DOI: 10.12382/bgxb.2022.0711

基于分层强化学习的无人机空战多维决策

张建东[1], 王鼎涵[1], 杨啟明[1*], 史国庆[1], 陆屹[2], 张耀中[1]
(1. 西北工业大学 电子信息学院, 陕西 西安 710072; 2. 沈阳飞机设计研究所, 辽宁 沈阳 110035)

摘要: 针对无人机空战过程中面临的智能决策问题,基于分层强化学习架构建立无人机智能空战的多维决策模型。将空战自主决策由单一维度的机动决策扩展到雷达开关、主动干扰、队形转换、目标探测、目标追踪、干扰规避、武器选择等多个维度,实现空战主要环节的自主决策;为解决维度扩展后决策模型状态空间复杂度、学习效率低的问题,结合 Soft Actor-Critic 算法和专家经验训练和建立元策略组,并改进传统的 Option-Critic 算法,设计优化策略终止函数,提高策略的切换的灵活性,实现空战中多个维度决策的无缝切换。实验结果表明,该模型在无人机空战全流程的多维度决策问题中具有较好的对抗效果,能够控制智能体根据不同的战场态势灵活切换干扰、搜索、打击、规避等策略,达到提升传统算法性能和提高解决复杂决策效率的目的。

关键词: 无人机空战; 多维决策; 分层强化学习; Soft Actor-Critic 算法; Option-Critic 算法
中图分类号: V279 **文献标志码**: A **文章编号**: 1000-1093(2023)06-1547-17

Multi-Dimensional Decision-Making for UAV Air Combat Based on Hierarchical Reinforcement Learning

ZHANG Jiandong[1], WANG Dinghan[1], YANG Qiming[1*], SHI Guoqing[1], LU Yi[2], ZHANG Yaozhong[1]
(1. School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China;
2. AVIC Shenyang Aircraft Design and Research Institute, Shenyang 110035, Liaoning, China)

Abstract: To solve the intelligent decision-making problem in the process of UAV air combat, a multi-dimensional decision-making model for UAV intelligent air combat based on the hierarchical reinforcement learning architecture is established, allowing the autonomous decision-making of air combat to be extended from a single-dimensional maneuver decision to a multi-dimensional one including radar switch, active jamming, formation conversion, target detection, target tracking, interference avoidance, weapon selection, etc., so that autonomous decision-making in the main steps of air combat is realized. In order to solve the problems of state-space complexity and low learning efficiency of the decision-making model after the dimension expansion, a meta-strategy group is trained and established with the Soft Actor-Critic algorithm and expert experience, and the traditional Option-Critic algorithm is improved. The strategy termination function is designed and optimized to improve the flexibility of strategy switching and realize seamless multi-dimensional decision-making switching in air combat.. The experimental results show that the proposed method has good countermeasure effectiveness for the multi-dimensional decision-making during the whole process of UAV air combat, which can control the agent to flexibly switch among interference, search, strike, and avoidance strategies according to different battlefield situations with the purpose of improving the performance of traditional algorithms and the efficiency of solving complex

收稿日期: 2022-08-13
基金项目: 陕西省自然科学基础研究计划项目(2022JQ-593);陕西省科技厅重点研发计划项目(2022GY-089)
* **通信作者邮箱**: yangqm@nwpu.edu.cn

decision-making processes.

Keywords: UAV air combat; multi-dimensional decision-making; hierarchical reinforcement learning; Soft Actor-Critic algorithm; Option-Critic algorithm

0 引言

现代空战以决策速度快、机动性能高、态势感知能力强、高鲁棒性等特点为核心,然而有人机受人类生理极限限制,无法发挥出战斗机的极限性能。无人机摆脱了人类生理极限,但机动控制由地面指挥,决策速度慢,若大幅延长观察、判断、决策、行动(OODA)环的时间则容易错失战机,因此智能化无人机空战自主决策成为当今的研究热点。

随着 OODA 3.0 概念的提出[1]以及人工智能技术的不断发展,无人机在机动决策等单一维度的决策方面已经实现了一定程度的自主化,并且在某些方面已经达到或者超越了人类飞行员的水平。然而,空战过程是一个复杂的多维决策过程,要完成空战的自主化决策,必须要实现多个维度的协同自主化决策。因此无人机多维空战决策一直是该领域亟需攻克的难关,其对实现完全无人化空战的终极目标至关重要。

当前对无人机自主决策的诸多研究都集中在机动决策方面,通过深度 Q 网络(DQN)[2]、深度确定性策略梯度(DDPG)[3-5]、Actor-Critic 等深度强化学习算法来实现对无人机的机动控制。但这些方法有着超参数敏感、策略选择单一、无法解决多维决策问题[6]等缺点,无法很好地满足无人机空战对于快速收敛、高鲁棒性及多维决策的要求。事实上,空战决策除了机动决策外,还包括传感器决策、武器决策、干扰决策等各方面多维度的决策。相比而言,分层强化学习凭借着其能够进行空间分解和分层训练的优势,有望使无人机具备充足的策略,从而完成复杂的作战任务。

目前,已经有很多学者使用分层强化学习方法对无人机多维决策的相关问题进行了探索性研究。王俊敏等[7]在空战编队协同上应用了分层策略,但关键的观测数据并未给出,无法进行有效训练。付跃文等[8]应用了分层优化方法解决了无人机之间协作任务规划模块设计,证实了空战决策空间建模的可行性。文永明等[9]研究了一种无人机机群对抗多耦合任务智能决策方法,采用分层强化策略训练方法,提出混合式深度强化学习架构,完成了无人机突防侦察任务及目标的协同分配任务,证实了分层架构的有效性。程先峰等[10]采用一种基于 MAXQ 的 Multi-agent 分层强化学习的无人机协调方法,增强了无人机在混合运行复杂环境下适应环境和自协调的能力。吴宜珈等[11]提出基于选项的近端策略分层优化算法,用来解决近端策略优化算法在空战智能决策过程中面临的动作空间过大、难以收敛的问题。通过对相关文献的分析可以看出,目前在无人机多维决策方面的研究还不够完善,所研究问题的规模都比较小,决策维度与现实差距较大,导致其应用环境过于简单。

与此同时,以美国为代表的军事强国正在紧锣密鼓地开展将人工智能技术应用于无人机复杂作战任务的相关实验验证。2021 年美国洛克希德·马丁公司于美国国防部高级研究计划局(DARPA)举办的 Alpha 狗斗(ADT)比赛中展示了其最新研发的分层强化学习算法适应性新颖策略生成的操作层级结构(PHANG-MAN[12]),成功地将分层强化学习方法应用到无人机空战决策中,实现了多维空战决策中的追击决策、规避决策、打击决策。该算法在 ADT 决赛中斩获第二,并击败了美国空军 F-16 武器教练课的毕业生。该算法充分体现了分层强化学习在解决多维空战决策问题中的策略模块化、智能化、去中心化的特点,这一实验结果表明美军在无人机多维决策方面已经达到了很高水平。此外,其他相关研究[13-21]均表明深度强化学习在空战中的理论可行性。因此,进行无人机多维自主决策的应用研究具有一定的理论意义和使用价值。

本文以无人机一对一(1v1)、集群四对四(4v4)的红蓝空战对抗任务为场景,基于分层强化学习的架构建立无人机智能空战的多维决策模型,采用 Soft Actor-Critic 算法训练底层单元策略,并结合专家经验建立元策略组,扩展了决策的维度。改进传统的 Option-Critic 算法,设计优化了策略终止函数,提高了策略切换的灵活性,实现了空战中多个维度决策的无缝切换。

为了较好地完成目标打击任务,设计雷达开关、主动干扰、队形转换、目标探测、目标追踪、干扰规避、武器选择与目标打击共 7 种元策略。以贪心算法作为顶层元策略选择策略,完成智能多维空战自主决策。仿真实验结果表明,训练完成后的无人机

可以灵活地完成元策略的切换调用,能够以丰富的元策略组合完成更高层次的作战决策,体现了分层强化学习算法在提升无人机自主决策维度上的应用潜力。

1　空战决策维度分解

根据空战 OODA 环的概念,第 1 步需要确定目标方位。本文设定双方雷达探测能力一致,为实现先敌发现,需要构建高效的搜索方法。

贯穿整个空战过程的雷达探测至关重要,它有着确定目标精确方位、攻击引导的作用。在打击前,应确保目标不丢失,因此需要我机雷达能够持续照射目标,同时规避目标的电磁干扰。

在目标探测过程中,被动雷达能够在电磁静默情况下确定目标方位。然而单架飞机的被动探测仅能确定目标方向,无法精确确定目标的坐标。若要完成精确探测,则需要至少两架飞机协同探测。

为降低因雷达开机暴露位置的风险,需要对雷达资源做合理的分配。在编队内,对于距离较近、航向差较小的我机,仅需开启其中一个雷达,因此需要给出合理分配雷达资源的数学模型和规则模型。

在打击目标前,需要判断目标的距离以及自身剩余的导弹数量和种类以选择合适的导弹类型。打击目标时,应该确保我机安全,采用合理的干扰策略,避免暴露位置。

在多机作战过程中,编队往往能够最大化作战能力,最小化作战损耗。常用的编队模型为长机-僚机编队。作战伊始通过合理的编队布局增强战力,作战过程中遇到队形破坏可以采用队形转变策略重组编队,维持整个作战过程中的战力。

