一种无人机集群对抗多耦合任务智能决策方法

文永明，石晓荣，黄雪梅，余 跃
（北京控制与电子技术研究所，北京 100038）

摘 要：针对复杂场景下无人机集群对抗中协同目标分配和突防轨迹规划等多耦合任务的决策问题，提出了一种集群对抗多耦合任务智能决策方法。首先，针对无人机集群对抗中耦合任务多和决策空间大难题，结合集中式和分层式架构的优点，设计了面向多耦合任务的混合式深度强化学习架构，可提升多耦合任务间的协同性和集群对抗效能；其次，针对轨迹规划序贯决策的稀疏奖励难题，设计了基于轨迹构造的一步式动作空间设计方法，可加快策略网络收敛速度；再次，针对强对抗条件下的场景不确定难题，基于无人机集群红蓝对抗仿真平台，设计了基于多随机场景的红蓝博弈训练方法，可增强策略网络的泛化性；最后，通过与传统方法、集中式架构方法和分层式架构方法进行对比，验证了此方法的有效性和先进性。

关键词：深度强化学习；智能决策；无人机集群对抗；协同目标分配；突防轨迹规划
中图分类号：V249.1 **文献标识码**：A **文章编号**：1000-1328(2021)04-0504-09
DOI：10.3873/j.issn.1000-1328.2021.04.011

An Intelligent Decision-Making Method for Multi-Coupling Tasks of UAV Cluster Countermeasure

WEN Yong-ming, SHI Xiao-rong, HUANG Xue-mei, YU Yue
(Beijing Institute of Control & Electronics Technology, Beijing 100038, China)

Abstract: Aiming at the decision-making problems of multi-coupling tasks such as cooperative target assignment and penetration trajectory planning in UAV cluster countermeasure in complex scenes, an intelligent decision-making method for multi-coupling tasks in UAV cluster countermeasure is proposed. Firstly, aiming at the problems of multi-coupling tasks and large decision-making space in UAV cluster countermeasure, combined with the advantages of centralized and hierarchical architectures, a hybrid deep reinforcement learning architecture for multi-coupling tasks is designed, which can improve the cooperation between the multi-coupling tasks and the effectiveness of cluster countermeasure. Secondly, for the sparse reward problem of sequential decision-making in trajectory planning, a trajectory construction method is designed. Thirdly, aiming at the scene uncertainty problem under the strong countermeasure conditions, based on the UAV cluster red blue countermeasure simulation platform, a red blue game training method based on multiple random scenes is designed, which can enhance the generalization of the strategy network. Finally, by comparing with the traditional method, the centralized architecture method and the hierarchical architecture method, the simulation results show that the effectiveness and the advanced nature of the proposed method are verified.

Key words: Deep reinforcement learning; Intelligent decision-making; UAV cluster countermeasure; Cooperative target assignment; Penetration trajectory planning

0 引 言

随着集群技术和人工智能的发展，基于群体智能的集群协同技术逐渐发展为未来智能化战争的发展方向[1-3]。无人机集群利用低成本、大规模和分布式的优势，协同侦查作战可以体现出显著的灵活性和智能性。无人机集群协同侦查在线决策主要包括协同目标分配和突防轨迹规划等多个相互耦合的

收稿日期：2021-02-13； 修回日期：2021-02-25

任务,无人机集群需要根据战场态势和作战任务决策出每架无人机的侦查目标和突防轨迹,以最大化集群对抗效能。随着对抗环境愈加复杂动态,对抗手段愈加多样智能,无人机集群对抗在线决策存在耦合任务多、决策空间大和场景不确定难题,导致传统基于专家知识和现代优化算法的决策方法难以同时满足在线决策的实时性、最优性和泛化性。

随着人工智能技术的发展与突破[4],尤其是深度强化学习在智能决策等方面得到了广泛关注与研究[5-7]。深度强化学习是深度学习和强化学习的有机结合,深度学习善于拟合,可通过深层神经网络表征复杂空间的非线性和泛化性,强化学习善于决策,可通过迭代学习使累计奖励最大化来获得高性能策略。深度强化学习基于大量离线训练得到智能策略网络,进行快速在线决策,可弥补无人机集群对抗程序化策略带来的局限性,提升应对复杂飞行环境和突发事件的适应能力[8]。

在深度强化学习架构方面,当决策问题由单个任务或少数简单任务构成时,通常采用集中式深度强化学习架构来解决。如图 1 所示,集中式架构的多耦合任务使用同一套策略网络、奖励函数和经验池,进行集中式耦合训练,在决策时一次同时输出各个任务的动作。集中式架构建模简单,并且在理论上可保证存在全局最优解。

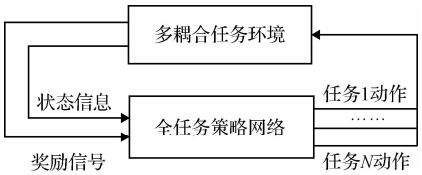

图 1　集中式架构

Fig. 1　Centralized architecture

文献[9]基于 DDPG 集中式架构优化一类变体飞行器外形,因其决策空间较小,故可以快速收敛到最优变外形策略。文献[10]采用 DQN(Deep Q-Network)算法对多个 Atari 小游戏(比如“乒乓球”、“打砖块”等)进行建模和训练,最终在多款游戏上的表现超越了人类玩家。然而,在“蒙特祖玛的复仇”这款游戏中,DQN 算法的胜率为 0%[11],其原因是这款游戏的任务较多且相互耦合(比如爬楼梯、躲避敌人、拿钥匙等),策略空间巨大,集中式架构在有限计算资源下难以收敛。为了解决多个耦合复杂任务所带来的决策空间爆炸等问题,分层式深度强化学习架构被提出[12]。如图 2 所示,分层式架构的多耦合任务使用多套对应的策略网络、奖励函数和经验池,按照任务间的逻辑关系进行分层单独训练,在决策时输出各自的动作进行组合来完成整个决策问题。分层式架构将多耦合任务进行解耦建模与分层单独训练,可以缩小整个决策问题的策略空间,使得各个任务的策略网络收敛速度加快。

