ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 6,Nov. 2024,pp. 1355－1369
DOI:10. 16337/j. 1004-9037. 2024. 06. 006

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

基于近端策略优化算法和Mask-TIT网络的多功能雷达干扰决策方法

娄雨璇，孙闽红，尹　帅

（杭州电子科技大学通信工程学院，杭州 310018）

摘　要：为应对愈加智能的多功能雷达给对抗方带来的挑战，本文提出一种基于近端策略优化（Proximal policy optimization，PPO）算法和Mask-TIT（Mask-Transformer in Transformer）网络的干扰决策方法。首先，从一种现实场景出发，将干扰机与雷达的对抗场景建模为部分可观察马尔可夫决策过程（Partially observable Markov decision process，POMDP），根据雷达工作原理设计了新的状态转移函数和奖励函数，并根据多功能雷达层级模型设计了观测空间。其次，利用Transformer对序列数据的表征能力和雷达干扰样式的特点设计了一种Mask-TIT网络结构，用于构建更强大的Actor-Critic网络架构。最后，使用近端策略优化算法进行优化学习。实验结果表明，该算法较现有方法收敛所需交互数据平均减少25.6％，并且收敛后的方差显著降低。

关键词：雷达干扰决策；部分可观察马尔可夫决策过程；强化学习；Transformer；近端策略优化
中图分类号：TN974　　**文献标志码：**A

A Multi-functional Radar Jamming Decision Method Based on Proximal Policy Optimization Algorithm and Mask-TIT Network

LOU Yuxuan，SUN Minhong，YIN Shuai

（School of Communication Engineering，Hangzhou Dianzi University，Hangzhou 310018，China）

Abstract：To cope with the challenges brought by increasingly intelligent multifunctional radars to the opposing side，this paper proposes an jamming decision-making method based on the proximal policy optimization（PPO）algorithm and the Mask-Transformer in Transformer（Mask-TIT）network. Firstly，starting from a realistic scenario，the adversarial scene between the jammer and the radar is modeled as a partially observable Markov decision process（POMDP）. A new state transition function and reward function are designed based on the working principles of the radar，and the observation space is designed according to the hierarchy of the multifunctional radar model. Secondly，a Mask-TIT network structure is designed using the Transformer's representation capacity for sequence data and the characteristics of radar jamming patterns，which is used to build a more powerful Actor-Critic network architecture. Finally，the PPO algorithm is used for optimization learning. Experimental results show that compared with existing methods，the proposed algorithm reduces the average amount of interactive data required for convergence by 25.6％，and the variance after convergence is significantly reduced.

Key words: radar jamming decision; partially observable Markov decision process（POMDP）; reinforcement learning; Transformer; proximal policy optimization（PPO）

收稿日期：2023-05-23；**修订日期：**2023-08-12

引　言

在电子战逐渐成为现代战争重心的今天，愈加智能的雷达给雷达对抗方带来严峻挑战。自Simon Haykin[1]于2006年提出“认知雷达”概念以来，相控阵天线技术和数字阵列雷达技术逐渐成熟推动着雷达技术迅猛发展，以相控阵雷达为代表的多功能雷达(Multi-function radar，MFR)在自适应波束成形技术的支持下，能够根据战场环境和任务需求实时改变工作参数，同时执行多种任务，并且具有较强的抗干扰能力[2-3]。如何有效地干扰敌方多功能雷达逐渐成为电子对抗技术关注的焦点之一。

准确的干扰决策是实施有效干扰的前提。传统的基于模板匹配的干扰决策方法[4]受限于先验知识的匮乏，已不适用于灵活多变的多功能雷达，而强化学习为基于学习的决策提供了一个通用的框架。强化学习理论并不受限于先验知识，并且能够在未知的环境中通过“试错”的方式不断迭代优化，因此成为目前干扰决策领域研究的热点。文献[5]首次将强化学习引入雷达干扰决策领域，将Q-Learning应用于雷达干扰决策。进一步地，文献[6-7]在此基础上针对未知雷达工作模式的情况进行了改进。文献[8]用模拟退火算法改进了Q-learning算法，加强了干扰策略的探索和利用。文献[9]将深度Q网络(Deep Q network，DQN)算法应用到雷达干扰决策中。文献[10]将双重深度Q网络(Double DQN，DDQN)算法应用于干扰资源分配；文献[11]将竞争双深度Q网络(Dueling double DQN，D3QN)算法应用于雷达干扰决策，并将高维干扰行为空间分解为两个低维子空间，通过建立两个相互作用的Q学习模型求得最优解；文献[12]将异步优势动作评估(Asynchronous advantage actor-critic，A3C)算法应用于雷达干扰决策中。以上研究主要集中在对强化学习算法的使用和改进上，未对构建更为强大的强化学习网络进行更深入的探讨。此外，上述文献大都将雷达与干扰机的对抗过程建模为马尔科夫决策过程(Markov decision processes，MDP)[13]，但在实际场景中，干扰方无法直接获取雷达的真实状态，只能通过分析雷达信号进行推测，这不满足MDP的“环境是完全可观测的”的前提假设。

