ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 1,Jan. 2024,pp. 15－30
DOI:10. 16337/j. 1004-9037. 2024. 01. 003

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于数字孪生和强化学习的低空智联网协同认知干扰

沈高青[1]，蔡圣所[2]，雷　磊[1,2]，贲　德[2]

(1. 南京航空航天大学公共实验教学部,南京 211106;2. 南京航空航天大学电子信息工程学院,南京 211106)

摘　要：针对低空智联网协同认知干扰决策过程中,多架电子干扰无人机对抗多部多功能雷达的干扰资源分配问题,提出了一种基于数字孪生和深度强化学习的认知干扰决策方法。首先,将协同电子干扰问题建模为马尔可夫决策问题,建立认知干扰决策系统模型,综合考虑干扰对象、干扰功率和干扰样式选择约束,构建智能体动作空间、状态空间和奖励函数。其次,在近端策略优化(Proximal policy optimization, PPO)深度强化学习算法的基础上,提出了自适应学习率近端策略优化(Adaptive learning rate proximal policy optimization, APPO)算法。同时,为了以高保真的方式提高深度强化学习算法的训练速度,提出了一种基于数字孪生的协同电子干扰决策模型训练方法。仿真结果表明,与已有的深度强化学习算法相比,APPO算法干扰效能提升30%以上,所提训练方法能够提高50%以上的模型训练速度。

关键词：多无人机协同;认知干扰决策;多功能雷达;深度强化学习;数字孪生

中图分类号：TP391　　**文献标志码**:A

Cooperative Cognitive Jamming in Low-Altitude Intelligent Network Based on Digital Twin and Reinforcement Learning

SHEN Gaoqing[1], CAI Shengsuo[2], LEI Lei[1,2], BEN De[2]

(1. Public Experimental Teaching Department, Nanjing University of Aeronautics & Astronautics, Nanjing 211106, China;
2. College of Electronic and Information Engineering, Nanjing University of Aeronautics & Astronautics, Nanjing 211106, China)

Abstract: To address the issue of resource allocation for multiple electronic jamming unmanned aerial vehicles (UAVs) against multiple multifunctional radars in the low-altitude intelligent network cooperative cognitive jamming decision-making process, a cognitive jamming decision-making approach based on digital twinning and deep reinforcement learning is proposed. Firstly, a cognitive jamming decision-making system model is established by treating the cooperative electronic jamming problem as a Markov decision process. Considering the constraints related to jamming target, jamming power, and jamming pattern selection comprehensively, the agents' action space, state space, and reward function are constructed. Secondly, an adaptive learning rate proximal policy optimization (APPO) algorithm is proposed based on the proximal policy optimization (PPO) algorithm. Additionally, to enhance the training speed of the deep reinforcement learning algorithm in a high-fidelity manner, a digital twin-based cooperative electronic jamming decision-making model training method is presented. Simulation results demonstrate that

基金项目:国家自然科学基金(62371232)。
收稿日期:2023-12-11;**修订日期**:2024-01-15

compared with existing deep reinforcement learning algorithms, the interference efficiency of the APPO algorithm is improved by more than 30%, and the proposed training method increases the model training speed by more than 50%.

Key words: multi-UAV cooperation; cognitive jamming decision-making; multifunctional radars; deep reinforcement learning; digital twin

引　　言

低空智联网作为一种新兴的智能网络体系架构，依托空天地海网络基础设施建设，未来将为边境侦查、电子干扰、农林植保及交通运输等军用和民用任务提供重要支撑[1]。近年来，无人机因其廉价、隐蔽性高和易部署等优点，被广泛应用于低空智联网协同认知干扰领域[2]。如图1所示，在由多架突防战斗机和电子干扰无人机构成的低空智联网系统中，将多架电子干扰无人机分散在有人机四周伴随飞行，协同干扰敌方雷达系统，是提高对敌干扰效果、保障飞行编队突防成功率的重要手段。多无人机协同电子干扰要求根据战场态势信息和实际作战需求，合理高效地分配干扰对象、干扰功率和干扰样式等干扰资源，避免单干扰机在时间、功率和频率等方面的限制，获得整个系统的最佳干扰效果。因此，研究具备认知能力的智能协同电子干扰方法具有重要意义。

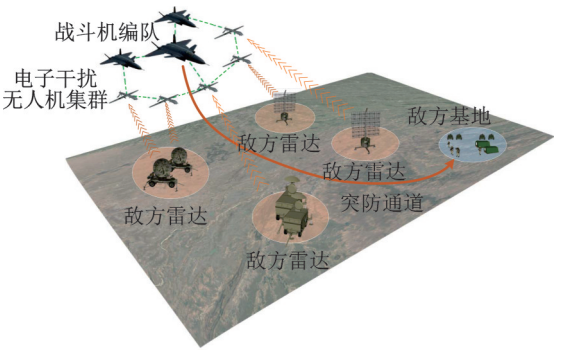

图1　有人/无人机协同突防作战场景

Fig.1　Manned/unmanned aerial vehicle collaborative penetration combat scenario

随着通信技术的发展，多功能体制雷达已成为对抗电子干扰的重要手段，具备波束扫描方向快速变化和灵活的多波束形成能力，能够根据侦察态势结果自适应实现搜索、跟踪、截获和制导等多种功能[3]。依赖于固定干扰样式的传统干扰决策方法决策效率低、准确率差，在面对多功能体制雷达时往往无法快速加载准确的干扰样式，干扰效能大打折扣。

强化学习算法作为一种交互式学习算法，通过智能体与环境之间的不断交互试错，学习积累经验，为马尔可夫决策问题提供更加有效的行为策略。尤其是和深度学习结合之后，即深度强化学习，大大加强了强化学习对复杂问题的求解能力[4-6]。更重要的是，深度强化学习是一种面向策略设计的方法，其策略的输出不依赖于特定的环境，在求解动态不确定性问题时具有突出的优势。多无人机协同电子干扰问题本质上也是一种马尔可夫决策问题，因此深度强化学习在解决动态环境下的多无人机协同电子干扰问题时具有突出的优势。作为一种数据驱动的算法，强化学习需要智能体与环境不断的交互产生大量经验数据从而提高策略性能。然而，就协同电子干扰问题而言，如果采用仿真的手段获取数据，由于电磁环境存在大量的非线性和时变因素，导致获取到的经验数据误差大，模型有效性低；如果采用真实环境来获取经验数据训练智能体，不仅实验成本高、难度大，而且数据获取效率极低，难以满足智能体决策模型的高效训练。如何高效、高保真地获取协同电子干扰经验数据，成为了将深度强化学习应用于多无人机协同电子干扰策略优化问题的关键瓶颈。

数字孪生(Digital twin, DT)作为近年来的新兴技术，为解决上述问题提供了新思路[7-8]。数字孪生被定义为一种多物理量、多尺度及高保真的仿真方式，它能够根据历史数据和实时传感器数据实时地反映物理实体的状态。借助数字孪生，机器学习算法可以轻松获得用于模型训练的真实世界的高保真

状态信息[9]。因此,可以针对多无人机协同电子干扰问题建立一个数字孪生训练环境,并在数据处理中使用深度强化学习算法,为现实世界中的协同电子干扰提供智能认知决策。通过感知真实环境的变化,数字孪生中运行的认知干扰决策模型可以实时更新,从而适应动态环境的变化。换言之,数字孪生可以解决深度强化学习应用于协同电子干扰时的数据获取效率低的问题。然而,如何在深度强化学习的框架中结合数字孪生来解决多无人机协同电子干扰问题还没有得到充分的研究。

本文围绕上述问题展开研究,主要工作和贡献如下:

(1) 建立雷达发现概率模型,分析不同的有源压制干扰样式对雷达发现概率的影响,构建高保真孪生式多无人机协同电子干扰环境。

(2) 将深度强化学习算法应用于多干扰机、多雷达的协同电子干扰策略优化问题中,在近端策略优化(Proximal policy optimization, PPO)算法[10]的基础上,提出一种基于自适应学习率近端策略优化(Adaptive learning rate proximal policy optimization, APPO)算法的多无人机协同电子干扰决策模型。

