知识数据协同的多对手智能空中博弈策略设计

冯锦元[1,2],陈 敏[1*],李俊影[1],陈加乐[3],蒲志强[1,2],陈敏杰[4],孙方义[4]
(1. 中国科学院自动化研究所,北京 100190;2. 中国科学院大学人工智能学院,北京 100049;
3. 南京信息工程大学计算机学院,江苏南京 210044;4. 华如研究院,北京 100193)

摘 要: 人工智能技术的迅速发展赋予了空战自主对抗策略超越人类专家的潜力. 现有智能空战对抗策略依据驱动方式主要包含两类:其中,基于知识规则的策略对应用场景和专家知识依赖性强,而以强化学习为代表的数据驱动的策略可解释性差、泛化性弱. 本文以全国智能空中博弈大赛多机协同空战为背景,提出了一种知识数据协同的多对手的空中博弈策略设计方法. 其中,知识驱动部分基于专家知识设计一种参数化、风格化的策略,以生成高质量离线数据及初始化策略;数据驱动部分基于图注意力网络对队友、对手的信息进行针对性表征,提升训练效率及收敛性能. 进一步,以动态对手匹配机制进行多对手强化学习训练,进一步提升策略泛化性. 该策略与大赛16强中的12支队伍对抗,达到70%以上的统计胜率,这些队伍均采用最新的知识或数据驱动方法,风格各异,同时具有较强的作战能力.

关键词: 强化学习;知识数据协同驱动;空中博弈;多对手;泛化性
基金项目: 国家自然科学基金(No.62073323);中国科学院战略性先导研究项目(No.XDA27030403);北京市科技新星计划项目(No.20220484077)
中图分类号: TP183 **文献标识码:** A **文章编号:** 0372-2112(2024)11-3809-14
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230985

Knowledge-Based and Data-Driven Integrating Design Methodology for Air Combat Strategy in Multi-Opponent Adversarial Game

FENG Jin-yuan[1,2], CHEN Min[1*], LI Jun-ying[1], CHEN Jia-le[3], PU Zhi-qiang[1,2], CHEN Min-jie[4], SUN Fang-yi[4]
(1. *Institute of Automation*, *Chinese Academy of Sciences*, *Beijing* 100190, *China*;
2. *School of Artificial Intelligence*, *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*;
3. *School of Computer Science*, *Nanjing University of Information Science & Technology*, *Nanjing*, *Jiangsu* 210044, *China*;
4. *Huaru Research Institute*, *Beijing* 100193, *China*)

Abstract: The rapid development of artificial intelligence technology has endowed autonomous air combat strategies with the potential to surpass human experts. Existing intelligent air combat strategies can be classified into two categories based on their driving methods: knowledge-based strategies, which heavily rely on application scenarios and expert knowledge; and data-driven strategies, represented by reinforcement learning, which have poor interpretability and weak generalization. In this study, focusing on the scenario of multi-agent cooperative air combat from the air intelligence game (AIG)—a knowledge-based and data-driven integrating strategy design method is proposed. The knowledge-based part utilizes expert knowledge to design a parameterized and stylized knowledge-based artificial intelligence (AI) system, which generates high-quality offline data and initializes the strategy. The data-driven part employs graph attention networks to selectively represent information about teammates and opponents, aiming to improve training efficiency and convergence performance. Furthermore, a dynamic opponent matching mechanism is introduced for multi-agent reinforcement learning training to enhance strategy generalization. The proposed strategy achieved a statistical winning rate of over 70% when competing against 12 teams from the top 16 teams in AIG. It is worth mentioning that these teams all adopt the latest knowledge-based or data-driven methods, with diverse styles, and at the same time, they have strong combat capabilities.

Key words: reinforcement learning; knowledge and data integrating; air combat; multi-opponent; generalization

收稿日期:2023-10-19;修回日期:2024-05-27;责任编辑:梅志强
*通讯作者:陈敏

Foundation Item(s): National Natural Science Foundation of China (No.62073323); Strategic Priority Research Program of Chinese Academy of Sciences (No.XDA27030403); Beijing Nova Program (No.20220484077)

1 引言

未来空战将朝着集群化和智能化方向发展. 其中,集群数量扩展带来了计算复杂度的指数增长,智能化对空战策略泛化性提出了高要求,自主分布式决策给关系复杂的集群带来了协同困难问题,这一系列挑战引起了各领域学者广泛关注.

现有智能空战策略依据驱动方式主要分为基于知识的方法和数据驱动方法. 其中,知识驱动方法[1]主要基于专家系统、模糊推理、进化计算和博弈论等. 专家系统将人类知识归纳为规则库,然而,在集群空战场景中,专家知识难以覆盖所有可能的情况[2]. 此外,专家系统规则库的弱点易在博弈与策略分析中被对手所挖掘并反复利用[3,4]. 因此,模糊推理、进化计算[5]等方法通过构建概率图网络或设计启发式目标函数,改进了基于规则的系统. 博弈论方法通过微分博弈建立数学模型进行求解,具有严谨的理论保证. 文献[6]将微分博弈模型和Markov决策扩展到多阶段决策,以有效计算多阶段博弈的均衡解. 但经典的博弈论方法的高计算复杂度难以满足高动态空战场景的实时决策要求. 总体而言,基于知识方法所设计的空中博弈策略水平有限且难以应用于大规模复杂协同空战场景,另一方面,其规则设计往往均依赖于特定问题场景,难以复用.

在数据驱动方法中,深度强化学习通过结合深度学习的表征学习能力及强化学习的自主探索能力,逐渐成为目前设计空战策略的一类重要方法. 在空战背景下,国防高级研究计划局(Defense Advanced Research Projects Agency,DARPA)开展了空战进化计划(Air Combat Evolution,ACE),旨在发展智能化有人/无人空战对抗技术[7]. DARPA开发人工智能(Artificial Intelligence,AI)在AlphaDogfight场景中以5∶0击败了专家级人类飞行员,开创了基于人工智能方法的空战智能体达到人类专家水平的先例[8]. 文献[9]提出了一种端到端的分层梯度算法进行策略学习,以自博弈方式学习策略,突破了专家知识的限制;同时,利用分层决策网络处理离散与连续混合的动作空间,有效地减轻了决策的负担. 文献[10]针对深度确定性策略梯度(Deep Deterministic Policy Gradient,DDPG)算法探索能力不足的问题,引入启发式探索策略. 文献[11]基于图神经网络解决了具有复杂关系的大规模自主空战问题. 文献[12]针对复杂的态势信息,利用无人机的同质性,设计了多智能体Transformer网络结构. 然而,上述方法在探索过程中缺乏专家知识的指导,难以在复杂场景下从零开始端到端地学习出达到人类水平的策略.

知识数据协同驱动算法为空战协同策略设计提供了新范式. 这类方法将知识或数据驱动部分中的某一方作为算法组件,融入到另一方中紧密结合成一个完整算法[13]. AlphaStar[14]所使用的星际争霸环境具有大量可操控的智能体以及庞大的状态和动作空间,基于知识的策略难以扩展到多个智能体以及设计出高水平的对抗策略,而数据驱动方法难以从零开始探索出达到人类水平的策略. 因此,AlphaStar首先采集了百万规模的高水平玩家数据集并加入大量的人类先验知识,后使用监督学习方法从知识中训练初始化策略网络,以降低后续数据驱动的强化学习探索的复杂度. 最终,AlphaStar在真实对战中达到宗师级水平,分位超过了99.8%的人类玩家. 文献[15]使用基于一致性理论的编队控制器和避碰策略组成的模型作为先验知识以产生高质量数据,进一步,分别采用模仿学习和强化学习进行预训练与精调,知识的引入大大加速了后续数据驱动方法的训练速度和效率. 文献[16]提出了一种基于多阶段Flipit博弈的欺骗策略选择方法. 该方法在扩展单阶段Flipit博弈模型到多阶段网络欺骗模型的基础上,考虑了时间和空间两个方面的防御过程. 此外,作者还设计了基于深度强化学习的近端策略优化(Proximal Policy Optimization,PPO)算法用于计算防御者的最优时空策略. 在知识驱动层面,利用已有的单阶段Flipit博弈模型,有严谨的理论保证,在数据驱动层面,基于强化学习算法来自适应调整欺骗策略的部署,以应对更加复杂动态的网络攻击. 在知识数据协同层面,现有方法将已有的模型知识融入到强化学习的策略学习之中,但是二者耦合性强,难以迁移.

另一方面,为提升对抗策略的泛化性,多任务训练和对手建模被广泛应用于强化学习训练. 其中,多任务训练方法对多个相似任务进行联合训练,避免陷入单个任务的局部最优;对手建模方法通过智能体与对手的交互,构建对手模型,针对性调整自身策略. GSCU(Greedy when Sure and Conservative when Uncertain)[17]算法离线学习对手策略表征,与基于变分自编码器的策略嵌入学习方法相比,该方法的策略表征学习与其他信息的表征学习分离,仅根据对手索引调节编码器. MATE(Multi-Agent Task Embeddings)[18]算法将每个轨迹步中的信息(观测、动作、奖励)和当前的隐变量作为输入更新策略表征. OMDDPG(Opponent Modeling DDPG)[19]算法使用变分自编码器对对手进行建模,仅基于我方智能体的局部观测识别潜在的对手模型,打

破了对手观测和动作信息可访问的假设. LIAM(Local Information Agent Modelling)[20]算法对智能体轨迹与对手轨迹的关联性进行了建模研究. 上述方法直接使用环境中原始的观测信息、动作空间,并未引入专家知识对观测信息、动作空间进行针对性分组或重构. 此外,现有的针对多机协同空中博弈场景下的多任务训练和对手建模方法仅适用于小规模对抗场景,且难以直接迁移至复杂的多机协同空中博弈场景.

