中国科学 : 信息科学 2025 年 第 55 卷 第 9 期: 2371–2396
SCIENTIA SINICA Informationis
论文

基于深度强化学习的分层协同干扰资源分配方案

景小荣[1,2*], 彭喆[1], 陈前斌[2]

1. 重庆邮电大学通信与信息工程学院, 重庆 400065
2. 重庆邮电大学移动通信技术重庆市重点实验室, 重庆 400065
* 通信作者. E-mail: jingxr@cqupt.edu.cn

收稿日期: 2025–01–06; 修回日期: 2025–03–18; 接受日期: 2025–07–01; 网络出版日期: 2025–09–08

国家自然科学基金 (批准号: U23A20279) 和长江学者和创新团队发展计划 (批准号: IRT16R72) 资助项目

摘要 针对 “多对多” 通信对抗场景中的差异化动态干扰资源分配问题, 本文基于深度强化学习 (deep reinforcement learning, DRL) 提出一种分层协同干扰资源分配 (hierarchical collaborative jamming resource allocation, HCJRA) 方案. 该方案包括波束分配和功率分配两个阶段: 首先, 通过构建 DRL 辅助的目标威胁度动态评估模型, 实时评估目标威胁度, 并动态调整波束资源分配; 其次, 将多干扰设备协同干扰功率分配任务建模为分布式局部可观测马尔可夫决策过程 (decentralized partially observable Markov decision process, Dec-POMDP), 各干扰设备基于波束资源配置独立进行功率资源分配, 解决了协同干扰中的干扰收益分配问题; 为降低功率决策复杂度, 设计了 “去中心化训练、分布式决策 (decentralized training and decentralized execution, DTDE)” 框架; 最后, 提出一种基于离散柔性演员 – 评论家 (discrete soft actor-critic, DSAC) 的干扰功率资源分配策略, 运用最大熵强化学习理论在策略探索与利用之间进行自适应平衡, 并通过优化设计 Dec-POMDP 模型来增强策略的泛化性. 仿真实验结果表明, 所提出的干扰资源分配方案相比现有方案在干扰效能和扩展性上具有显著优势.

关键词 通信对抗, 目标威胁度评估, 多智能体深度强化学习, 干扰资源分配, 分布式决策, 离散柔性演员 – 评论家

1 引言

机器学习与电子战的深度融合, 推动了电子战从 “人工认知” 向 “机器认知” 的跨越式发展 [1]. 在智能化电子装备威胁日益严峻的背景下, 对抗双方围绕电磁频谱控制权的争夺愈加激烈, 而资源受限成为制约双方对抗效能发挥的主要因素 [2]. 因此, 高效合理地分配有限干扰资源, 成为提升电子战干扰效能的关键.

深度强化学习 (deep reinforcement learning, DRL) 通过结合深度学习 (deep learning, DL) 和强化学习 (reinforcement learning, RL), 赋予智能体强大的自主决策能力和对环境的抽象表征能力 [3], 具有

引用格式: 景小荣, 彭喆, 陈前斌. 基于深度强化学习的分层协同干扰资源分配方案. 中国科学: 信息科学, 2025, 55: 2371–2396, doi: 10.1360/SSI-2025-0006
Jing X-R, Peng Z, Chen Q-B. Hierarchical cooperative jamming resource allocation scheme based on deep reinforcement learning. Sci Sin Inform, 2025, 55: 2371–2396, doi: 10.1360/SSI-2025-0006

求解速度快, 支持多维度决策, 泛化能力强等优点 [4]. 基于此, 部分学者将 DRL 引入到干扰资源分配领域. 以最小化基于 DRL 的动态信道接入代理的准确性为目标, 分别基于前馈神经网络和 DRL, 作者在文献 [5] 中提出两种干扰资源分配策略. 在目标通信功率及功率控制策略未知的条件下, 文献 [6] 提出一种基于 DRL 的动态自适应干扰功率分配策略, 显著缓解了干扰效费比低的问题. 在干扰机编队突防组网雷达系统场景中, 文献 [7] 基于近端策略优化 (proximal policy optimization, PPO) 算法, 提出了一种联合路径规划与干扰功率分配的优化方案. 文献 [8] 将电子雷达对抗场景中的联合干扰类型选择和功率控制任务建模为马尔可夫 (Markov) 决策过程, 设计了决斗双深度 Q 网络 (dueling double deep q-network, D3QN) 和混合近端策略优化 (hybrid proximal policy optimization, HPPO) 两种算法, 分别处理离散动作和连续动作, 实现对干扰类型和干扰功率的联合优化选择. 然而, 上述研究均基于单智能体 DRL 框架, 主要适用于较为简单的电子对抗场景.

随着电子对抗技术的不断发展, “多对多” 集群化对抗逐渐成为电子战的主流模式, 对干扰设备提出了同时干扰多个目标的能力要求. 为应对无线传感器网中对抗跳频扩频通信的挑战, 文献 [9] 基于元-DRL, 提出一种自适应干扰资源分配策略. 文献 [10] 综合考虑个体和整体干扰效果, 研究了多波束地对空雷达干扰系统中针对多目标的资源分配问题. 文献 [11] 探讨了组网雷达系统中干扰波束与功率的联合调度问题. 在干扰机欺骗干扰功率预算受限条件下, 文献 [12] 提出一种多干扰机系统协同对抗分布式雷达系统的欺骗干扰资源优化分配算法. 然而, 这些研究主要基于传统的多目标优化理论, 对高维干扰资源分配的复杂性缺乏充分考虑, 通常面临收敛速度慢和扩展性差的问题, 难以满足认知电子对抗技术的实际需求.

针对高维资源分配难题, 已有学者借助多智能体深度强化学习 (multi-agent deep reinforcement learning, MADRL) 方法, 在多个领域深入开展资源分配的研究探索. 文献 [13] 针对无线网络中的无线下载业务, 提出一种基于分布式深度 Q 网络的干扰策略, 该策略先通过中心节点集中训练, 而后将模型参数分发给各基站; 所提策略能够在节点之间只需交互少量信息的条件下, 根据干扰环境和业务需求的特点自适应调整传输策略. 文献 [14] 针对联合频谱域和功率域的无人机簇群抗干扰问题, 提出了一种基于多智能体双深度递归 Q 学习 (double deep recurrent q-network, DDRQN) 的协同多域节能抗干扰通信决策方案, 有效降低了功率资源消耗和跳频开销. 文献 [15] 针对干扰攻击下多无人机辅助的移动边缘计算场景, 提出一种基于 MADRL 的资源管理方法, 通过动态调整无人机的 CPU 频率, 通信带宽等资源, 从而增强系统抵御干扰攻击的能力. 文献 [16] 设计了一种连续动作注意力多智能体深度确定性策略梯度 (continuous action attention multi-agent deep deterministic policy gradient, CAA-MADDPG) 算法, 通过优化无人机轨迹与干扰功率, 实现了多无人机环境下的通信安全容量最大化. 文献 [17] 针对通信对抗中的协同干扰功率分配难题, 提出了一种多智能体分布式干扰功率分配 (multi-agent distributive jamming power allocation, MADJPA) 方案, 在集中训练和分布决策的网络架构下提高了对抗环境下干扰效率. 然而, 该研究未充分考虑多目标移动特性对资源分配的影响. 文献 [18] 在对抗地面组网雷达场景中, 基于分层多智能体强化学习 (hierarchical multi-agent reinforcement learning, HMARL) 提出了一种频域协同干扰资源分配方法, 通过对全局频域干扰任务进行分解, 有效应对较大干扰动作空间和状态空间下的干扰决策问题, 但未考虑干扰距离动态变化对资源分配的影响. 文献 [19] 在多智能体框架下研究了反无人机解决方案; 进一步, 文献 [20] 针对在敏感区域上空飞行的单架或多架失控无人机, 提出了一种协作多智能体干扰方案; 该方案通过优化追踪无人机的联合机动性与功率控制策略, 以最大化对失控无人机的干扰功率, 从而有效破坏其通信链路和传感电路.

综上所述, 目前多干扰设备对抗多干扰目标的干扰资源分配研究主要集中于单一资源, 对高维干扰资源协同分配的研究尚不充分. 表 1 [5~9,16~18,20] 给出了相关研究工作的对比, 经过分析, 现有研究主要存在以下问题: (1) 上述基于 MADRL 的研究多假设对抗双方位置固定, 相关资源分配方法难以扩展到对抗双方成员数量动态变化的场景; (2) 在干扰资源分配中通常简化或忽略了多目标威胁度评估环节. 多目标威胁度评估作为电子对抗过程中指挥、控制、决策等环节的核心组成部分, 是实现干

表 1　基于 DRL 的干扰资源分配方案对比.

Table 1　Comparison of DRL-based jamming resource allocation schemes.

Scheme	Multi agent	Paradigm	Action space	Resource	Movable target	Threat assessment
[5]	✗	Actor-critic	Discrete	Frequency	✗	✗
[6]	✗	Value-based	Discrete	Power	✗	✗
[7]	✗	Policy-based	Continuous	Power, path	✓	✗
[8]	✗	Value/ policy-based	Discrete, continuous	Jamming type, power	✗	✗
[9]	✗	Actor-critic	Discrete	Frequency	✗	✗
[16]	✓	Actor-critic	Continuous	Power	✓	✗
[17]	✓	Actor-critic	Continuous	Beam, power	✗	✓
[18]	✓	Value-based	Discrete	Frequency	✗	✗
[20]	✓	Value-based	Discrete	Power, path	✓	✗

扰资源分配的重要依据 [21, 22], 因此, 为了对具有不同威胁度的目标进行差异化资源分配, 建立合理有效的威胁度评估模型对提高整体干扰效能具有重要意义.

基于上述分析, 本文以对敌方无人机编队实施压制性干扰为目标, 针对 “多对多” 复杂对抗场景中的动态干扰资源分配难题, 提出一种基于 DRL 的分层协同干扰资源分配 (hierarchical collaborative jamming resource allocation, HCJRA) 方案, 其主要创新包括如下.

(1) 针对 “多对多” 动态协同对抗场景中对差异化干扰资源分配的需求, 本文提出了一种 DRL 辅助的目标威胁度评估模型与波束资源分配策略. 该策略通过实时分析无人机目标的状态信息, 动态评估目标威胁度, 自适应调整干扰波束资源的分配方案, 并据此指导干扰功率资源的合理配置. 通过该方法, 提升了资源分配的科学性和适应性.

(2) 为实现 “多对多” 对抗场景下干扰设备间协同, 通过构建分布式局部可观测马尔可夫决策过程 (decentralized partially observable Markov decision process, Dec-POMDP), 将干扰功率分配建模为多智能体完全协作模型. 鉴于多智能体完全协作中团队奖励的稀疏性问题, 依据干扰波束分配结果将干扰功率分配问题分解为若干子任务, 进而将整体干扰收益精准分配至各干扰设备, 以加速多智能体的学习过程.

(3) 为降低干扰设备在功率决策过程中的维度复杂度, 提高实时决策能力, 设计了 “去中心化训练, 分布式决策 (decentralized training and decentralized execution, DTDE)” 的多智能体框架. 在该框架下, 各干扰设备在训练和决策时无需知晓其他干扰设备的观测信息和决策动作, 仅依靠自身观测就可完成对干扰目标的功率分配, 降低了中心化训练成本以及干扰设备间信息交互所引起的决策时延.

(4) 为增强智能体对未知环境的探索性和适应性, 基于离散柔性演员 – 评论家 (discrete soft actor-critic, DSAC) 算法提出一种干扰功率资源分配策略; 该策略应用最大熵 RL 理论, 通过自适应调整熵温度系数以平衡智能体的探索与利用; 同时, 通过优化设计智能体的观测空间和动作空间, 不仅降低了训练成本和决策维度, 还有利于灵活扩展干扰设备与无人机目标数量, 使该策略能够适应复杂多变的复杂对抗环境.

2　通信对抗模型及问题描述

2.1　通信对抗描述

考虑如图 1 所示通信对抗场景: 干扰方包括 M 台干扰设备, 对敌方预警机指挥的 N 架无人机组成的无人机编队实施压制性干扰; 令 $N_e = \{1, 2, \ldots, N\}$ 和 $M_e = \{1, 2, \ldots, M\}$ 分别表示由无人机

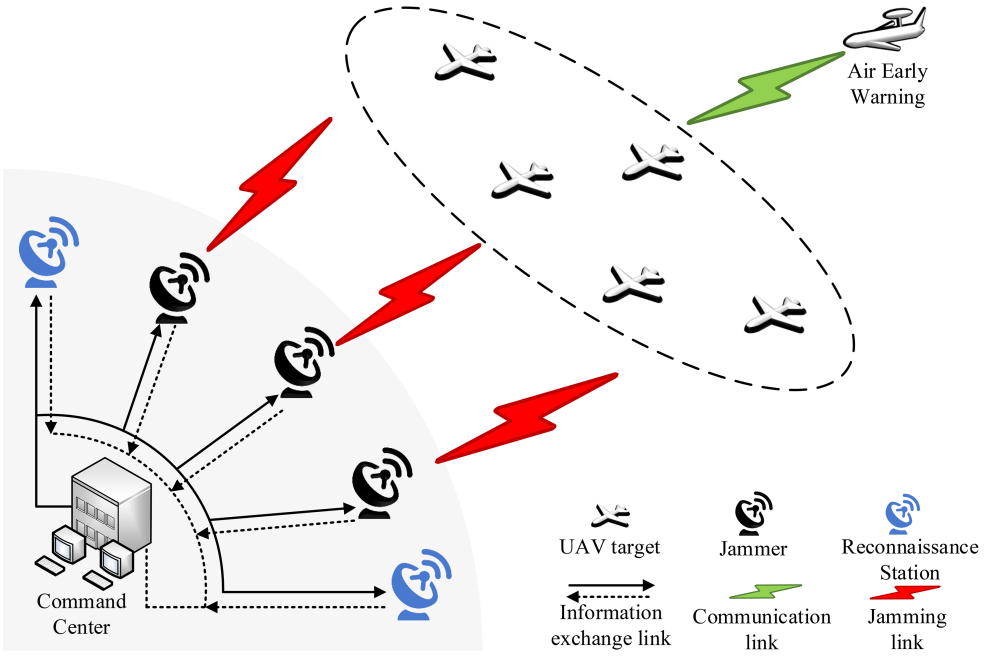

图 1　(网络版彩图) 通信对抗模型.

Figure 1　(Color online) Communication countermeasure model.

编队中的无人机和干扰设备构成的集合. 假设敌方预警机与 N 架无人机间存在 N 条互相正交的等带宽独立通信链路, 预警机通过这些链路将对干扰方阵地的持续性侦察情报和指挥信息下发至无人机编队. 干扰指挥中心通过信息交互链路收集无人机目标通信频率、带宽等情报信息, 并向干扰设备下达干扰任务指令; 在无人机编队向干扰方的飞行过程中, 干扰设备持续性地对敌方预警机与无人机编队间的所有通信链路实施压制干扰, 以期破坏敌方无人机编队与预警机间的联通.