(3) 针对多无人机协同电子干扰问题,提出了一种基于数字孪生的"集中式训练、分布/集中式执行、持续进化"的深度强化学习算法训练框架。

(4) 仿真结果表明本文提出的APPO算法相较于现有的深度强化学习算法,能够有效提升突防成功率,提出的孪生式训练框架能够有效提高模型的训练速度。

1 相关工作

近年来,越来越多的研究人员开始采用强化学习来解决多功能雷达干扰资源认知决策问题。文献[11]首先根据多功能雷达的信号特点建立多功能雷达信号层级结构,为认知干扰决策体系构建提供基础。然后,作者结合干扰决策体系中的干扰案例库以及干扰有效性分析,阐述了采用强化学习解决认知干扰决策问题的基本思路。文献[12]将Q学习算法应用于多功能雷达认知干扰决策问题中,分析多功能雷达工作状态和干扰样式对雷达工作状态的影响,构建雷达状态转移概率图,通过仿真验证了Q学习算法在雷达认知干扰决策问题中的有效性。文献[13]提出了雷达状态的联合表征方法,分析了雷达状态序列的马尔可夫性,系统阐述了将多功能雷达认知干扰决策问题建模为马尔可夫决策过程的步骤,并通过仿真实验证明了采用强化学习算法能够自主学习出最佳干扰策略。文献[14]研究了多个雷达共同探测目标的场景,建立自适应干扰机实现框架,并采用双Q学习算法实现干扰资源的分配优化,用于缓解Q学习中动作值函数的过估计问题。仿真结果证明了双Q学习在收敛速度和干扰成功概率等指标上比Q学习获得了更好的性能。

随着研究问题复杂性的提升,研究人员开始使用深度强化学习求解干扰资源分配优化问题。文献[15]以机载火控雷达的典型工作模式和干扰样式集为基础,构建了干扰决策矩阵,并提出了一种基于双深度Q网络(Double deep Q network, DDQN)的干扰样式选择算法,仿真实验证明了DDQN在解决过估计问题上的有效性。为了解决DQN算法在雷达干扰资源中存在的低效、不精确问题,文献[16]提出了一种基于竞争双深度Q网络(Dueling double deep Q network, D3QN)的多功能雷达干扰决策方法。D3QN将动作值函数网络用状态值函数网络和优势函数网络共同表示,能够更准确地估计值函数,选择更加合适的动作。仿真结果表明,D3QN完成决策任务的效率是DQN的2.1倍,决策精度比DQN提高了约10%。文献[17]研究了多干扰机对抗多功能组网雷达的场景,提出了一种基于优先经验回放双深度Q网络(Priority experience replay double deep Q network, PER-DDQN)的干扰资源决策算法,并对该算法进行了仿真分析。仿真结果表明,与DQN算法相比,PER-DDQN可以克服数据相关性问题并避免不必要的迭代,更适合稀疏奖励环境。文献[18]提出了一种基于异步优势演员-评论家(Asynchronous advantage actor-critic, A3C)算法的干扰资源决策方法。不同于DQN等基于价值的深度强化

学习算法，A3C能够处理连续动作空间问题，同时利用异步多线程的方式与环境进行交互加快模型训练速度。仿真结果表明，基于A3C的认知干扰决策方法与基于DQN的认知干扰决策算法相比，决策时间降低70%，决策准确性大大提高。

综合上述分析可以发现，目前研究人员使用强化学习解决认知干扰决策问题时考虑的场景多为单干扰机对抗单部多功能雷达的场景。然而，现代战场环境下的电子对抗通常是涉及多干扰机对抗多功能组网雷达系统的场景。因此，本文提出了一种基于深度强化学习的多无人机协同电子干扰决策算法。同时，不同于现有研究只考虑干扰对象和干扰样式的决策问题，本文将干扰功率的选择引入干扰资源决策中，进一步提高了干扰资源决策算法的可行性。

2 系统模型

2.1 雷达发现概率

根据雷达方程，雷达接收到来自目标回波信号的功率可以表示为

$$P_{\mathrm{r}}=\frac{P_{\mathrm{t}}G_{\mathrm{t}}^{2}\sigma\lambda^{2}}{(4\pi)^{3}R^{4}} \tag{1}$$

式中：P_{t}表示雷达发射功率；G_{t}表示雷达天线增益；σ表示探测目标的雷达散射截面积(Radar cross section, RCS)；λ表示雷达发射信号的波长；R表示目标到雷达的径向距离。

当受到干扰时，雷达接收到的干扰信号功率可以表示为

$$P_{\mathrm{rj}}=\frac{P_{\mathrm{j}}G_{\mathrm{j}}G(\theta)\lambda^{2}\xi_{\mathrm{j}}}{(4\pi)^{2}R_{\mathrm{j}}^{2}} \tag{2}$$

式中：P_{j}表示干扰机发射功率；G_{j}表示干扰机天线增益；$G(\theta)$为雷达在干扰机方向上的增益；ξ_{j}表示雷达信号与干扰信号之间的极化适配损失系数；R_{j}表示干扰机到雷达的径向距离。

在干扰机生成干扰信号时其信号带宽Δf_{j}往往要比雷达发射信号带宽Δf_{r}要大，因此只有一部分干扰信号能够被雷达接收机接收。假设干扰信号瞄准雷达发射信号中心频率，雷达接收机具有矩形频率响应，且干扰信号功率谱呈均匀分布，则实际能够进入雷达接收机的干扰功率为

$$P_{\mathrm{rj}}=\frac{P_{\mathrm{j}}G_{\mathrm{j}}G(\theta)\lambda^{2}\xi_{\mathrm{j}}}{(4\pi)^{2}R_{\mathrm{j}}^{2}}\cdot\frac{\Delta f_{\mathrm{r}}}{\Delta f_{\mathrm{j}}} \tag{3}$$

因此，在不考虑脉冲积累的情况下，雷达接收机的信噪比(Signal noise ratio，SNR)为

$$\mathrm{SNR}=\frac{P_{\mathrm{t}}G_{\mathrm{t}}^{2}R_{\mathrm{j}}^{2}\Delta f_{\mathrm{j}}\sigma}{4\pi P_{\mathrm{j}}G_{\mathrm{j}}G(\theta)R^{4}\Delta f_{\mathrm{r}}} \tag{4}$$

得出信噪比之后，即可按照信噪比经验公式推出发现概率。对于非起伏目标而言，雷达发现概率P_{d}、虚警概率P_{fa}和SNR之间的关系，满足如下经验公式[19]

$$\mathrm{SNR}=A+0.12AB+1.7B \tag{5}$$

式中：$A=\ln\left(0.62/P_{\mathrm{fa}}\right)$；$B=\ln\left(P_{\mathrm{d}}/(1-P_{\mathrm{d}})\right)$。重新整理式(5)，可得

$$P_{\mathrm{d}}=\frac{\exp^{\frac{\mathrm{SNR}-A}{0.12A+1.7}}}{1+\exp^{\frac{\mathrm{SNR}-A}{0.12A+1.7}}} \tag{6}$$

2.2 雷达有源压制干扰样式及干扰效果评估

经前文分析可知，多功能雷达处于不同的工作状态时，干扰机需要采取不同的干扰样式，从而获取

最佳的干扰效果。雷达有源干扰样式根据干扰目的不同可以分为压制干扰和欺骗干扰。其中压制干扰的基本原理是在雷达的回波信号中加入干扰信号,使得真实目标回波信号淹没在干扰信号中,降低敌方雷达接收机的信噪比,减弱雷达的检测性能。本文假设电子干扰机采用压制干扰对敌方雷达进行干扰。根据干扰信号调制方式的不同,压制干扰又可进一步分为噪声调幅干扰、噪声调频干扰、灵巧噪声干扰和密集假目标干扰等,其基本干扰原理分别如下。

(1) 噪声调幅干扰

定义广义平稳随机过程式(7)为噪声调幅干扰。

$$J(t)=\left[U_{\mathrm{j}}+U_{\mathrm{n}}(t)\right]\exp\left(\mathrm{j}2\pi f_{\mathrm{j}}t+\mathrm{j}\varphi\right) \tag{7}$$

式中:exp(·)表示以自然常数e为底的指数函数;U_{j}、f_{j}均为常数,分别表示信号的幅度和中心频率;$U_{\mathrm{n}}(t)$表示均值为0、方差为σ_{n}^2、分布区间为$[-U_{\mathrm{j}},\infty)$的广义平稳随机过程;初始相位φ为$[0,2\pi)$上均匀分布的随机变量,且与$U_{\mathrm{n}}(t)$相互独立。

(2) 噪声调频干扰

噪声调频干扰信号的时域表达式为

$$J(t)=U_{\mathrm{j}}\exp\left(\mathrm{j}2\pi\left(f_{\mathrm{j}}t+k_{\mathrm{j}}\int_0^t u(t)\,\mathrm{d}t\right)+\mathrm{j}\varphi\right) \tag{8}$$

