doi:10.11887/j.issn.1001-2486.24090024 http://journal.nudt.edu.cn

沙普利值分解的动态兵力部署策略规划方法

罗俊仁,张万鹏,苏炯铭,李胜强,陈 璟*
(国防科技大学 智能科学学院,湖南 长沙 410073)

摘 要:针对动态兵力部署问题,提出了一种基于沙普利值分解多智能体强化学习的策略规划方法。借助沙普利值分解来解释协作多智能体之间的奖励分配,利用基于沙普利分解强化学习方法求解马尔可夫凸博弈策略;针对海空跨域协同对抗场景,分析异构多实体协同对抗中空间域作战资源的分配,构建动态兵力部署策略规划模型,设计问题的状态空间、动作空间和奖励函数。围绕典型应用场景,利用兵棋推演系统对动态兵力部署问题组织了仿真实验验证,结果表明,与多类基线算法相比,所提方法在动态兵力部署策略规划方面性能优异,同时理论上具备可解释性,学到了"层层拦截、分区对抗,掩护核心、分层破击"长时域动态兵力部署策略。

关键词:沙普利值;多智能体;强化学习;兵力部署;策略规划
中图分类号:TP183 **文献标志码**:A **文章编号**:1001-2486(2025)04-123-09

论文拓展

Shapley value decomposition method in dynamic force deployment strategy planning

LUO Junren, *ZHANG Wanpeng*, *SU Jiongming*, *LI Shengqiang*, *CHEN Jing**
(College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China)

Abstract: Aiming at the dynamic force deployment problem, a multi-agent reinforcement learning strategy planning method based on SVD (Shapley value decomposition) was proposed. The reward distribution among cooperative multi-agents was explained by SVD, and the reward distribution was analysed by SVD reinforcement learning method to solve Markov convex game strategy. Secondly, based on the scenario of naval and air cross-domain cooperative confrontation, the allocation of space domain combat resources in heterogeneous multi-entity cooperative confrontation was analysed, a dynamic force deployment strategy planning model was built, and the state space, action space and reward function of the problem were designed. Finally, based on typical application scenarios, simulation experiments were organized to verify the dynamic force deployment problem with the military chess deduction system. Results show that compared with the multi-class baseline algorithm, the proposed method has excellent performance in strategic planning of dynamic force deployment, and it is theoretically interpretable. The proposed method learns the strategy of "layer upon layer interception, zone confrontation, core cover, and hierarchical breaking".

Keywords: Shapley value; multi-agent; reinforcement learning; force deployment; strategy planning

在复杂的强对抗环境中,实体感知信息不完整、实时响应要求高,对长时域、前瞻性动态兵力部署决策提出了挑战。如何通过可解释的有效奖励激励,实现策略的高效探索,是利用学习类方法驱动动态兵力部署策略规划的关键。根据分层抽象思想,兵棋推演中智能博弈对抗问题通常可约简为"排兵布阵"与"异步协同"两类子问题[1]。与多实体异步协同对抗不同,排兵布阵更关注空间域作战资源的分配,其中兵力资源的有效分配对于作战效能的有效发挥有着重要作用。面对强博弈对抗环境带来的非平稳性挑战,兵力部署[2]是研究排兵布阵的核心问题。与火力配置、武器目标分配、杀伤链设计等"资源分配"问题不同的是,兵力部署通常需要考虑全局、前瞻规划。近年

收稿日期:2024-09-14
基金项目:国家自然科学基金资助项目(61806212);湖南省研究生科研创新基金资助项目(CX20210011)
第一作者:罗俊仁(1989—),男,湖北大冶人,博士研究生,E-mail:luojunren17@nudt.edu.cn
***通信作者**:陈璟(1972—),男,江西南昌人,教授,博士,博士生导师,E-mail:chenjing001@sina.vip.com
引用格式:罗俊仁,张万鹏,苏炯铭,等. 沙普利值分解的动态兵力部署策略规划方法[J]. 国防科技大学学报,2025,47(4):123-131.
Citation: LUO J R, ZHANG W P, SU J M, et al. Shapley value decomposition method in dynamic force deployment strategy planning[J]. Journal of National University of Defense Technology, 2025, 47(4): 123-131.

来,一些研究尝试利用布洛托上校博弈[3]和网络阻断博弈[4]等模型来建模多方对抗条件下的资源分配问题。但这类模型面临多实体之间“长时、动态”交互过程建模难,“临机规划”动态响应要求高等挑战。姜龙亭等[5]围绕空战对抗中的攻击占位决策问题设计了近似动态规划方法;Mills 等[6]研究了空域敏捷作战部署问题;梁星星等[7]提出了基于预测编码的行动策略样本自适应规划方法;Li 等[8]围绕海空协同对抗,提出了基于 Q 值混合(Q mix, QMIX)多智能体强化学习的海空对抗决策方法。获得长时域动态兵力部署策略充满挑战。

伴随着人工智能技术的跨越式发展,智能博弈领域的相关问题求解也发展出新的解决方案和范式。利用人工智能技术辅助决策规划可提高规划的效率,应对动态环境的适应能力,确保规划的灵活性。当前,协作式多智能体深度强化学习(multi-agent deep reinforcement learning, MADRL)方法被广泛应用于求解多实体协同序贯决策类问题,学界展开了大量算法探索,主要包括基于反事实基线的反事实多智能体[9]和多智能体近端策略优化[10]等方法,价值分解网络[11]和 QMIX[12]等因子分解型值函数方法,Q 值变换[13](QTRAN)和 Q 值双决斗[14]等基于个体全局最大(individual global max, IGM)原则的值函数方法,Q 值注意力[15]和全局策略分解变换[16](universal policy decoupling transformer, UPDET)等基于注意力机制的值函数方法。其中关于值函数分解类方法通常关注各智能体的贡献,分配机制的可解释性是设计方法需要考量的重要维度。

