中国科学: 信息科学 2023 年 第 53 卷 第 12 期: 2493–2510
SCIENTIA SINICA Informationis
论文

舰载机保障作业人机协同决策方法

李亚飞[1,2,3], 高磊[1], 蒿宏杰[1], 靳远远[1,2,3], 王可[1,2,3], 徐明亮[1,2,3*]

1. 郑州大学计算机与人工智能学院, 郑州 450001
2. 智能集群系统教育部工程研究中心, 郑州 450001
3. 国家超级计算郑州中心, 郑州 450001
* 通信作者. E-mail: iexumingliang@zzu.edu.cn

收稿日期: 2023–01–31; 修回日期: 2023–04–10; 接受日期: 2023–05–10; 网络出版日期: 2023–12–13

国家自然科学基金重点项目 (批准号: 62036010)、国家自然科学基金面上项目 (批准号: 61972362) 和河南省自然科学基金 (批准号: 202300410378) 资助项目

摘要 舰载机保障作业是航空母舰航空保障系统的重要组成部分, 其调度效率不仅影响舰载机出动架次率, 而且严重制约航空母舰作战效能发挥. 在多舰载机保障的动态甲板作业环境下, 安全高效地为舰载机分配保障资源, 最大限度地减少舰载机因资源分配冲突产生的时间开销, 是提高舰载机保障作业调度效率的关键途经. 现有基于启发式、机器学习等方法舰载机保障作业调度策略, 存在计算量大、鲁棒性差、训练效率低等问题. 为此, 本文将舰载机保障作业调度问题建模为分布式多智能体协同控制的顺序决策问题, 构建了一种新颖的基于人机协同的多智能体作业调度决策框架 HCMTPF (human-machine collaborative multi-agent task planning framework), 有效地提高了保障作业调度决策模型的学习效率. 在此基础上, 提出了一种基于人类行为可信度的自适应作业分配方法, 进一步提高了智能体自主探索能力和人类指导经验利用率. 经大量仿真实验验证, 本文提出的舰载机保障作业人机协同决策方法比其他方法在计算性能和学习效率方面具有明显优势.

关键词 舰载机, 人机协同, 深度强化学习, 任务分配, 资源分配

1 引言

航空母舰 (简称 “航母”) 以舰载机作为主要攻防力量, 通常搭载数十甚至上百架不同类型的舰载机, 其作战效能的核心指标是舰载机出动架次率. 舰载机保障作业调度是保障舰载机在航母上安全起降和有序作业的关键, 是提升舰载机出动架次率的基础和保证 [1, 2]. 不同于陆基航空保障作业, 舰载机保障作业在不足陆基机场面积 1/10 的飞行甲板上进行, 保障过程以多类型、多架次、多波次舰载机为保障对象, 需要不同航保业务部门人员在资源与空间高度受限、作业工序繁巨且衔接紧凑等严苛条件下高效协作与配合, 保证舰载机能够在飞行甲板上安全、持续地进行起飞、着舰、调运、维修、挂弹、

引用格式: 李亚飞, 高磊, 蒿宏杰, 等. 舰载机保障作业人机协同决策方法. 中国科学: 信息科学, 2023, 53: 2493–2510, doi: 10.1360/SSI-2022-0403
Li Y F, Gao L, Hao H J, et al. Human-machine collaborative decision-making for carrier aircraft support operations (in Chinese). Sci Sin Inform, 2023, 53: 2493–2510, doi: 10.1360/SSI-2022-0403

加油、充气等保障作业. 航母飞行甲板拥有数十个保障作业站位, 其主要功能是为舰载机提供保障资源或临时停靠位 [3], 舰载机主要在这些保障站位中执行保障作业或等待保障资源. 因此, 提高舰载机保障作业调度效率的关键问题是如何高效地协调舰载机以最小移动代价和资源等待时间在各保障站位完成所需保障作业.

作业调度是目前航空保障、城市计算、军事指挥等应用领域重点研究的一类优化问题. 具体来说, 针对此类优化问题, 主要有集中式和分布式两类解决方法. 在航空保障领域, Wu 等 [4] 采用禁忌算法框架下的多重编码、工作日志表搜索等策略, 提出了一种包含资源分配及机位分配等因素的动态调度问题求解算法, 实现了舰载机甲板作业的动态调度. Meng 等 [5] 为了降低保障作业期间的不确定性风险, 提出了保障作业风险系数值和故障机会影响因子研究方案, 采用改进的禁忌算法提高了舰载机甲板保障调度稳定性. 为了进一步满足高动态作战场景下的实时保障作业调度需求, Li 等 [6] 基于深度强化学习模型实现舰载机的实时调度, 该模型收集智能体的全局状态信息用于处理决策, 有序地为每个舰载机选择保障站位. 为了提高舰载机机群甲板作业的保障效率和资源利用率, Su 等 [7] 基于边际 – 人工算法, 构建舰载机机群出动调度和保障人员配置的联合优化模型, 实现了调度决策和人员配置的集成, 从而提高了舰载机航空保障的 “减员增效”. 为了满足舰载机大规模着舰调度问题, Liu 等 [8] 基于人工蜂群算法, 引入遗传算法中的交叉算子、精英策略和自适应局部搜索策略, 解决了大规模舰载机着舰调度问题. 同时, Liu 等 [9] 将单侧辛伪谱方法与避碰策略相结合, 构建了一个舰载机架次调度模型, 实现了多机移动的协同轨迹规划. 虽然上述集中式求解方法有效解决了舰载机保障作业调度, 但是随着航母平台搭载舰载机数量的增加, 调度中心计算负载和传输数据开销急剧上升, 难以满足实时调度需求 [10].

近年来, 基于分布式多智能体方法在此类应用问题上也得到了广泛的研究. 在军事指挥领域, Liu 等 [11] 针对多战机场景, 基于数据链信息共享、协同作战中的多目标分配问题, 提出一种基于组合拍卖的协同多目标分配算法, 提高了协同多目标攻击空战决策算法的性能和稳定性. 在城市计算领域, Yu 等 [12] 为了完成城市监控或摄像等任务的数据收集, 提出了一种完全分散的多智能体近似策略优化算法和一个时空记忆增强神经网络, 使人与无人机能够在复杂场景中协同工作, 解决了无人机收集信息效率与空间信息传输保真度的平衡. 为了缓解城市交通高峰时期供需缺口, Li 等 [13] 采用分布式多智能体强化学习解决拼车中的订单调度问题, 通过智能体与环境间的相互作用, 捕获全局动态供需变化, 实时且高效地将用户拼车请求发送给合适的司机, 提高了交通管理效率.

虽然上述现有研究取得了一些成果, 但仍存不足之处, 如集中式算法执行时产生通信延时导致决策速度慢, 无法保证航母甲板的实时性, 分布式算法在训练时数据空间维度较大导致模型求解速度较慢. 因此, 为了解决大规模舰载机保障作业调度问题, 本文受演示学习和集中式训练 — 分布式执行框架 [14] 启发, 基于协作式多智能体强化学习构建了一种人机协同决策的舰载机保障作业调度模型. 该模型通过自适应人机交互机制, 在增强调度策略全局寻优能力的同时, 加快了模型训练的收敛速度. 此外, 本文通过设计可交互的舰载机保障作业仿真环境, 验证了所提算法的有效性.

本文主要贡献如下:

(1) 将舰载机保障作业调度建模为一个分布式协同控制的顺序决策问题, 并在多智能体强化学习框架下求解, 实现了一种高效的舰载机保障作业调度方法.

(2) 提出了一种基于人类行为可信度的自适应分配机制, 构建了一种人机协同决策的舰载机保障作业调度模型, 提高了模型的训练效率.

(3) 在仿真环境中进行了广泛实验, 实验结果验证了本文人机协同决策模型的有效性, 与其他方法相比训练效率也有明显的提升.

2 问题概述

舰载机离舰执行作战任务前需要完成一系列的航空保障作业, 不同保障作业间有着严苛的执行依赖顺序, 并且不同保障作业需要对应保障资源才可执行, 而这些保障资源不规则地分布在舰面的各个保障站位上. 值得注意的是, 一个保障站位可以拥有多种保障资源, 如油、电、弹药, 这也意味着一个保障站位可以执行多种保障作业. 舰载机保障作业调度以最小化单波次舰载机编队保障作业总完成时间为优化目标, 为每项保障作业分配合适保障站位的实时决策问题. 相关定义如下:

定义1 (保障站位) 一个保障站位可被表示为一个二元组 $p=(l,W)$, 其中 l 表示保障站位的位置, $W=\{w_1,w_2,\ldots,w_k\}$ 表示该站位所拥有的保障资源.

舰面上分布着若干个保障站位 $P=\{p_1,p_2,\ldots,p_n\}$, 每一个保障站位 p 都具有唯一的位置 l, 且拥有若干保障资源 W, 如一个保障站位拥有油和弹药资源, 那么舰载机既可以在该站位执行加油作业, 又可以执行挂弹作业.

定义2 (保障作业) 一个保障作业可被表示为一个二元组 $\tau=(w,d)$, 其中 w 表示保障作业所需的保障资源, d 表示作业执行时长.