式中:k_{j}为常数,表示噪声调频干扰的调频斜率;$u(t)$表示均值为0、方差为σ_{n}^2的高斯噪声。

(3) 灵巧噪声干扰

假定雷达发射信号为线性调频信号$s(t)$,干扰机产生的噪声调制信号为$n(t)$,则经过卷积调制后产生的干扰信号为

$$J(t)=s(t)\otimes n(t) \tag{9}$$

式中$\otimes$表示卷积。

干扰信号进入雷达接收机后,会经过匹配滤波处理,脉冲压缩后雷达信号处理机得到的输出为

$$J_{\mathrm{pc}}(t)=s(t)\otimes n(t)\otimes s^*(-t) \tag{10}$$

式中$s^*(t)$表示$s(t)$匹配滤波器的响应函数。假定$J_{\mathrm{pc}}(t)$、$s(t)$、$n(t)$的频谱函数分别为$J_{\mathrm{pc}}(f)$、$S(f)$、$N(f)$,则可以得出

$$J_{\mathrm{pc}}(f)=N(f)\left|S(f)\right|^2 \tag{11}$$

对频谱函数$J_{\mathrm{pc}}(f)$进行傅里叶逆变换,得到时域输出为

$$J_{\mathrm{pc}}(t)=F^{-1}\left[\left|S(f)\right|^2\right]\otimes n(t) \tag{12}$$

式中:F^{-1}表示傅里叶逆变换函数;$F^{-1}[|S(f)|^2]$表示点扩展函数。当雷达发射信号为线性调频信号时,$F^{-1}[|S(f)|^2]$为sinc函数。也就是说,噪声信号$s(t)$与$F^{-1}[|S(f)|^2]$卷积时同样会获得脉冲压缩处理增益。同噪声调幅干扰和噪声调频干扰相比,灵巧噪声干扰对带有脉冲压缩的雷达干扰效果更好。

(4) 密集假目标干扰

密集假目标干扰样式的基本原理是干扰机发射与真实目标回波信号相仿的多假目标信号,与真实目标回波信号一同进入雷达接收机,形成一系列假目标脉冲,影响敌方雷达对真实目标的探测。

本文采取的密集假目标干扰实现方式通过干扰机对已复制的多个假目标信号在时域上进行延迟并相互叠加,叠加后的假目标信号进入雷达接收机后再经过匹配滤波处理,得到脉冲宽度更小的回波信号,从而增大假目标的密集程度。

本文所考虑的干扰性能指标为雷达发现概率P_{d}。根据前文的介绍可知,在雷达虚警概率P_{fa}固定的前提下,信干比越大,发现概率P_{d}越大。事实上,在干扰功率相同的前提下,不同干扰样式的干扰效果与产生的干扰信号能够获得的雷达信号处理增益成反比。干扰样式获得的处理增益越大,雷达信号处理机最终能够获得的信干比也越大,雷达的发现概率越高,干扰效果越差。假设雷达脉冲信号脉宽为τ,带宽为Δf_r,噪声调幅干扰的功率谱密度为$G_{\mathrm{n}}(f)$,噪声调频干扰的有效调频带宽为f_{de},有效调频指数为m_{fe},灵巧噪声卷积干扰的视频噪声时宽为T_{n},密集假目标干扰脉冲占空比为β,脉冲积累数为n,表1总结了噪声调幅干扰、噪声调频干扰、灵巧噪声干扰和密集假目标干扰4种干扰样式在不同信号处理方式下信干比的增益大小[20]。

表1 不同信号处理方式对压制干扰信干比的影响

Table 1 Impact of different signal processing methods on the signal-to-interference ratio of suppression jamming

干扰样式	中频滤波	脉冲压缩	非相干积累	相干积累
噪声调幅干扰	$\dfrac{U_{\mathrm{j}}^2+R_{\mathrm{n}}(0)}{U_{\mathrm{j}}^2+\int_{\Delta f-\frac{\Delta f_{\mathrm{r}}}{2}}^{\Delta f+\frac{\Delta f_{\mathrm{r}}}{2}}G_{\mathrm{n}}(f)\mathrm{d}f}$	$\Delta f_{\mathrm{r}}\tau$	$\dfrac{\sqrt{n}}{\eta}$	$\dfrac{n^{0.8}}{\eta}$
噪声调频干扰	$\begin{cases}\dfrac{\sqrt{2\pi}\,f_{\mathrm{de}}}{f_{\mathrm{r}}} & m_{\mathrm{fe}}\gg 1\\ \dfrac{\pi f_{\mathrm{de}}^2}{2\Delta F_{\mathrm{n}}\Delta f_{\mathrm{r}}} & m_{\mathrm{fe}}\ll 1\end{cases}$	$\Delta f_{\mathrm{r}}\tau$	$\sqrt{n}$	$n^{0.8}$
灵巧噪声干扰	1	$\dfrac{\Delta f_{\mathrm{r}}\tau T_{\mathrm{n}}+\tau}{\tau+T_{\mathrm{n}}}$	$\sqrt{n}$	$n^{0.8}$
密集假目标干扰	1	$1/\beta^2$	1	1

表中η表示伪随机噪声的质量因素,其取值规则如下

$$\eta=\begin{cases}0.1 & \Delta f_{\mathrm{j}}/\Delta f_{\mathrm{r}}\approx 1\\ 0.2 & \Delta f_{\mathrm{j}}/\Delta f_{\mathrm{r}}\gg 1\end{cases}\tag{13}$$

从表1中可以看出,在干扰功率相同的前提下,噪声调幅、噪声调频、灵巧噪声和密集假目标几种干扰样式的压制干扰能力逐渐降低,但实现难度和对电磁态势感知能力的要求越来越高。

2.3 多功能雷达工作模式

多功能雷达具备波束扫描方向快速变化和灵活的多波束形成能力,能够根据侦察态势结果自适应地调整自身的工作模式,实现搜索、跟踪、截获和制导等多种功能。

假定我方有N架无人机需要掩护目标穿过由M部多功能雷达组成的敌方侦察系统。无人机可供选择的干扰方式包括噪声调幅干扰、噪声调频干扰、灵巧噪声干扰和密集假目标干扰4种。多功能雷达的工作状态包括搜索、跟踪、截获和制导4种,其中搜索模式是初始状态,制导模式是终止状态。每隔一段时间,无人机选择一种干扰方式对敌方雷达进行干扰,雷达的工作状态也会根据对目标的探测结果进行转变。

假设多功能雷达工作状态记为$S_{\mathrm{r,m}}$,当雷达处于搜索、跟踪、截获和制导工作状态时,$S_{\mathrm{r,m}}$分别取值1,2,3,4。如图2所示,多功能雷达的工作状态转变遵循如下规则:

(1) 搜索状态下,如果5次探测中至少有2次探测到目标,雷达进入跟踪状态;否则,保持搜索状态。

(2) 跟踪状态下,如果3次探测中至少有1次探测到目标,雷达进入截获状态;若3次探测,雷达均未探测到目标,雷达返回搜索状态。

(3) 截获状态下,如果2次探测中至少有1次探测到目标,雷达进入制导状态;如果连续2次均未探测到目标,雷达返回跟踪状态。

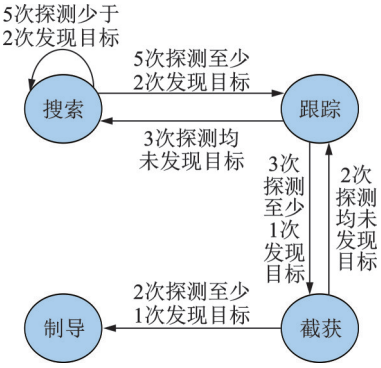

图2 多功能雷达工作状态转换规则

Fig.2 Transition rules of multifunctional-radar working state

3 基于深度强化学习的协同电子干扰方法

3.1 强化学习问题建模

在经典强化学习中,通常将要解决的问题描述为一个马尔可夫决策过程(Markov decision process, MDP)。MDP基于一组对象构建,即智能体与环境。MDP满足马尔可夫性,也就是智能体的下一个状态,仅取决于智能体的当前状态和智能体选择的动作。如果将电子干扰无人机看作智能体,将无人机感知到的电磁环境信息看作状态,无人机执行的干扰策略看作动作,则协同电子干扰问题也满足马尔可夫性,电磁环境的下一个状态仅取决于无人机的当前状态和无人机选择的动作。因此,可以使用强化学习来解决协同电子干扰问题。

MDP通常用元组$(\boldsymbol{S},\boldsymbol{A},\boldsymbol{P}^{a}_{s,s'},\boldsymbol{R}^{a}_{s,s'})$定义,其中$\boldsymbol{S}$为状态集合,称为状态空间;$\boldsymbol{A}$为动作集合,称为动作空间;$\boldsymbol{P}^{a}_{s,s'}$表示执行操作$\boldsymbol{a}$后,状态$\boldsymbol{s}$转换为状态$\boldsymbol{s}'$的概率;$\boldsymbol{R}^{a}_{s,s'}$表示执行动作$\boldsymbol{a}$,从状态$\boldsymbol{s}$转换为状态$\boldsymbol{s}'$后的及时奖励。