根据以上分析,基于知识的空中博弈策略利用已有的空中博弈对抗知识,包括博弈对抗模型、专家规则知识库等,具有完备的理论体系支撑. 然而,知识获取的代价高昂且难以扩展到复杂对抗场景. 此外,知识的质量限制了策略的性能上限,也难以支持空中博弈策略的持续学习与进化. 数据驱动的方法具有无需精确建模、自主探索以及从数据中持续学习、迭代进化的优点. 然而,此类方法缺乏专家知识指导,泛化性差且依赖高质量数据. 本文将知识方法和数据驱动方法相结合,利用各自优势,形成知识与数据协同驱动的策略设计新范式,旨在设计出高水平、泛化性强、可持续学习、迭代进化的对抗策略. 本文创新性地提出一种知识数据协同的多对手空战博弈对抗策略设计方法,其中包含风格化知识智能体(知识AI)设计、离线预任务训练、强化学习智能体设计(强化AI). 算法中使用模仿学习、多任务强化学习等人工智能方法,在协同空战场景中训练出高水平、泛化性强的策略. 具体来说,首先设计层次化、参数化、风格化的知识AI,通过其与多个目标对手的交互构建高质量离线数据库,并进行预任务训练产生初始化策略;进一步,基于分组图注意力网络针对性地表征队友、对手的信息,并采用动态多对手匹配机制进行强化学习训练,产生高水平、泛化性强的对抗策略. 本文的主要贡献如下:

(1)在问题建模层面,针对多机协同空战任务进行建模,一是引入专家知识对观测进行针对性分组重构,基于重构后的信息将飞机间关系建模为图,二是将离散与连续混合动作空间约简为绝对、相对耦合的离散动作空间,三是对智能体在强化学习随机探索阶段进行基于专家知识的探索约束.

(2)在知识数据协同层面,基于层次化、参数化、风格化知识智能体(知识AI)和离线预任务实现知识数据协同. 其中层次化是基于经典军事理论(Observe-Orient-Decide-Act,OODA)[21]解决了空战任务中的态势理解、群体协商及任务分配问题;参数化通过对策略进行表征实现了初始风格设定以及动态风格调整;预任务训练产生强化AI的初始化策略.

(3)在多对手对抗层面,基于分组图注意力机制与动态对手匹配机制训练强化AI的策略. 其中,图注意力网络对队友、对手的信息进行针对性表征,提高训练效率;动态对手匹配机制根据对战不同对手的胜率动态选择对手,进一步提升策略的收敛性能和泛化性.

2　问题描述

本文使用首届全国空中智能博弈大赛(http://cicc-aig.c2.org.cn/)中的空战对抗场景,该场景以异构、全透明态势空战为背景,作战想定为红蓝双方均为1架有人机携带4架无人机进行自由空战,并模拟雷达探测功能及空空导弹攻击过程. 其中有人机挂载4枚中距空空导弹,每架无人机挂载2枚中距空空导弹,且有人机的其余各项性能参数高于无人机. 对抗的具体场景如图1所示. 我方的有人/无人机需要在有全局信息的情况下获得最后的胜利. 胜负判定的规则优先级如下:

(1)一方若有人机战损,则战败,回合结束.

(2)一方若无弹,则战败,仿真结束.

(3)在双方战损数和剩余导弹数相同的情况下,控制作战中心区域时间较长的一方获胜.

实际空战中,不同对手策略风格不可避免地呈现出差异性,为提升空战策略在面对不同对手时的表现,本文采用首届大赛16强中的12支队伍作为训练和测试对手策略,这些队伍均采用最新的知识或数据驱动方法,风格各异,同时具有较强的作战能力. 为了保护队伍隐私,本文中使用OP1~OP12代替队伍名称.

本文多机协同空战场景被建模为特殊的马尔可夫博弈(N,S,A,P,r,γ)[22],N表示智能体数量,N个智能体组成我方队伍集合$I=\{1,2,\cdots,N\}$. 在马尔可夫博弈的标准假设之外,提供两组对手策略集合,训练集$T=\{I_{-1,m}\}_{m=1}^{m=M_T}$和测试集$G=\{I_{-1,m}\}_{m=1}^{m=M_G}$,$I_{-1,m}$代表对手策略,用于提供联合动作$A_{-1}$,$M_G$和$M_T$表示训练集和测试集中对手策略的个数. $S=S_1\times\cdots\times S_i\times\cdots\times S_N$是智能体的联合状态空间,$S_i$表示智能体$i$的状态空间. $A=A_1\times A_{-1}$为联合动作空间,其中$A_1=A_{1,1}\times\cdots\times A_{1,i}\times\cdots\times A_{1,N}$,$A_{1,i}$表示智能体$i$的动作空间. $P:S\times A\times S\to[0,1]$是状态转移函数. $r_i(s_t,a_{1,1},\cdots,a_{1,i},\cdots,a_{1,N},a_{-1,1},\cdots,a_{-1,N},s_{t+1})$表示奖励函数,$s_t$表示智能体$i$在时间步$t$的状态,$s_{t+1}$表示智能体$i$在时间步$t+1$的状态. 在每一幕的开始从$T$中挑选对手用于训练,从$G$中挑选对手进行测试. 使用强化学习训练智能体集合$I$,目标是最大化训练集$T$中的对手对抗的平均回报,即找到策略网络$\pi$的一组参数$\theta^*$满足:

$$\theta^*=\arg\max_{\theta}\mathrm{E}_{\pi}[\mathrm{E}_{T}[\sum_{t}\gamma^t r_t]] \tag{1}$$

3　任务建模

分析主流的游戏AI,例如AlphaStar[14]、AlphaGo/AlphaZero[23,24]、OpenAI Five[25]、腾讯觉悟[26]等,其任务的

图1 空战对抗场景示意图(双方各包含1驾有人机和4驾无人机)

针对性建模为后续智能体训练提供了重要支撑. 多机协同空战场景与上述游戏场景类似,解空间庞大且态势动态变化,需要近实时决策. 本文对多机协同空战场景的任务建模主要包括三个方面:(1)状态空间设计;(2)动作空间设计;(3)动作屏蔽设计.

3.1 状态空间设计

3.1.1 态势信息处理

本文基于首届全国空中智能博弈大赛的仿真推演平台XSimStudio(XSIM)仿真了现实中整个5v5空战对抗流程. 该平台模拟多种编队及平台级行为,具有自主可控、高扩展性等优点.

XSIM输出原始空战对抗实时态势信息,通过态势信息处理模块进行重构,获取对环境的观测信息. 原始态势信息经过提取、处理和观测信息重构后,输入到知识AI和强化AI中. XSIM仿真推演平台返回的态势信息内容包括仿真时间、红方态势和蓝方态势三部分. 以红方态势为例,红方态势包含武器平台信息、情报信息、地方发射的导弹信息. 为了更好地表征观测信息并降低计算复杂度,引入专家知识对观测信息进行针对性分组重构. 态势信息处理模块如图2所示.

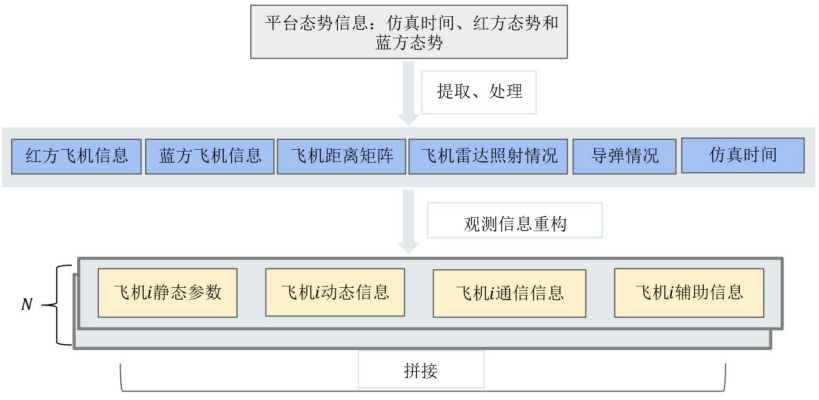


图2 态势信息处理模块

飞机(有人机/无人机)i在时间t重构后的观测信息表示为

$$\boldsymbol{O}_i^t=[\boldsymbol{O}_{i,\mathrm{sta}},\boldsymbol{O}_{i,\mathrm{dyn}}^t,\boldsymbol{O}_{i,\mathrm{com}}^t,\boldsymbol{O}_{i,\mathrm{oth}}^t] \tag{2}$$

$$\boldsymbol{O}_{i,\mathrm{com}}^t=[\boldsymbol{O}_{i,\mathrm{plane}}^t,\boldsymbol{O}_{i,\mathrm{missile}}^t] \tag{3}$$

静态参数向量$\boldsymbol{O}_{i,\mathrm{sta}}$包含飞机本身的固定参数,由于有人机和无人机共享策略网络,静态参数的差异化输入用于有效引导策略网络输出差异化行为. 静态参数包含飞机类型及飞机编号,通过归一化的方式减小飞机性能参数的差异,以稳定策略梯度. 以速度为例,将飞机i的速度大小转化为$v_i/v_i^{\max}$,其中v_i为当前速度,$v_i^{\max}$为最大飞行速度,从而有人机/无人机的速度参数区间表示一致,因此无需引入飞机性能参数差异.

动态信息$\boldsymbol{O}_{i,\mathrm{dyn}}^t$表示飞机$i$在时间$t$对自身和周围环境的实时认知,在强化学习中起着关键作用. $\boldsymbol{O}_{i,\mathrm{dyn}}^t$中包

括飞机 i 相对于边界的距离、自身是否被雷达照射、当前遭受的导弹锁定数量、剩余的导弹数量以及飞机位姿和速度等信息.

