控制与决策
Control and Decision

面向智能空中博弈的风险约束离线强化学习算法

李博文, 王臆淞, 赵铭慧, 骞晨旭, 程光权, 张雪波

引用本文:
李博文, 王臆淞, 赵铭慧, 等. 面向智能空中博弈的风险约束离线强化学习算法[J]. 控制与决策, 2026, 41(6): 1665-1675.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2025.0850

面向智能空中博弈的风险约束离线强化学习算法

李博文[1,2]，王臆淞[1,2]，赵铭慧[1,2†]，骞晨旭[1,2]，程光权[3]，张雪波[1,2]

(1. 南开大学 机器人与信息自动化研究所，天津 300350；
2. 天津市智能机器人技术重点实验室，天津 300350；
3. 国防科技大学 系统工程学院，长沙 410073)

摘　要: 复杂空中博弈场景同时面临两类核心挑战: 1) 在线交互采样效率低, 且高风险试错行为易导致严重后果; 2) 离线数据稀缺且覆盖有限, 策略对分布外状态的泛化能力不足. 针对这两类问题, 提出一种基于风险约束和保守值函数学习的离线博弈算法 CQL-Safe, 其中风险被定义为智能体在博弈对抗过程中遭受损毁或被对手击落的概率及其相关安全威胁指标. 首先, 设计一种融合示教学习与扩散模型的数据集构建方法, 能够有效缓解离线强化学习数据稀缺问题; 然后, 设计多维风险评估函数量化风险因素, 并将其嵌入保守值函数学习框架, 抑制分布外动作的过高估计; 最后, 引入拉格朗日乘子机制动态调节风险约束强度, 以实现奖励最大化与安全性保障间的自适应平衡. 所提出算法在多类空中博弈场景下具有较高的训练效率和显著的性能优势, 能够在保障智能体安全的同时大幅提升策略的泛化性和有效性.

关键词: 智能空中博弈；离线强化学习；示教学习；扩散模型；风险约束；保守值函数；拉格朗日法

中图分类号: TP18　　**文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.0850

引用格式: 李博文, 王臆淞, 赵铭慧, 等. 面向智能空中博弈的风险约束离线强化学习算法 [J]. 控制与决策, 2026, 41(6): 1665-1675.

Risk-constrained offline reinforcement learning for intelligent aerial combat

LI Bo-wen[1,2]，*WANG Yi-song*[1,2]，*ZHAO Ming-hui*[1,2†]，*QIAN Chen-xu*[1,2]，*CHENG Guang-quan*[3]，*ZHANG Xue-bo*[1,2]

(1. Institute of Robotics and Automatic Information System，Nankai University，Tianjin 300350，China；2. Key Laboratory of Intelligent Robotics，Tianjin 300350，China；3. College of Systems Engineering，National University of Defense Technology，Changsha 410073，China)

Abstract: Complex aerial combat scenarios face two major challenges: 1) low online sampling efficiency with potentially catastrophic high-risk exploration; 2) scarce offline data with limited coverage that undermines generalization to out-of-distribution states. To address these challenges, we propose CQL-Safe, a risk-constrained offline algorithm built on conservative value function learning, where risk is defined as the probability of agent damage or shoot-down and related safety-threat metrics. CQL-Safe constructs an augmented offline dataset combining learning from demonstration and diffusion model to alleviate data scarcity, embeds a multi-dimensional risk evaluation function into the conservative value function learning framework to curb overestimation of out-of-distribution actions, and adopts a Lagrangian multiplier to adaptively tune constraint strength, achieving a practical balance between reward maximization and safety. Experiments across diverse aerial combat scenarios demonstrate improved training efficiency and superior performance, substantially enhancing policy effectiveness and generalization while maintaining agent safety.

Keywords: intelligent aerial combat；offline reinforcement learning；learning from demonstration；diffusion model；risk constraint；conservative value function；Lagrangian approach

0　引　言

随着智能决策和自主控制技术的快速发展, 智能体在空中对抗环境中的应用日益广泛, 催生了对智能空中博弈问题的深入研究[1]. 相关研究面临诸多

收稿日期: 2025-08-18；**录用日期:** 2025-12-19.
基金项目: 国家自然科学基金项目 (62293510, 62293513)；天津市自然科学基金项目 (22JCZDJC00810).
†通信作者. E-mail: zhaomh@nankai.edu.cn.

难题: 一方面, 空中博弈场景操作精度要求极高, 还需要兼顾飞行控制的物理特性与博弈论中的决策过程, 因此, 建模难度极大[2]; 另一方面, 传统基于规则或优化的博弈方法难以应对持续变化的战场态势与高维动作空间的组合爆炸问题, 而在应用强化学习方法时, 在线强化学习需要与环境持续交互来收集数据, 样本采集效率低, 且探索过程中可能会采用潜在高风险策略, 不具备安全保障[3].

针对细粒度操控行为的建模挑战, 相关研究多采用分层框架设计方法. 通过将复杂的空中博弈决策分解为不同抽象层次的子问题, 可有效降低建模的复杂度, 提升飞行器细粒度动作模拟能力[4-5].

针对训练过程中样本交互效率和安全性问题, 相关研究逐渐转向离线强化学习范式. 然而, 离线强化学习性能高度依赖于预先收集的数据集, 面临显著的数据获取难度和质量问题. 数据集往往来源于有限的专家演示或历史交互记录, 覆盖范围受限, 获取难度较高, 且数据中可能包含次优行为、低质量样本或冗余信息[6]. 近年来, 相关研究主要从数据生成、筛选和增强等角度提出了多种改进策略: 一类方法引入示教学习 (LfD), 通过采集专家演示行为来提升数据集的初始质量和策略引导性[7]; 另一类研究则聚焦于数据增强技术, 采用模型生成方法 (如对抗生成网络 (GAN)[8] 或扩散模型 (diffusion model)[9]) 对现有数据分布建模, 并生成多样化、任务相关的合成样本, 以缓解关键状态空间的覆盖不足.

离线强化学习同时存在数据分布偏移的问题, 对分布外数据易出现价值过高估计, 导致学习过程不稳定. 为解决这一问题, 保守Q学习 (CQL) 算法[10]通过引入保守的Q函数, 避免策略对未知数据分布产生过度乐观估计, 在标准的贝尔曼误差目标上加入Q值正则化, 改善了传统离线 RL 方法在复杂控制领域的性能; 隐式Q学习 (IQL) 算法[11] 则通过不直接评估未见动作的Q值, 利用函数逼近器的泛化能力, 通过对状态值函数随机化处理以及优势加权行为克隆提取策略, 在多个基准测试中取得领先成果.

尽管上述方法在提升策略效率和性能方面取得了一定进展, 但是, 在面对动态约束复杂以及策略安全性要求较高的场景时仍然存在不足. 特别是在空中博弈这类高度对抗性、不可逆场景中, 智能体面临诸多风险, 一些试错行为可能会导致严重后果[12]. 本文将此类风险明确定义为智能体在博弈对抗过程中遭受损毁或被击落的概率及其相关安全威胁指标, 并将其作为策略优化过程中需要严格控制的核心变量. 因此, 如何在离线学习框架下引入安全强化学习 (SRL) 思想, 实现对该类风险的量化评估和动态约束, 成为当前研究的重要方向.

