控制与决策
Control and Decision

基于近端策略优化的动态武器目标分配

王晴, 王浩然, 辛斌, 张佳

引用本文:
王晴, 王浩然, 辛斌, 等. 基于近端策略优化的动态武器目标分配[J]. 控制与决策, 2026, 41(4): 919-930.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2025.0910

基于近端策略优化的动态武器目标分配

王 晴[1,2]，王浩然[1]，辛 斌[1,2†]，张 佳[1,2]

(1. 北京理工大学 自动化学院， 北京 100081；2. 自主智能无人系统全国重点实验室， 北京 100081)

摘 要: 现代战场环境下的动态传感器-武器-目标分配 (SWTA) 问题具有高动态、强对抗的特点, 传统静态分配方法难以适应战场态势的快速演化, 存在求解效率低、环境适应性差等局限. 鉴于此, 提出一种基于近端策略优化 (PPO) 的动态 SWTA 方法, 融合 OODA(观察-判断-决策-行动) 循环理论, 构建符合实际作战场景的传感器探测概率模型与武器毁伤概率模型, 通过 PPO 算法实现智能体与环境的持续交互与策略优化, 在决策过程中统筹作战效能与资源消耗. 实验结果表明, 该方法在多种弹药目标比场景下均表现出优越性能, 能够显著提升系统整体作战的效能与资源利用率. 所提出方法为动态 SWTA 问题提供了一种高效、自适应的智能决策框架, 推动了指挥决策的智能化进程, 具备较强的实际应用潜力.

关键词: 传感器-武器-目标分配；近端策略优化；OODA 环；动态决策；强化学习；Actor-Critic 架构

中图分类号: E91；TP18 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.0910

引用格式: 王晴, 王浩然, 辛斌, 等. 基于近端策略优化的动态武器目标分配 [J]. 控制与决策, 2026, 41(4): 919-930.

Dynamic weapon-target assignment based on proximal policy optimization

WANG Qing[1,2]，*WANG Hao-ran*[1]，*XIN Bin*[1,2†]，*ZHANG Jia*[1,2]

(1. School of Automation， Beijing Institute of Technology， Beijing 100081， China；2. State Key Laboratory of Autonomous Intelligent Unmanned Systems， Beijing 100081， China)

Abstract: The dynamic sensor-weapon-target assignment (SWTA) problem in modern battlefield environments is characterized by high dynamism and strong adversariality. Traditional static assignment methods struggle to adapt to the rapidly evolving battlefield situation due to their low solving efficiency and inadequate environmental adaptability. To address these challenges, this paper proposes a dynamic SWTA method based on proximal policy optimization (PPO). By integrating the OODA (observe–orient–decide–act) loop theory, the method constructs realistic sensor detection probability and weapon kill probability models. Through continuous interaction between the agent and the environment, the PPO algorithm optimizes the strategy while balancing operational effectiveness and resource consumption. Experimental results demonstrate that the proposed method achieves superior performance across various ammo-to-target ratio scenarios, significantly enhancing both resource utilization and overall operational effectiveness. This study provides an efficient and adaptive intelligent decision-making framework for the dynamic SWTA problem, advancing the process of command and decision-making, with strong potential for practical application.

Keywords: sensor-weapon-target assignment； proximal policy optimization； OODA loop； dynamic decision-making；reinforcement learning；Actor-Critic framework

0 引 言

传感器-武器-目标分配问题是现代战争中的关键决策环节, 其核心是在复杂多变的战场环境下, 将有限的传感器探测资源与武器打击资源进行协同整合, 从而实现对多类型来袭目标的最优拦截与打击效果[1-2]. 深入研究该问题对于最大化体系作战效能、推动指挥决策向自动化与智能化转型具有重大的理论与应用价值[3].

武器目标分配问题的研究经历了从静态模型到动态模型的演进[4-6]. 早期研究主要依赖于经典的组合优化算法, 例如匈牙利算法[7]. 20 世纪 80 年代, Lloyd 等[8] 从计算复杂性理论角度证明了武器目标

收稿日期: 2025-09-01；**录用日期:** 2025-10-16.

基金项目: 北京市自然科学基金面上项目 (4252050)；国家自然科学基金青年科学基金项目 (A 类)(62425304)；国家自然科学基金基础科学中心项目 (62088101).

责任编委: 王凌.

†通信作者. E-mail: brucebin@bit.edu.cn.

分配问题是一个多参数、多约束的NP完全问题, 这一定性深刻影响了后续研究方法的发展. 随着计算能力的提升, 枚举法、分支定界法[9]等精确算法被用于求解小规模静态问题. 20世纪后期, 智能优化算法逐渐成为主流, 遗传算法(GA)[10-11]、粒子群算法(PSO)[12-14]和蚁群算法(ACO)[15]等仿生启发式算法, 通过模拟自然界的进化机制与群体智能, 实现了对解空间的高效搜索. 这些方法不仅能够处理动态决策场景, 还在多目标优化与实时响应方面取得了显著进展[16-19].

然而, 随着现代战场呈现出高强度、快节奏、高度不确定性的特点, 传统方法的局限性日益凸显. 遗传算法在面对大规模目标来袭场景时, 时间复杂度呈指数级增长, 难以满足秒级甚至毫秒级的实时决策要求[20]. 粒子群算法的性能表现严重依赖参数设置, 在动态环境下需要引入复杂的扰动检测与响应机制, 面对突发威胁时适应性不足[21]. 现代信息化战争对决策系统的实时性提出了极高要求[22], 而现有方法大多着眼于单一决策时刻的静态最优解, 缺乏对持续演变战场态势的感知与应对能力, 难以支撑整个作战过程的优化[23].

为应对上述挑战, 近年来研究者们致力于将新兴人工智能技术应用于动态传感器-武器-目标分配问题. 观察-判断-决策-行动(OODA)循环理论, 为动态不确定环境下的敏捷决策提供了经典理论框架[24]. 在此背景下, 以深度Q网络[25]为代表的强化学习方法, 通过构建智能体与环境交互的“状态-动作-奖励”闭环学习机制, 为实现动态策略优化提供了新途径. 图神经网络[26]通过消息传递机制能够有效捕捉节点之间的高阶关系, 聚合邻居节点的特征信息, 从而增强节点的表示能力. 注意力机制[27]能够对高维战场信息进行筛选, 实现威胁等级的精准评估. 尽管如此, 如何实现跨域异构信息的快速融合与高效利用, 并对动态变化的战场态势做出实时、精准的认知与响应, 仍然是当前研究中亟待突破的核心难点[28-29].

近年来, 深度强化学习在资源分配问题中的应用正处于快速发展阶段. 研究者通过引入多头Q网络[30]、融合DQN (deep Q-learning network, DQN)与启发式规则[31]或借助元启发式策略[32]引导智能体探索, 显著提升了模型的决策精度与学习效率. 伍国华等[33]基于任务分解策略将复杂的决策任务分解为子目标平台选择和子平台火力分配两个阶段, 通过融合启发式算法和强化学习模型进行求解. 孙昕等[34]针对防空资源分配问题, 在基于多目标进化算法基础上进行改进, 种群进化过程中自适应调整交叉与变异的概率以提高个体的质量, 最终得到一组可供决策者使用的最优解集. 李泷鑫等[35]针对多机器人除草任务分配问题, 提出了一种新颖的基于分组策略的多目标离散人工蜂群算法进行高效求解, 在解的质量、收敛速度和鲁棒性方面优于多种先进算法. 然而, 现有方法在面对高度动态且对抗激烈的环境时, 仍普遍存在策略适应性弱、泛化能力有限的问题. 此外, 其计算复杂度随问题规模增长而急剧上升, 易导致维度灾难[36], 制约了其在大型作战体系中的实际应用.

针对上述挑战, 本文研究动机在于构建一个能够深度融合态势感知、决策推理与实时动作的智能决策框架, 以解决现有方法在动态适应性、计算效率与抗干扰能力方面的不足. 基于此, 本文提出一种融合OODA循环理论与近端策略优化算法的智能决策框架, 旨在通过强化学习智能体与环境的持续交互与策略优化, 实时生成高效的传感器-武器-目标分配方案, 显著提升系统在高动态、强对抗战场环境下的决策适应性、鲁棒性和运行效率.

