else}\end{cases}\tag{10}$$

3　基于深度强化学习的无人机反追击算法

值函数过估计的问题既在 DQN 中存在，也存在于“行动者-评论家”网络。在 DQN 中采用的双重深度 Q 学习[17]（double deep Q-learning network，DDQN）方法可以一定程度上降低过估计的误差，但在“行动者—评论家”网络中使用类似 DDQN 的方法是无效的，因此本文采用双延迟深度确定性策略梯度算法 TD3 来解决值函数过估计的问题；为了提高训练前期的效率和训练后期的稳定收敛，将经验回放区进行重构并改进传统的随机抽样策略。

3.1　无人机反追击算法框架

经验回放区重构将成功经验和失败经验分为两个经验回放区。如果 RL 无人机反追击任务满足式(10)中 $r(s,a)=1$，则被认为是暂时的成功经验储存在成功经验回放区中，记为 R_s；相反，满足 $r(s,a)=0$，则将失败经验储存在失败经验回放区中，记为 R_f。由于 RL 的奖励过程中存在着时间延迟，所以存储在 R_s 中的一些即将达到失败前的经验也与失败有关。因此，可以把这些经验从 R_s 以 η_f 的比例提取出来。具体来说，对每一个时间步，如果是成功

经验，$<s_t,a_t,r_t,s_{t+1}>$将被直接储存在 R_s 中；如果是失败经验，将$<s_t,a_t,r_t,s_{t+1}>$存放至 R_f，同时以 η_f 的比例从 R_s 中提取出上述的失败经验。

改进传统的随机采样策略：更新时，行动者和评论家同时从 R_s 以 ξ_s 比例以及从 R_f 抽取$(1-\xi_s)$的样本来进行优化。其中，考虑训练前期的效率和训练后期局部最优的制衡，ξ_s 应随着训练总迭代次数 M 衰减：

$$\xi_s = 0.9 \times e^{-\left(\frac{i}{M}\right)} \tag{11}$$

经验回放区重构的 TD3 方法见图 2。

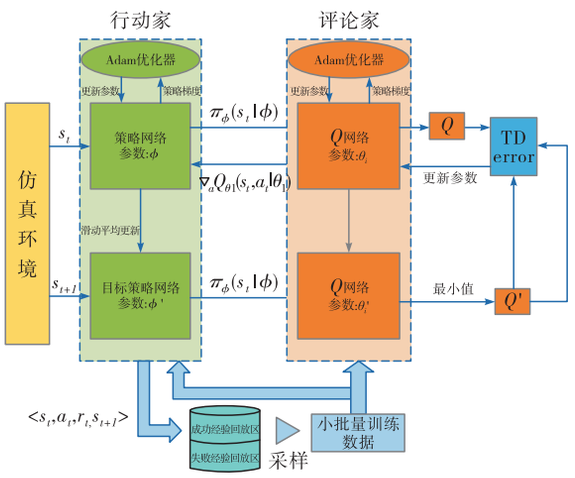

图 2 经验回放区重构的 TD3 方法框架图

传统 TD3 使用经验回放区解决训练数据间的相关性，在环境中探索得到的样本数据，以状态转换序列$<s_t,a_t,r_t,s_{t+1}>$为单元存放于回放区中，当记忆库满时，则删掉旧的样本数据，保证回放库中的容量不变。每次更新时，行动者和评论家都会从中随机的抽取一部分样本进行优化，来减少一些不稳定性。但是，随机采样会导致训练效率低，收敛性能差。本文提出的经验回放区重构可以一定程度上解决这一问题。

从重构经验回放区采样得到一个小批量的训练数据，TD3 通过梯度上升/下降算法更新当前网络的参数。然后再通过优化的滑动平均方法更新目标网络的参数，使得目标网络参数缓慢变化，以此提高学习的稳定性。