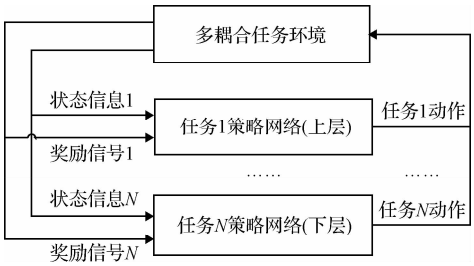

图 2　分层式架构

Fig. 2　Hierarchical architecture

文献[13]采用分层深度强化学习架构将“蒙特祖玛的复仇”抽象成多个不同层次的子任务进行建模,AI 可以完成游戏任务。文献[14]采用分层深度强化学习架构对一款篮球游戏建模,基于下层已熟练掌握的篮球技巧,智能体学到了上层的有效策略。文献[15]基于高斯过程回归与深度强化学习的分层人机协作控制方法,并以人机协作控制球杆系统为例检验该方法的高效性。然而,分层式架构的各个子任务的策略网络分离,即使各个子任务都收敛到各自的全局最优解,但是将它们组合后,得到的结果很可能不是整个任务的全局最优解。例如在无人机集群对抗中,目标分配结果是轨迹规划的输入,而轨迹规划性能是目标分配的依据,分层式架构将这两个子任务分开训练,没有充分考虑它们之间固有的耦合关系,因此多耦合任务间的协同性无法充分体现,集群对抗效能无法充分发挥。本文针对无人机集群对抗中耦合任务多和决策空间大难题,结合集中式和分层式架构的优点,设计了面向多耦合任务的混合式深度强化学习架构,通过构建多套相关联的多耦合任务分层策略网络进行集中耦合训练,可提升多耦合任务间的协同性和集群对抗效能。

在深度强化学习奖励函数设计方面,序贯动作导致的稀疏奖励问题是指在多步强化学习中,往往只在最后一步存在明确奖励,而中间过程的即时奖

励函数难以人为设计且存在主观性和经验性。例如无人机集群对抗的多步轨迹规划只在结束时才能得到是否被拦截或者侦查目标的结果,而中间过程很难根据当前的位置和速度等信息设计合适的即时奖励函数来引导突防和侦查目标。强化学习是求累计奖励期望最大时的最优策略,奖励函数不同将直接影响策略的性能,如果没有合适的即时奖励,稀疏奖励问题会导致策略网络难以快速且稳定收敛[16]。为了解决稀疏奖励问题,文献[17]提出逆向强化学习方法,即专家在完成某项任务时,其决策往往是最优或接近最优,可以假设,当所有的策略所产生的累积奖励期望都不比专家策略所产生的累积奖励期望大时,所对应的奖励函数就是根据示例学到的奖励函数。为了使逆向强化学习可以很好地扩展到具有大量状态的系统,将其与深度学习相结合,在神经网络中学习状态动作对的奖励,如基于最大边际法的深度逆向强化学习[18]和基于深度 Q 网络的深度学徒学习[19]等。然而,逆向强化学习和深度逆向强化学习都是从专家示例中学习奖励函数,在复杂场景下无人机集群对抗问题中难以获取足够的专家示例来支撑上述方法。本文针对轨迹规划序贯决策的稀疏奖励难题,设计了基于轨迹构造的一步式动作空间设计方法,回避了多步决策的中间过程,从而避免了稀疏奖励问题,可使策略网络稳定快速收敛。

在深度强化学习的泛化性研究方面,泛化性是指训练好的智能策略网络在未见过的场景中也具有一定的适应能力,其体现在深度神经网络对独立同分布数据强大的拟合和预测能力。因此,在深度强化学习训练过程中,使策略网络探索到尽可能大的决策空间,增加数据的多样性,是提升其泛化性的有效途径。2017 年,DeepMind 团队在《Nature》上推出了围棋人工智能 AlphaZero[20],AlphaZero 不需要人类专家知识,只使用纯粹的深度强化学习和蒙特卡洛树搜索,经过 3 天的自我博弈就以 100 比 0 的成绩完败了 AlphaGo,AlphaZero 强大的搜索能力和泛化性得益于海量且多样的自我博弈数据。文献[21]指出,AlphaZero 智能化方法框架可以启发人工智能在智能指挥决策等领域的应用。本文针对强对抗条件下的场景不确定难题,基于无人机集群红蓝对抗仿真平台,设计了基于多随机场景的红蓝博弈训练方法,通过随机变化对抗双方的初始位置和速度等,来设置每局的对抗态势,从而得到多样化的对抗训练数据;通过设计蓝方 AI,采用红蓝博弈的方式获得更加智能的蓝方策略作为红方 AI 的陪练,从而可以进一步提升红方 AI 的泛化性。

本文的主要创新点和贡献:1) 针对无人机集群对抗中耦合任务多和决策空间大难题,设计了面向多耦合任务的混合式深度强化学习架构,可提升多耦合任务间的协同性和集群对抗效能;2) 针对轨迹规划序贯决策的稀疏奖励难题,设计了基于轨迹构造的一步式动作空间设计方法,可加快策略网络收敛速度;3) 针对强对抗条件下的场景不确定难题,设计了基于多随机场景的红蓝博弈训练方法,可增强策略网络的泛化性。