通信信息 $\boldsymbol{O}^t_{i,\mathrm{com}}=[\boldsymbol{O}^t_{i,\mathrm{plane}},\boldsymbol{O}^t_{i,\mathrm{missile}}]$ 包含其他飞机信息 $\boldsymbol{O}^t_{i,\mathrm{plane}}$ 和针对飞机 i 的导弹信息 $\boldsymbol{O}^t_{i,\mathrm{missile}}$,用于促进多机协同场景下智能体间的合作与对抗. 其他飞机信息 $\boldsymbol{O}^t_{i,\mathrm{plane}}$ 包括队友和对手的相对位置、相对距离以及它们的静态参数和动态信息. 针对飞机 i 的导弹信息 $\boldsymbol{O}^t_{i,\mathrm{missile}}$ 包括导弹相对于飞机的位置和距离.

其他信息 $\boldsymbol{O}^t_{i,\mathrm{oth}}$ 是加速强化 AI 训练的信息,包括最近的队友、对手的飞机编号等信息,用于减轻智能体推理的负担.

3.1.2　状态空间设计

与 1v1 空战不同,集群空战需要考虑与其他无人机(队友和对手)的相对状态. 本文基于重构后的观测信息,将飞机之间的关系建模为相对关系图.

定义 1　相对关系图(Relative Relationship Graph, RRG):在 M vs N 空战场景中,每架飞机与其他飞机的关系建模为以自身为中心的图 $G=(V,E)$. V 代表图 G 中结点集合,包含自己、队友和对手在内的 $M+N$ 架飞机. E 代表中心结点指向 V 中其余节点的边的集合,包含 $M+N-1$ 条边. 每个结点对应一架飞机实体,包含飞机的静态参数和动态信息,每条边表示两架飞机之间的相对信息.

飞机 i 在时间 t 的结点向量在 RRG 中表示为

$$\boldsymbol{v}^t_i=[\boldsymbol{O}_{i,\mathrm{sta}},\boldsymbol{O}^t_{i,\mathrm{dyn}}] \tag{4}$$

静态参数向量 $\boldsymbol{O}_{i,\mathrm{sta}}$ 包含飞机本身的固定参数. 动态信息 $\boldsymbol{O}^t_{i,\mathrm{dyn}}$ 包含飞机相对于边界的距离、自身是否被雷达照射、当前被几枚导弹锁定、剩余的导弹数量、自身位姿和速度等信息. 结点 i 到结点 j 的边向量表示为

$$\boldsymbol{e}^t_{ij}=[\Delta x^t_{ij},\Delta y^t_{ij},d^t_{ij}] \tag{5}$$

其中,$\Delta x^t_{ij},\Delta y^t_{ij},d^t_{ij}$ 分别表示横向的相对距离、纵向的相对距离以及飞机 i 和飞机 j 的欧式距离. 将重构后的部分观测信息 $[\boldsymbol{O}_{i,\mathrm{sta}},\boldsymbol{O}^t_{i,\mathrm{dyn}},\boldsymbol{O}^t_{i,\mathrm{plane}}]$ 表示为相对关系图,飞机 i 在时间 t 的状态表示为

$$\begin{aligned}
&\boldsymbol{s}^t_i=[\boldsymbol{G}^t_i,\boldsymbol{O}^t_{i,\mathrm{missile}},\boldsymbol{O}^t_{i,\mathrm{oth}}]\\
&\boldsymbol{G}^t_i=[\boldsymbol{V}^t_i,\boldsymbol{E}^t_i]\\
&\boldsymbol{V}^t_i=[\boldsymbol{v}^t_i,\boldsymbol{v}^t_{j,j\in N^t_{i,\mathrm{f}}},\boldsymbol{v}^t_{k,k\in N^t_{i,\mathrm{o}}}]\\
&\boldsymbol{E}^t_i=[\boldsymbol{e}^t_{ij,j\in N^t_{i,\mathrm{f}}},\boldsymbol{e}^t_{ik,k\in N^t_{i,\mathrm{o}}}]
\end{aligned} \tag{6}$$

其中,$\boldsymbol{G}^t_i$ 表示飞机 i 在时间 t 时构建的相对关系图向量;$\boldsymbol{V}^t_i,\boldsymbol{E}^t_i$ 表示图 $\boldsymbol{G}^t_i$ 中的结点向量和边向量;$N^t_{i,\mathrm{f}},N^t_{i,\mathrm{o}}$ 表示飞机 i 在时间 t 的队友结点和对手结点.

3.2　动作空间设计

本文中的动作设计参考了 AlphaStar[14] 的动作设计方式,将有人/无人机的动作选择过程划分为三个维度:(1)动作类型;(2)动作对象;(3)执行方式. 整个动作空间被离散化为 133 个离散动作.

动作类型划分为六种,包含打击、正向移动、反向移动、垂直移动、占领中心和绝对方向移动. 动作对象包含四类 $v^t_i,N^t_{i,\mathrm{f}},N^t_{i,\mathrm{o}},N^t_{i,\mathrm{m}}$,分别表示飞机 i 本身、队友结点、对手结点及与飞机 i 最近的两枚导弹结点. 执行方式设计为加速和减速两种,动作空间设计见表 1.

表 1　动作空间设计

动作类型	动作对象	执行方式	动作总数
打击	$N^t_{i,\mathrm{o}}$	—	5
正向移动	$N^t_{i,\mathrm{f}}$、$N^t_{i,\mathrm{o}}$	加减速	22
反向移动	$N^t_{i,\mathrm{f}}$、$N^t_{i,\mathrm{o}}$、$N^t_{i,\mathrm{m}}$	加减速	22
垂直移动	$N^t_{i,\mathrm{f}}$、$N^t_{i,\mathrm{o}}$、$N^t_{i,\mathrm{m}}$	加减速	66
占领中心	v^t_i	加减速	2
绝对方向移动	v^t_i	加减速	16

垂直移动动作分为三类:顺时针垂直方向、逆时针垂直方向、选择与自身运动方向夹角较小的垂直方向. 如果移动动作的对象选择自身属于绝对移动,绝对移动包含两个动作类型:占领中心和绝对方向移动,绝对方向移动离散化为八个运动方向.

本文的动作空间设计将离散与连续混合动作约简为绝对动作、相对动作耦合的同时,实现了飞机的精细控制. 动作对象选择包含自身和其他结点,用于衍生出相对移动动作,克服只有绝对移动动作时决策推演计算量大的缺点.

动作对象的选择赋予飞机学习特定技巧的能力. 当相对移动的动作对象选择了对手,飞机主要学习追击、围捕以及规避、逃逸技巧;当相对移动动作的动作对象选择了队友,飞机学习编队和协同打击技巧;当相对移动动作的动作对象选择了导弹,智能体学习规避导弹的技巧. 通过补充绝对移动动作,弥补了三类相对动作无法实现的策略行为. 采用加减速的执行方式实现了在不改变雷达照射角度的前提下改变队形.

3.3　动作屏蔽设计

动作屏蔽充分利用专家知识,在强化 AI 随机探索阶段限制探索空间,加速收敛. 本文中共设计了三类动作屏蔽:非法屏蔽、经验屏蔽和备用屏蔽.

非法屏蔽是指屏蔽掉智能体的非法动作和不合理行为. 例如,智能体的动作对象不能选择已经阵亡的飞机和已经失去制导的导弹;当目标阵亡、没有进入射程、被照射时或智能体自身没有导弹时禁用打击动作.

经验屏蔽基于领域知识进一步约简动作空间,避免无效探索,本文针对有人机引入了两个经验屏蔽:在比赛开始时,禁用了最高速度运动,避免被对手集中攻击导致阵亡;有人机被导弹锁定时,采取基于专家知识的导弹规避策略.

备用屏蔽中包含一些可选的屏蔽方式,使用这类屏蔽有利于加速智能体的训练,但会限制智能体策略的性能上限,例如减速动作屏蔽、绝对动作屏蔽等.

4 知识数据协同的多对手空战博弈对抗策略设计

整体训练框架分为两部分:针对多机协同空战场景的任务建模和多对手空战博弈对抗策略设计.整体训练流程如图3所示.

任务建模部分包括状态空间设计、动作空间设计和动作屏蔽设计.状态空间设计基于平台全透明态势信息引入专家知识进行针对性分组重构,重构后的信息将飞机间关系建模为图.动作空间设计将原协同空战场景离散与连续混合的动作空间约简为绝对、相对耦合的离散动作空间.动作屏蔽设计基于专家知识对强化学习随机探索阶段进行约束.

多对手空战博弈对抗策略设计包含三个模块:风格化知识智能体(知识AI)、离线预任务、强化学习智能体(强化AI).首先,设计一种参数化、层次化的风格化知识智能体(知识AI),其基于专家知识完成整个策略的设计并通过参数调优提升策略水平.然后,通过知识AI与多个目标对手交互构建高质量离线数据库,基于离线数据库通过模仿学习初始化强化AI的策略.离线数据库也可用于训练结果预测、事件预测线预任务.最后,强化AI引入队友、对手的观测表征,使用强化学习提升策略水平,产生高水平、泛化性强的对抗策略.

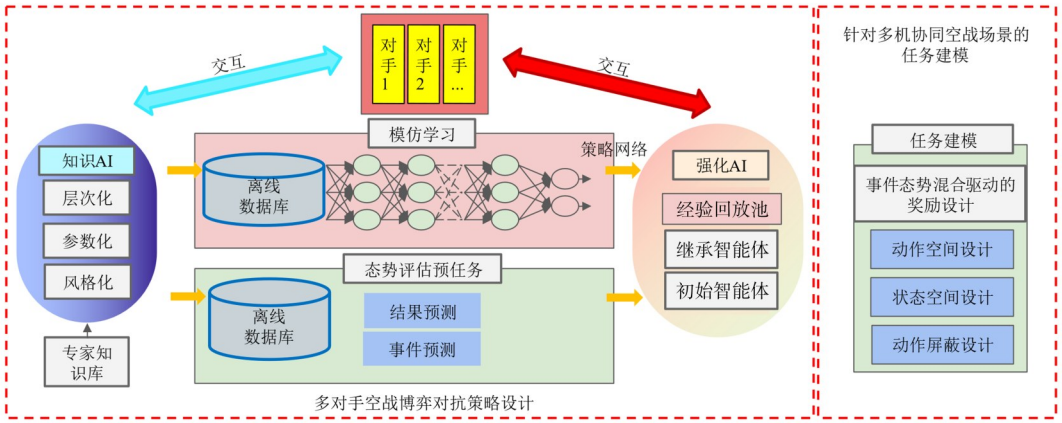


图3 策略训练整体流程

4.1 风格化知识智能体(知识AI)设计

本文提出了一种层次化、参数化的风格化知识智能体(知识AI)设计方法.其中,层次化是基于专家知识解决了空战任务中的态势理解、群体协商及任务分配问题.通过静态、动态价值计算度量可选动作的收益,同时将群体协商及任务分配问题转化为无约束的优化问题.参数化通过对策略进行表征实现了初始风格设定以及动态风格调整.通过风格参数调整,知识AI切换不同的作战风格,并通过参数优化实现高质量的人工策略.