一些研究利用多合作实体联盟博弈,尝试采用沙普利(Shapley)值来构建可解释性信度分配机制。其中,沙普利 Q 值深度确定性策略梯度(Shapley Q-value deep deterministic policy gradient, SQDDPG)[17]在智能体之间重新分配全局奖励,利用沙普利值来量化并公平分配每个实体对集成收益的边际贡献,沙普利 Q 学习(Shapley Q learning, SHAQ)[18]利用边际贡献来构建沙普利值分解强化学习方法,尝试解决全局奖励中的值分解问题,然而各类方法的应用场景不一致,性能表现不同,算法的可解释性也不一样。此外,一些研究聚焦可解释性利用稳定高效的反事实方式来计算沙普利值[19],从智能体重要性[20]角度来拓展沙普利值的相关应用,其中关于模型可解释方面,Heuillet 等[21]采用沙普利值来解释多智能体强化学习中的合作策略和个体贡献,通过蒙特卡罗采样近似沙普利值来降低计算成本。作为沙普利值在网络结构上的扩展,Angelotti 等[22]利用迈森值(Myerson value)来解释多智能体系统中单个个体或策略的贡献。

本文采用多智能体强化学习方法辅助动态兵力部署策略规划,首先构建兵力部署的马尔可夫凸博弈模型,其次设计满足多实体协同策略学习的多智能体强化学习方法及对应“状态空间、动作空间和奖励函数”,为兵力部署策略规划提供可解释性指导;最后根据典型场景组织仿真实验验证方法的有效性。

1 沙普利值分解强化学习方法

1.1 可转移效用博弈及公平分配

对于可转移效用(transferable utility, TU)合作博弈,即联盟博弈(N,v),其中 $N=\{1,\cdots,n\}$ 表示博弈局中人集合,$v:2^n\to\mathbb{R}$ 为实值特征函数,对于任意联盟 $C,D\subseteq N$,TU 博弈为凸博弈的充分条件为 $v(C\cup D)\geqslant v(C)+v(D)-v(C\cap D)$。记集合 N 的基数为 $|N|$,每个局中人的沙普利值为:

$$\phi_i(v)=\sum_{C\subseteq N\setminus\{i\}}\frac{|C|!(|N|-|C|-1)!}{|N|!}\cdot[v(C\cup\{i\})-v(C)]$$

其中,$C\subseteq N\setminus\{i\}$ 表示除 i 外的其他局中人。

对于一个公平的奖励分配方案,应满足:

1)对称性:对于任意联盟 $C\subseteq N\setminus\{i,j\}$,如果 $v(C\cup\{i\}\,|\,s,a)=v(C\cup\{j\}\,|\,s,a)$,则 $\phi_i(s,a)=\phi_j(s,a)$。

2)退化阶数(零度):对于任意联盟 $C\subseteq N$,如果 $v(C\cup\{i\}\,|\,s,a)=v(C\,|\,s,a)$,则 $\phi_i(s,a)=0$。

3)线性可加性:对于博弈局中人,两个特征函数满足 $\phi_i(s,a\,|\,v_1+v_2)=\phi_i(s,a\,|\,v_1)+\phi_i(s,a\,|\,v_2)$,其中 $\phi_i(s,a\,|\,v_1+v_2)$ 是可转移效用合作博弈(N,v_1+v_2)的沙普利值且满足$(v_1+v_2)(C\,|\,s,a)=v_1(C\,|\,s,a)+v_2(C\,|\,s,a)$。

博弈局中人数量越多,沙普利值的计算越复杂,一些研究尝试采用神经网络近似的方式来降低计算复杂度,但这些方式从不同角度忽略了沙普利值的某些属性。

1.2 马尔可夫凸博弈

对于由多个协作实体组建的合作联盟,其协同对抗场景可建模成为部分可观马尔可夫凸博弈,可用八元组$\langle N,S,A,T,\Lambda,\pi,R_i,\gamma\rangle$来表示。其中,$N$ 表示博弈局中人集合,S 表示所有状态,

$A=\times_{i\in N}A_i$表示所有实体的联合行动($\times$表示策略叉乘),$T(s,a,s')=Pr(s'|s,a)$表示状态转移,Λ 是所有联盟结构的集合,π 是行动策略,$R_t(s,a)$表示第 t 个时间步奖励,$\gamma\in(0,1)$为折扣因子。

定义:由多个实体组成的合作联盟结构 $CS=\{C_1,\cdots,C_n\}$,其中 $C_i\subseteq N$ 表示由多个具体实体构成的合作联盟。对于任意联盟,其合作联盟行动集为 $A_C=\times_{i\in C}A_i$,合作联盟行动策略 $\pi_C(\boldsymbol{a}_C|\boldsymbol{s})=\times_{i\in C}\pi_i(a_i|\boldsymbol{s})$。可定义无限长时域折扣奖励为:

$$V_{\pi_C}(\boldsymbol{s})=E_{\pi_C}\left[\sum_{t=1}^{\infty}\gamma^{t-1}R_t(\boldsymbol{s},\boldsymbol{a}_C)\,|\,\boldsymbol{S}_t=\boldsymbol{s}\right]\quad(1)$$

马尔可夫凸博弈中,任意两个不相交的联盟的合作将带来更多奖励,即对于任意 $C_i,C_j\subseteq N$,$C_k=C_i\cup C_j$,有:

$$\max_{\pi_{C_k}}V_{\pi_{C_k}}(\boldsymbol{s})\geqslant\max_{\pi_{C_i}}V_{\pi_{C_i}}(\boldsymbol{s})+\max_{\pi_{C_j}}V_{\pi_{C_j}}(\boldsymbol{s})\quad(2)$$

