

基于奖励塑造强化学习的智能导弹突防策略规划方法

罗俊仁, 刘果, 苏炯铭, 张万鹏, 陈璟

(国防科技大学智能科学学院, 湖南 长沙 410073)

摘要: 面向未来海上分布式作战需求, 以对抗环境下分布式作战场景中智能导弹齐射突防水面舰艇为背景, 首先, 分析了导弹突防策略规划问题; 其次, 设计了基于奖励塑造强化学习的智能导弹突防策略规划方法; 然后, 在墨子联合作战推演系统上构建导弹突防舰艇的作战想定, 实验结果表明本文方法的智能导弹突防打击成功率为79%, 验证了基于奖励塑造强化学习方法的有效性; 最后, 经过复盘发现, 奖励塑造实验涌现出智能导弹的4类对舰打击突防策略: 集中迂回攻击、分散突防多向攻击、分组延时攻击、巡弋探测指引攻击。

关键词: 穿透性反舰; 智能导弹; 齐射突防; 任务规划; 兵棋推演

中图分类号: TP39

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202411

Reward shaping based reinforcement learning for intelligent missile penetration attack strategy planning

LUO Junren, LIU Guo, SU Jiongming, ZHANG Wanpeng, CHEN Jing

College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China

Abstract: Facing the future requirements of distributed warfare at sea, the strategic planning of missile penetration is firstly analyzed based on the background of intelligent missile salvo penetration against surface ships in distributed warfare scenario. Secondly, a strategic planning method of intelligent missile penetration based on reward-shaping reinforcement learning is designed by using multi-class reward function. Then, the operation scenario of the missile penetration ship is constructed on the Mozi joint operation simulation system. The comparison experiment shows that the success rate of the intelligent missile penetration attack controlled by the model learned by the reward molding method is 79%, which verifies the effectiveness of the reward-based reinforcement learning method. Finally, after action review, it is found that there are emerging four kinds of penetration strategies of intelligent missiles in the reward shaping experiment, including concentrated and roundabout attack, scattered penetration multi-direction attack, group delay attack and cruise detection guide attack.

Key words: penetrating anti-ship, intelligent missile, salvo penetration, task planning, wargaming

0 引言

现代战争通常通过“空袭与反空袭”拉开序幕。面对“环境高动态、信息不完备、博弈强对

抗、响应高实时、边界不确定”的未来战场,“穿透突防”与“防空反导”作为一对非对称孪生问题,已然成为当前国防安全领域的热门研究方向。近年来,为应对日益突出的导弹威胁,部分国家积

收稿日期: 2023-10-22; 修回日期: 2024-04-30

通信作者: 苏炯铭, sujiongming07@nudt.edu.cn

基金项目: 国家自然科学基金项目 (No.61806212); 湖南省研究生创新项目 (No.CX20210011)

Foundation Items: The National Natural Science Foundation of China(No.61806212), The Graduate Innovation Project of Hunan Province(No.CX20210011)

极研发和部署导弹防御系统。同时,为突破竞争对手的导弹防御系统,部分国家展开了智能导弹相关技术研究。为提高智能导弹的突防能力,当前研究主要聚焦两个要点。一是采用多弹头技术或集群协同方法。例如,挪威海军打击导弹^[1]具备多目标打击能力,以协同作战的模式提高整体突防能力。有关反舰导弹突破“宙斯盾”系统的研究也指出,机动式多弹头与弹头集群的协同能显著提升打击效果^[2]。二是提升单枚导弹的隐身性能或智能化水平,例如,采用红外隐身、雷达隐身、激光隐身技术等降低被探测系统发现的可能性^[3],通过加装弹载传感器或增加与外部实体的无线通信增强主动感知战场态势的能力,自主研判后提供电子干扰、诱饵欺骗、大范围机动等突防方案,从而增强突防能力。例如,美国远程反舰导弹^[4]增加了许多智能化模块,减少了对中继信息的依赖,在网络被切断的情况下也能够“自主寻的”式打击目标;美国战斧巡航导弹^[5]具备一定的航迹自主规划能力,能有效规避防空拦截,进而提升突防效果。

智能导弹的突防是其作战运用无法回避的问题。随着导弹智能化程度的不断提高,导弹突防的手段也越来越多,如分导弹头突防、隐身突防、机动突防、电子干扰、多导弹协同突防等。对于预定的打击目标,利用确定数量的导弹,如何通过智能突防提升打击成功率是一个值得研究的问题。

2015年,DeepMind提出深度Q学习网络(deep Q-learning network, DQN)算法^[6],拉开了深度强化学习算法研究的序幕^[7]。当前,深度强化学习主要包括值函数、策略梯度和行动者-评价者(actor-critic, AC)三大类方法。基于值函数的强化学习方法主要有双重DQN(double DQN)方法^[8]、竞争架构DQN方法^[9]等。基于策略梯度的强化学习方法主要有近端策略优化(proximal policy optimization, PPO)方法^[10]、双延迟深度确定性策略梯度(twin delayed deep deterministic policy gradient, TD3)^[11]等。基于AC框架衍生的变体有深度确定性策略梯度(deep deterministic policy gradient, DDPG)^[12]、异步优势AC(asynchronous advantage actor-critic, A3C)^[13]、软AC(soft actor-critic, SAC)方法^[14]等。如何在对抗条件下增强导弹突防能力,一些研究利用强化学习方法在作战仿真推演平台进行探索。王国岩等^[15]利用墨子系统分析了对舰打击的突防策略、攻击策略和补射策略。肖友刚

等^[16]研究了基于深度强化学习的舰船防空导弹的目标分配问题。张斌等^[17]利用离散SAC强化学习算法和武器装备作战效能仿真系统(Weapon effectiveness simulation system, WESS)构建了导弹智能决策训练环境。黄卓等^[18]利用分区间强化学习研究了集群导弹任务的快速分配问题。梁星星等^[19]设计了基于预测编码强化学习的行动策略规划方法。

本文在“分布式杀伤”场景下,围绕穿透性反舰任务中智能导弹突防策略规划问题,利用墨子兵棋系统设计了典型场景想定。针对奖励稀疏问题,利用奖励塑造强化学习的方式,动态生成多组智能导弹突防航线,从中挖掘有效的突防策略。

1 导弹突防策略规划问题分析

1.1 导弹齐射问题概述

在分布式海战中,通常采用导弹齐射方式实施对敌舰艇的穿透式打击,其中齐射方式主要有两种:流式齐射和饱和齐射^[20]。于忠琳等^[20]研究了无(带)航路规划齐射方式的反舰导弹发射时间控制问题。

1.2 导弹毁伤与拦截建模

本文研究内容是导弹对舰艇的打击,打击成功与否受到诸多因素的影响,且存在一定程度的随机性。本节主要对导弹命中精度、导弹命中舰艇后对舰艇毁伤程度、导弹被拦截弹拦截概率的计算及判定进行形式化描述。