针对现有研究的不足，本文首先将干扰机与雷达的对抗过程建模为部分可观察马尔可夫决策过程(Partially observable Markov decision process，POMDP)[14]，将更契合实际的雷达信号作为观测输入。在此基础上，提出结合近端策略优化(Proximal policy optimization，PPO)算法[15]和Mask-TIT(Mask-Transformer in Transformer)网络的干扰决策方法，设计了一种Mask-TIT网络结构并将其作为PPO算法中Actor-Critic结构的骨干网络。Mask-TIT网络利用Transformer[16]对时序数据的强大表征力提取干扰机观测数据的特征，并通过Mask网络层结构将干扰样式中的先验信息融入网络，来降低动作空间的搜索维度，以此提升算法的收敛速度。

1　雷达干扰对抗场景建模

本文场景设定为：对位于雷达探测区域外的电子战飞机实施支援干扰，掩护无人机编队突防多功能雷达信号覆盖区域，场景如图1所示。

雷达(对应于强化学习中的“环境”)周期性地扫描临近空域，在发现目标后对其进行目标识别和跟踪，在跟踪一定时间后获得飞行目标的运动信息，即可对其进行武器制导、实施打击。

为完成突防任务，干扰机(对应于“智能体”)通过接收到的雷达信号获取信号参数，分析雷达状态，进行干扰决策，并生成相应的干扰信号，对目标雷达进行干扰。

实际对抗过程中，雷达的真实状态对于干扰方不可知，因此

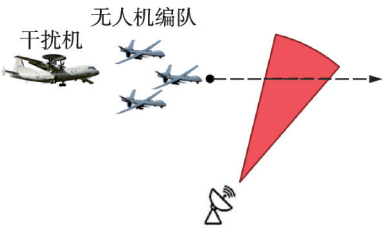


图1　突防任务场景

Fig.1　Assault mission scenario

本文将干扰机与雷达的交互过程建模为POMDP,由$\{\boldsymbol{S},\boldsymbol{A},P,R,\boldsymbol{O},\gamma\}$六元组表示,其中$\boldsymbol{S}$为状态空间,$\boldsymbol{A}$为动作空间,$P$为状态转移函数,$R:\boldsymbol{S}\times\boldsymbol{A}\to R$为奖励函数,$\boldsymbol{O}$为观测空间,$\gamma$为折现因子。由于环境的不确定性和不完全观测性质,智能体无法准确地估计未来的状态和奖励,引入折现因子γ可以降低未来不确定性对决策的影响。γ介于0和1之间,越接近1表示未来奖励的价值越高,越接近0表示未来奖励的价值越低。为简化模型此处不考虑观测概率函数。

1.1 状态空间

状态空间用于描述多功能雷达当前的状态,本文将其定义为雷达当前工作模式和雷达已获取目标的信息集合。参考文献[3]并结合本文提出的对抗场景,不失一般性,假设多功能雷达任务简化为3种:搜索、跟踪和制导,且其各工作模式对应的场景如表1所示。

表1 多功能雷达工作模式-场景对应

Table 1 Multifunctional radar working mode-scene correspondence

序号	雷达工作模式	对应场景
1	搜索	雷达未发现任何目标
2	边搜索边跟踪	雷达发现疑似目标并继续搜索
3	搜索加跟踪	雷达跟踪少量目标并继续搜索
4	单目标跟踪	雷达只跟踪单个目标
5	多目标跟踪	雷达跟踪多个目标
6	导弹制导	对跟踪的目标进行打击