鉴于此, 本文提出一种具备动态风险约束的保守值函数离线博弈算法. 首先, 构建融合示教学习与扩散模型的数据集生成方法, 以缓解离线强化学习中的数据稀缺和分布偏移问题. 然后, 设计 3 层层次化策略架构, 将整体博弈问题分解为底层机动控制、中层策略规划和顶层策略选择, 从而降低策略搜索的计算复杂度, 有效应对高维复杂任务. 在算法层面, 基于保守值函数学习算法框架, 构建多维风险评估函数, 对飞机生存概率、威胁指数以及机动能力等指标进行量化, 并引入拉格朗日乘子机制在策略更新过程中动态调节风险约束强度, 以实现奖励最大化与安全性保障间的自适应平衡. 大量对抗场景下的实验结果表明, 所提出方法在保证策略安全性的基础上能够显著提升训练效率和泛化性能, 展现出在复杂空中博弈环境中的广阔应用前景.

1 问题建模

本文基于安全离线强化学习方法来解决智能空中博弈问题, 以应对高风险环境中探索带来的不确定性, 避免智能体由于在未知或危险状态下做出过于激进的决策而导致损失. 本节包括安全离线强化学习问题描述与建模以及智能空中博弈问题形式化表示的内容.

1.1 安全离线强化学习问题描述与建模

安全离线强化学习通常会建模为约束马尔可夫决策过程 (CMDP)[13], 旨在最大化累积奖励的同时满足安全约束. CMDP 是标准马尔可夫决策过程 (MDP) 的扩展, 形式化定义为

$$M=(S,A,P,r,c,\gamma,d). \tag{1}$$

其中: S为状态空间, A为动作空间, $P(s'|s,a)$为状态转移概率, $r:S\times A\times S\to\mathbb{R}$为奖励函数, $c:S\times A\to\mathbb{R}$为成本函数, $\gamma\in[0,1]$为折扣因子, d为安全阈值.

安全强化学习的目标是在满足安全约束的情况下, 求解使得期望回报最大化的最优可行策略, 即

$$\pi^*=\arg\max_{\pi\in\Pi_C}J(\pi). \tag{2}$$

其中: Π_C为满足安全约束的安全策略集; $J(\pi)$为期望回报, 且

$$J(\pi)=\mathbb{E}_{\tau\sim\pi}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\Big], \tag{3}$$

$$\tau=(s_0,a_0,s_1,a_1,\ldots). \tag{4}$$

这里τ为样本轨迹, 表示从初始状态开始的一系列状

态-动作序列.

对于不同类型的决策任务, 安全策略集Π_C可以有不同的表达形式. 对于安全性要求严格的决策任务, 通常采用硬约束方式, 即在所有时刻均需要强制满足单步约束, 即

$$\Pi_C = \{\pi \in \Pi : c(s_t, a_t) < d_h, \forall t \in \{0, 1, \ldots\}\}. \tag{5}$$

在无模型情况下, 软约束方式有更广泛的应用, 即对折扣累积成本的期望进行约束, 有

$$\Pi_C = \left\{\pi \in \Pi : \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t c(s_t, a_t)\right] < d_s\right\}. \tag{6}$$

离线强化学习则通过固定的数据集D进行策略优化, 如下所示:

$$D = \{(s_i, a_i, r_i, s_i')\}_{i=1}^{N}, \tag{7}$$

其中数据来自于行为策略π_β. 与在线强化学习算法不同, 离线强化学习在训练过程中不允许与环境交互, 这保障了训练过程的安全性, 但是, 同时也引入了数据分布偏移问题, 给求解最优策略的过程带来了挑战.

安全离线强化学习通过 CMDP 框架建模, 利用静态数据集进行训练并进行约束优化, 确保了学习策略的安全性.

1.2　智能空中博弈问题形式化表示

在空中博弈场景中, 智能体需要对飞机的油门、方向舵、升降舵等多个控制变量进行精确操控, 以实现灵活机动飞行. 然而, 直接在细粒度控制层面进行博弈策略学习, 面临高维状态空间、复杂动力学约束以及长时间决策链等挑战. 为此, 本文在先前工作的基础上, 采用分层架构, 将智能空中博弈问题划分为底层、中层和顶层 3 个层级: 底层负责机动控制, 中层设计博弈策略, 顶层进行博弈策略选择, 以此降低问题复杂度, 提高训练效率.

底层控制器的主要目标是提供稳定的机动控制能力, 确保飞机能够抵达指定经度、纬度、高度的目标点, 进而支持中层策略设计. 本文中, 底层控制器采用 Li 等[14] 提出的基于近端策略优化算法训练的固定翼飞机控制器.

中层根据底层控制器的执行能力, 设计多种规则博弈策略, 以适应不同战术需求. 本文设计如下 4 种典型策略作为中层博弈策略:

a_a: 全力攻击策略, 己方飞机朝着选定敌方目标行进并进行全力攻击, 以歼灭敌方飞机;

a_d: 伪装战术策略, 己方飞机朝着选定敌方目标行进, 但是不发射导弹, 进行佯装攻击来迷惑对手;

a_p: 威胁保持策略, 己方飞机抬升高度, 调整对于选定敌方目标的角度, 以保持对敌方飞机的压制态势;

a_m: 导弹躲避策略, 己方飞机向远离选定敌方目标的方向执行机动规避操作并撤离博弈区域, 以提升存活概率.

顶层智能体基于强化学习方法, 通过对中层策略的动态选择, 实现最优博弈决策. 本文参考 Qian 等[15] 所提出的建模方式对顶层博弈智能体进行建模.

后续将基于此智能空中博弈问题建模方式, 进行基于风险约束的离线博弈算法研究.

2　面向智能空中博弈的风险约束离线强化学习算法

基于以上智能空中博弈问题建模, 本文提出一种面向智能空中博弈的离线强化学习算法. 算法框图如图 1 所示. 首先, 依托高仿真智能空中博弈平台, 采用示教学习与扩散模型相结合的数据集生成方法构建高质量离线数据集; 然后, 基于保守值函数学习框架, 综合考虑生存概率、威胁指数以及飞机机动能力设计风险评估函数, 同时, 引入动态调整的拉格朗日乘子优化博弈过程中的风险约束权重, 以实现高效且安全的博弈决策.

2.1　结合示教学习与扩散模型的离线数据集构建

2.1.1　示教数据集获取

示教学习是一种利用已有演示数据学习决策策略的方法, 目前, 已广泛应用于机器人控制、自动驾驶以及强化学习等领域. 在多智能体强化学习中, 示教学习允许多个智能体借助人类专家或其他高性能策略的演示数据进行学习, 避免从零开始探索, 从而提高学习效率并降低训练数据需求, 适用于高风险或高成本的环境[16].

本节在“智空仿真推演与训练平台”中构建示教数据集, 以支持示教学习模型的训练. 为获取丰富多样的示教数据, 本文采用两种数据来源:

1) 基于行为树的策略模拟生成的示教数据;

2) 由人类操作员通过演示采集的高质量示教数据.

行为树是一种广泛应用于指挥控制、机器人决策以及游戏 AI 的层次化任务建模方法. 其核心优势在于通过模块化结构清晰地描述复杂决策逻辑, 并提高系统的可扩展性和鲁棒性[17].