1.2.1 导弹命中精度

圆形公算误差,也叫圆概率误差(circular error probable, CEP),又称圆公算偏差,用于衡量导弹命中精度。在同一条件下,向同一个预定目标发射多枚导弹,受瞄准误差、制导系统误差以及气象条件等因素的影响,导弹的弹着点会在目标附近形成散布,其散布中心到瞄准点的距离称为导弹的系统误差,而每个弹着点到平均弹着点(散布中心)的距离称为随机误差。通常情况下,系统误差可以修正且比随机误差小,因此,瞄准点被近似视为平均弹着点。圆形公算误差的通俗描述是:以瞄准点为中心,包含50%弹着点的圆半径。半径越小,说明导弹的命中精度越高。如美国带常规弹头的BGM-109C型“战斧”巡航导弹采用了惯性加地形匹配以及景象匹配制导,其圆概率误差只有9 m。

可以推导出,在任意概率 P 条件下,某型导弹的散布半径 R_p 与散布半径标准差 σ 的关系如下:

$$\exp\left[-\frac{(R_p)^2}{2\sigma^2}\right] = 1 - P \quad (1)$$

通常圆形公算误差是指 $P=50\%$ 时, R_p 的值, 也就是 R_{50} :

$$R_{50} = \sqrt{2\ln 2} \sigma \quad (2)$$

1.2.2 导弹对舰艇的毁伤

导弹战斗部对舰艇这一作战平台的毁伤主要从毁伤程度和贯穿程度两个因素来考虑, 从而确定毁伤部件数量, 而毁伤部件数量与舰艇所有部件数量的比值代表舰船最终的系统毁伤比例, 当系统毁伤比例达到 100% 时, 判定舰艇被摧毁。

(1) 毁伤程度

当前战斗部造成的毁伤当量 DP 与平台初始可承受毁伤当量 DP_0 相除, 得到作战平台的结构毁伤程度 f :

$$f = \frac{DP}{DP_0} \quad (3)$$

不同型号的舰艇对应不同的值, 该系数取一位有效数字。

(2) 部件毁伤数量

根据战斗部当次对作战平台的贯穿程度 p , 计算部件毁伤数量:

$$n = \text{Fix}((1+p) \times \text{rand}(1\sim 7)) \quad (4)$$

其中, Fix 为取整函数, rand 为随机函数。根据作战平台的毁伤程度, 判断不同毁伤程度下的部件毁伤数量:

$$n = \begin{cases} n-6, & f < 0.1 \\ n-4, & f = 0.1 \\ n-3, & f = 0.2 \\ n-2, & f = 0.3 \\ n-1, & f = 0.4 \\ n, & f = 0.5 \\ n+1, & f = 0.6 \\ n+2, & f = 0.7 \\ n+3, & f = 0.8 \\ n+4, & f = 0.9 \\ n+5, & f = 1.0 \end{cases} \quad (5)$$

上述计算得到的部件毁伤数量不能超过作战平台包含的部件数量。

1.2.3 导弹的拦截

(1) 拦截流程

导弹打击的目标为具备体系化、层次化防空能力的舰艇, 其拦截流程可简要描述为“发现来袭目标-发射远程拦截导弹-观察拦截效果-仍存来袭目标-发射中程拦截导弹-观察拦截效果-仍存来袭目标-发射近程拦截导弹”。在每个阶段, 如果相应的拦截导弹有余量且射程与发射间隔时间在允许的情况下, 可以实施多次拦截, 对于同一目标, 拦截导弹遵循同一预设的作战条令, 如 1 枚拦截 1 枚或 2 枚拦截 1 枚。

(2) 拦截概率

防空系统拦截弹对导弹的拦截概率由 3 部分组成, 分别是对空 P_{oh} 、角度修正概率 P_a 、目标特征修正概率 P_{tf} 。其中, 对空 P_{oh} 表示该型导弹对空中目标进行拦截的可靠率, 即发射后不出现故障的概率, 可以理解为对空中待拦截目标的基准命中概率, 主要由该型导弹的统计数据生成。角度修正概率 P_a 指的是拦截弹与拦截目标的不同角度关系对命中概率的影响, 一般为负值。目标特征修正概率 P_{tf} 则代表拦截目标的隐身性能对命中概率的影响, 一般为负值。故而最终导弹的拦截概率 P_i 为:

$$P_i = P_{\text{oh}} + P_a + P_{\text{tf}} \quad (6)$$

采用多枚防空拦截弹来拦截导弹的情形, 可建模成多阶段拦截模型^[21]。

(3) 拦截是否成功的判断

主要采用随机数生成判定法, 根据拦截流程, 计算每枚拦截导弹的拦截概率 P_i , 然后在 $[0\sim 100]$ 抽取随机数 N , 如 $N\%$ 小于 P_i , 则判定命中目标, 否则判定为脱靶。

1.3 导弹突防策略规划

1.3.1 任务规划

智能导弹的任务规划是其作战运用中的关键问题, 对于以战斧巡航导弹建模的导弹及其集群来说, 无论是单一导弹还是多枚导弹, 其任务规划问题都十分复杂。

一般而言, 单一的导弹的任务规划问题通常包含发射时间、目标选择、航迹规划等。其任务规划的好坏, 直接影响该导弹的突防概率、命中概率、毁伤概率等。当导弹作为集群运用且目标为多目标时, 其任务规划问题愈加复杂, 除了打击波次选择、目标分配、发射时序确定、航迹规划外, 还存在与其他单元或武器平台作战协同相关的问题。然

而，随着未来智能化战争的来临与导弹的智能化发展，智能导弹的任务规划将增加自主决策时机、方式以及内容等，对未来智能导弹的任务规划也提出了更高的要求。其任务规划如图1所示。

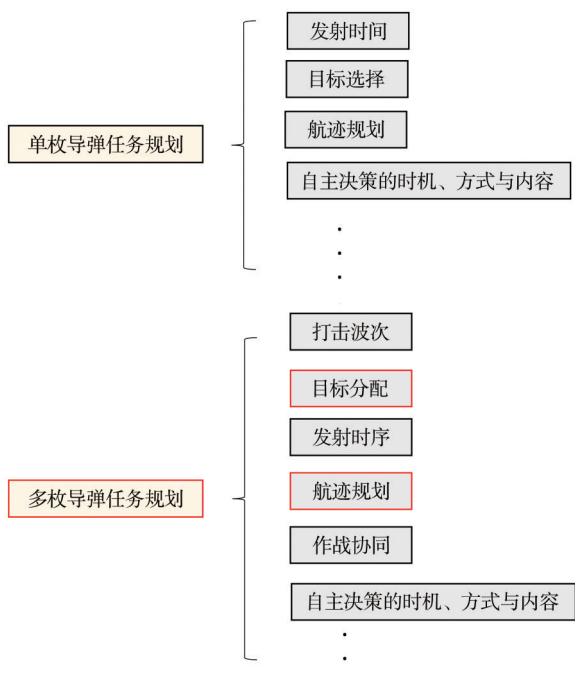


图1 任务规划

1.3.2 航迹规划

导弹的航迹规划问题可以理解成根据导弹预定的打击目标规划满足相应约束条件的飞行轨迹。

导弹的航迹规划需要确定导弹自控终点的位置，而导弹自控终点的位置与预打击目标的实际距离又将直接影响导弹捕捉目标的概率，从而影响打击成功率。同时，导弹不同的航迹会影响导弹的飞行时间，从而对导弹的发射时序产生影响。在受到地形限制，或者需要规避威胁、避开密集居民区、禁飞区，或是考虑与其他单元达成协同的情况时，也需要对导弹航迹进行规划。