为保证干扰效果, 干扰方应根据干扰任务合理分配有限的干扰资源. 假设每部干扰设备均能产生多个功率可控的干扰波束, 定义干扰设备 m 的波束指向向量 $\boldsymbol{b}_m = [b_{m,1}, b_{m,2}, \ldots, b_{m,N}]^{\mathrm{T}}$, 其中 $b_{m,n} \in \{0,1\}$ 表示波束指示变量, $m \in M_e$, $n \in N_e$. 当干扰设备 m 分配波束对无人机目标 n 实施干扰时, $b_{m,n} = 1$, 否则 $b_{m,n} = 0$. 符号 $[\cdot]^{\mathrm{T}}$ 表示转置运算. 基于此, 定义干扰波束分配矩阵

$$\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_M]^{\mathrm{T}}. \tag{1}$$

考虑到编队中各无人机目标在空间较为分散, 单个波束可能无法覆盖多个目标, 为此假设一个波束仅能有效干扰一个目标; 同时, 受干扰设备本身性能制约, 单部干扰设备最多可同时干扰 I 个目标, 即

$$\sum_{n=1}^{N} b_{m,n} \leqslant I \leqslant N. \tag{2}$$

为确保干扰资源得到有效利用, 限制每个无人机目标最多分配 $U \leqslant M$ 个干扰波束, 即

$$1 \leqslant \sum_{m=1}^{M} b_{m,n} \leqslant U. \tag{3}$$

根据干扰波束分配结果, 定义干扰设备 m 的功率分配向量 $\boldsymbol{p}_m = [p_{m,1}, p_{m,2}, \ldots, p_{m,N}]^{\mathrm{T}}$, 其中 $p_{m,n}$ 表示与 $b_{m,n}$ 对应的功率分配值. 为高效利用干扰功率, 仅当干扰波束照射到目标时, 干扰功率才

为非零值, 即

$$r_{m,p.t}=\begin{cases}0<p_{m,n}\leqslant P_{\max}, & b_{m,n}=1,\\ p_{m,n}=0, & b_{m,n}=0,\end{cases}\tag{4}$$

其中, $P_{\max}$ 表示干扰设备最大辐射功率. 因此, 对于干扰设备 m, 满足

$$\sum_{n=1}^{N}p_{m,n}\leqslant P_{\max}.\tag{5}$$

根据 M 个干扰设备的功率分配向量, 定义干扰功率分配矩阵

$$\boldsymbol{P}=[\boldsymbol{p}_1,\boldsymbol{p}_2,\ldots,\boldsymbol{p}_M]^{\mathrm{T}}.\tag{6}$$

当多个干扰设备协同对无人机编队实施干扰时, 可用 $\boldsymbol{B}$ 和 $\boldsymbol{P}$ 来表征干扰资源的分配情况.

2.2　问题描述

通过干扰协同使多个干扰设备辐射的干扰信号在目标处叠加, 使目标通信接收机丧失正常工作能力, 干扰效果通常用干信比 (jamming signal ratio, JSR) 来衡量. 假设无人机目标接收机采用全向天线, 同时忽略极化损失和带宽失配损耗等影响, 则在某一时刻, 无人机目标 n 接收机处的 JSR 为

$$\mathrm{JSR}_n=\frac{\sum_{m=1}^{M}p_{m,n}G_{m,n}G_{nr}L_{m,n}+p_{\mathrm{noise}}}{p_nG_{nt}G_{nr}L_{nr}},\tag{7}$$

其中 p_n 为预警机通过通信链路向无人机目标 n 发射的信号功率, G_{nt} 表示预警机向无人机目标 n 的发射天线增益, G_{nr} 为无人机目标 n 的接收天线增益, $L_{nr}=10^{L_{nr}(\mathrm{dB})/10}$ 表示预警机到无人机目标 n 的路径传播损耗, 其中 $L_{nr}(\mathrm{dB})=32.44+20\lg(f_n)+20\lg(R_{nr})$, f_n 为无人机通信频率, R_{nr} 表示预警机到无人机目标 n 的距离; $G_{m,n}$ 表示干扰设备 m 对无人机目标 n 实施干扰的发射天线增益, $L_{mn}=10^{L_{mn}(\mathrm{dB})/10}$ 表示干扰设备 m 到无人机目标 n 的路径传播损耗, 其中 $L_{mn}(\mathrm{dB})=32.44+20\lg(f_n)+20\lg(R_{m,n})$, 这里 $R_{m,n}$ 表示干扰设备 m 到无人机目标 n 的距离; p_{noise} 表示环境噪声功率 (单位: W).

当无人机目标 n 接收机处的 JSR_n 超过干扰压制系数 K_n 时, 则视干扰方对无人机目标 n 的干扰有效, 即 $\mathrm{JSR}_n\geqslant K_n$, 其中 $n\in N_e$. 从式 (7) 中可以看出, 在对无人机目标进行干扰压制, 满足空域准则和频域准则的前提下, 干扰方可通过调整干扰策略影响 JSR 的优化变量包括 $p_{m,n}$ 和 $G_{m,n}$. 考虑到干扰设备的硬件特性差异, 本文假设 $G_{m,n}$ 为固定值, 在此基础上设计波束资源和功率资源的联合分配方案. 实际上, 干扰方可采用相控阵技术实时调整波束, 确保波束主瓣持续对准高速机动无人机目标, 同时可采用目标跟踪算法和自适应波束成形技术补偿无人机目标位置误差, 从而使干扰设备天线增益的波动范围处于合理区间. 此外, 式 (7) 中影响 JSR 的变量还包括 G_{nt}, G_{nr} 等通信信号参数, 由于对抗双方属于非合作博弈, 干扰方通常无法直接获取相关参数, 但在对抗前期可通过侦察和情报分析对敌方通信信号频段、调制方式、干扰压制系数等通信参数建立干扰情报数据库, 而在对抗过程可依托侦察系统完成对敌方通信信号的实时检测、分析和识别. 基于这些技术, 干扰方可根据被干扰目标在干扰前后的态势变化实现干扰效果评估, 从而建立有效的闭环反馈机制. 因此, 本文假设干扰方可通过技术手段获知无人机目标 JSR, 并将其作为干扰效果评估指标.

在通信对抗中, 目标威胁度越大, 应越受到关注, 为此, 在分配干扰资源时, 需要考虑不同无人机目标威胁度的差异性. 定义干扰设备 m 的目标威胁度评估向量为 $\boldsymbol{w}_m=[w_{m,1},w_{m,2},\ldots,w_{m,N}]^{\mathrm{T}}$, 其中 $0\leqslant w_{m,n}\leqslant 1$; 综合所有干扰设备评估结果, 定义目标威胁度评估矩阵 $\boldsymbol{W}=[\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_M]^{\mathrm{T}}$.

为确保干扰设备的生存性, 各干扰设备通常采取分布式部署, 因而不同干扰设备对相同目标的威胁评估存在一定的差异. 为充分发挥各干扰设备的优势, 应优先由目标威胁度评估较高的设备对该目

标实施干扰, 即可依据 $\boldsymbol{W}$ 来确定干扰波束资源的分配. 为此, 定义无人机目标 n 的平均威胁度

$$\bar{w}_n = \frac{\sum_{m=1}^{M} b_{m,n} \cdot w_{m,n}}{\sum_{m=1}^{M} b_{m,n}}. \tag{8}$$

实际上, 式 (8) 亦表示对无人机目标 n 分配干扰波束所带来的收益. 如果将有限波束资源尽可能分配给威胁度评估较高的目标, 则其收益会更大. 同时, 为表征干扰功率分配的有效性, 定义 $\mathrm{sign}\left(\mathrm{JSR}_n - K_n\right)$ 为针对无人机目标 n, 实施干扰功率分配所带来的收益, 其中 $\mathrm{sign}\left(\cdot\right)$ 表示符号函数. 基于此, 定义针对无人机目标 n 分配干扰波束和干扰功率所带来的联合收益为 $\bar{w}_n \cdot \mathrm{sign}\left(\mathrm{JSR}_n - K_n\right)$. 以上干扰收益设计满足了干扰任务的实时性需求, 然而, 当敌方通信链路参数未知或缺乏高精度频谱检测手段时, 干扰方难以直接量化 JSR. 在此类非理想条件下, 无人机目标的威胁度变化量可作为干扰效能的替代指标, 这是由于无人机目标在受到干扰后可能采取飞行高度跃升、速度随机化调整等主动抗干扰策略, 导致干扰方对其威胁度评估结果发生变化. 同时, 无人机目标威胁度减少也间接验证了干扰策略的有效性.

在上述分析基础上, 考虑无人机目标威胁度的差异性, 将多干扰设备对无人机编队的干扰资源分配问题转化为最大化整体干扰收益 $\sum_{n=1}^{N} \bar{w}_n \cdot \mathrm{sign}\left(\mathrm{JSR}_n - K_n\right)$, 具体如下式所示:

$$\begin{aligned}
\mathrm{P}_1:\ & \max_{\boldsymbol{B},\boldsymbol{P}} \mathbb{F}\left(\boldsymbol{B},\boldsymbol{P}\right) = \max_{\boldsymbol{B},\boldsymbol{P}} \sum_{n=1}^{N} \bar{w}_n \cdot \mathrm{sign}\left(\mathrm{JSR}_n - K_n\right) \\
\text{s.t.}\ & \begin{cases}
C_1: b_{m,n} \in \{0,1\}, & \forall m \in M_e, \forall n \in N, \\
C_2: \begin{cases} 0 < p_{m,n} \leqslant P_{\max}, & b_{m,n} = 1, \\ p_{m,n} = 0, & b_{m,n} = 0, \end{cases} & \forall m \in M_e, \forall n \in N_e, \\
C_3: \sum\limits_{n=1}^{N} b_{m,n} \leqslant I, & I \leqslant N, \forall m \in M_e, \\
C_4: 1 \leqslant \sum\limits_{m=1}^{M} b_{m,n} \leqslant U, & U \leqslant M, \forall n \in N_e, \\
C_5: \sum\limits_{n=1}^{N} p_{m,n} \leqslant P_{\max}, & \forall m \in M_e,
\end{cases}
\end{aligned} \tag{9}$$

其中, 约束 C_1 和 C_2 表示干扰波束指示和干扰功率间的互耦限制; 约束 C_3 表示干扰设备能力约束; 约束 C_4 表示单个无人机目标最多同时被 U 个波束干扰; 约束 C_5 表示干扰功率限制.

3　分层协同干扰资源分配 (HCJRA) 方案

P_1 属于典型的带有非凸约束的组合优化问题; 由于优化变量 $b_{m,n}$ 和 $p_{m,n}$ 相互耦合, 且随对抗双方成员数量的增加, 解空间呈指数级扩张, 因而传统的组合优化方法将难以奏效. 庆幸的是, 干扰功率分配依赖于波束分配结果, 该依赖关系暗含波束指示变量与功率分配变量求解的先后顺序. 基于此, 文中提出一种 HCJRA 方案. 在该方案中, 将干扰资源分配任务细分为波束分配和功率分配两个阶段, 如图 2 所示. 在波束分配阶段, 每间隔 k 时间步, 各干扰设备将敌方目标威胁信息上传至指挥中心, 指挥中心通过信息融合评估目标威胁度, 并根据评估结果分配干扰波束, 然后将波束分配结果下发至各干扰设备; 在功率分配阶段, 干扰设备根据波束分配结果对无人机目标进行干扰功率分配; 该阶段完成后, 由各干扰设备执行具体的干扰任务. 接下来将给出波束分配和功率分配策略的具体设计过程.

3.1　干扰波束分配策略

实际应用场景中, 要对敌方目标实施有效干扰, 有限干扰波束资源应优先分配给威胁度较高的目标. 为此, 借鉴战术重要性标绘 (tactical significance map, TSM), 结合 DRL 辅助, 本小节构建了一种新的目标威胁度动态评估模型, 以此为基础对干扰波束资源进行分配.

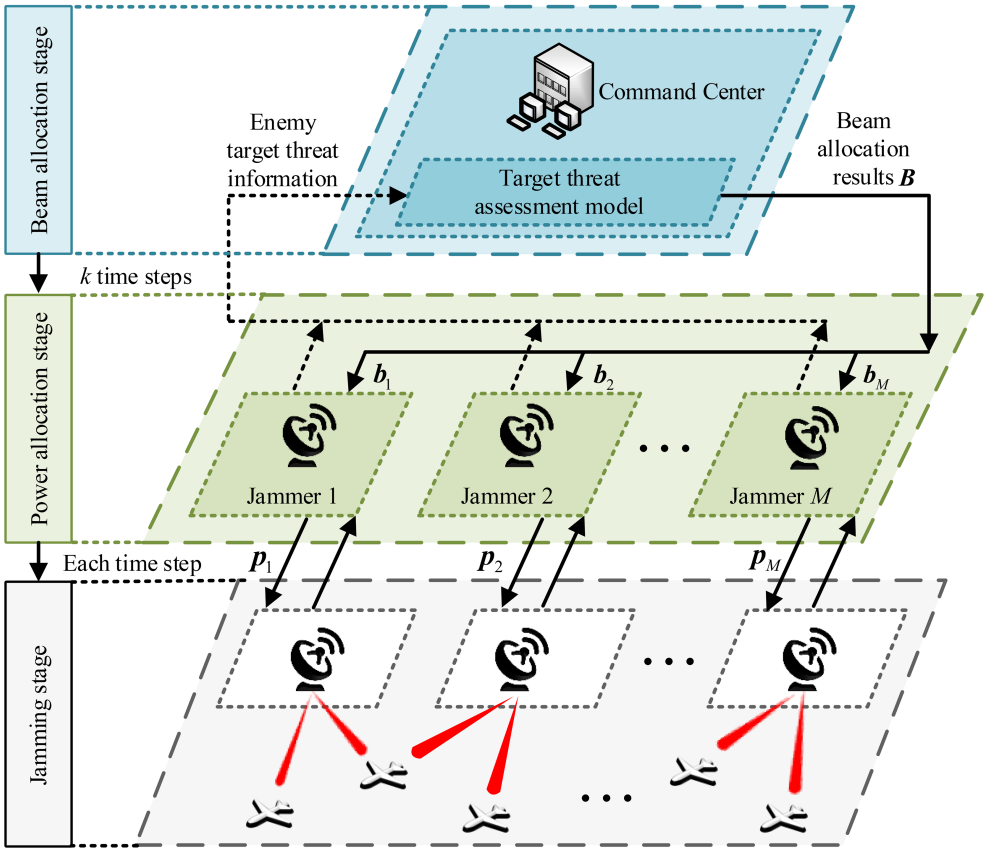

图 2　(网络版彩图) HCJRA 方案结构.

Figure 2　(Color online) HCJRA scheme structure.

3.1.1　基于 TSM 的目标威胁度评估模型

TSM 函数 [23] 作为经典的威胁度评估函数, 综合考虑了目标运动状态对威胁度的非线性影响, 而无需考虑目标本身特性 (如侦察能力、突防能力、攻击精度等), 因此, 基于 TSM 函数的目标威胁度评估模型更具普适性.

