基于深度强化学习的多机协同空战方法研究

施 伟[1] 冯旸赫[1] 程光权[1] 黄红蓝[1] 黄金才[1] 刘 忠[1] 贺 威[2,3]

摘 要 多机协同是空中作战的关键环节, 如何处理多实体间复杂的协作关系、实现多机协同空战的智能决策是亟待解决的问题. 为此, 提出基于深度强化学习的多机协同空战决策流程框架 (Deep-reinforcement-learning-based multi-aircraft cooperative air combat decision framework, DRL-MACACDF), 并针对近端策略优化 (Proximal policy optimization, PPO) 算法, 设计 4 种算法增强机制, 提高多机协同对抗场景下智能体间的协同程度. 在兵棋推演平台上进行的仿真实验, 验证了该方法的可行性和实用性, 并对对抗过程数据进行了可解释性复盘分析, 研讨了强化学习与传统兵棋推演结合的交叉研究方向.

关键词 多机协同空战, 智能决策, 深度强化学习, PPO 算法, 增强机制

引用格式 施伟, 冯旸赫, 程光权, 黄红蓝, 黄金才, 刘忠, 贺威. 基于深度强化学习的多机协同空战方法研究. 自动化学报, 2021, **47**(7): 1610−1623

DOI 10.16383/j.aas.c201059

Research on Multi-aircraft Cooperative Air Combat Method Based on Deep Reinforcement Learning

SHI Wei[1] FENG Yang-He[1] CHENG Guang-Quan[1] HUANG Hong-Lan[1]
HUANG Jin-Cai[1] LIU Zhong[1] HE Wei[2,3]

Abstract Multi-aircraft cooperation is the key part of air combat, and how to deal with the complex cooperation relationship between multi-entities is the essential problem to be solved urgently. In order to solve the problem of intelligent decision-making in multi-aircraft cooperative air combat, a deep-reinforcement-learning-based multi-aircraft cooperative air combat decision framework (DRL-MACACDF) is proposed in this paper. Based on proximal policy optimization (PPO), four algorithm enhancement mechanisms are designed to improve the synergistic degree of agents in multi-aircraft cooperative confrontation scenarios. The feasibility and practicability of the method are verified by the simulation on the wargame platform, and the interpretable review analysis of the antagonistic process data is carried out, and the cross research direction of the combination of reinforcement learning and traditional wargame deduction is discussed.

Key words Multi-aircraft cooperative air combat, intelligent decision, deep reinforcement learning, proximal policy optimization (PPO) algorithm, enhancement mechanism

Citation Shi Wei, Feng Yang-He, Cheng Guang-Quan, Huang Hong-Lan, Huang Jin-Cai, Liu Zhong, He Wei. Research on multi-aircraft cooperative air combat method based on deep reinforcement learning. *Acta Automatica Sinica*, 2021, **47**(7): 1610−1623

多机协同空战是指由两架或两架以上的作战飞机互相配合、相互协作, 完成对空作战任务的一种战争方式, 包括协同机动、协同打击及火力掩护等环节, 是现代海、陆、空、天、电一体化作战模式在多机空战中的具体体现[1]. 因此, 提高多机协同效率, 对于掌握战场制空权、提高对空作战任务成功率和减少作战伤亡都具有重大意义. 世界各国也越来越关注和重视有助于提高机群整体作战效能的协同空战的研究. 然而, 相较于单架战机的空战决策, 多机协同问题涉及的实体类型更多、决策空间更大、复杂程度更高.

目前, 自主空战决策的算法研究, 依据其核心内涵的不同, 主要分为数学求解、机器搜索以及数据驱动三类方法.

收稿日期 2020-12-24 录用日期 2021-03-19
Manuscript received December 24, 2020; accepted March 19, 2021
国家自然科学基金 (71701205, 62073333) 资助
Supported by National Natural Science Foundation of China (71701205, 62073333)
本文责任编委 魏庆来
Recommended by Associate Editor WEI Qing-Lai
1. 国防科技大学系统工程学院 长沙 410073 2. 北京科技大学人工智能研究院 北京 100083 3. 北京科技大学自动化学院 北京 100083
1. College of Systems Engineering, National University of Defense Technology, Changsha 410073 2. Institute of Artificial Intelligence, University of Science and Technology Beijing, Beijing 100083 3. School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083

第一类是基于数学求解的空战决策方法. 该方法最早可以追溯到上世纪 Isaacs[2] 提出的利用数学形式解解决追逐问题, 但 Isaacs 提出的方法缺乏严格的数学证明, 只适用于简单的空战场景[3]. 随着优化控制理论在 20 世纪 60 年代被提出, 学者们开始尝试用该理论解决空战决策问题. 早期的研究将空战问题简化为纯追逐问题[4] (即一方被指定为追逐者, 另一方为被追逐者, 空战过程中, 角色不发生更改), 在空战优化目标以及飞行动力学的限制下, 采用 Hamilton 方程求解. 进入 20 世纪 80 年代后, 战机与导弹性能显著提升, 传统纯追逐形式的空战被超视距空战替代, 敌我攻防角色转换频繁, 固定角色的纯追逐优化问题不再使用, 针对双目标优化的研究被大量开展[5−8]. 双目标分别是给定战场态势, 确定最终空战结局; 给定空战结局, 优化战机机动动作.

第二类是基于机器搜索的空战决策方法. 目前较为成熟可行的空战机动决策算法, 如影像图[9−10]、马尔科夫方法[11−12]、蒙特卡洛搜索[13]、矩阵决策[14−15]、决策树[16]、近似动态规划[17−18] 等, 均是基于类似思路展开的. 欧建军等[19] 引入偏好规划理论解决不确定环境下态势评估不准确的问题; 奚之飞等[20] 引入描述多目标威胁的威力势场理论来构建态势评价函数; 韩统等[21] 设计了一种协同威胁指数, 强调战机协同关系对战场态势的影响; 嵇慧明等[22] 结合距离、高度、速度、角度、性能要素构建战机综合优势函数; 王炫等[23] 建立进化式专家系统树框架; 周同乐等[24] 提出将战场态势与任务效益相结合的目标函数; 左家亮等[25] 利用深度神经网络的预测能力来启发决策序列搜索; 刘树林[26] 提出一种专家意见、会议判断与统计分析相结合的评价方法.

第三类是基于数据驱动的空战决策方法. 基于数据驱动的方法以神经网络技术为主, 该技术分为两类: 一类是将空战机动决策问题转变为分类 (模式识别) 问题, 输入实时战场态势, 输出战机采取的机动动作[27−28]; 另一类与前向搜索方法类似, 采用动态贝叶斯网络, 对不同战场态势下敌、我机机动动作的概率分布进行仿真、预测, 判定我机采取的动作[29−30]. 目前, 基于强化学习 (Reinforcement learning, RL) 的空战决策技术[31−33] 最为流行, 以 Q-learning 算法为例, 该技术重点研究 Q 值的设计方法, 目标是获得准确的战场态势到动作决策的映射关系.

上述三类研究方向也存在如下问题.

1) 基于数学求解的空战决策方法. 是最理想也是最难以实现的, 因为该方法要求严格的数学逻辑证明, 模型构建复杂. 仅针对较为简单的空战形式有效, 如规避导弹、拦截卫星等, 但当面临三维空间复杂机动的缠斗空战问题时, 适用性较为有限.

2) 基于机器搜索的空战决策方法. 本质在于解决任务规划、态势评估、目标分配等辅助决策问题, 遵循 “设计态势评估函数评价战场态势、使用智能优化算法搜索最优策略”的逻辑内核. 所以, 这类方法具有专家经验要求较高、态势评估函数设计复杂且粒度难以把握、机动动作策略库空间较小、优化算法搜索效率低、难以满足战场实时性决策的要求、场景简单且泛化性能差的通病.

3) 基于数据驱动的空战决策方法. 以强化学习为例, 很多研究只提到强化学习的概念, 本质上仍属于机器搜索的范畴, 仅利用神经网络的预测能力为优化搜索算法提供启发式经验; 一些研究仅适用于简单的一对一空战场景, 并且需要大量专家经验支撑, 如评价函数、态势估计、飞行动力学模型的设计等, 这类研究难以移植到复杂场景, 泛化性能较差; 一些研究虽然提出了多机协同的概念, 但只是简单地将多机问题分解为单机问题来解决, 较难提炼出协同战法.

鉴于上述不同方法的缺点, 本文提出一种 “集中式训练–分布式执行”的多机协同空战决策流程框架. 该框架不需要对空战环境以及战机飞行动力学进行建模、对专家经验的需求较小、具有实时决策的能力, 且本文提出的 4 种算法改进机制能有效提高模型训练的效率和稳定性, 实现了使用强化学习算法解决多机协同空战决策问题的技术途径.

本文首先从构建整个决策流程框架入手, 设计模型的训练与执行架构; 然后, 针对多机空战场景的特点, 设计了 4 种改进近端策略优化 (Proximal policy optimization[34], PPO) 算法的机制, 针对性提高了多机协同对抗场景下深度强化学习算法的效果; 最后, 在兵棋推演平台上仿真, 测试本文提出的决策流程框架以及改进算法的效果, 并总结模型涌现出的 5 种典型战法, 实验结果验证了本文方法的有效性和实用性.