导弹进入蓝方探测与拦截范围，可以被视为进入了突防阶段。在导弹的突防阶段，导弹到目标的直线航迹最短，用时最少，但往往也容易被预测，易被拦截。导弹若在突防阶段能够根据战场态势自主进行连续多向机动，则有望摆脱拦截弹的拦截，增加突防概率。但是，如何实现导弹在突防阶段的连续多向机动、有效进行动态航迹规划则是一个难题。相比于目标分配，导弹的航迹规划更偏向于战术层级，其动态航迹规划的效果将直接影响本波次的突防与打击成效。

2 智能导弹突防策略规划强化学习

2.1 框架构建

本节立足“分布式杀伤”场景，构建智能导弹突防策略规划的强化学习框架。按照OODA环的概念，智能导弹突防策略规划框架如图2所示，其具体运行的模式如下。

首先，将真实对抗环境中的“分布式杀伤”场

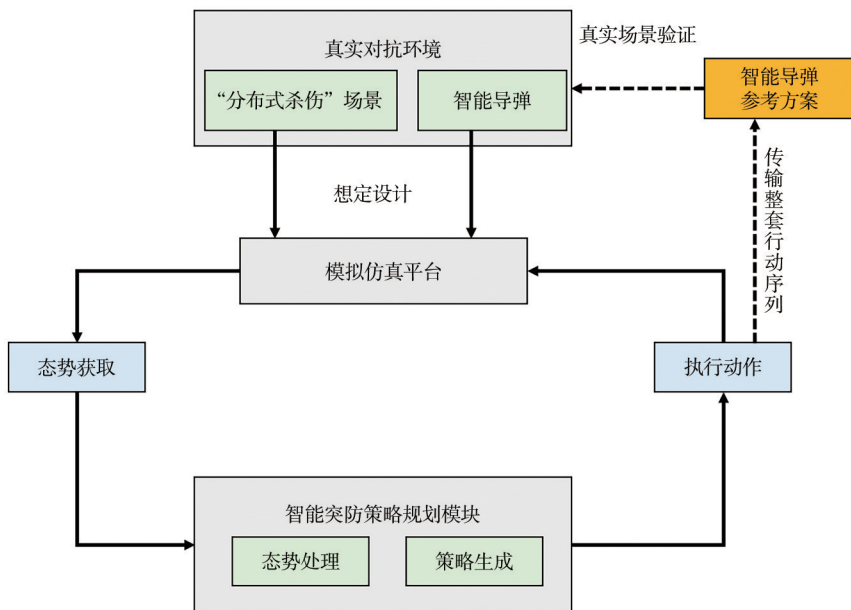


图2 智能导弹突防策略规划框架

景与智能导弹的必要信息通过想定设计导入模拟仿真平台，由模拟仿真平台实现智能导弹与“分布式杀伤”作战态势的映射，以模拟真实环境的战场态势。智能导弹的信息主要包括武器平台类型、发射架数、导弹种类与数量以及地理位置等，“分布式杀伤”作战态势的信息主要包括舰艇编队数量、编队队形样式、具体舰艇类型、携带导弹种类与数量、信息平台类型以及各单元位置、任务具体信息等。

其次，态势获取模块负责接收模拟仿真平台的态势信息，包括智能体自身信息与对蓝方的观测信息，将其输入智能决策模块。智能决策模块再通过态势处理，由已训练好的智能体决策模型生成特定的策略，即具体的行动方案。

最后，执行动作模块根据智能决策模块输出的具体行动方案生成具体的动作。一方面输入模拟仿真平台，更新平台态势，从而循环迭代；另一方面传输的整套行动序列可转化为现实中智能导弹打击参考方案，辅助指挥员决策。

2.2 奖励塑造

2.2.1 原理简介

在强化学习智能体与环境的交互过程中，智能体执行某个动作会获得一定的奖励或者惩罚。由于奖励稀疏问题和高维环境的复杂性，智能体获得的奖励或者惩罚可能会延时到达。当奖励延时到达时，智能体将难以判断是先前做的哪个动作引发的奖励。所以在进行强化学习的过程中，训练环境的奖励稀疏通常会导致长期信用分配问题（即难以确定哪一行为对最终奖励的影响较大）、智能体收敛缓慢以及训练效率低的问题，进而可能导致训练成本过高甚至训练失败。针对这一问题，许多学者和工程师进行了相关的应用。

奖励塑造是一种解决奖励稀疏问题的重要方法^[22]，它的核心思想是通过重新设计奖励函数，为对最终奖励起到关键性作用的行动重设奖励信号，从而使奖励更为稠密，这也对于加快智能体的训练至关重要。但同时，奖励的塑造也需要相关的领域知识或专家经验等，需要根据具体实例进行设计。

2.2.2 具体设计

在基于DQN的智能导弹目标分配的实验中，由于导弹打击过程需要较长时间，在设计具体方法时对获取的训练数据进行筛选与组合，确保收集的

每组训练数据都有确定的奖励值，即智能导弹在每一波次的打击完成后都会获得一定的奖励。故而，在目标分配实验数据中，所得到的奖励相对较为稠密。

本实验的动作设置是控制多个导弹分组的飞行方向，在一局作战推演中，需要多个决策步，导弹才可能击中预定目标，获得一定的分数。若仅考虑得分奖励 r_s ，即使导弹分组最终能够击中舰艇，仍需要多次动作才能获得奖励。若打击失败，则整个对局都没有奖励，而总体上奖励稀疏，又难以获得较好的训练效果。

根据上述分析，除了平台设置的得分奖励 r_s 外，本实验还进行了额外的奖励塑造。塑造的奖励包括两部分，分别是距离奖励 r_d 与导弹损失奖励 r_l 。

距离奖励 r_d 是指在一次决策中，当导弹靠近或者远离目标时所得的奖励。对于每一枚导弹，假设其上一时刻与目标的距离为 d ，下一时刻与目标的距离为 d' ，如果 $d'-d<0$ ，说明导弹在本决策步中在向目标靠近，如果 $d'-d>0$ ，则说明在远离目标。给每枚导弹都设置相应的距离奖励 r_{di} ，其具体大小与弹目（导弹和目标）距离成反比。若导弹数量为 n ，则每一决策步的距离奖励为：

$$r_d = \sum_{i=1}^n r_{di} \quad (7)$$

距离奖励将促使导弹靠近预定打击目标。显而易见，最大的距离奖励值通常在导弹直飞预定目标的情况下得到，但此航线往往存在较大的风险，易被敌预测拦截，故而考虑距离奖励的同时，还应考虑导弹损失奖励。