考虑通信对抗场景为 XY 二维空间, 干扰方、无人机目标和预警机处于同一坐标系. 假设当前时刻无人机目标 n 的状态向量可表示为 $\boldsymbol{h}_n = [x_n, \dot{x}_n, y_n, \dot{y}_n]^{\mathrm{T}}$, 其中 $[x_n, y_n]^{\mathrm{T}}$ 和 $[\dot{x}_n, \dot{y}_n]^{\mathrm{T}}$ 分别表示无人机目标 n 的位置矢量和速度矢量, 干扰方阵地中心坐标设为 $\boldsymbol{h}_c = [x_c, y_c]^{\mathrm{T}}$. 根据运动状态向量, 由 TSM 函数得到的无人机目标 n 对干扰方的威胁度计算表达式为 [24]

$$w_{c,n} = f_{\mathrm{TSM}}(\boldsymbol{h}_c, \boldsymbol{h}_n) \triangleq \exp\left(-\frac{(R_{c,n})^2}{2(1-\theta_{c,n}/\pi)^2(k_0 v_n + m_0)^2}\right), \tag{10}$$

其中, $R_{c,n} = \sqrt{(x_c - x_n)^2 + (y_c - y_n)^2}$ 表示无人机目标 n 与干扰方阵地中心的相对距离, $v_n = \sqrt{(\dot{x}_n)^2 + (\dot{y}_n)^2}$ 表示无人机目标 n 的绝对速度, $\theta_{c,n} = \cos^{-1}\left(\frac{(x_c - x_n)\dot{x}_n + (y_c - y_n)\dot{y}_n}{R_{c,n} \cdot v_n}\right)$ 表示无人机目标 n 的航向角, 即无人机目标 n 到干扰方阵地中心连线和无人机目标速度矢量之间的夹角, k_0 和 m_0 为预设常数.

由式 (10) 可知: $R_{c,n}$ 越小, v_n 越大, $\theta_{c,n}$ 越小, 则无人机目标 n 对干扰方的威胁度越大. 考虑到对抗环境中多部干扰设备通常采用分布式部署, 因此分别以各干扰设备作为参考点, 来构建目标威胁度评估矩阵 $\boldsymbol{W}$. 然而, 由于 TSM 函数为指数递减函数形式, 且对 $R_{c,n}$ 的变化高度敏感, 随着无人机目标与干扰设备的距离增大, 威胁度呈指数减小并趋于零. 这一特性导致在干扰设备与无人机目标距离

较远时, 威胁度评估值整体偏低, 且不同无人机目标之间的评估值差异性减弱, 从而使矩阵 $\boldsymbol{W}$ 中各行向量之间的相似度逐渐增强, 进一步加剧了资源分配的难度. 在波束资源分配阶段, 依据式 (9) 中 C_3 和 C_4 约束条件, 根据矩阵 $\boldsymbol{W}$ 中目标威胁度从高到低的顺序对无人机目标分配干扰波束. 然而, 由于矩阵 $\boldsymbol{W}$ 中的各行向量相互独立, 可能出现多个干扰设备同时对同一无人机目标赋予较高威胁度, 导致该无人机目标被多个干扰设备同时选中, 使得其波束分配数量在短时间内迅速达到上限 U; 导致后续波束分配过程中, 其他无人机目标可能无法获得任何波束资源, 从而造成干扰资源分配失衡, 降低干扰方整体干扰效率.

3.1.2　DRL 辅助的目标威胁度评估模型与波束资源分配策略

面对基于 TSM 的目标威胁度评估模型存在的问题, 本小节将提出一种 DRL 辅助的目标威胁度评估模型与干扰波束资源分配策略. 在众多 DRL 中, 演员 – 评论家 (actor-critic, AC) 模型通过 Actor 网络和 Critic 网络之间的交互, 在训练过程中能够高效地平衡探索和利用. 于是, 在该策略设计中, 将所有干扰设备针对同一无人机目标的运动状态信息作为 AC 模型中 Actor 网络的输入, 并基于 TSM 的目标威胁度评估结果来设计奖励函数; Actor 网络在处理离散动作时, 可通过 Softmax 函数输出和为 1 的概率向量, 实现了无人机目标原始运动状态到威胁度评估值的动态映射; 将 Actor 网络输出的归一化向量作为新的威胁度评估向量, 替代原始矩阵 $\boldsymbol{W}$ 中的列向量, 以重构威胁度评估矩阵. 此模型的优势在于, 其奖励函数设计借鉴了基于 TSM 函数的威胁度评估结果, 为智能体提供了有效反馈, 促使其在训练过程中学习到能够放大不同干扰设备威胁度评估值差异的策略. 因此, 即使在干扰设备和无人机目标距离较远的情形下, 威胁度归一化向量仍能保持良好的区分度, 从而缓解 TSM 函数在远距离下趋同的问题. 此外, 由于 DRL 辅助的目标威胁度评估模型以无人机目标为主导, 更倾向于选择最优干扰设备进行波束资源分配, 显著降低了多个干扰设备重复选择同一无人机目标的概率, 从而减少资源冲突, 提升了干扰资源分配的公平性和利用效率.

基于上述思路, 将目标威胁度评估矩阵的重构问题建模为马尔可夫决策过程 (Markov decision process, MDP), 对应的状态空间、动作空间和奖励函数分别定义如下.

(1) 状态空间. t 时刻无人机目标 n 针对 M 个干扰设备的运动状态信息为 $s_{n,b,t}=[o_{1,b,t},o_{2,b,t},\dots,o_{M,b,t}]$, 其中 $o_{m,b.t}=[R_{m,n,t},v_{n,t},\theta_{m,n,t}]$, 包含 t 时刻无人机目标 n 到干扰设备 m 的距离、速度和航向角信息. 基于 $s_{n,b,t}$, 定义 MDP 的状态

$$s_{b,t}=[s_{1,b,t},s_{2,b,t},\dots,s_{N,b,t}]\in\boldsymbol{S}_b. \tag{11}$$

(2) 动作空间. 动作 $a_{n,b,t}\in M_e$ 表示从 M 个干扰设备中选择其一对无人机目标 n 实施干扰, 动作空间为 M_e.

(3) 奖励函数. 从基于 TSM 的目标威胁度评估矩阵 $\boldsymbol{W}$ 中, 选择第 n 列, 即 $\boldsymbol{w}_n=[w_{1,n},w_{2,n},\dots,w_{M,n}]^{\mathrm{T}}$, 设计奖励函数

$$r_{n,b,t}=\frac{\mathrm{idx}\left(\boldsymbol{w}_n\left[a_{n,b,t}\right],\mathrm{sorted}\left(\boldsymbol{w}_n\right)\right)-1}{M-1}, \tag{12}$$

其中, $\boldsymbol{w}_n\left[a_{n,b,t}\right]$ 表示从 $\boldsymbol{w}_n$ 中获取动作 $a_{n,b,t}$ 对应的干扰设备 m 针对无人机目标 n 的目标威胁度评估值 $w_{m,n}$, $\mathrm{sorted}\left(\boldsymbol{w}_n\right)$ 表示对向量 $\boldsymbol{w}_n$ 中的元素从小到大进行排序, $\mathrm{idx}\left(\boldsymbol{w}_n\left[a_{n,b,t}\right],\mathrm{sorted}\left(\boldsymbol{w}_n\right)\right)$ 用于获取 $w_{m,n}$ 在排序后的向量 $\boldsymbol{w}_n$ 中的位置索引 (假设位置次序用 $(1,2,\dots,M)$ 表示). 该式表明, 当选取干扰设备对无人机目标 n 实施干扰时, 若选择的干扰设备对该目标的威胁度评估值 $w_{m,n}$ 越大, 则获得的奖励值越高.

在与环境交互过程中, 首先将无人机目标 n 相对于所有干扰设备的运动状态信息 $s_{n,b,t}$ 作为 Actor 网络的输入, 通过对网络输出的概率向量采样以决策动作 $a_{n,b,t}$, 并根据式 (12) 确定相应的奖励值 $r_{n,b,t}$, 随后将与之对应的经验样本 $(s_{n,b,t},a_{n,b,t},r_{n,b,t},s_{n,b,t+1})$ 存储到经验回放池 $\mathcal{D}_b$ 中. 当经验回放池中样

本数量累计达到预设阈值 Φ_b 时, 随机地从经验回放池中抽取批量样本进行训练; AC 模型中的 Critic 网络对所选动作进行策略评估, Actor 网络则依据评估结果进行策略改进, 同时更新 Actor 和 Critic 网络的参数. 随着训练的深入, 智能体逐渐倾向于选择能够带来更高奖励的最优动作. 同时, 随着最优动作被更频繁的选择, Actor 输出层概率向量的差距将逐渐扩大, 最优动作对应的概率值也会显著升高, 导致不同干扰设备针对同一目标的威胁度评估的差异性进一步得到放大. 训练结束后, 保存 AC 网络模型, 即可执行干扰波束分配.

理想情况下, 各无人机目标应由对其威胁度评估值最高的干扰设备对其实施干扰. 然而, 在实际对抗场景中, 受干扰波束资源的限制, 应对波束分配过程进行优化设计, 以确保在满足资源约束的前提下, 实现更为有效的波束资源分配. 下面结合上述分析, 给出具体波束分配过程.

在 t 时刻, 将 $s_{n,b,t}=[o_{1,b,t},o_{2,b,t},\dots,o_{M,b,t}]$ 输入 Actor 网络, 最后一层全连接层输出向量 $\hat{\boldsymbol{\xi}}_n=[\hat{\xi}_{1,n},\hat{\xi}_{2,n},\dots,\hat{\xi}_{M,n}]^{\mathrm{T}}$, 之后经过 Softmax 函数得到归一化概率向量 $\boldsymbol{\xi}_n=[\xi_{1,n},\xi_{2,n},\dots,\xi_{M,n}]^{\mathrm{T}}$, 其中 $\xi_{m,n}$ 可表示为

$$\xi_{m,n}=\frac{\exp(\hat{\xi}_{m,n})}{\sum_{m=1}^{M}\exp(\hat{\xi}_{m,n})}. \tag{13}$$

$\boldsymbol{\xi}_n$ 即为针对无人机目标 n 重构后的威胁度评估向量; 接着将与 N 个无人机目标对应的威胁度评估向量进行合并, 重构目标威胁度评估矩阵 $\boldsymbol{W}=[\boldsymbol{\xi}_1,\boldsymbol{\xi}_2,\dots,\boldsymbol{\xi}_N]$; 最后, 依据 $\boldsymbol{W}$, 分两阶段进行波束资源分配.

第 1 阶段. 循环遍历 $\boldsymbol{W}$, 按照降序依次选取 $\boldsymbol{W}$ 中威胁度评估值最大的位置并将其置 0, 同时将干扰波束分配矩阵 $\boldsymbol{B}$ 中相应位置的值映射为 1, 映射过程中需同时满足式 (9) 中约束 C_3 和 C_4, 直到每个无人机目标均被分配一个干扰波束, 则该阶段结束.

第 2 阶段. 以第 1 阶段结果为基础, 重复第 1 阶段的循环遍历映射过程, 保证每个干扰设备所分配的无人机数量达到其最多可同时干扰的无人机目标数, 则波束资源分配过程结束.

综上分析, DRL 辅助的目标威胁度评估模型与波束资源分配策略的具体流程如算法 1 所示.

3.2　功率资源分配策略

干扰波束分配结果通过指挥中心下发至各干扰设备后, 各设备根据分配结果自主决策其干扰功率分配. 鉴于真实场景中各干扰设备难以获得完备情报, 文中将多干扰设备协同干扰功率分配任务建模为 Dec-POMDP. 在该模型中, 每部干扰设备被视为独立的智能体; 基于此, 设计了基于 DTDE 的 MADRL 框架. 在此框架下, 提出基于 DSAC 的干扰功率分配策略. 以下将对干扰功率资源分配过程进行详细阐述.

3.2.1　分布式局部可观察马尔可夫决策过程 (Dec-POMDP)

Dec-POMDP 可用七元组 $\langle M_e,\boldsymbol{S}_p,\{A_m\},T,\{\boldsymbol{O}_m\},Z,r\rangle$ 表示, 其中 M_e 即为智能体集合; $\boldsymbol{S}_p$ 为状态空间; A_m 为智能体 m 的动作空间, $\boldsymbol{A}=A_1\times A_2\times\cdots\times A_M$ 为联合动作空间, $\times$ 表示笛卡尔积操作符; 状态转移函数 $T:\boldsymbol{S}_p\times\boldsymbol{A}\times\boldsymbol{S}_p\to[0,1]$; t 时刻智能体执行联合动作 $\boldsymbol{a}_{p,t}$, 根据 $T(s_{p,t},\boldsymbol{a}_{p,t},s_{p,t+1})=\Pr(s_{p,t+1}|s_{p,t},\boldsymbol{a}_{p,t})$, 状态由 $s_{p,t}\in\boldsymbol{S}_p$ 转移到 $s_{p,t+1}\in\boldsymbol{S}_p$; $\boldsymbol{O}_m$ 为智能体 m 的观测空间, $\boldsymbol{O}=\boldsymbol{O}_1\times\boldsymbol{O}_2\times\cdots\times\boldsymbol{O}_M$ 为联合观测空间; 观测函数定义为 $Z:\boldsymbol{O}\times\boldsymbol{A}\times\boldsymbol{S}_p\to[0,1]$, 表示执行联合动作 $\boldsymbol{a}_{p,t}$ 和状态转移后, 获得联合观测 $o_{p,t+1}$ 的概率, 即 $Z(o_{p,t+1},\boldsymbol{a}_{p,t},s_{p,t+1})=\Pr(o_{p,t+1}|\boldsymbol{a}_{p,t},s_{p,t+1})$; r 表示全局奖励函数, 定义为 $r:\boldsymbol{S}_p\times\boldsymbol{A}\to\mathbb{R}$, $r(s_{p,t},\boldsymbol{a}_{p,t})$ 表示了在状态 $s_{p,t}$ 执行联合动作 $\boldsymbol{a}_{p,t}$ 后获得的即时奖励.

在多干扰设备协同实施干扰任务中, 当干扰波束确定后, 干扰设备需要通过探索找到最优的联合功率分配策略 $\pi=\langle\pi_1,\dots,\pi_M\rangle$ 来最大化累计折扣奖励的期望值 $\mathbb{E}_{(s_{p,t},\boldsymbol{a}_{p,t})\sim\rho_\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_{p,t},\boldsymbol{a}_{p,t})\right]$,

Algorithm 1 DRL-assisted target threat assessment model and beam resource allocation strategy.

Input: The number of jammers: M; the number of UAV targets: N; the number of training episodes: max_episode; interaction time steps per episode: max_t; parameters of the Actor and Critic networks; experience replay pool: $\mathcal{D}_b$; the threshold at which the model starts training: Φ_b.

Initialize parameters of Actor and Critic networks and experience replay pool $\mathcal{D}_b$;
for episode = 1 to max_episode **do**
  Randomly initialize the countermeasure environment;
  for $t = 1$ to max_t **do**
    Construct target threat assessment matrix $\boldsymbol{W}$ based on TSM;
    for $n = 1$ to N **do**
      The Actor network inputs $s_{n,b,t}$, outputs action $a_{n,b,t}$, and obtains reward $r_{n,b,t}$; the experience $\left(s_{n,b,t}, a_{n,b,t}, r_{n,b,t}, s_{n,b,t+1}\right)$ is stored in the experience replay pool $\mathcal{D}_b$;
    end for
    if the number of experiences in $\mathcal{D}_b$ is greater than Φ_b **then**
      Randomly sample batch experiences, the Critic network and the Actor network perform strategy evaluation and strategy improvement respectively, and update the network parameters;
    end if
  end for
end for
Saving the Actor-Critic model;
for $n = 1$ to N **do**
  Actor network of agent inputs $s_{n,b,t}$ and saves the output probability vector ξ_n;
end for
Construct the target threat assessment matrix $\boldsymbol{W} = [\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_N]$, and let $\bar{\boldsymbol{W}} = \boldsymbol{W}$;
while the jamming beams do not cover all targets **do**
  Find the maximum value index $[x_k, y_k] = \arg\max \bar{\boldsymbol{W}}$;
  Let $\boldsymbol{B}[x_k, y_k] = 1$, $\bar{\boldsymbol{W}}[x_k, y_k] = 0$;
  if target y_k has already been assigned a beam or constraint C_3 is not satisfied **then**
    Let $\boldsymbol{B}[x_k, y_k] = 0$;
  end if
end while
According to the assigned beam result, the corresponding threat assessment value in $\boldsymbol{W}$ is set to 0: $\boldsymbol{W} = \boldsymbol{W} * (\boldsymbol{1} - \boldsymbol{B})$;
while $\boldsymbol{W}$ is a non-zero matrix **do**
  Find the maximum value index $[x_k, y_k] = \arg\max \boldsymbol{W}$;
  Let $\boldsymbol{B}[x_k, y_k] = 1$, $\boldsymbol{W}[x_k, y_k] = 0$;
  if C_3, C_4 is not satisfied **then**
    Let $\boldsymbol{B}[x_k, y_k] = 0$;
  end if
  if each jammer is assigned I beams **then**
    Break;
  end if
end while

Output: Beam allocator and beam resource allocation matrix $\boldsymbol{B}$.