行为树通常采用深度优先遍历, 确保智能体能够优先执行关键任务, 并在任务失败时动态调整策

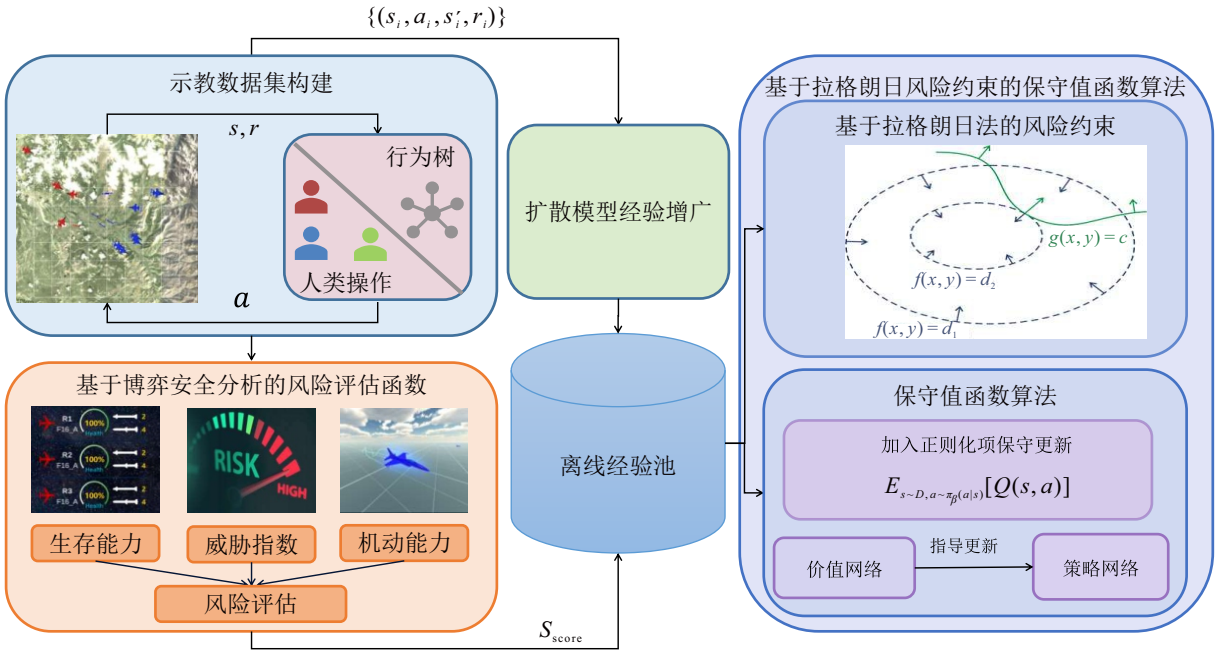

图1 面向智能空中博弈的风险约束离线强化学习算法

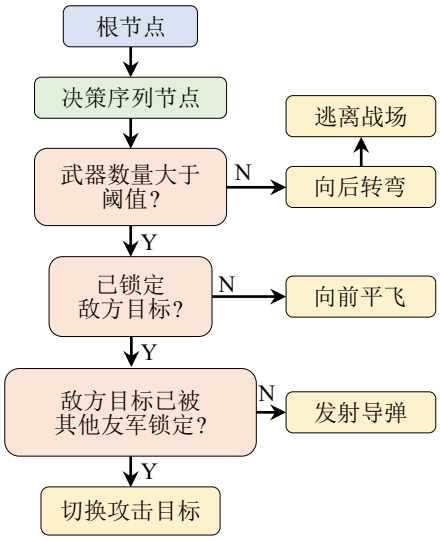

图2 行为树决策流程

略. 针对空中博弈任务特点, 本文设计如图 2 所示的多级行为树结构.

基于上述行为树策略, 本文在仿真环境中执行 100 幕对抗, 收集结构化的示教数据, 为后续模型训练提供基准数据支持.

虽然行为树能够提供结构化的决策数据, 但是, 人类专家在博弈决策中体现的灵活性、创造性和战术智慧是行为树难以完全模拟的. 因此, 本文设计人类操作数据采集环节, 以捕获这些高价值的策略特征.

本文邀请多名志愿者在“智空仿真推演与训练平台”上进行操作, 其数据采集流程如下:

1) 任务设定: 志愿者在多种空中博弈场景下进行智能体的行为决策.

2) 交互模式: 人类在交互界面通过鼠标和键盘交互的形式操控多架飞机与智能体进行博弈; 数据采集涵盖宏观任务规划、中层战术决策以及底层机动控制等多个层级.

3) 数据采集: 记录 100 幕示教数据, 主要涵盖博弈态势信息和飞机执行的动作, 以及环境返回的奖励信息.

通过行为树和人类操作两种方式, 本文构建一个包含 200 轮对抗数据的示教数据集. 尽管数据采集过程耗费了大量时间, 但是, 这些数据对于深度学习模型的训练仍显不足.

2.1.2 扩散模型经验增广

鉴于使用行为树以及人类操作进行示教数据采集的效率较低, 本节采用基于扩散模型的经验增广方法, 以生成高质量的合成示教数据. 该方法可有效扩充训练集, 提高模型的泛化能力, 增强其在不同博弈场景中的适应性和稳健性.

扩散模型是一类基于逐步去噪过程的生成模型, 近年来, 在计算机视觉、强化学习以及数据合成等领域得到了广泛应用. 模型示意图如图 3 所示. 其基本思想如下: 在前向扩散过程中, 逐步向真实数据中添加噪声, 直至数据近似服从标准高斯分布; 在逆向去噪过程中, 则利用参数化模型逐步去除噪声, 从而恢复数据分布, 实现高质量样本生成和分布学习[18].

前向扩散过程: 给定真实数据 $x_0 \sim q(x)$, 前向扩散过程通过在每个时间步 t 添加少量高斯噪声, 将数据逐步退化. 该过程定义为

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I). \quad (8)$$

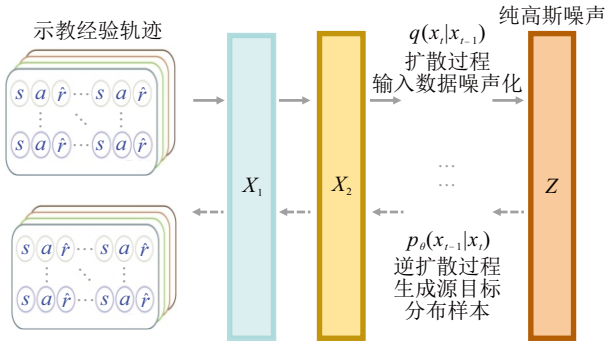

图3　扩散模型示意图

其中: β_t为时间步t的预定义噪声系数, I为单位矩阵. 经T个时间步后, 数据分布趋近于各向同性的高斯分布, 即

$$q(x_T) \approx \mathcal{N}(0, I). \tag{9}$$

逆向去噪过程: 逆向过程的目标是从噪声样本中逐步恢复出原始数据分布, 其转移概率由参数化模型给出, 如下所示:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{10}$$

根据重参数化定理, 定义$\alpha_t = 1 - \beta_t$和$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, 本节设定$\Sigma_\theta(x_t, t) = \sigma_t^2 I$, 选取$\sigma_t^2 = \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$, 则均值$\mu_\theta(x_t, t)$参数化为

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \beta_t\sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t, t)), \tag{11}$$

其中$\epsilon_\theta(x_t, t)$为模型预测的噪声.