分别用状态mode_1至mode_6表示表1中的6种雷达工作模式。雷达所获得的环境信息为雷达侦察到的目标信息(位置、速度等),表示为

$$\boldsymbol{I}_{\text{target}_i}=[x_t^i,y_t^i,z_t^i,v_{x,t}^i,v_{y,t}^i,v_{z,t}^i] \tag{1}$$

$$\boldsymbol{I}_{\text{Target}}=[\boldsymbol{I}_{\text{target}_i}]\quad i=1,2,\cdots,N \tag{2}$$

式中:$\boldsymbol{I}_{\text{target}_i}$为第$i$个目标位置与速度信息向量;$x_t^i,y_t^i,z_t^i$为第$i$个目标在$t$时刻位置的三维坐标;$v_{x,y}^i,v_{y,i}^i,v_{z,t}^i$为各个方向的速度分量。则状态空间表示为

$$\boldsymbol{S}=[\text{mode}_i,\boldsymbol{I}_{\text{Target}}]\quad i=1,2,3,4,5,6 \tag{3}$$

多功能雷达在对搜索到的目标进行跟踪的同时,将加强对目标附近空域的搜索。多功能雷达可以同时跟踪多个目标,但是受限于雷达数据处理能力,同一时段只能跟踪有限目标。因多功能雷达同时跟踪目标的数量大小对本文方法的结果没有根本性影响,不失一般性且为了方便计算,假定雷达同时跟踪目标的数量上限为5个。

1.2 动作空间

动作空间$\boldsymbol{A}$为干扰机可选择的干扰样式集合,干扰与符号的对应关系如表2所示。为便于实验验证,额外地定义a_0为不进行干扰,即干扰机不工作。

表2 干扰符号对应关系

Table 2 Jamming-symbol correspondence

符号	干扰样式
a_1	射频噪声干扰
a_2	噪声调幅干扰
a_3	噪声调频干扰
a_4	距离欺骗干扰
a_5	距离-速度联合假目标干扰
a_6	频谱弥散(Smeared spectrum, SMSP)干扰
a_7	切片组合(Chopping and interleaving, C&I)干扰
a_8	距离拖引干扰
a_9	距离-速度联合拖引干扰

1.3 状态转移函数

雷达状态或工作模式的转移与实际雷达任务以及雷达所探测到的信息相关,每一架无人机都被视为独立的目标,单独计算其是否被雷达搜索(m_1)、跟踪(m_2)或制导(m_3)。

雷达处于搜索模式时,会周期性地扫描附近区域,当搜索信号范围覆盖到无人机群时(如图2所示),雷达依检测概率判断目标的存在。检测概率主要由目标回波功率和信干比决定[17],为降低模型复杂度,本文将检测概率简化为一个仅与时间相关的函数,假定目标暴露在雷达信号覆盖范围内越久,被发现的概率越大。雷

达发现目标后为确认目标的存在,会加强对该区域的搜索,并依概率转换为跟踪任务,转换概率定义为

$$P(m_2|m_1,a_0)=\tanh(\omega_1 t) \tag{4}$$

$$P(m_1|m_1,a_0)=1-P(m_2|m_1,a_0) \tag{5}$$

式中:$\omega_1>0$为超参数,用于控制雷达的探测性能。ω_1越小,雷达探测性能越差。

当雷达跟踪到目标时,在保持跟踪一定时间后,将转为制导任务,对目标进行打击,任务转移概率随时间递增,定义为

$$P(m_3|m_2,a_0)=\tanh(\omega_2 t) \tag{6}$$

$$P(m_2|m_2,a_0)=1-\tanh(\omega_2 t) \tag{7}$$

式中:$\omega_2>0$为超参数,用于控制雷达的跟踪性能,ω_2越小,雷达跟踪性能越差。

图2 雷达搜索到目标

Fig.2 Radar finding the target

设雷达一旦进入制导,则判定该无人机被摧毁。若所有无人机都被摧毁,则判定整个突防任务失败。为完成突防任务,干扰机在飞行过程中向雷达发射干扰信号。压制性干扰通过干扰噪声降低雷达检测概率,进而影响雷达任务转移概率。例如,对执行搜索任务(m_1)的雷达实施第i种压制性干扰a_i,雷达任务转移为跟踪(m_2)的概率$P(m_2|m_1,a_i)$将降低$P(m_1,a_i)$。

欺骗性干扰通过发射虚假脉冲欺骗雷达,使得雷达跟踪或制导到假目标。设干扰机发射第i种欺骗干扰a_i时,雷达搜索并跟踪到假目标的概率为$P(m_1,a_i)$,则雷达由搜索转为跟踪的转移概率为

$$P(m_2|m_1,a_i)=\begin{cases}P(m_1,a_i) & \text{跟踪假目标}\\ 1-P(m_1,a_i) & \text{跟踪真目标}\end{cases} \tag{8}$$