1　深度强化学习背景知识

强化学习是机器学习的一个重要领域, 其本质是描述和解决智能体在与环境的交互过程中学习策略以最大化回报或实现特定目标的问题. 与监督学习不同, 强化学习中的智能体不被告知如何选择正确的动作, 而是通过智能体不断与环境交互试错, 从而学习到当前任务最优或较优的策略, 能够有效地解决在自然科学、社会科学以及工程应用等领域中存在的序贯决策问题.

现有强化学习方法利用马尔科夫决策过程 (Markov decision process, MDP) 从理论方面对 RL 问题进行基础建模. MDP 由一个五元组 $\langle S, A, R, T, \gamma\rangle$ 定义, 其中, S 表示由有限状态集合组成的环境; A 表示可采取的一组有限动作集; 状态转移函数 $T: S\times A\to\Delta(S)$ 表示将某一状态−动作对映射到可能的后继状态的概率分布, $\Delta(S)$ 表示状态全集的概率分布, 对于状态 s, $s'\in S$ 以及 $a\in A$, 函数 T 确定了采取动作 a 后, 环境由状态 s 转移到状态 s' 的概率; 奖赏函数 $R(s, a, s')$ 定义了状态转移获得的立即奖赏; γ 是折扣因子, 代表长期奖赏与立即奖赏之间的权衡.

近年来, 随着深度学习 (Deep learning, DL) 技术的兴起及其在诸多领域取得的辉煌成就, 融合深度神经网络和 RL 的深度强化学习 (Deep reinforcement learning, DRL) 成为各方研究的热点. 同基本的强化学习方法相比, DRL 将深度神经网络作为函数近似和策略梯度的回归函数. 虽然使用深度神经网络解决强化学习问题缺乏较好的理论保证, 但深度神经网络的强大表现力使得 DRL 的结果远超预期, 并在战略博弈[35−36]、无人机控制[37]、自主驾驶[38] 和机器人合作[39] 等领域取得了较大突破.

在非凸优化的情况下, 梯度可以用数值方法或抽样方法计算, 但很难确定适当的迭代学习率, 需要随时间变化以确保更好的性能. 早期的强化学习研究在使用基于梯度的优化技术时也遇到了这样的困境, 为规避瓶颈, Schulman 等[40] 提出一种处理随机策略的信任域策略优化 (Trust region policy optimization, TRPO) 算法. 该算法在目标函数中考虑了旧策略和更新策略之间的 Kullback-Leibler (KL) 发散, 并能对每个状态点的 KL 发散进行有界处理. 该方法跳出了对学习率的修正, 使策略改进过程更加稳定, 理论证明该方法单调地增加了累积奖赏. 考虑到 TRPO 中二阶 Hessian 矩阵计算的复杂性, Schulman 等[34] 进一步发展了一阶导数 PPO 算法.

图 1 描述 PPO 算法中神经网络的更新流程. 训练时从经验回放库 (Replay buffer) 中选择一批样本 (Sample) 供网络参数更新. PPO 算法采用的是 Actor-Critic (AC) 框架, 包含两个网络. Actor 网络更新部分, 同 TRPO 方法一样, 定义了 surrogate 目标:

$$\max L^{\mathrm{CPI}}(\theta)=\max\hat{\mathrm{E}}_t[r_t(\theta)\hat{A}_t],\quad r_t(\theta)=\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)} \tag{1}$$

其中, $\hat{\mathrm{E}}_t$ 表示对 0 ~ t 区间求均值, π_θ 代表当前时刻的策略, $\pi_{\theta_{\mathrm{old}}}$ 代表上一时刻的策略, $\hat{A}_t$ 估计了动作 a_t 在状态 s_t 下的优势函数.

在 PPO 中, 对上述代理目标进行了裁剪:

$$L^{\mathrm{CLIP}}(\theta)=\hat{\mathrm{E}}_t[\min(r_t(\theta)\hat{A}_t,\ clip(r_t(\theta),1-\varepsilon,1+\varepsilon)\hat{A}_t] \tag{2}$$

$$clip(x,\ x_{\mathrm{MIN}},\ x_{\mathrm{MAX}})=\begin{cases}x, & \text{if } x_{\mathrm{MIN}}\le x\le x_{\mathrm{MAX}}\\ x_{\mathrm{MIN}}, & \text{if } x<x_{\mathrm{MIN}}\\ x_{\mathrm{MAX}}, & \text{if } x_{\mathrm{MAX}}<x\end{cases} \tag{3}$$

该目标 $L^{\mathrm{CLIP}}(\theta)$ 实现了一种与随机梯度下降兼容的信赖域修正方法, 并通过消除 KL 损失来简化算法以及减小适应性修正的需求.

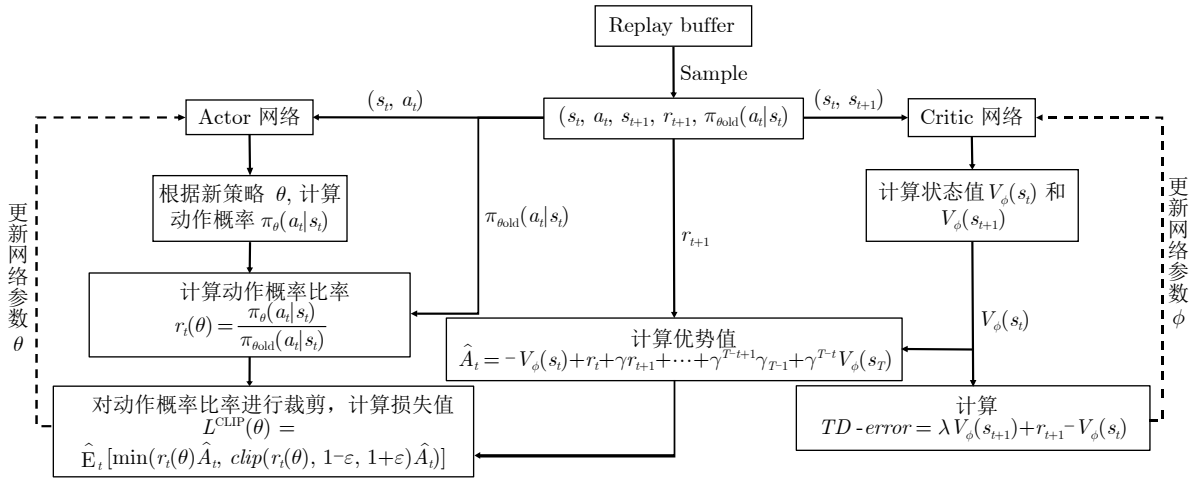


图 1 PPO 训练流程图

Fig. 1 PPO algorithm training flow chart

Critic 网络部分, 采用传统 *TD-error* 更新网络参数 ϕ, $V_\phi(s_t)$ 估计了状态 s_t 的状态价值函数.

2　多机协同空战决策流程设计

本节首先介绍多机协同空战决策流程的总体框架和 “训练–执行”架构, 之后在 PPO 算法基础上, 设计 4 种算法增强机制, 用于提升算法和整体框架的性能.

2.1　总体框架设计

图 2 是基于深度强化学习的多机协同空战决策流程框架 (Deep-reinforcement-learning-based multi-aircraft cooperative air combat decision framework, DRL-MACACDF). 整个框架共包括 5 个模块, 分别为态势信息处理模块、深度强化学习模块、策略解码模块、经验存储模块、神经网络训练模块.

框架的输入量是战场实时态势信息, 输出量是所控实体的动作决策方案. 原始战场态势信息输入框架后, 会首先经过态势信息处理模块进行加工, 数据经过提取、清洗、筛选、打包、归一化以及格式化表示后, 将传给深度强化学习模块; 深度强化学习模块接收态势信息数据, 输出动作决策; 策略解码模块接收深度强化学习模块的动作决策输出, 解码封装为平台环境可接受的操作指令, 对相应单元进行控制; 同时, 通过执行新动作获得的新的环境态势以及奖励值与本步决策的环境态势信息、动作决策方案一并被打包存储进经验存储模块; 待训练网络时, 再将这些样本数据从经验库中提取出来, 传入神经网络训练模块进行训练.