定义3 (作业环境) 航母甲板作业环境定义为一个无向图模型 $G=(P,E)$, 其中 P 表示保障站位集, $p\in P$ 表示航母甲板的一个保障站位. $e_{i,j}=(p_i,p_j,c_{i,j})\in E$ 表示连接保障站位 p_i 和 p_j 的路径, 其中 $p_i,p_j\in P$, 边的权重 $c_{i,j}$ 表示保障站位之间的移动代价.

本文将舰载机保障作业调度环境建模为一个无向图, 图的节点表示保障站位, 图的边表示舰载机可在保障站位间移动, 即目标保障站位可执行该舰载机的下一保障作业, 而移动代价便是该边的权重. 那么为一个保障作业寻找最佳保障站位便可映射为在无向图中寻找一条权重最小的边.

定义4 (作业调度) 给定一组关联保障作业集合 Γ 的舰载机 F 和一组保障站位 P, 作业调度是根据舰载机 $f\in F$ 的保障作业 $\tau\in\Gamma$ 所需保障资源, 为该架舰载机分配合适的保障站位 $p\in P$ 执行保障作业. 它的目标 T_{goal} 是最小化每一波次舰载机编队保障作业的总完成时间, 可被定义为

$$T_{\text{goal}}=\min\sum_{f\in F}\sum_{\tau\in\Gamma}\sum_{p\in P}T_{\text{exec}}(f,\tau)+T_{\text{wait}}(f,p)+T_{\text{move}}(f,p),\tag{1}$$

其中 T_{exec} 表示舰载机执行保障作业所需的时间, T_{wait} 表示舰载机等待分配保障站位的时间, T_{move} 表示舰载机在舰面上移动花费的时间.

需要说明的是现有研究中舰载机多波次编队出动主要是指, 航母甲板上的舰载机每波次按最大出动能力编成一个机队执行循环出动回收, 即当前波次的舰载机从执行保障作业到完成起飞任务后, 下一个波次舰载机才开始执行降落任务, 然后在甲板站位上进行作业保障, 待保障完毕后进入起飞就绪状态 [15,16]. 因此, 两个相邻波次的舰载机交替占用甲板站位资源, 不存在站位资源共享问题. 本文旨在通过设计高效人机协同舰载机保障作业决策模型, 在优化策略模型决策质量前提下, 解决当前学习类决策方法存在训练时间开销大的问题.

3 决策模型构建

在每个决策时刻, 舰载机按保障作业集顺序选择站位并执行保障作业. 调度方案选择只依赖于保障作业和保障站位的当前状态, 与历史数据无关, 这意味着舰载机选择决策是具有马尔科夫 (Markov) 性的顺序决策. 本节将舰载机保障作业调度问题建模为部分可观测马尔科夫决策过程, 使用一个七元

组 $\langle F, S, A, \Omega, O, R, \gamma\rangle$ 表示, 其中 F 表示有限个智能体的集合, 智能体的数量用 N 表示, S 表示全局状态空间, A 表示智能体的联合动作空间, Ω 表示观测函数, O 表示智能体的局部观测值, R 表示奖励函数, $\gamma \in [0,1)$ 表示折扣因子, 用于权衡未来奖励与当前奖励占比. 在不考虑多波次编队之间的相互影响下, 同一波次舰载机编队作为一组同构智能体集合 $F = \{f_1, f_2, \ldots, f_N\}$. 在决策时间 $t \in \{1,2,\ldots,T\}$ 时, 舰载机编队 F 根据舰载机联合局部观测值 O 和航母甲板全局状态 S, 选择出适合的保障站位.

状态 $S = \{s_1, s_2, \ldots, s_t\}$, $s_t \in S$ 表示决策时间 t 时航母甲板的全局状态. 具体来说, 航母甲板的全局状态可定义为一个二元组 $s_t = (\Gamma_t, P_t)$, 其中 Γ_t 表示在时间 t 舰载机编队所需的保障作业集, P_t 表示当前时间舰载机编队停靠的保障站位.

动作 $A = A_1 \times A_2 \times \cdots \times A_N$ 表示智能体的联合动作空间, $a_t = a_t^1 \times a_t^2 \times \cdots \times a_t^n$ 被定义为决策时间 t 时舰载机编队 F 的联合动作, 其中 $a_t^n \in A_n$ 表示舰载机 f_n 从甲板站位中选择的最佳保障站位. 舰载机 f_n 有 16 个动作, 每个动作对应航母甲板的保障站位 $p \in P$. 若舰载机 f_n 已经完成所需的保障作业集合 $\Gamma_{n,t}$, 则进入待飞区等待起飞.

观测值 $O = O_1 \times O_2 \times \cdots \times O_N$ 表示智能体联合局部观测值. 在时间 t 时, 舰载机 f_n 局部观测值被定义为一个三元组 $o_t^n = (\tau_t^n, p_t^n, P_n) \in O_n$, 其中 τ_t^n 表示时间 t 舰载机 f_n 准备执行的保障作业, p_t^n 表示舰载机停靠的保障站位, P_n 表示在时间 $t+1$ 时满足舰载机 f_n 待执行作业的保障站位集. 因此, 根据舰载机的联合动作 a_t 和全局状态 s_{t+1}, 由观测函数 $\Omega(o_{t+1}|a_t, s_{t+1})$ 计算舰载机编队 $t+1$ 时间的联合局部观测值 $o_{t+1} = \{o_{t+1}^1, o_{t+1}^2, \ldots, o_{t+1}^N\}$. 注意, 本文采用集中式训练框架, 训练阶段本波次舰载机共享局部观测值 [17].

奖励函数 R 表示智能体与环境交互获得的反馈信号. 本文以最小化舰载机编队完成时间为目标, 根据目标公式 (1), 将算法奖励函数分三部分, 舰载机的移动距离奖励、舰载机选择保障站位的无冲突奖励和舰载机的优先级奖励,

$$r_t = r_t^e + r_t^\sigma + r_t^h, \tag{2}$$

其中 r_t 表示时间 t 舰载机编队 F 获得 r^e, r^σ, r^h 奖励之和. 为了鼓励舰载机就近选择站位, r_t^e 奖励被定义为舰载机编队 F 选择保障站位移动距离的奖励:

$$r_t^e = C - \sum_{f \in F} c_{i,j}, \tag{3}$$

其中 C 表示航母甲板上所有站位的距离之和, 反映了舰载机编队最大移动距离的估值, $c_{i,j}$ 移动距离采用欧氏距离代替, $c_{i,j} = \sqrt{(l_x^i - l_x^j)^2 + (l_y^i - l_y^j)^2}$, (l_x^i, l_y^i) 和 (l_x^j, l_y^j) 分别表示舰载机移动过程中起始站位和结束站位的二维坐标. r_t^σ 表示舰载机编队选择保障站位无冲突的奖励:

$$r_t^\sigma = \begin{cases} \sum\limits_{f \in F} r_{\text{coll}}, & \text{if } p_t^n \neq p_t^m, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

其中 r_{coll} 表示单架舰载机选择保障站位无冲突的奖励. 为了缩短编队整体完成作业的时间, 通过引入优先级奖励 r_t^h, 鼓励 t 时刻剩余作业时间较长的舰载机优先选择保障站位, 从而避免单架舰载机执行作业时间影响编队的整体进度. 在训练阶段, 本文设置 $C = 3.5$, $r_{\text{coll}} = 6$, $r_t^h = 5$, $r_t^e \in [0, 3.5]$.

折扣因子 γ 表示对未来收益的重视程度. 当折扣因子较小时, 舰载机更关注即时回报, 否则舰载机更注重长期回报. 舰载机保障作业的执行通常包括多个决策时间 T, 因此本文将着重考虑舰载机编队的长期累积收益.

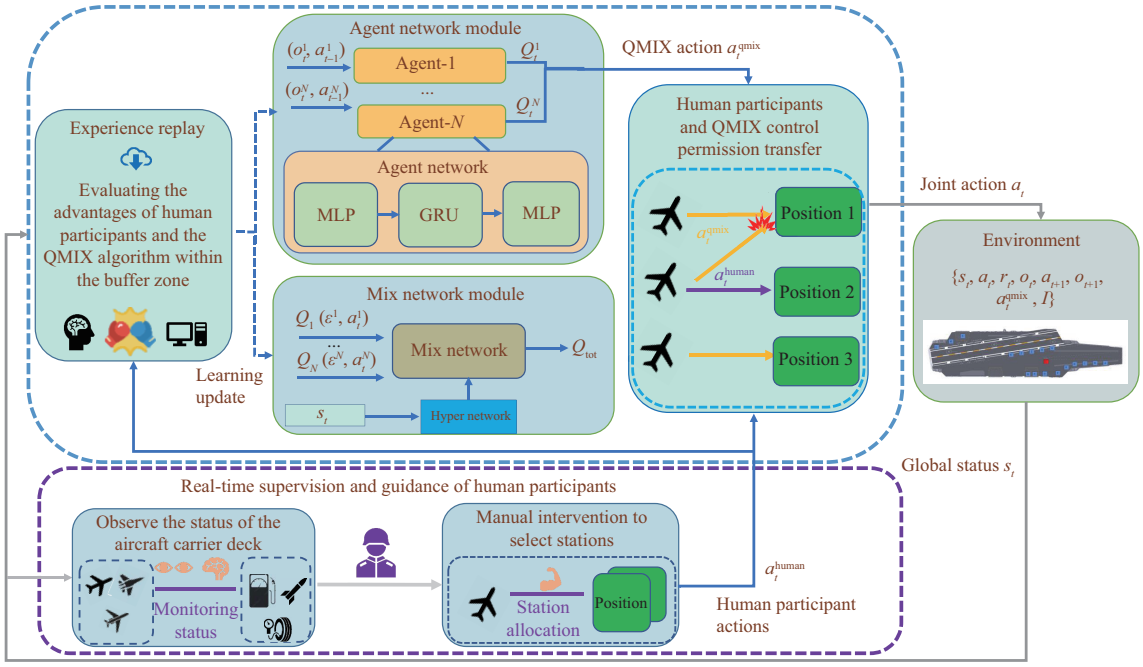


图 1 (网络版彩图) 人机协同多智能体作业规划决策框架

Figure 1 (Color online) Human-machine collaborative multi-agent task planning framework