知识AI的整体执行流程如图4所示,构成一个循环,不断地执行动作、接收反馈、修正参数.该流程帮助知识AI渐进地优化策略,提高策略水平.

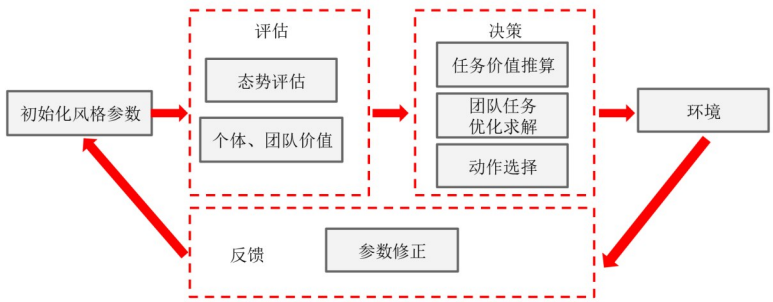


图4 知识AI交互流程图

初始化风格参数阶段设定了一些关键风格参数,包括有人机价值r_{MAV}、无人机价值r_{UAV}、导弹价值r_{missile}和打击距离β等.关键风格参数在后续的态势评估和动作选择中起重要作用.

参数初始化阶段实现了知识 AI 的初始化风格设定. 态势评估阶段根据当前对局状况评估我方优劣态势. 知识 AI 划分了三种态势:敌优我劣、敌劣我优、势均力敌. 态势划分通过以下指标反映当前的优劣态势:我方剩余无人机数 m_{UAV}、我方剩余导弹数 m_{missile}、敌方剩余无人机数 o_{UAV}、敌方剩余导弹数 o_{missile} 及积极因子 α. 具体而言,敌优我劣的判定标准如下:

$$(m_{\mathrm{missile}}-1)+\alpha<(o_{\mathrm{UAV}}-m_{\mathrm{UAV}}) \tag{7}$$

敌劣我优的判定标准如下:

$$(o_{\mathrm{missile}}-1)-\alpha<(m_{\mathrm{UAV}}-o_{\mathrm{UAV}}) \tag{8}$$

上述标准外的态势均判定为势均力敌. 在风格初始化的时候,预先指定了积极因子 α,其数值的大小反映了我方对态势把控的积极程度. 数值越大,表示对态势的评估越乐观. 通过与对手的交互,结合网格搜索、贝叶斯优化等方法修正相关参数如 β 等,帮助知识 AI 在不同态势下作出合理决策.

进一步地,通过个体、团队价值模块评估当前态势,个体团队价值分为两部分:静态价值与动态价值. 静态价值基于关键风格参数衡量当前时刻剩余的战斗力,静态价值 V_{static} 计算如下:

$$V_{\mathrm{static}}=r_{\mathrm{missile}}\times m_{\mathrm{missile}}+r_{\mathrm{UAV}}\times m_{\mathrm{UAV}} \tag{9}$$

同时,通过团队动态价值计算评估当前机群的协同能力和所处形势. 计算度量可选动作的收益动态价值计算多种因素的共同影响,如对手行为、环境变化等. 此处以飞机之间的协同能力为例,计算动态价值 V_{dynamic}:

$$V_{\mathrm{dynamic}}=\frac{(n^2-n)}{2}\times\frac{\gamma}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sqrt{(x_i-x_j)^2+(y_i-y_j)^2+(z_i-z_j)^2}} \tag{10}$$

其中,n 代表剩余的飞机数量;x_i,y_i,z_i 分别代表飞机 i 的三维坐标;γ 作为超参数,调节动态价值在决策中所占的比重. 动态价值的值越大,飞机之间的连结越紧密,协同能力越强.

在任务价值推算模块中,价值计算方式为执行不同任务给团队价值带来的改变,进而对每一架飞机评估当前的所有任务,并分别计算执行各个任务所能获得的价值. 团队任务优化求解是将任务选择转化为无约束的优化问题,通过选择最优的任务来确定执行的动作.

在求解过程中,通过求解任务最大值问题来寻找最佳的任务分配方案. 本文对知识 AI 离散地设定多组初始风格参数,在环境中执行选择动作后,根据环境给出的反馈使用网格搜索法来寻找最优参数设置,参数修正帮助知识 AI 逐渐优化策略,提高策略鲁棒性和泛化性.

4.2　离线数据及预任务训练

预任务训练基于离线数据集训练模型,旨在获得能够快速适应不同下游任务的网络模型. 这种思想也被应用到强化学习领域,例如模仿学习、自监督强化学习等方法. 通过预训练,给予智能体对环境或任务的初始知识,从而显著提高训练效率和收敛性能.

本文提出了两种预任务训练方法. (1)模仿学习:产生初始化策略;(2)态势评估预任务:作为可选方案,用于构建密集奖励、预测态势等任务.

4.2.1　模仿学习

本文采用了 12 支不同对手策略用于交互,这些对手策略均采用最新的知识或数据驱动方法,风格各异,同时具有较强的作战能力,标记为 OP1~OP12. 模仿学习基于知识 AI 与 OP1~OP12 交互的离线数据,交互的数据表示为智能体状态与所选择的动作构成的二元组 (s, a).

对上述离线数据模仿学习后,获得的策略具有一定的对抗能力,模仿学习获得的策略为强化 AI 提供了初始化策略.

本文使用行为克隆方法来对策略行为进行模仿,行为克隆[27]是模仿学习中的一种算法,通过监督学习来实现. 行为克隆最小化学习策略 π 和专家行为策略 π_{E} 之间的差异,即最小化目标:

$$\min_{\pi}\mathrm{E}_{s\sim d_{\pi_{\mathrm{E}}}}\left[D_{\mathrm{KL}}\left(\pi_{\mathrm{E}}(\cdot|s),\pi(\cdot|s)\right)\right]:=\mathrm{E}_{(s,a)\sim\rho_{\pi_{\mathrm{E}}}}\left[\log\left(\frac{\pi_{\mathrm{E}}(a|s)}{\pi(a|s)}\right)\right] \tag{11}$$

其中,$d_{\pi_{\mathrm{E}}}$ 和 $\rho_{\pi_{\mathrm{E}}}$ 分别是由策略 π_{E} 产生的状态分布和状态动作分布,定义如下:

$$d_{\pi}(s)=\mathrm{E}\left[\sum_{t=0}^{\infty}\gamma^t P(s_t=s)|s_0\sim\rho,\ a_t\sim\pi(\cdot|s_t),\forall t\geqslant 0\right] \tag{12}$$

$$\rho_{\pi}(s,a)=\mathrm{E}\left[\sum_{t=0}^{\infty}\gamma^t P(s_t=s,\ a_t=a)|a_t\sim\pi(\cdot|s_t),\forall t\geqslant 0\right] \tag{13}$$

用于行为克隆的网络结构与强化 AI 的 actor 网络一致. 网络输入为重构后的环境观测向量,输出为离散化后的动作选择向量.

4.2.2　态势评估预任务

态势评估预任务作为一种可选方案,包括构建密集奖励、预测态势等一系列任务,旨在在强化学习难以训练时提供帮助. 本节以结果预测任务为例,阐述态势评估预任务的设计事件与态势混合驱动的奖励设计,从事件和态势两个角度来构建密集奖励,但是在关键事件发生之外,大多数时间段奖励以态势奖励为主导,即引导飞机前往战场中央. 结果预测任务基于离线数据库,使用监督学习方法构建更为密集的奖励,引导智能体关注更丰富的态势信息. 结果预测任务如图 5 所示.