1.3　沙普利值分解及其近似

1.3.1　可解释性

当前围绕多智能体强化学习的相关研究很少关注可解释性,本文主要聚焦如何有效度量信度分配(credit assignment),即将全局奖励有效分给各智能体。

在合作博弈论领域,沙普利值正好可解决合作大联盟中收益分配问题,度量了联盟中每个实体成员对集体收益的贡献。

具体来看,在马尔可夫凸博弈中,$\Phi_i(C_i)$为集合 C_i 的贡献值,每个智能体的边际贡献可以定义为:

$$S_i=\sum_{C_i\subseteq N\setminus\{i\}}\frac{|C_i|!(|N|-|C_i|-1)!}{|N|!}\cdot\Phi_i(C_i)\quad(3)$$

1.3.2　沙普利值分解

在多智能体强化学习领域,类比值函数分解方法,可以定义状态值函数为:

$$V_i^{\Phi}(\boldsymbol{s})=\sum_{C_i\subseteq N\setminus\{i\}}\frac{|C_i|!(|N|-|C_i|-1)!}{|N|!}\cdot\Phi_i(\boldsymbol{s}|C_i)\quad(4)$$

如果智能体采用的是确定性策略,则可定义动作值函数 $Q_i^{\Phi}(\boldsymbol{s},a_i)$。如果定义最优动作值函数为 $Q_i^{\Phi^*}(\boldsymbol{s},a_i)=w_i(\boldsymbol{s},a_i)Q^{\pi^*}(\boldsymbol{s},\boldsymbol{a})-b_i(\boldsymbol{s})$且 $\sum_{i\in N}w_i(\boldsymbol{s},a_i)^{-1}b_i(\boldsymbol{s})=0$,则基于沙普利值的贝尔曼方程为:

$$\boldsymbol{Q}^{\Phi^*}(\boldsymbol{s},\boldsymbol{a})=\boldsymbol{w}(\boldsymbol{s},\boldsymbol{a})\sum_{s'\in S}Pr(s'|\boldsymbol{s},\boldsymbol{a})\cdot\left[R+\gamma\sum_{i\in N}\max_{a_i}Q_i^{\Phi^*}(\boldsymbol{s}',a_i)\right]-\boldsymbol{b}(\boldsymbol{s})\quad(5)$$

其权重向量为 $\boldsymbol{w}(\boldsymbol{s},\boldsymbol{a})=[w_i(\boldsymbol{s},a_i)]^{\mathrm{T}}\in\mathbb{R}_{+}^{|N|}$,基线函数为 $\boldsymbol{b}(\boldsymbol{s})=[b_i(\boldsymbol{s})]^{\mathrm{T}}\in\mathbb{R}_{\geqslant 0}^{|N|}$,动作值函数满足 $\boldsymbol{Q}^{\Phi^*}(\boldsymbol{s},\boldsymbol{a})=[Q_i^{\Phi^*}(\boldsymbol{s},a_i)]^{\mathrm{T}}\in\mathbb{R}_{\geqslant 0}^{|N|}$。为了便于求解,通常采用随机近似的方式来求解动作值函数:

$$\Delta(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')=R+\gamma\sum_{i\in N}\max_{a_i}Q_i^{\Phi}(\boldsymbol{s}',a_i)-\sum_{i\in N}\delta_i(\boldsymbol{s},a_i)Q_i^{\Phi}(\boldsymbol{s},a_i)\quad(6)$$

其中,$\delta_i(\boldsymbol{s},a_i)=|N|^{-1}w_i(\boldsymbol{s},a_i)^{-1}$,根据不同情况取值可记作:

$$\delta_i(\boldsymbol{s},a_i)=\begin{cases}1 & a_i=\arg\max_{a_i}Q_i^{\Phi}(\boldsymbol{s},a_i)\\ \alpha_i(\boldsymbol{s},a_i) & a_i\neq\arg\max_{a_i}Q_i^{\Phi}(\boldsymbol{s},a_i)\end{cases}\quad(7)$$

1.3.3　蒙特卡罗近似

多智能体强化学习方法中引入沙普利值可以提高算法的可解释性,有效缓解了信度分配问题,帮助理解每个智能体的贡献。但仍面临计算复杂性和如何模拟联盟不考虑某个智能体的挑战。对于$|N|=n$ 个智能体可能存在2^{n-1}个联盟状态。此外,可以三种机制来模拟计算联盟的沙普利值:假设智能体不执行动作,假设智能体执行随机动作,假设智能体执行联盟中随机选择的动作。但研究表明采用第一种机制可以获得更准确的近似。本文采用蒙特卡罗采样的方式来近似沙普利值。故沙普利值多智能体强化学习方法的损失函数如下:

$$L_\theta=\left\|R+\gamma\sum_{i\in N}\max_{a_i}\hat{Q}_i(\tau_i',a_i;\theta)-\sum_{i\in N}\hat{\delta}_i(\boldsymbol{s},a_i)\hat{Q}_i^{\text{target}}\right\|^2\quad(8)$$

其中,$\hat{\delta}_i(\boldsymbol{s},a_i)=\begin{cases}a_i=\arg\max_{a_i}Q_i^{\Phi}(\boldsymbol{s},a_i)\\ a_i\neq\arg\max_{a_i}Q_i^{\Phi}(\boldsymbol{s},a_i)\end{cases}$。

$$\hat{\alpha}_i(\boldsymbol{s},a_i)=\frac{1}{M}\sum_{k=1}^{M}f(\hat{Q}_{C_i^k}(\tau_{C_i^k},a_{C_i^k}),\hat{Q}_i(\tau_i,a_i))+1$$