4 方法描述

4.1 HCMTPF 决策框架

多智能体强化学习本质上是学习多智能体状态与行为之间的映射关系. 在多智能体强化学习领域, 随着智能体数量增加, 环境中状态空间维度和观测空间维度也随之呈指数型增长 [18, 19]. 这将导致智能体的探索空间较大, 在训练初期往往尝试很多无效探索, 进而影响算法的训练效率. 因此, 本文提出了一种人机协同的多智能体任务规划决策框架 HCMTPF (human-machine collaborative multi-agent task planning framework), 来解决传统多智能体求解方法效率低下的问题.

HCMTPF 主要包括智能体网络模块、混合网络模块和人机控制权限转移模块, 如图 1 所示. 其中智能体网络模块和混合网络模块是基于协作多智能体强化学习 QMIX [20] 算法, 来解决舰载机编队的作业调度问题. 人机控制权限转移模块旨在提高上述传统多智能体强化学习算法的训练效率 [21, 22]. 模型首先根据智能体网络模块对舰载机编队作出决策, 然后将人机控制权限转移模块中调整后的无冲突决策输入环境中, 得到反馈经验, 并存储在经验缓冲区, 最终通过经验回放机制取出经验以更新模型网络参数.

在智能体网络模块中, 每个智能体包含一个智能体网络. 智能体网络输入为一个二元组 $X_t^n = (o_t^n, a_{t-1}^n)$, 其中 o_t^n 和 a_{t-1}^n 分别表示时间 t 舰载机 f_n 的局部观测值和时间 $t-1$ 舰载机的站位选择. 然后, 数据经过两端的多层神经网络和中间层 GRU 门控循环单元网络 [23] 计算后, 输出时间 t 舰载机 f_n 动作效用函数, 即 $Q_t^n(\xi_t^n, a_t^n)$, 用于后续环境交互, 其中 $\xi_t^n = \{a_0^n, o_1^n, a_1^n, o_2^n, \ldots, a_{t-1}^n, o_t^n\}$ 表示舰载机 f_n 的历史动作和历史观测值 [24].

在人机控制权限转移模块中, 为了优化智能体自主决策的调度方案, 本文设计了一种智能体与人

类参与者动作决策的在线切换机制 [25]. $H(o_t)$ 表示人的策略, 人类参与者根据联合观测值 o_t 指导舰载机编队选择保障站位, $a_t^{\text{human}} \in A$. 人为干预作为一种随机事件 $I(a_t^{\text{qmix}})$, 表示人类参与者根据智能体自主探索的动作 a_t^{qmix} 来判断是否干预指导, 即人机协作决策联合动作 a_t:

$$a_t = I(a_t^{\text{qmix}}) \cdot a_t^{\text{human}} + (1 - I(a_t^{\text{qmix}})) \cdot a_t^{\text{qmix}}, \tag{5}$$

其中 a_t^{human} 表示人类参与者指导的动作, a_t^{qmix} 表示智能体网络选择的动作. 若 $I(a_t^{\text{qmix}}) = 1$ 表示舰载机选择站位冲突需要人类参与者干预指导; $I(a_t^{\text{qmix}}) = 0$ 表示舰载机选择站位无冲突或因人的注意力不集中而未能及时解决冲突. 在训练期间, 如果舰载机选择站位产生冲突未得到人工干预指导, 本文将已产生冲突的舰载机与剩余空闲站位随机匹配, 保证本轮次后续决策的正常执行. 在式 (5) 中, 当人类参与者选择干预指导时, 舰载机站位的选择将完全由人类参与者控制 [26, 27].

舰载机编队联合动作 a_t 传入环境中, 环境根据动作产生反馈信息, 并生成经验元组存入经验缓冲区中, 用于后续的策略更新. 由于人类策略和智能体策略的经验都存储在同一缓冲区中, 本文将经验元组定义为 $\zeta = (s_t, o_t, a_t, r_t, o_{t+1}, a_{t+1}, I(a_t^{\text{qmix}}))$ 来区分人类指导经验与智能体探索经验, 其中 I 用来区分人类策略与智能体策略的存储经验. 假设从经验缓冲区中取出数量为 M 的一批经验元组 $\Phi = \{\zeta_1, \zeta_2, \ldots, \zeta_M\}$, 智能体主动探索的经验元组集合为 Φ_{M1}, 人类指导经验元组集合为 $\Phi_{M2=M-M_1}$.

混合网络模块包含混合网络 [20] 和超网络 [28]. 该模块将多智能体的分散策略进行统一控制, 并加入全局状态辅助信息, 提高算法性能. 具体地, 每个舰载机的动作效用函数 $Q_t^n(\xi_t^n, a_t^n)$ 作为输入, 全局状态 s_t 作为辅助信息, 通过混合后生成联合动作 a_t 的效用函数 $Q_{\text{tot}}(\xi_t, a_t)$, 如式 (6) 所示. 注意, 本文以总奖励 r_t 作为最小化舰载机编队完成作业的时间, 最大化每个舰载机的动作效用函数 $Q_t^n(\xi_t^n, a_t^n)$.

$$\arg\max_{a_t} Q_{\text{tot}}(\xi_t, a_t) = \begin{pmatrix} \arg\max_{a_t^1} Q_t^1(\xi_t^1, a_t^1) \\ \vdots \\ \arg\max_{a_t^N} Q_t^N(\xi_t^N, a_t^N) \end{pmatrix}. \tag{6}$$

在执行决策过程中, 智能体根据部分观测值 o_t 计算出动作的局部效用函数, 根据选择站位对应的局部效用函数 $Q_t^n(\xi_t^n, a_t^n)$ 计算联合动作的效用函数 $Q_{\text{tot}}(\xi_t, a_t)$. 为了保持联合动作效用函数 $Q_{\text{tot}}(\xi_t, a_t)$ 与单个智能体动作效用函数 $Q_t^n(\xi_t^n, a_t^n)$ 的一致性, 本文对 $Q_{\text{tot}}(\xi_t, a_t)$ 和 $Q_t^n(\xi_t^n, a_t^n)$ 之间关系施加单调性约束, 如下所示:

$$\frac{\partial Q_{\text{tot}}(\xi_t, a_t)}{\partial Q_t^n(\xi_t^n, a_t^n)} > 0. \tag{7}$$

本文损失函数分为两部分, 人类参与者策略的损失函数和智能体算法策略的损失函数, 即 L^{human} 和 L^{qmix}. 从经验缓冲区中取出一批数量为 M 的经验元组 ζ 用于模型参数的更新, 则有

$$L^{\text{human}}(\theta) = \sum_{m=1}^{M_1} [(y^{\text{tot}} - Q_{\text{tot}}(\xi_t, a_t^{\text{human}}, s_t; \theta))^2], \tag{8}$$

$$L^{\text{qmix}}(\theta) = \sum_{m=1}^{M_2} [(y^{\text{tot}} - Q_{\text{tot}}(\xi_t, a_t^{\text{qmix}}, s_t; \theta))^2], \tag{9}$$

其中 θ 表示智能体网络参数, $y^{\text{tot}} = r_t + \gamma \max_{a_{t+1}} Q_{\text{tot}}(\xi_{t+1}, a_{t+1}, s_{t+1}; \theta^-)$ 表示更新参数的目标函数, θ^- 是目标网络参数 [14, 29]. 为了保证在迭代训练更新时拉近随机探索策略动作和专家演示系统动作的分布差异, 本文根据式 (8) 和 (9), 将最终损失函数定义如下:

$$L(\theta) = \frac{1}{M_1} L^{\text{qmix}} + \frac{1}{M_2} \omega L^{\text{human}} \cdot I(a_t^{\text{qmix}}), \tag{10}$$

其中 ω 是权重因子, 表示在学习过程中算法对人类指导的依赖程度. 本文为人类行为可信度设计了一种自适应分配机制如下:

$$\omega = \lambda^{\kappa} \cdot \left\{ \max_{(\xi_t, a_t, s_t) \in B} [\exp(Q_{\text{tot}}(\xi_t, a_t, s_t; \theta) - Q_{\text{tot}}(\xi_t, a_t^{\text{qmix}}, s_t; \theta)), 1] - 1 \right\}, \tag{11}$$

其中 λ 是略小于 1 的超参数, B 为经验缓存池, κ 是学习轮次. λ^{κ} 表示随着算法效果逐渐提高, 人类参与者干预动作的可信度逐渐降低. 在上述人类可信度自适应分配函数中, $\max(\cdot)$ 函数保证了算法模型只学习人类参与者 "良好" 动作的指导. $\exp(\cdot)$ 指数函数放大训练初期人类参与者指导 "良好" 动作的影响.