导弹损失奖励 r_l 是指当导弹被蓝方防空导弹拦截时，即导弹有数量损失时，所得的奖励，导弹损失奖励为负值。假设在一个决策步中有 n_l 枚导弹被拦截，则导弹损失奖励为：

$$r_l = \lambda \cdot n_l \quad (8)$$

其中， λ 为可变系数。导弹损失奖励会促使导弹在突防过程中尽量保存自己，不选择更易被拦截的航线。

综上所述，最终提供给训练模型的奖励是由得分奖励、距离奖励、导弹损失奖励进行组合后并进行归一化处理得到的。该奖励一方面会促使导弹不断靠近蓝方目标，另一方面又探索出不易被拦截的航线，最终获得理想的打击效果，其最终奖励 r 为：

$$r = r_s \cdot \lambda_s + r_d + r_l \quad (9)$$

其中, $\lambda_s=0.001$ 为得分奖励的权重系数。

2.3 奖励塑造强化学习方法

本方法的设计主体基于 DQN 方法, 并加入奖励塑造。设置墨子联合作战推演系统的决策间隔时间为 2 min, 最大决策步数设置为 20, 最大训练轮数为 5 000。针对每个决策步, 分别对状态、动作、奖励、下一状态进行获取并提供给智能体进行训练。其算法流程如下。

(1) 建立当前值网络 q_eval 与目标值网络 q_target , 初始化 DQN 方法的超参数。

(2) 判定开始训练的开始局数 $start_epoch$ 是否大于 0。如果是, 则加载之前相应 $start_epoch$ 的训练模型, 否则将重新开始训练。

(3) 进入单局推演循环, 进入环境 $step$ (这里的环境 $step$ 对应于墨子联合作战推演平台的每个决策间隔)。

(4) 智能体获取当前的状态信息 s_{step} , 在贪婪策略下输出执行动作 a_{step} , 即如果随机生成的数 $random$ 大于 $epsilon$ (依据贪婪策略大小逐渐变化), 则选择 q_eval 取最大值时对应的动作作为生成动作, 否则在动作空间中随机生成动作。获取下一状态信息 s_{step+1} , 以及环境的奖励信息 r_{step} (r_{step} 为奖励塑造后的奖励值, 由得分奖励、距离奖励、导弹损失奖励共同生成)。

(5) 将时间步 $step$ 采样所得的“状态, 行动, 奖励, 新状态” ($s_{step}, a_{step}, r_{step}, s_{step+1}$) 进行组合放入回放缓存中, 按照采样顺序插入, 超过的则覆盖存储。

(6) 判断推演局数 $episode$ 是否大于训练前数据采集局数 N , 一般 $episode$ 是 10 的倍数。如果是, 进入步骤 (7), 否则, 进入步骤 (3)。

(7) 从回放缓存中抽取大小为 $batch_size$ 的数据, 对 q_eval 网络进行训练与更新, 当训练次数达到目标网络替换迭代次数 $replace_target_iter$ 时, 对 q_target 网络进行更新, 并保存当前训练模型, 进入步骤 (3)。

(8) 判定训练循环是否达到最大训练局数 $MAX_EPISODES$, 如果否, 返回步骤 (3), 否则, 训练结束。

伪代码如下所示。

算法1 基于奖励塑造的智能导弹突防策略规划算法

输入: 当前值网络 q_eval 参数 θ , 目标值网络 q_target 参数 θ'

输出: 最优的行动策略网络参数 θ^*

初始化经验回放池内存 D 、训练前数据采集局数 N 、最大训练局数 $MAX_EPISODES$

随机初始化当前值网络 q_eval 参数 θ

初始化目标值网络 q_target 参数 $\theta' \leftarrow \theta$

For $episode = 1, 2, \dots, MAX_EPISODES$ do:

 环境初始化, 获得初始状态 s_0

 For $step = 1, 2, \dots$ do:

 依据 ϵ -greedy 策略选择动作:

$$a_{step} = \begin{cases} \text{随机动作} & , \text{random} < \epsilon \\ \arg \max_a Q(s_{step}, a) & , \text{otherwise} \end{cases}$$

 在环境中执行动作 a_{step} , 获得新的状态 s_{step+1} 与收益 r_{step} (由得分奖励、距离奖励、导弹损失奖励共同生成)

 存储经验 $(s_{step}, a_{step}, r_{step}, s_{step+1})$ 至重放内存 D

$episode > N$ and $episode \% 10 = 0$:

 If

 从内存 D 中采样数量为 $batch_size$ 的经验 $\{(s_j, a_j, r_j, s_{j+1}), \dots\}, j$ 为索引

 利用当前值网络获得预估值: $q(s_j, a_j)$

 利用损失函数 $E(q(s_j, a_j) - \hat{q}(s_j, a_j))^2$, 执行梯度下降更新行为网络参数 θ

 If $step \% replace_target_iter = 0$:

 更新目标网络参数: $\theta' \leftarrow \theta$

 End If

 End If

End For

End For

3 实验分析

3.1 实验想定设计

3.1.1 总体思路

受导弹本身命中概率与蓝方拦截武器的影响, 即使应用同一打击策略, 打击效果也可能有所波动, 有时难以达成预期的效果。本实验设置想定的思路是在一定数量的导弹和预定的打击目标的前提下, 验证能否通过智能导弹的动态航迹规划来提升突防概率, 进而提升打击效果。想定的设计仍参考智能导弹建模与“分布式杀伤”舰艇编队建模的具体内容, 以最大限度模拟未来海上高强度对抗为原则, 其想定设计思路如图3所示。