设雷达跟踪到真实目标后被第i种拖引干扰a_i成功欺骗的概率为$P(m_2,a_i)$,则雷达丢失跟踪的概率为

$$P(m_1|m_2,a_i)=\begin{cases}P(m_2,a_i) & \text{拖引成功}\\ 1-P(m_2,a_i) & \text{拖引失败}\end{cases} \tag{9}$$

上述的概率P由雷达方接收机信号处理结果决定,干扰方实际无法获取其准确值,本文实验参考文献[3]并根据已有经验进行合理假设。雷达任务转移概率函数可由矩阵$\boldsymbol{P}'$描述,其元素为所有转移概率,表示为$\boldsymbol{P}'=[P(m_i|m_j,a_k)]$,$i=1,2,3$,$j=1,2$,$k=0,1,\cdots,9$,是一个$3\times2\times10$的三维矩阵。

1.4 奖励函数

在雷达干扰对抗领域中,传统的干扰效能评估方法和准则(例如信息准则、功率准则)大都站在雷达方的角度,根据雷达接收机接收到的信号质量评估干扰效果,这并不符合实际对抗场景。实际上,干扰方仅能通过对比干扰前后雷达信号的变化来评估干扰效果。因此,本文从雷达信号角度和突防任务目标角度,建立了干扰效能评估层次模型。

从雷达信号角度,本文将雷达受干扰后可能发生的变化分为2类,并分别给出奖励函数。

(1)雷达信号参数变化。当雷达被干扰后,通常会改变雷达发射信号参数,如采取捷变频等抗干扰措施,将工作频段切换至未被干扰噪声所覆盖的频段。

$$r_1=\begin{cases}1 & \text{雷达信号参数大幅变动}\\ -2 & \text{雷达信号参数不变}\end{cases} \tag{10}$$

(2)雷达工作模式变化。有效的干扰可以影响雷达的正常工作,使得雷达从高威胁等级的工作模式逐步转为低威胁等级的工作模式。

$$r_2=\begin{cases}10 & \text{转移至更低威胁等级的工作模式}\\ -12 & \text{转移至更高威胁等级的工作模式}\end{cases} \tag{11}$$

从突防任务角度,以无人机编队通过雷达信号覆盖区域或全部被击落作为终止时刻($t=T$)。若顺利通过雷达覆盖区域,则根据成功突防的无人机的存活率σ作为奖励标准,给予$100\times\sigma$点奖励。若无人机全部被摧毁,则给予-100点惩罚。

$$r_T=\begin{cases}100\times\sigma & t=T\ ;\ \sigma\neq 0\\ -100 & \sigma=0\end{cases} \tag{12}$$

雷达工作模式的转变相较于其信号参数变化更能体现干扰效果,因此r_2比r_1的值高一个数量级。并且为保证无人机编队尽可能处于安全状态,r_1和r_2中的惩罚项略大于奖励项。综上,全局奖励函数定义为

$$R=\begin{cases}r_1+r_2 & t\neq T\\ r_T & t=T\end{cases} \tag{13}$$

1.5 观测空间

干扰机接收(观测)雷达的脉冲信号并提取其参数,主要包括脉宽(Pulse width,PW)、脉冲幅度(Pulse amplitude,PA)、脉冲重复间隔(Pulse repetition interval,PRI)和载频(Radio frequency,RF)。将所有参数组合得到脉冲描述字(Pulse description word,PDW)向量。

由于多功能雷达可以在各种工作模式中灵活切换,并且这些模式具有灵活的调制类型和可变的参数,仅靠雷达脉冲信号的PDW信息难以为干扰决策提供足够的信息,因此本文在此基础上进行目标雷达的工作模式识别,并根据多功能雷达的层级模型对雷达状态进行编码。

关于多功能雷达的工作模式识别和层级模型建模方法已有不少相关文献和成果[18-19],这里对其不作过多讨论。多功能雷达的层级模型最早由Visnevski等[20]提出,用于描述多功能雷达工作模式与雷达脉冲序列之间的映射关系。每种雷达工作模式包含多个雷达短语,雷达短语由多个雷达字组成,雷达字是对一组雷达脉冲序列的抽象。本文将雷达工作模式和雷达字的组合作为雷达状态编码,并将其定义为雷达的观测

$$\boldsymbol{o}\triangleq n_{\text{mode}}\circ\boldsymbol{p}_m\triangleq[n_{\text{mode}},w_i,w_j,w_k,w_l] \tag{14}$$