即求解

$$\pi* = \underset{\pi=\langle\pi_1,\ldots,\pi_M\rangle}{\arg\max}\ \mathbb{E}_{(s_{p,t},\boldsymbol{a}_{p,t})\sim\rho_\pi}\left[\sum_{t=0}^{\infty}\gamma^t r\left(s_{p,t},\boldsymbol{a}_{p,t}\right)\right]. \tag{14}$$

Dec-POMDP 模型中, 包括如下定义.

(1) 观测空间与状态空间. 在 t 时刻, 干扰设备 m 的局部观测信息为 $o_{m,p,t} = [o_{m,1,p,t}, o_{m,2,p,t}, \ldots, o_{m,N,p,t}] \in \boldsymbol{O}_p$, 包含针对 N 个无人机目标的观测信息; $o_{m,n,p,t} = [m, n, R_{m,n,t}, R_{nr,t}, f_{n,t}, p_{m,n,t-1}, \mathrm{JSR}_{n,t}]$ 代表干扰设备 m 获取的关于无人机目标 n 的观察信息, $R_{m,n,t}$ 和 $R_{nr,t}$ 分别代表干扰设备 m

和预警机与无人机目标 n 之间的距离, $f_{n,t}$ 为通信频率, $p_{m,n,t-1}$ 为前一时刻干扰设备 m 对干扰无人机目标 n 的干扰功率, $\mathrm{JSR}_{n,t}$ 为 t 时刻无人机目标 n 处的 JSR; $\boldsymbol{O}_p$ 表示干扰设备的观测空间. 由所有干扰设备局部观测信息定义全局状态 $s_{p,t}=[o_{1,p,t},o_{2,p,t},\dots,o_{M,p,t}]\in\boldsymbol{S}_p$. 为确保各干扰设备独立实施干扰, 所设计模型除整合局部观测信息外, 还引入了干扰设备索引和目标无人机索引, 这一设计允许模型可针对各干扰设备和其分配的目标执行个性化处理, 以提高干扰决策的精准性.

(2) 动作空间. 以 $P_{\max}$ 为基准, 将干扰功率划分为 L_p 个等级, 以 $P_L=\{0,1,2,3,\dots,L_p\}$ 表示干扰功率等级集合, 0 表示不分配干扰功率; 同时, 兼顾式 (9) 中约束 C_2, C_3 和 C_4, 将 Actor 网络输出动作 $a_{m,p,t}$ 表示为 I 个无人机目标的干扰功率等级的组合, 其对应动作空间为

$$A_p=\underbrace{P_L\times P_L\times\cdots\times P_L}_{I}=\{(l_1,\dots,l_I)\,|l_1+\cdots+l_I\leqslant L_p,\forall i\in\boldsymbol{b}_m,l_i\in P_L,l_i\neq 0\},\tag{15}$$

其中, $\boldsymbol{b}_m$ 为第 m 个干扰设备的波束分配向量.

(3) 奖励函数. 分布式架构要求将功率干扰任务分配至各干扰设备, 因而需要对功率干扰任务的全局奖励进行分解; 考虑到资源效率, 将干扰功率消耗作为惩罚项引入奖励函数, 定义 $\mathrm{cost}=\sum_{n\in\boldsymbol{b}_m}p_{m,n}/P_{\max}$ 表示干扰设备 m 的归一化功率消耗. 综合上述考虑, 为确保各干扰设备能完成各自的干扰子任务, 定义如下奖励函数:

$$r_{m,p.t}=\begin{cases}-\rho_1, & \mathrm{JSR}_n<K_n,\forall n\in\boldsymbol{b}_m,\\ \rho_2-\rho_3*\mathrm{cost}, & \mathrm{JSR}_n\geqslant K_n,\forall n\in\boldsymbol{b}_m,\end{cases}\tag{16}$$

其中, ρ_1, ρ_2 和 ρ_3 均为正常数, ρ_3 用于控制训练过程中对功率消耗的重视程度.

3.2.2　去中心化训练, 分布式决策 (DTDE) 框架

尽管 “中心化训练, 分布式决策 (centralized training and decentralized execution, CTDE)” [25] 框架在一定程度上缓解了 MADRL 模型的环境非平稳问题, 但随智能体数量的增加, 训练时将面临 “维度诅咒” 挑战, 同时多智能体信用分配问题将更加突出 [26]. 考虑到在波束分配阶段已实现了对整体干扰任务的分解, 干扰设备在决策干扰功率时无需确知其他干扰设备的决策行动. 为此, 文中设计一种基于 DTDE 的 MADRL 框架, 如图 3 所示. 在 DTDE 框架下, 智能体在训练和决策时仅依赖自身的决策, 因此, DTDE 框架可有效地减少训练成本. 在文中所探讨的对抗环境, 各干扰设备具备相同的物理和功能属性, 它们可视为同构智能体. 为此, 这些设备可共享相同的训练参数, 即可采用单智能体 DRL 方法, 对多智能体网络参数进行训练, 从而显著提高训练效率和决策效果.

3.2.3　基于 DSAC 的功率资源分配策略

柔性演员 – 评论家 (soft actor-critic, SAC) 算法基于 AC 架构和最大熵强化学习理论, 旨在最大程度地提高期望收益和策略的熵; 同时, SAC 能够降低算法对模型与估计误差的敏感性, 使得算法更加健壮. 本文假设干扰功率为有限离散值, 因此, 本小节在 DTDE 框架下基于离散柔性演员 – 评论家 (DSAC) 算法提出一种高效的干扰功率分配策略.

为兼顾干扰收益和联合功率分配策略的随机性, 在式 (14) 中添加熵正则项, 形成如式 (17) 所示优化问题:

$$\pi^*=\underset{\pi=\langle\pi_1,\dots,\pi_M\rangle}{\arg\max}\sum_{t=0}^{\infty}\mathbb{E}_{(s_{p,t},\boldsymbol{a}_{p,t})\sim\rho_\pi}\left[\gamma^t\left(r(s_{p,t},\boldsymbol{a}_{p,t})+\alpha\mathcal{H}(\pi(\cdot\mid s_{p,t}))\right)\right],\tag{17}$$

其中, $\mathcal{H}(\pi(\cdot|s_{p,t}))=-\log\pi(\cdot|s_{p,t})$ 为联合功率分配策略 π 在状态 $s_{p,t}$ 的熵, α 代表熵温度系数, γ 为折扣因子.

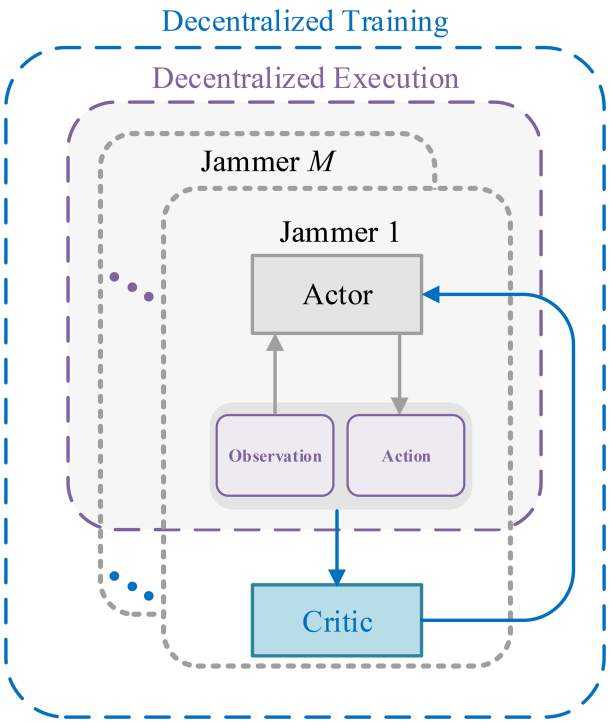

图 3　(网络版彩图) DTDE 框架.

Figure 3　(Color online) DTDE framework.

求解式 (17) 需要在最大熵框架下基于柔性策略采用交替迭代法进行策略评估和策略改进, 其中在柔性策略评估阶段, 干扰设备 m 的柔性状态价值函数为

$$V_m\left(o_{m,p,t}\right)=\pi_m(o_{m,p,t})^{\mathrm{T}}\left[Q_m\left(o_{m,p,t},a_{m,p,t}\right)-\alpha\log\left(\pi_m\left(a_{m,p,t}|o_{m,p,t}\right)\right)\right]. \tag{18}$$

DTDE 框架下, 各干扰设备仅针对所分配无人机目标执行干扰任务. 因而, 为了提升策略评估/改进效率, 设计一提取器 (见 3.3 小节) 对 $o_{m,p,t}$ 进行预处理, 提取与所分配目标相关的内容, 记为 $\tilde{o}_{m,p,t}$. 于是, 上式改写为

$$V_m\left(\tilde{o}_{m,p,t}\right)=\pi_m(o_{m,p,t})^{\mathrm{T}}\left[Q_m\left(\tilde{o}_{m,p,t},a_{m,p,t}\right)-\alpha\log\left(\pi_m\left(a_{m,p,t}|\tilde{o}_{m,p,t}\right)\right)\right]. \tag{19}$$

根据 Bellman 方程定义柔性动作价值函数 (柔性 Q 函数)

$$Q_m\left(\tilde{o}_{m,p,t},a_{m,p,t}\right)=r\left(\tilde{o}_{m,p,t},a_{m,p,t}\right)+\gamma\left(Q_m\left(\tilde{o}_{m,p,t+1},a_{m,p,t+1}\right)-\alpha\log\left(\pi_m\left(a_{m,p,t+1}|\tilde{o}_{m,p,t+1}\right)\right)\right). \tag{20}$$

上式确定了智能体当前策略的价值, 之后通过求解以下优化问题实现柔性策略改进:

$$\pi_{\mathrm{new}}=\arg\min_{\pi'_m\in\Pi}D_{\mathrm{KL}}\left(\pi'_m(\cdot\mid\tilde{o}_{m,p,t})\,\middle\|\,\frac{\exp\left(\alpha^{-1}Q_m^{\pi_{\mathrm{old}}}(\tilde{o}_{m,p,t},\cdot)\right)}{Z_m^{\pi_{\mathrm{old}}}(\tilde{o}_{m,p,t})}\right), \tag{21}$$

其中, D_{KL} 表示 Kullback-Leibler 散度, $Q_m^{\pi_{\mathrm{old}}}\left(\tilde{o}_{m,p,t},\cdot\right)$ 表示当前策略下的柔性 Q 函数; $Z_m^{\pi_{\mathrm{old}}}\left(\tilde{o}_{m,p,t}\right)=\sum_{a_{m,p,t}}\exp\left(\alpha^{-1}Q_m^{\pi_{\mathrm{old}}}\left(\tilde{o}_{m,p,t},a_{m,p,t}\right)\right)$ 为归一化因子, Π 表示候选策略集合. 式 (21) 通过不断使用新策略 π_{new} 代替旧策略 π_{old} 来实现柔性策略改进, 当动作空间有限 (即 $|A_p|<\infty$) 时, 对于任意的观测 $o_{m,p,t}$ 和动作 $a_{m,p,t}$, 可满足

$$Q_m^{\pi_{\mathrm{new}}}\left(o_{m,p,t},a_{m,p,t}\right)\geqslant Q_m^{\pi_{\mathrm{old}}}\left(o_{m,p,t},a_{m,p,t}\right). \tag{22}$$

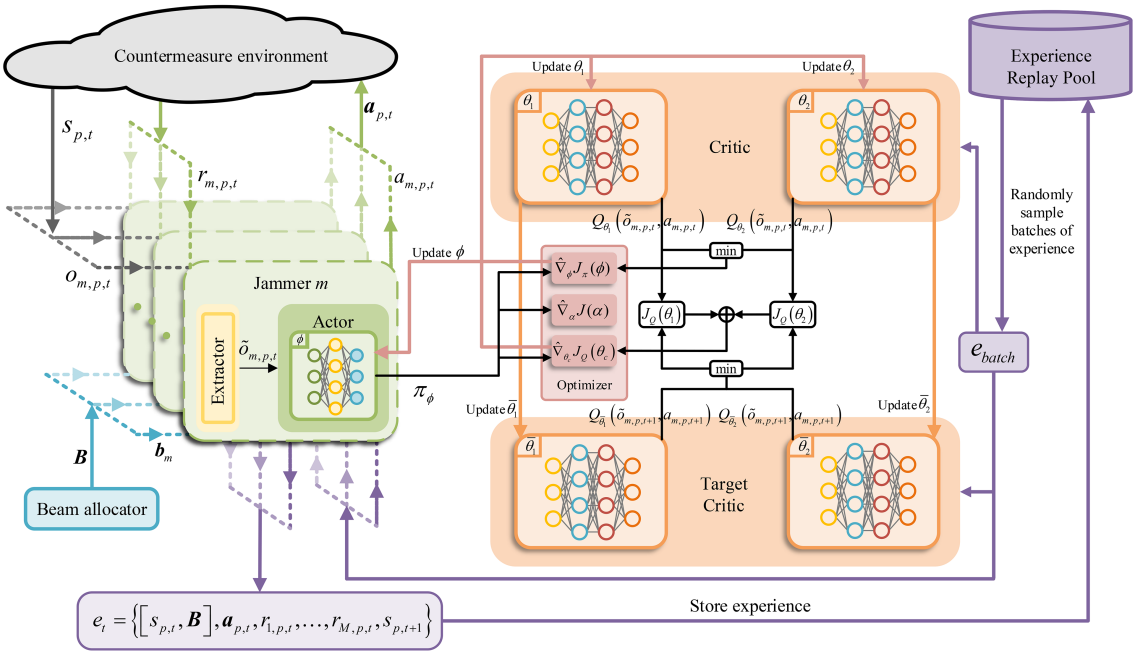

图 4　(网络版彩图) HCJRA 方案中功率分配策略训练原理图.

Figure 4　(Color online) Schematic diagram of power allocation strategy training in HCJRA scheme.

基于柔性策略的更新方法能够保证每次策略更新都会带来策略值函数的提升, 从而确保策略改进过程的单调性与稳定性. 通过交替进行柔性策略评估与改进, 当 $|A_p| < \infty$ 时, 对于任意的观测 $o_{m,p,t}$ 和动作 $a_{m,p,t}$, 策略 $\pi \in \Pi$ 可收敛到一最优策略 π^* [27].