深度神经网络模块是整个框架的核心, 因为 PPO 算法收敛稳定、性能好, 并且其使用的一阶优化与剪切概率比率的方法操作简便, 适合在兵棋推演平台上进行多机协同对抗实验, 所以该模块选取 PPO 算法进行验证性实验. 本文重点在于对多机协同对抗问题进行抽象建模, 验证兵棋推演平台上使用强化学习算法解决该类问题的有效性, 所以文章没有对比众多算法的性能差异, 只是选取了其中较

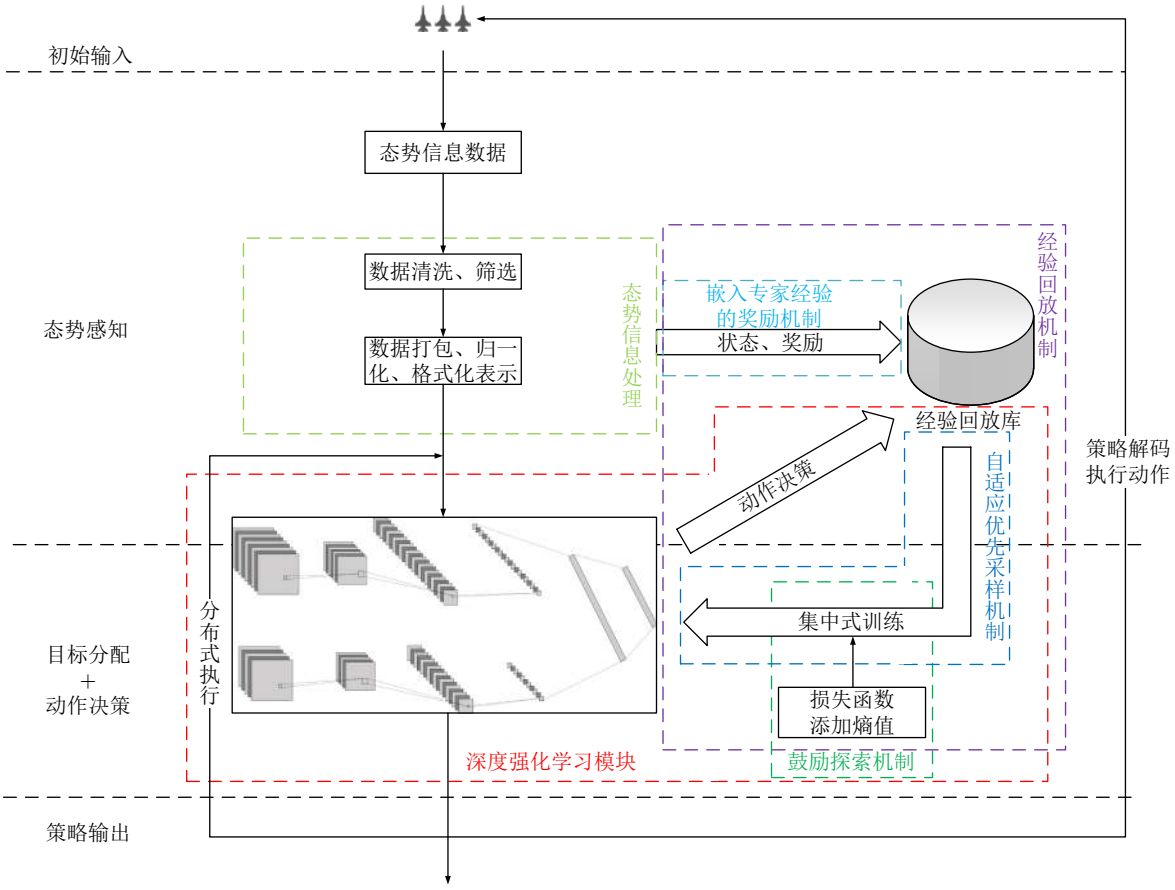


图 2　多机协同空战决策流程框架

Fig. 2　Multi-aircraft collaborative air combat decision framework

为先进的 PPO 算法举例.

2.2 集中式训练–分布式执行架构设计

在单智能体强化学习中, 环境的状态转移只与单智能体的动作有关, 而多智能体环境的状态转移依赖于所有智能体的动作; 并且, 在多智能体系统中, 每个智能体所获得的回报不只与自身的动作有关, 还与其他智能体有关. 通过学习改变其中一个智能体的策略将会影响其他智能体最优策略的选取, 且值函数的估计也将不准确, 这样将很难保证算法的收敛性. 因此, 我们采用集中式训练–分布式执行的架构, 如图 3.

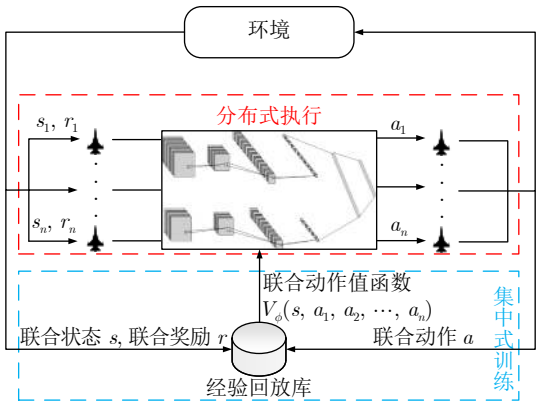


图 3 集中式训练–分布式执行架构

Fig. 3 Framework of centralized training and decentralized execution

“集中式训练”是指在训练中使用联合状态–动作值函数 $V_\phi(s, a_1, a_2, \cdots, a_n)$ 对智能体进行训练. 与分布式训练 (使用局部动作值函数 $V_\phi(s_i, a_i)$ 训练, 输入单个智能体的局部状态 s_i 和动作 a_i) 相比, 联合动作值函数输入的是全局态势信息 s 和所有实体的动作信息 $a_1 \sim a_n$, 是真正对于联合状态–策略的评估. 其优点在于所有实体共享一套网络参数, 在决策动作时能够考虑实体间的耦合关系, 因为整个系统的转移和回报函数的产生都与联合动作有关, 这样能有效防止一个实体的策略影响其他实体的策略, 解决算法较难收敛的问题.

然而, 在真正执行的时候, 单个智能体只能观测到部分信息 (包括部分的状态信息和动作信息), 无法获得其他智能体的动作, 甚至无法获得联合状态. 这种情况下, 输入决策网络的是单智能体的局部观测信息, 输出的是单智能体的决策动作, 这就是 “分布式执行”. 这种决策方式可以弥补不同实体间的动作区分度不大、模型探索性不强的缺点.

2.3 嵌入式专家经验奖励机制

空战决策问题是专业要求高、系统性较强的研究领域, 用于空战的战法战术及策略复杂且丰富, 所以专家经验在解决该问题过程中往往具有十分关键的作用. 传统奖励函数通常根据实体间交战的输赢判定得分, 将战损分值 $score_{\text{total}}$ 作为奖励值 r 用于网络训练.

$$r = score_{\text{total}} \tag{4}$$

但是, 这样设置的最大问题是奖励过于稀疏, 算法很难收敛. 为解决这一问题, 对奖励函数进行改进, 将专家经验嵌入奖励函数中. 考虑到战机靠近目标点时, 神经网络收到的奖励反馈应该变大, 所以在传统奖励函数基础上增加一个额外奖励项 $score_{\text{encourage}}$.

$$score_{\text{encourage}} = dis_{\text{cur}} - dis_{\text{next}} \tag{5}$$

其中, dis_{cur} 表示当前时刻战机距离目标点的距离, dis_{next} 表示下一时刻战机距离目标点的距离. 经过改进后的奖励函数则变为:

$$r = (1-\tau)\cdot score_{\text{total}} + \tau \cdot score_{\text{encourage}} \tag{6}$$

$$\tau = \frac{\tau - \tau_{\text{step}}}{\tau_{\text{temp}}} \tag{7}$$

式中, τ 是衰减系数, 随着训练的进行, 该值按照式 (7) 逐渐减小; τ_{step} 是递减步长; τ_{temp} 是衰减基数. 实验中的具体设置见附录表 A1.

嵌入专家经验的奖励函数, 在训练初期以额外奖励部分占主导, 引导战机飞往目标点. 随着训练迭代次数增加, 传统奖励渐渐占据主导, 侧重探索空战的战法战术.

使用强化学习解决问题, 很难设计一种放之四海而皆准的奖励函数, 需要具体问题具体分析. 本节提出的专家经验奖励机制的核心思想, 是在原有稀疏奖励的基础上, 人为添加一个稠密的奖励, 从而稠密化智能体获得的奖赏值, 加快智能体的训练速度. 上述专家经验奖励函数只是一种参考, 真正使用算法框架时, 还需要根据实际问题进行具体分析与设计.

2.4 自适应权重及优先采样机制

在经验回放库中采样时, 如果使用传统随机采样操作, 不仅无法有效利用高质量的样本, 还有可能导致模型陷入局部最优解. 另外, 回放库中的样本数量不断变化, 也不利于训练的收敛. 因此, 提出一种自适应权重以及优先采样的机制, 解决上述问题.

考虑到神经网络的损失函数受优势值影响, 在

设计自适应权重过程中, 提高优势值对采样权重的影响. 将参与采样的每个智能体产生的样本分别按照优势值的绝对值, 由大至小、从 1 到 N 进行排序. 考虑到全部样本的采样概率之和为 1, 设计如下样本自适应权重计算公式:

$$P_j = \frac{\frac{1}{j}}{\sum_{j=1}^{N} \frac{1}{j}} \tag{8}$$

其中, j 表示样本排序序号, P_j 表示第 j 号样本的采样概率, N 表示一个智能体包含的样本数量. 提出的自适应权重计算公式, 既增加了优势值绝对值较大样本的采样概率, 使奖励值极大或极小的样本都能影响神经网络的训练, 加快算法收敛速度; 又能充分发挥探索与利用的关系, 平衡不同样本采样概率.