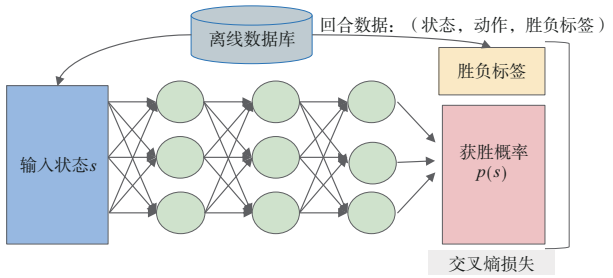


图5 结果预测任务示意图

模仿学习需要动作数据，因此学习所需要的数据库只包含知识AI交互的数据．与模仿学习使用的离线数据库不同，构建结果预测任务的离线数据库需要对记录的回合数据额外添加胜负标签，而不需要使用动作数据，所以数据集可以包含任意的对战数据．引导性密集奖励需要先基于领域知识预先定义一个势能函数$\varphi(s)$，使用势能函数为智能体在每个时间步上生成$r(s_t)$．由于领域知识难以以精确的函数形式表示，因此使用离线数据来预测获胜概率，并基于此构建密集奖励[28]：

$$\begin{aligned}\varphi(s_t)&=2\times p(s_t)-1\\ r(s_t)&=[\gamma\varphi(s_{t+1})-\varphi(s_t)]\times \text{scale}\end{aligned}\tag{14}$$

其中，$p(s_t)$表示t时刻状态s最后获胜的概率，γ是折扣因子，超参数scale用于调节设计的密集奖励的权重．当智能体接近或进入胜率更高的状态时，给予正向的奖励，以鼓励智能体继续采取类似的行为．相反，当智能体处于不利状态时，给予负向的奖励，以强调需要避免的行为．态势评估预任务基于离线数据库，充分利用知识AI与多个对手的交互数据，以提高智能体的决策能力．

4.3 强化学习智能体（强化AI）设计

本文提出一种基于分组图注意力机制与动态对手匹配的强化学习（强化AI）训练方法，主要包括图注意力网络模块、多任务学习模块和强化学习算法模块，如图6所示．图注意力模块基于空战场景中的关系图，对智能体间的关系进行建模．多任务学习模块采用动态对手匹配的启发式方法，根据不同对手的对战胜率动态调整对手的选择，进一步提升策略泛化性，避免某个对手在训练过程中占据主导地位．强化学习算法模块使用MAPPO（Multi-Agent Proximal Policy Optimization）[29]算法优化策略，MAPPO采用价值归一化等技巧，有助于提高训练策略的效率和稳定性．

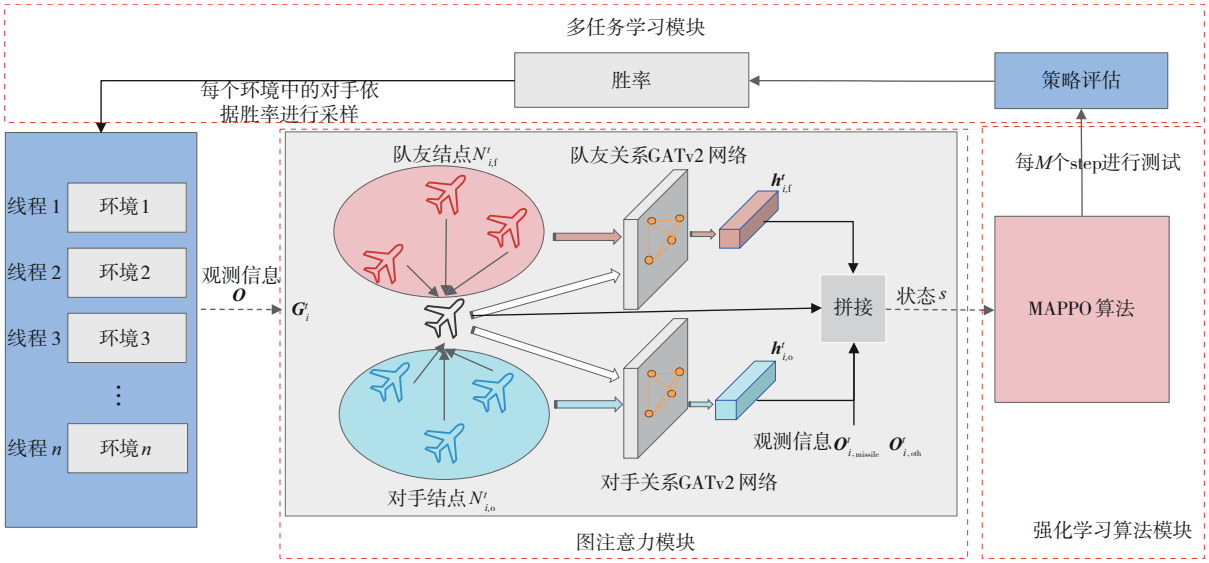


图6 强化AI设计示意图

4.3.1 图注意力模块

在现实的多机协同空战场景中，飞行员需要识别重要的队友和对手信息，并过滤掉不相关的无效信息．类似地，智能体可以选择性地关注周围实体的重要信息，以更好地了解周围环境并把握实体之间的关系．强化AI利用GATv2[30]高效地聚合周围飞机的信息，有选择性地关注队友、对手．

GATv2与GAT[31]在应用场景上具有相似性，但GATv2通过修改注意力计算的方式解决了GAT中的静态注意力问题．首先，将$\boldsymbol{G}_i^t$输入到GATv2中，以获得飞机i的空间特征．然后可以通过下式计算结点对$(i,j)_{j\in N_{i,\mathrm{f}}^t}$之间的注意力系数：

$$a_{ij}^t=\frac{\exp(\boldsymbol{a}_{\mathrm{f}}^{\mathrm{T}}\mathrm{LeakyReLU}([\boldsymbol{W}_{\mathrm{lf}}\boldsymbol{v}_i^t\|\boldsymbol{W}_{\mathrm{rf}}(\boldsymbol{v}_j^t\|\boldsymbol{e}_{ij}^t)]))}{\sum_{j'\in N_{i,\mathrm{f}}^t}\exp(\boldsymbol{a}_{\mathrm{f}}^{\mathrm{T}}\mathrm{LeakyReLU}([\boldsymbol{W}_{\mathrm{lf}}\boldsymbol{v}_i^t\|\boldsymbol{W}_{\mathrm{rf}}(\boldsymbol{v}_{j'}^t\|\boldsymbol{e}_{ij'}^t)]))}\tag{15}$$

其中，$\|$表示拼接操作，T表示转置，a_{ij}^t代表t时刻飞机结点i对飞机结点j的关注系数，$\boldsymbol{W}_{\mathrm{lf}}$和$\boldsymbol{W}_{\mathrm{rf}}$是可训练的权重矩阵，$\boldsymbol{a}_{\mathrm{f}}^{\mathrm{T}}$是单层神经网络的权重向量并选择

LeakyReLU作为激活函数. 对于结点集合$(i,j)_{j\in N_{i,\mathrm{o}}^t}$的注意力系数,可基于不同的训练参数$\boldsymbol{W}_{\mathrm{lo}}$、$\boldsymbol{W}_{\mathrm{ro}}$和$\boldsymbol{a}_{\mathrm{o}}^{\mathrm{T}}$通过式(15)计算得出.

获得归一化的注意力系数之后,通过加权分别聚合队友和对手的特征表示. 飞机结点i的两个图注意力层的输出如下:

$$\boldsymbol{h}_{i,\mathrm{f}}^t=\delta\left(\sum_{j\in N_{i,\mathrm{f}}^t}\alpha_{ij}^t\boldsymbol{W}_{\mathrm{rf}}\left(\boldsymbol{v}_j^t\|\boldsymbol{e}_{ij}^t\right)\right)\tag{16}$$

$$\boldsymbol{h}_{i,\mathrm{o}}^t=\delta\left(\sum_{j\in N_{i,\mathrm{o}}^t}\alpha_{ij}^t\boldsymbol{W}_{\mathrm{ro}}\left(\boldsymbol{v}_j^t\|\boldsymbol{e}_{ij}^t\right)\right)\tag{17}$$

其中,$\boldsymbol{h}_{i,\mathrm{f}}^t$和$\boldsymbol{h}_{i,\mathrm{o}}^t$是通过图注意力网络聚合的飞机结点$i$的队友和对手的高阶特征表示,$\delta$是非线性激活函数. 飞机$i$的高阶状态$\boldsymbol{s}_i^t$表征为

$$\boldsymbol{s}_i^t=[\boldsymbol{v}_i^t,\boldsymbol{h}_{i,\mathrm{f}}^t,\boldsymbol{h}_{i,\mathrm{o}}^t,\boldsymbol{O}_{i,\mathrm{missile}}^t,\boldsymbol{O}_{i,\mathrm{oth}}^t]\tag{18}$$

4.3.2　多任务学习模块

强化学习容易出现严重的过拟合问题[32],多机协同空战场景下表现为对手的特定弱点被反复利用,学习到的策略缺乏泛化能力,难以对抗新的对手. 为解决策略的过拟合问题,引入多任务学习机制,以提高策略的泛化性.

多机协同空战场景下的多任务学习体现为智能体从对手集合$T=\{I_{-1,m}\}_{m=1}^{m=M_T}$中采样对手进行对抗交互来学习策略. 首先,智能体以相同的概率与各个对手进行对抗,并定期评估当前策略在与各个对手对抗时的胜率. 然后,基于评估结果给予当前智能体更大的概率与难度更大的对手进行交互,以提高智能体的训练效果. 动态对手匹配的启发式方法与平均抽样不同,而是使用匹配机制来提供良好的学习信号,有侧重地与对手进行交互. 对手匹配概率的计算如式(17)所示,给定一个智能体正在学习的策略A,从对手池C中抽取一个对手B的概率为

$$\frac{f(P[\text{A beats B}])}{\sum\limits_{c\in C}f(P[\text{A beats C}])}\tag{19}$$

其中,$f(P[\text{A beats B}])$表示策略A与对手B对抗的胜率,f是一种映射函数. AlphaStar[14]将映射函数设计为$f(x)=(1-x)^p$,目的是训练过程中关注胜率较低的对手,其中p用于控制结果分布的熵大小. $f(1)=0$即不会考虑与已经完全击败的对手进行交互. 本文强化AI的动态对手匹配机制采用基于Softmax的对手匹配方式,即$f(x)=\mathrm{e}^{-x}$. 基于Softmax的对手匹配方式与难易结合的课程学习方式类似,策略与能力极强的对手交互时,也能以一定的概率与能力稍弱的对手进行交互,加速策略的学习过程. 与AlphaStar的设计方式不同,$f(x)=\mathrm{e}^{-x}$与$f(x)=(1-x)^p$相比,即使是胜率非常高的对手也会有较小的概率被选择,这种引导在对抗能力较强的对手时起到促进作用,且能够有效避免由部分对手采样概率过低带来的灾难性遗忘问题. 此外,本文用于训练的对手策略能力较强,训练时测试胜率的波动较大,基于Softmax的对手匹配方式能够使得训练更为平稳.