本文的主要贡献如下:

1) 针对动态适应性与抗干扰能力不足的问题, 与文献[5]相比, 本文构建一个由OODA循环理论指导的闭环决策框架. 该框架通过强化学习智能体与高仿真环境的持续交互, 能够实现对动态战场态势的在线感知与实时决策, 提升模型在突发威胁、目标信息变化下的策略适应性和鲁棒性.

2) 针对泛化能力有限的问题, 与文献[37]相比, 本文基于近端策略优化算法实现策略的自主与持续优化. 该算法通过其良好的探索-利用平衡特性, 能够学习到一个更具通用性的高层决策策略, 有效提升模型在不同场景下的泛化能力.

3) 针对计算效率低的问题, 本文通过设计高效的神经网络决策器替代传统基于搜索或迭代优化的决策机制, 能够在大规模问题中避免维度灾难, 在保证决策质量的同时满足现代战场对算法实时性的要求, 具备应用于大型作战体系的潜力.

1 问题定义与环境建模

本文研究的动态SWTA问题旨在通过合理分配传感器和武器资源, 在动态场景中最大化总体毁伤效能, 同时兼顾资源成本. 作战想定叙述如下: 在一定时间T内, 进攻方有E个目标来袭, 这些目标具有不同的初始位置、速度、威胁值等, 防御方有W个武器(如导弹)拦截目标, 有S个传感器跟踪和照亮

目标, 以便引导武器, 构成协同防御网络. 每个传感器与武器单元被赋予特定的性能指标, 如探测概率、毁伤概率和成本. 本文考虑的资源和目标是异构的, 不同类别的资源和目标属性不同. 因此, 武器和传感器的选择与分配不仅取决于它们的性能, 也与目标类别有关. 此外, 假设在每个决策时刻, 各目标只能分配一个传感器和武器[38], 若打击事件完成后该目标依旧存活, 则可再次进行分配.

防御方需在离散作战周期内建立决策模型, 并满足以下假设.

假设 1　作战资源假设. 假设问题模型由S个传感器、W个武器和E个目标组成, 传感器、武器和目标的数量随机生成.

假设 2　打击规则假设. 所提出传感器、武器均为最小分配单元, 实际场景中能探测多个目标的传感器以及能打击多个目标的武器, 可以被等效为多个独立的传感器和武器[5]. 在每个决策时刻, 各武器或传感器仅作用于一个目标[1], 且每个目标只能分配一个传感器和武器[38].

假设 3　作战事件独立性假设. 目标被传感器探测与武器打击时, 探测事件与打击事件在概率空间中相互独立[5,39].

结合实际作战场景, 既要考虑武器和传感器对目标的联合毁伤概率, 最大化整体作战效能, 还需考虑资源使用成本, 兼顾毁伤效能最大化与资源消耗最小化. 毁伤效能函数[1,39] 为

$$f_1(X)=\sum_{t=1}^{T}\sum_{e=1}^{E}\left(1-\prod_{w=1}^{W}\prod_{s=1}^{S}(1-p_{se}(t)\,q_{we}(t)\,x_{ews}(t))\right)h_e. \tag{1}$$

其中: T为时间区间; E、W、S分别为目标、武器和传感器的总数; $p_{se}(t)$为传感器s在时刻t对目标e的探测概率; $q_{we}(t)$为武器w在时刻t对目标e的毁伤概率; $x_{ews}(t)$为决策变量, 代表传感器s与武器w是否在时刻t分配给目标e; $x_{ews}(t)=x_{se}(t)\cdot x_{we}(t)$, $x_{se}(t)$代表传感器s是否在时刻t分配给目标e, $x_{we}(t)$代表武器w是否在时刻t分配给目标e; h_e为目标e的威胁值.

资源消耗函数为

$$f_2(X)=\sum_{t=1}^{T}\sum_{e=1}^{E}\sum_{w=1}^{W}\sum_{s=1}^{S}c_{ews}(t)x_{ews}(t), \tag{2}$$

其中$c_{ews}(t)$为时刻t传感器s与武器w对目标e的资源消耗.

综上, 本文建立的动态 SWTA 模型的目标函数为

$$\begin{cases}\max f_1(X),\\ \min f_2(X).\end{cases} \tag{3}$$

约束条件包括以下 4 类约束:

$$\sum_{e=1}^{E}x_{we}(t)\leqslant 1,\ \forall w\in\{1,2,\ldots,W\}, \tag{4}$$

$$\sum_{e=1}^{E}x_{se}(t)\leqslant 1,\ \forall s\in\{1,2,\ldots,S\}, \tag{5}$$

$$\sum_{w=1}^{W}x_{we}(t)\leqslant 1,\ \forall e\in\{1,2,\ldots,E\}, \tag{6}$$

$$\sum_{s=1}^{S}x_{se}(t)\leqslant 1,\ \forall e\in\{1,2,\ldots,E\}. \tag{7}$$

其中: 约束 (4) 表示在时刻t一个武器只能分配给一个目标, 约束 (5) 表示在时刻t一个传感器只能分配给一个目标, 约束 (6) 表示在时刻t一个目标只能分配一个武器, 约束 (7) 表示在时刻t一个目标只能分配一个传感器.

对该问题的求解依赖于以下先验知识和实时输入:

1) 目标信息. 目标总数E, 每个目标的属性包括其类型、实时位置、速度、威胁值.

2) 防御方资源信息. 包括传感器资源和武器资源: 传感器总数S, 每种传感器资源包括其类型、探测概率模型、使用成本; 武器总数W, 每种武器资源包括其类型、毁伤概率模型、使用成本.

3) 作战规则. 每个决策时刻的资源分配约束, 如式 (4) ∼ (7) 所示.

4) 实时态势. 在动态环境中, 以上大部分信息 (如目标位置、资源状态) 会随时间变化, 作为算法在每个决策时刻的输入状态.

模型求解的输出是一个分配方案, 具体表现为一个四维决策变量矩阵, 即

$$X=[x_{ews}(t)]_{E\times W\times S\times T}. \tag{8}$$

其中: $x_{ews}(t)$为决策变量, $x_{ews}(t)=1$代表在时刻t传感器s和武器w协同分配给目标e. 输出方案必须满足约束条件 (4) ∼ (7).

该问题是一个高维离散组合优化问题, 解空间随目标数、武器数、传感器数和时间呈指数级增长, 且具有强动态时变性, 战场态势 (如目标位置、资源状态) 实时变化, 要求算法能够在线实时决策, 而非离线一次性计算, 对该问题的求解已被证明是 NP 完全问题[8]. 综上, 动态 SWTA 问题是一个具有高维、强动态、强约束特性的组合优化问题, 求解难度较

高, 传统优化方法难以使用, 需采用智能决策算法来寻找高效、可行的近似最优解.

2 算法设计与实现

本文提出一种基于 PPO 算法的智能决策框架, 采用基于 Actor-Critic 的网络结构, 在保证性能提升的同时限制更新幅度以稳定训练过程. 算法框架如图 1 所示.

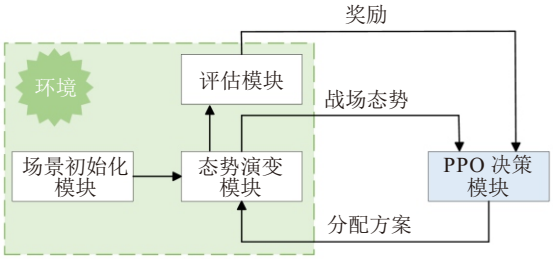


图1 智能决策框架

如图 1 所示, 本文提出的算法主要由环境模块与基于 PPO 算法的决策模块两部分组成. 下文将分别对这两个核心模块进行详细阐述.

2.1 环境设计与构建

本文基于 Python 构建了一个动态 SWTA 仿真环境, 环境的主要组成部分包括:

1) 场景初始化模块. 负责快速生成保卫要地并确定其位置, 配置传感器系统, 部署武器单元, 并设定目标的动态属性, 进而构建完整的战场态势与基础作战能力.