1 混合式深度强化学习架构

混合式架构将集中式架构和分层式架构进行结合。多耦合任务使用多套与子任务对应的执行者 - 评估者(Actor-Critic, AC)神经网络与奖励函数分层构建网络,且多个经验池中的经验相互关联。在策略网络训练控制器的调度下,多个策略网络按照多任务间的分层关系进行集中耦合训练。在训练过程中,每个评估者(Critic)网络收集所有任务的状态和动作信息作为评价的输入,从而为策略更新提供准确且稳定的信号,更充分的状态和动作信息有助于提高耦合任务间的协同性;在策略执行过程中,各任务只需根据自己的状态和执行者(Actor)网络,进行决策控制,如图 3 所示。

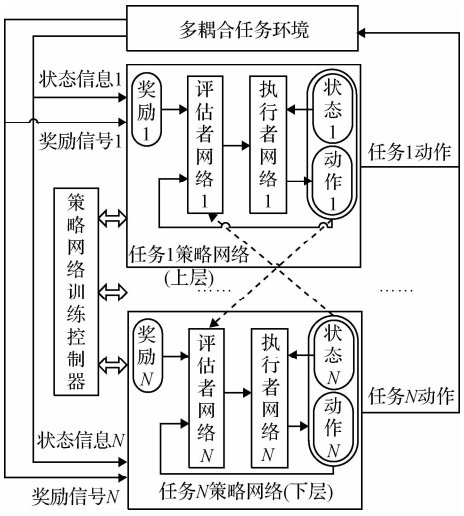

图 3 混合式架构

Fig. 3 Hybrid architecture

混合式架构保留了集中式和分层式架构的主要优点,又克服了它们的突出缺点,既保证了各个耦合任务之间相对稳定的训练环境,有利于得到多任务协同下的全局最优解,又使得策略空间规模可接受,有利于策略网络快速收敛。三种深度强化学习架构特点对比如表 1 所示。

混合式深度强化学习架构主要由多任务策略网络和策略网络训练控制器组成,多任务策略网络利用多套相关联的 AC 网络对子任务进行建模并分层,策略网络训练控制器按照多任务间的分层关系进行集中耦合训练。混合式架构的建模和训练流程如图 4 所示。

表 1　三种架构特点对比

Table 1　Comparison of three architectures

项目	集中式	分层式	混合式
网络构建	1 套网络	多套分层的 独立网络	多套分层的非 独立网络
网络训练	1 个经验池 集中训练	多个独立经验池, 多任务分层 迭代训练	多个相关经验池, 在训练控制器的 调度下集中训练
策略空间	大	子任务小	较小
收敛性能	慢且难	子任务快且易, 迭代训练慢	速度较快且较易收敛
策略性能	理论存在全 局最优解	易收敛到 局部最优	易得到全局最优解,提 升多任务间协同能力

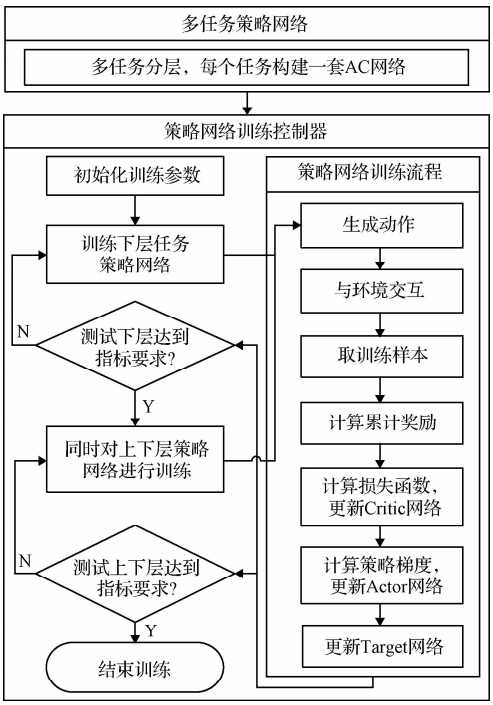

图 4　混合式架构建模与训练流程图

Fig. 4　Hybrid architecture modeling and training flow chart

1.1　多任务策略网络

多耦合任务 M 由 N 个子任务 m_i 组成,即 $M=\{m_i\}$ (i 表示子任务编号且 $i=1,2,\cdots,N$),根据多耦合任务之间的逻辑关系,将 N 个子任务进行分层。任务 m_i 基于 AC 架构构建执行者(Actor) 神经网络 A_i 和评估者(Critic) 神经网络 C_i 。任务 m_i 的状态空间为 s_i ,动作空间为 a_i ,奖励值为 r_i 。任务 m_i 的经验池设计为:

$$\boldsymbol{e}_i=\{s_1,s_2,\cdots,s_N,a_1,a_2,\cdots,a_N,s_1',s_2',\cdots,s_N',r_i,d_1,d_2,\cdots,d_N\} \tag{1}$$

式中: s_i'为任务 m_i 下一步的状态, d_i 为任务 m_i 结束标志,且当任务 m_i 结束时, $d_i=1$,反之, $d_i=0$ 。

任务 m_i 的评估者神经网络 $\boldsymbol{C}_i$ 的输入层为所有任务的状态 $\boldsymbol{S}=\{s_1,s_2,\cdots,s_N\}$ 和所有任务的动作 $\boldsymbol{A}=\{a_1,a_2,\cdots,a_N\}$, $\boldsymbol{C}_i$ 的输出层为 1 维的全局评估值。任务 m_i 的执行者神经网络 $\boldsymbol{A}_i$ 的输入层为任务 m_i 的状态 s_i , $\boldsymbol{A}_i$ 的输出层为任务 m_i 的动作 a_i 。