对前向过程重新参数化为

$$x_t(x_0, \epsilon) = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon,\ \epsilon \sim \mathcal{N}(0, I). \tag{12}$$

可以证明, 变分下界中关于x_{t-1}的损失项可写为

$$L_{t-1} = \mathbb{E}_{x_0,\epsilon}\Big[\frac{1}{2\sigma_t^2}\Big\|\frac{1}{\sqrt{\alpha_t}}(x_t(x_0,\epsilon) - \beta_t\sqrt{1-\bar{\alpha}_t}\epsilon) - \mu_\theta(x_t(x_0,\epsilon), t)\Big\|^2\Big] + C, \tag{13}$$

其中C为与参数θ无关的常数. 进一步地, 式 (13) 中损失可重写为

$$\mathbb{E}_{x_0,\epsilon}\Big[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2\Big]. \tag{14}$$

该目标与多尺度噪声下的去噪分数匹配相似.

为简化训练过程并提升采样质量, 本节采用以下目标函数:

$$L_s(\theta) = \mathbb{E}_{t,x_0,\epsilon}[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2], \tag{15}$$

其中t均匀采样于$\{1, 2, \ldots, T\}$. 当$t = 1$时, 该损失近似对应于离散解码器中积分的高斯近似; 而对于$t > 1$, 该目标与式 (14) 中的无权重版本等价. 实验结果表明, 该简化目标能够提高采样质量, 同时降低实际应用难度.

2.2　基于博弈安全分析的风险评估函数设计

在复杂的博弈对抗环境中, 如何衡量飞机在复杂对抗环境中的安全性, 并将其作为强化学习过程中的优化目标, 是确保决策系统稳定性和有效性的关键. 为量化空中博弈中的安全性, 本节设计风险评估函数, 综合考虑飞机生存概率、威胁指数以及机动能力, 为强化学习模型中的策略优化提供约束目标.

2.2.1　生存概率

生存概率是衡量飞机在空中博弈中生存能力的关键指标. 在传统的生存概率计算中, 常常忽略敌方和己方飞机的数量, 但是, 在实际博弈中, 敌我双方飞机的数量对于局势具有决定性影响. 因此, 本节设计的生存概率的计算公式为

$$P_s = \frac{N_r}{N_r + N_b} \cdot \mathrm{e}^{-k \cdot T}. \tag{16}$$

其中: N_r和N_b分别为己方和敌方存活飞机的数量; k为一个常数, 表示博弈对抗持续时间对于生存概率的影响; T为当前博弈的持续时间. 该公式的含义为己方飞机数量越多, 敌方飞机数量越少, 生存概率越高; 同时, 随着博弈时间的增加, 生存概率会逐渐降低.

2.2.2　威胁指数

威胁指数部分由两大核心因素组成: 敌方威胁 (R_t) 和己方防御能力 (R_d). 通过加权平均这两个因素, 可得到总的威胁指数 (R_{total}).

首先, 对敌方威胁进行评估. 敌方飞机对我方的威胁来自其锁定的目标以及发射导弹的数量. 若敌方飞机锁定目标, 则其威胁评分增加. 威胁评分为

$$R_t = \frac{\sum_{i=1}^{N_b} t_i}{N_b}, \tag{17}$$

其中t_i为敌方飞机的威胁评分. 若敌方锁定目标, 则该评分为 1; 否则, 为 0. 总威胁分数对敌方飞机的总数进行归一化处理. 然后, 对己方防御能力进行评估. 己方防御能力通过可用导弹的数量 (如近程红外导弹、雷达导弹) 来评估. 具体公式为

$$R_d = \frac{\sum_{i=1}^{N_r}\left(\frac{S_c}{S_t} + \frac{A_c}{A_t}\right)}{N_r}. \tag{18}$$

这里: S_c和A_c分别为当前可用的近程红外导弹和中

程雷达导弹的数量, S_t和A_t分别为近程红外导弹和中程雷达导弹的总量. 最后, 通过敌方威胁与己方防御能力的加权平均, 可计算得到当前态势总威胁指数 (R_{total}), 具体公式为

$$R_{\text{total}} = \lambda_r \cdot R_t + (1-\lambda_r)\cdot(1-R_d), \tag{19}$$

其中λ_r为敌方威胁的权重, 取值范围为 [0, 1].

威胁因子 (r_f) 由总威胁指数反向计算得出, 如下所示:

$$r_f = 1 - R_{\text{total}}, \tag{20}$$

其中r_f值越高, 威胁越低.

2.2.3 机动能力

机动能力评估反映了飞机在博弈中的机动能力和战术效果. 通过计算己方存活飞机的平均飞行速度和飞行高度来评估战术效果. 其计算公式为

$$U_t = \lambda_u \cdot \frac{v_{\text{avg}}}{v_{\max}} + (1-\lambda_u)\cdot\frac{h_{\text{avg}}}{h_{\max}}. \tag{21}$$

其中: λ_u为飞行速度的权重, 取值范围为 [0, 1]; v_{avg}为己方存活飞机的平均飞行速度; h_{avg}为己方存活飞机的平均飞行高度; $v_{\max}$和$h_{\max}$分别为己方存活飞机的最大飞行速度和最大飞行高度, 这两个指标分别被归一化到 [0, 1] 的区间.

最终的风险函数综合考虑了生存概率、威胁指数和机动能力 3 个因素. 综合风险度量 (S_{score}) 通过加权求和的方式得出, 公式为

$$S_{\text{score}} = \lambda_{P_s}\cdot P_s + \lambda_{r_f}\cdot r_f + \lambda_{U_t}\cdot U_t. \tag{22}$$

其中: λ_{P_s}、λ_{r_f}和λ_{U_t}分别为生存概率、威胁指数和机动能力的权重, 取值范围为 [0, 1]. 该函数的值范围在 [0, 1] 之间, 越接近 0, 风险越高; 越接近 1, 风险越低.

2.3 基于拉格朗日法的风险约束保守值函数算法

在离线学习场景下, 强化学习算法可能会由于对分布外动作 (OOD Actions) 的Q值过高估计而导致性能下降. 针对这一问题, 本文提出基于风险约束和保守值函数学习的离线强化学习算法 CQL-Safe, 通过引入拉格朗日法的风险约束机制, 能够有效抑制分布外动作Q值的过高估计, 可显著提高模型在复杂对抗环境下的鲁棒性和可靠性. 如算法 1 所示.

算法 1 基于拉格朗日法的风险约束保守值函数算法.

input: 扩散离线数据集$\mathcal{D}$; 风险约束阈值c_{thresh}; 拉格朗日乘子初始值λ_{init}; 学习率$\eta_Q, \eta_\pi, \eta_\lambda$.

1. 初始化: Q函数Q_θ, 策略π_ϕ, 拉格朗日乘子$\lambda \leftarrow \lambda_{\text{init}}$
2. for 每$t \in \{1, 2, \ldots, N\}$ do
3. 从离线数据集$\mathcal{D}$中采样一批数据(s, a, r, s')
4. 计算贝尔曼误差损失, 如下所示:

$$\mathcal{L}_{\text{BE}} = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}[(Q_\theta(s,a) - (r+ \gamma\mathbb{E}_{a'\sim\pi_\phi(\cdot|s')}[Q_\theta(s',a')]))^2]$$

5. 计算CQL正则化损失, 如下所示:

$$\mathcal{L}_{\text{CQL}} = \mathbb{E}_{s\sim\mathcal{D}}\Big[\log\sum_a \exp(Q_\theta(s,a)) - \mathbb{E}_{a\sim\hat{\pi}_\beta(\cdot|s)}[Q_\theta(s,a)]\Big]$$

6. 计算风险约束$c(s,a)$
7. 计算约束违反程度, 如下所示:

$$\delta_c = \mathbb{E}[c(s,a)] - c_{\text{thresh}}$$

8. 得到总损失, 如下所示:

$$\mathcal{L} = \mathcal{L}_{\text{BE}} + \mathcal{L}_{\text{CQL}} + \alpha\cdot\delta_c$$

9. 更新Q网络, 即

$$\theta_t \leftarrow \theta_{t-1} - \eta_Q\nabla_\theta\mathcal{L}$$

10. 更新策略网络, 如下所示:

$$\phi_t \leftarrow \phi_{t-1} + \eta_\pi\mathbb{E}_{s,a}[Q_\theta(s,a) - \log\pi_\phi(a|s)]$$

更新拉格朗日乘子, 如下所示:

$$\lambda \leftarrow \text{clip}(\lambda + \eta_\lambda\cdot\delta_c, 0, \lambda_{\max})$$

11. end

output: 优化后的策略网络π_ϕ和Q函数Q_θ.