对于多无人机协同电子干扰而言,如果无人机的干扰决策指令均由有人机发出,无人机仅作为执行干扰决策的载体,则可将有人机看作智能体,将上述问题建模为马尔可夫决策过程。

此时,状态$\boldsymbol{s}$定义为

$$\boldsymbol{s}=\left[\boldsymbol{S}_r,\boldsymbol{d},\boldsymbol{q},\boldsymbol{a}_{\text{pre}}\right] \tag{14}$$

式中:$\boldsymbol{S}_r=[S_{r,1},S_{r,2},\cdots,S_{r,M}]$表示$M$部雷达当前时刻的工作状态;$\boldsymbol{d}=[d_1,d_2,\cdots,d_M]$表示$N$架无人机分别与$M$部雷达当前时刻的距离;$\boldsymbol{q}=[q_1,q_2,\cdots,q_M]$表示$M$部雷达当前时刻的威胁系数;$\boldsymbol{a}_{\text{pre}}$表示上一时刻采取的干扰策略。

动作$\boldsymbol{a}$定义为

$$\boldsymbol{a}=\left[m_1,m_2,\cdots,m_N,P_1,P_2,\cdots,P_N,f_1,f_2,\cdots,f_N\right] \tag{15}$$

当每架无人机的干扰样式固定时,干扰样式将不再作为动作选项之一,动作退化为

$$\boldsymbol{a}'=[m_1,m_2,\cdots,m_N,P_1,P_2,\cdots,P_N] \tag{16}$$

合理的设置奖励函数可以帮助智能体在与环境交互的过程中快速学习到有效策略。结合协同电子干扰的目标,本文从5个方面设置奖励函数:干扰有效性奖励、干扰功率奖励、干扰样式奖励、突防奖励和干扰覆盖奖励,分别定义如下:

(1) 干扰有效性奖励:此奖励函数为协同电子干扰的直接目标,即尽可能使得敌方雷达处于低威胁的工作状态。雷达的工作状态升高,予以惩罚;雷达的工作状态降低,则予以奖励。因此,t时刻的干扰有效性奖励定义如下

$$r_{1,t}=\sum_{m=1}^{M}\begin{cases}-20 & S^{t}_{r,m}=3,S^{t-1}_{r,m}=2\\ -5 & S^{t}_{r,m}=S^{t-1}_{r,m}+1,S^{t-1}_{r,m}<2\\ 5 & S^{t}_{r,m}=S^{t-1}_{r,m}-1,0<S^{t-1}_{r,m}<3\\ 0 & S^{t}_{r,m}=S^{t-1}_{r,m}=0\end{cases} \tag{17}$$

(2) 干扰功率奖励:此奖励函数主要用于引导电子干扰无人机尽可能用较小的干扰功率实施电子干扰从而减小干扰资源消耗,提高干扰能效比。因此,t时刻的干扰功率奖励定义为

$$r_{2,t}=\frac{3}{N}\sum_{n=1}^{N}\left(\frac{P_{\max}-P_n}{P_{\max}-P_{\min}}-0.5\right) \tag{18}$$

式中$P_{\max}$和$P_{\min}$分别表示最大和最小干扰功率。

(3) 干扰样式奖励:干扰样式编号从0到3分别表示噪声调幅、噪声调频、灵巧噪声和密集假目标干扰,其干扰实现难度越来越大,干扰机的复杂程度也越来越高。为了尽可能采用低复杂度的干扰样式完成干扰任务,降低干扰机复杂度,t时刻的干扰样式奖励定义为

$$r_{3,t}=\frac{3}{N}\sum_{n=1}^{N}\left(\frac{f_{\max}-f_n}{f_{\max}-f_{\min}}-0.5\right) \tag{19}$$

式中$f_{\max}$和$f_{\min}$分别表示可用干扰样式的最大和最小编号。若每架电子干扰无人机的干扰样式固定,则此项为0。

(4) 突防奖励:此奖励函数是协同电子干扰的最终目标,即尽可能地突破敌方雷达系统,掩护突防编队朝目标位置行进。为了引导无人机获得更多的突防奖励,每过一个时间步就予以一个正奖励。因此,t时刻的突防奖励定义为

$$r_{4,t}=5 \tag{20}$$

(5) 干扰覆盖奖励:为了确保协同电子干扰的效果,应该在每一个时间步尽可能的提高敌方雷达被干扰的比例。假设t时刻共有J_t部雷达被干扰,则干扰覆盖奖励定义为

$$r_{5,t}=2J_t/M \tag{21}$$

综上所述,时间步t的完整奖励函数定义为

$$r_t=\omega_1 r_{1,t}+\omega_2 r_{2,t}+\omega_3 r_{3,t}+\omega_4 r_{4,t}+\omega_5 r_{5,t} \tag{22}$$

式中ω_1~ω_5表示权重因子,取值范围为[0,1]。

3.2 基于深度强化学习的协同电子干扰决策算法

APPO算法属于基于策略梯度的强化学习算法。策略梯度法的基本思想是策略π接受环境的输入$\boldsymbol{s}$,输出动作的概率分布,并在动作的概率分布中进行采样得到动作$\boldsymbol{a}$,智能体在环境中执行动作$\boldsymbol{a}$,得到回报r并进入下一个状态。假设用π_θ表示带有参数θ的策略,$J(\pi_\theta)$表示该策略下智能体的长期累积回报,则策略梯度法的梯度更新公式为

$$\nabla_\theta J(\pi_\theta)=\underset{\tau\sim\pi_\theta}{E}\left[\nabla\log\pi_\theta(\boldsymbol{a}|\boldsymbol{s})A^{\pi_\theta}(\boldsymbol{s},\boldsymbol{a})\right] \tag{23}$$

式中:τ表示智能体的轨迹;$A^{\pi\theta}$表示当前策略的优势函数;$E[\cdot]$表示数学期望。

策略梯度法是一种同策略深度强化学习方法,即收集样本的行为策略和需要被学习更新的目标策略必须为同一策略。使用策略梯度法收集到一批样本,利用这些样本更新策略之后,上述收集到的样本就不再能够使用了,必须使用更新后的策略$\pi_{\theta'}$重新收集样本数据,造成策略梯度法数据利用率极低。为了解决上述问题,继续使用老样本数据进行策略学习;为提高样本利用率,研究人员提出使用重要性采样来对上述过程进行修正。假设更新前的策略为$\pi_{\theta'}$,更新后的策略为π_θ,则使用旧数据表示的梯度为

$$\nabla_\theta^{\theta'}J(\pi_\theta)=\underset{\tau\sim\pi_{\theta'}}{E}\left[\frac{\pi_\theta(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\theta'}(\boldsymbol{a}|\boldsymbol{s})}A^{\pi_{\theta'}}(\boldsymbol{s},\boldsymbol{a})\nabla\log\pi_\theta(\boldsymbol{a}|\boldsymbol{s})\right] \tag{24}$$

此时,损失函数可以表示为

$$L(\theta)=\underset{\tau\sim\pi_{\theta'}}{E}\left[\frac{\pi_\theta(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\theta'}(\boldsymbol{a}|\boldsymbol{s})}A^{\pi_{\theta'}}(\boldsymbol{s},\boldsymbol{a})\right] \tag{25}$$

尽管经过修正后的梯度公式能够改善数据的相关性,但当新旧策略分布的差距较大时,最终会导致对梯度期望估计不准确。因此,必须对新旧策略之间的差异进行约束。PPO算法使用了裁剪操作对策略的更新幅度进行约束,其损失函数定义为

$$L(\theta)_{\text{PPO}}=\underset{\tau\sim\pi_{\theta'}}{E}\left[\min\left(\frac{\pi_\theta(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\theta'}(\boldsymbol{a}|\boldsymbol{s})}A^{\pi_{\theta'}}(\boldsymbol{s},\boldsymbol{a}),\text{clip}\left(\frac{\pi_\theta(\boldsymbol{a}|\boldsymbol{s})}{\pi_{\theta'}(\boldsymbol{a}|\boldsymbol{s})},1-\varepsilon,1+\varepsilon\right)A^{\pi_{\theta'}}(\boldsymbol{s},\boldsymbol{a})\right)\right] \tag{26}$$