智能导弹经过初期发射, 按照预定航线向目标飞行。在高强度的对抗环境下, 由于蓝方远程拦截

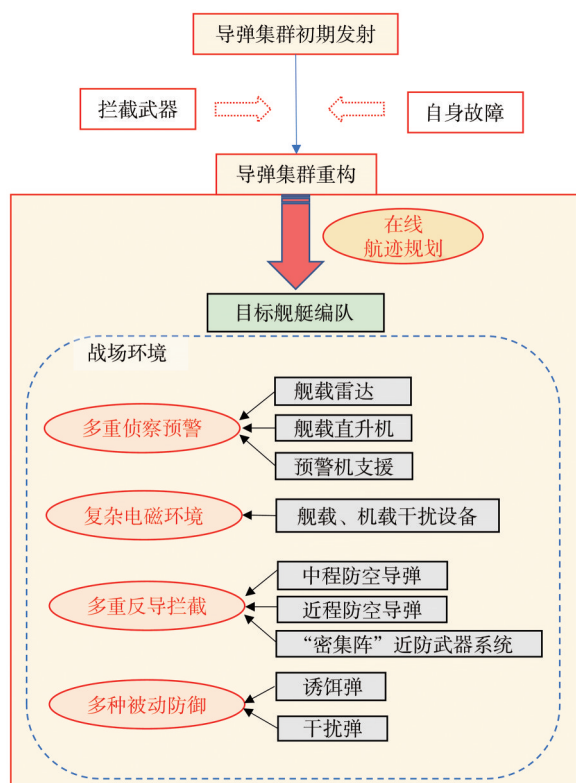


图3 突防策略规划实验想定设计

武器的拦截以及部分导弹出现故障，导弹数量有所损失。智能导弹经过群组重构后，在线改变了打击方案，对蓝方舰艇编队进行打击。战场环境具体体现为多重侦察预警、复杂电磁环境、多重反导拦截、多种被动防御，在此环境下，拟采用智能导弹动态航迹规划的方法来形成突防策略，提升打击效果。

3.1.2 具体设置与实现

想定持续总时间为40 min，以包含红方智能导弹与某舰艇编队的矩形海域为想定主要区域，想定区域大小约240 km×180 km。红方智能导弹由舰艇编队发射后，经过预定航线，在蓝方舰艇发射的海基中段拦截导弹的拦截下已有所损失，弹群剩下

3组共9枚导弹，距离舰艇编队的直线距离约为200 km。蓝方舰艇编队成一字纵队，呈防御态势，如图4所示。

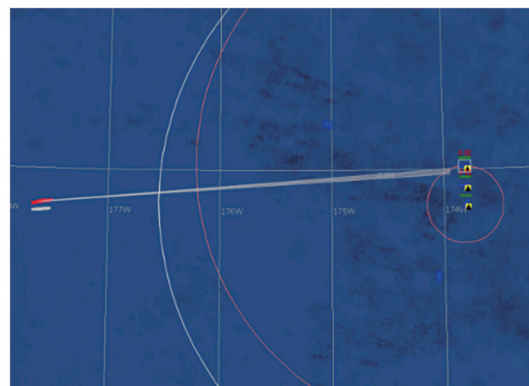


图4 突防策略规划实验想定

在整个想定中，由于双方实体相距较远，红方智能导弹掌握的蓝方舰艇编队的位置信息主要来源于其他信息平台（如卫星），具有间断性、滞后性。只有当到达预定航迹终点，近距离靠近蓝方编队时，智能导弹才能打开主动雷达，当搜索到蓝方舰艇的具体位置信息时，才能再次刷新位置信息。蓝方编队具有较强的对空探测能力与对空打击能力，可对红方弹群进行拦截，如图4所示，红色圆圈代表蓝方各型导弹的攻击范围，白色圆圈代表蓝方对空探测范围。此外，蓝方舰载干扰设备处于开机状态，战场区域能得到后方预警机的信息支援。红方、蓝方具体兵力情况分别见表1、表2。

3.2 状态空间与动作空间

本实验从己方所感知的全局态势中抽取2类数

表1 红方兵力情况

类别	装备型号	数量	备注
智能导弹	战斧-3000亚声速反舰/对陆导弹	9	分为3个编组，每组3枚
卫星	海军海洋监控卫星[白云型Blk III]	1	—

表2 蓝方兵力情况

类别	装备型号	数量	携带主要防御性武器及数量
水面舰艇编队	DDG-1000朱姆沃尔特级驱逐舰	1	RIM-162A型“海麻雀”舰对空导弹32枚、155 mm/52舰炮超高速炮弹460发、SM-3 block3海基中段拦截导弹6枚、Mk214海蚊箔条弹40枚、Mk234纳尔卡主动干扰弹24枚
	阿利·伯克Ⅲ驱逐舰	1	SM-6 block4远程防空导弹18枚、海拉姆-3000近程防空导弹40枚、RIM-162型海麻雀导弹40枚、密集阵近程防御武器系统、Mk214海蚊箔条弹60枚、Mk245巨人红外干扰弹36枚
	美军独立级濒海战斗舰	1	RIM-116C滚体导弹(拉姆导弹)11枚、Mk214海蚊箔条弹20枚、Mk234纳尔卡主动干扰弹12枚
信息支援力量	E-2D型“先进鹰眼”舰载预警机	1	—
卫星	海军海洋监控卫星[白云型Blk III]	2	—

据进行状态空间的描述，主要包含了红方和蓝方的实体位置信息，具体见表3，其中位置信息是指将想定中导弹与舰艇的实际经纬度进行归一化处理得到的位置信息，如果目标被摧毁则补零处理。

表3 状态空间信息

实体	信息
红方导弹	经度、纬度
蓝方舰艇	经度、纬度

本文决策的是3组导弹的飞行方向，通过对每组导弹设置下一航路点来实现。为了简化实验，将导弹飞行方向沿目标方向离散化为3个角度，而多次设置下一航路点就形成了各弹群分组的突防航线，其动作空间维度为 $3 \times 3 \times 3$ 共27维，动作空间信息具体见表4。

表4 动作空间信息

类别	取值
导弹飞行方向	60° 、 90° 、 120°

在本次想定中，为了确保导弹打击流程顺利进行，根据导弹武器的特点，加入了3个规则约束导弹分组。

作战区域的限定。如果某一导弹分组离预定打击目标过远，下一步动作按预定航路终点飞行，不再重新设置新的航路点，防止导弹飞离预定的作战区，从而丢失目标。

对预定目标的锁定。当导弹离预定目标的距离快接近雷达的探测距离时，主动打开雷达进入搜索攻击阶段，防止导弹与目标“擦肩而过”。

对打击目标的转换。当导弹预定目标被其他导弹分组摧毁时，则按照优先级重新分配下一个打击目标。重新分配打击目标时，以阿利伯克级驱逐舰为首要攻击目标，朱姆沃尔特级驱逐舰次之，最后是濒海战斗舰。

3.3 实验设置与效果

3.3.1 实验设置

本节实验同样在墨子联合作战推演平台上设置想定，通过作战推演产生数据以供深度神经网络学习。墨子平台设置的仿真决策间隔时间为2 min。用训练好的 q_target 网络作为智能导弹突防策略规划的模型。实验中的参数和超参数的设置如表5所示。

表5 模型训练参数

参数	符号	值
最大训练局数	MAX_EPISODES	5 000
最大步数	MAX_STEPS	20
持续时长	DURATION_INTERVAL	120
动作	n_actions	27
特征	n_features	21
贪婪增长因子	e_greedy_increment	0.0005
学习率	learning_rate	0.01
贪婪因子	e_greedy	0.9
奖励衰减	reward_decay	0.9
目标网络替换迭代次数	replace_target_iter	10
缓存大小	memory_size	6 000
批大小	batch_size	30