式中:"$\circ$"为向量拼接符号;n_{mode}为第n种工作模式;$\boldsymbol{p}_m$,$m=1,2,\cdots,M$为雷达工作模式n_{mode}所包含的雷达短语,共有M种;$\{w_i,w_j,w_k,w_l\}$,$i,j,k,l=1,2,\cdots,9$表示雷达短语中的雷达字,共有9种雷达字。

2 Mask-TIT网络

Actor-Critic架构是目前强化学习最先进的架构,它由两个网络构成:Actor网络用于学习最佳策略函数,策略函数$\pi(a|\boldsymbol{o})$将观测空间映射到动作空间,它计算在得到观测$\boldsymbol{o}$时所采取动作的概率分布,实际决策时智能体(即干扰机)通过对其采样得到最终的动作。Critic网络用于学习价值函数$V(\boldsymbol{o})$,智能体通过价值函数评估所做的决策优劣,为策略网络的更新提供依据。

在POMDP中,往往一个动作的后果只有在环境状态多次转换后才会体现,因此智能体当前的观测一定程度上取决于其之前的动作,具有较强的时间相关性,所以智能体必须能够处理长期的时间依赖。这被称为时间性的信用分配问题[21]。Transformer[16,22-23]和Vision Transformer[24-26]是自然语言处理和计算机视觉领域最杰出的突破,它们已经证明了其处理连续单词和图像块(更一般地说,是连续数据)的有效性和可扩展性。其次,强化学习的目标是基于顺序观察的顺序决策,这与Transformer和Vision

Transformer的能力完全匹配。因此本文设计了一种基于Transformer的强化学习网络结构——Mask-TIT,其结构如图3所示。

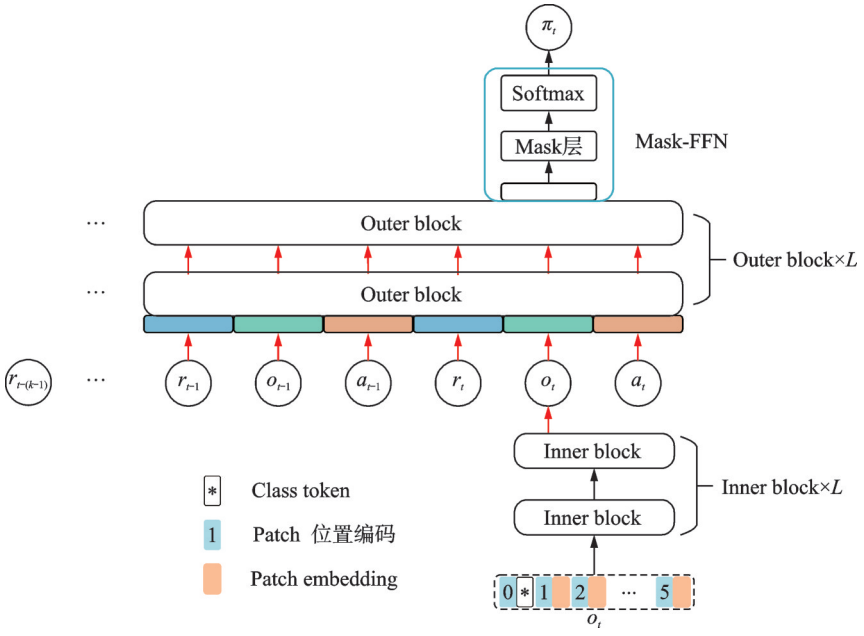


图3 Mask-TIT网络结构

Fig.3 Mask-TIT network structure

Mask-TIT以Transformer作为骨干网络,并根据雷达干扰样式的特性设计了一种Mask网络层结构。TIT由两种Transformer组成:Inner Transformer负责处理当前单次观测的数据,学习一个有效的观测表示,以捕获观测中的重要空间信息;Outer Transformer负责处理连续的多个历史观测,以捕捉跨越多个观测的重要时间信息。二者相结合,能够提取干扰机和雷达交互序列数据的时空表征,从而做出更好的决策。