综上所述,整个空战流程涵盖了雷达开关、主动干扰、队形转换、目标探测、武器选择、目标打击、目标追踪、干扰规避策略,空战中的主要决策环节如图 1 所示。

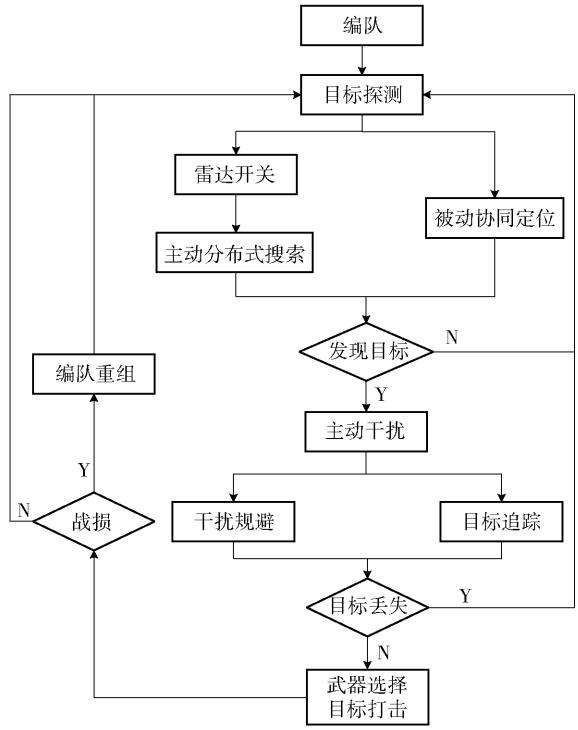


图 1　空战全流程分析

Fig. 1　Analysis of the whole air combat process

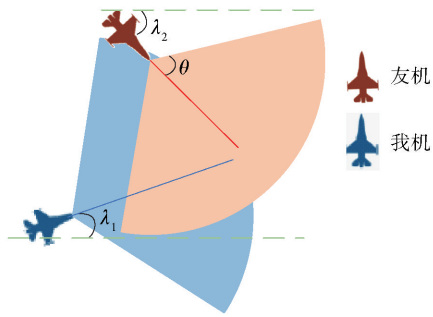


图 2　雷达探测重叠区域分析

Fig. 2　Overlapping area analysis of radar detection

1.1　雷达开关策略模型

为了降低因雷达开机暴露位置的风险,飞机往往会在非必要时刻关闭雷达,处于电磁静默状态。本文构建了雷达开关模型,分析探测重叠区域,给出了雷达开关判定规则。

为避免探测资源浪费,并降低暴露位置风险,分析了雷达探测重叠区域,如图 2 所示。图 2 中,θ 表示雷达的探测半角,λ_1 和 λ_2 分别表示两架飞机的航向角。

设 d 表示两机的间距,则无人机进入判决区域的条件如式(1)所示:

$$\begin{cases} d \leqslant r\sin\theta \\ \lambda_1 - \lambda_2 \leqslant \theta \end{cases} \tag{1}$$

式中:r 为雷达探测距离。

式(1)表述了两机间距及两机航向角度差值小于阈值时,两机处于判决区域,需要关闭其中一架飞机的雷达。

设定判决状态变量 p,如果满足判决公式,则判决变量 p 置为 1,否则置为 0,具体的判定规则如下:

1)若 $p=1$,则关联判决友机编号(id)为 i_p,本机 id 为 i_m,根据全局判定列表[$(p,i_p,i_m)\cdots$],观察是否存在重复 i_p,若存在则不开启 $i_m=i_p$ 飞机的雷达,开启 $i_p \neq i_m$ 飞机的雷达。否则开启长机雷达。

2)所有不在全局列表中的无人机全部开启雷达。

模型输入为我机的坐标、航向、雷达开关状态,输出为雷达的开机频点,0 表示关机,非 0 表示开机相应频点。

1.2 主动干扰策略模型

为了实现瞄准式干扰,本文构建了主动干扰模型,分析了干扰区域,给出了干扰规则。

实施干扰前,我机需要确定被干扰目标的雷达频点,记为 r_t。若目标处于我机主动雷达的照射范围内且不受目标干扰时,则我机可以获取到目标雷达开机频点的观测信息。此时仅需将我机的干扰频点 r_j 设置为目标的雷达频点即可完成瞄准式干扰,即满足:

$$r_t = r_j \tag{2}$$

模型的输入为目标的开机频点,未探测到时奖励记为 0,探测到 n 个目标干扰频点,奖励记为 n。输出为我机的开机频点。

1.3 队形转换策略模型

为了提高协同效能,构建队形转换模型,建立长机-僚机编队模型,考虑到作战过程中被破坏的情况给出了编队重组方案。

初始时刻我方编队为两两一组,以长机-僚机形式编队,长机执行搜索-攻击任务,僚机进行探测干扰任务,掩护长机。若长机被击毁,僚机将接替长机位置完成攻击与目标探测等任务。长机 id 记为 id_1,僚机 id 记为 id_f。构建编队列表与全局编队 $\{[id_1, id_f]\cdots\}$,若作战过程中因战损导致编队结构被破坏,则可以通过判断编队列表进行编队重组。例如,编队 1 长机被击毁,记 $[-id_1, id_f]$。若整队成员全部被击毁,则将该编队列表移出全局编队。

编队重组通过遍历所有编队,根据编队列表中是否存在负值筛选不完整编队,不完整编队数量记作 N,重组编队数记作 T,有

$$T = N\%2 \tag{3}$$

无法重组编队数记为 L,有

$$L = N - 2T \tag{4}$$

重组的编队根据遍历顺序赋予长机或僚机职能,无法重组的单机单独完成作战任务。

模型的输入为我机编队的位置坐标、航向及我机的存活状态,输出为我机的航向。

1.4 目标探测策略模型

为实现目标的快速定位,本文构建目标探测模型,提出基于人工势场的主动搜索方法,构建搜索圆域模型,设计被动搜索方案。

为确保主动搜索时编队的分布式搜索,采用人工势场维持我方无人机之间的距离,主要采用人工势场中的斥力场,我机编队在分布式搜索过程中应避免搜索区域的重复。通过定义势场函数,当友机间距离过近时,势场的斥力趋近无穷;当友机间距离超过指定值时,势场的斥力减少到 0 N。定义 $\rho(q)$ 为我机到其他友机自定义可调圆形边界 QO 的距离:

$$\rho(q) = \min_{q' \in \partial QO} \| q - q' \| \tag{5}$$

式中:q 为我机当前位形;q' 为边界位形;∂QO 表示空间障碍区域的边界。定义 ρ_0 为一个障碍物影响的距离,当我机 q 距离障碍(即友机)距离大于 ρ_0 时,不会排斥 q。符合上述标准的势函数描述为

$$U_{\text{rep}}(q) = \begin{cases} \dfrac{1}{2}\eta\left(\dfrac{1}{\rho(q)} - \dfrac{1}{\rho_0}\right)^2, & \rho(q) \leqslant \rho_0 \\ 0, & \rho(q) > \rho_0 \end{cases} \tag{6}$$

式中:η 为比例系数。排斥力为 $U_{\text{rep}}(q)$ 的负梯度,当 $\rho(q) \leqslant \rho_0$ 时,排斥力为

$$F_{\text{req}}(q) = \eta\left(\frac{1}{\rho(q)} - \frac{1}{\rho_0}\right)\frac{1}{\rho^2(q)}\nabla\rho(q) \tag{7}$$

如果 QO 为凸函数,b 是 QO 边界上最接近 q 的点,则

$$\rho(q) = \| q - b \| \tag{8}$$

其梯度为

$$\nabla\rho(q) = \frac{q - b}{\| q - b \|} \tag{9}$$

被动探测方面,被动雷达通过吸收敌方电磁波照射获取目标相对于自身的方位。被动探测的优点是能够在不发射电磁波的情况下对目标进行探测,缺点是精度较差,单架飞机仅能测得辐射来源的粗略方向,需要至少两架无人机协同被动探测目标才能实现目标位置的准确计算。

多机协同作战可利用被动雷达定位目标位置,当编队内有我机被动接收到目标信号时,友机配合支援,从不同方向进行同步雷达搜索,可以快速定位目标,并进行打击(干扰,打击协同一体化),但前提是目标不丢失。目标丢失分两种情况:

1)目标被其他友机摧毁;

2)目标雷达照射区域脱离被动探测区域(例如突然改变方向等)。

针对第 1 种情况,可以通过设计并检查全局摧毁列表来解决;针对第 2 种情况,放弃被动探测方法,直接开启主动雷达搜寻目标。

具体的搜索方法为:我机 1 被动探测到目标,主动雷达并没有探测到;我机 1 根据自身坐标位置及航向确定假想目标最远位置(被动探测能够确定目标方向,因此可以确定目标在该方位线上最远距离 d_{max} 到最近距离 d_{min} 之间),第 1 次记录的点记为 $p_v(x_v,y_v)$,此时调动距离最近的友机前来支援,但是最近的友机也可能受到目标的干扰,此时应跟随我机 1 一同朝向目标行进,并调动其他距离最近的友机。如果在判断圆域外,直接向 p_v 点航行(在中轴线友机侧),或者向我机 1 所在的位置航行(在中轴线友机另一侧)。如图 3 所示,友机在我机同侧时朝向 p_v 航行,友机雷达探测区域将覆盖目标位置,进而探测到目标具体坐标及方位;友机在我机对侧时朝我机(我机 1)方向航行,同样可以覆盖目标所有可能的位置。

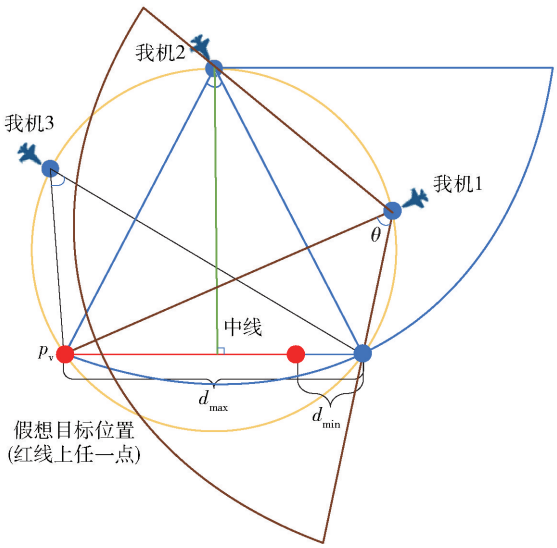


图 3　我机位于判断圆域外分析图

Fig. 3　Analysis of our UAVs located outside the judgment circle

如果在判断圆域内且位于我机 1 一侧,同样直接朝向 p_v 航行,到达中线位置仍未探测到,则掉头朝向我机 1 航行。反之亦然,按照该策略一定能够快速探测到目标。图 4 中深蓝色扇形表明初始位置友机的探测区域,由于目标处于探测区域外,为了全覆盖对侧目标可能存在的区域,需要飞到中线,如果没有探测到,折返朝向我机 1 航行。

已知 $\theta=60°$,我机 2 飞行到中线再折返的原因在于中线与判断圆域的交点 Q 距离 p_v 恰好为最大探测距离 d_{max},此时朝向 p_v 能够覆盖目标所在弦。若我机 2 在圆域内 Q 点与我机 1 构成的弦内接以 p_v 为圆心、p_vQ 为半径的部分圆弧,在此圆弧外时距