3.3.2 模型效果

对比实验使用DQN和DDPG两个基类算法，分别对只有得分奖励与包含塑造奖励的两种情况进行训练，其余条件和参数相同。每种情况均在累计对战20局后开始训练，每推演10局再次进行训练，并保存一个模型，共对战2 420局，保存240个模型。训练完成后用每个模型对战10局，平滑后得到的训练模型得分如图5所示。其中，无塑造奖励的DQN方法训练效果较差，得分曲线上升不明显，很难判定最终能否训练出较好的决策模型；有塑造奖励的DQN_RS方法得分曲线上升趋势明显，最后平均得分能稳定在1 000分左右，得到了有效的决策模型。

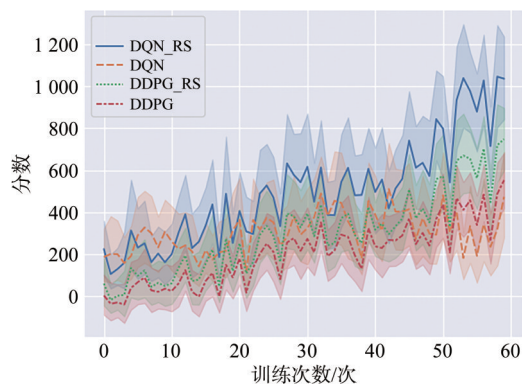


图5 训练模型得分

将加载训练模型动态控制突防航线的方法与采用原航线的方法进行实验对比。2种方法分别对战100局，采集相应数据，对比验证训练的有效性，如表6所示。

表6中的高得分率是指在一次推演中总得分大

表6 训练效果对比

决策方法	最高分	最低分	平均分	高得分率	打击成功率
训练模型控制	2 830	0	1 087.50	63%	79%
原航线	1 480	0	100.60	5%	13%

于或等于1 480分的概率，相当于彻底摧毁蓝方一艘主力舰艇（“朱姆沃尔特”级驱逐舰或“阿利伯克”级驱逐舰）的概率。打击成功率是指在一次推演中总得分大于或等于160分，相当于击伤蓝方一艘主力舰艇的概率。

实验结果表明，训练模型控制能够生成有效的突防策略。对比不采取动态航迹规划的原航线，由决策模型控制的3个智能导弹分组最终平均得分由100.60提高至1 087.50，高得分率由5%提高至63%，打击成功率由13%提高至79%。

综上所述，面对一个舰艇编队时，即使只有9枚导弹，即使整体战场态势具有一定的“战争迷雾”，在智能突防阶段采用该航迹规划的方法也能够生成特定的突防策略，提升打击效果。

3.4 奖励塑造突防策略分析

复盘实验数据，分析训练推演过程中由决策模型控制的智能导弹在打击成功时执行的具体行为，总结出以下几项航迹规划的突防策略：集中迂回攻击、分散突防多向攻击、分组延时攻击以及巡弋探

测指引攻击，可以为现实中导弹的战法运用提供一定的参考。

3.4.1 集中迂回攻击

在防空反导作战中，大多数防空导弹通过预测拦截目标的航迹实现对来袭导弹的拦截。导弹在对舰艇进行攻击时，如果一直按固定的航线飞行，被拦截的概率较大。故而，智能导弹在突防时，可采用大范围迂回策略，如图6所示。

导弹群组沿蓝方舰艇编队某一侧向连续变更航线，以绕飞的形式规避防空导弹的拦截，在一定程度上降低防空导弹的拦截概率，进而伺机进入蓝方编队防空的不利阵位，从蓝方编队的侧面进行打击。同时，绕飞的过程中，也会诱发蓝方进行多波次的反导拦截，能消耗大量蓝方的防空武器，而且弹群采取集中突防的行动模式还能形成几乎同时到达目标的打击态势，以冗余攻击的方式来确保对敌高价值目标的有效毁伤。

3.4.2 分散突防多向攻击

在训练调试与模型测试过程中发现，分散突防多向攻击也是智能导弹较好的突防策略，其突防效果如图7所示。

3个弹群分组分别向蓝方舰艇编队的3个主要方向分散飞行，同时在多个方向上造成一定的打击

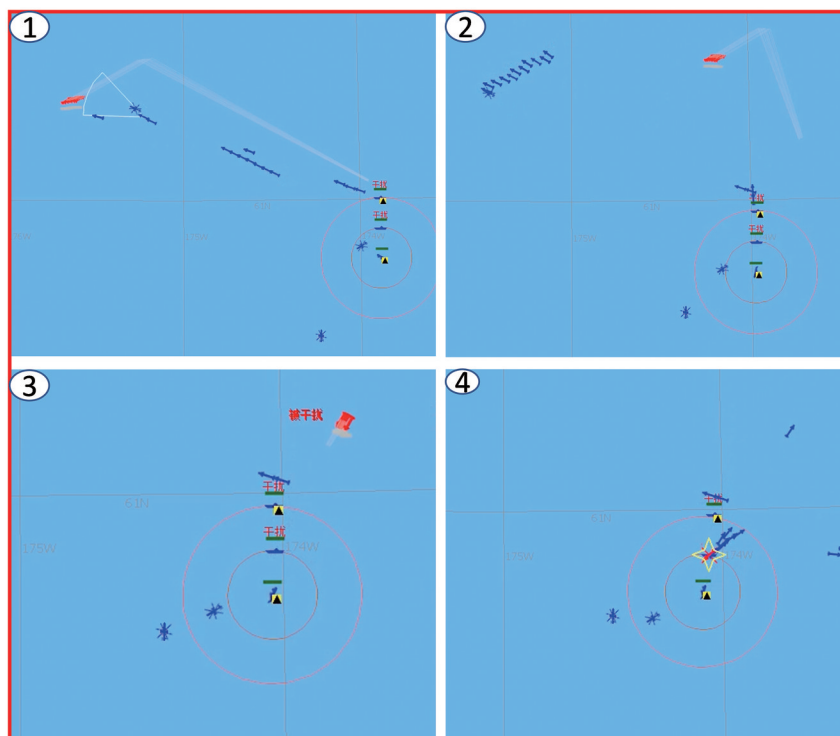


图6 集中迂回攻击

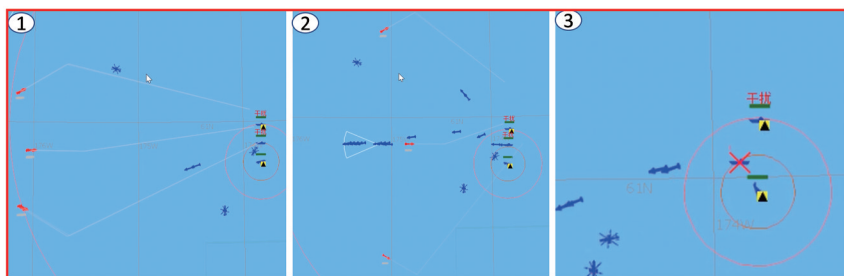


图7 分散突防多向攻击

态势。弹群从多个不同方向接近蓝方目标,使得蓝方舰艇编队不得不同时应对多个方向上的导弹威胁,这样能有效分散其防空火力与精力,进而在整体上增加弹群最终的命中概率与对舰艇的毁伤概率。此外,由于本实验对导弹的飞行方向进行了离散化处理,按照想定的设计几乎难以达成多组导弹多个方向且同时到达的攻击态势,但在现实导弹作战运用中,可以通过对航迹的进一步调整来弥补本实验的缺陷。