式中:clip(·)表示截断函数;ε表示一个很小的超参数,用来衡量新旧分布之间的差距。式(26)将新旧策略之间的比值限制在了(1−ε,1+ε)之间,保证了策略更新的幅度不会太大,大大简化了计算量,提高了学习效率。

然而,PPO算法经常会由于学习率参数α设置不合适导致策略陷入局部最优,无法收敛到全局最优策略。因此,本文在PPO算法的基础上添加了自适应学习率的约束,提出了APPO算法,进一步提高了学习效率。假设t表示当前仿真步长,T表示仿真总步长,APPO算法的参数更新遵循如下规则

$$\theta_{\text{new}}=\theta-\alpha\left(1-\frac{t}{T}\right)L_{\text{PPO}}(\theta|\theta') \tag{27}$$

本文将APPO算法应用于多无人机协同电子干扰任务,决策过程如图3所示。首先,无人机通过传感器感知敌方雷达工作状态以及自身位置等环境状态信息$\boldsymbol{s}$;然后,将状态信息输入到APPO算法的策略网络中,生成无人机的干扰样式$\boldsymbol{a}$;无人机执行干扰样式$\boldsymbol{a}$,我方进行干扰效果评估后获得环境奖励r并观测下一时刻的环境状态信息$\boldsymbol{s}'$。不断执行上述过程,直到一个完整的训练回合结束后,将轨迹数据$(\boldsymbol{s}_0,\boldsymbol{a}_0,\boldsymbol{r}_0,\boldsymbol{s}_1,\cdots)$用于更新APPO算法的策略网络和价值网络,循环往复,直到达到终止条件。

APPO算法本质上是一种单智能体算法。就多无人机协同电子干扰任务而言,当有人机与无人机之间的通信畅通时,采用APPO算法执行协同电子干扰任务是可行的。但另一方面,在日益复杂的作战背景下,必须考虑当有人机无法实时获取集群无人机状态信息时,协同电子干扰的决策问题。此时,为了提高系统的鲁棒性,电子干扰无人机还应当具备分布式协同干扰的能力。每个无人机可根据局部观测信息,采用部署在本地的决策模型分布式输出干扰策略,保证干扰有效性。

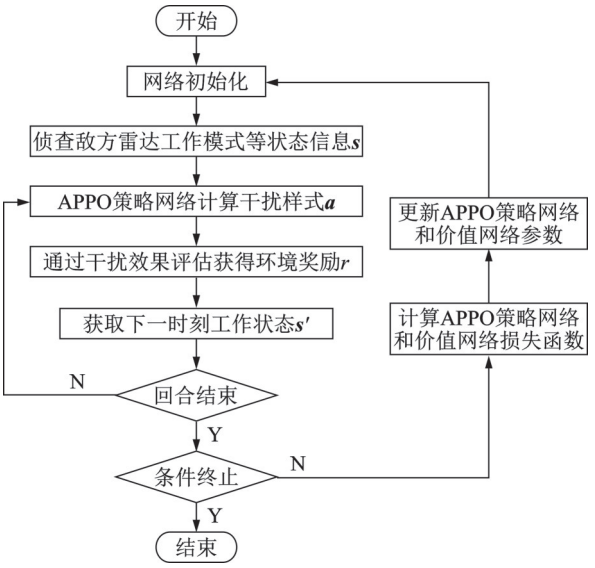

图3 基于APPO算法的认知干扰决策流程

Fig.3 Cognitive jamming decision-making process based on APPO algorithm

多智能体近端策略梯度(Multi-agent proximal policy optimization, MAPPO)算法[21]是一种基于PPO算法的多智能体强化学习的算法,可以有效地解决多智能体博弈中的合作和对抗问题。MAPPO算法的网络结构与APPO算法相同,其主要区别在于MAPPO算法价值网络输入为所有智能体的联合观测,并遵循“集中式训练、分布式执行”的框架。在训练阶段,价值网络将全局状态信息作为输入用于评价每个智能体的策略网络;在执行阶段,每个智能体只需

根据自身的局部观测,即可生成行为策略。

3.3 基于数字孪生的训练执行框架

如前所述,借助数字孪生仿真技术,可以在信息域建立高保真的物理实体孪生镜像,从而获取更加真实的环境状态信息,用于深度强化学习模型训练,提高模型有效性。如图4所示,本文在建立多无人机协同电子干扰数字孪生系统的基础上,以多智能体深度强化学习算法的训练过程为例,为认知干扰决策模型提出了一个数字孪生驱动的“集中式训练、分布/集中式执行、持续进化”的训练执行框架。

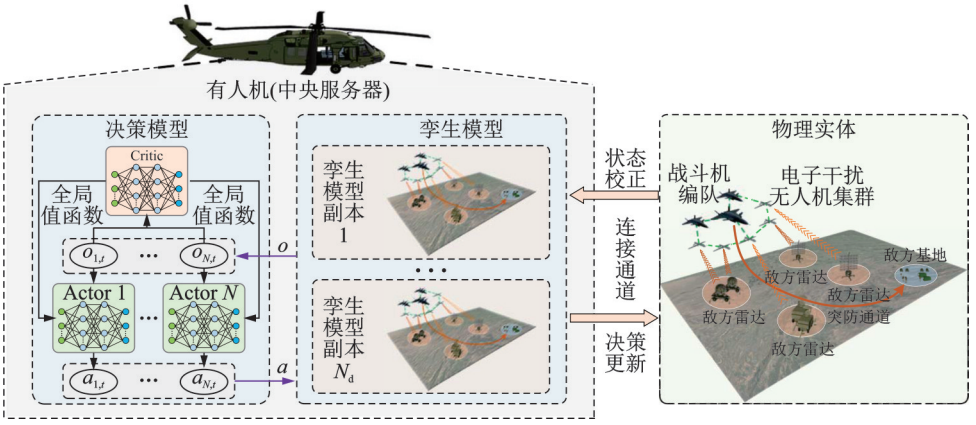

图4 基于数字孪生的认知干扰决策模型训练执行框架

Fig.4 Training and execution framework of the cognitive jamming decision-making model based on DT

多无人机协同电子干扰数字孪生系统由物理实体、孪生模型、决策模型和连接通道组成,其中孪生模型和决策模型均运行于计算资源更加丰富的中央服务器(有人机)中,确保满足孪生模型和决策模型的仿真和决策实时性要求。具体来说,多无人机协同电子干扰数字孪生系统各部分的定义分别如下:

物理实体:电子干扰无人机编队、突防战斗机编队、敌方雷达系统以及电磁环境共同构成物理实体。其中电子干扰无人机能够实施不同的干扰样式,并且具备一定的计算能力,能够完成本地化的认知干扰决策。突防战斗机具备更强的计算能力,可以通过机间数据链向电子干扰无人机编队发送协同干扰指令。

孪生模型:物理实体的高保真镜像称为孪生模型。中央服务器利用仿真和建模技术构建电磁环境孪生模型。为了提高认知干扰决策模型的样本数据获取效率,中央服务器共构建了N_d个孪生模型副本。每个孪生模型副本独立运行,互不干扰。在连接通道正常的条件下,孪生模型定期更新校正后的物理实体参数和状态。

决策模型:多无人机协同电子干扰的控制中心称为决策模型。决策模型既可以是单智能体强化学习算法PPO和APPO,也可以是多智能体强化学习算法MAPPO。决策模型的输入是多个孪生模型副本采集到的样本数据集合,输出是智能体在下一时刻的干扰策略。

连接通道:物理实体和孪生模型之间的通信连接称为连接通道。连接通道可以通过多种方式建立,如5G、自组网或卫星等。连接通道的质量将直接影响孪生模型的保真度和决策的实时性。为了便于分析,本文认为连接通道是完美的。

在训练阶段,多个孪生模型副本同时运行,每个孪生模型副本中的智能体均接受来自决策模型的输出作为认知干扰策略。此时,无论智能体是无人机还是有人机,认知干扰策略都是由一个共同的决策模型输出,因此训练阶段是集中式的。在执行阶段,将训练好的决策模型部署到物理实体上。如果

采用单智能体强化学习求解认知干扰决策问题,则将决策模型部署到有人机上,此时执行方式为集中式。否则,将决策模型部署到无人机上,执行方式变为分布式。因此,执行阶段既可以是集中式,也可以是分布式。同时,在连接通道畅通的基础上,孪生模型将持续不断地获取最新的物理实体样本数据,继续对中央服务器中的决策模型进行训练。每隔一段时间后,再将更新后的决策模型通过连接通道重新部署到智能体上,实现决策模型的持续进化。综上,数字孪生驱动的认知干扰决策模型训练执行框架具有"集中式训练、分布/集中式执行、持续进化"的特点。