3.2 基于改进 TD3 的无人机反追击决策算法

TD3 采用行动者-评论家框架，包含 6 个神经网络，见表 2。

表 2 TD3 中的神经网络

网络位置	行动者	评论家
当前网络	策略网络	Q_1 网络 Q_2 网络
目标网络	目标策略网络	目标 Q_1 网络 目标 Q_2 网络

拟合策略函数的策略网络 π_φ，参数为 φ，输入为当前状态 s_t，输出无人机的动作：

$$a_t = \pi_\varphi(s_t|\varphi) \tag{12}$$

策略网络图见图 3，网络参数见表 3。

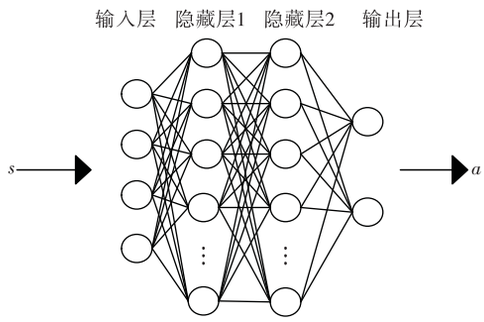

图 3 策略网络

表 3 策略网络的参数

参数名称	描述
输入层	输入无人机的状态
隐藏层 1	400 个神经元
隐藏层 2	300 个神经元
输出层	输出当前状态下的动作
激活函数	隐藏层为线性整流函数(ReLU)，输出层为 Sigmoid 函数和双曲正切函数(Tanh)
优化方法	梯度上升

网络参数通过确定性策略网络梯度定理更新：

$$\nabla_\varphi J(\varphi) = \frac{1}{N}\sum_i \nabla_a Q_{Q_1}(s,a \mid Q_1)\Big|_{s=s_i,a=\pi_\varphi} \cdot \nabla_\varphi \pi(\varphi)(s \mid \varphi)\Big|_{s=s_i} \tag{13}$$

目标策略网络 $\pi_{\varphi'}$ 的参数为 φ'，输入为下一状态，输出下一状态的动作：

$$a_{t+1} = \pi_{\varphi'}(s_{t+1}|\varphi') \tag{14}$$

拟合状态动作值函数的 Q_1 网络 Q_{θ_1} 和 Q_2 网络 Q_{θ_2}，参数分别为 θ_1 和 θ_2，输入为当前状态 s_t 和实际执行的动作 a_t，输出为状态动作值即 Q_1 值和 Q_2 值：

$$Q_i = Q_{\theta_i}(s_t,a_t|\theta_i) \tag{15}$$

Q_1 网络还输出状态动作值函数对动作的梯度 $\nabla_a Q_{\theta_1}(s_t,a_t|\theta_1)$ 用于式(13)的参数更新。

状态动作值网络见图 4，网络参数见表 4。

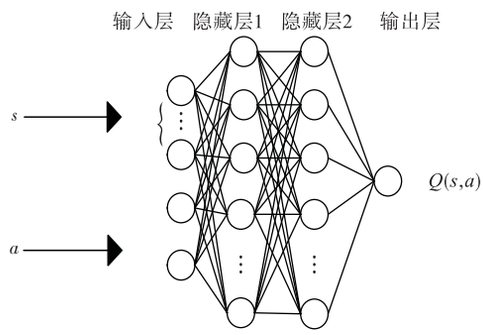

图 4 状态动作值网络

表 4　状态动作值网络的参数

参数名称	描述
输入层	当前状态、实际执行的动作
隐藏层 1	400 个神经元
隐藏层 2	300 个神经元
输出层	Q 值
激活函数	隐藏层为线性整流函数(ReLU)，输出层为双曲正切函数(Tanh)
优化方法	梯度下降