4.2 模型训练

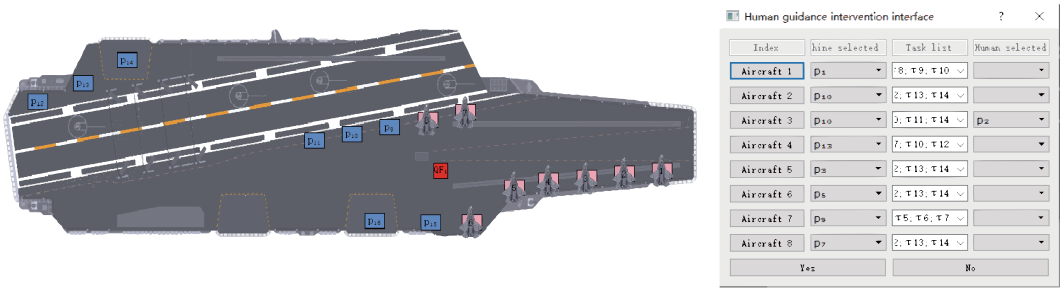
集中式训练过程如算法 1 所示. 首先, 初始化智能体网络和仿真环境的状态参数 (详见第 5 节). 智能体网络根据输入数据 X_t^n, 获得动作效用函数 $Q_t^n(\xi_t^n, a_t^n)$ (第 4~7 行). 收集舰载机编队动作 a_t^{qmix} 作为人类介入的预估动作, 根据式 (5), 判断人类参与者是否干预指导, 最终将智能体主动探索动作 a_t^{qmix} 或人类指导动作 a_t^{human} 作为联合动作 a_t, 并与环境交互获得反馈经验. HCMTPF 采用经验回放方法 [30], 将反馈信息存储在经验缓存池 B 中 (第 8~11 行). 从经验缓存池中采样 M 组经验元组, 根据式 (10), 使用 Adam 优化器 [31] 反向传播更新智能体网络和混合网络的参数 (第 13 和 14 行). 目标网络参数 θ^- 是智能体网络参数的一个副本, 每隔 d 个轮次与 θ 同步一次 (第 15~17 行), 最终输出智能体网络参数 θ_n, 用于舰载机选择站位的在线决策. 在分布式执行阶段, 模型根据各自已训练的网络利用局部观测信息选择决策序列, 直至本波次舰载机完成保障作业.

算法 1 HCMTPF training

Input: Initialize agent network parameters θ, experience buffer B, weight parameter λ, the discount factor γ, the maximum training episode number K, and the number of agents N;

1: **while** $\kappa = 1, K$ **do**
2: Observe the initial global state s and local observations o;
3: **while** $t = 1, T$ **do**
4: **while** agent $n = 1, N$ **do**
5: Input agent network X_t^n, and select RL action a_t^n;
6: Calculate $a_t^n = \arg\max Q_t^n(\xi_t^n, a_t^n)$;
7: **end while**
8: Collect the joint action $a_t^{\text{qmix}} = \{a_t^1, a_t^2, \ldots, a_t^N\}$;
9: Compute the joint action of human-machine cooperative decision by (5);
10: Calculate team reward r_t for action a_t and update environment;
11: Store transition tuple $(s_t, o_t, a_t, r_t, s_{t+1}, o_{t+1}, I(a_t^{\text{qmix}}))$ into replay buffer B;
12: **end while**
13: Sample a batch of empirical tuples ζ with the number of M from buffer B;
14: Update agent network using (10);
15: **if** κ mod d **then**
16: Update the parameters of the target network $\theta^- = \theta$;
17: **end if**
18: **end while**

Output: Agent network parameters θ_n $(n = 1, 2, \ldots, N)$.



(a) (b)

图 2 (网络版彩图) 航母舰载机保障作业仿真环境

Figure 2 (Color online) Simulation environment for carrier aircraft support operations. (a) Working environment on deck; (b) human conflict intervention interface

5 实验评估

5.1 仿真环境

本小节以俄罗斯库兹尼佐夫号航母为原型, 构建了一个典型的舰载机保障作业优化调度仿真环境, 如图 2(a) 所示. 本文主要从保障站位、保障作业和调度流程等方面介绍仿真环境. 根据真实场景舰载机保障作业调度, 仿真环境需要遵循以下约束:

次序约束. 同一架舰载机当前保障作业完成后才能开始下一项保障作业.

站位约束. 同一个保障站位内最多容纳一架舰载机停靠.

其他约束. 不考虑突发故障和其他干扰因素.

仿真环境包含 16 个保障站位和 1 个起飞站位, 其中不同数字蓝色矩形表示不同保障站位, 红色矩形表示起飞站位. 仿真环境使用欧氏距离计算保障站位之间的距离, 舰载机移动速度假定为匀速 $v = 0.2$ 单位/秒. 根据图 1 舰载机保障作业的调度流程, 该仿真环境涉及了 15 种保障作业, 具体描述如表 1 所示. 为了简化表达, 用 ALL 表示环境中全部保障站位的集合.

舰载机保障作业的调度流程可分为 3 个阶段, 部分作业之间存在依赖关系, 如图 3 所示. 图中实线表示前后作业存在依赖关系, 虚线表示作业之间无依赖关系. 第 1 阶段表示着舰后舰载机的基础检查, 第 2 阶段表示根据舰载机实时作战任务不同, 组合排列不同的保障作业序列, 第 3 阶段表示舰载机起飞前的过程.

在训练阶段, 舰载机根据作战指令获取实时的保障作业, 通过观察航母甲板的局部观测值 o_t 和全局状态 s_t 选择保障站位. 若智能体自主探索选择的保障站位导致舰载机发生冲突, 人类参与者将根据弹出的冲突警告窗口选择性对调度方案控制干预, 以解决舰载机的站位冲突, 如图 2(b) 所示. 当舰载机完成保障作业集的所有作业后, 仿真环境认定舰载机进入待飞区等待起飞, 同波次舰载机全部保障作业完成后表示一个训练周期 [24].

5.2 实验设置

为了验证 HCMTPF 的有效性, 本文设计了 3 种不同舰载机数量的仿真场景, 舰载机数量分别为 4, 8, 12 架次. 在训练阶段, 舰载机编队训练轮次为 $K = 5000$ 轮, 每批次处理数据大小为 32. 网络优化器为 Adam 算法 [31], 激活函数为 ReLU, 算法的折扣因子为 $\gamma = 0.999$, 经验缓存池 B 的大小为 5000. 仿真环境和算法使用 Python 3.7 实现, 网络模型训练使用 Pytorch 1.5, 对比实验运行在 Intel(R)

表 1 保障作业详细操作

Table 1 Details of support operations

Task	Execution time (min)	Available stations
τ_1	2	ALL
τ_2	3	ALL
τ_3	4	$\{p_1, p_4, p_5, p_6, p_8, p_9, p_{10}, p_{11}, p_{12}, p_{15}\}$
τ_4	3	$\{p_2, p_3, p_4, p_5, p_6, p_8, p_9, p_{12}, p_{15}, p_{16}\}$
τ_5	5	ALL
τ_6	1	$\{p_3, p_4, p_5, p_6, p_7, p_9, p_{11}, p_{12}, p_{13}, p_{15}\}$
τ_7	6	$\{p_1, p_3, p_4, p_5, p_8, p_9, p_{10}, p_{12}, p_{13}, p_{14}, p_{16}\}$
τ_8	4	$\{p_1, p_2, p_3, p_4, p_5, p_7, p_8, p_{10}, p_{13}, p_{14}, p_{15}, p_{16}\}$
τ_9	2	ALL
τ_{10}	4	$\{p_1, p_2, p_3, p_6, p_7, p_{10}, p_{12}, p_{13}, p_{14}, p_{15}\}$
τ_{11}	1	$\{p_1, p_2, p_3, p_5, p_6, p_9, p_{10}, p_{11}, p_{12}\}$
τ_{12}	5	$\{p_3, p_4, p_6, p_8, p_9, p_{10}, p_{11}, p_{15}\}$
τ_{13}	7	ALL
τ_{14}	8	ALL
τ_{15}	2	$\{p_1, p_2, p_4, p_6, p_7, p_9, p_{10}, p_{12}, p_{14}, p_{16}\}$

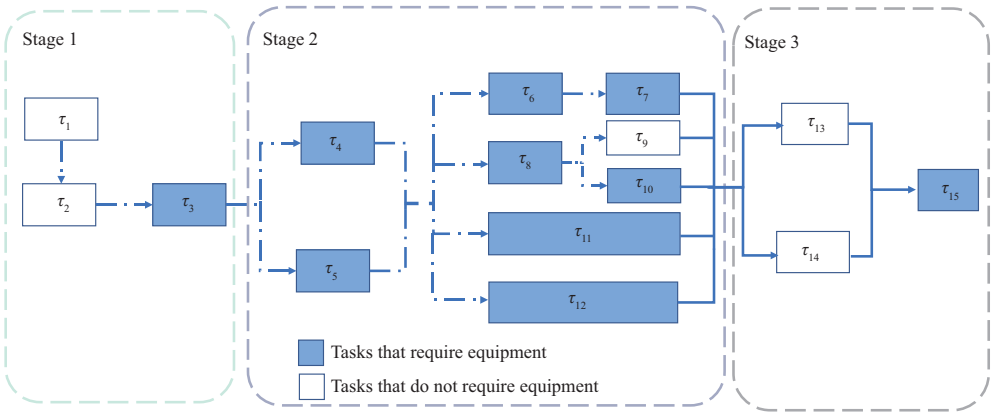


图 3 (网络版彩图) 仿真环境作业依赖关系

Figure 3 (Color online) Job dependency in the simulation environment

Core(TM) i9-11900 CPU, 64 G 运行内存, 单张 GPU (GeForce RTX 3060, 显存 12 G), Windows 10 操作系统的机器上.