1.2　策略网络训练控制器

为了多耦合任务 M 的整个策略网络能够快速稳定收敛,下层任务需要给上层任务创造良好的学习环境基础,故策略网络训练控制器设计为先训练下层任务,达到设计指标后,再耦合训练上一层任务,即上下层集中训练。

策略网络训练控制器设计训练流程如下:

1) 初始化: 设置多任务策略网络和策略网络训练控制器参数;

2) 生成下层动作: 根据下层执行者神经网络 A_i 的策略生成动作:

$$a_i=A_i(s_i)+\delta_i \tag{2}$$

式中: $A_i(\cdot)$ 为任务 m_i 的执行者神经网络策略函数,输入为任务 m_i 的状态,输出为任务 m_i 的动作, θ_{Ai} 为执行者神经网络 A_i 的神经网络参数; $\delta_i\sim N(0,\sigma_i^2)$ 为服从均值为 0,标准差为探索贪婪率 σ_i 正态分布的随机数;

3) 生成上层动作: 上层任务随机生成动作:

$$a_i=\xi_i \tag{3}$$

式中: ξ_i 为服从均匀分布的随机数;

4) 与仿真环境交互: 将得到动作集合 $\boldsymbol{A}=\{a_1,a_2,\cdots,a_N\}$ 在仿真环境中执行,得到奖励值集合 $\boldsymbol{R}=\{r_1,r_2,\cdots,r_N\}$,下一个状态集合 $\boldsymbol{S}'=\{s_1',s_2',\cdots,s_N'\}$ 和任务是否结束标志集合 $\boldsymbol{D}=\{d_1,d_2,\cdots,d_N\}$;

5) 保存经验: 将经验 $\boldsymbol{e}_i=\{S,A,S',r_i,D\}=\{s_1,s_2,\cdots,s_N,a_1,a_2,\cdots,$

$$a_N, s_1', s_2', \cdots, s_N', r_i, d_1, d_2, \cdots, d_N\} \quad (4)$$

存入任务 m_i 的经验池 E_i；

6) 策略网络训练: 当任务 m_i 的经验池 E_i 总经验数达到开始训练的条件时，开始对任务 m_i 的策略网络进行训练:

(1) 取训练样本: 从任务 m_i 的经验池 E_i 中随机取出批处理规模 S_{B_i} 个经验 $\{S^j, A^j, S^{j}, r_i^j, D^j\}$，$j = 1, 2, \cdots, S_{B_i}$ 表示批处理经验中的序号;

(2) 定义累计奖励函数: 令任务 m_i 的累计奖励为:

$$y_i^j = r_i^j + (1 - d_i^j)\gamma_i \cdot C_i'(S^{j}, a_1', a_2', \cdots, a_N') \mid_{a_k' = A_k'(s_k^j)} \quad (5)$$

式中: d_i^j 为第 j 个经验中任务 m_i 的结束标志; γ_i 为任务 m_i 的累计奖励折扣因子; $C_i'(\cdot)$ 为任务 m_i 在神经网络参数为 θ_{Ci}' 下的目标(Target)评估者神经网络价值函数，输入为所有任务的下一时刻状态和动作，输出为任务 m_i 的下一时刻评估值; S^{j} 为第 j 个经验中所有任务的下一时刻状态; $a_k' = A_k'(s^{j}_k)$ 为任务 m_k 的下一时刻动作，$A_k'(\cdot)$ 为任务 m_k 在神经网络参数为 θ_{Ak}' 下的目标(Target)执行者神经网络策略函数，输入为任务 m_k 的下一时刻状态，输出为任务 m_k 的下一时刻动作，$k = 1, 2, \cdots, N$;

(3) 定义损失函数: 令任务 m_i 的损失函数为:

$$L(\theta_{Ci}) = \frac{1}{S_{B_i}} \sum_j [y_i^j - C_i'(S^j, a_1, a_2, \cdots, a_N)]^2 \quad (6)$$

式中: $C_i(\cdot)$ 为任务 m_i 在神经网络参数为 θ_{Ci} 下的评估者神经网络价值函数。通过求 $L(\theta_{Ci})$ 的极小值来更新 θ_{Ci};

(4) 定义采样策略梯度函数: 令任务 m_i 的采样策略梯度为:

$$\nabla_{\theta_{Ai}} J_i \approx \frac{1}{S_{B_i}} \sum_j \{ [\nabla_{\theta_{Ai}} A_i(s_i^j)] \cdot [\nabla_{a_i} C_i'(S^j, a_1^j, \cdots, a_i, \cdots, a_{N_m}^j) \mid_{a_i = A_i(s_i^j)}] \} \quad (7)$$

式中: $\nabla_{\theta_{Ai}} J$ 表示任务 m_i 的执行者代价函数 J_i 对执行者神经网络参数 θ_{Ai} 的梯度;

(5) 更新策略网络参数: 根据式(7)估计的策略梯度通过深度学习优化器来更新任务 m_i 的执行者神经网络参数 θ_{Ai};

(6) 更新目标网络参数: 满足一定条件时，按照式(8)来更新任务 m_i 的目标执行者神经网络参数 θ_{Ai}' 和目标评估者神经网络参数 θ_{Ci}':

$$\begin{cases} \theta_{Ai}' \leftarrow \tau_i \theta_{Ai} + (1 - \tau_i)\theta_{Ai}' \\ \theta_{Ci}' \leftarrow \tau_i \theta_{Ci} + (1 - \tau_i)\theta_{Ci}' \end{cases} \quad (8)$$