在文中各干扰设备视为同构智能体, 各智能体使用同一网络模型, 分别利用 Critic 网络和 Actor 网络对柔性 Q 函数 $Q_m\left(\tilde{o}_{m,p,t}, a_{m,p,t}\right)$ 和策略函数 $\pi_m\left(\cdot|\tilde{o}_{m,p,t}\right)$ 进行参数化, 图 4 展示了所提 HCJRA 方案中功率分配策略的训练流程图.

假设在时刻 t 环境状态为 $s_{p,t}$, 指挥中心下发干扰波束分配指令, 经提取器预处理, 各干扰设备得到所分配无人机目标的局部观测信息, 接着将其输入至各自 Actor 网络进行决策, 形成联合动作 $\boldsymbol{a}_{p,t} = [a_{1,p,t}, a_{2,p,t}, \ldots, a_{M,p,t}]$; 执行 $\boldsymbol{a}_{p,t}$ 后, 状态由 $s_{p,t}$ 转移至 $s_{p,t+1}$, 干扰设备 m 获得奖励反馈 $r_{m,p,t}$; 随后, 构建经验样本 $e_t = \{[s_{p,t}, \boldsymbol{B}], \boldsymbol{a}_{p,t}, r_{1,p,t}, \ldots, r_{M,p,t}, s_{p,t+1}\}$, 将其存入经验回放池 $\mathcal{D}_p$ 中. 当经验回放池中累计样本数达到阈值 Φ_p 后, 随机从中抽取批量样本, 对 Actor 和 Critic 网络进行训练. 文中采用与 Critic 网络结构相同的目标 Critic 网络来缓解 Q 值过估计问题, 目标 Q 值为获得的奖励值与目标 Critic 网络输出的柔性 Q 值之和; 同时, 为保证训练的稳定性, Critic 网络和目标 Critic 网络中均采用 Double 结构 [28]. 具体训练过程如下.

在柔性策略评估过程中, 基于贝尔曼 (Bellman) 残差定义损失函数

$$J_Q\left(\theta_i\right) = \mathbb{E}_{\left(\tilde{o}_{m,p,t}, a_{m,p,t}\right)\sim\mathcal{D}}\left[\frac{1}{2}\left(Q_{\theta_i}\left(\tilde{o}_{m,p,t}, a_{m,p,t}\right) - y_{m,p,t}\right)^2\right], i \in \{1,2\}, \tag{23}$$

其中, θ_i 为 Critic 网络的参数, $y_{m,p,t}$ 为在状态 $\tilde{o}_{m,p,t+1}$ 下采取动作 $a_{m,p,t}$ 的柔性目标 Q 值, 可表示为

$$y_{m,p,t} = r\left(\tilde{o}_{m,p,t}, a_{m,p,t}\right) + \gamma\left(\min_{i=1,2} Q_{\bar{\theta}_i}\left(\tilde{o}_{m,p,t+1}, a_{m,p,t+1}\right) - \alpha\log\left(\pi_\phi\left(a_{m,p,t+1}|\tilde{o}_{m,p,t+1}\right)\right)\right), \tag{24}$$

其中, $\bar{\theta}_i$ 为目标 Critic 网络的参数, ϕ 为 Actor 网络的参数.

以最小化 $J_Q\left(\theta_i\right)$ 为目标, 基于随机梯度下降法 (stochastic gradient descent, SGD) 对 θ_i 进行更新, 即

$$\hat{\nabla}_{\theta_i} J_Q\left(\theta_i\right) = \nabla_{\theta_i} Q_{\theta_i}\left(\tilde{o}_{m,p,t}, a_{m,p,t}\right)\left(Q_{\theta_i}\left(\tilde{o}_{m,p,t}, a_{m,p,t}\right) - y_{m,p,t}\right), i \in \{1,2\}, \tag{25}$$

$$\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q\left(\theta_i\right), i \in \{1,2\}, \tag{26}$$

其中, λ_Q 为 Critic 网络的学习率.

在策略改进阶段, 定义如式 (27) 所示损失函数, 并基于 SGD 对 Actor 网络参数进行更新

$$J_\pi(\phi) = E_{(\tilde{o}_{m,p,t}, a_{m,p,t}) \sim \mathcal{D}} \left[\pi_\phi(\tilde{o}_{m,p,t})^{\mathrm{T}} \left[\alpha \log\left(\pi_\phi\left(a_{m,p,t} | \tilde{o}_{m,p,t}\right)\right) - \min_{i=1,2} Q_{\theta_i}\left(\tilde{o}_{m,p,t}, a_{m,p,t}\right) \right] \right], \tag{27}$$

$$\phi \leftarrow \phi - \lambda_\phi \hat{\nabla}_\phi J_\pi(\phi). \tag{28}$$

为了增强智能体在训练初期探索环境的随机性, 并在训练后期确保收益最大化, 需要在不同训练阶段动态调整熵温度系数 α. 当 α 较大时, 智能体的策略表现出更强的随机性; 随着训练的深入, 智能体逐渐学习到更优的策略, 此时 α 逐步减小, 智能体更倾向于采取能够最大化累计奖励的策略. 为此, 设定目标熵 $\bar{H}$, 定义损失函数 $J(\alpha)$, 然后采用 SGD 对 α 进行自适应优化更新

$$J(\alpha) = \pi_\phi(\tilde{o}_{m,p,t})^{\mathrm{T}} \left[-\alpha \left(\log\left(\pi_\phi\left(\tilde{o}_{m,p,t}\right)\right) + \bar{H} \right) \right], \tag{29}$$

$$\alpha \leftarrow \alpha - \lambda_\alpha \hat{\nabla}_\alpha J(\alpha), \tag{30}$$

其中, λ_α 为熵温度系数 α 的学习率.

当 Critic 网络参数更新后, 通过设定柔性更新系数 τ, 更新目标 Critic 网络的参数

$$\bar{\theta}_i \leftarrow \tau \theta_i + (1-\tau)\,\bar{\theta}_i, i \in \{1,2\}, \tag{31}$$

其中, λ_ϕ 为 Actor 网络的学习率.

在上述分析的基础上, 算法 2 给出了基于 DSAC 的干扰功率资源分配策略的整体训练流程.

3.3　HCJRA 方案模块化结构

图 5 展示了 HCJRA 方案的模块化结构, 包括波束分配器和功率分配器. 波束分配器位于指挥中心, 负责收集每个无人机目标的运动状态信息 $s_{n,b,t}$, 并将其作为输入, 生成干扰波束分配矩阵 $\boldsymbol{B}$, 之后将干扰波束分配结果下发至各干扰设备. 干扰设备 m 在接收波束分配结果 $\boldsymbol{b}_m$ 和相应的局部观测信息 $o_{m,p,t}$ 后, 利用提取器剔除与未分配无人机目标相关的信息, 得到优化后的观测数据 $\tilde{o}_{m,p,t}$; 之后, 基于 $\tilde{o}_{m,p,t}$, 功率分配器对分配的无人机目标进行干扰功率决策.

经过优化处理的局部观测信息 $\tilde{o}_{m,p,t}$ 的维度不再与无人机目标总数相关, 而仅与干扰设备 m 分配的无人机目标数量有关, 这种优化方式在干扰设备性能限制下, 支持对任意数量的无人机目标进行功率资源分配, 不仅有效降低了神经网络输入的维度, 而且显著提高了方案的扩展性; 同时, 通过从观测数据中提取关键信息, 减少了无关信息对决策过程的干扰, 从而提升了系统的决策效率和精度.

3.4　复杂度分析

DRL 的计算复杂度与状态空间、动作空间、奖励函数, 以及策略本身的结构有关, 而且通常以网络参数量级来衡量计算复杂度. 文中波束分配器和功率分配器中的网络模型均基于 AC 架构, 相关参数均为全连接层神经元之间的权重和偏置量. 因而, 为便于分析, 假设训练单个参数的计算复杂度为 χ.

假设波束分配器 Critic 网络的隐藏层数为 $\rho_{b,c}$, 第 i 层包含 $k_i^{b,c}$ 个神经元; Actor 网络的隐藏层数为 $\rho_{b,a}$, 第 i 层包含 $k_i^{b,a}$ 个神经元. Critic 网络输入层的神经元数与干扰设备数 M 和目标威胁度信息维度相关, 其输出层神经元数为 M, 则 Critic 网络的参数量为 $(3M+1)k_1^{b,c} + \sum_{i=1}^{\rho_{b,c}-1}(k_i^{b,c}+1)k_{i+1}^{b,c} + (k_{\rho_{b,c}}^{b,c}+1)M$. Actor 网络输入层和输出层的神经元数与 Critic 网络相同, 则 Actor 网络的参数量可表

Algorithm 2 Jamming power resource allocation strategy based on DSAC.

Input: The number of jammers: M; the number of UAV targets: N; Critic network parameters: $Q_{\theta_1}, Q_{\theta_2}$; target Critic network parameters: $Q_{\bar{\theta}_1}, Q_{\bar{\theta}_2}$; $\bar{\theta}_1 \leftarrow \theta_1$, $\bar{\theta}_2 \leftarrow \theta_2$; Actor network parameters: ϕ; the number of training episodes: max_episode; interaction time steps per episode: max_t; experience replay pool: $\mathcal{D}_p$; the threshold at which the model starts training: Φ_p; the beam allocator.

Initialize parameters of Actor and Critic networks and experience replay pool $\mathcal{D}_p$;
for episode = 1 to max_episode **do**
  Randomly initialize the countermeasure environment;
  for $t = 1$ to max_t **do**
    The beam allocator generates the beam resource allocation matrix $\boldsymbol{B}$;
    for $m = 1$ to M **do**
      The jammer m obtains the beam allocation $\boldsymbol{b}_m$, extract the observation information to get $\tilde{o}_{m,p,t}$, the power allocator inputs $\tilde{o}_{m,p,t}$, and outputs action $a_{m,p,t}$;
    end for
    Execute the joint action $\boldsymbol{a}_{p,t}$, the state is transferred to $s_{p,t+1}$, each jammer obtains reward value feedback $r_{m,p,t}$ and gets the next observation, store experience $e_t = \left\{[s_{p,t}, \boldsymbol{B}], \boldsymbol{a}_{p,t}, r_{1,p,t}, \ldots, r_{M,p,t}, s_{p,t+1}\right\}$ into $\mathcal{D}_p$;
    if the number of experiences in $\mathcal{D}_p$ is greater than Φ_p **then**
      Randomly sample batch experiences from $\mathcal{D}_p$;
      for $m = 1$ to M **do**
        Update Critic network parameters $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_c} J_Q(\theta_c)$ for $i \in \{1, 2\}$;
        Update Actor network parameters $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$;
        Update entropy temperature coefficient $\alpha \leftarrow \alpha - \lambda_\alpha \hat{\nabla}_\alpha J(\alpha)$;
        Update target critic network parameters softly $\bar{\theta}_i \leftarrow \tau\theta_i + (1-\tau)\bar{\theta}_i$ for $i \in \{1, 2\}$;
      end for
    end if
  end for
end for

Output: Power allocator.

示为 $(3M+1)k_1^{b,a} + \sum_{i=1}^{\rho_{b,a}-1}(k_i^{b,a}+1)k_{i+1}^{b,a} + (k_{\rho_{b,a}}^{b,a}+1)M$. 因此, 干扰波束分配策略进行一次前向传播的计算复杂度为

$$O\left(\chi\left((3M+1)\,k_1^{b,c} + \sum_{i=1}^{\rho_{b,c}-1}\left(k_i^{b,c}+1\right)k_{i+1}^{b,c} + \left(k_{\rho_{b,c}}^{b,c}+1\right)M + (3M+1)\,k_1^{b,a} + \sum_{i=1}^{\rho_{b,a}-1}\left(k_i^{b,a}+1\right)k_{i+1}^{b,a} + \left(k_{\rho_{b,a}}^{b,a}+1\right)M\right)\right). \tag{32}$$

假设功率分配器 Critic 网络的隐藏层数为 $\rho_{p,c}$, 第 i 层包含 $k_i^{p,c}$ 个神经元; Actor 网络的隐藏层数为 $\rho_{p,a}$, 第 i 层有 $k_i^{p,a}$ 个神经元. 同样地, Critic 网络输入层的神经元数与干扰设备最多可同时干扰的无人机目标数量 I 和单个无人机目标的观测信息维度有关; 其输出层神经元数与 A_p 维度相同, 则 Critic 网络的参数量为 $(7I+1)k_1^{p,c} + \sum_{i=1}^{\rho_{p,c}-1}(k_i^{p,c}+1)k_{i+1}^{p,c} + (k_{\rho_{p,c}}^{p,c}+1)|A_p|$. Actor 网络输入层和输出层的神经元数与 Critic 网络相同, 则 Actor 网络的参数量为 $(7I+1)k_1^{p,a} + \sum_{i=1}^{\rho_{p,a}-1}(k_i^{p,a}+1)k_{i+1}^{p,a} + (k_{\rho_{p,a}}^{p,a}+1)|A_p|$. 因此, 基于 DSAC 的功率资源分配策略进行一次前向传播的计算复杂度为

$$O\left(\chi\left((7I+1)\,k_1^{p,c} + \sum_{i=1}^{\rho_{p,c}-1}(k_i^{p,c}+1)\,k_{i+1}^{p,c} + \left(k_{\rho_{p,c}}^{p,c}+1\right)|A_p| + (7I+1)\,k_1^{p,a} + \sum_{i=1}^{\rho_{p,a}-1}(k_i^{p,a}+1)\,k_{i+1}^{p,a} + \left(k_{\rho_{p,a}}^{p,a}+1\right)|A_p|\right)\right). \tag{33}$$

CTDE 框架和 DTDE 框架在共用波束分配器的前提下进行功率分配, 接下来对 CTDE 框架下功

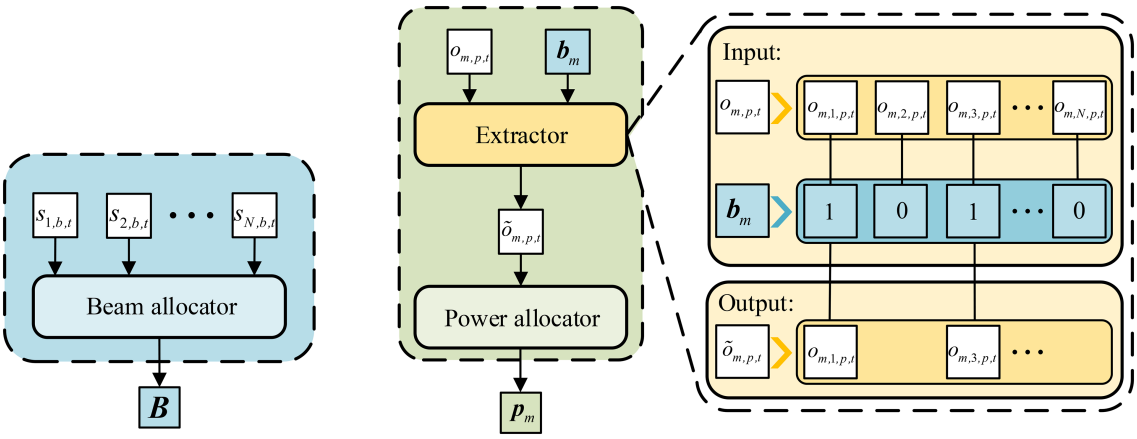

图 5　(网络版彩图) HCJRA 方案模块结构 (波束分配器和功率分配器).