在本文的人机协同机制中, 人类仅在智能体决策产生冲突并请求人类指导时参与指导, 此时人类可以选择参与或不参与指导. 为了探究人类实际指导次数对模型训练效率和性能的影响, 本文利用伯努利 (Bernoulli) 分布对人类指导智能体决策的事件进行建模, 将人类实际指导智能体决策的概率表示为 μ, 即为人类参与率; 而人类因特殊因素不指导智能体决策的概率为 $1-\mu$. 当智能体请求人类指导干预时, 本文将人类实际指导智能体决策的概率设置为 25%, 45%, 85%, 探究不同人类指导次数对模型性能和效率的影响.

5.3 对比方法和评价指标

本文将 HCMTPF 与以下算法进行对比:

• IQL [14] 方法将单智能体 Q-learning 算法扩展至多智能体强化学习环境, 算法把其余的智能体看作环境的一部分, 每个智能体将探索各自最优策略.

• VDN [24] 方法是一种价值分解的多智能体协作强化学习算法. 该算法利用整体效用值 Q_{tot} 进行反向传播, 隐式训练每个智能体网络进而提高团队的奖励.

• QMIX [20] 方法是一种价值分解的多智能体协作强化学习算法. 算法通过约束保证全局与每个智能体局部效用值函数的一致性 Q_{tot}, 使用混合网络构建出具有更强表达能力的全局 Q 效用函数.

• IGM-DA [32] 方法将值分解多智能体强化学习算法与模仿学习策略相结合, 通过将有损分解与 Bellman 迭代分离避免错误累积, 提高了协作式多智能体强化学习算法的性能.

• DQN [6] 该方法是一种单智能体强化学习算法, 通过集中协调的方式确定多架舰载机和保障站位之间的匹配.

为了评估 HCMTPF 的性能, 本文使用以下两个评价指标:

(1) 编队决策成功率. 该指标定义如下:

$$\rho = \frac{1}{K}\sum_{\kappa=1}^{K}\frac{\sum_{t=1}^{T_\kappa} J(a_t)}{T_\kappa}, \tag{12}$$

其中 T 表示舰载机编队完成作业集花费的时长. $J(a_t)$ 是分段函数, 判断舰载机之间选择站位是否产生冲突. 若有冲突 $J(a_t)=0$, 否则为 $J(a_t)=1$. ρ 越大意味着舰载机编队选择保障站位越有效, 即协同多智能体算法效果越好.

(2) 舰载机编队奖励. 该指标表示本波次舰载机从开始执行作业到完成作业所获得的奖励之和, 其具体描述如下:

$$R_{\text{total}} = \sum_{t=1}^{T}\sum_{n=1}^{N} r_t^n, \tag{13}$$

其中 N 表示舰载机的数量, T 表示本波次舰载机完成保障作业所需的决策时间, r_t^n 表示单架舰载机 n 在时间 t 与环境交互所获得的奖励. 若舰载机编队奖励越大, 则编队调度方案越好.

5.4 实验结果分析

为了更好地验证提出方法的模型训练效率和模型决策质量, 本文分别对 QMIX, VDN, DQN, HCMTPF, IQL 和 IGM-DA 算法在上述仿真环境中进行了实验, 具体实验结果如下.

5.4.1 模型训练效率评估

为了验证本文构建人机协同决策舰载机保障作业调度模型的训练效率, 本小节从这 3 个方面进行讨论: 舰载机编队奖励、编队决策成功率和人机协同机制的有效性进行对比实验.

舰载机编队奖励. 图 4 描述了舰载机编队奖励的收敛趋势, 可以发现协作多智能体强化学习 VDN, QMIX 和 HCMTPF 比 IQL 算法获得更高的平均总奖励值. 因为 IQL 在训练期间编队舰载机没有相互通信, 导致策略网络选择决策缺少协同, 每架舰载机在作出决策时只考虑了自身最大奖励, 而忽略了其他舰载机对整体环境的影响. 这种策略导致舰载机之间选择相同站位的冲突概率上升, 对算法编队奖励造成了负面影响. 相比于传统分布式 QMIX 和 VDN 算法, HCMTPF 求解速度更快, 较少的训练轮次便可以选择较优的决策效果, 模型通过融入人类指导决策, 为智能体提供了有目标的导向, 而

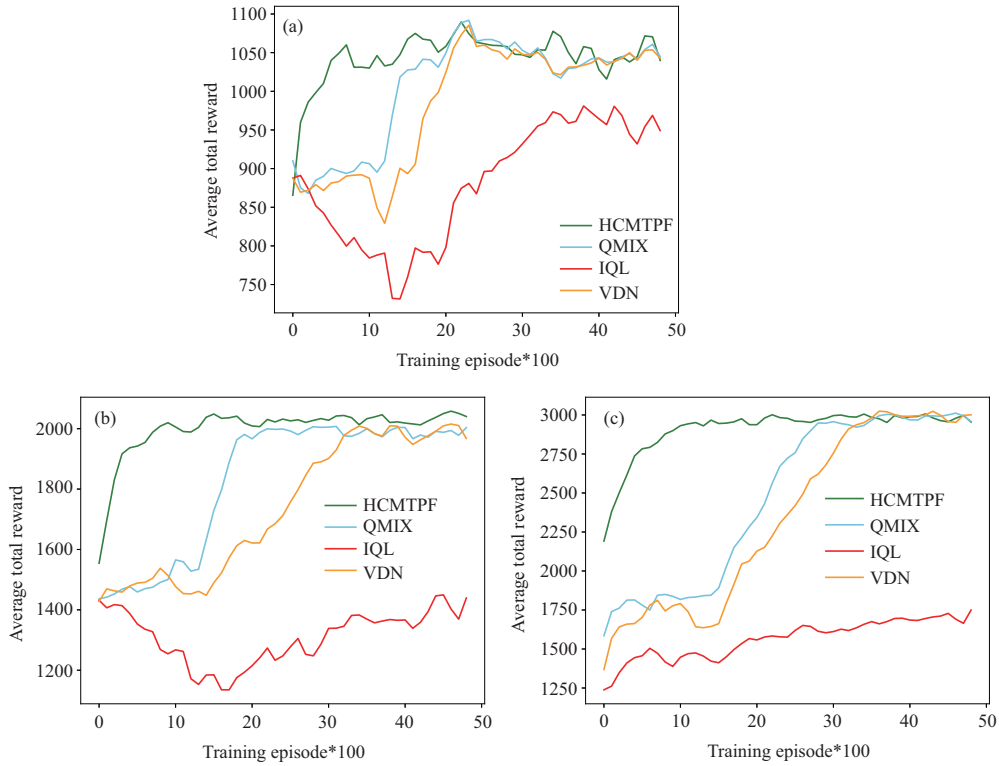


图 4 (网络版彩图) 不同算法舰载机编队奖励对比

Figure 4 (Color online) Comparison of rewards for carrier aircraft formation using different algorithms. (a) Environment of four carrier aircraft; (b) environment of eight carrier aircraft; (c) environment of twelve carrier aircraft

非盲目的探索导向, 进而利用高效的决策经验训练模型, 加快了模型的收敛速度. 同时在 4, 8, 12 架舰载机场景中, HCMTPF 平均不到 1000 轮次就达到了最佳奖励值. 相比于 QMIX 算法, HCMTPF 模型在不同场景中的训练效率分别提高了 46.7%, 135.6%, 187.4% 以上. 其原因是在舰载机保障作业问题中, 航母甲板的保障站位数量是固定的, 舰载机数量的增加会导致训练期间舰载机选择相同保障站位概率更大, 即产生更高的冲突概率, 模型学习低效的经验会降低训练效率, 即舰载机数量越多模型的训练效率提升越明显.

编队决策成功率. 图 5 展示了不同算法的编队决策成功率, 从实验结果可以观察到, HCMTPF 模型以较少训练轮次达到了较高决策成功率. 同时, 在训练后期舰载机编队整体决策成功率最高. 这种趋势主要原因在于, 人类可以依据先验知识快速对甲板上突发状况作出响应, 减少了因智能体盲目探索而产生的大量试错代价, 提高了模型学习效率. 同时, 在训练后期智能体不仅能够自主探索高质量经验, 还能从人类先验经验中学习部分较优经验, 增强了智能体自主决策能力, 提高了模型性能.

为了更进一步探究人机协同机制对多智能体强化学习算法的影响, 下面以 8 架舰载机仿真环境为例, 从有/无人机协同模型对比和不同人类干预次数实验进行分析.