式中: τ_i 为神经网络参数更新频率，“←”表示赋值。

(7) 测试与训练层级递进: 训练一定次数后，测试当前层对应的所有任务是否都达到设计指标，如果是，则开始上一层任务的训练; 否则，继续本层任务的训练;

(8) 循环: 重复流程(1)至流程(8)，直至多耦合任务 M 训练结束，且测试达到预定指标。

1.3 无人机集群对抗混合式架构建模

1) 多任务策略网络。上层: 协同目标分配，决策红方无人机集群中每架无人机的侦查目标，以最大化集群对抗效能(侦查总得分); 下层: 突防轨迹规划，决策红方无人机的突防和侦查轨迹，既要进行躲避机动又要保留足够的机动能力对目标进行侦查，以最大化突防概率(突防成功的红方无人机数量除以红方无人机总数量)和侦查成功率(侦查成功的红方无人机数量除以红方无人机总数量)。

2) 状态空间。目标分配策略网络的状态主要包括: 红方无人机数量、位置、速度和蓝方待侦查目标数量、位置、价值等; 轨迹规划策略网络的状态主要包括: 红方无人机位置、速度和蓝方待侦查目标位置等。

3) 动作空间。目标分配策略网络的动作为: 红方无人机侦查目标的编号; 轨迹规划策略网络的动作为: 红方无人机轨迹构造函数的参数。

4) 奖励函数。确定 3 个元奖励分别为突防元奖励 r_{o_tf}、侦查元奖励 r_{o_zc} 和效能元奖励 r_{o_xn}。红方无人机突防成功，则 $r_{o\_tf} = 1$，否则 $r_{o\_tf} = -1$; 红方无人机成功侦查目标，则 $r_{o\_zc} = 1$，否则 $r_{o\_zc} = -1$; 集群对抗效能归一化作为效能元奖励 r_{o_xn}。为了进一步体现各个耦合任务之间的协同性，采用元奖励加权的方式使目标分配和轨迹规划的奖励函数相互关联。根据目标分配对各个元奖励的影响确定目标分配的突防权重 $w_{o\_tf\_mb}$、侦查权重 $w_{o\_zc\_mb}$ 和效能权重 $w_{o\_xn\_mb}$，且满足 $w_{o\_tf\_mb} + w_{o\_zc\_mb} + w_{o\_xn\_mb} = 1$。同理，根据轨迹规划对各个元奖励的影响确定轨迹规划的突防权重 $w_{o\_tf\_gj}$、侦查权重 $w_{o\_zc\_gj}$ 和效能权重 $w_{o\_xn\_gj}$，且满足 $w_{o\_tf\_gj} + w_{o\_zc\_gj} + w_{o\_xn\_gj} = 1$。则目标分配奖励函数为:

$$r_{mb} = w_{o\_tf\_mb} r_{o\_tf} + w_{o\_zc\_mb} r_{o\_zc} + w_{o\_xn\_mb} r_{o\_xn} \quad (9)$$

轨迹规划的奖励函数为:

$$r_{gj} = w_{o\_tf\_gj} r_{o\_tf} + w_{o\_zc\_gj} r_{o\_zc} + w_{o\_xn\_gj} r_{o\_xn} \quad (10)$$

5) 策略网络训练控制器。先训练下层轨迹规划策略网络。当突防概率和侦查成功率达到指标要求后,再训练上层目标分配策略网络,两个任务进行集中耦合训练,直至突防概率、侦查成功率和集群对抗效能达到指标要求后,训练完毕。

2　基于轨迹构造的一步式动作空间设计方法

在突防轨迹规划中,红方无人机通过在线生成机动指令来达到躲避拦截和侦查目标的目的。通常采用多步序贯决策方式会带来稀疏奖励问题,它是指在每个决策周期都生成无人机的机动指令,但只在最后一步存在明确的奖励,而过程奖励难以设计,会导致策略网络难以快速稳定收敛。针对上述问题,设计了基于轨迹构造的一步式动作空间设计方法。

根据红方无人机机动特性和蓝方拦截无人机的拦截特点确定突防轨迹构造函数表示为:

$$n_c(t) = F(P,t) + a_0(t) \tag{11}$$

式中: $n_c(t)$ 表示 t 时刻无人机的机动指令。$a_0(t)$ 表示 t 时刻无人机的比例导引指令,引导无人机飞向目标。$F(P,t)$ 表示 t 时刻无人机的附加机动指令函数,控制机动突防, P 为函数参数集合。$F(P,t)$ 的具体表达形式可以根据无人机的机动特性和拦截无人机的拦截特点确定,比如无人机的动态性能良好且蓝方拦截策略简单, $F(P,t)$ 可确定为方波函数; 无人机的动态性能一般且蓝方拦截策略简单, $F(P,t)$ 可确定为正弦函数; 蓝方拦截策略复杂, $F(P,t)$ 可确定为多项式函数。

从函数参数集合 $\boldsymbol{P}$ 中确定待优化的参数,表示为:

$$\boldsymbol{P} = \boldsymbol{C} \cup \boldsymbol{X} \tag{12}$$

式中: $\boldsymbol{C} = \{c_1, c_2, \cdots, c_m\}$ 表示 m 个常值参数集合, $\boldsymbol{X} = \{x_1, x_2, \cdots, x_n\}$ 表示 n 个待优化参数集合。

确定深度强化学习的动作空间表示为:

$$\boldsymbol{A} = [x_1, x_2, \cdots, x_n]^{\mathrm{T}} (x_{i_{\min}} \leqslant x_i \leqslant x_{i_{\max}}, i = 1, 2, \cdots, n) \tag{13}$$