4 仿真结果与分析

4.1 仿真环境与参数设置

为了测试和评估基于深度强化学习的认知干扰决策模型的性能,图5展示了本文设定的仿真场景。假定任务区域为边长100 km的方形区域,突防编队沿着底部中线位置穿过任务区域。4部敌方雷达随机布设在航线两侧(其中灰色圆圈表示雷达的最大探测距离)。假定掩护目标仅有一个,其RCS为$\sigma=10\ \mathrm{m}^2$,6架电子干扰无人机的位置与掩护目标重合。突防编队的速度为100 m/s,每隔10 s,突防编队进行一次认知干扰决策。

雷达发射机和接收机天线增益$G_t=40$ dB,发射信号波长$\lambda=0.1$ m,噪声系数$F=3$ dB,脉冲积累数$n=32$,虚警概率$P_{fa}=10^{-6}$。雷达编号为1~4,其位置及发射功率如表2所示。干扰机的干扰功率共有10档可以选择,从10~50 W按照干扰机的编号呈等差数列分布。干扰样式共有4种选择,分别为噪声调幅、噪声调频、灵巧噪声和密集假目标干扰。

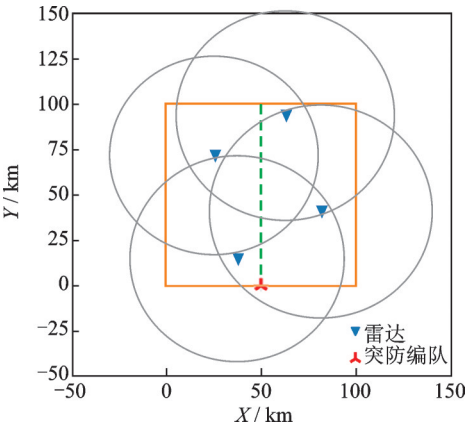

图5 有人/无人机协同突防仿真场景

Fig.5 Manned/unmanned aerial vehicle collaborative penetration simulation scenario

训练认知干扰决策模型所用的深度强化学习算法均是基于Python3.6和PyTorch1.5.1实现的。突防编队的初始位置为任务区域的底部中线处,每一幕中智能体的最大决策步数为100步。模型训练中所用到的其他仿真参数如表3所示。

表2 雷达位置及发射功率参数信息

Table 2 Parameter information of radar position and transmit power

雷达编号	X轴坐标/km	Y轴坐标/km	发射功率/kW
1	26	72	300
2	38	15	333
3	63	94	367
4	82	41	400

表3 仿真参数

Table 3 Simulation parameters

参数名	参数值	参数名	参数值
学习率α	0.000 7	折扣因子γ	0.99
隐藏层神经元数量	64	权重因子ω_1~ω_5	1
训练线程	1	激活函数	ReLU
孪生副本数量	128	每轮训练批次	10

4.2 训练结果分析

对基于深度强化学习的认知决策模型训练结果的分析非常重要。通过分析可以看出不同算法的收敛速度和收敛性能。通常,研究人员希望算法的收敛速度越快越好,收敛性能越高越好。为了测试本文所提算法的有效性,本节分别对比了当电子干扰无人机干扰样式固定和可选时,APPO算法和PPO算法的模型收敛结果。此外,本节还同时测试了MAPPO算法在相同环境下的收敛性能,以验证本文所设计的状态空间及奖励函数的合理性。

4.2.1 干扰样式可选时训练结果收敛性分析

电子干扰机干扰样式可选时，PPO、MAPPO、APPO三种算法的训练结果如图6所示，此时智能体的动作空间包含干扰样式选择。图6(a)展示了PPO、MAPPO和APPO三种算法的平均幕奖励变化趋势。可以看到，PPO算法和APPO算法相较于MAPPO算法收敛速度更快，且APPO算法的收敛值最高。图6(b)展示了3种算法的平均干扰样式奖励随训练步数的变化情况，其趋势基本与平均幕奖励一致。在模型收敛时，APPO算法获得的平均干扰样式奖励比MAPPO和PPO算法提升50%以上。如图6(c)所示，3种算法最终的平均突防成功率均在95%左右。换句话说，APPO算法用更低的干扰代价实现了与MAPPO和PPO算法相同的干扰效果。

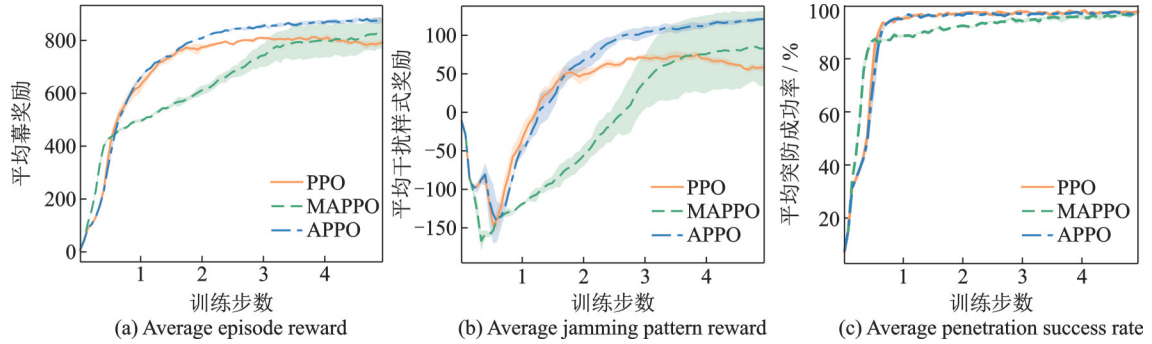

图6 干扰样式可选时认知决策模型训练结果

Fig.6 Cognitive decision-making model training results with optional jamming pattern

4.2.2 干扰样式固定时训练结果收敛性分析

电子干扰机干扰样式固定时，PPO、MAPPO、APPO三种算法的训练结果如图7所示，此时智能体的动作空间仅包含干扰对象选择和干扰功率选择。假设6部干扰机的编号为1～6，干扰样式分别为[0,1,2,2,3,3]。如图7(a)所示，PPO算法和APPO算法的平均幕奖励收敛值要明显高于MAPPO算法，且APPO算法的收敛值略高于PPO算法。这是因为PPO和APPO均是单智能体算法，其输入为全局状态观测，决策相对容易，而MAPPO算法为多智能体算法，输入为智能体局部观测，其决策难度更大。图7(b)展示了3种算法的平均干扰功率奖励变化趋势。可以看到，APPO算法获得了最高的平均干扰功率奖励，相较于MAPPO和PPO算法，提升幅度超30%。图7(c)展示了3种算法的突防成功率变化趋势，在模型稳定时，APPO算法和PPO算法有将近100%的突防成功率，而MAPPO算法突防成功率只有90%左右。

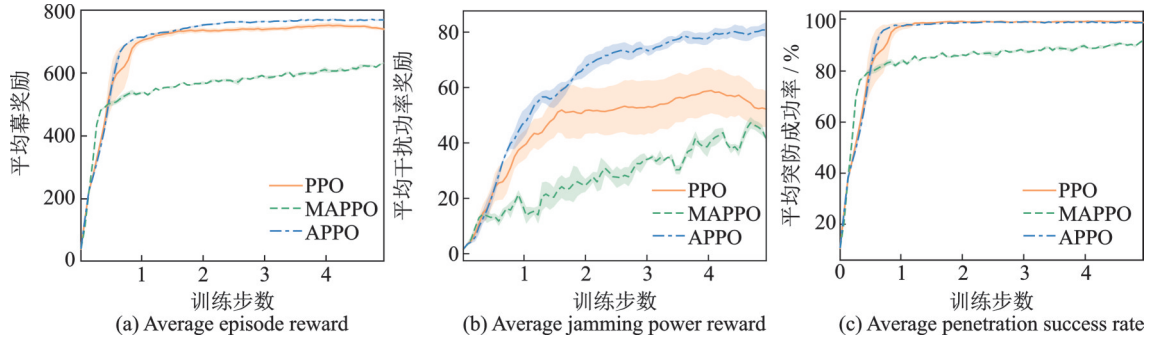

图7 干扰样式固定时认知决策模型训练结果

Fig.7 Cognitive decision-making model training results with fixed jamming pattern

综合图6和图7的仿真结果可以得出如下结论:在电子干扰无人机能够与突防目标保持连接的情况下,将协同电子干扰问题看作为一个单智能体问题,通过突防目标收集环境信息进行集中决策,能够获得更好的干扰效果;当电子干扰无人机和突防目标之间的连接不可靠,只能依赖电子干扰无人机独立决策时,将协同电子干扰问题看作为多智能体问题,采用分布式决策时,在本文所设计的协同电子干扰状态空间和奖励函数的保障下,分布式决策算法MAPPO也能达到将近90%的突防成功率。因此,本文提出的认知干扰决策模型能够胜任复杂环境下的认知干扰决策问题。