2.3.1 保守值函数学习算法

保守值函数学习算法[10] 作为一种专门应对离线强化学习挑战的算法, 通过引入保守约束, 可有效抑制分布外动作Q值的过高估计.

CQL 算法的核心思想在于通过在Q值更新过程中引入保守正则化项, 学习一个对目标策略价值形成下界估计的Q函数. 具体而言, CQL 通过最小化某个状态-动作分布下的Q值, 同时, 最大化在行为策略分布下的Q值, 确保学习到的Q函数能够为真实策略价值提供保守估计, 从而避免由于分布外动作导致的过高估计问题.

CQL 算法的目标是最小化损失函数, 有

$$\mathcal{L}_{\text{CQL-total}} = \mathcal{L}_{\text{BE}} + \alpha\cdot\mathcal{L}_{\text{CQL}}. \tag{23}$$

其中: $\mathcal{L}_{\text{BE}}$为标准的贝尔曼误差损失, 用于逼近目标Q值; $\mathcal{L}_{\text{CQL}}$为保守正则化项; α为权重系数, 用于控制保守性约束的强度.

CQL 算法的执行流程包括以下几个关键步骤: 首先, 构建如下贝尔曼误差损失函数:

$$\mathcal{L}_{\text{BE}} = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}[(Q(s,a) - (r+ \gamma\mathbb{E}_{a'\sim\pi(\cdot|s')}[Q'(s',a')]))^2]. \tag{24}$$

其中: (s,a,r,s')为从离线数据集$\mathcal{D}$中采样的状态-动作-奖励-下一状态样本; γ为折扣因子; $Q'(s',a')$为目标网络的Q值, 用于保持训练稳定性. 然后, 为了抑制 OOD 动作的高估值, CQL 引入以下正则化项 (CQL(H) 变体):

$$\mathcal{L}_{\mathrm{CQL}}=\mathbb{E}_{s\sim\mathcal{D}}\Big[\log\sum_{a}\exp(Q(s,a))-\mathbb{E}_{a\sim\hat{\pi}_{\beta}(\cdot|s)}[Q(s,a)]\Big]. \tag{25}$$

这里: 第 1 项$\log\sum_{a}\exp(Q(s,a))$基于Q值的概率分布进行策略选择; 第 2 项$\mathbb{E}_{a\sim\hat{\pi}_{\beta}(\cdot|s)}[Q(s,a)]$对行为策略下的$Q$值进行最大化, 确保对行为策略内的动作值不会被低估. 这种设计增大了数据分布内动作与分布外动作间的Q值差距, 从而有效抑制了 OOD 动作的选择. 最后, 每次更新后, Q值目标网络通过软更新与主网络进行一致性同步, 即

$$\theta'\leftarrow\tau\theta+(1-\tau)\theta', \tag{26}$$

其中τ为更新步长.

CQL 算法从理论上提供了安全策略改进保证. 通过学习Q函数下界, CQL 确保了最终策略至少能够达到估计价值, 避免了过于乐观的性能预测. 此外, CQL 算法使得策略偏向于选择数据分布内的动作, 有效防止了分布偏移带来的负面影响.

2.3.2　拉格朗日法

拉格朗日法是一种常用于解决约束优化问题的数学方法, 在机器学习、控制理论等领域具有广泛应用. 其核心思想为通过引入拉格朗日乘子, 将约束条件与目标函数相结合, 从而转化为无约束优化问题进行求解.

在经典的优化问题中, 假设希望优化的目标函数为$F(x)$, 且该问题受到约束$g(x)=0$的限制. 拉格朗日法通过引入拉格朗日乘子λ, 构建拉格朗日函数, 如下所示:

$$\mathcal{L}(x,\lambda)=F(x)-\lambda\cdot g(x). \tag{27}$$

在这个过程中, λ被称为拉格朗日乘子, 用于表示约束条件$g(x)=0$对目标函数$F(x)$优化的影响. 通过对拉格朗日函数$\mathcal{L}(x,\lambda)$分别关于x和λ求偏导数并设为 0, 可得到最优解.

本节采用拉格朗日法处理强化学习中的约束优化问题. 在实际的强化学习任务中, 智能体不仅需要最大化累积奖励, 还需要确保在学习和执行过程中满足各种约束. 这种多目标优化问题可通过拉格朗日法进行有效处理. 通过拉格朗日法, 算法可将有约束的策略优化问题转化为无约束的极小极大问题, 从而在优化过程中动态调整安全性要求的权重. 拉格朗日法的核心思想是通过拉格朗日乘子λ对约束违反程度进行惩罚, 使得优化过程在最大化奖励的同时满足约束条件.

在该框架下, 选取前文中的综合风险度量S_{score}作为风险约束函数$c(s_t,a_t)$. 作为衡量博弈风险的指标, $c(s_t,a_t)$综合考虑了生存概率、威胁指数以及飞机机动能力等因素, 反映了当前策略在环境中面临的风险. 风险约束阈值设置为c_{thresh}, c_{thresh}与理论建模中的d_s含义一致, 仅用于区分仿真参数设定.

2.3.3　基于风险约束的保守值函数算法

基于风险约束的保守值函数算法在优化过程中明确区分了奖励最大化目标和风险约束条件, 通过拉格朗日法原理将两者统一在一个优化框架中, 进一步增强了强化学习算法的安全性保障.

本文中所需满足约束为$\mathbb{E}[c(s_t,a_t)]\leqslant c_{\mathrm{thresh}}$. 算法的拉格朗日函数结合了 CQL 损失与风险约束惩罚项, 具体如下所示:

$$\mathcal{L}(\pi,\lambda)=\mathcal{L}_{\text{CQL-total}}+\lambda\cdot(\mathbb{E}[c(s_t,a_t)]-c_{\mathrm{thresh}}). \tag{28}$$

其中: $\mathcal{L}_{\text{CQL-total}}$为完整的保守$Q$学习损失函数, 包含贝尔曼误差和保守正则化项; $\lambda\geqslant 0$为对应约束的拉格朗日乘子. 该问题可建模为以下 minmax 形式:

$$\min_{\pi}\max_{\lambda\geqslant 0}\mathcal{L}(\pi,\lambda)=\min_{\pi}\max_{\lambda\geqslant 0}[L_{\text{CQL-total}}(\pi)+\lambda(\mathbb{E}[c(s_t,a_t)]-c_{\mathrm{thresh}})]. \tag{29}$$

对应的拉格朗日对偶形式为

$$\max_{\lambda\geqslant 0}\min_{\pi}\mathcal{L}(\pi,\lambda)=\max_{\lambda\geqslant 0}\min_{\pi}[L_{\text{CQL-total}}(\pi)+\lambda(\mathbb{E}[c(s_t,a_t)]-c_{\mathrm{thresh}})]. \tag{30}$$

算法通过交替更新策略参数和拉格朗日乘子来实现对该原-对偶问题鞍点的近似求解, 从而在训练过程中自适应地平衡奖励最大化与风险约束.