Figure 5　(Color online) HCJRA scheme module structure diagram (beam allocator and power allocator).

率资源分配策略的复杂度进行分析. 在 CTDE 框架下, 功率资源分配策略的 Actor 网络计算复杂度与 DTDE 框架相同, 但 Critic 网络的输入包括所有干扰设备的观测和动作, 输入层的神经元个数为 $7MI+MN$, 则 Critic 网络参数量为 $(7MI+MN+1)\,k_1^{p,c}+\sum_{i=1}^{\rho_{p,c}-1}(k_i^{p,c}+1)\,k_{i+1}^{p,c}+(k_{\rho_{p,c}}^{p,c}+1)\,|A_p|$. 因此, CTDE 框架下的功率资源分配策略进行一次前向传播的计算复杂度为

$$O\left(\chi\left(\begin{aligned}&(7MI+MN+1)\,k_1^{p,c}+\sum_{i=1}^{\rho^{p,c}-1}(k_i^{p,c}+1)\,k_{i+1}^{p,c}+\left(k_{\rho^{p,c}}^{p,c}+1\right)|A_p|\\&+(7I+1)\,k_1^{p,a}+\sum_{i=1}^{\rho^{p,a}-1}(k_i^{p,a}+1)\,k_{i+1}^{p,a}+\left(k_{\rho^{p,a}}^{p,a}+1\right)|A_p|\end{aligned}\right)\right). \tag{34}$$

对比可见, 随着干扰设备和无人机目标数量的增加, CTDE 框架下的功率分配策略计算复杂度显著上升; 而在 DTDE 框架中, 功率分配策略的计算复杂度则不受干扰设备和无人机目标数量的影响.

4　数值仿真分析

4.1　仿真参数设置

仿真中, 设干扰方指挥控制中心下辖 3 部干扰设备, 分别位于 XY 平面 (20, 0), (40, 0) 和 (60, 0) km 处, 对位于 (40, 400) km 处的预警机指挥的 5 架无人机组成的飞行编队实施干扰压制, 5 架无人机分别位于 (10, 300), (25, 290), (40, 300), (55, 290) 和 (70, 300) km 处. 为增强干扰设备对环境的适应性, 每个对抗回合开始, 无人机水平和垂直飞行速度分别在 $[-0.2, 0.2]$ 和 $[-0.4, -0.6]$ km/s 范围内随机初始化, 设定预警机飞行速度为 $(0, -0.5)$ km/s; 每个对抗回合无人机目标飞行时间设定为 200 s, 回合结束后重新初始化环境; 无人机目标通信频率介于 225~300 MHz, 以 1 MHz 为间隔随机选择, 并在下个回合更新; 本文假设干扰设备释放瞄准式干扰, 干扰方在侦察系统辅助下, 运用实时频谱感知与窄带参数估计技术, 动态生成匹配无人机目标通信信号时频特征的瞄准式干扰波形, 其干扰中心频率自适应跟踪目标通信频点, 且干扰带宽与无人机目标通信信号带宽实现严格对齐; 假设各无人机目标具有相同的干扰压制系数; 训练中随机抽样批量大小设为 64. 其他仿真参数设置如表 2 所示.

4.2　仿真分析

4.2.1　训练仿真与分析

为评估本文所提 HCJRA 方案的性能, 将其与 MADJPA 方案 [17]、D3QN 方案 [8]、双深度 Q 网

表 2　仿真参数.

Table 2　Simulation parameters.

Parameter name	Value	Parameter name	Value
The maximum number of UAV targets one jammer can jam I	2	The maximum number of beams assigned to each UAV target U	2
Jamming power levels L_p	10	Jamming link transmit antenna gain $G_{m,n}$	5 dB
Maximum jamming power of each jammer $P_{\max}$	70 dBm	Communication link transmit antenna gain G_{nt}	2 dB
Communication link transmit power p_n	55 dBm	Communication link receive antenna gain G_{nr}	3 dB
Noise power σ^2	−85 dBm	Jamming suppression coefficient K_n	3
Size of experience replay pool $\mathcal{D}_b$, $\mathcal{D}_p$	1e6, 1e6	Soft update factor τ	0.01
Discount factor γ	0.99	Entropy temperature coefficient α	3

表 3　波束分配器与功率分配器网络模型参数.

Table 3　Network model parameters of beam allocator and power allocator.

Model	Parameter name	Value	Parameter name	Value
Beam allocator	Input layer	$3 \times M$	Output layer	M
	Hidden layer 1	128	Hidden layer 2	32
	Critic network learning rate	0.001	Actor network learning rate	0.0001
	Activation function	ReLU	Optimizer	Adam
Power allocator	Input layer	$7 \times I$	Output layer	45
	Hidden layer 1	256	Hidden layer 2	128
	Critic network learning rate	0.0002	Actor network learning rate	0.0001
	Activation function	ReLU	Optimizer	Adam

络 (double deep q-network, DDQN) 方案 [29] 以及随机干扰资源分配方案 (以 Random 标识) 进行对比. MADJPA 方案原本基于 AC 模型和 CTDE 架构, 在训练过程中需要对多个智能体进行训练, 为了与本文所提方案进行对比, 将 MADJPA 方案拓展至 DSAC 模型, 以 MADJPA-D 标识; 考虑到干扰设备通常为同构智能体, 因而将训练单个智能体的 MADJPA-D 方案以 MADJPA-DSA 标识; D3QN 方案与 DDQN 方案均采用基于价值的 DRL 算法模型, 基于一致性考虑, 将这两种方案推广至 DTDE 框架下进行训练. 为保证对比公平性, 对比方案均采用相同的提取器. HCJRA 方案中, 尽管干扰波束分配和干扰功率分配相对独立, 但为保持整体方案的一致性, 波束分配器亦采用 DSAC 模型, 每种方案均在 3 个不同的随机数种子环境下进行仿真实验. 波束分配器和功率分配器各自的 Critic 和 Actor 网络结构相同, 参数调优后的网络参数配置见表 3.

图 6 展示了训练过程中波束分配器累计奖励值的变化情况. 在每个对抗回合, 通过对 Actor 网络输出概率向量进行采样, 分别为 5 个无人机目标确定对应的干扰设备. 根据式 (12), 每回合最高累计奖励值应为 5, 该仿真结果恰好与其一致. 同时, 波束分配器大约在 200 回合后趋于收敛.

图 7 展示了训练过程中功率分配器的平均干扰成功率随对抗回合数的变化情况. D3QN 和 DDQN 方案采用贪心策略, 训练前期具有较高的探索率, 但收敛慢, 最终平均干扰成功率较低; 随机干扰方案, 干扰设备在每个时刻随机选择无人机目标和干扰功率, 其干扰成功率始终维持在约 67% 左右. MADJPA-DSA 方案具有更快的收敛速度, 其干扰成功率收敛水平可达到 100%; HCJRA 方案大约在

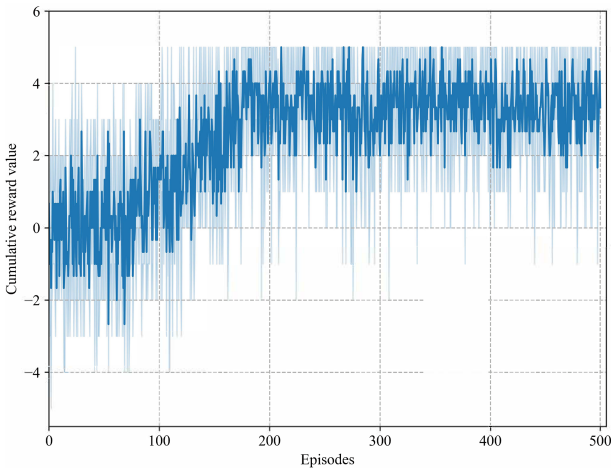

图 6　(网络版彩图) 波束分配器训练过程累计奖励值.

Figure 6　(Color online) Cumulative reward value during the beam allocator training.

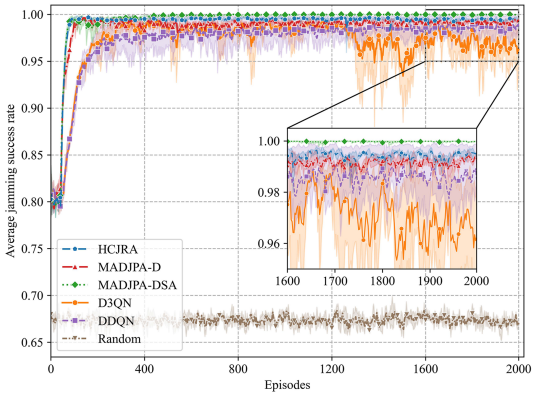

图 7　(网络版彩图) 功率分配器训练过程平均干扰成功率.

Figure 7　(Color online) Average jamming success rate during power allocator training.

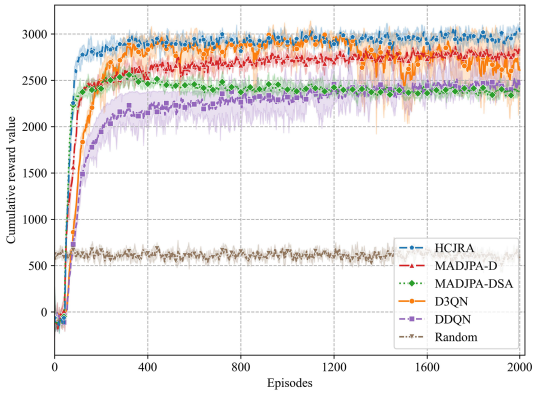

图 8　(网络版彩图) 功率分配器训练过程累计奖励值.

Figure 8　(Color online) Cumulative reward value during power allocator training.

100 回合时迅速达到峰值干扰成功率, 而且后期基本维持在 99% 以上, 虽然 MADJPA-DSA 的干扰成功率略高于 HCJRA, 但其干扰功率消耗明显更高 (见图 8), HCJRA 方案相比 MADJPA-DSA 方案在干扰效费比指标方面具有显著优势.

图 8 展示了训练过程中功率分配器的奖励值随对抗回合数的变化情况. 训练前期 MADJPA-D 和 MADJPA-DSA 方案的累计奖励值增长较为迅速, 但受多智能体信用分配问题影响, 在训练后期这两种方案的奖励值难以得到进一步提升; 尤其是 MADJPA-DSA 方案, 由于其在 CTDE 框架仅对单个智能体进行训练, 导致其探索能力非常有限, 各干扰设备难以从全局干扰收益中准确获得自身干扰决策的收益. 因此, 无法通过进一步的策略探索提升干扰效费比, MADJPA-DSA 方案最终收敛的干扰功率分配策略相对保守, 虽然干扰成功率非常可观, 但由于干扰功率资源的过度消耗导致累积奖励值呈现先上升后下降的趋势. D3QN 和 DDQN 方案在训练过程中稳定性较差, 曲线波动较为明显. HCJRA 方案基于 DTDE 框架由于实现了对整体干扰收益的分配, 有效缓解了信用分配问题, 相比其他方案, 累计奖励值收敛速度最快, 且最终收敛奖励值也最高.

图 9 展示了训练过程中功率分配器的归一化干扰功率消耗随对抗回合数的变化情况. MADJPA-D

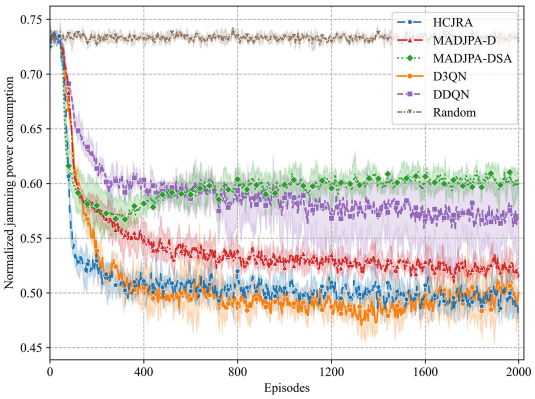

图 9　(网络版彩图) 功率分配器训练过程归一化干扰功率消耗.

Figure 9 (Color online) Normalized jamming power consumption during power allocator training.

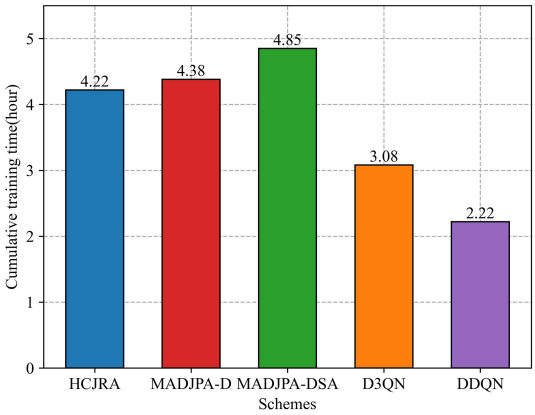

图 10　(网络版彩图) 不同方案平均训练时间.

Figure 10 (Color online) Average training time for different schemes.

和 MADJPA-DSA 方案虽然具有较高的干扰成功率, 但受 CTDE 框架下信用分配问题的影响, 难以学习到更优的策略. MADJPA-DSA 方案实际上通过牺牲干扰功率资源换取了极高的干扰成功率. D3QN 方案虽然在训练过程中出现功率消耗低于 HCJRA 方案的情况, 但由于其训练不稳定无法有效平衡智能体的探索和利用. 相比之下, HCJRA 方案在收敛速度和训练稳定性方面优势明显, 且能够以更少的功耗实现对无人机目标的有效干扰压制.

图 10 展示了不同方案的平均训练时间对比. D3QN 和 DDQN 方案仅由 Q 值网络和目标 Q 值网络组成, 因而其平均训练时间较短, 但其训练稳定性和干扰压制性能均较差; 相比之下, 其他方案的网络结构更复杂, 因而训练时间较长. 具体而言, HCJRA 方案中, 智能体的 Critic 网络输入仅包含自身观测信息, 而 MADJPA-D 和 MADJPA-DSA 方案的 Critic 网络输入包含所有智能体的观测和动作, 因此 HCJRA 所需训练时间低于 MADJPA-D 和 MADJPA-DSA 方案; 此外, MADJPA-D 方案需要同时训练 3 个智能体, 而 HCJRA 方案在相同时间内仅需针对单个智能体进行训练, 因而训练成本低. 这里仅给出不同方案在离线训练阶段的时间消耗对比. 需要注意的是, 网络模型的训练时间和推理速度受网络模型本身复杂度、软硬件平台 (如 CPU, GPU) 限制等多种因素影响, 实际工程应用中可通过模型压缩、硬件加速等技术降低模型训练和推理的复杂度.