2) 态势演变模块. 该模块以$\Delta t = 1\,\mathrm{s}$作为演变基本单位, 实时更新目标与导弹位置、各类资源的状态等信息.

3) 决策模块. 该模块以基于 Actor-Critic 的网络结构作为基础, 使用 PPO 算法进行训练, 实时评估态势, 并对目标进行分配或采取“等待策略”, 提供灵活多样的决策支持.

4) 毁伤评估模块. 该模块包括传感器探测概率模型和武器毁伤概率模型.

① 传感器探测概率模型: 传感器探测概率结合目标与传感器的相对位置和目标速度方向进行计算[40], 公式如下:

$$P_{\text{final}} = P_s \cdot \min\left(1, \frac{T_{\text{stay}}}{T_{\text{radar}}}\right). \tag{9}$$

其中初始探测概率P_s表示为

$$P_s = \exp\left(-\left(\frac{R}{R_{\max}}\right)^4\right).$$

这里:R为目标与传感器的距离, $R_{\max}$为传感器的最大探测半径. 目标在传感器探测范围内的停留时间T_{stay}通过计算目标从当前点出发到探测圆边界的路径长度及速度决定. T_{radar}是传感器的准备时间, 通常与扫描周期和处理时间有关.

② 武器毁伤概率模型: 武器毁伤概率结合脱靶量ΔS与预期拦截时间t_{go}进行计算[41], 其公式如下:

$$\Delta S = \left| t_{\text{go}}^2 \cdot \exp\left(-\frac{t_{\text{go}}}{\tau}\right) \cdot \left(0.5 - \frac{t_{\text{go}}}{6\tau}\right)\right|.$$

其中: τ为导弹特性的时间参数, 与导弹的性能有关; t_{go}为导弹到目标的预期拦截时间. 最终的毁伤概率表示为

$$P_w = 1 - \exp\left(-\frac{\delta_0^2}{\Delta S^2}\right), \tag{10}$$

其中δ_0为目标毁伤特性参数, 与目标易损性相关.

③ 综合毁伤概率评估: 目标e的综合毁伤概率P_e通过传感器探测概率P_{final}和武器毁伤概率P_w计算, 综合毁伤概率表示为

$$P_e = P_w \cdot P_{\text{final}}. \tag{11}$$

在上述模块的基础上, 本文构建了一个完整作战仿真闭环流程, 其结构如图 2 所示. 该流程通过各模块间的协同与信息交互, 有效保障了实时决策过程的高效性与系统稳定性.

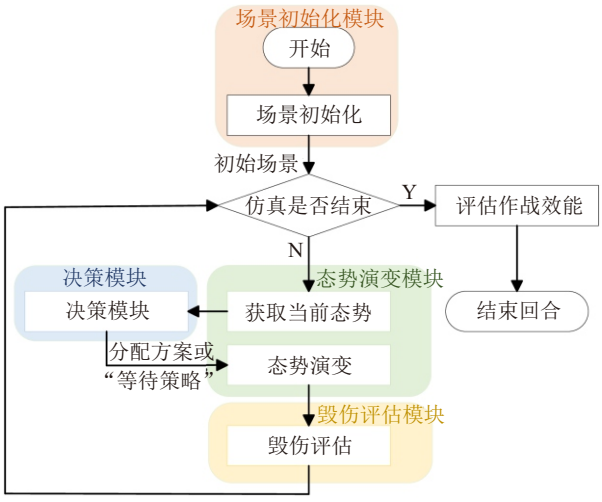


图2 仿真运行流程

2.2 强化学习算法设计

传统方法将 SWTA 视为一个静态的、一次性的组合优化问题, 本文的核心思路是借鉴 OODA 循环理论, 将问题重新定义为一个动态的决策过程. 具体而言, 将整个作战过程离散化为多个决策时刻t, 在每个决策时刻, 智能体观察当前战场状态$s(t)$, 并选择一个动作$a(t)$ (即一个分配指令或“等待策略”), 环境执行该动作后转移到新状态, 并给予智能体一个奖励$r(t)$. 智能体的目标是学习一个策略π, 以最大化整个回合的累积奖励.

这一转变使算法能够实时响应战场变化, 不再

追求“最优静态解”, 而是学习一个“最优决策策略”, 可以根据当前的态势 (如哪些目标威胁最大、哪些资源可用) 做出当前最有利的决策, 从而实现全局自适应优化. 算法的核心要素包括状态空间、动作空间和奖励机制的设计.

1) 状态空间.

状态向量通过多维特征全面描述战场环境, 包括目标特征、武器状态、传感器状态和战场态势等关键信息. 在某个决策时刻t_k, 状态空间定义为

$$s(t_k)=(\bar{P}_s(t_k),\bar{P}_w(t_k),M(t_k),N(t_k),Q(t_k)). \tag{12}$$

$\bar{P}_s(t_k)$代表当前已使用的传感器的平均能力, 其计算方式为当前时刻t_k已使用的传感器的探测概率和与已分配的目标个数Z的比值, 即

$$\bar{P}_s(t_k)=\sum_{t=1}^{t_k}\sum_{s=1}^{S}\sum_{e=1}^{E}p_{se}(t)x_{se}(t)\Big/Z.$$

其中: $p_{se}(t)$为时刻t传感器s对目标e的探测概率, $x_{se}(t)$表示传感器s是否在时刻t分配给目标e.

$\bar{P}_w(t_k)$代表当前已使用的武器的平均能力, 其计算方式为当前时刻t_k已使用的武器的毁伤概率和与已分配的目标个数Z的比值, 即

$$\bar{P}_w(t_k)=\sum_{t=1}^{t_k}\sum_{w=1}^{W}\sum_{e=1}^{E}q_{we}(t)x_{we}(t)\Big/Z.$$

其中: $q_{we}(t)$为时刻t武器w对目标e的毁伤概率, $x_{we}(t)$表示武器w是否在时刻t分配给目标e.

$M(t_k)=(m_{ac}(t_k))_{N_s\cdot N_e}$代表当前传感器目标的分类别分配情况, N_s、N_e代表传感器和目标的类型数目, $m_{ac}(t_k)$代表当前时刻t_k, 第a类传感器分配给第c类目标的个数与第a类传感器总数的比值, 即

$$m_{ac}(t_k)=\sum_{t=1}^{t_k}\sum_{s=1}^{S_a}\sum_{e=1}^{E_c}x_{se}(t)\Big/S_a.$$

其中: $x_{se}(t)$代表传感器s是否在时刻t分配给目标e, S_a为第a类传感器的总数, E_c为第c类目标的总数.

$N(t_k)=(n_{bc}(t_k))_{N_w\cdot N_e}$代表当前武器目标的分类别分配情况, N_w、N_e代表武器和目标的类型数目, $n_{bc}(t_k)$代表当前时刻t_k, 第b类武器分配给第c类目标的个数与第b类武器总数的比值, 即

$$n_{bc}(t_k)=\sum_{t=1}^{t_k}\sum_{w=1}^{W_b}\sum_{e=1}^{E_c}x_{we}(t)\Big/W_b.$$

其中: $x_{we}(t)$代表武器w是否在时刻t分配给目标e, W_b为第b类武器的总数, E_c为第c类目标的总数.

$Q(t_k)$表示当前资源的情况, 有

$$Q(t_k)=(c_s(t_k),c_w(t_k),c_e(t_k),h_e,\text{stage},u_w(t_k),u_s(t_k)).$$

其中: $c_s(t_k)=(1,0,\ldots,1)_{N_s}$代表每种传感器的剩余情况, 若某种传感器有剩余则对应位置为“1”; $c_w(t_k)=(1,0,\ldots,1)_{N_w}$代表每种武器的剩余情况, 若某种武器有剩余则对应位置为“1”; $c_e(t_k)$为当前决策目标的 one-hot 编码; h_e为当前目标的威胁值; stage为当前的态势时刻; $u_w(t_k)$为当前每种武器对该目标的毁伤概率; $u_s(t_k)$为每种传感器对该目标的探测概率.

状态向量考虑了实时态势 (如当前目标威胁值、资源剩余情况)、历史信息 (如各类传感器或武器对各类目标的已分配比例) 和物理特性 (如当前分配资源的毁伤或探测概率), 使智能体能够感知战场态势, 并捕捉资源与目标类型的匹配关系, 为智能体做出明智决策提供信息基础.