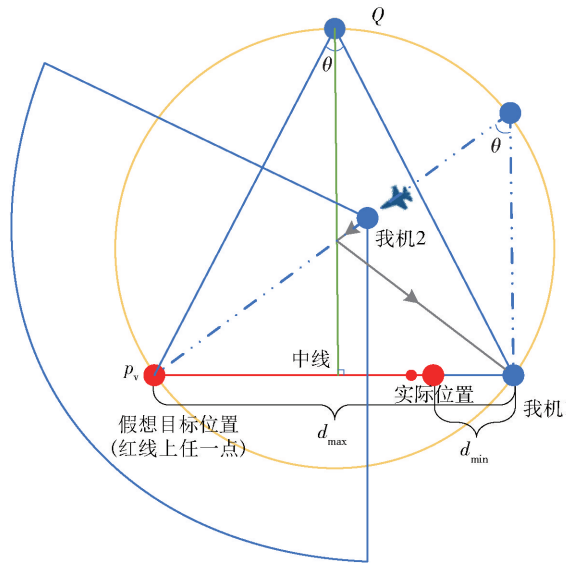


图 4　我机位于判断圆域内分析图

Fig. 4　Analysis of our UAVs located in the judgment circle

离 p_v 大于最大探测距离 d_{max},需要飞到中线附近才能够全覆盖。这个极限在于 Q 点,越趋近于 Q 点,意味着越需要朝着中线行进,才能全覆盖。为了便于处理,我机 2 没有在弦的不同侧采取不同策略,而是统一按照先到达中线再折返这一思路。实际上,当我机 2 在圆域内由 Q 点与我机 1 构成的弦右侧圆弧内时,只需朝向 p_v 进行瞬时探测,若没有发现目标即可折返。

模型输入为我机的位置坐标及航向,输出为我机的航向。

1.5　武器选择与目标打击策略模型

为实现先敌打击,构建武器选择与目标打击模型,建立打击目标分配策略,分析导弹攻击区,给出打击策略。导弹攻击区如图 5 所示。

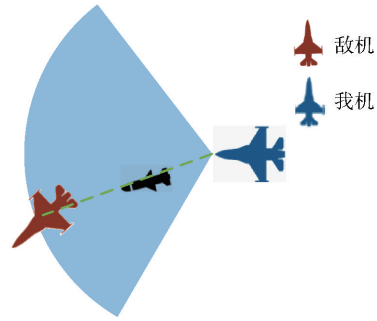


图 5　导弹攻击区

Fig. 5　Missile attack zone

整个作战 OODA 环中先敌打击至关重要。显然,当目标位于武器极限攻击距离时立即开火即为最优打击策略。武器的种类需要根据距离进行选

择,首选远程导弹,远距探测到即打击,无远距导弹可贴近用中距导弹,近距离则选中距导弹。

此外,当编队作战时,应考虑打击目标的分配问题。打击目标 id 放入全局打击列表中,每次迭代到相应无人机时查询本机打击列表是否在全局打击列表中,若存在,则具有相同打击目标的无人机不打击此目标。若打击无人机阵亡,则将目标 id 从全局打击列表中移除。打击目标按照我机与目标个体间距离大小进行分配,距离近的个体优先执行对应 id 的目标打击任务,如果目标在全局打击列表中,友机选择除此机之外探测到的目标进行打击。

模型输入为探测到的目标位置坐标及航向,输出为我机的航向。

1.6 目标追踪策略模型

为实现探测到目标后的目标追踪,本文构建目标追踪模型,构建其观测值与奖励函数,最后基于最大化熵软演员-评论家(SAC)算法训练模型。

模型输入为我机位置坐标、航向及探测到的目标位置坐标,输出为我机航向。

1.7 干扰规避策略模型

为了避免追踪过程中因目标干扰导致目标丢失,本文构建了干扰规避模型,构建其观测值与奖励函数,最后基于 SAC 算法训练模型。

模型输入为我机的位置坐标、探测到的目标位置坐标及我机航向,输出为我机的航向。

2 空战多维决策模型

为实现空战多维决策,需要构建空战多维决策模型。本文基于分层结构,将底层决策模型分为依靠专家知识的经验模型和基于 SAC 算法决策的训练模型。针对决策模型何时结束的问题,本文基于 Option-Critic 算法,摒弃策略训练,取而代之使用已有的策略模型,仅训练策略的终止函数,实现策略的灵活切换。顶层策略选择器基于贪心算法,选择期望回报最高的策略作为当前状态下的决策。

2.1 元策略模型训练算法

对于由雷达开关、主动干扰、队形转换、目标探测、武器选择与目标打击元策略构成的经验模型基于专家知识无需训练。对于由目标追踪和干扰规避策略构成的训练模型,训练采用 SAC 算法。其在传统的 Actor-Critic 方法引入最大化熵的思想,采用与 PPO[19] 类似的随机分布式策略函数,且是 Off-Policy、Actor-Critic 的算法。SAC 算法区别于其他算法的明显之处在于 SAC 同时最大化了回报和策略的熵值。在实际应用中,SAC 在各种常用的 benchmark 以及真实的机器人控制任务中表现稳定、性能优秀,具有极强的抗干扰能力。针对 DDPG 算法选择确定性策略问题,SAC 引入了最大化熵方法,能够让策略尽可能随机,智能体可以充分探索状态空间,避免策略过早陷入局部最优,并且可以探索到多个可行的方案来完成制定任务,提高了抗干扰能力。此外,为提高算法性能,采用 DQN 中的技巧,引入两个 Q 网络以及目标网络,为表述最大化熵值的重要程度,引入自适应温度系数 α,针对不同问题温度系数的调节,将其构造成一个带约束的优化问题,即最大化期望收益的同时,保持策略的熵大于一个阈值。SAC 训练模型算法的伪代码如图 6 所示。

SAC 训练模型算法

初始化网络参数 θ_1,θ_2,ϕ,初始化目标网络权重 $\bar{\theta}_1\leftarrow\theta_1$,$\bar{\theta}_2\leftarrow\theta_2$,初始化一个空经验池 $\mathscr{D}\leftarrow\varnothing$
  循环迭代
  对于环境中的每一步,循环迭代
    从策略中采样动作,即 $a_t\sim\pi_\phi(a_t|s_t)$
    从环境中通过转移函数采样下一时刻状态,即 $s_{t+1}\sim p(s_{t+1}|s_t,a_t)$
    存储经验到经验池中,即 $\mathscr{D}\leftarrow\mathscr{D}\cup\{(s_t,a_t,r(s_t,a_t),s_{t+1})\}$
  结束循环
  对于网络参数梯度,循环迭代
    更新 Q 函数网络参数 $\theta_i\leftarrow\theta_i-\lambda_Q\hat{\nabla}_{\theta_i}J_Q(\theta_i)$ for $i\in\{1,2\}$
    更新策略网络权重 $\phi\leftarrow\phi-\lambda_\pi\hat{\nabla}_\phi J_\pi(\phi)$
    调整温度参数 $\alpha\leftarrow\alpha-\lambda\hat{\nabla}_\alpha J(\alpha)$
    更新目标网络参数权重 $\bar{\theta}_i\leftarrow\tau\theta_i+(1-\tau)\bar{\theta}_i$ for $i\in\{1,2\}$
  结束循环
结束循环
输出优化后参数 θ_1,θ_2,ϕ

图 6 SAC 训练模型算法伪代码

Fig. 6 Pseudocode of SAC training model algorithm

2.2 空战多维决策算法

2.2.1 决策结构分解

为构建整体作战策略,需要确定作战流程以及作战逻辑,整体作战的分层决策结构图如图 7 所示。

决策选择层作为策略选择器负责在当前状态下进行元策略的挑选,初始编队及需要编队重组时选

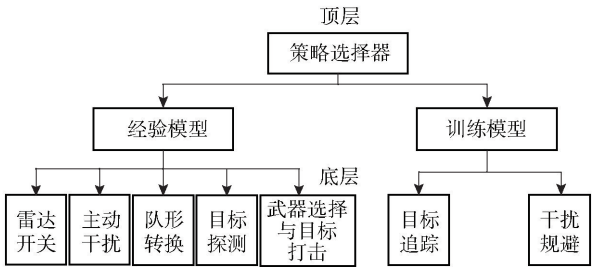


图 7　整体作战的分层决策结构

Fig. 7　Hierarchical decision-making structure for operations

择队形转换策略;在雷达未发现目标阶段应选择目标探测策略进行目标搜索(分布式);搜索过程中要合理分配雷达资源选择雷达开关策略;发现目标选择目标追踪策略对目标展开追击,追踪目标过程中避免目标丢失与反击应该采取主动干扰策略对目标雷达干扰,并采取干扰规避策略;目标进入攻击区时采用武器选择与目标打击模型完成对敌打击。

整个作战策略由 7 部分元策略构成:训练和干扰规避 2 个训练策略;雷达开关、主动干扰、队形转换、目标探测及武器选择与目标打击 5 个固定策略。对于训练策略基于 Actor-Critic 框架分别构建执行和评估神经网络。记录状态空间、动作空间和奖励值,最终为这两个策略设计经验池。

2.2.2　改进 Option-Critic 方法

由于基于传统 Option-Critic 的分层强化学习方法很难引入专家的经验知识且只能输入元策略的个数,其余均由 Option-Critic 算法训练每个元策略的策略函数和终止函数。而选项方法虽能引入经验知识,但要求人为设计终止函数,无法实现元策略的灵活切换。为了更好解决复杂空战问题,引入现有效果较好的专家经验模型十分必要,且具有明显的策略含义。本文基于传统的 Option-Critic 算法并做出改进,为引入自定义模型,首先为Option-Critic指定现有元策略模型的个数,将每个自定义策略模型和 Option-Critic 框架下的模型一一对应起来,在执行 Option-Critic 框架训练时,对于选中的策略仅训练其终止函数,策略函数由自定义模型提供。

上层策略选择一个选项 $\omega \in \Omega$,选项包含 3 部分:策略 $\pi_\omega(a|s)$ 表示选项中的策略,终止条件 β 表示状态 s 有 $\beta_\omega(s)$ 概率结束当前选项,初始集 I_ω 表示选项的初始状态集合。