采样时并非将经验回放库中的所有样本统一计算权重并采样, 而是不同智能体分别计算各自产生的样本的采样权重, 并按照该权重值分别采集预先设定数量的样本, 用于更新网络参数. 这种优先采样机制能够采集不同智能体产生的样本, 体现不同智能体间的合作关系, 促使不同智能体逐渐收敛到相同目标.

2.5　经验共享机制

由于多机空战场景的状态、动作空间庞大, 单个智能体能够探索的空间有限, 样本使用效率不高. 另外, 作为典型的多智能体系统, 多机协同空战问题中, 单个智能体的策略不只取决于自身的策略和环境的反馈, 同时还受到其他智能体的动作及与其合作关系的影响. 所以, 设计经验共享机制, 该机制包含共享样本经验库和共享网络参数两个方面.

所谓共享样本经验库, 是将全局环境态势信息 s_t、智能体的动作决策信息 a_t、智能体执行新动作后的环境态势信息 s_{t+1} 和环境针对该动作反馈的奖励值 r_{t+1} 按照四元组 $(s_t,\ a_t,\ s_{t+1},\ r_{t+1})$ 的形式存储进经验回放库, 每一个智能体的信息均按照该格式存储进同一个经验回放库中.

在更新网络参数时, 按照第 2.4 节所述机制从经验回放库中提取样本, 分别计算不同智能体产生的样本在 Actor 网络和 Critic 网络下的损失值, 进而求得两个神经网络的更新梯度 J_i. 将不同智能体的样本计算出的梯度值 J_i 进行加权, 可以得到全局梯度公式为:

$$J = \frac{1}{n}\sum_{i=0}^{n} w_i \cdot J_i \tag{9}$$

其中, J_i 表示第 i 个智能体样本计算出的梯度, n 表示样本总数, w_i 表示智能体 i 对全局梯度计算的影响权重. 这种不同智能体的样本共同更新同一套网络参数的机制称为 “共享网络参数”.

本文实验只涉及同构智能体, 可以使用所有实体的样本对共享策略网络进行训练. 当环境中存在异构实体时, 依然可以所有实体共享一套网络参数与经验池, 但需要在状态空间输入端、动作空间输出端、经验回放池采样方法上作出一些针对性处理. 例如, 定义总的状态空间维度, 囊括不同类别实体的所有状态, 每类实体只在其包含的状态维度上填充数据, 其余状态维度补零, 从而统一所有类别实体的网络输入维度; 同理, 输出端也定义总的动作空间维度, 囊括不同类别实体的所有动作, 每类实体进行决策时, 在输出端添加 mask 操作, 实体具有的动作维度正常输出, 不具有的动作维度补零, 再对非零维度的输出进行 softmax 操作, 按照其概率选择动作; 经验回放池中的样本则可以添加实体类别的标签, 在采样时, 均匀采集不同类别实体的样本. 理论上, 按照上述方法, 深度神经网络能够具备决策不同类别实体的能力. 由于篇幅有限, 本文不对其进行详细建模.

2.6　鼓励探索机制

多机交战的策略与战术战法构成丰富、种类多样、风格多变, 即便在有限动作空间下, 依旧具有涌现出丰富战法的潜力. 如果采用传统 PPO 算法的损失函数, 训练中后期智能体的探索能力会显著下降. 如何在算法收敛速度与智能体探索能力之间权衡是值得思考的问题.

为解决上述问题, 设计一种基于策略熵的鼓励探索机制, 增强智能体的探索能力, 并加快执行器网络的收敛速度.

不同智能体添加策略熵后的损失函数定义为:

$$L_i = L^{\mathrm{CLIP}}(\theta_i) + H_{\theta_i}(\pi(\cdot|s_t)) \tag{10}$$

其中, 下标 i 表示第 i 个智能体; θ_i 表示网络参数; $L^{\mathrm{CLIP}}(\theta_i)$ 为传统 PPO 算法的损失函数, 计算方法如式 (2) 所示; $H_{\theta_i}(\pi(\cdot|s_t))$ 表示在参数 θ_i 下策略 $\pi(\cdot|s_t)$ 的策略熵, 具体计算方法如式 (11):

$$H_{\theta_i}(\pi(\cdot|s_t)) = -\sum_{a_t \in A} \pi_{\theta_i}(a_t|s_t)\ln \pi_{\theta_i}(a_t|s_t) \tag{11}$$

本文出现的策略熵权重默认为 1, 因此没有在公式中另行标注.

本文针对的问题背景是多机协同空战决策, 主要强调不同实体间的配合协作. 所以在计算损失函

数时, 不是直接计算全局损失, 而是结合优先采样机制和经验共享机制, 求解不同智能体各自产生的样本的损失值. 相应地, 其策略熵也单独计算, 最后计算均值 E_i 作为全局损失函数值.

包含鼓励探索机制的损失函数如式 (12):

$$L = \mathrm{E}_i[L_i] = \frac{1}{M}\sum_{i=1}^{M} L_i \tag{12}$$

其中, M 是智能体总数.

3 仿真实验及结果

本文实验平台为 "墨子•未来指挥官系统 (个人版)[1]", 该平台支持联合作战背景下的制空作战、反水面作战等多种作战样式的仿真推演, 适用于作战方案验证、武器装备效能评估、武器装备战法研究等. 实验台式机搭载的 CPU 为 i9-10900K、显卡为 NVIDIA GeForce RTX 3090、内存为 64 GB.

3.1 实验想定

实验想定如图 4 所示, 该想定中红蓝兵力配置相等, 各自包含 3 架战斗机和一个可起降飞机的基地, 想定范围为长 1 400 km、宽 1 000 km 的长方形公海区域.

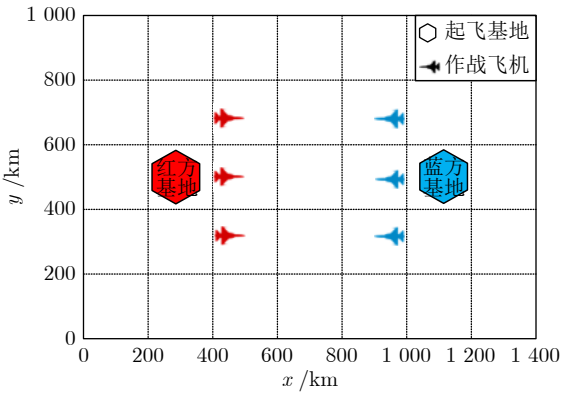


图 4 想定示意图

Fig. 4 Scenario diagram

想定推演的过程为飞机从基地起飞, 对己方基地进行护卫, 同时对敌方的战斗机和基地进行摧毁. 各个实体的具体型号和损失得分见附录表 A2 和表 A3.

3.2 模型构建

使用 PPO 算法构建强化学习智能体, 按照第 1 节所述方法, 对强化学习的要素进行定义.

[1] 版本号: v1.4.1.0

1) 状态设计

状态包含己方和敌方两部分实体信息, 己方实体信息包含己方飞机和导弹的信息, 敌方实体信息包含敌方飞机和导弹的信息. 由于战场迷雾, 己方和敌方同类型实体的状态信息可能不一致, 其中缺失的信息补零处理, 数据全部按照去量纲的方式进行缩放. 具体见附录表 A4.

2) 动作设计

本文决策的实体控制包含航向、高度、速度、自动开火距离、导弹齐射数量 5 个类. 由于武器数量有限, 当弹药耗尽时, 自动开火距离以及导弹齐射数量的决策将失效. 为降低决策动作的维度, 本文对航向、高度、速度和自动开火距离进行了离散化处理, 具体见附录表 A5. 动作空间维度为 6 × 3 × 3 × 6 × 2 共 648 维.

3) 奖励设计

奖励包含两个部分, 一部分是稀疏的战损奖励, 另一部分是嵌入式专家经验奖励. 如第 2.3 节所述的额外奖励思想, 本文采用的嵌入式奖励是战斗机距离敌方基地的距离减少量. 本文将战损得分与嵌入式专家经验奖励进行归一化, 防止变量量纲对计算结果的影响.

4) 网络设计

本文 Actor 网络与 Critic 网络结构大致相同. 其中, 全局态势信息以及实体个体态势信息分别经过多层归一化层、卷积层对特征进行压缩与提取, 将两部分获得的中间层信息进行拼接, 再经过全连接层后输出. Actor 网络输出 648 维动作概率分布, Critic 网络输出 1 维状态评价值. 神经网络示意图见附录图 A1.

5) 超参数设计

实验过程中涉及的各种超参数设置见附录表 A1.