3.4.3 分组延时攻击

重点观察实验的高分对局发现,分组延时攻击有时会收到很好的打击效果。原计划9枚导弹击沉一艘敌主力舰艇的目标不仅能够达成,还能获得额外的战果。经分析,可以将该策略简单理解为小型的分波次打击,其具体打击过程如图8所示。

多个弹群的分组延时攻击是,当突防到距离舰艇编队较近的不同方位时,由其中一个分组对预定目标进行先期攻击,另外的分组在一定区域外巡弋。由于弹群之间存在信息交互,预先打击的导弹分组会将打击效果传递给其他分组。如果未击中目标,其他导弹分组则会对目标进行接续攻击。如果目标已被摧毁,剩下的分组则会搜寻其他高价值目标,将战果最大化。

回看导弹群组的集中突防,虽然该策略能造成同时到达的态势和比较稳定的毁伤效果,但如果被敌雷达精准锁定,也存在被集中拦截的风险。此外,由于集中突防几乎同时到达,往往会造成一定的冗余攻击。例如,用3枚导弹能击沉一艘舰艇,结果却花费了5枚导弹,这在一定程度上浪费了武器资源。相比之下,分组延时攻击的突防策略的打击效果往往较好,但延时攻击也会给蓝方防空力量更多的喘息机会,存在导弹被先后尽数拦截的风险。

3.4.4 巡弋探测指引攻击

需要说明的是,本节实验主体考虑巡航类导弹

的原因是其速度在1马赫以内,比较容易对其航线进行即时调整,从而规避敌防空威胁。而超声速导弹,尤其是高超声速导弹,其速度超过了5马赫,目前一般的防空系统无法对其进行拦截,故而较少考虑其突防问题,而更多考虑其能否命中目标。然而,在本实验想定中,相隔近1 000 km仅依靠卫星探测的手段获取的蓝方舰艇的具体位置与实际位置存在较大偏差,难以为高超声速导弹提供精确指示。若考虑用超声速导弹对舰艇编队进行打击,则需要更为精确的信息指引,而本实验中先期突防的巡航导弹与在战区巡弋的导弹为这种精确的信息指引提供了可能。

巡弋探测指引攻击的打击策略是本实验进行拓展探索时偶然发现的,是原定打击策略效果不佳时的补救方案,具体策略演示如图9所示。

先期打击的弹群突防效果不佳,多数导弹已被蓝方防空导弹拦截,但部分导弹在被拦截前获得了蓝方舰艇的精确信息,传递给了后方的弹群分组。巡弋探测指引攻击就是在蓝方防空火力较强或者己方弹群损失较大的情况下,判定单靠本波次的导弹打击难以完成任务时,使在战场的智能导弹进入巡弋探测阶段,为后方补充打击的超声速导弹群提供蓝方目标的精确指示,同时与超声速导弹群进行协同打击。但超声速导弹的加入,也大大增加了自身的打击成本。该策略也为我们带来了一定的启发,即在未来的智能导弹或其他智能集群的设计过程中,能否构建出特定构型的智能作战集群,使该集群有多种不同侧重功能的作战单元,比如探测指引、协同指挥、火力打击等,在智能化的基础上,通过集群的多元化提升综合战斗力。

综上所述,集中迂回攻击、分散突防多向攻击、分组延时攻击以及巡弋探测指引攻击的智能导弹突防策略,不仅可以为现实中导弹的战法运用提供参考,也能为未来智能集群的发展提供一定的思路。

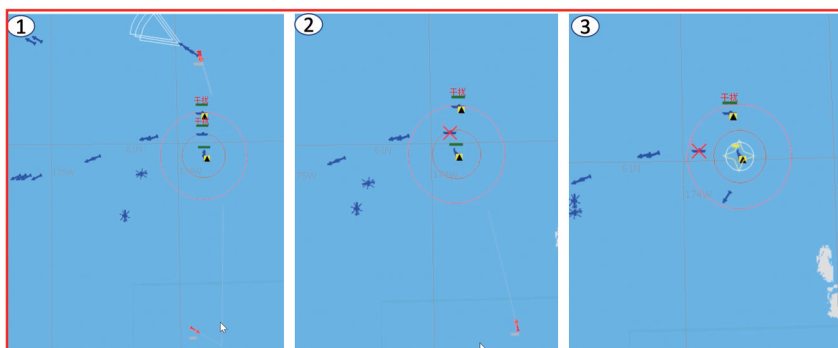


图8 分组延时攻击

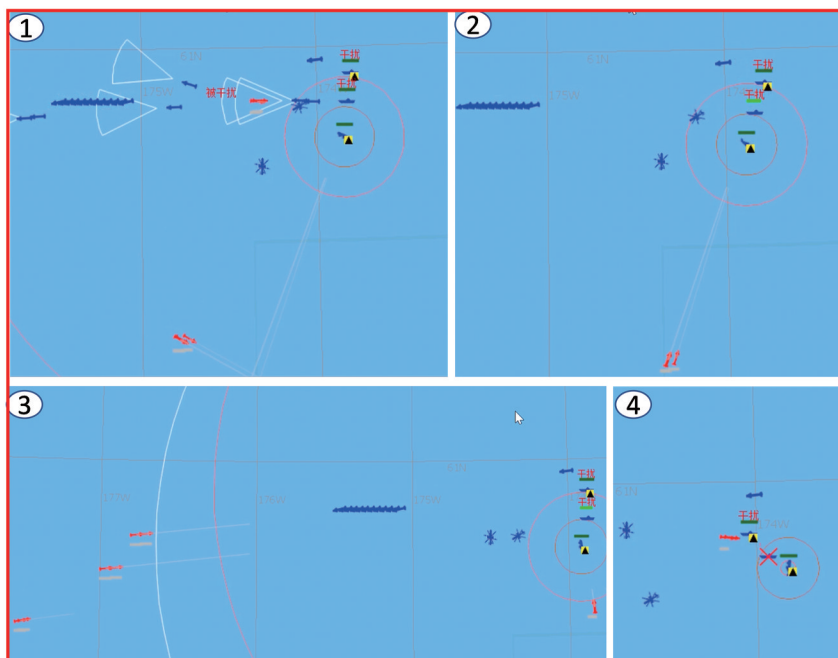


图9 巡逻探测指引攻击

4 结束语

本文重点分析智能导弹突防策略规划问题，设计了实验想定，构建了少量智能导弹分组，针对某一“分布式杀伤”舰艇编队进行突防的作战态势，通过墨子联合作战仿真平台实现了对想定兵力部署、武器配备等强对抗环境具体内容的设置。针对奖励稀疏问题，本实验进行了奖励塑造的具体设计。本实验基于DQN主体方法，加入奖励塑造内容，进行了智能导弹突防策略规划方法的设计与相关的对比实验。从实验效果来看，对于此智能导弹突防策略规划实验场景，基于奖励塑造的智能导弹突防策略规划方法具有一定的适用性。