式中: $x_{i_{\min}}$ 表示待优化参数 x_i 的最小值, $x_{i_{\max}}$ 表示待优化参数 x_i 的最大值。

基于轨迹构造的一步式动作空间设计方法只需决策一次突防轨迹构造函数的参数就可以规划出完整的轨迹,对抗仿真后即可得到一次明确的奖励,即一个动作对应一个奖励,因此避免了序贯动作的稀疏奖励问题,使收敛速度和稳定性有效提升。

3　基于多随机场景的红蓝博弈训练方法

针对强对抗条件下的场景不确定难题,基于无人机集群红蓝对抗仿真平台,设计基于多随机场景的红蓝博弈训练方法。

红方无人机集群的作战任务为最大化侦查覆盖蓝方目标编队,红方无人机在飞行过程中会受到蓝方拦截无人机的拦截,在红方无人机突防后,需要飞到待侦查目标附近且保留一定的机动能力进行侦查。如图 5 所示,无人机集群红蓝对抗的主要场景及设计要素如下: 1) 红方侦查无人机集群: 由 N_H 架侦查无人机组成; 2) 蓝方待侦查目标编队: 由 N_L 个待侦查目标组成,五角星表示主要待侦查目标(需要 3 架红方无人机侦查保证覆盖目标) ,三角形表示次要目标(需要 2 架红方无人机侦查保证覆盖目标) ; 3) 蓝方拦截无人机: 针对 1 架红方无人机最多可用 2 架蓝方无人机进行拦截; 4) 集群对抗效能: 1 架红方无人机成功侦查目标得 1 分,成功侦查主要目标最多得 3 分,成功侦查次要目标最多得 2 分,所得总分即为集群对抗效能; 5) 集群对抗效能比: 为了对比不同想定之间的效能,定义集群对抗效能比为集群对抗效能除以理论最大效能。想定的名称用 "N_HV N_L" 表示。

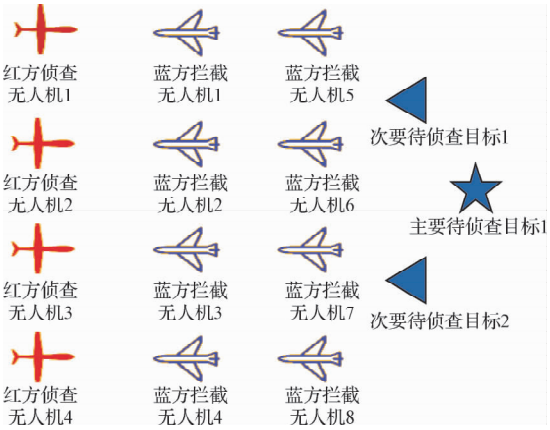

图 5　典型对抗场景示意图

Fig. 5　Typical confrontation scenarios

设置多个典型无人机集群对抗想定(如 8V5、8V7、12V10、18V12、18V14 等) 训练策略网络,设定红蓝对抗双方的初始位置和速度等参数的合理变化范围,每一局对抗训练随机选取一个想定和一组参数来设置对抗态势,则通过大量对抗仿真可得到多样化的对抗训练数据。

蓝方的对抗模型和策略通常采用基于专家知识的方式进行建模，然后进行红蓝对抗仿真对红方策略网络进行单方面训练，而基于蓝方单一策略对红方策略网络进行训练容易过拟合，导致红方策略单一且对蓝方策略的变化缺乏泛化性，难以适应高动态的实际战场环境。

设计蓝方策略网络，智能决策蓝方拦截无人机的拦截目标和起飞时机，红蓝策略网络在无人机集群红蓝对抗仿真平台上采用红蓝博弈方式进行训练。红蓝博弈训练方法流程如图 6 所示，在每个并行的博弈环境中，红蓝策略网络视对方为环境进行学习。为增强博弈训练中策略学习的稳定性，在每个博弈周期的训练中，固定红蓝双方中一方的策略，训练另一方。在每一个博弈周期结束后，根据红蓝方策略的表现进行优胜劣汰，将实力相当的红蓝策略网络配对，进行下一周期的博弈，如此往复，不断提升红方策略网络对不同蓝方策略的泛化性。

多平台分布式红蓝博弈训练场景如图 7 所示。

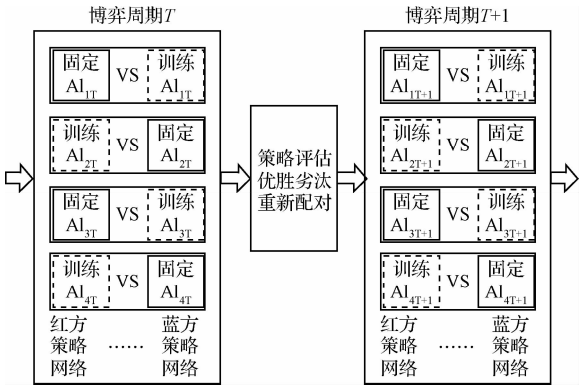

图 6 红蓝博弈训练流程

Fig. 6 Red blue game training process

图 7 多平台分布式红蓝博弈训练场景

Fig. 7 Multi platform distributed red blue game training scenario

4 仿真校验

4.1 有效性校验

采用基于多随机场景的红蓝博弈训练方法对红方和蓝方策略网络进行训练，得到最优的红方策略网络(红 AI)，以 18 架无人机集群侦查 14 个蓝方目标编队(18V14)为例来阐述仿真与测试结果。红方按照遗传算法决策，得到的典型红蓝对抗平面轨迹如图 8(a)所示；红方按照策略网络决策，得到的典型红蓝对抗平面轨迹如图 8(b)所示。