3.3 算法有效性检验

为验证本文所提出的算法框架的有效性, 根据上述设计方法进行对比实验, 分别记录 DRL-MACACDF 模型、传统 PPO 算法模型、人类高级水平的模型与传统规划方法对战的得分曲线, 如图 5. 其中, 传统 PPO 算法也采用了 "集中式训练–分布式执行" 框架; 传统规划方法采用的是人工势场避障算法与 0-1 规划相结合的规则模型; 人类高级水平数据来源于第三届全国兵棋推演大赛决赛前三名选手的比赛模型的平均得分.

从图 5 可以看出, 随着训练次数的增多, DRL-MACACDF 模型的得分曲线逐步上升, 经过大约

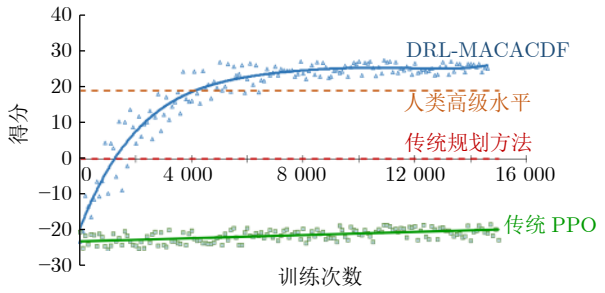


图 5　算法有效性对比图

Fig. 5　Algorithm effectiveness comparison diagram

1 000 轮训练后, 超越了传统规划算法水平. 模型大约在 6 000 轮左右开始收敛, 得分达到最大值. 相比而言, 传统 PPO 算法效果较差, 得分始终为负, 远不及传统规划算法水平线. 表 1 是 DRL-MACACDF 和传统 PPO 算法的实验数据统计.

表 1　算法有效性实验数据统计

Table 1　Experimental statistics of algorithm effectiveness

算法	平均得分	得分标准差	平均胜率 (%)
DRL-MACACDF	18.929	10.835	91.472
PPO	−21.179	1.698	0

从表 1 可以看出, 在 15 000 轮训练中, 本文提出的 DRL-MACACDF 算法平均胜率高达 91.472 %, 而传统 PPO 算法平均胜率仅为 0, 性能结果对比十分鲜明. 虽然 DRL-MACACDF 模型的得分标准差偏高, 但模型在经过训练后, 对战能力迅速提升, 比赛得分快速变化, 所以造成了高标准差. 当经过 6 000 轮训练, DRL-MACACDF 模型开始收敛后, 重新计算 DRL-MACACDF 模型的得分标准差则仅有 1.313, 反映出该模型性能稳定, 波动较小. 实验结果证明, 未加改进且缺乏专家经验的传统 PPO 算法难以解决多机协同对抗决策问题, 算法效果比传统规划算法效果还差. 相较而言, 本文提出的 DRL-MACACDF 算法及决策框架, 实验效果超过了传统强化学习算法和传统规划算法, 性能良好且效果稳定, 验证了算法框架的有效性.

3.4　消融实验

本文提出的算法框架包含 4 种针对多机协同对抗对策问题背景的改进机制. 为研究不同机制对算法性能的影响, 设计消融实验, 通过在传统 PPO 算法上增减 4 种改进机制, 比较不同模型的效果. 经过简单试验发现, 在未使用嵌入式专家经验奖励机制的情况下, 不同模型的得分都很低, 其他机制对算法性能的影响效果难以观察. 因此, 消融实验改为在 DRL-MACACDF 模型基础上分别去除某一机制, 根据实验结果间接比较不同机制的作用. 4 种对比算法的设置如表 2 所示.

表 2　消融实验设置

Table 2　The setting of ablation experiment

模型	嵌入式专家经验奖励机制	经验共享机制	自适应权重及优先采样机制	鼓励探索机制
DRL-MACACDF	●	●	●	●
DRL-MACACDF-R	○	●	●	●
DRL-MACACDF-A	●	○	●	●
DRL-MACACDF-S	●	●	○	●
DRL-MACACDF-E	●	●	●	○

注: ● 表示包含该机制, ○ 表示不包含

图 6 是消融实验算法性能对比曲线, 在传统 PPO 算法基础上增加任意三种增强机制对实验性能均有一定程度的提高, 由于作用机制不同, 其影响程度也存在差别. 具体来看, 未添加嵌入式专家经验奖励机制的 DRL-MACACDF-R 模型性能最差, 仅稍优于传统 PPO 算法, 所以说专家经验在强化学习中的指导意义巨大, 可以给实验性能带来显著提升; 未添加经验共享机制的 DRL-MACACDF-A 模型与 DRL-MACACDF 模型学习曲线大致相当, 但收敛速度相对较慢, 且最终收敛得分稍低. 无自适应权重及优先采样机制和无鼓励探索机制的模型性能依次降低, 其中未添加鼓励探索机制的 DRL-MACACDF-E 模型前期性能提升较快, 但大约在 6 000 轮左右就开始收敛并陷入局部最优, 最终落后于未添加自适应权重及优先采样机制的 DRL-MACACDF-S 曲线; DRL-MACACDF-S 模型, 前期收敛速度很

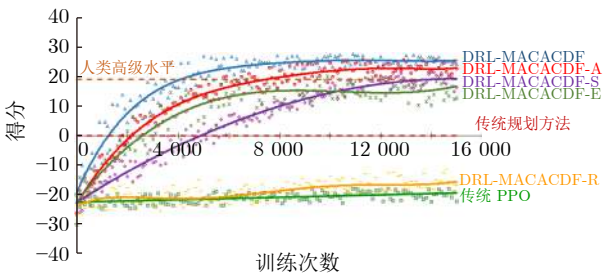


图 6　消融实验算法性能对比图

Fig. 6　Performance comparison diagram of ablation experimental algorithm

慢, 但模型性能一直在提高, 训练到 10 000 轮左右时, 反超 DRL-MACACDF-E 模型. 表 3 统计了 4 种对比算法相较于传统 PPO 算法平均得分提高的百分比.

消融实验证明, 在解决本文设计的 3V3 多机协同空战背景的问题时, 本文提出的 DRL-MACACDF 算法框架中添加的 4 种创新增强机制均能提高算法性能, 适用于解决多机协同空战决策问题.

3.5 算法效率分析

算法效率的高低是评价算法优劣的重要指标, 模型训练过程中的累计胜率曲线, 反映了算法的学习效率. 其导数为正值时, 代表模型性能正在提高, 胜利次数不断增多; 曲线斜率越大, 则学习效率越高. 如图 7 可以看出, 实验开始时算法更新迅速, 模型性能提升较快, 经过 2 500 轮左右的训练, 累计胜率就达到了 50 %; 至 6 000 轮左右时, 已经基本完成训练, 更新效率开始下降, 模型趋于收敛.

进一步, 分别抽取经过 500 轮、1 000 轮、2 000 轮、5 000 轮以及 10 000 轮训练的模型进行交叉对抗, 统计 100 局对抗的平均胜率, 绘制胜率分布图 (如图 8 所示).

从图 8 的渐变可以看出, 随着训练进行, 模型性能呈现明显的变化趋势. 以最左侧列为例, 从

表 3 消融实验数据统计

Table 3 Statistics of ablation experimental results

模型	平均得分	平均得分比传统 PPO 提高百分比 (%)	平均胜率 (%)
RL-MACACDF-R	−19.297130	8.327	0
RL-MACACDF-A	13.629237	154.019	86.774
RL-MACACDF-S	5.021890	115.934	66.673
RL-MACACDF-E	8.973194	133.417	82.361

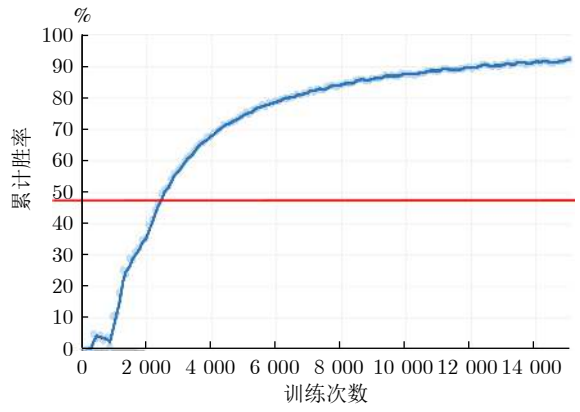


图 7 累计胜率曲线

Fig. 7 Cumulative winning rate curve

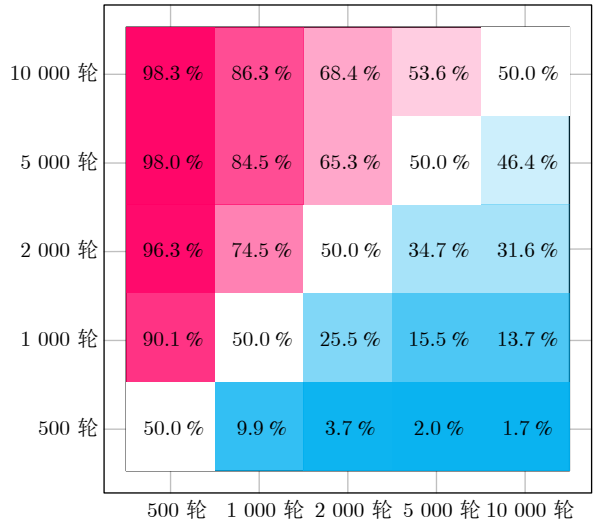


图 8 胜率分布图

Fig. 8 Winning rate distribution map

500 轮训练增加到 1 000 轮训练, 新模型就能以高达 90.1 % 的概率赢得胜利, 相较于传统强化学习算法, 本文模型能够以很快的速度提升决策能力, 随着训练次数增加, 模型基本上能以接近 100 % 的概率获胜. 由第 4 列可知, 当训练从 5 000 轮增加到 10 000 轮, 模型仅有 53.6 % 的概率获胜, 此时胜负基本上是按照相等概率随机分布的. 这说明当达到最优解时, 模型收敛稳定, 且性能不会有大幅度的波动.