拉格朗日乘子λ的更新基于约束违反程度按照如下方式进行动态调整:

$$\lambda_{t+1}=\mathrm{clip}(\lambda_t+\eta_{\lambda}(c(s_t,a_t)-c_{\mathrm{thresh}}),0,\lambda_{\max}). \tag{31}$$

其中: η_{λ}为乘子更新的步长, $\lambda_{\max}$为拉格朗日乘子的最大限制值.

当满足约束时, 拉格朗日乘子减小, 从而允许策略更加关注奖励最大化; 当违反约束时, 拉格朗日乘子增大, 导致总体损失增加, 从而促使策略更加关注安全性.

在经典约束马尔可夫决策过程 (CMDP) 理论中, 若将随机策略表示为状态-动作占用度量, 则所有可行策略在该占用度量空间中构成凸集, 且期望累积奖励和约束成本均可表示为该空间上的线性泛函, 因此, 原优化问题等价于一个线性规划问题. 在适当的可行性假设及 Slater 条件等正则性条件下, 可证明原问题与拉格朗日对偶问题间不存在对偶间隙,

即强对偶性成立[19]. 本文在这一理论框架的启发下, 将风险约束并入 CQL 损失函数, 通过近似的原始-对偶更新来自适应调节风险约束惩罚系数. 在基于深度神经网络的实际实现中, 由于策略参数空间不再严格凸, 无法直接沿用上述强对偶性结论, 但是, 该拉格朗日更新机制仍然可视为一种合理的自适应惩罚方法, 后续实验结果也表明该机制在复杂空中博弈场景中具有较好的约束满足能力和性能表现.

3 实验设计与结果分析

3.1 实验设计

本文在“智空仿真推演与训练平台”(以下简称智空平台) 上开展实验. 该平台具有六自由度飞机和弹体动力学, 确保模拟过程能够精确再现实际的飞行力学特性, 为智能体提供与实战环境接近的训练条件. 此外, 平台配备了完善的数据接口, 支持多种数据格式, 便于数据的采集和后续分析.

实验采用 8 v 8 对抗场景, 红蓝双方各部署 8 架飞机. 其中: 红方为己方智能体, 蓝方为采用高级防守策略的对抗智能体. 实验设置 batch_size = 32, train_steps = 20 M. 蓝方智能体具备动态攻防转换能力, 采用迂回战术, 在完成导弹发射或发现敌方目标后, 能够迅速进行战术转换, 通过阵型优化以确保攻防间的连续性, 博弈性能卓越. 图 4 为面对高级防守策略对手的博弈序列.

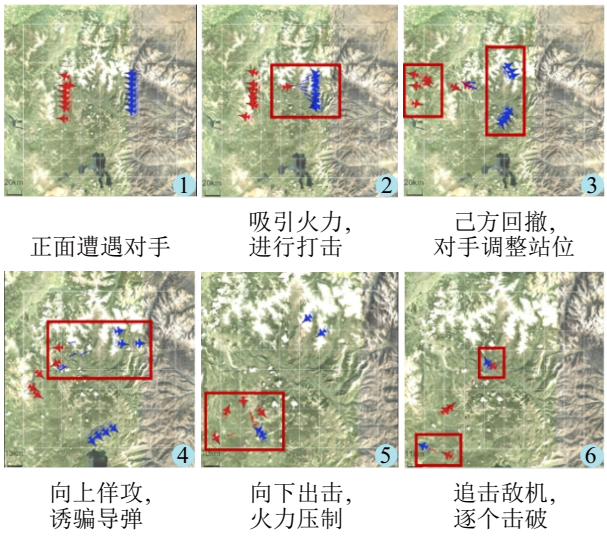

图4 面对高级防守策略对手博弈序列

为评估所提出方法的性能, 本节将其与在复杂对抗领域表现较为出色的 4 种不同的算法进行对比, 具体如下:

1) 遗传模糊树[20]: 遗传模糊树 (GFT) 方法结合了遗传算法的全局搜索能力和模糊逻辑处理不确定性的优势. GFT 将问题空间划分为树状结构, 每个节点表示模糊逻辑规则, 通过遗传算法优化这些规则. 本文中, GFT 以导弹数量、与敌机的角度和距离作为输入变量, 通过隶属度函数将其抽象为模糊规则, 并使用遗传算法进行训练优化, 输出第 1 节中所提出的 4 种中层策略的选择概率. 实验设置 population_size = 50, epochs = 1 000.

2) 多智能体近端策略优化算法[21]: 多智能体近端策略优化算法 (MAPPO) 是一个广泛应用于多智能体环境中的在线强化学习方法. MAPPO 通过智能体间的协作框架, 使得各智能体可以共享信息并共同更新策略, 从而在动态环境中持续进行学习. 在本文中, MAPPO 算法采用第 1 节所提出的建模方式, 训练顶层博弈智能体. 实验设置 batch_size = 128, train_steps = 10 M.

3) 软动作-评价算法[22]: 软动作-评价算法 (SAC) 是一种稳定高效的无模型离线策略最大熵强化学习方法. SAC 通过在目标函数中引入策略熵, 增强了智能体的探索能力, 使其能够捕捉到多种近似最优的行为模式. 该算法结合了离线更新和稳定的随机动作-评价框架, 在连续状态和动作空间中表现出色. 在本文中, SAC 算法也采用第 1 节所提出的建模方式, 训练顶层博弈智能体. 实验设置 batch_size = 256, train_steps = 10 M.

4) 隐式Q学习算法[11]: 隐式Q学习算法 (IQL) 是一种适用于离线强化学习的高效策略学习方法. IQL 通过将策略学习与价值函数估计解耦, 在不依赖动作生成的情况下直接从离线数据集中提取最优行为, 从而有效避免了对分布外动作的依赖. 在本文中, IQL 算法的离线数据集选用扩散模型经验增广后生成的合成示教数据集, 同样采用第 1 节所提出的分层建模方式, 用于训练顶层博弈智能体, 以验证其在多机空中对抗场景下的离线策略学习能力. 实验设置 batch_size = 32, train_steps = 20 M.

3.2 算法训练过程分析

3.2.1 对比分析

下面对所提出面向智能空中博弈的风险约束离线强化学习算法 CQL-Safe 进行算法性能验证.

如图 5 所示, 本节首先对比分析所提出 CQL-Safe 算法与主流在线学习算法及离线学习算法在训练过程中的性能差异. 其中: 纵轴为归一化后的平均回报值, 取 $[-1, 1]$; 横轴为训练时间. 由图 5 的训练曲线可以清晰地看出, CQL-Safe 算法在训练初期便展现出明显的性能优势. CQL-Safe 在训练的前 2 h 内回报值迅速由初始的−0.8 提升至约 0.2, 并在随后的训练过程中持续提升, 最终于 12 h 后稳定在约 0.8

的高水平. 整个训练过程中, 训练曲线变化平缓, 波动幅度较小, 性能表现稳定.