目标 Q_1 网络 $Q_{\theta'_1}$ 和目标 Q_2 网络 $Q_{\theta'_2}$ 的参数分别为 θ'_1 和 θ'_2。输入是下一状态 s' 和目标策略网络输出的下一状态的行为 a'，输出下一状态动作值即 Q_1' 值和 Q_2' 值：

$$Q_1' = Q_{\theta'_i}(s_{t+1}, \pi_{\varphi'}(s_{t+1} \mid \varphi') \mid \theta'_i) \tag{16}$$

TD3 在两个目标 Q 网络中选择较小的 Q 值，防止 DDPG 中评论家网络对动作 Q 值过估计的问题：

$$Q' = \min\{Q_1', Q_2'\} \tag{17}$$

对于 Q_1 网络和 Q_2 网络，定义损失函数：

$$L = N^{-1} \sum_i (y - Q_{\theta_j}(s, a \mid \theta_j)))^2 \Big|_{s=s_i, a=a_i} \quad (j = 1,2) \tag{18}$$

通过损失函数的反向传播算法更新得到 Q_1 网络和 Q_2 网络的参数。其中 y 表示时序差分(temporal-difference，TD)目标值：

$$y = r + \gamma \min_{i=1,2} Q_{\theta'_i}(s', \tilde{a}) \tag{19}$$

式中：$\tilde{a}$ 在目标策略网络的输出加上噪声剪枝增加了算法稳定性：

$$\tilde{a} \leftarrow \pi_{\varphi'}(s') + \varepsilon, \varepsilon \sim clip(\mathrm{N}(0, \tilde{\delta}), -c, c) \tag{20}$$

Q_1 网络和 Q_2 网络的 TD 误差分别为：

$$TD_{errori} = y - Q_{\theta_i}(s_t, a_t \mid \theta_i) \tag{21}$$

对评论家更新 2 次后，行动家再进行更新，策略网络 π_φ、Q_1 网络和 Q_2 网络的参数通过滑动平均分别得到目标策略 $\pi_{\varphi'}$ 网络和目标 Q_1 网络和 Q_2 网络的参数：

$$\begin{cases} \theta_i = \tau\theta_i + (1-\tau)\theta'_i \\ \varphi' = \tau\varphi + (1-\tau)\varphi' \end{cases} \tag{22}$$

无人机反追击决策算法训练流程如下：

初始化经验回放库 R_f、R_s、策略网络 π_φ、Q_1 网络和 Q_2 网络，并将它们的参数复制给目标策略网络 $\pi_{\varphi'}$ 和目标 Q_1 网络和目标 Q_2 网络。

For episode = 1,2,…,M：

$a \leftarrow \pi_{\theta_\mu}(s) + \varepsilon$，其中 $\varepsilon \sim N(0, \sigma)$，给行为添加噪声；

获取无人机飞行仿真环境的初始状态。

For $t = 1,2,\cdots,T$：

根据当前策略和探索噪声，获得行为 a；

执行行为 a，获得回报 $r(s,a)$ 和下一个状态 s'；

状态转换序列存储于回放记忆库 R_f、R_s 中；

R_f、R_s 中分别以 ξ_s 和 $(1-\xi_s)$ 的比例抽取 N 个状态转换序列，作为策略网络和 Q_i 网络的训练数据；

根据式(20)计算 $\tilde{a}$；

根据式(19)计算时序差分值 y；

根据式(18)更新 Q_1 网络和 Q_2 网络参数；

IF t mod 2：

根据式(13)计算样本策略梯度，更新策略网络；

根据式(22)更新目标策略网络和目标 Q_1 网络和 Q_2 网络。

End if

End for

End for

输出最优策略网络参数以及最优策略。

4　仿真验证与分析

设置训练集为 $M=10\ 000$，随机初始化两架无人机初始位置与姿态信息。其中 ATTACK 无人机的初始位置在原点，偏航角在[0，2π]内均匀分布；RL 无人机的初始位置是以原点为中心的横纵坐标 x、y 变量呈正态分布的随机分布，其中 x、y 方向标准差均为 0.5 km。这样的随机初始化可以做到让 RL 无人机在一个时间步长后摆脱追击并进行反制，实际上加快了收敛速度。超参数设置见表 5。