当终止函数返回 0 时,下一步还会由当前选项来控制;当终止函数返回 1 时,该选项的任务暂时完成,控制权交还给上层策略。把每个选项的终止函数都用神经网络进行函数近似来参数化表示,即 $\beta_{\omega,\vartheta}(s)$,$\vartheta$ 表示网络参数,策略选取构建好的模型策略 $\pi_\omega(a|s)$。在这些选项之间做选择的上层策略,用 $\pi_\Omega(\omega|s)$ 表示,即在状态 s 时策略选择选项 ω 的概率。在此基础上,可以定义某状态下选择某个选项后产生的总收益。选择某个选项时,采取某行动之后产生的总收益和在使用某选项到达某状态之后产生的总收益。

选项内部仅更新为各选项的终止函数 $\beta_{\omega,\vartheta}(s)$。根据总折扣回报相对其参数的导数,可以利用如 policy gradient 的方法更新其参数。改进的 Option-Critic 算法结构如图 8 所示,与原算法相比,本文将训练策略改成了自定义策略。图 8 中,A_Ω 为选项之间的优势函数,a_t 为 t 时刻的动作,ω_t 为 t 时刻选择的选项,r_t 为 t 时刻的奖励,Q_U 为最优选项-价值函数。

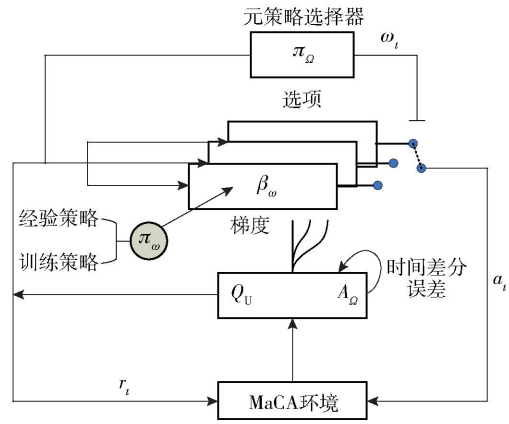


图 8　改进 Option-Critic 算法结构图

Fig. 8　Diagram of improved Option-Critic algorithm structure

2.2.3　多维空战决策算法构建

策略选择器采用贪婪策略,相应的单步离线策略更新目标 $g_t^{(1)}$ 为

$$g_t^{(1)} = r_{t+1} + \gamma((1-\beta_{\omega_t,\vartheta}(s_{t+1}))\cdot \sum_a \pi_{\omega_t,\theta}(a|s_{t+1})Q_U(s_{t+1},\omega_t,a) + \beta_{\omega_t,\vartheta}(s_{t+1})\max_\omega \sum_a \pi_{\omega,\theta}(a|s_{t+1})Q_U(s_{t+1},\omega,a)) \tag{10}$$

式中:γ 为折扣率;s_t 为 t 时刻状态;ω、a 为尚未观测到的随机变量。

多维空战决策算法(简称 Beta 算法)伪代码如图 9 所示。图 9 中,$Q_\Omega(s',\omega)$ 表示状态 s' 下选项 ω 的价值函数,$V_\Omega(s')$ 表示状态 s' 的价值函数,α 为软更新系数,δ 为时间差分误差,α_ϑ 为更新参数 ϑ 的学习率,$\overline{\omega}$ 为随机变量。模型的输入为我机的所有

状态及探测到目标的所有状态构成的状态池,模型输出更新状态池。元策略网络根据自身输入需要从状态池中获取输入值。空战全流程单元模型构建内容及方法、单元模型训练流程以及分层智能体训练方法如图 10 所示。

多维空战决策算法

1. 加载经验模型与训练模型的策略函数 $\pi_\omega(a|s)$,随机初始化改进 Option-Critic 网络参数 ϑ,元策略数量设为 7,初始状态 $s\leftarrow s_0$
2. 顶层策略选择器根据贪心策略 $\pi_\Omega(s)$ 选择元策略 ω
3. 重复以下步骤
4. 底层根据经验、训练策略 $\pi_\omega(a|s)$ 选择动作 a
5. 在状态 s 下采取动作 a,获取下一时刻观测值和奖励值 s',r
6. 元策略评估
7. 定义 $\delta\leftarrow r-Q_U(s,\omega,a)$
8. 如果 s'不是终止状态,则 $\delta\leftarrow\delta+\gamma(1-\beta_{\omega,\vartheta}(s'))\cdot Q_\Omega(s',\omega)+\gamma\beta_{\omega,\vartheta}(s')\max\limits_{\bar{\omega}}Q_\Omega(s',\bar{\omega})$
9. 结束
10. 元策略价值 $Q_U(s,\omega,a)\leftarrow Q_U(s,\omega,a)+\alpha\delta$
11. 元策略提升
12. $\vartheta\leftarrow\vartheta-\alpha_\vartheta\dfrac{\partial\beta_{\omega,\vartheta}(s')}{\partial\vartheta}(Q_\Omega(s',\omega)-V_\Omega(s'))$
13. 如果 $\beta_{\omega,\vartheta}$ 在 s'状态下终止,并跟据 $\pi_\Omega(s)$ 选择新的 ω
14. $s\leftarrow s'$
15. 直到 s'为终止状态
16. 顶层策略提升
17. $g_t^{(1)}=r_{t+1}+\gamma((1-\beta_{\omega_t,\vartheta}(s_{t+1}))\cdot\sum\limits_a\pi_{\omega_t,\theta}(a|s_{t+1})\cdot Q_U(s_{t+1},\omega_t,a)+\beta_{\omega_t,\vartheta}(s_{t+1})\max\limits_\omega\sum\limits_a\pi_{\omega,\theta}(a|s_{t+1})Q_U(s_{t+1},\omega,a))$

图 9 Beta 算法伪代码

Fig. 9 Pseudocode of Beta algorithm

3 仿真环境与仿真结果

3.1 实验环境设定

3.1.1 软件平台

选用文献[22]推出的 MaCA 环境对本文建立的模型进行仿真验证。MaCA 环境支持作战场景和规模自定义,智能体数量和种类自定义,智能体特征和属性自定义,支持智能体行为回报规则和回报值自定义等。

MaCA 环境中提供了一个电磁空间对抗的多智能体实验环境,环境中预设了探测单元和攻击单元两种智能体类型:探测单元可模拟 L、S 波段雷达进行全向探测,支持多频点切换;攻击单元具备侦察、探测、干扰、打击等功能,可模拟 X 波段雷达进行指向性探测,模拟 L、S、X 频段干扰设备进行阻塞式和瞄准式电子干扰,支持多频点切换,攻击单元还可对对方智能体进行导弹攻击,同时具有无源侦测能力,可模拟多站无源协同定位和辐射源特征识别。

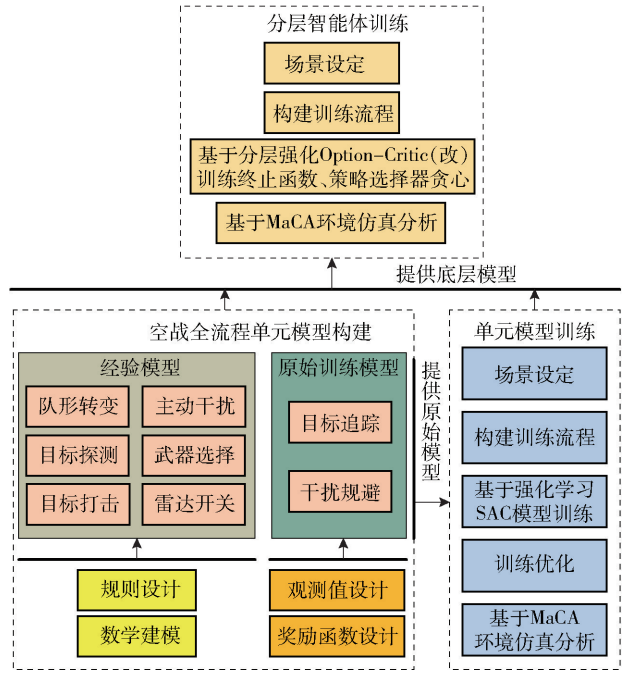


图 10 多维空战的构建方法及流程

Fig. 10 Construction method and process of multi-dimensional air combat

MaCA 环境为研究利用人工智能方法解决大规模多智能体分布式对抗问题提供了很好的支撑,专门面向多智能体深度强化学习开放了 RL-API 接口。环境支持使用 Python 语言进行算法实现,并可调用 Tensorflow、Pytorch 等常用深度学习框架。

3.1.2 硬件环境

CPU 采用 Intel i7-10700KF,GPU 采用 Nvidia RTX 3070 加速深度神经网络训练过程,显存大小为 8 GB,内存 16 GB。

3.2 定义想定任务

假定红蓝双方功能完全一致。双方在指定地图大小的二维环境中完成整个探测-干扰-规避-协同-打击作战流程。蓝方为规则驱动,规则未知。双方任务为在规定作战步数内尽可能少地消耗导弹去歼灭更多的目标,取得数量优势。单机 1v1 对抗场景地图修改双方战机数量为 1,远程导弹与近程导弹各 4 发,地图尺寸设置为 500×500。目标开启阻塞干扰,算法采用 MaCA 环境中的 fix_rule_no_att

黑盒算法;我机采用多维决策算法。共执行 20 回合,每回合最大运行步数为 5 000。

多机 4v4 对抗场景地图修改双方战机数量为 4,远程导弹与近程导弹各 4 发,地图尺寸设置为 500 × 500。目标开启阻塞干扰雷达,算法采用 MaCA 环境中的 fix_rule_no_att 黑盒算法;我机采用 Beta 算法。共执行 10 回合,每回合最大运行步数为 5 000。

3.3　跟踪元策略训练

3.3.1　训练方法

我机当前状态下的航向角为 α_1,坐标为(x_0,y_0);目标当前状态下的方位角为 α_2,坐标为(x_1,y_1)。记下一个状态我机航向角为 α_{1n},坐标为(x_{0n},y_{0n});下个状态目标方位角为 α_{2n},坐标为(x_{1n},x_{2n}),设偏航角为下个状态我机航向角与当前状态目标方位角的差值,记作 Δ,有 $\Delta=\alpha_{1n}-\alpha_2$。目标追踪模型为纯追踪,问题模型为最小化 Δ。Δ 和 α_2 作为神经网络的输入观测值。目标追踪问题模型最小化 Δ,因此可以构造二次函数 $R=-\Delta^2$ 作为问题的奖励函数,Δ 越小,奖励值越大,越接近 0。