有/无人机协同模型对比. 图 6 描述了算法训练时间对比, 虚线表示算法在收敛到最优奖励时模型所需的总时间, 从图中可以发现, 训练初期加入人机模块后算法训练耗时高于未加入人机模块的算法, 但模型整体的训练时间相对减少. 因为在训练初期, 智能体自主探索能力较低, 盲目决策往往会产生较多的冲突导致人类干预频繁, 从而产生额外的人类调控时间. 模型加入人类指导经验后, 智能体

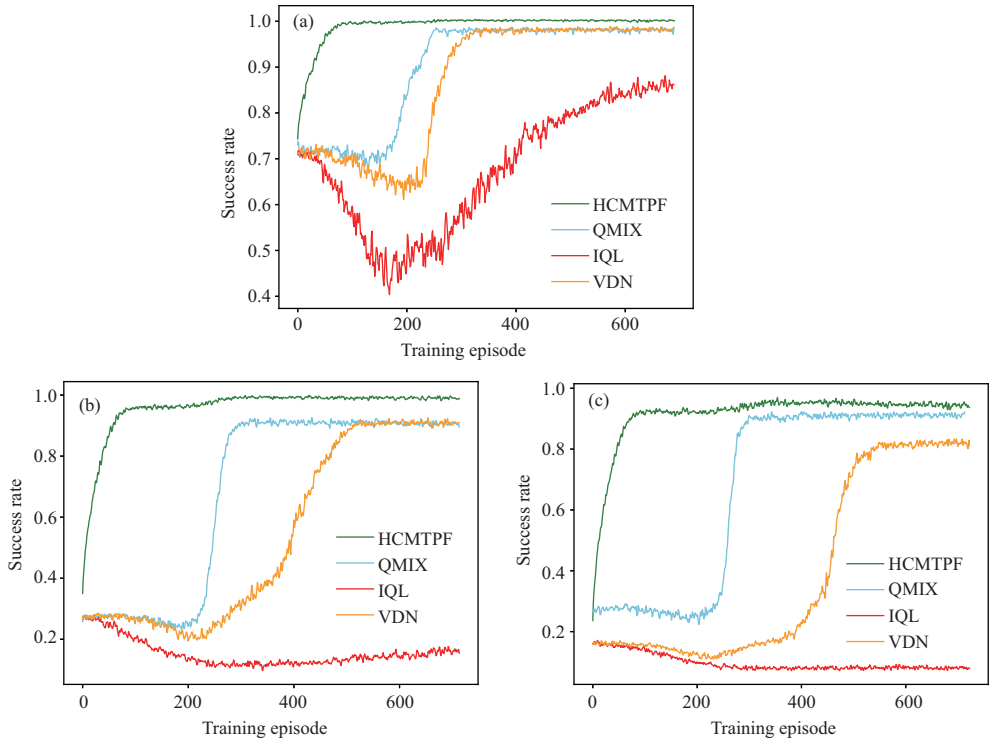


图 5 (网络版彩图) 不同算法编队决策成功率对比

Figure 5 (Color online) Comparison of formation decision success rates of different algorithms. (a) Environment of four carrier aircraft; (b) environment of eight carrier aircraft; (c) environment of twelve carrier aircraft

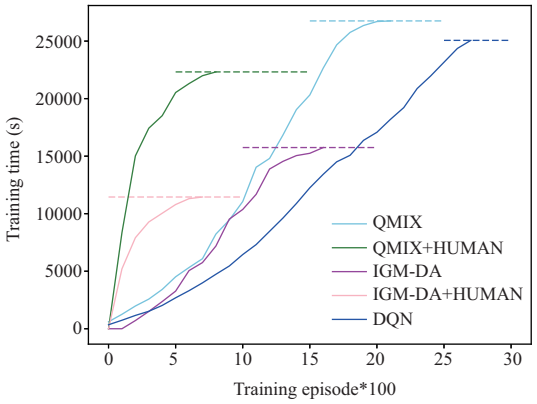


图 6 (网络版彩图) 训练时间对比

Figure 6 (Color online) Comparison of training time

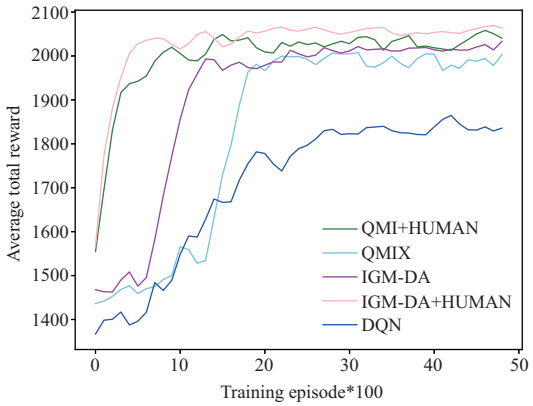


图 7 (网络版彩图) 舰载机编队奖励对比

Figure 7 (Color online) Comparison of carrier aircraft formation rewards

能够从人类经验中获取自主探索决策的导向型, 使智能体学习到高效且合理的解, 进而加快了模型的收敛速度, 减少了模型的整体训练时间, 提高了模型训练效率. 从表 2 算法训练总时间也可以看出, 加入人机模块的算法比未加入模块的算法训练总时间大幅减少, 即模型的训练效率普遍提升了 15.5% 以上, 其中图 6 和表 2 中 QMIX+HUMAN 方法即本文 HCMTPF 模型.

表 2 训练总时间

Table 2 Cumulative training time

	VDN (s)	QMIX (s)	IGM-DA (s)
–	31087	26755	15784
+HUMAN	25480	22322	11450

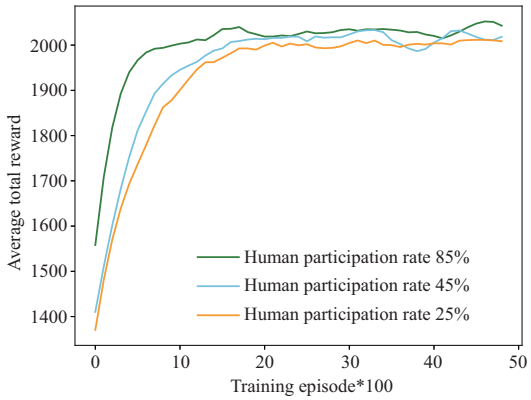


图 8 (网络版彩图) 不同人类参与率的舰载机编队奖励

Figure 8 (Color online) Carrier aircraft formation rewards with different human participation rates

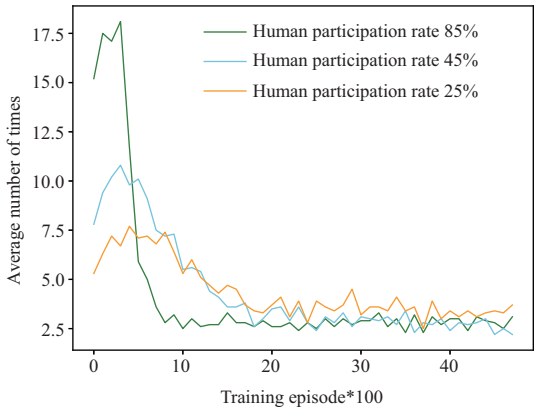


图 9 (网络版彩图) 人类平均指导次数

Figure 9 (Color online) Average number of human guides

图 7 描述了不同算法舰载机编队奖励, 结合图 6 可以看出, 相较于 QMIX 算法, IGM-DA 算法不仅训练时间更短, 而且总奖励更高; 加入人机机制后, 相比于 QMIX 算法, QMIX+HUMAN 的收敛速度和解的质量都获得了明显提升. 同时, 尽管 QMIX+HUMAN 的训练总时间大于 IGM-DA 算法, 但算法的奖励要优于 IGM-DA 算法. 这主要归因于本文提出的人类可信度自适应机制既能让模型在训练前中期提高训练效率, 又能够保障模型在训练中后期自主探索能力不被限制. 相比之下, IGM-DA 算法采用模仿学习和专家演示方法提高了训练效率, 但随着训练的增加, 专家经验的局限性限制了智能体自主探索解的能力, 导致解的质量不如 QMIX+HUMAN. 此外, DQN 算法的训练奖励低于协作式多智能体算法, 并且算法训练总时间也小于 QMIX 算法; 在加入人机机制后, QMIX+HUMAN 训练总时间小于 DQN 算法. 主要原因在于, 利用 DQN 算法训练舰载机编队决策时, 需根据预先设定顺序依次为舰载机分配最佳保障站位, 忽略了舰载机之间协同关系的变化. 尽管这种方法相比较于分布式算法处理更简单且训练速度较快, 但其决策效果较差. 而 HCMTPF 算法能够将人类指导经验融入 QMIX 算法训练过程中, 加强了智能体在训练初期的探索学习能力, 极大地提高了模型训练效率. 图 7 中 QMIX+HUMAN 方法即本文 HCMTPF 模型.