表示采样 M 次。

基于沙普利值分解强化学习方法的训练主要包括四个部分:利用贪婪策略收集样本数据,并存储至回放缓存中;从回放缓存中小批量采集样本;计算动作值函数;依最小化误差方式优化神经网络参数。具体流程伪代码见算法 1。

2　动态兵力部署策略规划问题建模

根据任务式指挥框架(集中式指挥、分布式控制、分散式执行),动态兵力部署主要聚焦在分布式控制。海空协同对抗中动态兵力部署主要是指如何将多个合作实体合理地配置在多个可能交战区域,各任务实体采用一线自主协同组织对抗。

算法 1 沙普利值分解强化学习

Alg. 1 Shapley value decomposition reinforcement learning

输入:实体集合 N,动作值函数 $\hat{Q}_i(\tau_i', a_i; \theta)$

输出:目标网络 $\hat{Q}_i^{\text{target}}$

1. **for each** 历史片段(*episode*) **do**
2. 初始化行动探索
3. **for each** 时间步(t) **do**
4. 随机采样联盟,每个实体选择行动,执行行动并观测奖励和新状态,可得($\boldsymbol{s}, \boldsymbol{a}, C, \boldsymbol{r}, \boldsymbol{s}'$)
5. 将历史状态存储进回放缓存 D
6. **end for each**
7. **for each** 实体(agent) **do**
8. 从回放缓存中小批次采样
9. 更新动作值函数网络
10. 计算沙普利值
11. **end for each**
12. 根据式(8)构建损失函数
13. 根据最小化损失函数更新神经网络参数
14. **end for each**

2.1 动态兵力部署场景描述

海空跨域协同对抗场景中,蓝方不顾多方反对,决定武力接管某海域,红方进行区域拒止与反介入作战。本小节采用墨子系统设计对抗想定:双方兵力相近,战斗力基本平衡。蓝方在舰载机和舰船上有优势,兵力采用分布式部署。双方兵力编成设计见表 1 和表 2,合计总分均为 595 分。红蓝双方的总分值相同,保证了分值上的公平。同时,可以看出蓝方和红方在兵力编成上的侧重点不同。红方在保有一支规模可观的航母编队的前提下,还有反舰弹道导弹用于承担部分对海对舰打击任务。蓝方作为一支可以在海外独立遂行作战任务的完整舰队,其本身具有较为完备对海对空打击能力。

表 1 红方兵力编成

Tab. 1 Red force organization

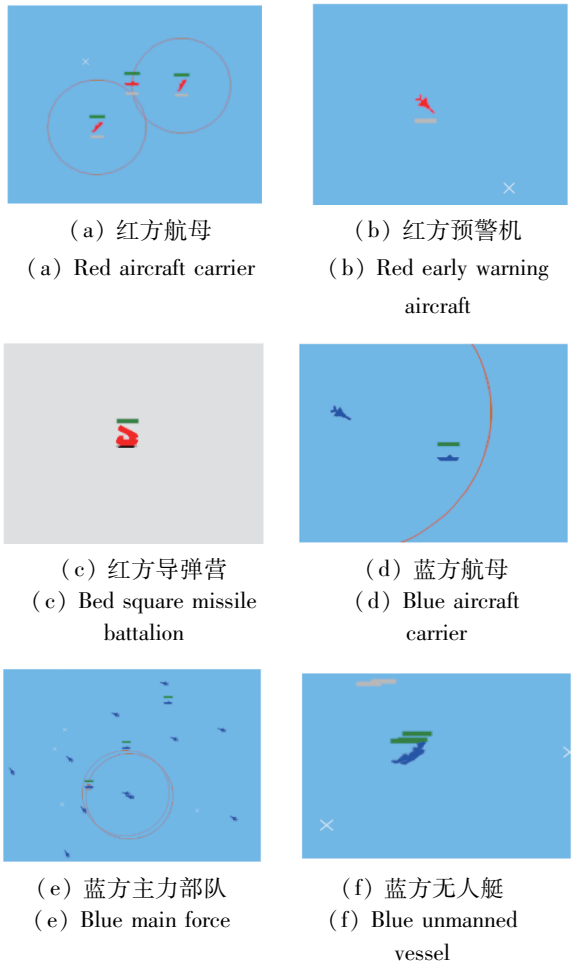
类别	装备型号	分值	数量	总分
地面	红方反舰导弹	75	1	75
水面舰艇	红方航空母舰	100	1	100
	红方驱逐舰 B	200	1	200
	红方驱逐舰 A	25	1	25
航空兵	红方战斗机	10	8	80
	红方攻击机	10	4	40
	红方预警机	75	1	75

表 2 蓝方兵力编成

Tab. 2 Blue forces organization

类别	装备型号	分值	数量	总分
水面舰艇	蓝方驱逐舰 A	100	1	100
	蓝方驱逐舰 B	100	1	100
	蓝方战斗舰	75	1	75
	蓝方无人艇	25	3	75
航空兵	蓝方战斗机	20	6	120
	蓝方电子战飞机	15	3	45
	蓝方攻击机	15	2	30
	蓝方预警机	45	1	45

对抗场景中,红方派出一支航母编队占领了 F 国东部海域作战阵位,同时起飞的红预警机也在编队的后方提供电子支援。位于大陆腹地的一个导弹发射营也占领发射阵位随时准备发射。蓝方的无人水面艇编队前出到了红方航母编队的附近,意图使用导弹对红方的航母编队进行打击。相关主要兵力部署如图 1 所示。