4.2.2　测试仿真与分析

为验证本文所提 DRL 辅助的目标威胁度评估模型的性能, 在设置相同随机数种子的条件下, 表 4 展示了基于 TSM 的目标威胁度评估模型与 DRL 辅助的目标威胁度评估模型在单阶段和两阶段波束资源分配策略下的平均测试结果对比. 表中数据基于 100 个对抗回合、共执行 20000 次波束资源分配策略后取平均所得. 从表中可知, 在单阶段波束分配策略下, 基于 TSM 的目标威胁度评估模型几乎无法实现波束资源的合理分配, 而 DRL 辅助的目标威胁度评估模型可达到 0.745 的成功率, 展现出了更强的分配能力和适应性; 在两阶段波束资源分配策略下, 两种模型均可 100% 实现波束资源的分配. DRL 辅助的目标威胁度评估模型在第 1 阶段波束分配过程中矩阵 $\boldsymbol{W}$ 平均遍历次数为 7.313, 明显低于基于 TSM 的目标威胁度评估模型的 13.005, 表明其有效缓解了干扰设备在波束资源分配过程中竞争同一无人机目标的问题. 同时, 从两阶段总遍历次数来看, DRL 辅助的目标威胁度评估模型较基于 TSM 的目标威胁度评估模型减少 26.9%, 显著提升了资源分配效率. 仿真结果表明, DRL 辅助的目标威胁度评估模型在波束资源有限、无人机目标分布复杂的环境中具备更高的分配效率.

在训练过程中, 功率分配器为每个干扰设备分配 I 个无人机目标的干扰功率, 以确保在设备性能

表 4　目标威胁度评估模型与波束资源分配策略测试性能对比.

Table 4　Comparison of testing performance of target threat assessment models and beam resource allocation strategies.

Model	Single-stage	Two-stage			
	Success rate	Stage 1 (passes)	Stage 2 (passes)	Total (passes)	Success rate
TSM-based	5.00×10^{-5}	13.005	1.654	14.659	1.000
DRL-assisted	0.745	7.313	3.401	10.714	1.000

表 5　不同方案平均干扰压制性能对比.

Table 5　Comparison of average jamming suppression performance of different schemes.

Scheme	Average jamming success rate	Average normalized jamming power consumption	Average jamming cost-effectiveness ratio
HCJRA	0.9994	0.3881	2.5800
MADJPA-D	0.9965	0.4148	2.4050
MADJPA-DSA	1.0	0.4665	2.1439
D3QN	0.9751	0.4188	2.3362
DDQN	0.9551	0.5038	1.9056

限制范围内, 能够对不超过 I 个无人机目标进行功率分配. 根据干扰波束分配结果, 一个无人机目标可能同时受到多个干扰设备的干扰. 为此, 基于 DTDE 框架设计了以干扰子任务为核心的奖励函数, 并对训练环境进行随机初始化, 从而使功率分配器能够学习并获得稳健的功率分配策略. 在该策略的指导下, 干扰设备之间无需交换功率分配信息, 即可独立完成分配目标的功率分配任务, 并确保干扰子任务的成功执行. 特别是在功率资源充足的条件下, 对于同时受到多个干扰设备干扰的无人机目标, 仅需其中一个干扰设备即可实现对目标的有效压制. 因此, 在干扰无人机目标时可能存在干扰功耗冗余问题. 为此, 本文在完成对功率分配器的训练后, 保存相应的神经网络模型, 并使用与训练过程不同的功率分配策略对模型进行测试. 具体而言, 各干扰设备首先依据第 2 阶段的波束资源分配结果生成功率分配动作, 随后按照第 1 阶段分配的无人机目标执行对应的功率分配操作. 该策略在保证对无人机目标形成有效干扰压制的前提下, 可进一步提升干扰功率资源的利用效率.

表 5 展示了不同方案的干扰压制性能测试对比结果. 其中, 平均干扰效费比定义为 100 个回合中干扰效费比的均值, 而单个回合干扰效费比则定义为单个回合内干扰成功率与归一化功率消耗的比值. 由表 5 可知, HCJRA 方案在实现有效干扰的前提下, 功率资源消耗更低, 在平均干扰效费比上具有显著的性能优势.

为了进一步直观评估 HCJRA 方案中波束分配器与功率分配器的性能, 图 11 展示了单个对抗回合中预警机与无人机的飞行轨迹. 图中显示, 为了应对干扰方的压制性干扰, 预警机指挥的无人机编队中目标的飞行速度和方向持续动态变化. 从而说明对抗环境的复杂性.

图 12 给出了在图 11 所示对抗回合中, 干扰设备的功率分配策略和对无人机目标实施压制干扰的情况 (颜色深浅代表功率等级高低, 背景颜色表示功率为 0 级). 由图 12 可知, 干扰设备在不同时间点上根据波束分配结果, 对无人机目标执行自适应功率分配决策. 在对抗初期, 由于干扰距离较远, 达到干扰压制所需功率较大, 因而图 12(a) 中干扰设备 1 分配给无人机目标 1 和 2 的功率等级为 3 级; 随着目标逐渐向干扰设备靠近, 达到干扰压制所需的功率逐渐减小, 功率分配器随之调整功率分配策略; 到对抗结束时, 干扰设备 1 分配给无人机目标 3 和 5 的干扰功率等级已降至 1 级. 其他干扰设备也具有类似的结果, 限于篇幅, 文中没有给出干扰设备 2 和 3 的具体结果. 此外, 在对抗过程中, 存在波束分配结果保持不变的情况. 比如在 12~200 s 期间, 干扰设备 1 持续干扰无人机目标 3 和 5. 在实际对抗中, 频繁切换波束指向可能导致干扰资源的无为消耗; 为此, 干扰设备在接收到波束分配指令后, 直至下次波束分配更新期间, 都会对所分配的无人机目标持续进行功率分配的自适应调整, 即指挥中

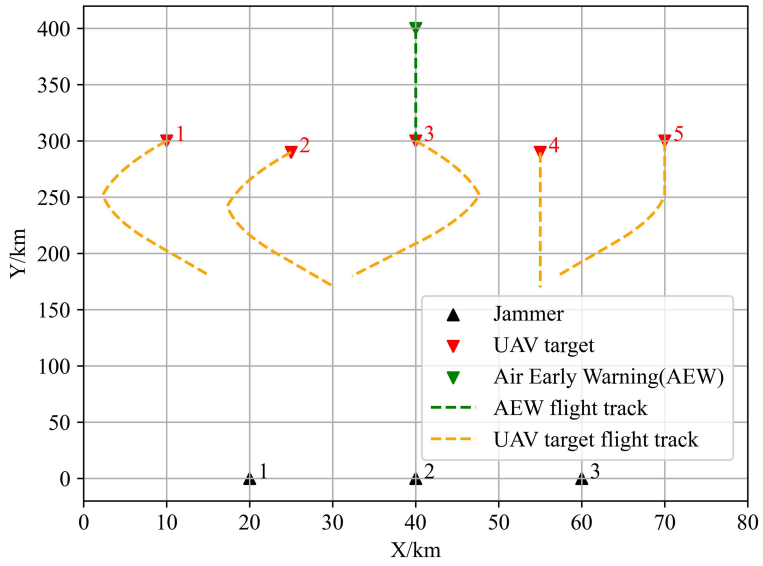

图 11　(网络版彩图) 对抗回合示意图.

Figure 11　(Color online) Diagram of confrontation round.

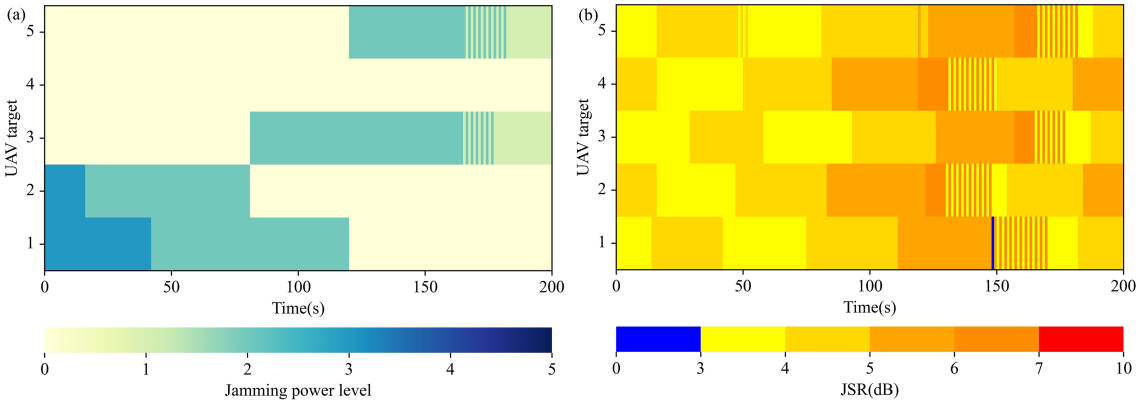

图 12　(网络版彩图) 对抗回合干扰功率分配与干扰压制效果热力图. (a) 干扰机 1 的功率分配结果; (b) 各目标 UAV 的 JSR 变化结果.

Figure 12　(Color online) Heat map of jamming power allocation results and jamming suppression effect in the confrontation round. (a) Power allocation results of jammer1; (b) results of JSR changes for each UAV target.

心更新干扰波束分配指令的频率对干扰功率分配决策并无决定性影响; 指挥中心只需根据对抗态势发展, 灵活地下达波束分配指令.

图 12(b) 展示了在对抗回合中, 不同无人机目标处的 JSR 随时间的变化情况. 当 JSR 低于干扰压制系数时, 相应区域用蓝色标示. 当 JSR 大于干扰压制系数时, 相应区域为黄色, 且随着 JSR 的升高颜色逐渐加深. 由图 12(b) 可知, 在无人机编队向干扰方靠近过程中, 对抗环境处于动态变化中, 导致无人机目标的 JSR 并非恒定. 随着无人机编队逐步逼近干扰方, 功率分配器持续地接收与无人机目标相关的观测信息, 综合考虑通信距离、干扰距离以及干扰功率等多方面因素, 通过自适应决策调整干扰功率, 使得无人机目标的 JSR 维持在接近或略高于干扰压制阈值的水平. 说明 HCJRA 方案不但提高了干扰功率资源的利用率, 而且确保在动态环境下依然可取得良好的干扰效果.

为验证本文所提 HCJRA 方案的泛化性, 在干扰设备数量固定为 3, 无人机目标数量分别设置为 3, 4, 5, 6 时, 图 13 给出了不同方案在 100 个对抗回合的平均成功率、平均干扰功率消耗和平均干扰效费比. 在对抗场景中, 当无人机目标数量发生变化时, 基于 CTDE 框架的模型需要重新训练. 然而,

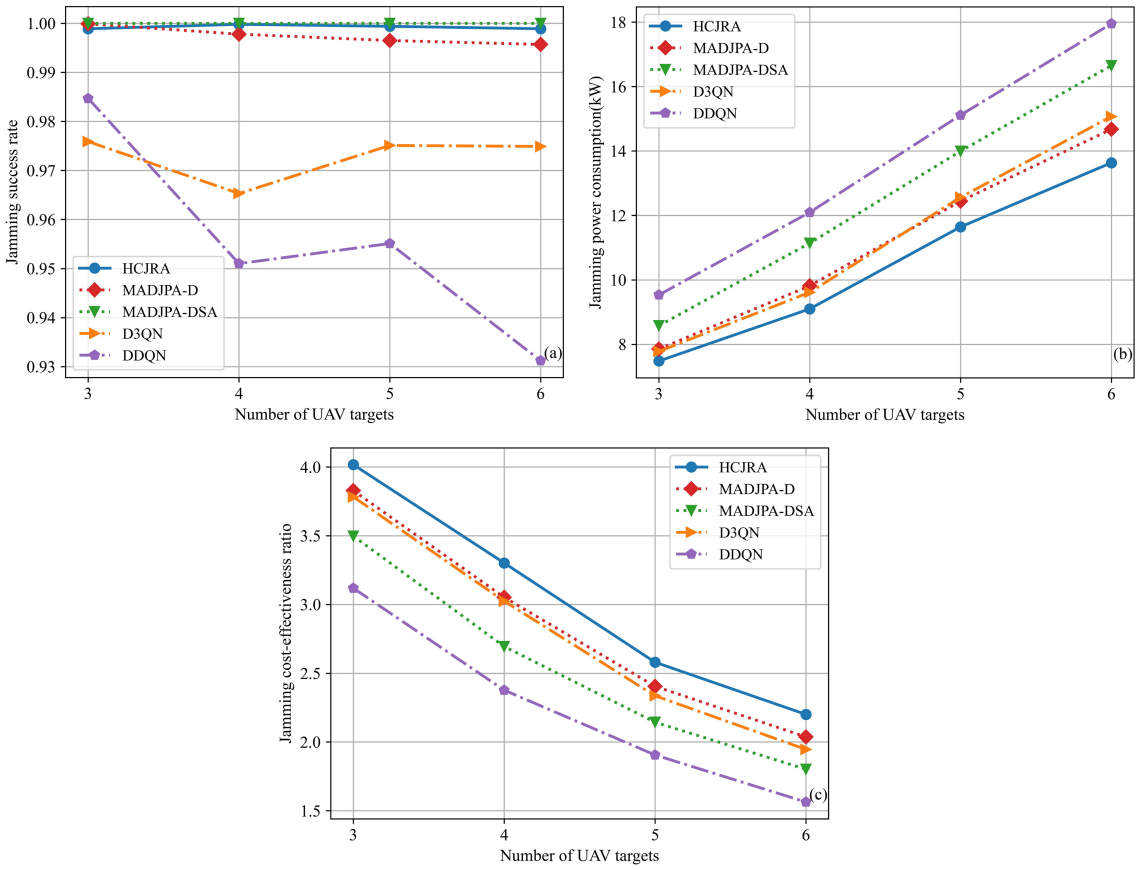

图 13　(网络版彩图) 干扰设备数为 3, 无人机目标数为 3, 4, 5, 6 对抗场景下的干扰性能对比. (a) 干扰成功率; (b) 干扰功率消耗; (c) 干扰效费比.

Figure 13　(Color online) Comparison of jamming performance in confrontation scenarios with 3 jammers and 3, 4, 5, and 6 UAV targets. (a) Jamming success rate; (b) jamming power consumption; (c) jamming cost-effectiveness ratio.

MADJPA-D 和 MADJPA-DSA 方案采用了文中设计的提取器, 使得功率分配器的 Actor 网络输入维度仅与分配到的无人机目标数相关, 而与无人机目标总数无关, 因此可用于无人机目标数量可变的场景. 由图 13 可知, 随着无人机目标数量的增多, 整体干扰功率消耗也相应增加, 对应的干扰效费比也在下降. HCJRA 方案将每个干扰设备的功率决策与干扰子任务绑定, 而非依赖无人机目标总数, 因此, 即使无人机目标数量发生变化, HCJRA 方案相较对比方案仍能保持显著的干扰效费比优势; 相比 MADJPA-DSA 方案, 干扰成功率略有逊色, 但其干扰功率消耗却明显高于 HCJRA 方案; 而 D3QN 和 DDQN 方案由于训练过程中的稳定性不足, 导致测试阶段干扰成功率难以得到保障.