表 5　超参数

参数	数值
学习率	10^{-4}
L_2 正则化权重衰减率	10^{-5}
折扣率	0.99
批量大小	100
经验回放库容量	50 000
探索噪声	0.005
最大探索次数	10 000
最大时间步长	100
策略延迟更新参数	2
目标策略更新参数	0.995
η_f	0.2
R_s	40 000
R_f	10 000

分别使用面向连续动作空间的确定性策略方法 TD3 算法和 DDPG 算法进行训练，每 100 次训练记录当前 100 次训练的胜率。训练效果见图 5。

可以看出，与基准 DDPG 算法相比，本文方法的胜率约高出 10%左右。改进后的 TD3 算法虽然在训练的后期与传统的 TD3 算法能达到的胜率相

差不大,但是由于经验回放区的重构,新的采样策略代替原始的随机采样策略,使得在训练伊始可以更多学习到任务成功经验序列,使改进后的 TD3 算法在训练前期的收敛速度较快,波动也较弱。

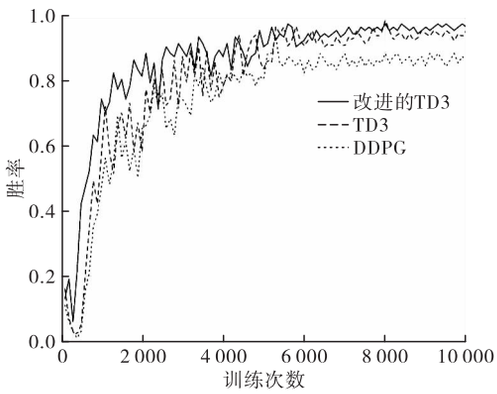

图 5 训练效果对比图

算法训练完成后,对训练集进行大量仿真测试,在训练集中进行 4 000 次对抗,每次包括 100 个时间步长,该时间步之内未能分出胜负则为平局。仿真表明,RL 无人机成功实施反追击的次数为 3 761 次,成功率为 94.025%,达到了预定目的。与 TD3 算法和 DDPG 算法的对比见表 6。

表 6 测试效果对比表

算法	成功反追击次数	成功率/%	决策用时/s	平均每场决策总时间/ms
改进的 TD3	3 789	94.28	22.77	5.69
TD3	3 685	92.13	21.09	5.53
DDPG	3 346	83.65	14.82	3.71

可以看到,改进后的 TD3 算法胜率略高于 TD3 算法,明显高于 DDPG 算法,但是由于整个算法当中比 DDPG 多了两个神经网络的参数,所以从决策时间来看,决策时间均略长于 DDPG 算法。

在测试集中,RL 无人机使用本文训练好的策略进行反追击的胜率也很难低于 93%。图 6～10 展示了测试集中具有代表性的双机轨迹,从中可以观察到典型的 RL 反追击策略。图 6～9 显示了 RL 无人机为了获胜所采用的最常见的策略,图 10 展示了 RL 平局时的大部分场景。