2) 动作空间.

动作空间定义了决策网络可采取的所有离散动作, 包括选择哪一种类型的武器和传感器分配给某一种目标, 以及“等待策略”. 不同类型的资源对目标的能力不同, 可以使智能体具备不同的决策偏好, 以适应多样环境. 策略网络输出不同种类武器和传感器的分配方式和“等待策略”, 即$N_s\cdot N_w+1$维的概率分布. 根据当前状态, 网络选择概率最高的策略生成分配方案, 若选择“等待策略”, 则不进行分配. 此设计将高维的、指数级的分配决策, 简化为一个可遍历的离散选择问题, 并通过“等待策略”赋予智能体决策时机选择的灵活性, 以适应动态不确定性.

3) 奖励设计.

奖励函数结合即时奖励和终止奖励, 综合评估智能体的表现, 兼顾最大化作战效能与最小化资源消耗.

即时奖励指当前分配方案消灭的目标威胁度, 智能体每选择一个动作, 代表选择了一个传感器和武器分配给一个确定的目标, 可以根据对应的分配方案得到此次分配的即时奖励值. 假定在某个决策时刻t_k, 本次动作选择了第s个传感器、第w个武器和第e个目标, 目标总数为E, 则即时奖励计算公式为

$$R_{ews}(t_k)=(h_e p_{se}(t_k)q_{we}(t_k)-c_{ews}(t_k))/E. \tag{13}$$

其中: h_e为目标e的威胁值, $p_{se}(t_k)$为时刻t_k传感器s对目标e的探测概率, $q_{we}(t_k)$为时刻t_k武器w对目标e的毁伤概率, $c_{ews}(t_k)$为时刻t_k传感器s与武器w对目标e的资源消耗.

终止奖励指作战结束后, 摧毁的目标奖励和未摧毁的目标奖励, 计算公式为

$$R_{\text{final}} = \sum_{e=1}^{E_{\text{destroy}}} h_e P_e - \sum_{e=1}^{E_{\text{live}}} h_e k_e. \tag{14}$$

其中: E_{destroy}为摧毁的目标集合; E_{live}为击中要地的目标集合; h_e为目标e的威胁值; P_e为目标e的综合毁伤概率; k_e为惩罚系数, 由专家经验给定; $\sum_{e=1}^{E_{\text{destroy}}} h_e P_e$为战果奖励, 根据目标的威胁值和综合毁伤概率对成功拦截行为给予正向激励, 鼓励智能体优先打击高威胁值目标; $\sum_{e=1}^{E_{\text{live}}} h_e k_e$为漏网惩罚项, 对未能拦截的目标按其威胁值施加惩罚, 迫使智能体兼顾整体防御完整性, 避免因局部优化而导致全局任务失败.

该多层次奖励机制紧密契合动态作战任务的内在要求, 即时奖励激励系统持续高效拦截来袭目标, 终止奖励则强化其对整体战局的理解, 全面增强智能体在复杂对抗环境下的决策与整体作战能力.

在网络架构方面, 本文采用基于 Actor-Critic 的网络框架, 该框架包含策略网络与价值网络两个核心部分, 采用 PPO 算法进行网络训练, 如图 3 所示. 策略网络由 4 层全连接层构成, 输入为状态向量$s(t)$, 经 3 层隐藏层 (每层包含 64 个神经元) 进行非线性变换, 激活函数选用 Tanh, 最终输出层通过 Softmax 函数将动作映射为概率分布, 在增强模型表达能力的同时保障了梯度可微性, 以支持策略梯度优化过程. 价值网络结构与之类似, 但其输出为单一标量, 用于估计当前状态的价值, 从而为策略更新提供稳定且可靠的目标基准.

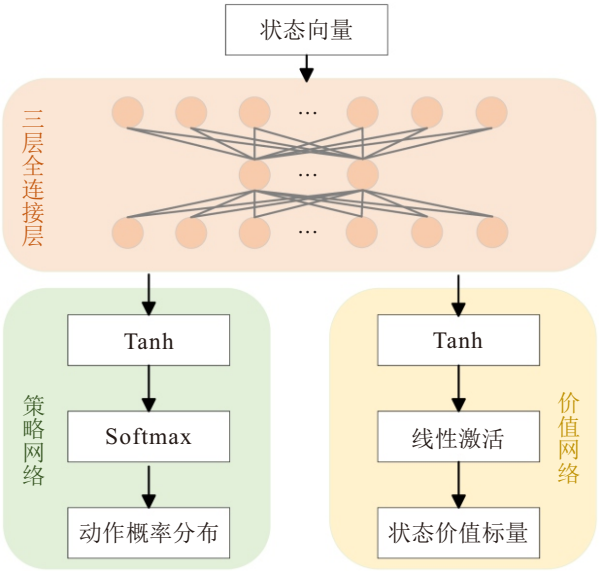


图3 网络架构

针对训练稳定性与效率, 通过消融实验确定关键超参数 (学习率和裁剪参数) 的最优组合, 避免了策略更新振荡或过慢的问题, 确保了训练过程的稳定性和收敛速度. 此外, 对输入状态进行归一化处理, 有效提升了训练的数值稳定性, 加快了收敛过程. 针对策略性能与泛化能力, 采用深度全连接网络作为策略函数和价值函数的近似, 无需依赖人工特征工程, 且使用 PPO 裁剪机制, 通过限制每次策略更新的幅度, 保证训练过程的稳定性, 避免因单个坏样本导致策略性能崩溃. 通过离线训练在线应用的方式, 满足实时性需求.

本研究通过强化学习框架, 将 SWTA 问题转化为动态决策问题, 并设计了与战场态势紧密耦合的状态、动作和奖励函数. 在此基础上, 通过调整的超参数、网络结构和训练机制, 确保了算法能够稳定、高效地学习到一个既能应对实时变化、又能统筹全局需求的高性能策略, 从而有效解决了现有方法在动态环境和求解效率上的瓶颈问题.

3 实验设计与结果分析

3.1 强化学习训练过程

本文采用基于 Actor-Critic 的网络结构, 通过 PPO 算法进行策略优化, 在训练过程中, 采用多项优化策略以提升算法的稳定性与收敛效率. 首先, 对全部奖励值进行归一化处理, 以减小奖励尺度波动带来的影响, 同时对网络输入也进行归一化, 以提升训练数值稳定性. 其次, 引入折扣因子$\gamma = 0.99$, 基于回合制从后向前计算累积回报. 为约束策略更新幅度、避免剧烈变动, 使用 PPO 裁剪机制有效限制每次策略更新的步长. 此外, 采用 Adam 优化器对策略网络与价值网络进行迭代优化, 并借助 PPO 的小批量梯度更新机制, 每批采样数据训练 4 个训练轮次, 以实现稳定且高效的收敛.

针对 PPO 算法关键超参数的设置, 本文重点分析了学习率α和裁剪参数ε的取值, 选取学习率$\alpha \in \{3\times10^{-5}, 3\times10^{-4}, 3\times10^{-3}\}$及裁剪参数$\varepsilon \in \{0.1, 0.2, 0.5\}$构成多组对照实验. 每组超参数组合均在相同初始条件下独立运行, 以滑动平均奖励值及收敛稳定性作为核心评估指标. 实验结果如图 4 所示.