4.3 超参数设置分析

为了探究不同超参数对APPO算法性能的影响,本节选择了孪生副本数量N_d、学习率α以及折扣因子γ作为测试对象,平均幕奖励作为测试指标,进行了一系列性能对比实验,仿真结果如图8和表4所示。

图8测试了不同孪生副本数量条件下,基于APPO算法的认知干扰决策而模型训练完成时间随孪生副本数量的变化趋势。可以看到,随着孪生副本数量的提升,模型训练完成所需的时间越来越短,证明了本文所提训练框架的有效性。孪生模型副本数量为64时所需的训练时长比孪生副本数量为16时的训练时长缩短了约36%。但随着孪生模型副本数量进一步增加到128,相比于孪生模型副本数量64时,训练时长仅缩短了20%左右。这是因为模型训练所需时长不仅取决于获取样本数据的时长,还取决于深度神经网络权重参数的更新时长。提高孪生模型副本数量仅能缩短获取样本数据的时长,当获取样本数据时长比深度神经网络参数更新时长小很多时,提升孪生模型副本数量的收益将迅速下降。

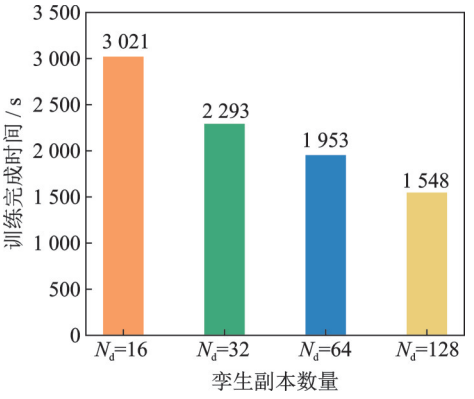

图8 孪生副本数量对APPO算法性能的影响

Fig.8 Impact of the number of DT threads on performance of the APPO algorithm

表4 学习率和折扣因子对APPO算法性能的影响

Table 4 Impact of learning rate and discount factor on performance of the APPO algorithm

γ	α							均值
	6×10^{-4}	7×10^{-4}	8×10^{-4}	9×10^{-4}	1×10^{-3}	2×10^{-3}	3×10^{-3}	
0.70	807.1	824.0	815.8	820.4	824.9	811.1	750.9	807.7
0.75	821.6	825.6	835.5	825.2	816.9	794.1	758.2	811.0
0.80	832.0	815.3	833.5	823.2	835.8	817.6	763.0	817.2
0.85	824.6	821.4	816.2	829.9	820.5	807.1	760.5	811.5
0.90	812.3	826.6	831.5	836.4	833.6	807.0	596.4	792.0
0.95	705.7	703.9	688.1	683.5	659.0	618.9	583.6	663.3
0.99	702.9	702.2	661.0	677.2	632.1	611.8	556.6	649.1
均值	786.6	788.4	783.1	785.1	774.7	752.5	681.3	—

表4展示了不同学习率和折扣因子的取值组合对APPO算法性能的影响。学习率的大小与每一轮训练模型的更新幅度成正相关,折扣因子的大小与智能体对未来奖励的关心程度成正相关。观察表4的最后1行可以看到,随着学习率由6×10^{-4}增加到3×10^{-3},平均幕奖励的均值总体呈先增大后减小的

趋势。这是因子学习率过小容易导致模型陷入局部最优,而学习率过大容易导致模型在最优值附近徘徊。观察表4最后1列可以看到,随着折扣因子由0.7增加到0.99,平均幕奖励的均值同样呈先增大后减小的趋势。这是因为折扣因子越大,智能体对未来的考虑就越深远,能够获得更准确的动作值函数估计,从而获得更高的奖励期望,但过高的折扣因子也会导致模型训练难度增大,无法学习到有效策略,反而使得奖励期望降低。在本次测试实验中,最佳学习率与折扣因子的组合为[9×10^{-4}, 0.9]。

4.4 干扰效果分析

为了进一步分析基于APPO算法的认知干扰决策模型的具体干扰效果,本节使用训练好的模型测试在电子干扰无人机干扰样式固定时,多无人机协同突防过程中干扰策略的变化情况。

图9展示了协同突防过程中雷达与目标距离随突防目标行进距离的变化情况。可以看到,随着目标行进距离的变化,4部雷达与目标之间的距离均是先减小后增大,并且分别在70、20、90和40 km左右时达到最小值,其中2号雷达和3号雷达与目标之间的最小距离均小于20 km,发现目标的概率较大,属于高威胁干扰对象。

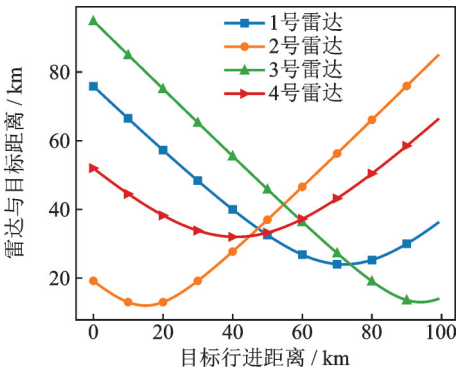

图9 雷达与目标距离随突防目标行进距离变化情况

Fig.9 Distance between the radar and the target changes with the travel distance of the penetration target

图10展示了协同突防过程中干扰策略的变化情况。图中的Ja_i代表干扰机的编号,Ra_m代表雷达的编号,方块中的数字代表干扰机采用的干扰功率等级,范围为1~10,方块的填充色代表干扰机的干扰样式,其中绿色表示噪声调幅干扰,紫色表示噪声调频干扰,蓝色表示灵巧噪声干扰,橙色表示密集假目标干扰。1~6号干扰机的干扰样式为[0,1,2,2,3,3]。图中带有填充色的方块组合即表示协同干扰策略,例如图10(a)中1~6号电子干扰

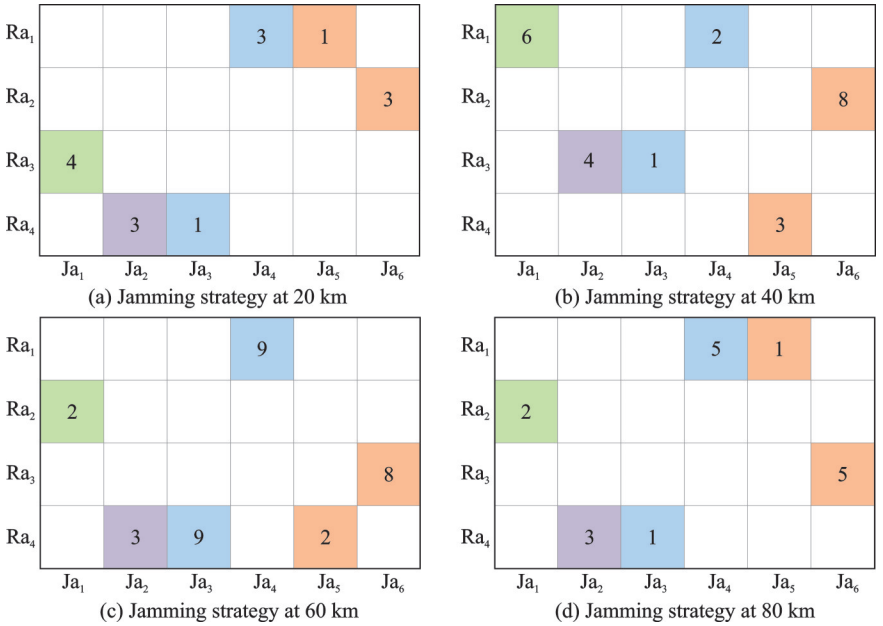

图10 协同突防过程中干扰策略变化

Fig.10 Changes in jamming strategies during cooperative penetration

无人机的干扰对象选择为[3,4,4,1,1,2],采用的干扰功率为[4,3,1,3,1,3],干扰样式为[0,1,2,2,3,3]。图10分别展示了突防目标行进至20、40、60和80 km处时,电子干扰无人机的协同干扰策略。在目标行进至20 km时,2号雷达距离目标最近,其威胁等级最高,因此6号干扰机采用密集假目标对2号雷达进行干扰,能够获得最佳的干扰效果。在目标行进至40 km时,2号雷达与目标距离最近的同时,4号雷达与目标的距离达到最小值。因此,拥有密集假目标干扰能力的5号和6号干扰机分别选择干扰4号和2号雷达。在目标行进至60 km时,2号雷达逐渐远离目标,而1号、3号和4号雷达距离目标较近,但均大于20 km。此时,灵巧噪声干扰和密集假目标干扰均能达到有效干扰效果。因此,4号、5号和6号干扰机分别选择干扰1号、4号和3号雷达。为了提高干扰效果,4号干扰机选择使用了第9档干扰功率。随着目标继续行进至80 km处,3号雷达与目标的距离达到最小值且小于20 km,威胁最大,其次是1号雷达。因此,5号和6号干扰机分别选择干扰1号和3号雷达。又因为1号雷达距离目标相对较远,因此5号干扰机选择了第1档干扰功率,从而提升干扰能效比。