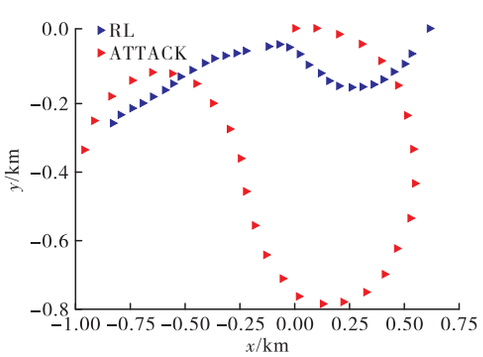

图 6 轨迹 1

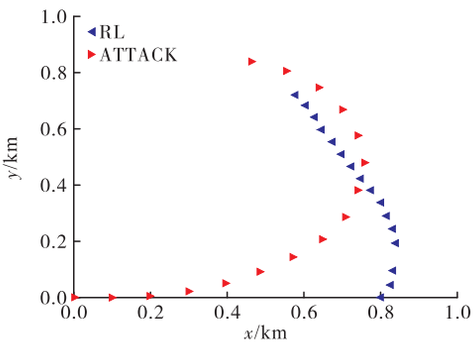

图 7 轨迹 2

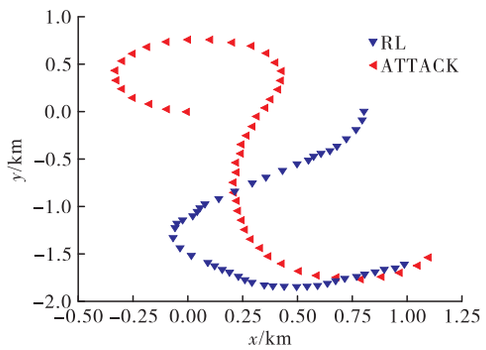

图 8 轨迹 3

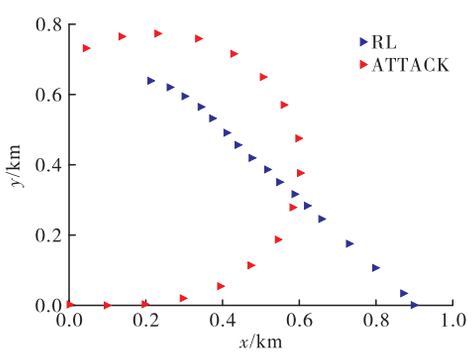

图 9 轨迹 4

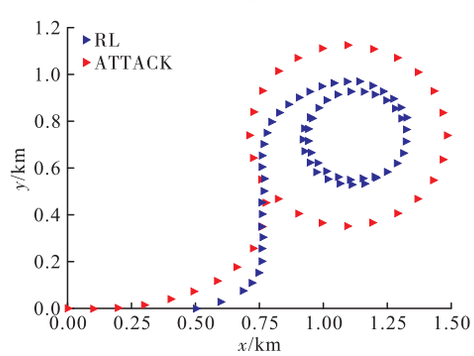

图 10 轨迹 5

如图 6 所示,RL 无人机通过在被追击的早期改变速度同时调整转向过载值来做出长期决策:一开始加速是防止在前期就被击中,随后进行减速和转弯机动,诱骗对手也进行转弯大机动,从而拉开与对手的距离,再之后采用的策略类似于图 7。

如图 7 所示,RL 无人机在 ATTACK 无人机前面先进行转弯机动,随即降低速度,由于 ATTACK 无人机按照纯追踪法以不小于 RL 无人机的恒定速

度向其方向移动，最终，ATTACK 无人机最终被锁定在 RL 无人机的前方的火力打击区域内。

如图 8 所示，RL 无人机通过调整速度和转向过载围绕 ATTACK 无人机轨迹两侧蜿蜒的方式进行机动，逐渐缩小与对手的距离，最后同样采用类似图 7 的策略，使 ATTACK 无人机飞行至自己的前方，被锁定在自己的火力打击范围内。

如图 9 所示，RL 无人机还可以学习到的策略是机会性的，不需要做过多的机动即可以在较短的时间步长内取得对抗的胜利而非依靠上述提及的策略。根据一些合适的初始条件，RL 无人机基本不调整转向过载地径直飞行，只是在前期需要采用类似于图 6 的策略调整速度防止前期被攻击。

如图 10 所示，还可以学到一种在规定时间步长内平局的策略，即 RL 无人机诱导 ATTACK 无人机一起做圆周运动，以此让 ATTACK 无人机的扇形火力区域无法攻击自己。

5　结论

本文针对无人机近距空战的自主机动反追击问题，提出经验回放区重构 TD3 算法。该方法将经验回放区重构为成功、失败两个经验回放区，取代传统的随机采样使用基于成功、失败经验区的采样策略。仿真结果表明，RL 无人机学到的策略在训练集上兼顾了训练前期的学习效率与训练后期的稳定收敛，在测试集上展示了较好的性能。

本文研究基于无人机的状态全局可观测这一假设条件，而真实空战环境下，受我机感知范围限制，敌机位置等态势信息并不能时刻被精确获取。针对不完全信息博弈条件进行空战决策研究，更具挑战性和实用性，将是本文下一步研究的重点。

参考文献

[1] 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.