设定无人机目标数量为 6、干扰设备数量分别设定为 3, 4, 5 时, 图 14 给出不同方案的平均性能对比结果, 其中干扰设备间隔均匀地部署在同一水平线上. 由于 MADJPA-D 和 MADJPA-DSA 基于 CTDE 框架, 其采用全局奖励对模型进行训练; 当干扰设备数量变化时, 全局奖励也随之改变, 此时模型需要重新设计和训练, 因而二者无法适应干扰设备数动态变化场景, 为此图 14 中未给出这两种方案的仿真结果. 从图 14(a) 可看出, 尽管干扰设备数量有所变化, 但在干扰相同数量无人机的场景中, HCJRA 方案的干扰成功率几乎不受干扰设备数量变化的影响, 表现出更为稳健的性能, 表明文中提出的 DTDE 框架下的 HCJRA 方案具有更强的环境适应能力. 在图 14(b) 中, 由于干扰设备数量的变化导致其部署位置发生了变化, 最终对干扰功率的决策产生了一定影响. 然而, 在实现对无人机目标的整体干扰压制时, 无论干扰设备数量如何变化, 总功率消耗仍维持在相对稳定的水平; 而相比之下, D3QN 和 DDQN 方案不但消耗了更多的功率资源, 而且干扰成功率均明显低于 HCJRA 方案, 导致其

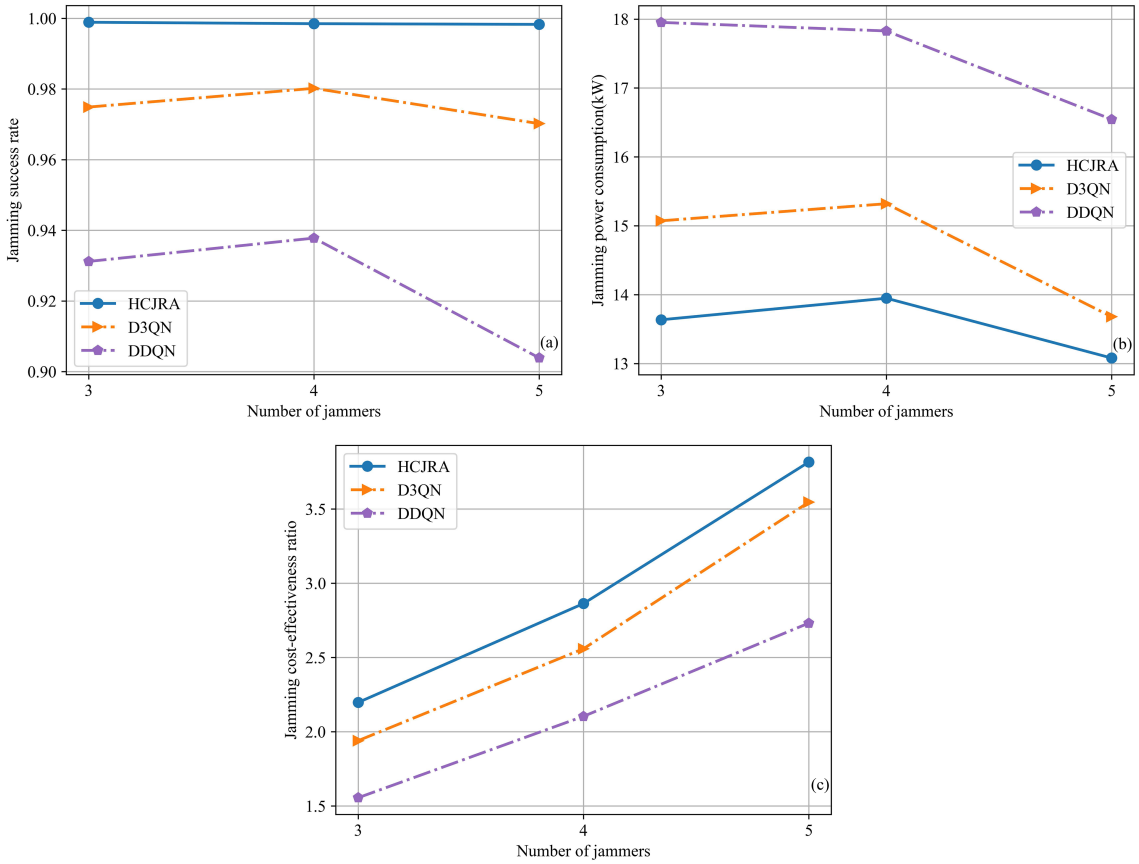

图 14　(网络版彩图) 无人机目标数为 6，干扰设备数为 3，4，5 对抗场景下的干扰性能对比．(a) 干扰成功率; (b) 干扰功率消耗; (c) 干扰效费比.

Figure 14　(Color online) Comparison of jamming performance in confrontation scenarios with 6 UAV targets and 3, 4, and 5 jammers. (a) Jamming success rate; (b) jamming power consumption; (c) jamming cost-effectiveness ratio.

干扰效费比性能亦显著低于 HCJRA 方案.

5　结论

针对 “多对多” 通信对抗场景中的动态差异化干扰资源分配问题, 本文提出了一种基于 DRL 的 HCJRA 方案. 该方案首先构建了多干扰设备协同对抗敌方无人机编队的干扰模型, 将整体干扰资源分配问题分解为波束分配和功率分配两个阶段; 接着, 利用 DRL 辅助进行目标威胁度的动态评估, 并优化干扰波束资源分配; 随后, 将多干扰设备的功率资源分配问题建模为 Dec-POMDP, 设计了 DTDE 多智能体框架, 并基于 DSAC 算法提出了功率资源分配策略, 实现了多干扰设备的分布式协同功率决策; 最后, 对所提方案的训练和测试过程进行了仿真分析. 仿真结果表明, 在多干扰设备协同对抗无人机编队的场景下, 与 CTDE 框架和传统基于价值的 DRL 方案相比, 所提 HCJRA 方案在收敛速度、精度及稳定性方面具有显著优势. 该方案不仅能够以更少的干扰资源完成任务, 还能有效应对对抗双方成员数量变化的复杂通信对抗挑战, 展现出良好的扩展性和适应性.

需要指出的是, 当前模型在设计时基于系统复杂度与训练成本的考虑, 对部分参数 (如通信方天线增益、干扰压制系数等) 作了已知且固定的假设, 未将其纳入观测空间与状态建模中. 因而, 在实际通信对抗中, 当这些参数动态变化时, 将导致模型泛化能力下降. 未来的研究我们将聚焦于扩展智能体的观测空间, 适当引入这些动态变化的参数, 在不显著增加模型复杂度的前提下, 探索具备自适应性

的高效 DRL 架构, 以进一步增强模型在动态通信环境下的泛化与迁移能力. 同时, 本研究在满足战术需求的前提下, 采用学界通行做法对波形参数进行合理简化, 未来我们将在现有研究基础上, 结合更为符合实际场景中干扰信号波形参数, 对基于多维度的干扰资源联合资源优化方案展开深入研究. 此外, 目前通信感知一体化 (integrated sensing and communication, ISAC) 技术的突破性发展为动态电磁资源协同提供了新的研究范式, 特别是在多域联合优化方面已取得显著进展 [30]. 该多域联合设计思想可进一步延伸至侦察干扰一体化场景, 通过深入探索多域耦合优化机制, 将为未来多域联合干扰资源分配研究开辟新的方向.

参考文献

1 Wang S F, Bao Y F, Li Y. The architecture and technology of cognitive electronic warfare. Sci Sin Inform, 2018, 48: 1603–1613 [王沙飞, 鲍雁飞, 李岩. 认知电子战体系结构与技术. 中国科学: 信息科学, 2018, 48: 1603–1613]

2 Sharma P, Sarma K K, Mastorakis N E. Artificial intelligence aided electronic warfare systems—recent trends and evolving applications. IEEE Access, 2020, 8: 224761

3 Luong N C, Hoang D T, Gong S, et al. Applications of deep reinforcement learning in communications and networking: a survey. IEEE Commun Surv Tut, 2019, 21: 3133–3174

4 Wang C, Deng D, Xu L, et al. Resource scheduling based on deep reinforcement learning in UAV assisted emergency communication networks. IEEE Trans Commun, 2022, 70: 3834–3848

5 Zhong C, Wang F, Gursoy M C, et al. Adversarial jamming attacks on deep reinforcement learning based dynamic multichannel access. In: Proccedings of IEEE Wireless Communications and Networking Conference (WCNC), 2020. 1–6

6 Peng X, Xu H, Jiang L, et al. A dynamic adaptive jamming power allocation method based on deep reinforcement learning. Acta Electron Sin, 2023, 51: 1223–1234 [彭翔, 许华, 蒋磊, 等. 一种基于深度强化学习的动态自适应干扰功率分配方法. 电子学报, 2023, 51: 1223–1234]

7 Li S, Liu G, Zhang K, et al. DRL-based joint path planning and jamming power allocation optimization for suppressing netted radar system. IEEE Signal Process Lett, 2023, 30: 548–552

8 Pan Z, Li Y, Wang S, et al. Joint optimization of jamming type selection and power control for countering multifunction radar based on deep reinforcement learning. IEEE Trans Aerosp Electron Syst, 2023, 59: 4651–4665

9 Rao N, Xu H, Qi Z, et al. Fast adaptive jamming resource allocation against frequency-hopping spread spectrum in wireless sensor networks via meta-deep-reinforcement-learning. IEEE Trans Aerosp Electron Syst, 2024, 60: 7676–7693

10 Cui Z M, Peng S R, Ren M Q, et al. Research on multi-target interference decision in multi-beam interference system. Fire Control Command Control, 2021, 46: 149–155 [崔哲铭, 彭世蕤, 任明秋, 等. 多波束干扰系统多目标干扰决策研究. 火力与指挥控制, 2021, 46: 149–155]

11 Zhang D, Sun J, Yi W, et al. Joint jamming beam and power scheduling for suppressing netted radar system. In: Proceedings of IEEE Radar Conference (RadarConf21), 2021. 1–6

12 Sun J, Yuan Y, Sabrina Greco M, et al. Coordinated deception jamming power scheduling for multijammer systems against distributed radar systems. Trans Rad Sys, 2024, 2: 1076–1088

13 Liu T T, Luo Y N, Yang C Y. Distributed interference coordination based on multi-agent deep reinforcement learning. J Commun, 2020, 41: 38–48 [刘婷婷, 罗义南, 杨晨阳. 基于多智能体深度强化学习的分布式干扰协调. 通信学报, 2020, 41: 38–48]

14 Wu Z J, Lin Y, Zhang Y J, et al. Multi-agent collaboration based UAV clusters multi-domain energy-saving antijamming communication. Sci Sin Inform, 2023, 53: 2511–2526 [吴志娟, 林艳, 张一晋, 等. 基于多智能体协同的无人机簇群多域节能抗干扰通信. 中国科学: 信息科学, 2023, 53: 2511–2526]

15 Shao Z, Yang H, Xiao L, et al. Deep reinforcement learning-based resource management for UAV-assisted mobile edge computing against jamming. IEEE Trans Mobile Comput, 2024, 23: 13358–13374

16 Zhang Y, Mou Z, Gao F, et al. UAV-enabled secure communications by multi-agent deep reinforcement learning. IEEE Trans Veh Technol, 2020, 69: 11599–11611

17 Rao N, Xu H, Jiang L, et al. Allocation algorithm of distributed cooperative jamming power based on multi-agent deep reinforcement learning. Acta Electron Sin, 2022, 50: 1319–1330 [饶宁, 许华, 蒋磊, 等. 基于多智能体深度强化学习的分布式协同干扰功率分配算法. 电子学报, 2022, 50: 1319–1330]

18 Zhang W, Zhao T, Zhao Z, et al. An intelligent strategy decision method for collaborative jamming based on hierarchical multi-agent reinforcement learning. IEEE Trans Cogn Commun Netw, 2024, 10: 1467–1480

19 Valianti P, Papaioannou S, Kolios P, et al. Multi-agent coordinated close-in jamming for disabling a rogue drone. IEEE Trans Mobile Comput, 2022, 21: 3700–3717
20 Valianti P, Malialis K, Kolios P, et al. Cooperative multi-agent jamming of multiple rogue drones using reinforcement learning. IEEE Trans Mobile Comput, 2024, 23: 12345–12359
21 Su Q, Zhong Y F, Cao Z Q, et al. Target threat assessment model based on operational situation and improved CRITIC-TOPSIS. J Syst Eng Electron, 2023, 45: 2343–2352 [苏倩, 钟元芾, 曹志钦, 等. 基于作战态势和改进 CRITIC-TOPSIS 的目标威胁评估模型. 系统工程与电子技术, 2023, 45: 2343–2352]
22 Liu J, Wang G, Fu Q, et al. Task assignment in ground-to-air confrontation based on multiagent deep reinforcement learning. Defence Tech, 2023, 19: 210–219
23 Song X C, Li Z, Ren H W, et al. Threat-driven resource allocation algorithm for distributed netted phased array radars. J Radar, 2023, 12: 629–641 [宋晓程, 李陟, 任海伟, 等. 目标动态威胁度驱动的分布式组网相控阵雷达资源优化分配算法. 雷达学报, 2023, 12: 629–641]
24 El-Fallah A, Zatezalo A, Mahler R, et al. Unified Bayesian situation assessment sensor management. In: Signal Processing, Sensor Fusion, and Target Recognition XIV. Bellingham: SPIE, 2005. 253–264
25 Lowe R, Wu Y, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), 2017. 6379–6390
26 Chen R, Tan Y. Credit assignment with predictive contribution measurement in multi-agent reinforcement learning. Neural Netws, 2023, 164: 681–690
27 Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Proceedings of the 35th International Conference on Machine Learning (ICML), 2018. 1861–1870
28 Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods. In: Proceedings of the 35th International Conference on Machine Learning (ICML), 2018. 1587–1596
29 Huang C, Chen G, Gong Y, et al. Joint buffer-aided hybrid-duplex relay selection and power allocation for secure cognitive networks with double deep Q-network. IEEE Trans Cogn Commun Netw, 2021, 7: 834–844
30 Zhao J R, Yang P, Xiao Y, et al. A new FMCW-based multi-domain joint modulation dual-function radar communication technology. Sci Sin Inform, 2023, 53: 1802–1821 [赵嘉荣, 杨平, 肖悦, 等. 基于 FMCW 的新型多域联合调制双功能雷达通信技术. 中国科学: 信息科学, 2023, 53: 1802–1821]

Hierarchical cooperative jamming resource allocation scheme based on deep reinforcement learning

Xiao-Rong JING[1,2*], Zhe PENG[1] & Qian-Bin CHEN[2]

1. *School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China*
2. *Chongqing Key Lab of Mobile Communications Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China*
* Corresponding author. E-mail: jingxr@cqupt.edu.cn

Abstract In this study, we propose a hierarchical cooperative jamming resource-scheme based on deep reinforcement learning (DRL) to address the issue of differentiated dynamic jamming resource allocation in any-to-many communication countermeasure scenarios. The proposed scheme includes two stages: beam allocation and power allocation. Initially, a DRL-assisted dynamic threat assessment model is constructed to assess in real-time the degree of threat of targets and dynamically adjust the allocation of jamming beam resources. Next, cooperative jamming power allocation among multiple jamming devices is modeled as a decentralized partially observable Markov decision process (Dec-POMDP); in this process, each jamming device allocates power resources independently based on a beam resource configuration to optimize their distribution in cooperative jamming scenarios. To reduce the dimension of power in the decision-making process, we develop a decentralized training and decentralized execution (DTDE) framework. Finally, we propose a jamming power resource-allocation strategy based on the discrete soft actor-critic algorithm; this algorithm utilizes the maximum entropy reinforcement learning theory to balance investigation and exploitation in the strategy. Additionally, the Dec-POMDP model is optimized to improve the generalizability of the strategy. The simulation results show that the proposed jamming resource-allocation scheme substantially outperforms existing methods in terms of jamming effectiveness and scalability.

Keywords communication countermeasures, target threat assessment, multi-agent deep reinforcement learning, jamming resource allocation, decentralized execution, discrete soft actor-critic