由图 4 可知, 学习率过高易导致策略更新波动剧烈, 难以收敛; 而过低则会显著延缓训练进程. 裁剪参数设置过小会限制策略更新幅度, 降低探索效率; 过大则易诱发训练不稳定. 最终, 综合比较各项性能指标, 确定以 $\alpha = 3\times10^{-4}$, $\varepsilon = 0.2$ 作为最优

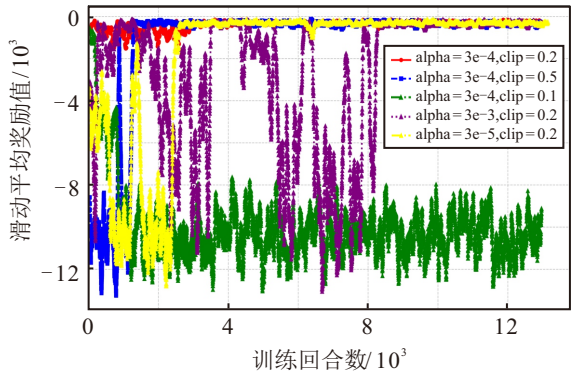


图4　训练过程中奖励变化曲线

组合, 其在取得较高累积奖励的同时, 兼具良好的收敛速度和鲁棒性. 通过不断采样与优化策略, 智能体可以获得更高的累积奖励, 同时网络也逐步收敛. 由奖励曲线可见, 平均奖励约在 4 000 个交互回合后趋于稳定, 这表明策略网络已基本收敛, 策略也逐渐由随机探索转向有效利用. 尽管整体趋势上升, 但仍存在波动, 主要源于环境随机性, 如部分场景不可避免失败, 从而影响奖励值. 此外, 根据专家经验设置部分参数. 其中: 折扣因子$\gamma = 0.99$, 批次大小为 32, 每个交互回合最大步数为 100. PPO 算法的超参数设置如表 1 所示.

表1　PPO 算法参数设置

参数	值
折扣因子γ	0.99
PPO裁剪参数ε	0.2
学习率α	3×10^{-4}
每个交互回合最大步数	100
批次大小	32
优化器	Adam

3.2　对比实验

为验证所提出基于 PPO 的动态传感器-武器-目标分配算法的有效性, 设计 5 种不同的弹药与目标比值 (ammo-to-target ratio, ATR) 场景, 分别为 1.0、1.2、1.5、1.8 和 2.0, 覆盖资源匮乏到资源充足的多种场景. 以两种武器、两种传感器、两种目标举例, 在 3 种场景下的探测概率、毁伤概率和联合概率分布如图 5 ～ 图 7 所示. 为了进一步对比其性能, 本文设计若干对比算法, 包括:

1) Random: 该算法在每次决策时, 从可行的传感器-武器组合或“等待策略”中随机选择一种动作, 用于评估系统在无先验知识条件下的随机适应能力.

2) 基于知识的构造启发式算法 (knowledge-based constructive heuristic, KCH)[39]: 该方法以贪婪策略选择当前即时奖励最大的分配方案, 能够快速生成可行解, 但由于缺乏长期规划机制, 在动态环境中的适应能力有限.

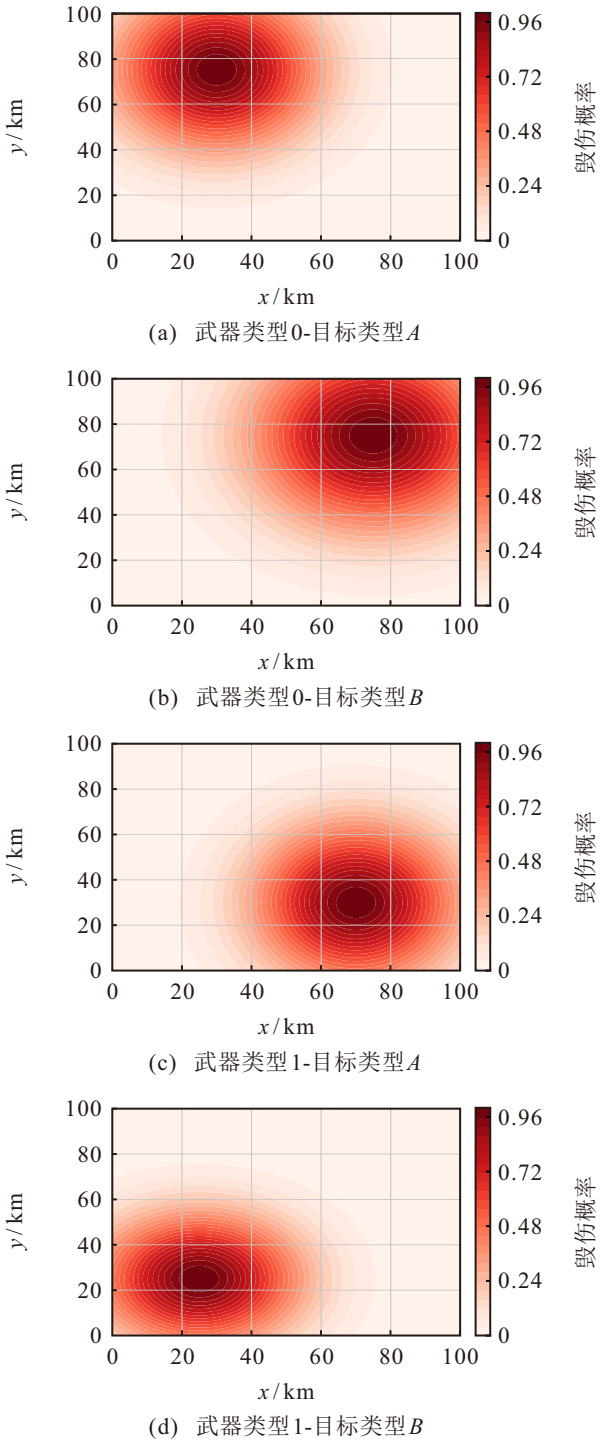
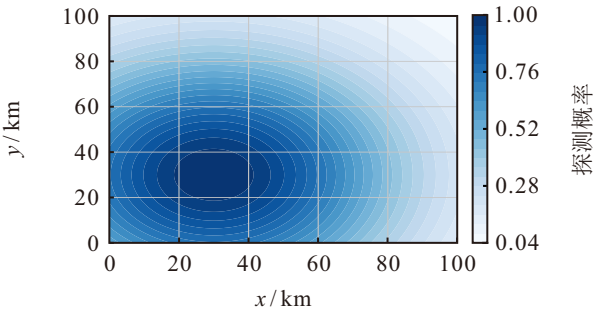


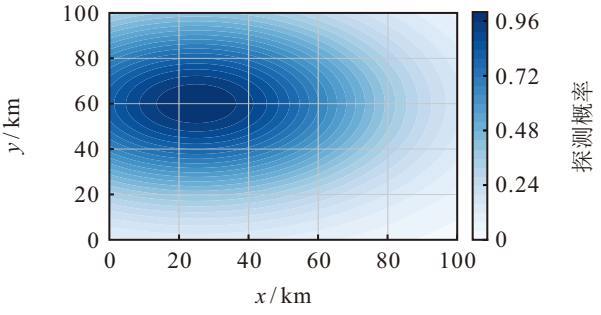
图5　不同类型武器对不同类型目标的毁伤概率分布

3) GA[10]: 算法根据当前可用传感器与武器资源, 生成初始分配方案种群, 通过选择、交叉和变异等遗传操作迭代进化, 最终选取适应度最高的个体所对应的分配方案作为决策结果.

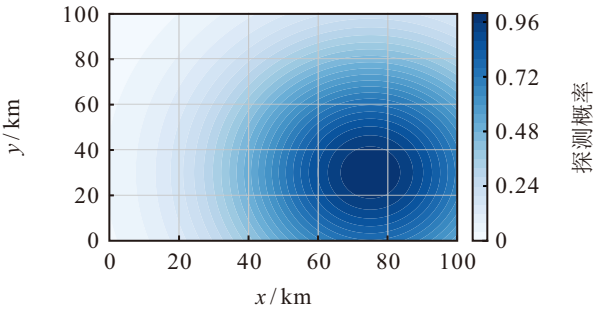
4) PSO[14]: 算法根据当前可用传感器与武器资源, 随机初始化一组粒子表示可能的分配方案, 每个粒子根据个体与群体最优位置更新状态, 最终选取