3.6 行为分析

复盘实验数据, 总结交战过程中 DRL-MACACDF 模型涌现出的作战意图、策略、战术与战法.

1) 双机与三机编队战术

智能体涌现出自主编队能力, 如图 9、图 10 所示, 从基地起飞后, 智能体会随机采取双机编队或者三机编队前往作战区域. 当使用双机编队时, 通常智能体会选择从南北两路分别前往作战区域包围敌方飞机; 而采用三机编队时, 智能体更倾向于从中路挺进, 高速机动至交战区主动迎敌.

2) 包夹战术

如图 11 所示, 在与敌方飞机对抗时, 智能体常常会使用包夹战术. 两架战斗机同时从两个方向对敌方飞机发起攻击, 充分发挥飞机数量优势, 与敌方战机进行缠斗. 这种包夹战术表明, 智能体已经具备控制多机、探索和实现复杂战法的能力.

3) 充分发挥武器射程优势

如图 12 所示, 经过训练的智能体学会充分利

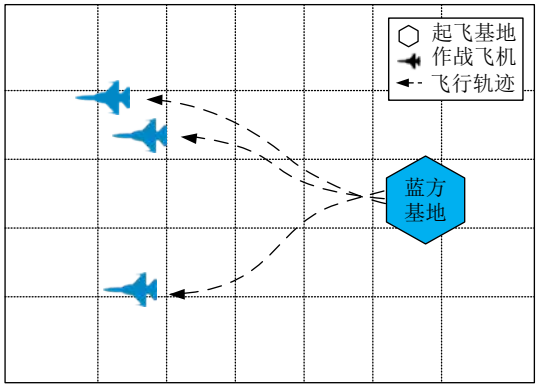


图 9　双机编队

Fig. 9　Two-plane formation

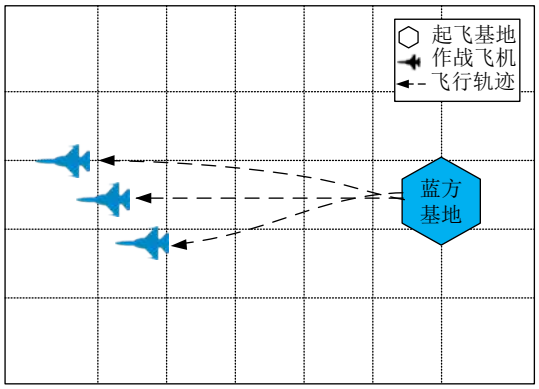


图 10　三机编队

Fig. 10　Three-plane formation

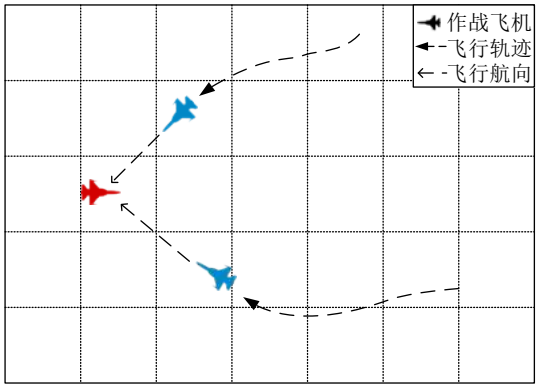


图 11　包夹战术

Fig. 11　Converging attack

用武器的有效射程, 在敌方飞机进入导弹射程后, 立即发射导弹进行攻击, 随后调头脱离敌方飞机攻击范围. 如果导弹未击落敌机, 则再次靠近敌方飞机, 重新组织进攻. 该战术动作既能有效节约弹药, 充分发挥导弹效能, 又能最大限度减少己方伤亡.

4) 快速机动避弹动作

如图 13 所示, 经过仔细复盘战斗机空战中的机动动作, 发现智能体的行为涌现出一种明显的快速机动主动避弹的战术动作. 当敌方导弹临近己方战斗机时, 战斗机会迅速向垂直于导弹瞄准基线的方向机动, 之后再重新飞往目标点. 采用突然变向的战术动作, 大幅降低了战机被击落的概率, 经过统计, 初始模型中击落一架战机平均需要 1 ~ 2 枚弹, 使用经过训练的智能体进行避弹, 平均需要 4 ~ 5 枚弹.

5) 诱骗敌方弹药战法

另一个明显的战法是诱骗敌方弹药, 如图 14 所示, 智能体控制多架战机在敌方火力范围边界试探, 引诱敌方进行攻击. 当探测到敌方发射导弹对己方飞机攻击后, 会机动至敌方攻击范围外, 超出敌方导弹射程; 待失去导弹攻击的威胁后, 会再次进入敌方火力覆盖范围. 该策略可以同时控制多架战机诱骗敌方弹药, 能够在短时间内大量消耗敌方导弹.

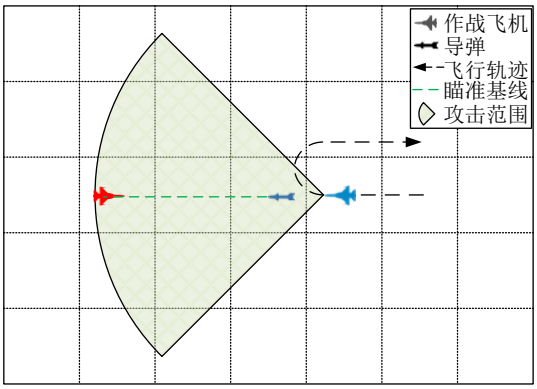


图 12　发挥射程优势

Fig. 12　Usage of maximum attack range

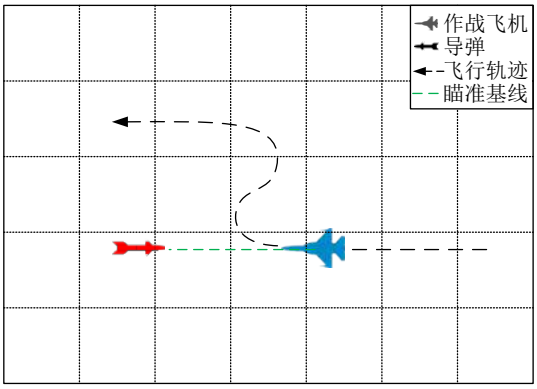


图 13　快速机动避弹

Fig. 13　Fast maneuvers to avoid attack

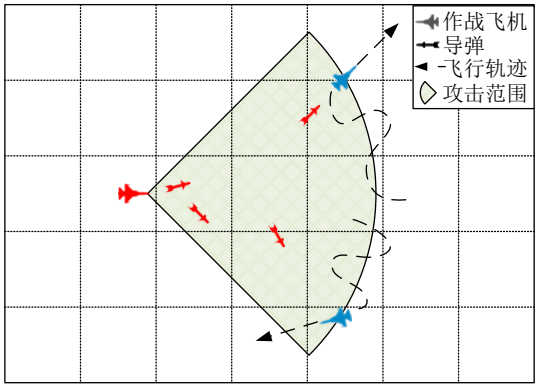


图 14 诱骗敌方弹药

Fig. 14 Consume enemy ammunition

4 结论

针对多机协同空战决策的问题背景, 本文提出一种 “集中式训练–分布式执行”的多机协同空战决策框架. 该框架内置深度强化学习模块, 并针对多机空战场景的特点, 设计了 4 种 PPO 算法改进机制, 针对性提高了多机协同对抗场景下深度强化学习算法的效果, 有效解决了多机协同空中作战实体类型众多、状态和动作空间巨大、协同合作关系复杂程度高等问题. 实验结果证明, 本文方法相较于传统规划算法和 PPO 算法具有明显优势, 进一步, 消融实验验证 4 种性能提升机制都不同程度地增强了算法性能, 并且算法效率较高, 能在有限的训练次数下达到良好的效果. 模型在训练过程中涌现出的大量鲜明的战术战法表明, 本文决策流程框架具有良好的探索能力, 能充分挖掘、利用多机空战场景下不同实体间协同合作的机制以及合作与竞争的战术战法, 在战场辅助决策领域具有巨大的应用价值.

本文重心在于抽象多机协同对抗问题, 构建适合强化学习算法求解的模型, 验证技术路径的可行性, 所以并未对不同强化学习算法进行对比分析. 在未来的工作中, 可以进一步拓展框架下的算法种类, 包括连续控制任务或者离散控制任务算法.