(a) 红方航母 (a) Red aircraft carrier

(b) 红方预警机 (b) Red early warning aircraft

(c) 红方导弹营 (c) Bed square missile battalion

(d) 蓝方航母 (d) Blue aircraft carrier

(e) 蓝方主力部队 (e) Blue main force

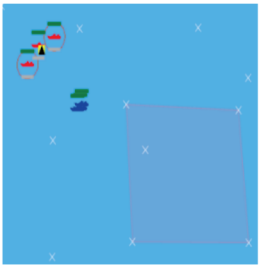
(f) 蓝方无人艇 (f) Blue unmanned vessel

图 1 海空对抗双方兵力部署

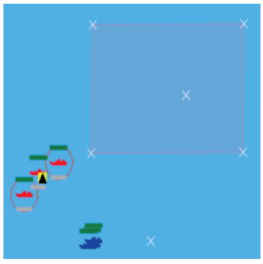
Fig. 1 Naval and air forces are deployed against both sides

2.2　动态兵力部署策略规划

在以上海空跨域协同对抗场景中,围绕红方行动一共可设置 7 个空间区域,包括空战 1、空战 2、空战 3、预警、导弹、反舰等。其中,3 个空战行动主要是战机进行空战巡逻,其区别在于巡逻区域的不同。预警是预警机进行电子支援的航线。导弹和反舰两个行动都以打击蓝方舰艇为目的,区别在于反舰行动由攻击机执行,导弹行动由反舰弹道导弹执行。编队会向航渡任务划定的区域机动,沿途执行巡逻警戒任务。相关关键区域如图 2 所示。

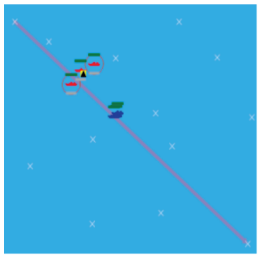


(a) 空中交战区域 1
(a) Air engagement zone 1

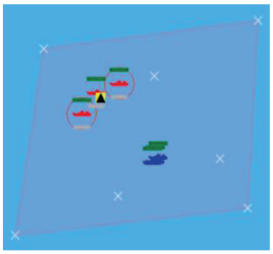


(b) 空中交战区域 2
(b) Air engagement zone 2

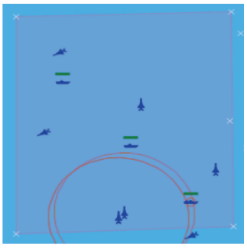
(c) 空中交战区域 3
(c) Air engagement zone 3



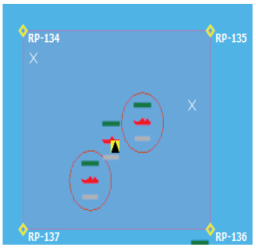
(d) 空中预警路线
(d) Air early warning route



(e) 对抗反舰导弹
(e) Counter anti-ship missiles



(f) 导弹打击区
(f) Missile strike zone



(g) 航渡区
(g) Ferry area

图 2　海空对抗中红方主要行动
Fig. 2　Main action of the red side in the naval and air confrontation

2.3　动态兵力部署特征工程

围绕海空跨域协同对抗典型场景,根据多智能体强化学习方法模块化需求,展开状态空间、动作空间和奖励函数的设计。

2.3.1　状态空间设计

状态空间主要涉及战场上的态势及双方智能体的状态。为了更加清楚地描述战场态势,将战场分成了三个部分,分别为红方编队区、空中交战区、蓝方编队区。各个区域内统计红蓝双方作战单元的数量及其状态(包括经度坐标、纬度坐标、单位类型)。一共由 $3 \times n \times 3$ 维状态(n 为区域内单位数量的总和)。

2.3.2　动作空间设计

结合多智能体模型,对应上文定义 7 个行动,设计多智能体强化学习的 7 个离散动作,并给出限制执行的条件(掩码),其中空战 1、空战 2、空战 3 限制条件均为战斗机、攻击机执行,反舰动作由攻击机执行、航渡由航母编队执行、导弹打击由导弹营执行、预警行动由预警机执行。

2.3.3　奖励函数设计

智能体奖励函数的设计反映了指挥员在制定作战行动方案时的决策偏好,如保存自己优先,消灭敌人其次,此时的奖励函数设计将对己方的损失更为敏感;如不惜一切代价消灭敌人,此时奖励函数设计将认为敌方的损失更为重要。奖励函数设计通常考虑以下因素:己方智能体的存活数量;己方智能体的损伤程度;敌方智能体的被消灭数量;敌方智能体的损伤程度;战场态势。在本想定中,异构多实体协同配合得分;攻击机应该尽量靠近敌方的舰队进行攻击,驱逐舰和弹道导弹也应积极对敌方发动进攻;尽量保存己方舰队。表 3 给出了策略奖励设计。

表 3　兵力部署策略规划奖励设计
Tab. 3　Reward design for force deployment strategy planning

策略	奖励值
获得更高的分数	$\Delta score$(595)
攻击机靠近目标	$\Delta d <$ 攻击机,舰船 > :10 000
拦截敌方攻击机	拦截一架: +0.01
发射导弹	发射一枚: +0.75
驱逐舰保护航母	航母每被击中一次: -0.01
驱逐舰发射反舰导弹	发射一枚反舰导弹: +0.001
舰队战力保存	存活一艘战舰: +0.04 一艘战舰毁伤 1%: -0.001