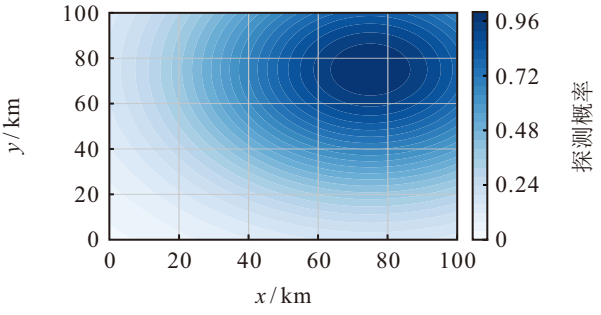
(a) 传感器类型0-目标类型A



(b) 传感器类型0-目标类型B

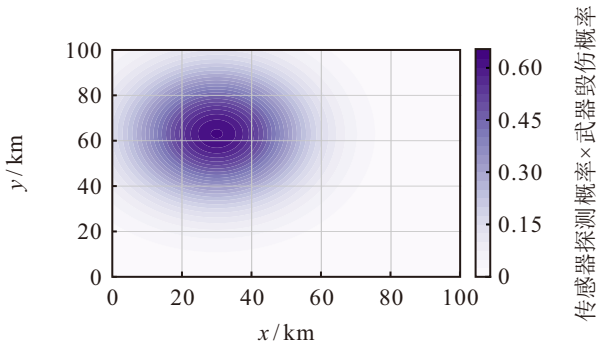


(c) 传感器类型1-目标类型A

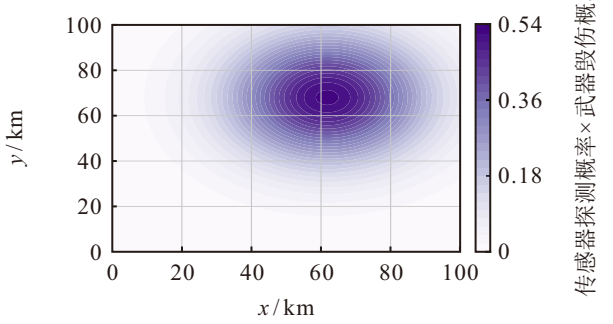


(d) 传感器类型1-目标类型B

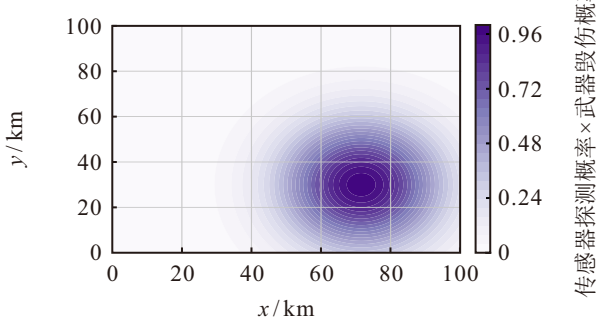
图6 不同类型传感器对不同类型目标的探测概率分布



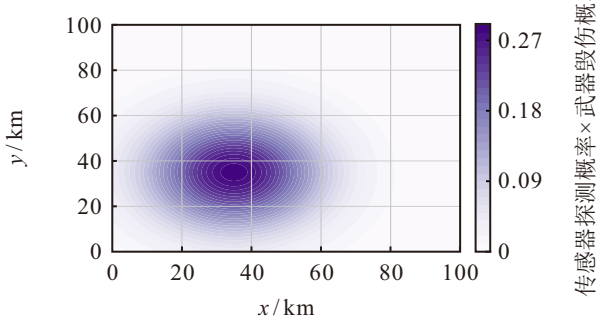
(a) 武器类型0+传感器类型0-目标类型A



(b) 武器类型0+传感器类型0-目标类型B



(c) 武器类型1+传感器类型1-目标类型A



(d) 武器类型1+传感器类型1-目标类型B

图7 武器传感器联合概率分布

全局最优适应度粒子所对应的方案作为决策结果.

5) DQN[37]: 算法将当前可用武器、传感器与目标组合, 作为状态输入至 Q 网络中, 并计算各动作的 Q 值, 选择 Q 值最大的分配方案作为决策结果.

为验证算法的稳定性与鲁棒性, 在不同的 ATR 值下对每种算法分别测试 50 个场景, 每种场景的传感器、武器和目标由场景初始化模块随机生成, 并评估胜率 (任务完成率)、奖励值 (整体表现) 和运行时间 (从作战开始至作战结束的整体用时), 每个场景的实验结果单独列出表格进行展示, 如表 2 ～ 表 6 所示.

实验结果表明, 本文所提出的 PPO 算法在不同弹药目标比场景下表现出优越且稳定的性能. 在资源极度匮乏的场景 (ATR = 1.0) 中, 所有算法胜率均为零, 但 PPO 算法所获得的平均奖励值显著高于其他对比算法, 表明其在严重资源约束下仍能做出更有效的决策, 实现更高的毁伤效能.

在资源适中场景 (ATR = 1.2, 1.5) 中, PPO 算法在胜率和平均奖励值两方面均优于所有对比算法,

表2　不同算法在 ATR = 1.0 下的性能对比

算法	胜率/%	奖励值(均值±标准差)	运行时间/s
Random	0	−11220.1728±11896.9925	0.0072±0.0057
KCH	0	−7363.9976±7838.9726	**0.0049±0.0037**
GA	0	−6528.2212±7890.1517	1.6806±0.9318
PSO	0	−6853.8251±7764.6513	1.5829±0.8415
DQN	0	−6254.7516±6240.6851	0.5124±0.3448
PPO	0	**−5074.6468±5775.3287**	0.1862±0.2096

表3　不同算法在 ATR = 1.2 下的性能对比

算法	胜率/%	奖励值(均值±标准差)	运行时间/s
Random	0.2500	−9742.7832±10133.5652	0.0071±0.0050
KCH	0	−7578.1956±8022.5144	**0.0046±0.0033**
GA	0	−6478.3717±7806.1357	2.1833±1.5678
PSO	3.1250	−7123.0062±7797.8800	2.0509±1.5311
DQN	0	−6118.6445±7133.5887	1.3668±0.5631
PPO	**3.1250**	**−5789.5074±6735.4704**	0.2338±0.2263

表4　不同算法在 ATR = 1.5 下的性能对比

算法	胜率/%	奖励值(均值±标准差)	运行时间/s
Random	15.1515	−6221.3273±5779.7198	0.0098±0.0066
KCH	9.0909	−4717.0414±5737.9165	**0.0067±0.0050**
GA	12.1212	−3073.4914±3413.6015	3.0864±2.2005
PSO	18.1818	−3786.9229±4677.4333	2.8582±2.1117
DQN	21.2121	−3177.8533±4155.6889	1.7112±1.8781
PPO	**27.2727**	**−2561.4540±3853.4781**	0.6109±0.6467

表5　不同算法在 ATR = 1.8 下的性能对比

算法	胜率/%	奖励值(均值±标准差)	运行时间/s
Random	34.4827	−3191.1078±3886.2820	0.0086±0.0063
KCH	20.6896	−3507.9056±4067.1703	**0.0060±0.0047**
GA	17.2413	−2615.7234±4086.6076	8.1740±7.1355
PSO	27.5862	−2224.7589±3530.1961	7.9839±7.0597
DQN	42.8874	−1688.4923±2477.5456	2.3348±2.6447
PPO	**62.0689**	**−586.2856±1240.2522**	0.7382±0.7702

表6　不同算法在 ATR = 2.0 下的性能对比

算法	胜率/%	奖励值(均值±标准差)	运行时间/s
Random	27.5862	−2901.4234±4120.1801	0.0103±0.0079
KCH	41.3793	−1457.1858±3036.2538	**0.0073±0.0059**
GA	55.1724	−1007.3281±2016.2758	11.0482±9.9802
PSO	44.8275	−869.5423±1464.2832	12.9499±8.0757
DQN	63.2774	−733.6448±1244.6642	2.8332±2.8874
PPO	**72.4137**	**−390.4540±743.3363**	1.0046±1.2594

且算法运行时间满足实时性需求, 显示出其在资源条件改善时仍保持良好的决策效率与资源利用能力. 在资源进一步充足的场景 (ATR = 1.8, 2.0) 中, PPO 算法在胜率与奖励值两方面继续表现出明显优势, 表明其能够在多数场景中取得胜利, 体现出稳定且高质量的决策性能.