另外, 实验规模局限在 3V3 飞机空战, 还未验证大规模复杂场景下的算法性能. 下一步的研究可以将想定设计的更加贴合实战、更加复杂, 比如增添实体种类、增加实体数量、丰富作战任务等.

致谢

特别感谢梁星星、马扬对本文实验及文章撰写工作的支持.

附录 A

表 A1 实验超参数设置

Table A1 Experimental hyperparameter setting

参数名	参数值	参数名	参数值
网络优化器	Adam	经验库容量	3000 (个)
学习率	5×10^{-5}	批大小	200 (个)
折扣率	0.9	τ 初始值	1.0
裁剪率	0.2	τ_{step}	1×10^{-4}
训练开始样本数	1400 (个)	τ_{temp}	50000

表 A2 想定实体类型

Table A2 Entity type of scenario

单元类型	数量	主要作战武器
F/A-18 型战斗机	2	4 × AIM-120D 空空导弹 2 × AGM-154C 空地导弹
F-35C 型战斗机	1	6 × AGM-154C 空地导弹
基地	1	2 × F/A-18 型战斗机 1 × F-35C 型战斗机

表 A3 推演事件得分

Table A3 The score of deduction events

推演事件	得分
击毁一架飞机	139
损失一架飞机	−139
击毁基地	1843
损失基地	−1843

表 A4 状态空间信息

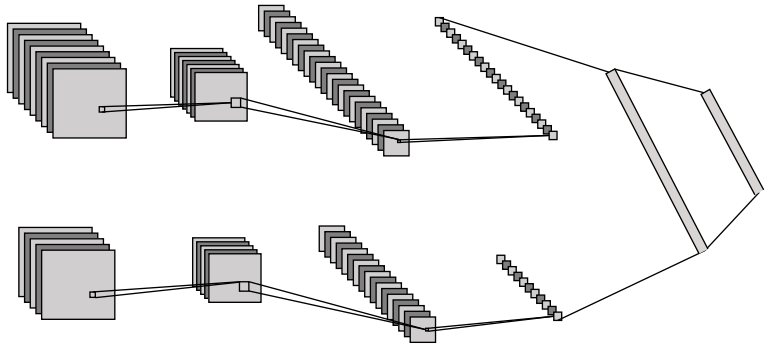
Table A4 State space information

实体	信息
己方飞机	经度、纬度、速度、朝向、海拔、目标点经度、目标点纬度等 7 维信息
己方导弹	经度、纬度、速度、朝向、海拔、打击目标的经度、打击目标的纬度等 7 维信息
敌方飞机	经度、纬度、速度、朝向、海拔等 5 维信息
敌方导弹	经度、纬度、速度、朝向、海拔等 5 维信息

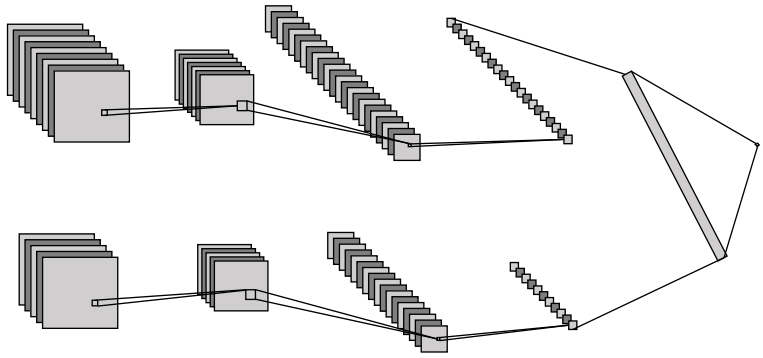
表 A5 动作空间信息

Table A5 Action space information

类别	取值范围
飞行航向	0°、60°、120°、180°、240°、300°
飞行高度	7620 米、10973 米、15240 米
飞行速度	低速、巡航、加力
自动开火距离	35 海里、40 海里、45 海里、50 海里、60 海里、70 海里
导弹齐射数量	1 枚、2 枚



(a) Actor 网络
(a) Actor neural network



(b) Critic 网络
(b) Critic neural network

图 A1 神经网络示意图

Fig. A1 Diagrams of neural network

References

1 Li Qing-Ying. Overview of collaborative air combat technology development and operational mode. *Science and Technology and Innovation*, 2020 (07): 124−126
(李卿莹. 协同空战技术发展概况及作战模式. 科技与创新, 2020 (07): 124−126)

2 Isaacs R. *Differential Games: A Mathematical Theory With Applications to Warfare and Pursuit, Control and Optimization*. North Chelmsford: Courier Dover Publications, 1999.

3 Yan T, Cai Y, Bin X U. Evasion guidance algorithms for air-breathing hypersonic vehicles in three-player pursuit-evasion games. *Chinese Journal of Aeronautics*, 2020, **33**(12): 3423−3436

4 Karelahti J, Virtanen K, Raivio T. Near-optimal missile avoidance trajectories via receding horizon control. *Journal of Guidance Control and Dynamics*, 2015, **30**(5): 1287−1298

5 Oyler D W, Kabamba P T, Girard A R. Pursuit-evasion games in the presence of obstacles. *Automatica*, 2016, **65**: 1−11

6 Li W. The confinement-escape problem of a defender against an evader escaping from a circular region. *IEEE Transactions on Cybernetics*, 2016, **46**(4): 1028−1039

7 Sun Q L, Shen M H, Gu X L, Hou K, Qi N M. Evasion-pursuit strategy against defended aircraft based on differential game theory. *International Journal of Aerospace Engineering*, 2019 (2019): 1−12

8 Scott W L, Leonard N E. Optimal evasive strategies for multiple interacting agents with motion constraints. *Automatica*, 2018, **94**: 26−34

9 Shao Jiang, Xu Yang, Luo De-Lin. Cooperative combat decision-making research for multi UAVs. *Information and Control*, 2018, **47**(03): 347−354
(邵将, 徐扬, 罗德林. 无人机多机协同对抗决策研究. 信息与控制, 2018, **47**(03): 347−354)

10 Virtanen K, Karelahti J, Raivio T. Modeling air combat by a moving horizon influence diagram game. *Journal of Guidance Control and Dynamics*, 2006, **29**(5): 1080−1091

11 Feng C, Yao P. On close-range air combat based on hidden markov model. In: Proceeding of the 2016 IEEE Chinese Guidance, Navigation and Control Conference. Piscataway, USA: IEEE, 2016. 687−694

12 Feng Chao, Jing Xiao-Ning, Li Qiu-Ni, Yao Peng. Theoretical research of decision-making point in air combat based on hidden markov model. *Journal of Beijing University of Aeronautics and Astronsutics (Natural Science Edition)*, 2017, **43**(3): 615−626
(冯超, 景小宁, 李秋妮, 姚鹏. 基于隐马尔科夫模型的空战决策点理论研究. 北京航空航天大学学报 (自然科学版), 2017, **43**(3): 615−626)

13 He Xu, Jing Xiao-Ning, Feng Chao. Air combat maneuver decision based on MCTS method. *Journal of Air Force Engineering University (Natural Science Edition)*, 2017, **18**(5): 36−41
(何旭, 景小宁, 冯超. 基于蒙特卡洛树搜索方法的空战机动决策. 空军工程大学学报 (自然科学版), 2017, **18**(5): 36−41)

14 Nelson R L, Rafal Z. Effectiveness of autonomous decision making for unmanned combat aerial vehicles in dogfight engagements. *Journal of Guidance Control and Dynamics*, 2018, **41**(4): 1021−1024

15 Xu Guang-Da, Lv Chao, Wang Guang-Hui, Xie Yu-Peng. Research on UCAV autonomous air combat maneuvering decision-

making based on bi-matrix game. *Ship Electronic Engineering*, 2017, **37**(11): 24−28
(徐光大, 吕超, 王光辉, 谢宇鹏. 基于双矩阵对策的 UCAV 空战自主机动决策研究. 舰船电子工程, 2017, **37**(11): 24−28)

16 Amnon K. Tree lookahead in air combat. *Journal of Aircraft*, 2015, **31**(4): 970−973

17 Ma Y F, Ma X L, Song X, Fei M R. A case study on air combat decision using approximated dynamic programming. *Mathematical Problems in Engineering*, 2014 (2014): 183401

18 Chen M, Zhou Z Y, Tomlin C J. Multiplayer reach-avoid games via low dimensional solutions and maximum matching. In: Proceeding of the 2014 American Control Conference. Piscataway, USA: IEEE, 2014. 1443−1449

19 Ou Jian-Jun, Zhang An. Target distribution model in cooperative air combat under uncertain environment. *Fire Control and Command Control*, 2020, **45**(5): 115−118
(欧建军, 张安. 不确定环境下协同空战目标分配模型. 火力与指挥控制, 2020, **45**(5): 115−118)