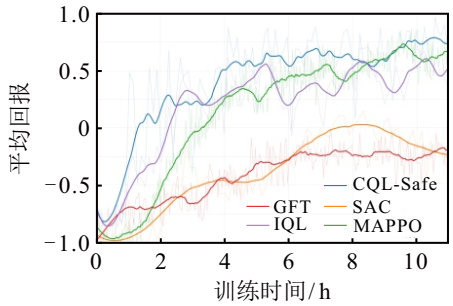

图5　算法性能对比

相比之下, 对比算法的训练表现存在明显不足. 尽管 MAPPO 算法在最终性能上也能达到较高水平, 但是, 其初期训练进程明显缓慢. 在前 4 h 内回报值仅缓慢提升, 直至训练 8 h 后才逐渐接近 CQL-Safe 的表现; SAC 算法则表现更为曲折, 在训练 6 ～ 8 h 区间内回报值短暂达到约 0.1 的峰值后, 性能出现显著波动并有所下降; 而 GFT 算法在整个训练过程中始终表现较为保守, 回报值始终维持在较低水平, 最终仅达到约−0.2; 而 IQL 算法虽然在训练初期提升较快, 但是中后期波动较大, 最终性能也与 CQL-Safe 算法存在一定差距.

3.2.2　消融实验分析

为进一步验证算法中各模块的作用, 本节对比基于风险约束和保守值函数学习的离线博弈算法 CQL-Safe 与未加入风险评估函数的 CQL 算法 (CQL-NoSafe) 以及数据集未进行扩散模型经验增广的 CQL 算法 (CQL-NoDiffusion) 的性能表现, 并展示消融实验, 如图 6 所示.

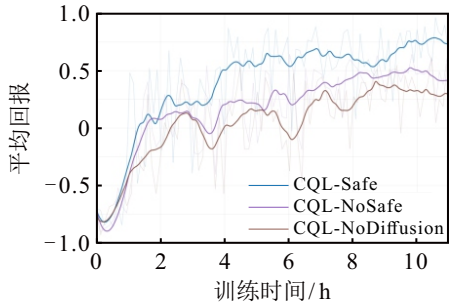

图6　消融实验

由图 6 可知, CQL-Safe 算法在训练稳定性和最终性能方面均具有明显优势. 从训练曲线可以观察到, 三者在训练初期均经历快速的性能提升, 前 2 h 回报值均迅速从−0.8 提升. 然而, 在训练中后期阶段, 3 种算法的表现产生明显分化: CQL-Safe 算法在 4 h 左右已达到约 0.5 的回报值, 并于后期稳定在 0.7 ～ 0.8 之间, 展现出显著的性能优势和稳定性. 相比之下, 其余二者性能提升较慢, 最终均未达到 0.5 的回报值水平.

这种性能差异是由于 CQL-Safe 算法的数据集采用扩散模型进行经验增广, 生成了高质量的合成示教数据, 从而有效扩充了训练集, 提高了模型的泛化能力. 同时, 算法引入了自适应风险更新机制, 该机制根据实时博弈环境的风险评分动态调整策略更新强度. 该机制综合考虑了博弈持续时间、敌我飞机数量对比以及威胁程度等多个因素, 使得训练过程更加安全和高效.

3.3　测试胜率分析

在测试阶段, 为了更直观地展示所提出 CQL-Safe 算法与其他方法在面对高级防守对手策略时的表现, 对不同算法对战高级防守策略的胜负情况进行分析, 具体如下.

表 1 为不同算法在面对高级防守对手策略时的胜率对比情况.

表1　不同算法对阵高级防守对手策略的胜平负率统计

算法	胜率/%	平率/%	负率/%
CQL-Safe	**94.0**	2.5	3.5
CQL-NoSafe	92.5	0.0	7.5
CQL-NoDiffusion	91.5	0.0	8.5
IQL	90.0	4.0	6.0
MAPPO	91.0	5.0	4.0
SAC	42.5	18.0	39.5
GFT	32.0	18.0	50.0

由表 1 可见: 所提出 CQL-Safe 算法胜率达到了最高的 94%; 与此同时, CQL-NoSafe 算法、CQL-NoDiffsuion 算法、IQL 算法和 MAPPO 算法的胜率分别为 92.5%、91.5%、90% 和 91%, 虽然 3 种算法在博弈对抗性能上与 CQL-Safe 算法相差不大, 但是负率均有所增加.

相较之下, SAC 算法胜率仅为 42.5%, 同时, 表现出较高的平局率 18% 和负率 39.5%. 由此可见, SAC 算法在面对复杂且动态的高级防守对手时难以做出有效的决策, 出现较多的平局和失败情况; 而 GFT 算法胜率只有 32%, 这表明遗传模糊树决策方法在处理动态变化较快的空中博弈环境时, 存在明显的局限性.

本节还设计了不同水平和风格的多种对手策略, 这些对手策略均由专家基于领域知识和任务规则精心构建, 用于全面评估所提出算法在不同博弈场景下的对抗性能和泛化能力. 各对手策略的特点如下:

高级进攻对手策略: 构建了一个整合智能火力分配与机动协同的进攻体系, 通过高效打击战略有

效压制对手, 同时, 在保持高强度进攻态势下具备灵活调整战术的能力.

中级对手策略: 采用标准战术执行方案, 能够完成标准攻防动作, 较好地应对典型博弈对抗场景.

初级对手策略: 采用基础战术实施方案, 能够完成基本的攻防动作, 执行基本的战术流程.

生存优先策略: 侧重于提升博弈生存能力, 通过最优规避算法来实现高效闪避, 生存效果优异.

本文针对高级进攻对手策略、中级对手策略、初级对手策略以及生存优先策略等典型对抗模式展开测试. 表 2 为不同算法对阵各类对手策略的胜率分布情况.

表2 不同算法对阵各类对手策略的胜平负率统计

算法	高级进攻对手			中级对手		
	胜率/%	平率/%	负率/%	胜率/%	平率/%	负率/%
CQL-Safe	**89.5**	5.0	5.5	**100.0**	0.0	0.0
CQL-NoSafe	84.0	0.0	16.0	99.5	0.0	0.5
CQL-NoDiffsuion	85.5	2.0	12.5	98.5	0.0	1.5
IQL	82.0	5.0	13.0	99.0	0.0	1.0
MAPPO	**90.0**	2.5	7.5	**100.0**	0.0	0.0
SAC	45.0	15.0	40.0	78.0	2.0	20.0
GFT	35.0	15.0	50.0	72.5	3.0	24.5

算法	初级对手			生存优先		
	胜率/%	平率/%	负率/%	胜率/%	平率/%	负率/%
CQL-Safe	**100.0**	0.0	0.0	**46.5**	53.5	0.0
CQL-NoSafe	**100.0**	0.0	0.0	36.5	63.5	0.0
CQL-NoDiffsuion	**100.0**	0.0	0.0	40.5	59.5	0.0
IQL	**100.0**	0.0	0.0	38.0	62.0	0.0
MAPPO	**100.0**	0.0	0.0	41.5	58.5	0.0
SAC	95.0	0.0	5.0	22.5	77.5	0.0
GFT	80.5	4.5	15.0	18.0	82.0	0.0

表 2 测试数据显示: CQL-Safe 算法在应对高级进攻策略时展现出 89.5% 的胜率, 与 MAPPO 算法 90% 的性能水平相当; CQL-NoSafe 算法、CQL-NoDiffsuion 算法与 IQL 算法以 84%、85.5% 和 82.0% 的胜率稍逊一筹; 而 SAC 和 GFT 算法的胜率均不足 50%. 在中级和初级对抗场景中: CQL-Safe 算法保持 100% 的完胜记录, 与 MAPPO 算法表现相当; CQL-NoSafe 算法、CQL-NoDiffsuion 算法和 IQL 算法在中级对抗中以 99.5%、98.5% 和 99.0% 的胜率稍显逊色; SAC 与 GFT 算法虽然较前有所提升, 但是仍然显著落后于其他几类算法.