随机初始状态,我机开启雷达对目标进行探测,目标干扰雷达关闭。训练环境采取 1v1 方式,首先固定我机进行跟踪训练。双方观测规则均采用 MaCA 环境中的 raw 规则,输入状态维度为 2,动作维度为 1,Actor-Critic 网络中 Actor 策略网络学习率设置为 3×10^{-4},Critic 策略网络学习率设置为 3×10^{-3},温度参数设置为 3×10^{-4},神经网络隐含层单元数为 512,共两层。回报折扣率设为 0.99,软更新参数设置为 0.005。经验池大小设置为 100 000,最小存储数据量设为 1 000 条。一次喂入神经网络的 batch 大小为 64,总回合数为 100,每回合最大步数设置为 500。将环境 Render 设置为可见。

为了加快训练并丰富样本,提出了训练优化方法。设置目标高速移动,我机固定,设置目标移动策略为每隔 10 步随机改变航向,缩小地图尺寸为 50 × 50。频繁的方位改变能够让我机充分探索各个航向。

整体训练过程收敛迅速,通过观察可见我机成功锁定目标,如图 11 所示。

图 12 和图 13 展示了总训练轮数为 100 和 1 000 次的回报曲线,可见,使用 SAC 算法在目标运动,我机固定时的环境下训练效果显著,该目标追踪问题在每回合 1 000 步的迭代中能够在第 10 回合左右达到收敛,收敛效果很好。

在 1 000 次的训练中,为了避免过拟合问题,丰富训练样本,本文采用了动态改变环境的方法,通过

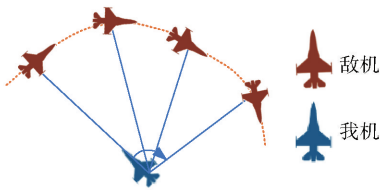


图 11　目标运动、我机固定时的追踪训练示意图

Fig. 11　Schematic diagram of tracking training when the enemy UAVs are moving and our UAV is fixed

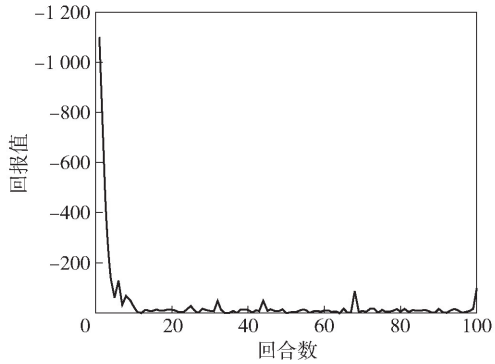


图 12　SAC 跟踪训练的原始回报曲线(100 轮)

Fig. 12　Original reward curve of SAC tracking training (100 rounds)

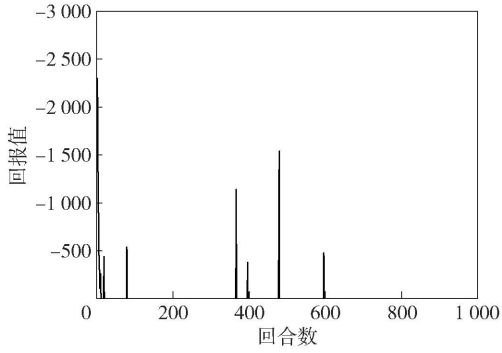


图 13　SAC 跟踪训练的原始回报曲线(1 000 轮)

Fig. 13　Original reward curve of SAC tracking training (1 000 rounds)

动态改变地图的大小与无人机位置保证初识状态的不同。从图 13 中可见,算法收敛后仍有一些回合回报值较低,但很快便达到收敛状态。

我机固定训练完成,将经过大量回合训练好的模型保存,改变地图参数,让我机具有速度并扩大地图尺寸,目标速度降低为与我机速度一致,验证跟踪模型的有效性。在 MaCA 环境中,对目标速度的改变并不会影响整体的代码结构,仅需在 map 地图中设置 speed 参数即可。使用图 14 展示了总验证回合为 10、目标移动时的验证回报曲线。

由图 14 可见,当我机运动时回报值依旧较小,

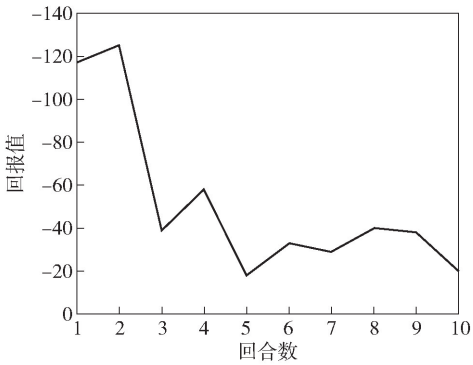


图 14 我机移动时跟踪验证的回报曲线

Fig. 14 Reward curve for tracking verification when our UAV is moving

通过 10 回合的验证(非训练),如图 15 所示,发现我机能在目标转向时完美同步追踪,跟踪效果显著,验证了训练模型的有效性。

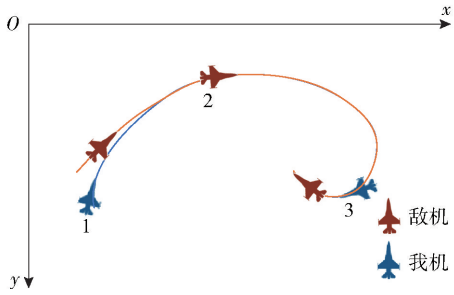


图 15 验证演示过程示意图

Fig. 15 Schematic diagram of verification demonstration

3.3.2 对照实验

采用 DDPG 算法作为对照,环境设置相同,通过调整超参数得到我机运动的情况下最优跟踪训练的回报曲线如图 16 所示。

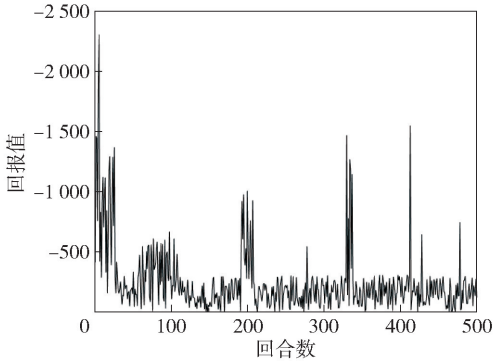


图 16 基于 DDPG 的跟踪训练原始回报曲线

Fig. 16 Original reward curve of tracking training based on DDPG

DDPG 算法处理该问题同样具有收敛趋势,但相较于 SAC 算法曲线波动大,回报值没有稳定在 0 附近。这是由于 DDPG 算法对超参数极其敏感,略微调整学习率就可能导致训练结果不收敛。此外,DDPG 算法只选择一个最优策略,而不考虑等优策略,这影响了算法稳定性及迁移性。而 SAC 算法同时最大化策略熵与回报期望,能够学到等优策略,此外,SAC 算法对于不同的随机数种子等超参数能够达到同样的收敛效果,稳定性能也更加出色。

3.4 干扰规避元策略训练

我机相对于目标的方位角为 α_1,目标的航向角为 α_2,我机的航向角为 α_3,目标航向角与方位角度差值为 Δ_1,方位角与我机的航向角差值为 Δ_2,两机间距记为 d。可以根据目标航向角与方位角差值以及中值 π 来构造观测值和奖励函数。同样地,先对观测值进行预处理,取 $\cos(|\Delta_1-\pi|)$ 作为观测值,由于 $|\Delta_1-\pi|\leqslant\pi$,当 Δ_1 距离中值 π 越近时,观测值越大且附近越平缓,表明状态很好。相反当距离中值越远时,观测值越小,且附近的观测值变化不大,表明状态很差。

奖励函数的构建需要记录当前状态下的目标航向 α_2,目标坐标 (x_2,y_2),下一状态下的方位角 α_{1s},我机坐标 (x_{1s},y_{1s})。计算两个状态下的坐标距离衡量采取的动作是否减小了距离,记作

$$d=(x_{1s}-x_2)^2+(y_{1s}-y_2)^2 \tag{11}$$

奖励函数第 1 部分构造为

$$R=-(\alpha_{2s}-\alpha_1)^2\cdot d \tag{12}$$

奖励值越大,表明我机航向角与目标方位角之间夹角越小,我机与目标之间的距离越近。

奖励函数第 2 部分构造为区域奖励,判断当前状态目标的航向角与下一状态方位角之间的差距是否在[175°,185°]和[-175°,-185°]内,如果在该范围内,则奖励函数为

$$R=-1\,000\cdot|\cos(10(\alpha_2-\alpha_{1s}))| \tag{13}$$

该奖励值越小,表明距离干扰中线越近,并且在远离干扰中线的方向奖励值梯度很大,引导我机优先进行干扰区的规避。最终观测值构造为 $\cos(|\Delta_1-\pi|)$、Δ_2、α_2、α_1、d。超参数设定同3.3 节。

通过图 17 的回报曲线可见随着训练的进行,整体训练过程达到收敛状态。观察整个追踪加躲避干扰的演示过程,如图 18 所示,可以明显看到我机在初始位置受到目标干扰时,选择快速逃离干扰区,逃离成功后对目标进行追击。