此外, 设计 5 种 ATR 值与 4 种目标数 (目标数分别为 6、20、30 和 50), 对应 20 种算例, 表 7 展示了不同场景下的 Wilcoxon 秩和检验结果, 该检验用于评估 PPO 算法与其他方法在 5% 显著性水平下的性能差异是否具有统计显著性. p 值表示在原假设成立的情况下观察到当前差异的概率, 其中较小的 p 值表明对原假设的反对证据更强. 显著性列用于指示差异是否具有统计显著性, 其中“Yes”代表差异显著, “No”代表差异不显著. 由实验结果可知, PPO 算法在大多数测试场景中表现出优于对比算法的性能, 特别是在相对简单的场景中优势很显著. 随着任务难度的增加, PPO 算法的优势逐渐减弱, 但仍保持竞争力. 在大规模场景中, PPO 算法与传统优化算法的显著性差异显著, 表明在这些极端条件下, 传统方法均面临较大挑战. 实验结果进一步验证了 PPO 算法在多种 ATR 场景下的优势, 表明其在复杂

表7　PPO 算法与其他方法的 Wilcoxon 秩和检验结果 (p 值及显著性)

算例	Random	KCH	GA	PSO	DQN
1	0.1094/Yes	0.0391/No	0.8438/No	0.0234/Yes	0.6388/No
2	0.0156/Yes	0.0156/Yes	0.0781/Yes	0.0156/Yes	0.0844/Yes
3	0.0078/Yes	0.0078/Yes	0.0547/Yes	0.0156/Yes	0.0547/Yes
4	0.0156/Yes	0.0156/Yes	0.0781/Yes	0.0156/Yes	0.0156/Yes
5	0.5625/No	0.0312/Yes	0.0625/No	0.6875/No	0.0334/Yes
6	0.0078/Yes	0.0117/Yes	0.0078/Yes	0.0117/Yes	0.0117/Yes
7	0.0020/Yes	0.0059/Yes	0.2324/No	0.0098/Yes	0.0084/Yes
8	0.0078/Yes	0.0469/Yes	0.3750/No	0.1562/No	0.2488/No
9	0.6875/No	0.0781/No	0.0469/Yes	0.2969/No	0.8774/No
10	0.0078/Yes	0.0039/Yes	0.1641/No	0.0078/Yes	0.1641/No
11	0.0059/Yes	0.0059/Yes	0.0371/Yes	0.0098/Yes	0.0087/Yes
12	0.0156/Yes	0.0469/Yes	0.8125/No	0.3750/No	0.1641/No
13	0.6875/No	0.4375/No	0.0938/No	0.3125/No	0.0156/Yes
14	0.0125/Yes	0.0781/Yes	0.1562/No	0.0312/Yes	0.0312/Yes
15	0.0078/Yes	0.0156/Yes	0.0547/Yes	0.7422/No	0.0547/Yes
16	0.0078/Yes	0.0078/Yes	0.0078/Yes	0.0781/Yes	0.0078/Yes
17	0.0312/Yes	0.1562/No	0.2188/No	0.1562/No	0.1562/No
18	0.0078/Yes	0.5781/No	0.8125/No	0.6875/No	0.6331/No
19	0.0234/Yes	0.0234/Yes	0.7422/No	0.1484/No	0.1573/No
20	0.0156/Yes	0.0547/Yes	0.0148/Yes	0.0781/Yes	0.0312/Yes

问题中的稳定性和优越性.

本文所提出的基于 PPO 的智能决策框架在多种弹药目标比场景下均展现出优越性能, 是由其方法论上的根本创新性所决定的. 传统方法本质上是在求解一个静态的、离线的优化模型, 它们假设所有信息已知, 并试图给出当前决策时刻的最优方案, 然而真实场景是高度动态的, 全局最优并不等同于每个决策时刻的分配最优. 本文所提出方法通过 OODA 循环不断感知-决策-执行-学习, 在每个决策时刻根据当前最新态势做出分配决策, 并通过即时奖励和终止奖励机制, 不断调整优化其策略, 学会长远规划和精细的资源管理策略, 其动态性和自适应性是在不同 ATR 场景下都能保持稳定高性能的根本原因.

4 结 论

本文针对动态战场环境下传感器-武器-目标分配问题的高复杂性、强实时性要求, 提出了一种基于近端策略优化的智能决策框架. 该方法将 OODA 循环理论融入强化学习训练流程, 通过智能体与仿真环境的持续交互实现策略自主优化, 显著提升了系统在动态对抗条件下的决策适应性和资源利用效率. 实验结果表明, 本文方法在多种弹药目标比场景下均表现出优异的综合性能. 在多种资源高度受限、资源适中和资源充足的场景下, PPO 算法在胜率和奖励值上均优于对比算法, 表明了其良好的泛化能力和稳定性.

本研究仍存在一定局限性, 主要包括: 电磁干扰条件下传感器探测能力易受削弱, 不同作战单元之间的通信延迟问题, 以及边缘设备计算能力有限所带来的部署瓶颈. 未来的研究工作将围绕以下方面展开: 一是引入多智能体协同与对抗机制, 增强系统在分布式决策场景下的鲁棒性; 二是结合在线学习与元强化学习技术, 提升模型在未知战场环境中的快速适应能力; 三是探索轻量化网络结构及模型压缩方法, 以满足边缘计算单元的实际部署需求.

参考文献 (References)

[1] 王晴, 王雨珏, 王浩然, 等. 融合深度强化学习与图神经网络的动态武器目标分配优化[J]. 控制理论与应用, DOI: 10.7641/CTA.2025.50065.
(Wang Q, Wang Y J, Wang H R, et al. Dynamic weapon-target assignment optimization integrating deep reinforcement learning and graph neural network[J]. Control Theory & Applications, DOI: 10.7641/CTA.2025.50065.)

[2] 李梦杰, 常雪凝, 石建迈, 等. 武器目标分配问题研究进展: 模型, 算法与应用[J]. 系统工程与电子技术, 2023, 45(4): 1049-1071.
(Li M J, Chang X N, Shi J M, et al. Developments of weapon target assignment: Models, algorithms, and applications[J]. Systems Engineering and Electronics, 2023, 45(4): 1049-1071.)

[3] Li J R, Wu G H, Wang L. A comprehensive survey of weapon target assignment problem: Model, algorithm, and application[J]. Engineering Applications of Artificial Intelligence, 2024, 137: 109212.

[4] Silav A, Karasakal E, Karasakal O. Bi-objective dynamic weapon-target assignment problem with stability measure[J]. Annals of Operations Research, 2022, 311(2): 1229-1247.

[5] 王艺鹏, 辛斌, 陈杰. 多阶段传感器-武器-目标分配问题的建模与优化求解[J]. 控制理论与应用, 2019, 36(11): 1886-1895.
(Wang Y P, Xin B, Chen J. Modeling and optimization of multi-stage sensor-weapon-target assignment[J]. Control Theory & Applications, 2019, 36(11): 1886-1895.)

[6] Xin B, Chen J, Peng Z H, et al. An efficient rule-based constructive heuristic to solve dynamic weapon-target assignment problem[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, 2011, 41(3): 598-606.

[7] 常雪凝, 石建迈, 陈超, 等. 基于匈牙利-模拟退火算法的多阶段武器目标分配方法[J]. 系统工程与电子技术, 2023, 45(11): 3516-3523.
(Chang X N, Shi J M, Chen C, et al. Multi-stage weapon target assignment method based on Hun-garian simulated annealing algorithms[J]. Systems, Engineering & Electronics, 2023, 45(11): 3516-3523.)

[8] Lloyd S P, Witsenhause H S. Weapon allocation is NP-complete[C]. Proceeding of the IEEE Summer Simulation Conference. Reno: IEEE, 1986: 1054-1058.

[9] Bertsimas D, Paskov A. Solving large-scale weapon target assignment problems in seconds using branch-price-and-cut[J]. Naval Research Logistics: NRL, 2025, 72(5): 735-749.

[10] Luo R N, Zhao Y. Improved genetic algorithm for weapon target assignment problem[C]. 2021 International Symposium on Computer Technology and Information Science. Guilin, 2021: 19-23.

[11] Zhou T, An R, Gao C, et al. A multi-stage target assignment method based on improved genetic algorithm[C]. International Conference on Advanced Unmanned Aerial Systems. Singapore: Springer Nature Singapore, 2023: 123-131.