图9,10结果证明,本文提出的APPO深度强化学习算法能够高效合理地完成干扰资源分配。

5 结束语

本文针对低空智联网协同认知干扰展开研究。首先,对多功能雷达系统发现概率进行建模,研究了4种雷达有源干扰样式的干扰机理,评估了不同干扰样式的干扰效果。然后,提出了一种基于深度强化学习的认知干扰决策模型,并在PPO算法的基础上提出了能够自适应改变学习率的APPO算法。同时,针对协同电子干扰问题,将数字孪生与深度强化学习结合,充分利用数字孪生能够获得高保真样本数据以及拥有强大计算资源的优势,提出了“集中式训练、分布/集中式执行、持续进化”的认知决策干扰模型训练框架。最后,通过仿真实验对比了PPO、APPO和MAPPO三种算法在不同场景下的性能,并分析了模型超参数对实验性能的影响。实验结果证明了本文所提出的认知决策干扰模型的有效性。

参考文献:

[1] 董超, 经宇骞, 屈毓锛, 等. 面向低空智联网频谱认知与决策的云边端融合体系架构[J]. 通信学报, 2023, 44(11): 1-12.
DONG Chao, JING Yuqian, QU Yuben, et al. Cloud-edge-device fusion architecture oriented to spectrum cognition and decision in low altitude intelligence network[J]. Journal on Communications, 2023, 44(11): 1-12.

[2] 吴中伟, 宋振之, 洪学尧. 电子对抗无人机在边境空中突防作战中的运用研究[J]. 战术导弹技术, 2022(2): 52-58.
WU Zhongwei, SONG Zhenzhi, HONG Xueyao. Research on application of electronic countermeasure UAV in border air penetration operation[J]. Tactical Missile Technology, 2022(2): 52-58.

[3] JIANG N, ZHANG H. Classification and model reconstruction method of non-cooperative multifunctional radar waveform unit [J]. IET Radar, Sonar & Navigation, 2023, 17(3): 408-421.

[4] ARULKUMARAN K, DEISENROTH M P, BRUNDAGE M, et al. Deep reinforcement learning: A brief survey[J]. IEEE Signal Processing Magazine, 2017, 34(6): 26-38.

[5] 李梓瑜, 葛芬, 张劲东, 等. 基于深度强化学习的雷达智能抗干扰决策FPGA加速器设计[J]. 数据采集与处理, 2023, 38(5): 1151-1161.
LI Ziyu, GE Fen, ZHANG Jindong, et al. Design of FPGA accelerator for radar intelligent anti-jamming decision-making based on deep reinforcement learning[J]. Journal of Data Acquisition and Processing, 2023, 38(5): 1151-1161.

[6] 王心一, 陈志江, 雷磊, 等. 多无人机网络边缘智能计算卸载算法[J]. 数据采集与处理, 2023, 38(6): 1286-1298.
WANG Xinyi, CHEN Zhijiang, LEI Lei, et al. Computation offloading algorithm for multi-UAV network based on edge intelligence[J]. Journal of Data Acquisition and Processing, 2023, 38(6): 1286-1298.

[7] WANG Z, GUPTA R, HAN K, et al. Mobility digital twin: Concept, architecture, case study, and future challenges[J]. IEEE Internet of Things Journal, 2022, 9(18): 17452-17467.

[8] MIHAI S, YAQOOB M, HUNG D V, et al. Digital twins: A survey on enabling technologies, challenges, trends and future prospects[J]. IEEE Communications Surveys & Tutorials, 2022, 24(4): 2255-2291.

[9] LEI L, SHEN G, ZHANG L, et al. Toward intelligent cooperation of UAV swarms: When machine learning meets digital twin[J]. IEEE Network, 2020, 35(1): 386-392.

[10] GUAN Y, REN Y, LI S E, et al. Centralized cooperation for connected and automated vehicles at intersections by proximal policy optimization[J]. IEEE Transactions on Vehicular Technology, 2020, 69(11): 12597-12608.

[11] 张柏开, 朱卫纲. MFR认知干扰决策体系构建及关键技术[J]. 系统工程与电子技术, 2020, 42(9): 1969-1975.
ZHANG Bokai, ZHU Weigang. Construction and key technologies of cognitive jamming decision-making system against MFR [J]. Systems Engineering and Electronics, 2020, 42(9): 1969-1975.

[12] 张柏开, 朱卫纲. 基于 Q-Learning 的多功能雷达认知干扰决策方法[J]. 电讯技术, 2020, 60(2): 129-136.
ZHANG Bokai, ZHU Weigang. A cognitive jamming decision method for multi-functional radar based on Q-learning[J]. Telecommunication Engineering, 2020, 60(2): 129-136.

[13] 朱霸坤, 朱卫纲, 李伟, 等. 基于马尔可夫的多功能雷达认知干扰决策建模研究[J]. 系统工程与电子技术, 2022, 44(8): 2488-2497.
ZHU Bakun, ZHU Weigang, LI Wei, et al. Research on decision-making modeling of cognitive jamming for multi-functional radar based on Markov[J]. Systems Engineering and Electronics, 2022, 44(8): 2488-2497.

[14] 黄星源, 李岩屹. 基于双Q学习算法的干扰资源分配策略[J]. 系统仿真学报, 2021, 33(8): 1801.
HUANG Xingyuan, LI Yanyi. The allocation of jamming resources based on double Q-learning algorithm[J]. Journal of System Simulation, 2021, 33(8): 1801.

[15] 陈泽盛, 杨承志, 曹鹏宇, 等. 一种基于双DQN的空战干扰样式选择方法[J]. 电讯技术, 2021, 61(11): 1371-1377.
CHEN Zesheng, YANG Chengzhi, CAO Pengyu, et al. An air combat jamming style selection method based on double DQN [J]. Telecommunication Engineering, 2021, 61(11): 1371-1377.

[16] FENG L W, LIU S T, XU H Z. Multifunctional radar cognitive jamming decision based on dueling double deep Q-network[J]. IEEE Access, 2022, 10: 112150-112157.

[17] ZHANG W, ZHAO T, ZHAO Z, et al. Performance analysis of deep reinforcement learning-based intelligent cooperative jamming method confronting multi-functional networked radar[J]. Signal Processing, 2023, 207: 108965.

[18] 邹玮琦, 牛朝阳, 刘伟, 等. 基于 A3C 的多功能雷达认知干扰决策方法[J]. 系统工程与电子技术, 2023, 45(1): 86-92.
ZOU Weiqi, NIU Chaoyang, LIU Wei, et al. Cognitive jamming decision-making method against multifunctional radar based on A3C[J]. Systems Engineering and Electronics, 2023, 45(1): 86-92.

[19] 胡小全, 刘钦, 孙建军. 雷达组网协同探测范围研究[J]. 雷达科学与技术, 2015, 13(3): 223-227.
HU Xiaoquan, LIU Qin, SUN Jianjun. Study on cooperative detection coverage of radar network[J]. Radar Science and Technology, 2015, 13(3): 223-227.

[20] 黄星源. 基于强化学习的雷达干扰策略分配技术研究[D]. 哈尔滨:哈尔滨工程大学, 2021.
HUANG Xingyuan. Research on allocation tactics for radar jamming based on reinforcement learning[D]. Harbin: Harbin Engineering University, 2021.

[21] YU C, VELU A, VINISKY E, et al. The surprising effectiveness of PPO in cooperative multi-agent games[J]. Advances in Neural Information Processing Systems, 2022, 35: 24611-24624.

作者简介:

沈高青(1994-),男,讲师,研究方向:飞行器智能组网与协同,E-mail: shengaoqing@nuaa.edu.cn。

蔡圣所(1988-),男,实验师,研究方向:飞行器智能组网与协同。

雷磊(1981-),**通信作者**,男,教授,博士生导师,研究方向:飞行器智能组网与协同,E-mail: leilei@nuaa.edu.cn。

贲德(1938-),男,教授,中国工程院院士,研究方向:雷达信号处理。

(编辑:张黄群)