针对生存优先策略的对抗测试中, 各算法平局率普遍较高, 这体现了生存优先策略侧重生存回避的特性. CQL-Safe 算法以 46.5% 的胜率展现出较强的主动制敌能力; 相较之下, GFT 与 SAC 算法分别仅取得了 18% 和 22.5% 的胜率, 追击能力明显不足.

综上, CQL-Safe 算法在各类对抗测试场景中均展现出显著优势, 相较于其他算法, CQL-Safe 在保持高胜率的同时, 通过安全机制有效控制了模型参数更新的幅度, 为复杂博弈场景下的智能决策提供了可靠解决方案.

4 结 论

本文针对智能空中博弈场景提出了一种结合风险约束与改进保守值函数学习的离线强化学习算法(CQL-Safe). 通过融合示教学习与生成式扩散模型的数据集构建算法, 有效突破了传统离线强化学习的数据困境; 通过设计风险评估函数并引入拉格朗日法, 在保证保守值函数算法性能的同时实现了风险约束的动态平衡, 为智能体在对抗环境中的稳健决策提供了新思路. 实验验证表明, 所提出的算法在多种对抗场景下均展现出了较快的收敛速度和较高的策略质量, 其安全性能和泛化能力相较于现有方法提升显著.

参考文献 (References)

[1] 王国岩, 赵旭华, 解宇轩, 等. 基于态势感知的无人机空战协同决策方法[J]. 控制与决策, 2025, 40(6): 1847-1854.
(Wang G Y, Zhao X H, Xie Y X, et al. A collaborative decision-making method for unmanned aerial vehicles in aerial combat based on situational awareness[J]. Control and Decision, 2025, 40(6): 1847-1854.)

[2] 施伟, 冯旸赫, 程光权, 等. 基于深度强化学习的多机协同空战方法研究[J]. 自动化学报, 2021, 47(7): 1610-1623.

(Shi W, Feng Y H, Cheng G Q, et al. Research on multi-aircraft cooperative air combat method based on deep reinforcement learning[J]. Acta Automatica Sinica, 2021, 47(7): 1610-1623.)

[3] Zhu J Y, Kuang M C, Zhou W Q, et al. Mastering air combat game with deep reinforcement learning[J]. Defence Technology, 2024, 34: 295-312.

[4] 孙辉辉, 胡春鹤, 张军国. 事件触发式多智能体分层安全强化学习运动规划[J]. 控制与决策, 2024, 39(11): 3755-3762.
(Sun H H, Hu C H, Zhang J G. Multi-agent event triggered hierarchical security reinforcement learning[J]. Control and Decision, 2024, 39(11): 3755-3762.)

[5] Qian C X, Zhang X B, Li L, et al. A partial joint optimization algorithm for autonomous air combat based on hierarchical reinforcement learning[J]. IEEE Transactions on Cybernetics, 2025, 55(9): 4145-4157.

[6] 王雪松, 张恒瑞, 张佳志, 等. 基于优势约束扩散策略的离线强化学习[J]. 控制与决策, 2025, 40(6): 1903-1912.
(Wang X S, Zhang H R, Zhang J Z, et al. Offline reinforcement learning based on advantage-constrained diffusion policy[J]. Control and Decision, 2025, 40(6): 1903-1912.)

[7] Mandlekar A, Xu D, Wong J, et al. What matters in learning from offline human demonstrations for robot manipulation[J/OL]. 2021, arXiv: 2108.03298.

[8] Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: A review[J]. Medical Image Analysis, 2019, 58: 101552.

[9] Oh J, Kim S, Kim G, et al. Diffusion-based episodes augmentation for offline multi-agent reinforcement learning[J/OL]. 2024, arXiv: 2408.13092.

[10] Kumar A, Zhou A, Tucker G, et al. Conservative *Q*-learning for offline reinforcement learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 1179-1191.

[11] Kostrikov I, Nair A, Levine S. Offline reinforcement learning with implicit *Q*-learning[J/OL]. 2021, arXiv: 2110.06169.

[12] 王臆淞, 赵铭慧, 张雪波. ASM^2: 面向海空联合场景的多对手多智能体博弈算法[J]. 控制理论与应用, 2025, 42(7): 1275-1284.
(Wang Y S, Zhao M H, Zhang X B. ASM^2: Multi-agent multi-opponent game algorithm for joint sea-air scenarios[J]. Control Theory & Applications, 2025, 42(7): 1275-1284.)

[13] Gu S D, Yang L, Du Y L, et al. A review of safe reinforcement learning: Methods, theories, and applications[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 11216-11235.

[14] Li L, Zhang X B, Qian C X, et al. Basic flight maneuver generation of fixed-wing plane based on proximal policy optimization[J]. Neural Computing and Applications, 2023, 35(14): 10239-10255.

[15] Qian C X, Zhang X B, Li L, et al. H3E: Learning air combat with a three-level hierarchical framework embedding expert knowledge[J]. Expert Systems with Applications, 2024, 245: 123084.

[16] Chen J. Reinforcement learning and swarm intelligence for cooperative aerial navigation and payload transportation[D]. Sheffield: University of Sheffield, 2024.

[17] 刘瑞峰, 王家胜, 张灏龙, 等. 行为树技术的研究进展与应用[J]. 计算机与现代化, 2020(2): 76-82.
(Liu R F, Wang J S, Zhang H L, et al. Research progress and application of behavior tree technology[J]. Computer and Modernization, 2020(2): 76-82.)

[18] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851.

[19] Paternain S, Chamon L, Calvo-Fullana M, et al. Constrained reinforcement learning has zero duality gap[J]. Advances in Neural Information Processing Systems, 2019, 32: 7491-7501.

[20] Ernest N, Cohen K, Kivelevitch E, et al. Genetic fuzzy trees and their application towards autonomous training and control of a squadron of unmanned combat aerial vehicles[J]. Unmanned Systems, 2015, 3(3): 185-204.

[21] Yu C, Velu A, Vinitsky E, et al. The surprising effectiveness of PPO in cooperative multi-agent games[J]. Advances in Neural Information Processing Systems, 2022, 35: 24611-24624.

[22] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]. Proceedings of the 35th International Conference on Machine Learning. Stockholm, 2018: 1861-1870.

作者简介

李博文 (2003−), 男, 硕士生, 主要研究方向为强化学习与智能博弈, E-mail: 2120240629@mail.nankai.edu.cn;

王臆淞 (2000−), 男, 硕士, 主要研究方向为多智能体强化学习, E-mail: 2120220504@mail.nankai.edu.cn;

赵铭慧 (1995−), 女, 中级实验师, 硕士, 主要研究方向为强化学习与智能博弈, E-mail: zhaomh@nankai.edu.cn;

骞晨旭 (1999−), 男, 博士生, 主要研究方向为强化学习与智能博弈, E-mail: qianchenxu@mail.nankai.edu.cn;

程光权 (1982−), 男, 研究员, 博士, 主要研究方向为复杂网络分析与决策支持技术、链路预测, E-mail: cgq299@nudt.edu.cn;

张雪波 (1984−), 男, 教授, 博士, 主要研究方向为机器人与人工智能、移动机器人视觉控制, E-mail: zhangxuebo@nankai.edu.cn.