4.3.3　事件与态势混合驱动的奖励设计

现有的工作依赖于专家设计的密集奖励信号[33],本文使用的事件与态势混合驱动的奖励设计方法基于具体的空战胜负规则,无需大量专家知识即可克服环境中稀疏奖励[9]. 这里的关键事件与态势和空战对抗的胜负判定有强相关性,该奖励设计方法避免了过度拟合专家设计的密集奖励信号,有助于降低策略学习的偏差. 奖励函数设计由两部分构成:

(1)　态势奖励

态势奖励r_{situ}分为两部分,态势评估预任务奖励r_{pred}和占领中心任务奖励r_{occu}. 态势评估预任务基于知识AI与多对手交互构建的离线数据库,使用监督学习构建引导性密集奖励. 态势奖励计算如下:

$$\begin{aligned}&r_{\mathrm{situ}}=r_{\mathrm{pred}}+r_{\mathrm{occu}}\\&r_{\mathrm{occu}}=r_{\mathrm{occu}}^{\max}\left(1-\sqrt{(\Delta x_i)^2+(\Delta y_i)^2}\right)\end{aligned}\tag{20}$$

其中,Δx_i和Δy_i表示横向和纵向上飞机i相对战场中心的距离. 占领中心任务奖励r_{occu}的大小取决于飞机相对于战场中心的距离,当飞机位于战场边界时无法获取到占领奖励,当处于战场正中心时会获得最大的奖励$r_{\mathrm{occu}}^{\max}$.

(2)　事件奖励

事件奖励大小设计综合考虑实体价值和事件价值两部分. 实体价值和事件价值从大到小为有人机价值r_{MAV}、无人机价值r_{UAV}、导弹价值r_{missile},占领中心能获得的最大价值$r_{\mathrm{occu}}^{\max}$. 价值的计算方式为

$$\begin{aligned}&r_{\mathrm{missile}}=n\times r_{\mathrm{occu}}^{\max}\times\mathrm{max\_step}\\&r_{\mathrm{UAV}}=2.5\times r_{\mathrm{missile}}\\&r_{\mathrm{MAV}}=n_{\mathrm{UAV}}\times r_{\mathrm{UAV}}+n_{\mathrm{missile}}\times r_{\mathrm{missile}}\end{aligned}\tag{21}$$

其中,n为飞机数量,max_step为一场对战最大步长. 导弹奖励r_{missile}大小设计的含义为即使所有飞机在一整场比赛都去占领中心,获得的最大奖励也不会超过一枚导弹的奖励r_{missile}. 由于双方的弹药总量是12发,有人机的打击难度略高于无人机,因此设计为用2发导弹击毁一架无人机是值得的,如果用3发则是对弹药的浪费,最终将无人机价值r_{UAV}设计为$r_{\mathrm{UAV}}=2.5\times r_{\mathrm{missile}}$. 有人机价值被定义为$r_{\mathrm{MAV}}=n_{\mathrm{UAV}}\times r_{\mathrm{UAV}}+n_{\mathrm{missile}}\times r_{\mathrm{missile}}$,其中$n_{\mathrm{UAV}}$为无人机数量,$n_{\mathrm{missile}}$为导弹数量. 事件驱动的奖励设计如表2所示.

4.3.4　强化学习算法模块

强化学习中策略网络的初始化采用离线预任务中的模仿学习策略. 在与多个对手进行对抗的过程中,采用MAPPO算法[29]进行强化学习训练. MAPPO算法是一种将单智能体的PPO算法推广到多智能体场景的方

表 2 事件奖励设计

奖励事件	奖励
占据中心奖励	r_{occu}
有人机损毁奖励	$-r_{\text{MAV}}$
无人机损毁奖励	$-r_{\text{UAV}}$
逃逸导弹奖励	r_{missile}
击毁有人机奖励	r_{MAV}
击毁无人机奖励	r_{UAV}
未击中奖励	$-r_{\text{missile}}$

法. 本文使用多智能体强化学习算法MAPPO来训练智能体策略,MAPPO算法使用两个神经网络进行训练:动作网络actor和评论家网络critic. 在该算法中,智能体共享参数,并使用智能体的聚合轨迹来更新策略参数. 对于每个智能体i的actor网络,目标是最大化下式:

$$J^{\text{CLIP}}(\theta^i)=\text{E}[\min(\eta_t^i(\theta^i)A_t^i,\text{clip}(\eta_t^i(\theta^i),1-\varepsilon,1+\varepsilon)A_t^i)] \tag{22}$$

其中,

$$\eta_t^i(\theta^i)=\frac{\pi_{\theta^i}(a_t^i|o_t^i)}{\pi_{\theta_{\text{old}}^i}(a_t^i|o_t^i)} \tag{23}$$

$$A_i^t=\sum_{l=0}^{h}(\gamma\lambda)^l\delta_{t+l}^i \tag{24}$$

$$\delta_t^i=r_t^i+\gamma V_{\varphi^i}(s_{t+1})-V_{\varphi^i}(s_t) \tag{25}$$

式中,$\pi_{\theta^i}(a_t^i|o_t^i)$代表智能体$i$在观测$o_t^i$下的新策略概率分布,$\pi_{\theta_{\text{old}}}$同理,表示旧的策略概率分布;$\eta_t^i(\theta^i)$代表新策略与旧策略的概率比,clip()函数对$\eta_t^i(\theta^i)$进行裁剪,根据参数$\varepsilon$来控制裁剪的区间为$[1-\varepsilon,1+\varepsilon]$;$A_i^t$是使用Generalized Advantage Estimation(GAE)计算的优势函数,h是轨迹的长度;δ_t^i是智能体i在时间步t的Temporal Difference(TD)误差,其中$V_{\varphi^i}(s_{t+1})$是状态s_{t+1}的状态价值函数. MAPPO中的critic网络训练目标为最小化损失:

$$L(\varphi)=\text{E}\left[\max\left[(V_\varphi(s_i^{(k)})-R_{\text{d}})^2,(\text{clip}(V_\varphi(s_i^{(k)}),V_{\varphi_{\text{old}}}(s_i^{(k)})-\varepsilon,V_{\varphi_{\text{old}}}(s_i^{(k)})+\varepsilon)-R_{\text{d}})^2\right]\right] \tag{26}$$

其中,R_{d}是折扣的奖励,V_φ是状态价值函数.

5 仿真实验及结果

本文实验环境为首届全国空中智能博弈大赛的仿真推演平台XSIM,该环境能够模拟多种编队及平台级行为,具有自主可控、高扩展性等优点. 实验中采用了12支不同对手策略用于交互,这些对手策略均采用最新的知识或数据驱动方法,风格各异,同时具有较强的作战能力.

5.1 风格化知识智能体(知识AI)

以XSIM仿真环境中的5v5空战对抗场景为例,通过分析五个指标值:协同能力、防守能力、打击能力、编队能力和攻击距离来有效地评估空战策略的整体强度和战术风格. 这可以说明知识AI的合理性、泛化性且具有一定的对抗强度.

通过调节知识AI的关键风格参数,可以观察到知识AI能具备一定的对抗强度,如图7所示. 将知识AI策略的五项指标与人类高级水平策略OP1~OP6进行对比,表明知识AI的合理性. 通过参数调优,知识AI能够达到与人类高级水平策略相当甚至超越的对抗强度.

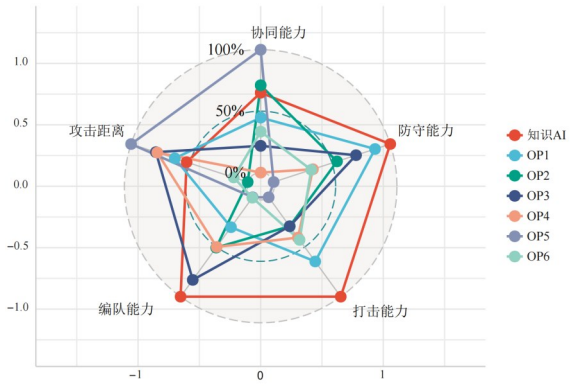


图7 队伍风格雷达图

通过对图7中5个指标进行分析,可以揭示知识AI策略和人类高级水平策略OP1~OP6的一些风格差异. 以OP1和OP2为例,可以观察到OP1在攻击距离和防守能力方面表现出色,而协同能力、打击能力和编队能力较弱,可以看出OP1的风格偏激进,更注重个体作战能力. 相比之下,OP2在各方面能力较均衡,防守能力较强但攻击能力较弱,其风格偏保守. 知识AI在各方面的能力较强而且相对均衡,实现了个体能力和团队合作间的权衡. 它在保持自身防守能力的同时,尽可能提升攻击距离和打击能力. 总体而言,知识AI的风格偏向保守.

不同队伍策略相互对抗后的胜率统计结果如图8所示. 对对抗胜率图分析可知,在空战场景中,很可能不存在一种完美的策略能够在面对所有对手时都表现出强大的能力. 相反,不同风格的队伍之间可能存在相互的克制关系. 这意味着某些策略可能在对抗某些对手时表现出色,而在面对其他对手时可能相对较弱. 这种相互克制的关系使得空战对抗变得更加复杂和多样化.