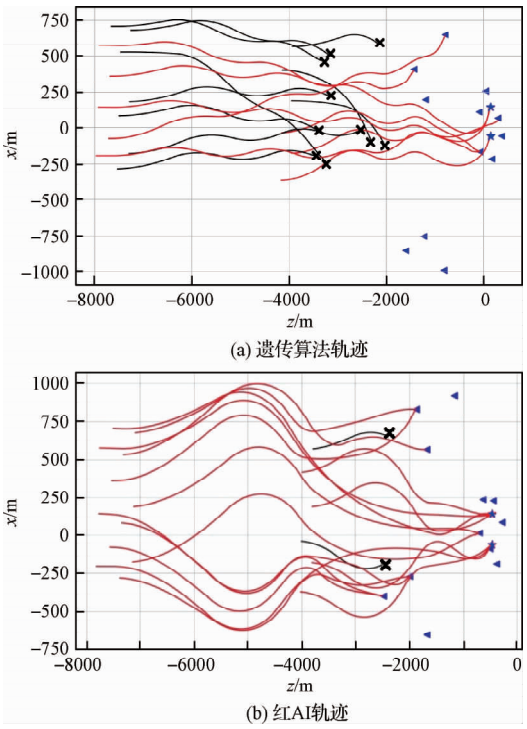

图 8 典型平面轨迹

Fig. 8 Typical plane trajectory

图 8 中，轨迹末端“×”表示红方无人机被蓝方无人机拦截或机动能力不足导致侦查失败。由图 8 可得遗传算法的突防概率为 8 ÷ 18 = 44%，集群对抗效能比为 7 ÷ 18 = 39%；红 AI 的突防概率为 16 ÷ 18 = 89%，集群对抗效能比为 15 ÷ 18 = 83%。通过对比可知：红 AI 可以为红方无人机集群分配合理的侦查目标和规划有效的突防和成功侦查目标轨迹，有效提高了集群对抗效能。

红 AI 训练过程曲线如图 9 所示。

图 9 中的训练曲线为单平台训练过程，仿真次数为 200 时，红 AI 收敛。采用 60 个无人机集群红蓝对抗仿真平台进行多平台分布式红蓝博弈训练，因此红 AI 的训练收敛次数约为 200 × 60 = 12000 次。

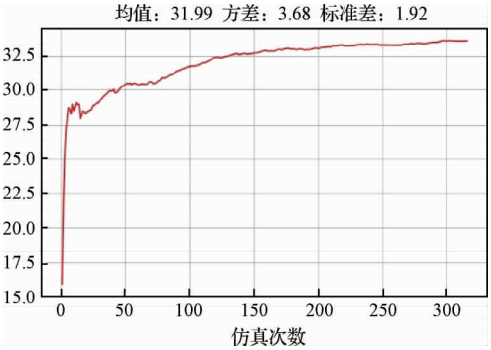

图 9　集群对抗效能训练曲线

Fig. 9　Efficiency training curve of group confrontation

遗传算法和红 AI 测试得到的性能对比如表 2 所示。由表 2 可得,红 AI 相比基于遗传算法在集群对抗效能上提升了约 95%,说明了本文方法的有效性。

表 2　遗传算法与红 AI 性能对比

Table 2　Performance comparison between genetic algorithm and red AI

性能指标	遗传算法	红 AI
突防概率/%	55	80
侦查成功率/%	51	78
集群对抗效能比/%	41	80

4.2　泛化性校验

通过对 12V10、18V14 等场景进行随机训练,得到的策略网络在未训练过的场景上(13V10、17V15)进行泛化性测试,得到的结果如表 3 所示。由表 3 可得,策略网络在未训练过场景上的适应性平均大于 90%,说明红 AI 具有一定的泛化性。

表 3　泛化性测试

Table 3　Generalization testing

性能指标	训练场景		未训练场景泛化	
	12V10	18V14	13V10	17V15
突防概率/%	89	80	88	80
侦查成功率/%	85	78	85	76
集群对抗效能比/%	87	80	87	79

4.3　先进性校验

将集中式架构训练得到的集中式 AI、分层式架构训练得到的分层式 AI 分别在无人机集群红蓝对抗仿真平台测试,得到的性能对比结果如表 4 所示。

从表 4 中可以得到以下结论: 1) 集中式 AI 在有限计算资源条件下难以收敛; 2) 分层式AI多任务迭代训练耗时大,且未得到全任务最优策略; 3) 混合式 AI 学到了多耦合任务间的协同能力,得到了全任务最优策略,相比分层式 AI 在集群对抗效能上提升了约 31%; 混合式 AI 策略网络收敛速度较快,相比分层式 AI 收敛速度提升 567%。上述结果表明: 在多耦合任务决策问题上,混合式深度强化学习架构相比集中式和分层式架构,具有较强的先进性。

表 4　三种架构性能对比

Table 4　Performance comparison of three architectures

性能指标	集中式 AI	分层式 AI	混合式 AI
突防概率/%	31	62	80
侦查成功率/%	30	60	78
集群对抗效能比/%	30	61	80
收敛次数	未收敛	80000	12000

5　结　论

本文针对复杂场景下无人机集群对抗中协同目标分配和突防轨迹规划等多耦合任务的决策问题,提出了一种集群对抗多耦合任务智能决策方法。设计了面向多耦合任务的混合式深度强化学习架构、基于轨迹构造的一步式动作空间设计方法和基于多随机场景的红蓝博弈训练方法,解决了无人机集群对抗在线决策耦合任务多、决策空间大和场景不确定等难题,增强了策略网络的收敛性能和泛化性,提升了无人机集群对抗多耦合任务间的协同性、集群对抗效能。通过与传统方法、集中式架构方法和分层式架构方法进行对比,验证了本文提出方法的有效性和先进性。

参　考　文　献

[1] 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 Xing-xing, Feng Yang-he, Ma Yang, et al. Deep multi-agent reinforcement learning: A survey [J]. Acta Automatica Sinica, 2020, 46(12): 2537 – 2557.]