3.5 分层框架构建

3.5.1 回报奖励设计

模型训练需要的回报奖励依据不同的元策略分

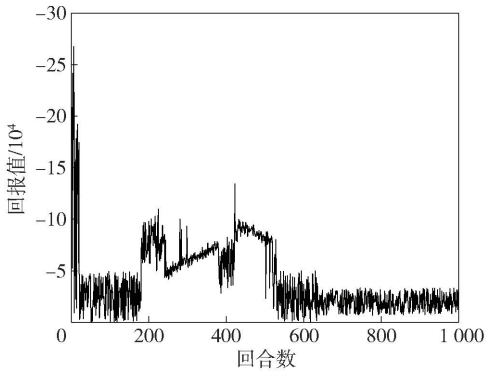


图 17　跟踪加规避干扰的回报曲线

Fig. 17　Tracking plus distraction avoidance payoff curve

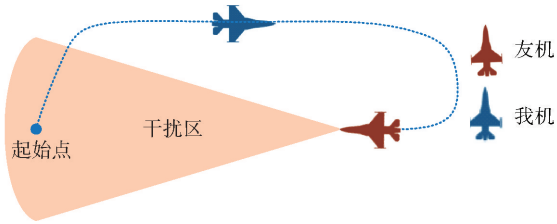


图 18　干扰规避与追踪演示图

Fig. 18　Interference avoidance and tracking demo

别设置。7 种元策略的具体奖励设置方案如下：

1)针对目标探测策略，设置固定时间步长 τ_1，发现目标则应该由 $\beta(s)$ 控制结束该选项。发现目标给予奖励 $R_o=10$。

2)针对武器选择与目标打击策略，成功打击一次目标记奖励为 $R_h=100$，若打击列表仍存在打击目标，则应给予与终止函数概率呈反比的惩罚项：

$$R_\beta=-\frac{1}{0.001+\beta(s)} \tag{14}$$

3)针对队形转换策略，在对局开始阶段与我机战损时完成策略执行记一次集群奖励 $R_g=30$。

4)针对雷达开关策略，成功完成开关切换动作记一次开关奖励 $R_s=20$。

5)针对主动干扰策略，探测到目标频点并成功设定干扰频点记一次干扰奖励 $R_j=100$。

6)对于目标追踪策略，发现目标时提供最大策略奖励，记作 $R_c=100$。

7)对于干扰规避策略，我机位于目标干扰区域时提供给最大策略奖励，记为 $R_m=100$。

3.5.2　决策模型构建

选项模型构建结构与一般 Option-Critic 结构不同，在训练时采取之前已经构建好的元策略模型，通过封装好的元策略模型，输入环境状态返还动作，即放弃 Option-Critic 自动训练策略，这是因为采取 Option-Critic 方法训练的元策略无法将动作序列具体化为符合人类逻辑与经验的策略，且容易过拟合到某个动作，而非序列动作。改进后的方法只通过 Actor-Critic 网络训练得到终止函数，给出元策略在不同状态下完成执行的时间。Critic 网络根据状态及奖励更新选项价值函数 Q 值以及选项间的优势函数，梯度反向传递 Actor 网络完成参数更新给出终止函数。决策模型算法流程与示意图如图 19 与图 20 所示。

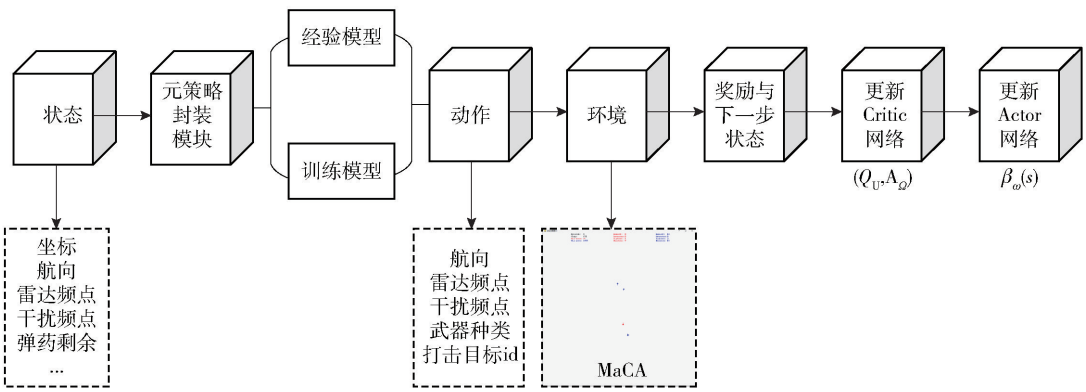


图 19　决策模型算法流程

Fig. 19　Process of decision-making model algorithm

3.6　仿真实验结果分析

3.6.1　1v1 对战结果

由于 MaCA 环境截图效果无法显示作战流程，这里通过记录实际数据绘制具有说明意义的局部作战轨迹，给出作战示意图。

图 21 展示了我机探测到目标之后选择目标追踪策略对目标展开追踪，能够精准锁定目标方位并进行追踪；在追踪过程中干扰开机且干扰频点发生

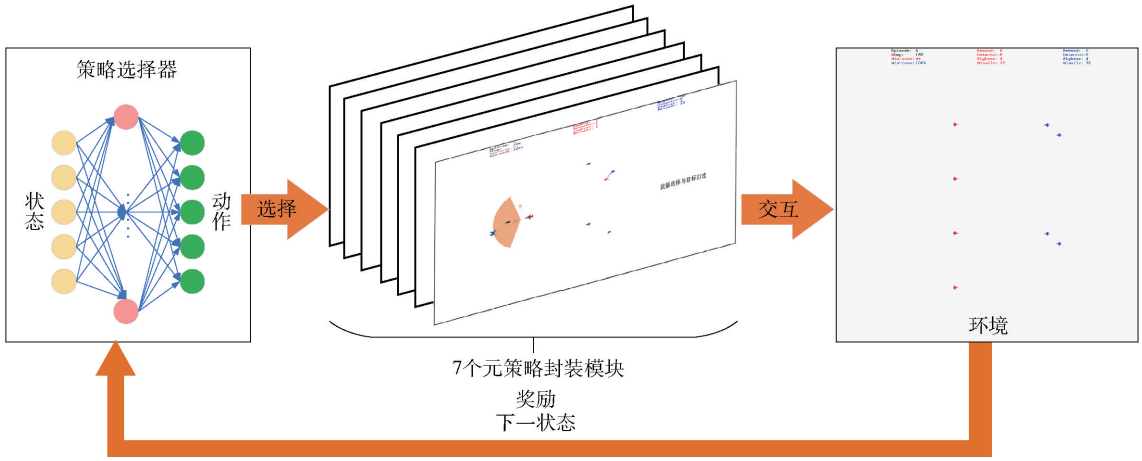


图 20 多维决策算法执行示意图

Fig. 20 Schematic diagram of multi-dimensional decision-making algorithm execution

转变,说明采取了主动干扰策略;与此同时,开启干扰规避策略规避目标干扰,从示意图中可见我机绕开了目标的干扰区域成功规避目标干扰。

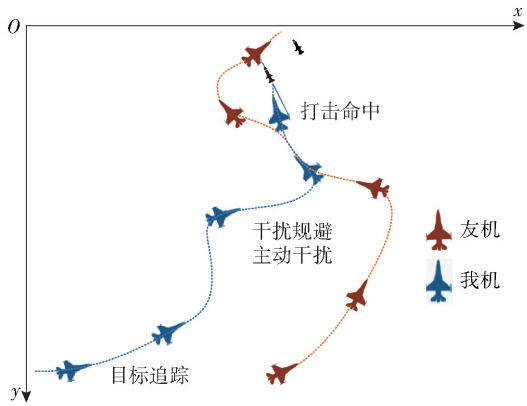


图 21 分层强化 1v1 的仿真验证示意图

Fig. 21 Schematic diagram of simulation verification of hierarchical reinforcement 1v1

与多机不同的是,在 1v1 中,由于没有友机提供目标的位置坐标信息,我机判断是否进入干扰区的条件是目标上一时刻所处探测区域的角度以及目标是否丢失,如果处于探测区域的边缘且目标丢失,则需要重新搜索,被干扰概率很低。如果处于探测区域中央且下一时刻目标丢失,则大概率由于目标干扰造成,从而判断出我机位于目标区,此时将假定目标在上一时刻的方位且朝向我机,我机需要躲避该固定干扰区域;将目标引导到攻击区,在适当的距离选择攻击动作并成功打击目标。

各个策略的执行动作均以合适的时机完成切换,且策略执行的终止函数能够正确地在不同状态终止正在执行的策略。通过 20 轮的算法演示验证,我机(红方)均取得胜利,以 100% 全胜的战绩战胜了目标算法,验证了该分层强化决策的有效性。

3.6.2 4v4 对战结果

首先对雷达开关策略有效性验证,Beta 智能体执行全流程空战时,禁用雷达开关策略使雷达常开,统计目标被动雷达探测到我机的频次,记为 N_1。计算全部 10 回合目标被动探测到我机频次的总数记为 N,计算平均每局被动发现次数,记为 $\overline{N}$,有 $\overline{N}=N/10$。

另外执行 10 回合使能雷达开关策略的 Beta 智能体全流程空战,统计目标被动探测到我机的频次,记为 $\mathcal{N}_1$。计算全部 10 回合目标被动探测到我机频次的总数记为 $\mathcal{N}$,计算平均每局被动发现的次数,记为 $\overline{\mathcal{N}}$,有 $\overline{\mathcal{N}}=\mathcal{N}/10$。

表 1 展示了 10 回合中雷达开关与否的被动发现次数统计,可见采用雷达开关策略能够大幅减少目标被动发现我机的频次,减少因目标被动探测暴露我机位置的概率。这点对于在执行异构的情况下我机执行的掩护任务尤为重要。

图 22 展示了 10 回合中雷达开关与否被动发现次数的平均统计柱形图,可以直观看出采用雷达开关策略目标被动探测到我机平均次数为 13 次,而禁用雷达开关策略被动探测次数上升到平均 20 次,证明了在分层决策下雷达开关策略的有效性。

图 23 所示为 Beta 智能体可以在回合开始完成初始化编队,以长机-僚机编队进行目标的探测与干扰,这种初始编队能够优先取得信息优势,实现压制干扰,在我方无损毁的情况下歼灭两架目标,在开始便掌握数量优势。我机编队在歼灭了受压制干扰的目标之后终止函数停止编队决策,

表 1　10 回合雷达开关与否的被动发现次数统计

Table 1　Data of the number of passive discoveries with or without radar switch in 10 rounds

回合	N_1	$\mathcal{N}_1$	$N_1-\mathcal{N}_1$
1	16	15	1
2	17	8	9
3	22	12	10
4	18	14	4
5	26	16	10
6	24	10	14
7	21	15	6
8	19	13	6
9	17	15	2
10	20	12	8