5.2 强化AI

5.2.1 强化学习奖励设计

强化AI使用MAPPO算法以多任务的形式对12个对手进行训练,并分析训练过程中奖励值和胜率之间的关系. 这一实验旨在验证任务建模、强化AI设计以及强化学习框架在解决集群空战问题上的适用性和有效性. 训练的结果如图9所示,从奖励曲线、胜率曲线可以观察到,随着训练的进行,强化AI获得的奖励值逐渐上升并趋于稳定,这验证了强化AI框架的合理性. 这意味着强

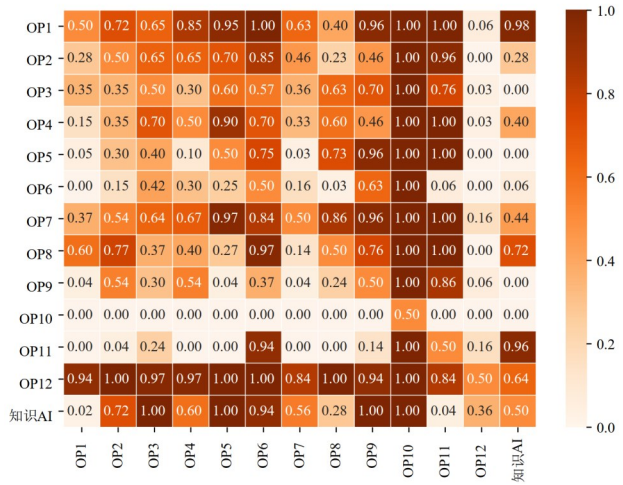


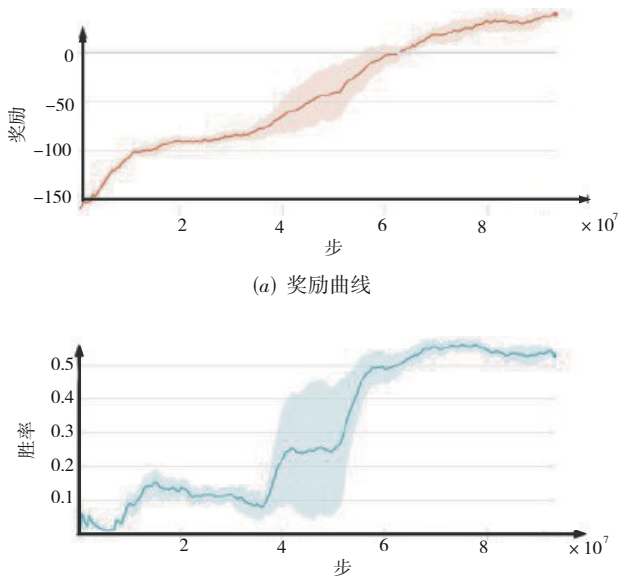
图 8　队伍对抗胜率图

化 AI 能够通过学习改进自身策略,从而获得更高的奖励值. 同时,强化 AI 的胜率随着奖励的增加而增加,这验证了奖励设计的合理性. 值得一提的是,该奖励设计无需大量的专家知识,而是通过优化的目标来引导学习过程,从而避免对智能体策略学习的误导.

5.2.2　单对手强化学习训练

为了更好地分析人类高级水平策略 OP1~OP12 的风格,首先进行单对手的强化学习训练.

图 10 展示了对不同对手分别进行强化学习训练的结果,从图中可以看出,对于能力较强的对手 OP2,奖励在 3×10^7 步时能达到 50 左右. 相比之下,对于稍弱的队伍如 OP6,能够在较短时间内训练到约 200 的奖励. 通过雷达图的分析可知,OP4 和 OP3 的能力相当,但是 OP3 存在明显漏洞,强化学习擅长于挖掘对手的策略弱点,所以最终在对抗 OP3 时能获得 100 左右的奖励,而在与 OP4 对抗时只能达到 50 左右.



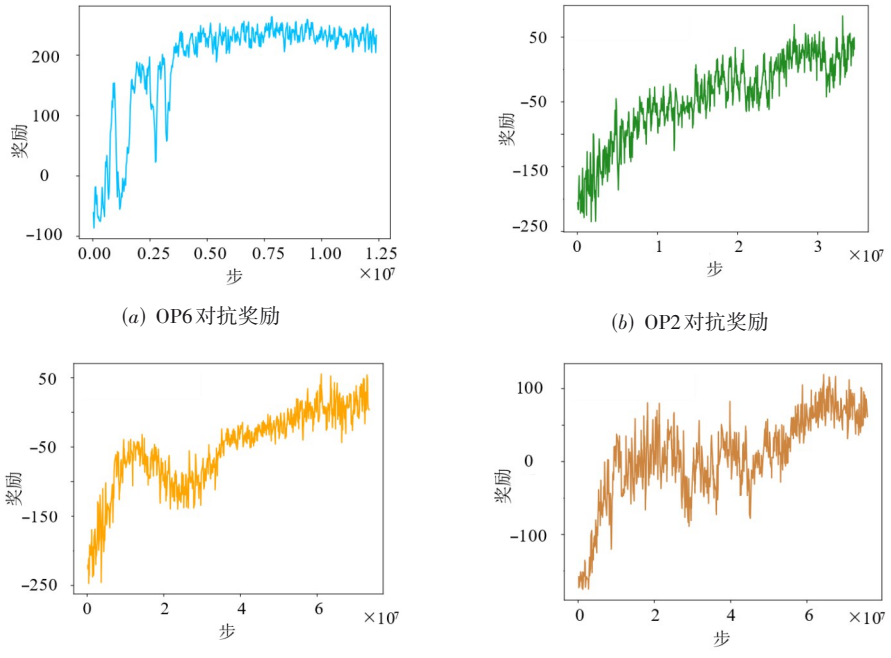
(*a*) 奖励曲线

(*b*) 胜率曲线

图 9　奖励与胜率关系图

5.2.3　多对手强化学习训练

为了验证多任务强化 AI 设计的有效性和高时间效



(*a*) OP6 对抗奖励

(*b*) OP2 对抗奖励

(*c*) OP4 对抗奖励

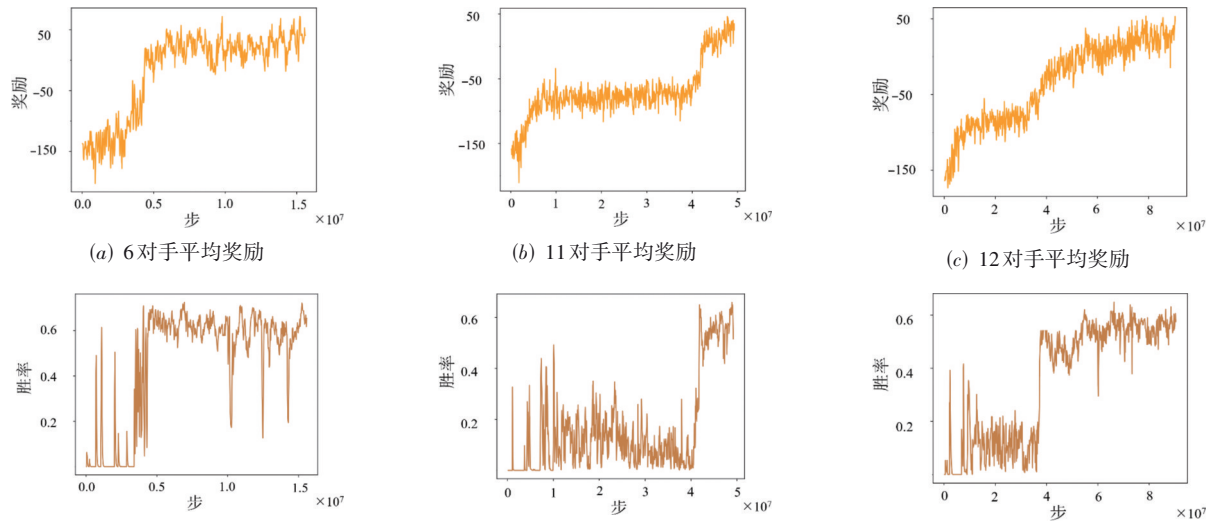
(*d*) OP3 对抗奖励

图 10　单对手强化学习训练

率,实验中训练单个策略,该策略同时与多个对手进行对抗交互学习. 在学习过程中,依据动态对手匹配机制动态调整对抗不同对手的概率. 下面对多对手训练曲线进行分析,将训练对手集合分别设置为6对手、11对手和12对手并进行训练,结果如图11所示.

从6对手的训练过程来看,在使用了约5 000场的对战数据后,强化AI的胜率就能超过70%,对于11对手和12对手在使用了约50 000场对战数据后,胜率能够达到60%以上.

根据实验结果,训练对手集合中对手的数量越多,能力越强,训练所需要的时间就越长. 与单对手强化AI的训练结果相比,多对手强化AI使用了丰富的对手池,有助于避免智能体挖掘特定对手的弱点,实现稳定的学习过程,并提升智能体策略的泛化能力. 进一步对强化AI训练结果进行分析,如表3所示,与知识AI相比,强化AI在保持策略泛化性和鲁棒性的同时,具有更高的策略强度,并且在与绝大多数对手对抗时表现更好.



(a) 6对手平均奖励 (b) 11对手平均奖励 (c) 12对手平均奖励

(d) 6对手平均胜率 (e) 11对手平均胜率 (f) 12对手平均胜率

图11 多对手强化学习训练

表3 知识AI和强化AI胜率表

队伍	OP1	OP2	OP3	OP4	OP5	OP6	OP7	OP8	OP9	OP10	OP11	OP12	平均
知识AI	0.02	0.72	1	0.6	1	0.94	0.56	0.28	1	1	0.04	0.36	0.62
强化AI	1	0.53	1	0.41	1	1	1	0.95	0.67	1	0.6	0.21	0.78

5.3 策略行为分析

在训练初期,有人/无人机采取随机动作,没有表现出任务有意义的战术行为. 由于事件与态势混合驱动的奖励设置,有人/无人机在发生关键事件,如攻击、逃逸等等动作后,逐步学会编队、攻击、拉扯等策略. 下面选取了几个多任务训练后出现的典型空战对抗策略进行分析.

在多任务场景下,有人机在对抗初期学会了一种编队策略,该策略让有人机与无人机保持一定的队形,能够避免被集中攻击导致失败. 在面对对抗强度大、攻击性强的对手策略时学会了一种拉扯策略. 在该策略中有人机迂回,采用拉扯的打法,边打边后退,我方三架无人机吸引敌方所有无人机注意,我方有人机和剩余无人机对敌方有人机形成多打少.