[2] 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 Chang-yin, Mu Chao-xu. Important scientific problems of multi-agent deep reinforcement learning [J]. Acta Automatica Sinica, 2020, 46(7): 1301 – 1312.]

[3] 冉惟之. 基于群体智能的无人机集群协同对抗协同的设计与实现[D]. 电子科技大学, 2020. [Ran Wei-zhi. Design and implementation of cooperative adversarial system based on swarm intelligence for unmanned aerial vehicles [D]. University of Electronic Science and Technology of China, 2020.]

[4] 赵冬斌,邵坤,朱圆恒,等. 深度强化学习综述:兼论计算机围棋的发展[J]. 控制理论与应用, 2016, 36(6): 701 - 717. [Zhao Dong-bin, Shao Kun, Zhu Yuan-heng, et al. Review of deep reinforcement learning and discussions on the development of computer Go[J]. Control Theory & Applications, 2016, 36(6): 701 - 717.]

[5] Van H H, Guez A, Silver D. Deep reinforcement learning with double Q-learning [C]. Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, Arizona, USA, February 12 - 17, 2016: 1813 - 1819.

[6] Cuccu G, Luciw M, Schmidhuber J, et al. Intrinsically motivated neuroevolution for vision-based reinforcement learning [C]. Proceedings of the IEEE International Conference on Development and Learning, Frankfurt am Main, Germany, August 24 - 27, 2011.

[7] Narasimhan K, Kulkarni T, Barzilay R. Language understanding for text-based games using deep reinforcement learning [C]. Proceedings of the Conference on Empirical Methods for Natural Language Processing, Lisbon, Portugal, September 11 - 13, 2015.

[8] 马卫华. 导弹/火箭制导、导航与控制技术发展与展望[J]. 宇航学报, 2020, 41(7): 860 - 867. [Ma Wei-hua. Review and prospect of missile/launch vehicle guidance, navigation and control technologies[J]. Journal of Astronautics, 2020, 41(7): 860 - 867]

[9] 温暖,刘正华,祝令谱,等. 深度强化学习在变体飞行器自主外形优化中的应用[J]. 宇航学报, 2017, 38(11): 1153 - 1159. [Wen Nuan, Liu Zheng-hua, Zhu Ling-pu, et al. Deep reinforcement learning and its application on autonomous shape optimization for morphing aircrafts[J]. Journal of Astronautics, 2017,38(11): 1153 - 1159.]

[10] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning[J]. Computer ence, 2013.

[11] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518(7540): 529 - 533.

[12] 刘全,翟建伟,章宗长,等. 深度强化学习综述[J]. 计算机学报, 2018, 41(1): 1 - 28. [Liu Quan, Zhai Jian-Wei, Zhang Zong-chang, et al. A survey on deep reinforcement learning[J]. Chinese Journal of Computers, 2018, 41(1): 1 - 28.]

[13] Kulkarni T, Narasimhan K, Saeedi A, et al. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation [C]. Advances in neural information processing systems, Barcelona, Spain, December 5 - 10, 2016.

[14] Tang H Y, Hao J Y, Lv T J, et al. Hierarchical deep multiagent reinforcement learning with temporal abstraction[C]. CoRR abs/1809.09332, 2018.

[15] 金哲豪,刘安东,俞立. 基于GPR和深度强化学习的分层人机协作控制[J]. 自动化学报, 2020, 46(x): 1 - 11. [Jin Zhe-hao, Liu An-dong, Yu Li. Hierarchical human-robot cooperative control based on GPR and DRL[J]. Acta Automatica Sinica, 2020, 46: 1 - 11.]

[16] 陈希亮,曹雷,何明,等. 深度逆向强化学习研究综述[J]. 计算机工程与应用, 2018, 54(5): 24 - 35. [Chen Xi-liang, Cao Lei, He Ming, et al. Overview of deep inverse reinforcement learning[J]. Computer Engineering and Applications, 2018, 54(5): 24 - 35.]

[17] Ng A Y, Russell S J. Algorithms for inverse reinforcement learning[C]. Seventeenth International Conference on Machine Learning, Stanford University, Stanford, CA, USA, June 29 - July 2, 2000.

[18] Shi Q, Cheng L, Wang L, et al. Human action segmentation and recognition using discriminative semi-markov models [J]. International Journal of Computer Vision, 2011, 93(1): 22 - 32.

[19] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]. The 26th Annual Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, USA, December 3 - 6, 2012.

[20] Sliver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge[J]. Nature, 2017, 550(Oct. 19 TN. 7676): 354 - 359.

[21] 唐川,淘业荣,麻曰亮. AlphaZero原理与启示[J]. 航空兵器, 2020, 27(3): 27 - 36. [Tang Chuan, Tao Ye-rong, Ma Yue-liang. Principle and enlightenment of alphaZero[J]. Aero Weaponry, 2020, 27(3): 27 - 36.]

作者简介:

文永明(1988-),男,博士,主要从事智能协同控制方面的研究。

通信地址:北京西城区木樨地北里甲51号(100038)

电话:(010) 63301514

E-mail: wenyongming_buaa@foxmail.com