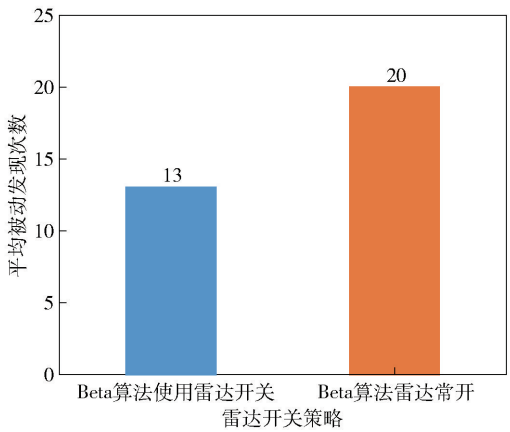


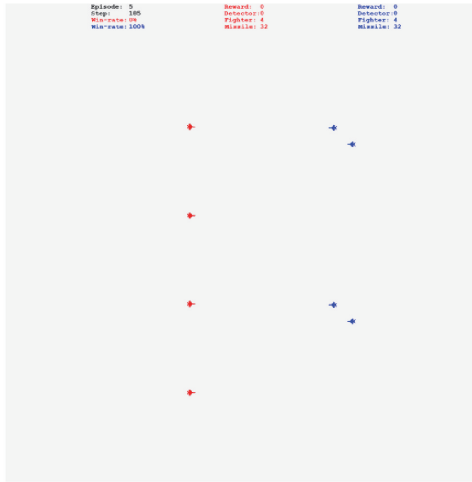
图 22　使用雷达开关策略与否的平均被动发现次数对比

Fig. 22　Comparison of average passive discoveries with and without radar switch strategy

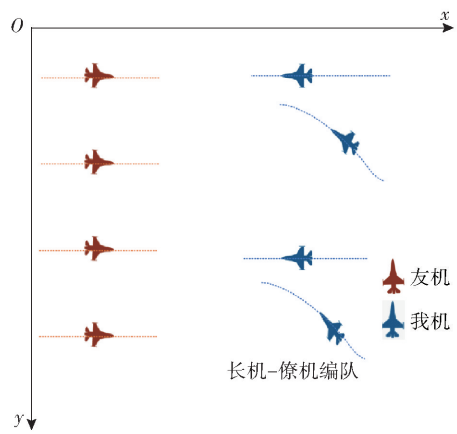
策略选择器根据当前状态选择其他元策略。队形转换策略在作战过程中能够在我机队形被破坏的情况下完成编队重组，重组后以编队形式对目标进行压制干扰，避免了因单独作战敌我同时发射导弹而互换的情况。

为验证队形转换策略的有效性，统计了 10 回合内编队重组后该编队 50 步内的打击情况以及损失情况。记编队成功打击目标数量为 S，编队成员损失数量为 D，每局的数量统计如表 2 所示，其中，S 与 D 均为 0 时代表本局编队结构未损坏。编队重组后协同打击仿真如图 24 所示。

从表 2 中的数据分析可知，除第 2、5、6、8 回合我机编队结构没有破坏外，其他回合重组的编队均完成了目标的压制打击任务，且仅有第 7 回合在编



(a) 程序运行截图(MaCA环境内)
(a) Screenshot of program operation (inMaCA)



(b) 决策示意图
(b) Diagram of decision-making

图 23　初始编队

Fig. 23　Initial formation

表 2　10 回合队形转换 50 步内打击与损失统计

Table 2　Data of strike and loss within 50 steps in 10 rounds of formation change

回合	S	D	回合	S	D
1	1	0	6	0	0
2	0	0	7	2	1
3	2	0	8	0	0
4	1	0	9	1	0
5	0	0	10	1	0

队执行搜索任务过程中损失 1 架战机，其他回合均无战机损失，验证了分层决策下队形转换策略的有效性。

图 25 展示了 Beta 算法能够在友机阵亡时指导我机向阵亡方位进行搜索，由于友机阵亡时代表目标在附近活动，前往该区域附近进行搜索能够大幅

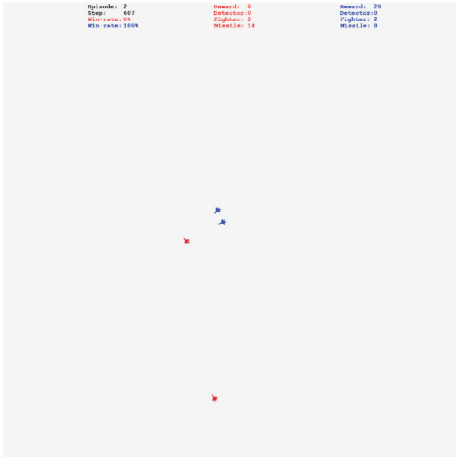
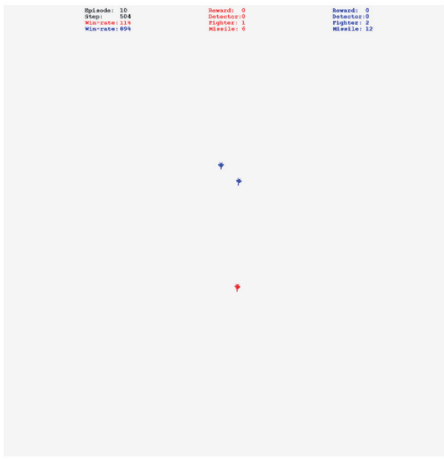


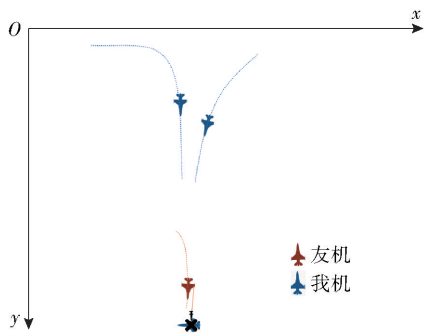
图 24 编队重组后协同编队打击目标

Fig. 24 Cooperative formation striking the target after formation reorganization

增加截获概率,充分展现了决策的智能性。



(a) 程序运行截图(MaCA环境内)
(a) Screenshot of program operation (inMaCA)

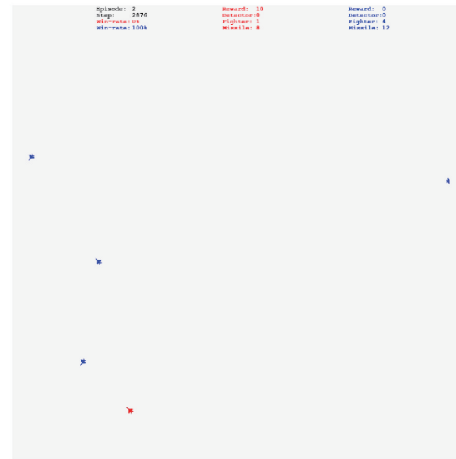


(b) 决策示意图
(b) Diagram of decision-making

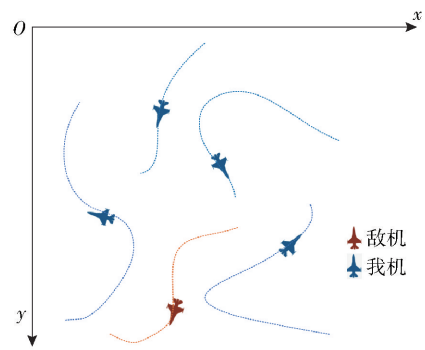
图 25 友机阵亡我机搜索决策

Fig. 25 Our UAV's search decision when friendly UAVs are killed

目标探测策略中的分布式搜索采用了人工势场中的斥力场,能够维持友机间距,避免探索重复区域。图 26 展示了我机分布式搜索策略,当友机相互靠近时会采取相互远离的动作。分层决策下,Beta 成功在我机雷达未发现目标时选择目标探测策略完成分布式搜索,验证了目标探测模型的有效性。



(a) 程序运行截图(MaCA环境内)
(a) Screenshot of program operation (inMaCA)



(b) 决策示意图
(b) Diagram of decision-making

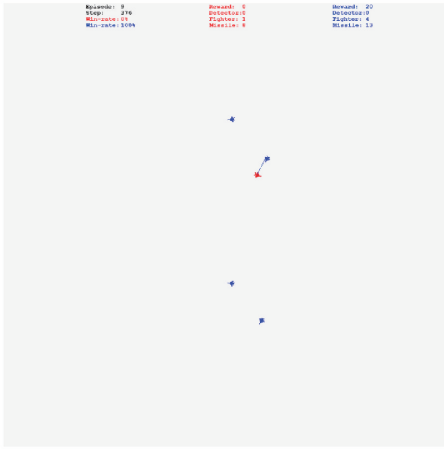
图 26 分布式搜索

Fig. 26 Distributed search

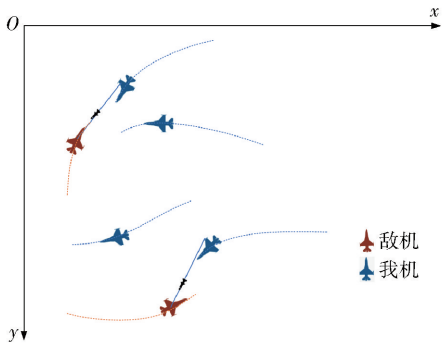
图 27 蓝色实线表示导弹发射并命中,展示了我机武器选择和打击决策。为验证分层决策结构是否可以准确完成先敌打击,需要考量我机采取武器选择与目标打击策略时我机与目标的距离是否恰为导弹攻击区的边缘。为此,统计了 10 轮全流程作战过程中选择打击时我机与目标平均距离,记为 D_1,给出该距离与攻击区边缘距离的差距,记为 Δ_1,相关数据记录如表 3 所示。

从表 3 中的数据可见,平均每局 Beta 采取打击决策时距离误差不超过 20 m,充分展现了分层决策在武器选择与目标打击的及时性与有效性,同时说明了终止函数训练的有效性。

此外,该图还展示了我机编队对目标进行压制干扰,让目标探测系统瘫痪从而无法对我机做出打



(a) 程序运行截图(MaCA环境内)
(a) Screenshot of program operation (inMaCA)



(b) 决策示意图
(b) Diagram of decision-making

图 27　打击演示
Fig. 27　Strike demo

表 3　10 回合队形转换 50 步内打击与损失统计

Table 3　The gap between the current distance and the edge of the attack zone in 10 rounds

回合	D_1/m	Δ_1/m	回合	D_1/m	Δ_1/m
1	119.3	0.7	6	108.2	11.8
2	114.7	5.3	7	103.9	16.1
3	100.2	19.8	8	114.2	5.8
4	118.9	1.1	9	119.1	0.9
5	117.3	2.7	10	119.3	0.7