根据得分得到的奖励值如下：

$$R_{分} = \Delta score/595 \tag{9}$$

根据攻击机与目标的欧几里得距离 Δd 得到的奖励值如下：

$$R_{机} = \Delta d < 攻击机,舰船 > /10\ 000 \tag{10}$$

根据对进入我方舰队区域的飞机的拦截数量 $n_{拦}$ 得到奖励值如下：

$$R_{拦} = n_{拦} \times 0.01 \tag{11}$$

根据发射导弹的数量 $n_{导}$ 得到的奖励值如下：

$$R_{导} = n_{导} \times 0.75 \tag{12}$$

根据航母被命中的次数 $n_{中}$ 得到的奖励值如下：

$$R_{保} = -n_{中} \times 0.01 \tag{13}$$

根据驱逐舰反舰导弹的发射数量 $n_{反舰}$ 得到的奖励值如下：

$$R_{反舰} = n_{反舰} \times 0.001 \tag{14}$$

根据舰艇存活数量 $n_{存活}$ 和毁伤值变化 Δdam 舰队状况得到的奖励值如下：

$$R_{舰} = n_{存活} \times 0.04 - \Delta dam \times 0.001 \tag{15}$$

总奖励函数设计如下：

$$R = R_{分} + R_{机} + R_{拦} + R_{导} + R_{保} + R_{反舰} + R_{舰} \tag{16}$$

3 仿真设计与实验分析

3.1 学习框架与参数配置

本文利用墨子联合作战兵棋推演系统输出全局状态、全局奖励和全局观测，各实体根据观测输出行动，模拟仿真环境输入每个实体的行动并更新当前全局状态，从而完成一次迭代循环。在每次采样交互中，状态数据存储进经验回放池中，沙普利值强化学习方法利用经验回放池中的数据进行学习，更新当前各实体行动策略。

为了验证本文提出方法的有效性和实用性。本文重点围绕海空对抗兵力部署，设计适用于墨子推演系统的代码框架。如图 3 所示。框架中包括：智能体控制器 mac，经验回放池 buffer，智能体学习器 Learner，环境运行器 Runner 和墨子环境类 MultiAgentEnv。墨子联合作战推演系统则负责按时间推进想定中行动的实施，并将当前态势推送至客户端中进行实时可视化输出。智能体控制器 mac 的主要功能是控制智能体的动作，包括前向 forward( ) 和行动选择 select_actions( ) 两个关键方法。forward( ) 输出单个片段内每个时刻的观测对

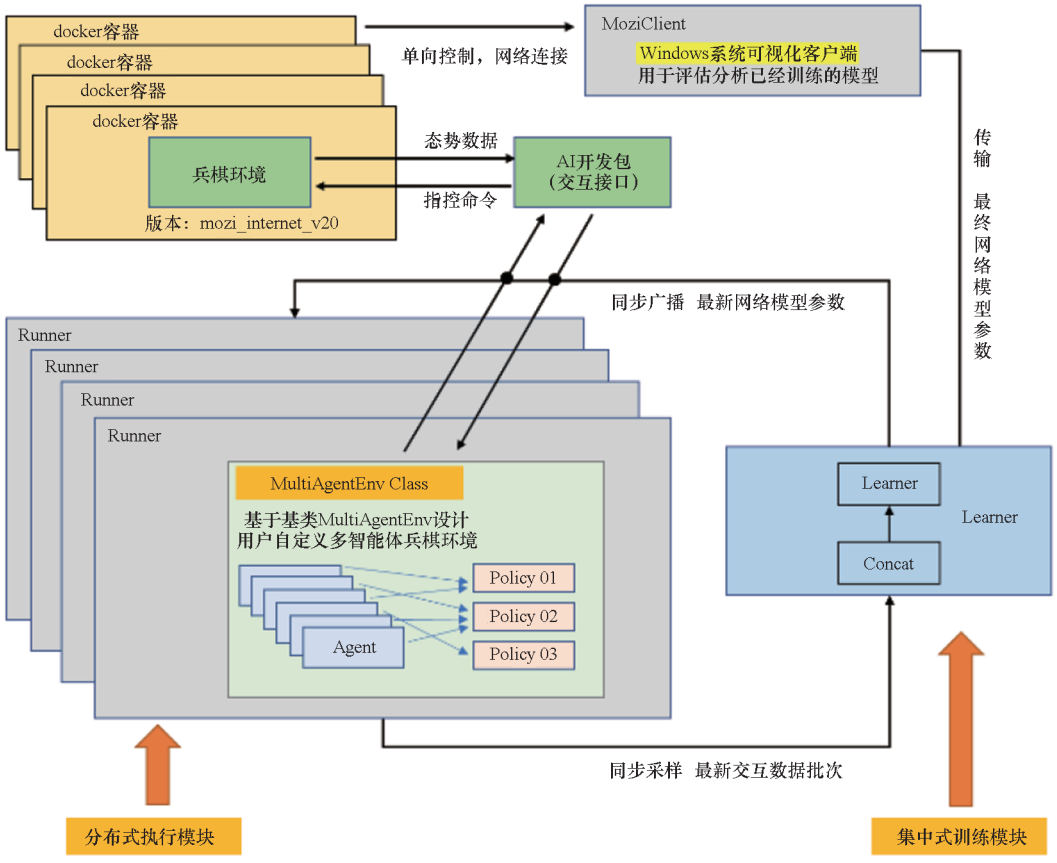


图 3 基于兵棋系统的多智能体强化学习训练框架

Fig. 3 Multi-agent reinforcement learning training framework with wargame system

应的所有动作的 Q 值与隐层变量。select_actions( ) 则是在每个时刻为智能体选择动作。经验回放池 buffer 的主要功能是存储样本和采样样本,insert_episode_batch( ) 方法用于存储样本,sample( ) 方法用于采样。智能体学习器 learner 对象利用沙普利多智能体强化学习方法对智能体的参数进行训练。Learner 对象的训练方法 train( ) 用于训练智能体。环境运行器 Runner 主要是运行环境并产生训练样本供经验回放池 buffer 收集和采样。Runner. run( ) 方法主要负责调用运行环境并产生样本。

实验相关超参数中折扣因子为 0. 99,训练集大小为 32,经验回放池大小为 1 000,学习率为 0. 000 5。

3.2　实验结果与分析

本节包含两组实验,旨在回答以下问题:

问题 1:沙普利值分解强化学习方法的有效性如何? 本文设计了三套兵力部署策略用于对比分析。其中策略 1 中将战斗机全部部署到空战 1 任务当中,且只有 3 架攻击机参加反舰任务。策略 2 将战斗机全部部署到空战 2 中,4 架攻击机参加反舰任务。策略 3 则是全部战斗机和 2 架攻击机参加空战 3 任务,2 架攻击机参加反舰任务。另外,还设置了一组红方不制定作战方案的对照组。

问题 2:较其他基线方法性能水平如何? 本文将与 QMIX、QTRAN 等值函数方法进行性能对比。

为了分析本文方法的有效性,本文一共设置了四组对照。每组对照都运行了 115 次。从实验结果可以看出,使用沙普利值方法的结果相比较于其他实验组来说更加均衡,其绝大部分位于 0 分以上。也就是说,本文方法在大部分想定中都能获得胜利。而无策略运行的情况下绝大部分在 0 分以下,虽然有在 0 分以上的情况,但是其分数也没有特别出众。其他三个策略运行的结果相较用于无策略运行的情况要好,但是相较于本文方法的结果来说也比较差。相关结果如表 4 所示。

为了验证本文所提算法的性能水平,分析多类多智能体强化学习方法学习到策略的能力水平,设置了五组对比测试实验,每组运行了 96 次,统计各组实验结果,各类算法的性能如图 4 所示。整体上看本文所提沙普利值分解强化学习方法的对抗收益比较集中,在多类算法中整体性能最优;

表 4　海空对抗兵力部署统计结果

Tab. 4　Naval and air counterforce deployment statistics

实验组	胜率/%	平均值	标准差
Shapley	75	36. 20	72
QTRAN	74	35. 80	91
QMIX	73	33. 93	105
策略 3	58	-15. 84	114
策略 1	56	-1. 89	104
策略 2	50	-11. 68	113
无策略	41	-36. 29	105

QMIX 和 QTRAN 方法的最终平均奖励比较高,但分布区间比较大,即方差比较大。使用沙普利值多智能体强化学习方法辅助动态兵力部署可有效地提高海空对抗的胜率、得分、稳定性等各项指标;该方法较 QMIX 和 QTRAN 方法具有较好的胜率和较小的方差。

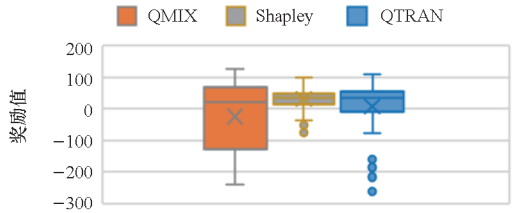


图 4　海空对抗兵力部署方法性能分析

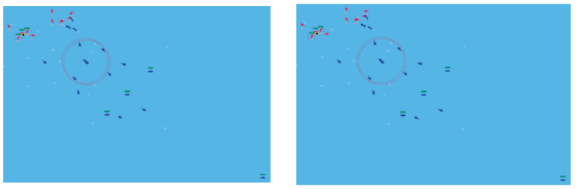
Fig. 4　Performance analysis of naval and air counterforce deployment methods

3.3　长时域策略分析

为了深刻理解红蓝双方的博弈对抗过程,可以采用长时域复盘分析各方所采用的行动策略。本小节聚焦分析双方对抗过程中各方作战意图、行动策略及对应战术战法,提炼典型阶段性特征,以期利用所得知识反向辅助、引导博弈对抗奖励设置、塑造学习算法。海空跨域对抗过程如图 5 所示。蓝方作战行动聚焦发挥制空优势和无人艇的作战优势,首先利用无人艇对红方编队进行首轮打击,之后红方战斗机进入进攻阵位清理红方负责防空巡逻的战斗机,为蓝方攻击机打击红方编队开辟通路。最后利用蓝方攻击机发射反舰导弹实施第二轮打击。

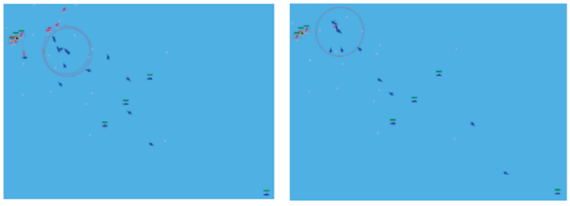
红方作战行动呈现“层层拦截、分区对抗”形态,首发打击蓝方无人艇,集中优势摧毁敌方作战力量,把握关键时刻定战局,凸显出火力控制方面的有效释放能力;其次围绕三个空中交战区争取有效占位,多样化的编队阵形有效增加了空间适

应性,增加了敌方的预测难度,提升了对抗的复杂性和不确定性;分区协同拦截蓝方飞机和来袭导弹,保持灵活机动,提高了适应战场的生存能力;此类“掩护核心、分层破击”行动策略有效响应了区域拒止与反介入作战需要。



(a) 层层拦截 (a) Layer-by-layer interception

(b) 分区对抗 (b) Partition confrontation



(c) 掩护核心 (c) Cover the core

(d) 分层破击 (d) Hierarchical breakthrough

图5 海空跨域协同对抗过程

Fig. 5 Naval and air cross-domain cooperative confrontation process diagram

4 结论

沙普利值提供了一种解释复杂的多主体交互作用的方法,使每个主体的决策和行为如何对系统的整体动态和结果做出贡献变得更加清晰。本文提出了基于沙普利值分解的强化学习方法,可用于海空对抗兵力部署。本文首先给出沙普利值多智能体强化学习方法的原理,然后围绕兵力部署策略规划问题,具体设计状态空间、动作空间和奖励函数,最后设计策略求解流程。