人类干预次数. 图 8 表示在不同人类参与率设定下观察舰载机编队奖励的变化情况. 在人类参与率达到 85% 时, 模型收敛速度明显优于其他两种人类参与情况. 可以发现, 随着人类参与率增加, 训练初期人类能够利用先验经验及时改善智能体的随机探索行为, 进而增强智能体自主探索的能力, 提高模型找到有效解的速度. 同时, 人类过度干预指导会限制智能体的自主探索能力, 图 9 设计了在不同人类参与率下, 采样每个训练轮次的人类平均指导次数, 来验证模型训练后期是否人为过度干预. 由图 8 和 9 可得, 在不同人类参与率下, 人工指导次数普遍随着算法训练轮次的增加而减少. 因为在决策模型训练前中期, 人类先验指导经验往往比智能体盲目探索经验更加有效 (即 $Q_{\text{human}} > Q_{\text{qmix}}$), 这个

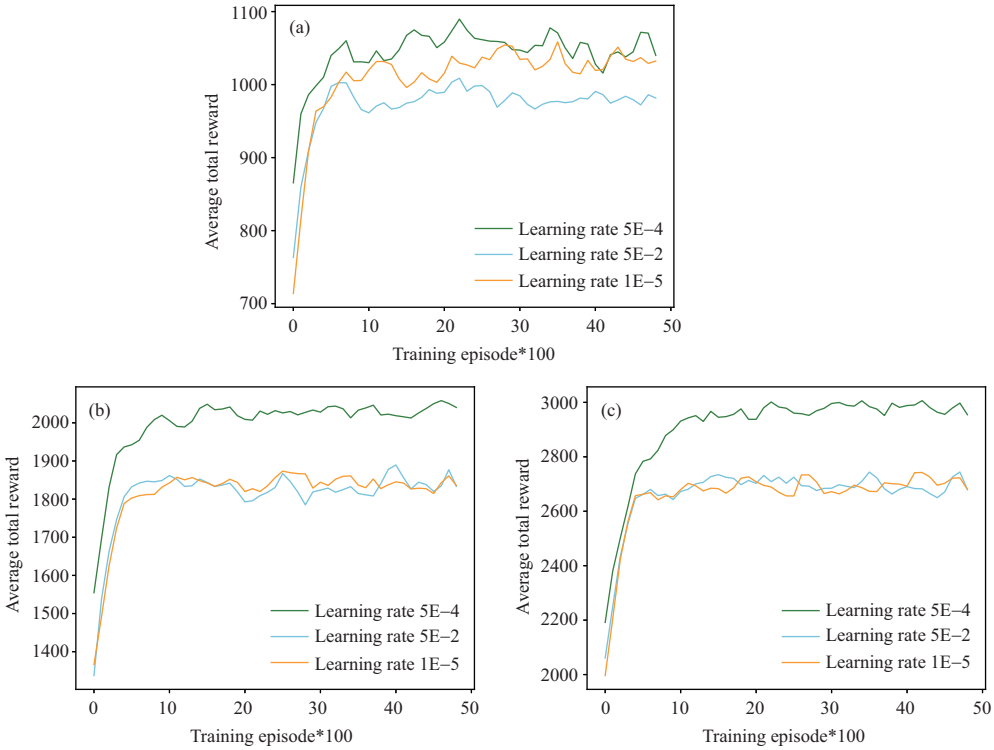


图 10 (网络版彩图) HCMTPF 不同学习率对比

Figure 10 (Color online) Comparison of different learning rates of HCMTPF. (a) Environment of four carrier aircraft; (b) environment of eight carrier aircraft; (c) environment of twelve carrier aircraft

阶段智能体策略网络更需要人类经验的指导, 进而避免探索过多的无效决策经验. 因此, 模型训练初中期人类指导会频繁干预智能体的自主探索决策. 然而, 随着决策模型训练到达中后期, 智能体策略网络已经学习到了人类有限的指导经验并额外探索了人类未知的高效决策经验 (即 $Q_{\text{human}} < Q_{\text{qmix}}$), 需要人类指导的次数随之减少, 这也是实验评估中后期人类指导频率逐渐降低的主要原因.

5.4.2 模型决策质量评估

为了探究模型的决策质量, 本小节从两个方面讨论: 网络模型参数和舰载机保障作业完成时间.

网络模型超参数. 图 10 描述了 HCMTPF 在不同学习率下舰载机编队的奖励, 从结果可以看出, 虽然网络参数学习率为 5E−4 时, 舰载机编队决策效果最好, 但不同参数的算法平均总奖励都趋近于收敛, 说明了本文建模的部分可观测马尔科夫决策过程适合舰载机保障作业应用环境.

保障作业执行时间. 图 11 描述了在训练期间舰载机保障作业完成时间的变化情况, 在 4, 8, 12 架不同舰载机场景中, 随着训练轮次的增加, 采样获得的舰载机保障作业平均完成时间逐渐减少, 即舰载机等待分配保障站位的时间和舰面移动花费的时间逐渐减少, 说明了模型能够不断地学习经验来选择最优的决策行为, 进而减少了执行决策时产生的冲突概率, 验证了模型能够有效地解决舰载机保障作业调度问题. 在实时决策阶段, 表 3 描述了不同算法舰载机保障作业平均完成时间的对比实验, 算法在加入人机协同模块后总完成时间比未加入人机协同模块算法的时间普遍减少了 0.2% 以上, 即加入人机协同模块后能够有限地提高模型决策质量. 因为在线决策时模型已融入了人类的思想, 使模型有一定的泛化能力, 能够选择更全面的高效决策, 进而提高了模型决策质量. 同时, 对于舰载机保障作

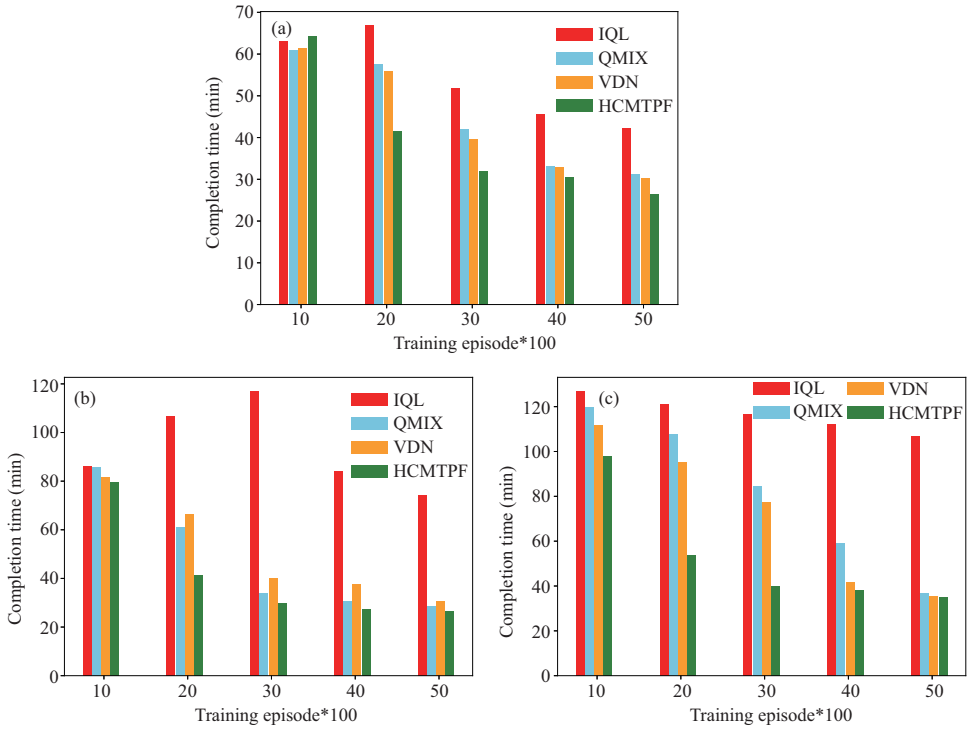


图 11 (网络版彩图) 舰载机执行保障作业完成时间

Figure 11 (Color online) Completion time comparison of different algorithms for carrier aircraft support operations. (a) Environment of four carrier aircraft; (b) environment of eight carrier aircraft; (c) environment of twelve carrier aircraft

表 3 舰载机保障作业平均完成时间

Table 3 Average completion time of carrier aircraft support operations

	VDN (min)	QMIX (min)	IGM-DA (min)
–	28.40	27.82	27.76
+HUMAN	28.12	27.49	26.72

业问题, 保障站位数量和舰载机的保障作业所需资源数量是有限的, 导致智能体的动作空间和观测信息也是有限的. 训练后期智能体策略网络已经学到了几乎所有观测信息对应的最佳决策策略, 即决策性能已经逼近上限. 其中, 表 3 中 QMIX+HUMAN 方法即本文 HCMTPF 模型.

6 总结与展望

本文以最小化完成舰载机保障作业时间为目标, 构建了一种人机协同的多智能体调度决策框架, 实现了舰载机高效协同的调度规划. 首先, 将舰载机保障作业调度问题抽象为部分可观测马尔科夫决策过程, 通过集中式训练、分布式执行训练框架, 解决了传统多智能体框架中环境不稳定导致的训练难收敛及局部最优值问题. 其次, 本文构建了一种新颖的基于人机协同的多智能体作业调度决策框架, 有效地提高了保障作业调度决策模型的学习效率. 最后, 大量仿真实验结果表明, 与其他算法相比 HCMTPF 求解速度更快. 为了验证人机协同机制在多智能体强化学习算法的有效性, 本文在实验设置中假定了人类指导机器的决策经验均为准确无误. 然而在真实复杂场景中, 人类的决策质量受到注

意力、情绪等因素的影响, 可能会导致人类的决策质量产生差异, 成为限制提升模型训练效率和性能的瓶颈. 因此, 未来将着重研究不同环境因素影响下的人机协同决策问题.