[12] Tunga H, Kar S, Giri D, et al. Efficacy analysis of NSGAII and multi-objective particle swarm optimization (MOPSO) in agent based weapon target assignment (WTA) model[J]. International Journal of Information Technology, 2024, 16(3): 1347-1356.

[13] Xu S L, Liu Y T, Zhang H C, et al. A distributed collaborative dynamic weapon-target assignment method based on improved binary particle swarm optimization algorithm[C]. Proceedings of 2022

International Conference on Autonomous Unmanned Systems. Singapore: Springer Nature Singapore, 2023: 3128-3142.

[14] Zhai H R, Wang W H, Li Q Z, et al. Weapon-target assignment based on improved PSO algorithm[C]. 2021 33rd Chinese Control and Decision Conference. Kunming, 2021: 6320-6325.

[15] Xing H X, Xing Q H. An air defense weapon target assignment method based on multi-objective artificial bee colony algorithm[J]. Computers, Materials & Continua, 2023, 76(3): 2685-2705.

[16] 刘庆国, 刘新学, 武健, 等. 基于改进 NSGA-Ⅲ 的多 SGSW 火力分配优化[J]. 系统工程与电子技术, 2020, 42(9): 1995-2002.
(Liu Q G, Liu X X, Wu J, et al. Optimization of fire distribution for multiple SGSW based on improved NSGA-III[J]. Systems Engineering & Electronics, 2020, 42(9): 1995-2002.)

[17] 于博文, 吕明. 基于改进 NSGA-Ⅲ 算法的动态武器协同火力分配方法[J]. 火力与指挥控制, 2021, 46(8): 71-77.
(Yu B W, Lv M. Dynamic weapon cooperative fire allocation method based on improved NSGA-III algorithm[J]. Fire Control & Command Control, 2021, 46(8): 71-77.)

[18] 梅海涛, 华继学, 王毅, 等. 基于 IF-HPSO 算法的防空作战 WTA 问题研究[J]. 计算机科学, 2017, 44(5): 263-267.
(Mei H T, Hua J X, Wang Y, et al. Research on WTA problem in air defense operations based on IF-HPSO algorithm[J]. Computer Science, 2017, 44(5): 263-267.)

[19] Peng Z, Lu Z F, Mao X, et al. Multi-ship dynamic weapon-target assignment via cooperative distributional reinforcement learning with dynamic reward[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2025, 9(2): 1843-1859.

[20] Alhijawi B, Awajan A. Genetic algorithms: Theory, genetic operators, solutions, and applications[J]. Evolutionary Intelligence, 2024, 17(3): 1245-1256.

[21] Kong L R, Wang J Z, Zhao P. Solving the dynamic weapon target assignment problem by an improved multiobjective particle swarm optimization algorithm[J]. Applied Sciences, 2021, 11(19): 9254.

[22] Zheng X J, Wu Q X, Gao J. Dynamic weapon-target assignment of armored units based on improved MOPSO algorithm[C]. 2021 3rd International Academic Exchange Conference on Science and Technology Innovation. Guangzhou, 2021: 151-156.

[23] Song G B, Qiang Y G, Liu T, et al. The present situation and progress of dynamic weapon target assignment[J]. Journal of Ordnance Equipment Engineering, 2022, 43(12): 83-88.

[24] Moon T, Kruzins E, Calbert G. Analyzing the OODA cycle[J]. Phalanx, 2002, 35(2): 9-35.

[25] Qin J B, Wang Y, Ding Z Y. Weapon target assignment based on deep Q-learning[C]. 2024 IEEE 9th International Conference on Data Science in Cyberspace. Jinan, 2024: 308-313.

[26] Wu Z H, Pan S R, Chen F W, et al. A comprehensive survey on graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(1): 4-24.

[27] Guo J, Xia W, Hu X X, et al. A spatiotemporal attention-based neural network to evaluate the route risk for unmanned aerial vehicles[J]. Applied Intelligence, 2022, 52(14): 15735-15750.

[28] 郭建国, 胡冠杰, 许新鹏, 等. 基于强化学习的多对多拦截目标分配方法[J]. 空天防御, 2024, 7(1): 24-31.
(Guo J G, Hu G J, Xu X P, et al. Multi-to-many interception target allocation method based on reinforcement learning[J]. Air & Space Defense, 2024, 7(1): 24-31.)

[29] Luo W, Lv J, Liu K, et al. Learning-based policy optimization for adversarial missile-target assignment[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 52(7): 4426-4437.

[30] Li S, He X Y, Xu X, et al. Weapon-target assignment strategy in joint combat decision-making based on multi-head deep reinforcement learning[J]. IEEE Access, 2023, 11: 113740-113751.

[31] Wang X C, Zhang Y, Wang G. Target assignment for multiple stages of weapons systems using a deep *Q*-learning network and a modified artificial bee colony method[J]. Computers and Electrical Engineering, 2024, 118: 109378.

[32] Na H, Ahn J, Moon I C. Weapon–target assignment by reinforcement learning with pointer network[J]. Journal of Aerospace Information Systems, 2023, 20(1): 53-59.

[33] 伍国华, 李冰洁, 袁于斐, 等. 基于任务分解与强化学习的多平台协同火力分配方法[J]. 控制与决策, 2024, 39(5): 1727-1735.
(Wu G H, Li B J, Yuan Y F, et al. Multi-platform collaborative firepower allocation method based on task decomposition and reinforcement learning[J]. Control and Decision, 2024, 39(5): 1727-1735.)

[34] 孙昕, 邢立宁, 王锐, 等. 基于多目标进化算法的防空导弹武器目标分配[J]. 系统仿真学报, 2024, 36(6): 1298-1308.
(Sun X, Xing L N, Wang R, et al. Air defense missile weapon target assignment based on multi-objective evolutionary algorithm[J]. Journal of System Simulation, 2024, 36(6): 1298-1308.)

[35] 李泷鑫, 桑红燕, 孟磊磊, 等. 基于分组策略的多除草机器人任务分配多目标优化[J]. 控制理论与应用, DOI: 10.7641/CTA.2025.50125.
(Li L X, Sang H Y, Meng L L, et al. Multi-objective optimization of task allocation for multiple weeding robots based on grouping strategy[J]. Control Theory & Applications, DOI: 10.7641/CTA.2025.50125.)

[36] Avci I, Yildirim M. Solving weapon-target assignment problem with salp swarm algorithm[J]. Tehnicki Vjesnik-Technical Gazette, 2023, 30(1): 17-23.

[37] Li C, Xin B, He Y M, et al. Dynamic weapon target assignment based on deep Q network[C]. 2023 42nd Chinese Control Conference. Tianjin, 2023: 1773-1778.

[38] Bogdanowicz Z R. A new efficient algorithm for optimal assignment of smart weapons to targets[J]. Computers & Mathematics with Applications, 2009, 58(10): 1965-1969.

[39] Xin B, Wang Y P, Chen J. An efficient marginal-return-based constructive heuristic to solve the sensor-weapon-target assignment problem[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2019, 49(12): 2536-2547.

[40] David K B. Radar system analysis and modeling[C]. Nanjing Research Institude of Electronics Technology. Beijing: Publishing House of Electronics Industry, 2017: 326-328.

[41] 张晓今, 张为华, 江振宇. 导弹系统性能分析[M]. 北京: 国防工业出版社, 2013: 146-148.
(Zhang X J, Zhang W H, Jiang Z Y. Performance analysis of missile system[M]. Beijing: National Defense Industry Press, 2013: 146-148.)

作者简介

王晴 (1990−), 女, 实验师, 博士, 主要研究方向为多智能体系统分布式协同控制、事件触发控制, E-mail: wangqing1020@bit.edu.cn;

王浩然 (1999−), 男, 硕士生, 主要研究方向为多目标优化与决策, E-mail: 1046367525@qq.com;

辛斌 (1982−), 男, 教授, 博士, 主要研究方向为多智能体系统与多机器人系统协同控制、智能优化与决策, E-mail: brucebin@bit.edu.cn;

张佳 (1980−), 女, 副教授, 博士, 主要研究方向为多目标优化与多目标决策、多智能体系统分布式协同, E-mail: zhangjia@bit.edu.cn.