击动作,虽然目标的探测能力与武器与我机完全一致,但由于受到干扰无法第一时间向我方发射导弹,从而可以使我机编队完成优先打击,主动干扰策略的有效性也体现于此。

对于武器的使用情况,为避免我机对同一目标发射多发导弹从而导致武器弹药的浪费,采取了打击目标分配方法。为验证该方法的有效性,统计了每局作战结束导弹的使用情况。结果表明,每局导弹消耗量均为 4,即一发导弹打击一个目标,利用效率为 100%。

为分析全流程作战分层决策 Beta 模型的有效性,本文定义了平均战损,描述为平均每局对决的无人机损失数量,由表 4 与图 28 可见在 10 场对局中 Beta 算法平均每局损失 1.4 架战机,黑盒算法 fix_rule_no_att平均损失 3.8 架,充分说明了结果表明 Beta 算法以全胜的战绩击败了黑盒算法。

表 4　作战结果

Table 4　Combat result

算法	回合数	胜率/%	平均战损
fix_rule_no_att 算法	10	0	3.8
Beta 算法	10	100	1.4

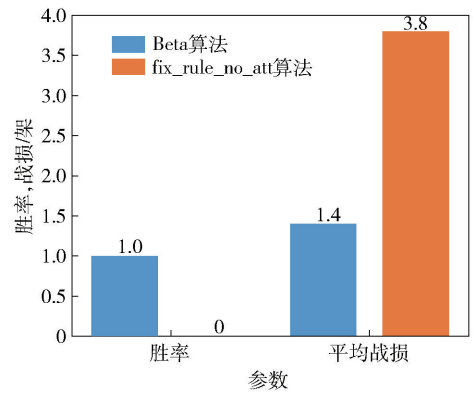


图 28　算法对比
Fig. 28　Algorithm comparison

4　结论

本文对无人机多维空战自主决策展开研究,分析了现有研究成果以及不足。针对相关研究缺乏多维决策的问题,提出了采用分层强化学习完成无人机多维决策模型构建以及训练方法的构建。本文提出的多维空战自主决策分层算法通过改变传统 Option-Critic 框架,结合传统选项框架能够引入专家经验的优势,同时无需设计终止函数,通过训练完成元策略的灵活切换,算法扩展性强、层次分明、策略意义显著。在 MaCA 环境完成红蓝对抗仿真,对空战战术决策以及自主空战的多元化发展具有重要意义。得到主要结论如下:

1)元策略训练收敛,仿真验证能够很好地完成目标跟踪、干扰规避等任务。

2)该算法能够在不同状态下完成最优策略切换,并以全胜的战绩击败黑盒算法 fix_rule_no_att。

下一步工作重点为:将训练环境迁移到三维,提

升状态动作空间的复杂度,引入动力学模型,引入侦察机将同构问题转化为异构问题,将该算法应用到更贴近真实空战的环境中进行测试。此外,该算法部分元策略基于专家经验,后续将完全去经验化,建立全训练模型。

参考文献(References)

[1] 杨伟. 关于未来战斗机发展的若干讨论[J]. 航空学报,2020,41(6):524337.
YANG W. Development of future fighters[J]. Acta Aeronautica et Astronautica Sinica, 2020, 41(6): 524377. (in Chinese)

[2] 刘冰雁, 叶雄兵, 周赤非, 等. 基于改进 DQN 的复合模式在轨服务资源分配[J]. 航空学报, 2020, 41(5): 323630.
LIU B Y, YE X B, ZHOU C F, et al. Allocation of composite mode on-orbit service resource based on improved DQN[J]. Acta Aeronautica et Astronautica Sinica, 2020, 41(5): 323630. (in Chinese)

[3] DAVID S, GUY L, NICOLAS H et al. Deterministic policy gradient algorithms[C]// Proceedings of the 31st International Conference on Machine Learning. Beijing, China: IEEE, 2014, 32(1): 387-395.

[4] 张耀中, 徐佳林, 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):324000.
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): 324000. (in Chinese)

[5] SHI H B, SUN Y R, LI G Y. Model-based DDPG for motor control[C]// Proceedings of 2017 International Conference on Progress in Informatics and Computing. Nanjing, China: IEEE, 2017: 284-288.

[6] KULKARNI T D, NARASIMHAN K R, SAEEDI A, et al. Hierarchical deep reinforcement learning: integrating temporal abstraction and intrinsic motivation[C]// Proceedings of the 30th Conference on Neural Information Processing Systems. Barcelona, Spain: Neural Information Processing Systems, 2016: 1826.

[7] 王俊敏, 姜青山, 罗泽明. 预警机指挥编队协同空战分层决策模型[J]. 海军航空工程学院学报,2014,29(5):491-496.
WANG J M, JIANG Q S, LUO Z M. A hierarchical decision-making model for cooperative air combat of early warning aircraft command formations[J]. Journal of Naval Aeronautical and Astronautical University, 2014, 29(5): 491-496. (in Chinese)

[8] 付跃文, 王元诚, 陈珍, 等. 基于多智能体粒子群的协同空战目标决策研究[J]. 系统仿真学报,2018,30(11):4151-4157.
FU Y W, WANG Y C, CHEN Z, et al. Research on target decision-making of cooperative air combat based on multi-agent particle swarm[J]. Journal of System Simulation, 2018, 30(11): 4151-4157. (in Chinese)

[9] 文永明, 石晓荣, 黄雪梅, 等. 一种无人机集群对抗多耦合任务智能决策方法[J]. 宇航学报,2021,42(4):504-512.
WEN Y M, SHI X R, HUANG X M, et al. An intelligent decision-making method for UAV swarms against multi-coupling tasks[J]. Journal of Astronautics, 2021, 42(4): 504-512. (in Chinese)

[10] 程先峰, 严勇杰. 基于 MAXQ 分层强化学习的有人机/无人机协同路径规划研究[J]. 信息化研究,2020,46(1):13-19.
CHENG X F, YAN Y J. Research on collaborative path planning of manned and unmanned aerial vehicles based on MAXQ hierarchical reinforcement learning[J]. Informatization Research, 2020, 46(1): 13-19. (in Chinese)

[11] 吴宜珈, 赖俊, 陈希亮, 等. 强化学习算法在超视距空战辅助决策上的应用研究[J]. 航空兵器,2021,28(2):55-61.
WU Y J, LAI J, CHEN X L, et al. Research on the application of reinforcement learning algorithm in decision-making assistance in over-the-horizon air combat[J]. Aero Weaponry, 2021, 28(2): 55-61. (in Chinese)

[12] POPE A P, IDE J S, MICOVIC D, et al. Hierarchical reinforcement learning for air-to-air combat[C]// Proceedings of 2021 International Conference on Unmanned Aircraft Systems. Athens, Greece: IEEE, 2021: 275-284.

[13] 冷鹏飞, 徐朝阳. 一种深度强化学习的雷达辐射源个体识别方法[J]. 兵工学报, 2018, 39(12):2420-2426.
LENG P F, XU Z Y. A deep reinforcement learning method for individual identification of radar radiation sources[J]. Acta Armamentarii, 2018, 39(12): 2420-2426. (in Chinese)

[14] 朱建文, 赵长见, 李小平, 等. 基于强化学习的集群多目标分配与智能决策方法[J]. 兵工学报, 2021, 42(9):2040-2048.
ZHU J W, ZHAO C J, LI X P, et al. Cluster multi-objective assignment and intelligent decision-making method based on reinforcement learning[J]. Acta Armamentarii, 2021, 42(9): 2040-2048. (in Chinese)

[15] 陈中原, 韦文书, 陈万春. 基于强化学习的多发导弹协同攻击智能制导律[J]. 兵工学报, 2021, 42(8):1638-1647.
CHEN Z Y, WEI W S, CHEN W C. Intelligent guidance law for cooperative attack of multiple missiles based on reinforcement learning[J]. Acta Armamentarii, 2021, 42(8): 1638-1647. (in Chinese)

[16] 高昂, 董志明, 叶红兵, 等. 基于深度强化学习的巡飞弹突防控制决策[J]. 兵工学报, 2021, 42(5):1101-1110.
GAO A, DONG Z M, YE H B, et al. Penetration control decision of cruise missile based on deep reinforcement learning

[J]. Acta Armamentarii, 2021, 42(5): 1101 - 1110. (in Chinese)

[17] 刘冰雁, 叶雄兵, 岳智宏, 等. 基于多组并行深度 Q 网络的连续空间追逃博弈算法[J]. 兵工学报, 2021, 42(3):663 - 672.
LIU B Y, YE X B, YUE Z H, et al. A continuous space chase-escape game algorithm based on multiple parallel deep Q-networks[J]. Acta Armamentarii, 2021, 42(3):663 - 672. (in Chinese)

[18] CHAKROVORTY J, WARD P N, ROY J, et al. Option-critic in cooperative multi-agent systems [C] // Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems Virtual. Auckland, New Zealand: IEEE, 2020: 1792 - 1794.

[19] 惠俊鹏, 汪韧, 俞启东. 基于强化学习的再入飞行器"新质"走廊在线生成技术研究[J]. 航空学报, 2022,43(9):623 - 635.
HUI J P, WANG R, YU Q D. Research of generating new quality flight corridor for reentry a ircraft based on reinforcement learning[J]. Acta Aeronautica et Astronautica Sinica, 2022, 43(9):623 - 635. (in Chinese)

[20] 罗杰,董志岩,翟鹏,等. 基于强化学习算法的智能飞控开发系统[J]. 计算机系统应用,2022,31(7):93 - 98.
LUO J, DONG Z Y, ZHAI P, et al. Intelligent flight control development system based on reinforcement learning algorithm [J]. Computer Systems & Applications, 2022, 31(7):93 - 98. (in Chinese)

[21] 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 H. Research of UCAV air combat based on reinforcement learning[D]. Harbin: Harbin Institute of Technology, 2015. (in Chinese)

[22] 中国电子科技集团公司认知与智能技术重点实验室. MaCA 环境说明[R]. 北京:中国电子科技集团公司第五十一研究所,2019:1 - 20.
China Electronics Technology Group Corporation Key Laboratory of Cognitive and Intelligent Technology. MaCA environment description[R]. Beijing: The 51st Research Institute of China Electronics Technology Group Corporation, 2019: 1 - 20. (in Chinese)