本文研究的对象是智能导弹，主要聚焦研究基于分组的流式齐射。总结的智能导弹的4种突防策

略，不仅能为现实的导弹作战提供战法参考，也能为未来智能导弹或其他智能集群的发展提供新的思路。

参考文献：

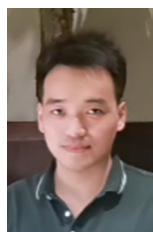
- [1] 于雪泳, 朱清浩. “海军打击导弹”(NSM)武器系统综述[J]. 舰船电子工程, 2013, 33(5): 27-29.
YU X Y, ZHU Q H. Summary of “navy strike missile” weapon system[J]. Ship Electronic Engineering, 2013, 33(5): 27-29.
- [2] 韩光松, 王忠, 宋志峰. 弹头集群协同制导在海战场的应用[J]. 战术导弹技术, 2021(3): 35-41.
HAN G S, WANG Z, SONG Z F. Application of warhead cluster cooperative guidance in sea battlefield[J]. Tactical Missile Technology, 2021(3): 35-41.
- [3] 许学春. 隐身导弹与隐身技术的应用[J]. 飞航导弹, 2013(5): 87-90.
XU X C. Stealth missile and the application of stealth technology[J]. Aerodynamic Missile Journal, 2013(5): 87-90.
- [4] 杨宁, 吴艳梅. 智能化的新型反舰导弹 LRASM 综述[J]. 军事文摘,

- 2020(23): 39-43.
- YANG N, WU Y M. Summary of intelligent new anti-ship missile LRASM[J]. Military Digest, 2020(23): 39-43.
- [5] 李文胜, 王然辉. 美国战斧 Block 4 巡航导弹的改进及未来发展[J]. 飞航导弹, 2016(4): 28-31.
- LI W S, WANG R H. Improvement and future development of American Tomahawk Block 4 cruise missile[J]. Aerodynamic Missile Journal, 2016(4): 28-31.
- [6] MNH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.
- [7] 刘朝阳, 穆朝絮, 孙长银. 深度强化学习算法与应用研究现状综述[J]. 智能科学与技术学报, 2020, 2(4): 314-326.
- LIU Z Y, MU C X, SUN C Y. An overview on algorithms and applications of deep reinforcement learning[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(4): 314-326.
- [8] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double Q-Learning[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. New York: ACM, 2016: 2094-2100.
- [9] WANG Z, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning [C]//Proceedings of the International Conference on Machine Learning. New York: ACM, 2016: 1995-2003.
- [10] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms [EB]. arXiv preprint, 2017, arXiv: 1707.06347.
- [11] BAI S X, SONG S M, LIANG S Y, et al. UAV maneuvering decision-making algorithm based on twin delayed deep deterministic policy gradient algorithm[J]. Journal of Artificial Intelligence and Technology, 2021, 2(1): 16-22.
- [12] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[EB]. arXiv preprint, 2015, arXiv: 1509.02971.
- [13] MNH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning[EB]. arXiv preprint, 2016, arXiv: 1602.01783.
- [14] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor[EB]. arXiv preprint, 2018, arXiv: 1801.01290.
- [15] 王国岩, 曹红松, 刘鹏飞, 等. 基于墨子系统的对舰打击策略设计与分析[J]. 测试技术学报, 2023, 37(4): 304-309.
- WANG G Y, CAO H S, LIU P F, et al. Design and analysis of warship strike strategy based on Mozi system[J]. Journal of Test and Measurement Technology, 2023, 37(4): 304-309.
- [16] 肖友刚, 金升成, 毛晓, 等. 基于深度强化学习的舰船导弹目标分配方法[J]. 控制理论与应用, 2024, 已录用.
- XIAO Y G, JIN S C, MAO X, et al. Missile-target assignment method of naval ship based on deep reinforcement learning[J]. Control Theory & Applications, 2024, accepted.
- [17] 张斌, 雷永林, 李群, 等. 基于强化学习的导弹突防决策建模研究[J]. 系统仿真学报, 2024, 已录用.
- ZHANG B, LEI Y L, LI Q, et al. Reinforcement learning modeling of missile penetration decision based on combat simulation[J]. Journal of System Simulation, 2024, accepted.
- [18] 黄卓, 徐振, 郭健, 等. 基于分区间强化学习的集群导弹快速任务分配[J]. 控制理论与应用, 2023, 40(6): 1129-1139.
- HUANG Z, XU Z, GUO J, et al. Fast task allocation for missile swarm based on sectioned reinforcement learning[J]. Control Theory & Applications, 2023, 40(6): 1129-1139.
- [19] 梁星星, 马扬, 冯晓赫, 等. 基于预测编码的样本自适应行动策略规划[J]. 软件学报, 2022, 33(4): 1477-1500.
- LIANG X X, MA Y, FENG Y H, et al. Sample adaptive policy planning

based on predictive coding[J]. Journal of Software, 2022, 33(4): 1477-1500.

- [20] 于忠琳, 张海峰, 段金升, 等. 基于饱和态势的反舰导弹齐射问题研究[J]. 兵器装备工程学报, 2021, 42(8): 46-50.
- YU Z L, ZHANG H F, DUAN J S, et al. Research on anti-ship missile salvo based on saturation state[J]. Journal of Ordnance Equipment Engineering, 2021, 42(8): 46-50.
- [21] GAO K Y, YAN X B. Study on the optimal strategy of missile interception[J]. IEEE Access, 2021, 9: 22239-22252.
- [22] DONG Y L, TANG X C, YUAN Y. Principled reward shaping for reinforcement learning via Lyapunov stability theory[J]. Neurocomputing, 2020, 393: 83-90.

[作者简介]



罗俊仁 (1989-), 男, 国防科技大学智能科学学院博士生, 主要研究方向为不完备信息博弈、多智能体学习等。



刘果 (1990-), 男, 国防科技大学智能科学学院硕士生, 主要研究方向为智能导弹攻防。



苏炯铭 (1984-), 男, 博士, 国防科技大学智能科学学院副研究员, 主要研究方向为智能博弈。



张万鹏 (1981-), 男, 博士, 国防科技大学智能科学学院研究员、博士生导师, 主要研究方向为大数据智能、智能演进等。



陈璟 (1972-), 男, 博士, 国防科技大学智能科学学院教授、博士生导师, 主要研究方向为认知决策博弈、智能任务规划等。