参考文献

1 Wan B, Han W, Sun X C, et al. Carrier-based aircraft departure scheduling optimization based on CE-PF algorithm. J Beihang Univ, 2022, 48: 771–785 [万兵, 韩维, 苏析超, 等. 基于 CE-PF 算法的舰载机离场调度优化问题. 北京航空航天大学学报, 2022, 48: 771–785]

2 Liu A, Liu K. Research progress on scheduling problems of carrier-based aircraft support operations. System Eng Theor Prac, 2017, 37: 49–60 [刘翱, 刘克. 舰载机保障作业调度问题研究进展. 系统工程理论与实践, 2017, 37: 49–60]

3 Wang X, Liu J, Su X C, et al. A review on carrier aircraft dispatch path planning and control on deck. Chin J Aeronautics, 2020, 33: 3039–3057

4 Wu J, Dai M Q, Wang J J, et al. Aircraft carrier based aircraft support operation scheduling based on Apprenticeship algorithm. Chin J Ship Res, 2021, 17: 1–10 [吴靳, 戴明强, 王俊杰, 等. 基于学徒制算法的航母舰载机保障作业调度. 中国舰船研究, 2021, 17: 1–10]

5 Meng Y K, Wang Z, Fan J L. Research on scheduling optimization of uncertain carrier aircraft support based on tabu algorithm. J Syst Simu, 2021, 33: 2363–2371 [孟杨凯, 王正, 范加利. 基于禁忌算法对不确定性舰载机保障的调度优化研究. 系统仿真学报, 2021, 33: 2363–2371]

6 Li Y F, Wu Q S, Xu M L, et al. Real-time scheduling for carrier-borne aircraft support operations: a reinforcement learning approach. Sci Sin Inform, 2021, 51: 247–262 [李亚飞, 吴庆顺, 徐明亮, 等. 基于强化学习的舰载机保障作业实时调度方法. 中国科学: 信息科学, 2021, 51: 247–262]

7 Su X C, Wu H, Cui R W, et al. Joint optimization method for carrier-based aircraft fleet sortie support personnel configuration and scheduling based on marginal-ABC algorithm. J Beihang Univ, 2020, 46: 2056–2068 [苏析超, 伍恒, 崔荣伟, 等. 基于边际 – 人工蜂群算法的舰载机机群出动保障人员配置 – 调度联合优化方法. 北京航空航天大学学报, 2020, 46: 2056–2068]

8 Liu Y J, Wan B, Su X C, et al. Shipborne aircraft landing scheduling based on IABC algorithm. Control Decision, 2021, 37: 1810–1818 [刘玉杰, 万兵, 苏析超, 等. 基于 IABC 算法的舰载机着舰调度. 控制与决策, 2022, 37: 1810–1818]

9 Liu J, Han W, Li J, et al. Integration design of sortie scheduling for carrier aircrafts based on hybrid flexible flowshop. IEEE Syst J, 2020, 14: 1503–1511

10 Peng K M, Lin F, Chen B M. Online schedule for autonomy of multiple unmanned aerial vehicles. Sci China Inf Sci, 2017, 60: 072203

11 Liu B, Zhang X P, Wang R, et al. Decision algorithm for cooperative multi-target attack in air combat based on combinatorial auction. Acta Aeronaut ET Astronaut Sin, 2010, 31: 1433–1444 [刘波, 张选平, 王瑞, 等. 基于组合拍卖的协同多目标攻击空战决策算法. 航空学报, 2010, 31: 1433–1444]

12 Yu W, Liu C H, Ye Y, et al. Human-drone collaborative spatial crowdsourcing by memory-augmented and distributed multi-agent deep reinforcement learning. In: Proceedings of the 38th International Conference on Data Engineering, 2022. 459–471

13 Li M, Qin Z, Jiao Y, et al. Efficient ridesharing order dispatching with mean field multi-agent reinforcement learning. In: Proceedings of the World Wide Web Conference, 2019. 983–994

14 Lowe R, Wu Y, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments. In: Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems, 2017. 6379–6390

15 Wei C Q, Chen C L, Wang B R. Study of aircraft support scheduling of wavily launching aircraft on carrier. Control Eng China, 2012, 19: 108–115 [魏昌全, 陈春良, 王保乳. 分波出动舰载机航空保障调度研究. 控制工程, 2012, 19: 108–115]

16 Han W, Liu Z X, Su X C, et al. Deck path planning algorithm of carrier-based on heuristic and optimal control. Systems Eng Electron, 2023, 45: 1098–1110 [韩维, 刘子玄, 苏析超, 等. 结合启发式与最优控制的舰载机甲板路径规划算法. 系统工程与电子技术, 2023, 45: 1098–1110]

17 Wan B, Su X C, Guo F, et al. A study for proactive robust scheduling of aircraft carrier flight deck operations with uncertain activity durations. Acta Aeronaut ET Astronaut Sin, 2019, 40: 1000–6893 [万兵, 苏析超, 郭放, 等. 不确定性工时下甲板作业的前摄性鲁棒调度. 航空学报, 2019, 40: 1000–6893]

18 Shi W, Huang H L, Cheng G Q, et al. Research on multi-aircraft cooperative air combat method based on deep reinforcement learning. Acta Automat Sin, 2021, 47: 1610–1623 [施伟, 黄红蓝, 程光权, 等. 基于深度强化学习的多机协同空战方法研究. 自动化学报, 2021, 47: 1610–1623]
19 Liu Z, Chen B, Zhou H, et al. MAPPER: multi-agent path planning with evolutionary reinforcement learning in mixed dynamic environments. In: Proceedings of International Conference on Intelligent Robots and Systems, 2020. 11748–11754
20 Tabish R, Mikayel S, Christian S, et al. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning. In: Proceedings of the 35th International Conference on Machine Learning, 2018. 4292–4301
21 Wang J, Guo B, Chen L. Human-in-the-loop machine learning: a macro-micro perspective. 2022. ArXiv:2202.10564
22 Amir O, Kamar E, Kolobov A, et al. Interactive teaching strategies for agent training. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence, 2016. 804–811
23 Chen C, Zhuo R, Ren J. Gated recurrent neural network with sentimental relations for sentiment classification. Inf Sci, 2019, 502: 268–278
24 Sunehag P, Lever G, Gruslys A, et al. Value-decomposition networks for cooperative multi-agent learning. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, 2018. 2085–2087
25 Wu J, Huang Z, Huang W, et al. Prioritized experience-based reinforcement learning with human guidance for autonomous driving. IEEE Trans Neural Netw Learn Syst, 2022. doi: 10.1109/TNNLS.2022.3177685
26 Chai C, Cao L, Li G, et al. Human-in-the-loop outlier detection. In: Proceedings of the International Conference on Management of Data, 2020. 19–33
27 Yang J, Zhao X, Fan J, et al. A human-in-the-loop approach to social behavioral targeting. In: Proceedings of the 37th International Conference on Data Engineering, 2021. 277–288
28 David H, Andrew M D, Le Q V. HyperNetworks. In: Proceedings of the 5th International Conference on Learning Representations, 2017
29 Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning. 2013. ArXiv:1312.5602
30 Wu J, Huang Z, Huang C, et al. Human-in-the-loop deep reinforcement learning with application to autonomous driving. 2021. ArXiv:2104.07246
31 Kingma D P, Ba J. Adam: a method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations, 2015. 1–13
32 Hong Y, Jin Y, Tang Y. Rethinking individual global max in cooperative multi-agent reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems, 2022

Human-machine collaborative decision-making for carrier aircraft support operations

Yafei LI[1,2,3], Lei GAO[1], Hongjie HAO[1], Yuanyuan JIN[1,2,3], Ke WANG[1,2,3] & Mingliang XU[1,2,3*]

1. *School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou* 450001*, China*;
2. *Engineering Research Center of Intelligent Swarm Systems, Ministry of Education, Zhengzhou* 450001*, China*;
3. *National Supercomputing Center in Zhengzhou, Zhengzhou* 450001*, China*
* Corresponding author. E-mail: iexumingliang@zzu.edu.cn

Abstract The carrier aircraft support operation is an essential part of an aircraft carrier's aviation support system. Its scheduling efficiency not only affects the sortie rate of carrier aircraft but also severely restricts the operational effectiveness of the aircraft carrier. In a dynamic deck operation environment with multiple carrier aircraft, the scheduling efficiency must be improved by safely and efficiently allocating support resources and minimizing the time overhead caused by conflicting resource allocation. Existing heuristics and machine learning-based scheduling strategies have shortcomings, such as large computational size, poor robustness, and low training efficiency. Therefore, in this paper, we model the scheduling problem of a carrier aircraft support operation as a sequential decision-making problem for distributed multi-agent cooperative control. Specifically, we propose a novel framework, HCMTPF (human-machine collaborative multi-agent task planning framework), which effectively improves the learning efficiency of the task planning model. Furthermore, an adaptive task assignment method based on the credibility of human decision-making is proposed, further improving the independent exploration ability of agents and the use rate of human guidance experience. Through sufficient simulation experiments, the obvious advantages of our proposed method over comparative methods are verified in computing performance, learning efficiency, and robustness.

Keywords carrier aircraft, human-machine collaboration, deep reinforcement learning, task allocation, resource allocation