20 Xi Zhi-Fei, Xu An, Kou Ying-Xin, Li Zhan-Wu, Yang Ai-Wu. Decision process of multi-aircraft cooperative air combat maneuver. *Systems Engineering and Electronics*, 2020, **42**(2): 381−389
(奚之飞, 徐安, 寇英信, 李战武, 杨爱武. 多机协同空战机动决策流程. 系统工程与电子技术, 2020, **42**(2): 381−389)

21 Han Tong, Cui Ming-Lang, Zhang Wei, Chen Guo-Ming, Wang Xiao-Fei. Multi-UCAV cooperative air combat maneuvering decision. *Journal of Ordnance Equipment Engineering*, 2020, **41**(04): 117−123
(韩统, 崔明朗, 张伟, 陈国明, 王骁飞. 多无人机协同空战机动决策. 兵器装备工程学报, 2020, **41**(04): 117−123)

22 Ji Hui-Ming, Yu Min-Jian, Qiao Xin-Hang, Yang Hai-Yan, Zhang Shuai-Wen. Application of the improved BAS-TIMS algorithm in air combat maneuver decision. *Journal of National University of Defense Technology*, 2020, **42**(04): 123−133
(嵇慧明, 余敏建, 乔新航, 杨海燕, 张帅文. 改进 BAS-TIMS 算法在空战机动决策中的应用. 国防科技大学学报, 2020, **42**(04): 123−133)

23 Wang Xuan, Wang Wei-Jia, Song Ke-Pu, Wang Min-Wen. UAV air combat decision based on evolutionary expert system tree. *Ordnance Industry Automation*, 2019, **38**(01): 42−47
(王炫, 王维嘉, 宋科璞, 王敏文. 基于进化式专家系统树的无人机空战决策技术. 兵工自动化, 2019, **38**(01): 42−47)

24 Zhou Tong-Le, Chen Mou, Zhu Rong-Gang, He Jian-Liang. Attack-defense satisficing decision-making of multi-UAVs cooperative multiple targets based on WPS Algorithm. *Journal of Command and Control*, 2020, **6**(03): 251−256
(周同乐, 陈谋, 朱荣刚, 贺建良. 基于狼群算法的多无人机协同多目标攻防满意决策方法. 指挥与控制学报, 2020, **6**(03): 251−256)

25 Zuo Jia-Liang, Yang Ren-Nong, Zhang Ying, Li Zhong-Lin, Wu Meng. Intelligent decision-making in air combat maneuvering based on heuristic reinforcement learning. *Acta Aeronautica et Astronautica Sinica*, 2017, **38**(10): 217−230
(左家亮, 杨任农, 张滢, 李中林, 邬蒙. 基于启发式强化学习的空战机动智能决策. 航空学报, 2017, **38**(10): 217−230)

26 Liu Shu-Lin. A new method of evaluation. *Systems Engineering-Theory and Practice*, 1991, **11**(4): 63−66
(刘树林. 一种评价的新方法. 系统工程理论与实践, 1991, **11**(4): 63−66)

27 Zhang H P, Huang C Q, Zhang Z R, Wang X F, Han B, Wei Z L, et al. The trajectory generation of UCAV evading missiles based on neural networks. *Journal of Physics Conference Series*, 2020, **1486**(2020): 022025

28 Teng T H, Tan A H, Tan Y S, Yeo A. Self-organizing neural networks for learning air combat maneuvers. In: Proceeding of the 2012 International Joint Conference on Neural Networks. Piscataway, USA: IEEE, 2012. 2858−2866

29 Meng Guang-Lei, Ma Xiao-Yu, Liu Xin, Xu Yi-Min. Situation assessment for unmanned aerial vehicles air combat based on hybrid dynamic Bayesian network. *Command Control and Simulation*, 2017, **39**(04): 1−6, 39
(孟光磊, 马晓玉, 刘昕, 徐一民. 基于混合动态贝叶斯网的无人机空战态势评估. 指挥控制与仿真, 2017, **39**(04): 1−6, 39)

30 Yang Ai-Wu, Li Zhan-Wu, Xu An, Xi Zhi-Fei, Chang Yi-Zhe. Threat level assessment of the air combat target based on weighted cloud dynamic Bayesian network. *Flight Dynamics*, 2020, **38**(04): 87−94
(杨爱武, 李战武, 徐安, 奚之飞, 常一哲. 基于加权动态云贝叶斯网络空战目标威胁评估. 飞行力学, 2020, **38**(04): 87−94)

31 Yang Q, Zhang J, Shi G, Wu Y. Maneuver decision of UAV in short-range air combat based on deep reinforcement learning. *IEEE Access*, 2019, **PP**(99): 1−1

32 Liu P, Ma Y. A deep reinforcement learning based intelligent decision method for UCAV air combat. In: Proceeding of the 2017 Asian Simulation Conference. Berlin, Germany: Springer, 2017. 274−286

33 Zhou Y N, Ma Y F, Song X, Gong G H. Hierarchical fuzzy ART for Q-learning and its application in air combat simulation. *International Journal of Modeling Simulation and Scientific Computing*, 2017, **8**(04): 1750052

34 Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms [Online], available: https://arxiv.org/abs/1707.06347v2, August 28, 2017

35 Mnih V, Kavukcuoglu K, Silver D, Rusu A A, Veness J, Bellemare M G, et al. Human-level control through deep reinforcement learning. *Nature*, 2015, **518**(7540): 529−533

36 Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of go without human knowledge. *Nature*, 2017, **550**(7676): 354−359

37 Conde R, Llata J R, Torre-Ferrero C. Time-varying formation controllers for unmanned aerial vehicles using deep reinforcement learning [Online], available: https://arxiv.org/abs/1706.01384, June 5, 2017

38 Shalev-Shwartz S, Shammah S, Shashua A. Safe, multi-agent, reinforcement learning for autonomous driving [Online], available: https://arxiv.org/abs/1610.03295, October 11, 2016

39 Su P H, Gasic M, Mrksic N, Rojas-Barahona L, Ultes S, Vandyke D, et al. On-line active reward learning for policy optimization in spoken dialogue systems [Online], available: https://arxiv.org/abs/1605.07669v2, June 2, 2016

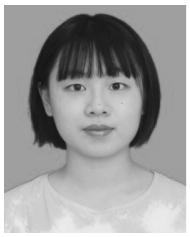
40 Schulman J, Levine S, Abbeel P, Jordan M, Moritz P. Trust region policy optimization [Online], available: https://arxiv.org/abs/1502.05477, April 20, 2017

施 伟 国防科技大学系统工程学院硕士研究生. 2019 年获得国防科技大学学士学位. 主要研究方向为层次强化学习, 多 agent 智能规划和多 agent 深度强化学习.
E-mail: shiwei15@nudt.edu.cn

(**SHI Wei** Master student at the College of Systems Engineering, National University of Defense Technology. He received his bachelor degree from National University of Defense Technology in 2019. His research interest covers hierarchical reinforcement learning, multi-agent intelligence planning, and multi-agent deep reinforcement learning.)

冯旸赫 国防科技大学系统工程学院副教授. 获得国防科技大学硕士、博士学位. 主要研究方向为因果发现与推理, 主动学习和强化学习. 本文通信作者.
E-mail: fengyanghe@nudt.edu.cn
(**FENG Yang-He** Associate professor at the College of Systems Engineering, National University of Defense Technology. He received his master and Ph. D. degrees from National University of Defense Technology. His research interest covers the casual discovery and inference, active learning, and reinforcement learning. Corresponding author of this paper.)

程光权 国防科技大学系统工程学院副研究员. 主要研究方向为链路预测.
E-mail: cgq299@163.com
(**CHENG Guang-Quan** Associate research fellow at the College of Systems Engineering, National University of Defense Technology. His main research interest is link prediction.)



黄红蓝 国防科技大学系统工程学院博士研究生. 主要研究方向为主动学习, 元学习.
E-mail: huanghonglan17@nudt.edu.cn
(**HUANG Hong-Lan** Ph. D. candidate at the College of Systems Engineering, National University of Defense Technology. Her research interest covers active learning and meta learning.)

黄金才 国防科技大学系统工程学院教授. 主要研究方向为智能调度与控制. E-mail: huangjincai@nudt.edu.cn
(**HUANG Jin-Cai** Professor at the College of Systems Engineering, National University of Defense Technology. His main research interest is intelligent scheduling and control.)

刘 忠 国防科技大学系统工程学院教授. 主要研究方向为智能规划与决策, 深度强化学习和多智能体系统.
E-mail: liuzhong@nudt.edu.cn
(**LIU Zhong** Professor at the College of Systems Engineering, National University of Defense Technology. His research interest covers intelligent planning and decision-making, deep reinforcement learning, and multi-agent system.)

贺 威 北京科技大学人工智能研究院、自动化学院教授. 主要研究方向为机器人学, 振动控制和智能控制系统. E-mail: weihe@ieee.org
(**HE Wei** Professor at the Institute of Artificial Intelligence and School of Automation and Electrical Engineering, University of Science and Technology Beijing. His research interest covers robotics, vibration control, and intelligent control system.)