对于多实体协同对抗中“排兵布阵”与“异步协同”两类子问题,既可以采用跨层联动分析,也可以采用双层耦合分析。未来将考虑构建分层决策框架,利用分层强化学习方法学习整体行动策略。

参考文献(References)

[1] 尹奇跃, 赵美静, 倪晚成, 等. 兵棋推演的智能决策技术与挑战[J]. 自动化学报, 2023, 49(5): 913 – 928.
YIN Q Y, ZHAO M J, NI W C, et al. Intelligent decision making technology and challenge of wargame [J]. Acta Automatica Sinica, 2023, 49(5): 913 – 928. (in Chinese)

[2] 郑斌, 赵瑾, 张庆捷. 兵力部署相关概念综述[J]. 火力与指挥控制, 2017, 42(1): 5 – 8.
ZHENG B, ZHAO J, ZHANG Q J. Review of distribution of troops related concepts [J]. Fire Control & Command Control, 2017, 42(1): 5 – 8. (in Chinese)

[3] 罗俊仁, 邹明我, 陈少飞, 等. 布洛托上校博弈模型及求解方法研究进展[J]. 计算机科学, 2024, 51(1): 84 – 98.
LUO J R, ZOU M W, CHEN S F, et al. Research progress on colonel Blotto game models and solving methods [J]. Computer Science, 2024, 51(1): 84 – 98. (in Chinese)

[4] 项寅. 网络阻断问题研究热点及发展方向[J]. 运筹与管理, 2022, 31(1): 128 – 134.
XIANG Y. Network interdiction: emerging research topics and progress [J]. Operations Research and Management Science, 2022, 31(1): 128 – 134. (in Chinese)

[5] 姜龙亭, 寇雅楠, 王栋, 等. 改进近似动态规划法的攻击占位决策[J]. 火力与指挥控制, 2019, 44(7): 135 – 141.
JIANG L T, KOU Y N, WANG D, et al. Attack placeholder decision based on improved approximate dynamic programming[J]. Fire Control & Command Control, 2019, 44(7): 135 – 141. (in Chinese)

[6] MILLS P, COSTELLO R, SARGENT M, et al. Assessing agile combat employment for the pacific air forces[R]. Santa Monica, CA: Rand Cooperation, 2024.

[7] 梁星星, 马扬, 冯旸赫, 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. (in Chinese)

[8] LI S Q, SU J M, SHI Q, et al. A MARL-based approach to intelligent strategic decision-making for air-sea confrontation[C]// Proceedings of 2023 11th China Conference on Command and Control, 2023.

[9] FOERSTER J N, FARQUHAR G, AFOURAS T, et al. Counterfactual multi-agent policy gradients[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 2974 – 2982.

[10] YU C, VELU A, VINITSKY E, et al. The surprising effectiveness of PPO in cooperative, multi-agent games[EB/OL]. (2022 – 11 – 04)[2024 – 02 – 01]. https://arxiv.org/abs/2103.01955v4.

[11] SUNEHAG P, LEVER G, GRUSLYS A, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward[EB/OL]. (2017 – 06 – 16)[2024 – 02 – 01]. https://arxiv.org/abs/1706.05296.

[12] RASHID T, SAMVELYAN M, DE WITT C S, et al. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning[EB/OL]. (2018 – 06 – 06)[2024 – 02 – 01]. https://arxiv.org/abs/1803.11485v2.

[13] SON K, KIM D, KANG W J, et al. QTRAN: learning to factorize with transformation for cooperative multi-agent reinforcement learning [C]// Proceedings of the 36th

International Conference on Machine Learning, 2019.

[14] WANG J H, REN Z Z, LIU T, et al. QPLEX: duplex dueling multi-agent *Q*-learning[EB/OL]. (2021 - 10 - 04)[2024 - 02 - 01]. https://arxiv.org/abs/2008.01062.

[15] YANG Y D, HAO J Y, LIAO B, et al. Qatten: a general framework for cooperative multiagent reinforcement learning[EB/OL]. (2020 - 06 - 09)[2024 - 02 - 01]. https://arxiv.org/abs/2002.03939v2.

[16] HU S Y, ZHU F D, CHANG X J, et al. UPDeT: universal multi-agent reinforcement learning via policy decoupling with transformers[EB/OL]. (2021 - 02 - 07)[2024 - 02 - 01]. https://arxiv.org/abs/2101.08001.

[17] WANG J H, ZHANG Y, KIM T K, et al. Shapley *Q*-value: a local reward approach to solve global reward games[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 7285 - 7292.

[18] WANG J H, ZHANG Y, GU Y J, et al. SHAQ: incorporating shapley value theory into multi-agent *Q*-learning[EB/OL]. (2023 - 01 - 09)[2024 - 02 - 01]. https://arxiv.org/abs/2105.15013.

[19] LI J H, KUANG K, WANG B X, et al. Shapley counterfactual credits for multi-agent reinforcement learning[C]// Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021: 934 - 942.

[20] MAHJOUB O, DE KOCK R, SINGH S, et al. Efficiently quantifying individual agent importance in cooperative MARL[EB/OL]. (2024 - 01 - 26)[2024 - 02 - 02]. https://arxiv.org/abs/2312.08466v2.

[21] HEUILLET A, COUTHOUIS F, DÍAZ-RODRÍGUEZ N. Collective eXplainable AI: explaining cooperative strategies and agent contribution in multiagent reinforcement learning with shapley values[J]. IEEE Computational Intelligence Magazine, 2022, 17(1): 59 - 71.

[22] ANGELOTTI G, DÍAZ-RODRÍGUEZ N. Towards a more efficient computation of individual attribute and policy contribution for post-hoc explanation of cooperative multi-agent systems using Myerson values[J]. Knowledge-Based Systems, 2023, 260: 110189.