6 结论和未来展望

本文提出了一种知识数据协同驱动的多对手空战博弈对抗策略设计方法,该方法训练的策略与首届全国智能空中博弈大赛中的12支队伍对抗,达到70%以上的统计胜率. 实验结果表明,本文提出的任务建模方法综合考虑了空战对抗场景中的各个因素,能够有效地对空战对抗问题进行建模,辅助知识AI、离线预任务和强化AI的设计和训练. 通过调节知识AI的关键风格参数,知识AI能够达到甚至超越与人类高水平策略的对抗强度. 多对手强化学习设计能够在不降低策略强度的前提下,进一步提高智能体策略泛化性和鲁棒性. 本文提出的策略设计方法仅需修改任务建模便可迁移至任意空战场景. 例如,将该策略应用于无人机编队可以提升编队的鲁棒性和对复杂环境的适应能力. 另外,该策略还可以应用于空中交通管制,通过对空中交通

流的建模和优化,提高空中交通的安全性和效率. 研究这些应用领域将进一步展示该策略的发展前景和价值,并提供智能化和高效的解决方案.

参考文献

[1] 赵静萌, 黄宁, 朱杰, 等. 多参数关联的机载系统空战业务可靠性评估方法[J]. 电子学报, 2022, 50(9): 2060-2067.
ZHAO J M, HUANG N, ZHU J, et al. Airborne system air combat application reliability evaluation method based on multi-parameter coupling[J]. Acta Electronica Sinica, 2022, 50(9): 2060-2067. (in Chinese)

[2] 傅莉, 谢福怀, 孟光磊, 等. 基于滚动时域的无人机空战决策专家系统[J]. 北京航空航天大学学报, 2015, 41(11): 1994-1999.
FU L, XIE F H, MENG G L, et al. An UAV air-combat decision expert system based on receding horizon control[J]. Journal of Beijing University of Aeronautics and Astronautics, 2015, 41(11): 1994-1999. (in Chinese).

[3] KENNETH H. A High-Fidelity, Six-Degree-of-Freedom Batch Simulation Environment for Tactical Guidance Research and Evaluation[R]. Washington: National Aeronautics and Space Administration, Office of Management, Scientific and Technical Information Program, 1993.

[4] GOODRICH K, MCMANUS J. Development of a tactical guidance research and evaluation system (TGRES)[C]//Proceedings of the Flight Simulation Technologies Conference and Exhibit. Reston: AIAA, 1989: 350-356.

[5] 张清华, 黄志康, 高满, 等. 基于不确定性与错误分类率博弈的序贯三支决策模型[J]. 电子学报, 2022, 50(5): 1033-1041.
ZHANG Q H, HUANG Z K, GAO M, et al. Sequential three-way decision model based on the game between uncertainty and error classification rate[J]. Acta Electronica Sinica, 2022, 50(5): 1033-1041. (in Chinese)

[6] 张恒巍, 黄世锐. Markov 微分博弈模型及其在网络安全中的应用[J]. 电子学报, 2019, 47(3): 606-612.
ZHANG H W, HUANG S R. Markov differential game model and its application in network security[J]. Acta Electronica Sinica, 2019, 47(3): 606-612. (in Chinese)

[7] 孙智孝, 杨晟琦, 朴海音, 等. 未来智能空战发展综述[J]. 航空学报, 2021, 42(8): 28-42.
SUN Z X, YANG S Q, PIAO H Y, et al. A survey of air combat artificial intelligence[J]. Acta Aeronautica et Astronautica Sinica, 2021, 42(8): 28-42. (in Chinese)

[8] HAMBLING D. AI outguns a human fighter pilot[J]. New Scientist, 2020, 247(3297): 12.

[9] SUN Z X, PIAO H Y, YANG Z, et al. Multi-agent hierarchical policy gradient for air combat tactics emergence via self-play[J]. Engineering Applications of Artificial Intelligence, 2021, 98: 104112.

[10] WANG Y, ZHANG X W, ZHOU R, et al. Research on UCAV maneuvering decision method based on heuristic reinforcement learning[J]. Computational Intelligence and Neuroscience, 2022, 2022: 1477078.

[11] PIAO H Y, HAN Y, CHEN H C, et al. Complex relationship graph abstraction for autonomous air combat collaboration: A learning and expert knowledge hybrid approach [J]. Expert Systems with Applications, 2023, 215: 119285.

[12] JIANG F L, XU M Q, LI Y Q, et al. Short-range air combat maneuver decision of UAV swarm based on multi-agent transformer introducing virtual objects[J]. Engineering Applications of Artificial Intelligence, 2023, 123: 106358.

[13] 蒲志强, 易建强, 刘振, 等. 知识和数据协同驱动的群体智能决策方法研究综述[J]. 自动化学报, 2022, 48(3): 627-643.
PU Z Q, YI J Q, LIU Z, et al. Knowledge-based and data-driven integrating methodologies for collective intelligence decision making: A survey[J]. Acta Automatica Sinica, 2022, 48(3): 627-643. (in Chinese)

[14] VINYALS O, BABUSCHKIN I, CZARNECKI W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning[J]. Nature, 2019, 575(7782): 350-354.

[15] SUI Z Z, PU Z Q, YI J Q, et al. Formation control with collision avoidance through deep reinforcement learning using model-guided demonstration[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(6): 2358-2372.

[16] HE W Z, TAN J L, GUO Y F, et al. Flipit game deception strategy selection method based on deep reinforcement learning[J]. International Journal of Intelligent Systems, 2023, 2023: 5560416.

[17] FU H B, TIAN Y, YU H X, et al. Greedy when sure and conservative when uncertain about the opponents[C]//International Conference on Machine Learning. Baltimore: ICML, 2022: 6829-6848.

[18] SCHÄFER L, CHRISTIANOS F, STORKEY A, et al. Learning task embeddings for teamwork adaptation in multi-agent reinforcement learning[EB/OL]. (2023-11-20) [2024-01-27]. https://arxiv.org/abs/2207.02249v2.

[19] PAPOUDAKIS G, ALBRECHT S V. Variational autoencoders for opponent modeling in multi-agent syste-

ms[EB/OL]. (2020-01-29)[2024-01-27]. https://arxiv.org/abs/2001.10829.

[20] PAPOUDAKIS G, CHRISTIANOS F, ALBRECHT S V. Agent modelling under partial observability for deep reinforcement learning[EB/OL]. (2021-11-09) [2024-01-27]. http://arxiv.org/abs/2006.09447.

[21] RICHARDS C. Boyd's OODA loop[J]. Necesse, 2020, 5(1):142-165.

[22] LITTMAN M L. Markov games as a framework for multi-agent reinforcement learning[M]//Machine Learning Proceedings 1994. Amsterdam: Elsevier, 1994: 157-163.

[23] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[24] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of go without human knowledge[J]. Nature, 2017, 550(7676): 354-359.

[25] BERNER C, BROCKMAN G, CHAN B, et al. Dota 2 with large scale deep reinforcement learning[EB/OL]. (2019-12-13)[2024-01-27]. https://arxiv.org/abs/1912.06680.

[26] YE D H, CHEN G B, ZHANG W, et al. Towards playing full moba games with deep reinforcement learning[EB/OL]. (2020-12-31) [2024-01-27]. http://arxiv.org/abs/2011.12692.

[27] ROSS S, GORDON G J, BAGNELL J. A reduction of imitation learning and structured prediction to no-regret online learning[EB/OL]. (2011-03-16) [2024-01-27]. http://arxiv.org/abs/1011.0686v3.

[28] CHEN M, PU Z Q, PAN Y, et al. All for goals: A stylized automated analysis framework in football matches [C]//2023 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE, 2023: 1-8.

[29] YU C, VELU A, VINITSKY E, et al. The surprising effectiveness of PPO in cooperative, multi-agent games[EB/OL]. (2022-11-04) [2024-01-27]. http://arxiv.org/abs/2103.01955v4.

[30] BRODY S, ALON U, YAHAV E. How attentive are graph attention networks? [EB/OL]. (2022-01-31) [2024-01-27]. https://arxiv.org/abs/2105.14491v3.

[31] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[EB/OL]. (2018-02-04) [2024-01-27]. https://arxiv.org/abs/1710.10903v3.

[32] ZHANG C Y, VINYALS O, MUNOS R, et al. A study on overfitting in deep reinforcement learning[EB/OL]. (2018-04-20) [2024-01-27]. http://arxiv.org/abs/1804.06893v2.

[33] KURNIAWAN B, VAMPLEW P, PAPASIMEON M, et al. An empirical study of reward structures for actor-critic reinforcement learning in air combat manoeuvring simulation[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019: 54-65.

作者简介

冯锦元 男，2000 年 8 月出生，江苏泰州人. 现为中国科学院自动化研究所博士研究生. 主要研究方向为强化学习泛化性、多智能体强化学习.

E-mail: fengjinyuan2022@ia.ac.cn

陈 敏 男，1997 年 2 月出生，福建福清人. 硕士，现为中国科学院自动化研究所助理工程师. 主要研究方向为多智能体系协同决策.

E-mail: chenmin161@mails.ucas.ac.cn

李俊影 女，2000 年 6 月出生，黑龙江鹤岗人. 现为中国科学院自动化所硕士研究生. 主要研究方向为强化学习.

E-mail: 1124327661@qq.com

陈加乐 男，1998 年 8 月出生，安徽阜阳人. 毕业于南京信息工程大学软件工程专业，现为南京信息工程大学在读研究生. 主要研究方向为群体智能决策.

E-mail: 202212490361@nuist.edu.cn

蒲志强 男. 中国科学院自动化研究所研究员，博士生导师. 主要研究方向为群体智能、多智能体强化学习等.

E-mail: zhiqiang.pu@ia.ac.cn

陈敏杰 男，1978 年 12 月出生，浙江杭州人. 高工，现为北京华如科技股份有限公司副总裁. 研究方向为建模仿真与智能决策.

E-mail: chenminjie@huaru.com.cn

孙方义 男，1985 年 5 月出生，吉林桦甸人. 硕士，现为北京华如科技股份有限公司 AI 部门经理. 研究方向为深度强化学习.

E-mail: sunfangyi@huaru.com.cn