[2] 黄长强，赵克新，韩邦杰，等. 一种近似动态规划的无人机机动决策方法[J]. 电子与信息学报，2018，40(10)：2447-2452.

[3] SUN W, TSIOTRAS P, LOLLA T, et al. Pursuit-Evasion Games in Dynamic Flow Fields via Reachability Set Analysis[C]//2017 American Control Conference (ACC). [S. l.]: IEEE, 2017: 4595-4600.

[4] VIRTANEN K, KARELAHTI J, RAIVIO T. Modeling Air Combat by a Moving Horizon Influence Diagram Game[J]. Journal of Guidance, Control, and Dynamics, 2006, 29(5): 1080-1091.

[5] CHANGQIANG H, KANGSHENG D, HANQIAO H, et al. Autonomous Air Combat Maneuver Decision Using Bayesian Inference and Moving Horizon Optimization[J]. Journal of Systems Engineering and Electronics, 2018, 29(1): 86-97.

[6] 国海峰，侯满义，张庆杰，等. 基于统计学原理的无人作战飞机鲁棒机动决策[J]. 兵工学报，2017，38(1)：160-167.

[7] 唐振韬，邵坤，赵冬斌，等. 深度强化学习进展：从 AlphaGo 到 AlphaGo Zero[J]，控制理论与应用，2017，34(12)：1529-1546.

[8] VINYALS O, EWALDS T, BARTUNOV S, et al., StarCraft II: A New Challenge for Reinforcement Learning[Z]. ArXiv: 1708.04782, 2017.

[9] VINYALS O, BABUSCHKIN I, CZARNECKI W M, et al., Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning [J]. Nature, 2019, 575(7782): 350-354.

[10] YE D, LIU Z, SUN M, et al., Mastering Complex Control in MOBA Games with Deep Reinforcement Learning[Z]. ArXiv: 1912.09729, 2019.

[11] ANDERSON L, SENAPATHY S. On Solving the 2-Dimensional Greedy Shooter Problem for UAVs[Z]. ArXiv Preprint ArXiv:1911.01419, 2019.

[12] VLAHOV B, SQUIRES E, STRICKLAND L, et al. On Developing a UAV Pursuit-Evasion Policy Using Reinforcement Learning[C]//2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). [S. l.]: IEEE, 2018: 859-864.

[13] WANG M, WANG L, YUE T. An Application of Continuous Deep Reinforcement Learning Approach to Pursuit-Evasion Differential Game[C]//3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). [S. l.]: IEEE, 2019: 1150-1156.

[14] MASSON W, RANCHOD P, KONIDARIS G D. Reinforcement Learning with Parameterized Actions[C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2016. [S. l.]: AAAI, 2016: 1934-1940.

[15] 张堃，李珂，时昊天，等. 基于深度强化学习的 UAV 航路自主引导机动控制决策算法[J]. 系统工程与电子技术，2020，42(7)：1567-1574.

[16] SHAW R L. Fighter Combat[R]. Annapolis, MD, USA: Naval Institute Press, 1985.

[17] LIM S H, FURUKAWA T, DISSANAYAKE G, et al. A Time-Optimal Control Strategy for Pursuit-Evasion Games Problems[C]//IEEE International Conference on Robotics and Automation, Proceedings. ICRA'04. [S. l.]: IEEE, 2004: 3962-3967.

（编辑：徐敏）