Transformer需要对数据进行如下步骤的预处理。

(1)生成patch:Inner Transformer的输入为一个包含D个元素的观测数组$\boldsymbol{o}\in\mathbf{R}^{D}$,将数组中每个元素作为一个patch,得到patch序列$\boldsymbol{o}^{p}\in\mathbf{R}^{N\times 1}$,其中$N=D$为Inner Transformer的上下文长度。

(2)将每个patch通过可训练的线性映射$\boldsymbol{E}^{p}$提取嵌入特征(Embedding)

$$\hat{\boldsymbol{z}}_0=[\boldsymbol{o}_1^{p}\boldsymbol{E}^{p},\boldsymbol{o}_2^{p}\boldsymbol{E}^{p},\cdots,\boldsymbol{o}_N^{p}\boldsymbol{E}^{p}]\in\mathbf{R}^{N\times D^{p}} \tag{15}$$

式中$\boldsymbol{E}^{p}\in\mathbf{R}^{1\times D^{p}}$,$D^{p}$是patch嵌入特征的维度。

(3)参考自然语言处理领域的BERT[22]和计算机视觉领域的ViT[24],添加一个可训练的Class token $\boldsymbol{z}_0^{\text{class}}\in\mathbf{R}^{1\times D^{p}}$

$$\check{\boldsymbol{z}}_0=[\boldsymbol{z}_0^{\text{class}};\hat{\boldsymbol{z}}_0]\in\mathbf{R}^{(N+1)D^{p}} \tag{16}$$

(4)再添加一个位置编码,用一个可训练的$\boldsymbol{E}_{\text{pos}}^{p}$参数来保留位置信息。

$$\boldsymbol{z}_0=\check{\boldsymbol{z}}_0+\boldsymbol{E}_{\text{pos}}^{p}\quad \boldsymbol{E}_{\text{pos}}^{p}\in\mathbf{R}^{(N+1)\times D^{p}} \tag{17}$$

最终结果$\boldsymbol{z}_0$作为Inner Transformer的输入。

2.1 Inner Transformer

Inner Transformer由L个Inner block堆叠而成,每个Inner block都是一个Transformer编码器(图4(a)),因为每个patch都有辅助决策的意义,第l个Inner block T_l^{in}的操作是

$$\tilde{\boldsymbol{z}}_l=\boldsymbol{z}_{l-1}+\text{MSA}(\text{LN}(\boldsymbol{z}_{l-1}))\quad l=1,2,\cdots,L \tag{18}$$

$$\boldsymbol{z}_l=\boldsymbol{z}_{l-1}+\text{FFN}(\text{LN}(\tilde{\boldsymbol{z}}_{l-1}))\quad l=1,2,\cdots,L \tag{19}$$

式中:MSA、LN和FFN分别为多头自注意力(Multiheaded self-attention,MSA)、层规范化(Layer normalization,LN)和前馈网络(Feed-forward network,FFN)。这样,每个块的输出$\boldsymbol{z}_l$通过计算任意两个观察patch之间的相互作用,建立起单个观测内各观测patch之间的空间关系,最终得到$\boldsymbol{z}_L\in\mathbf{R}^{(N+1)\times D^p}$。参考BERT和ViT,将第1个元素$\boldsymbol{z}_L[0]\in\mathbf{R}^{1\times D^p}$(Class token $\boldsymbol{z}_L^{\text{class}}$)作为所有patch的综合表示。

2.2 Outer Transformer

Inner Transformer的输出是Outer Transformer的输入,即跨越K个时间步的所有$\boldsymbol{z}_L[0]$的级联,表示为$\boldsymbol{y}_0=\left\{\boldsymbol{z}_L[0]_t\right\}_{t-(K-1)}^{t}\in\mathbf{R}^{K\times D^p}$,这里$K$为Outer Transformer的上下文长度。把干扰机与雷达的交互数据中的观测、动作和奖励三元组作为Outer Transformer的输入,其中的观测替换成Inner Transformer的输出。

Outer Transformer同样由L个Outer block堆叠而成,每个Outer block都是一个Transformer解码器(图4(b)),因为实际过程中决策时无法得到未来的观测,所以训练时将未来时间步的观测进行屏蔽。第l个块T_l^{out}的操作是

$$\tilde{\boldsymbol{y}}_l=\boldsymbol{y}_{l-1}+\text{MSA}(\text{LN}(\boldsymbol{y}_{l-1}))\quad l=1,2,\cdots,L \tag{20}$$

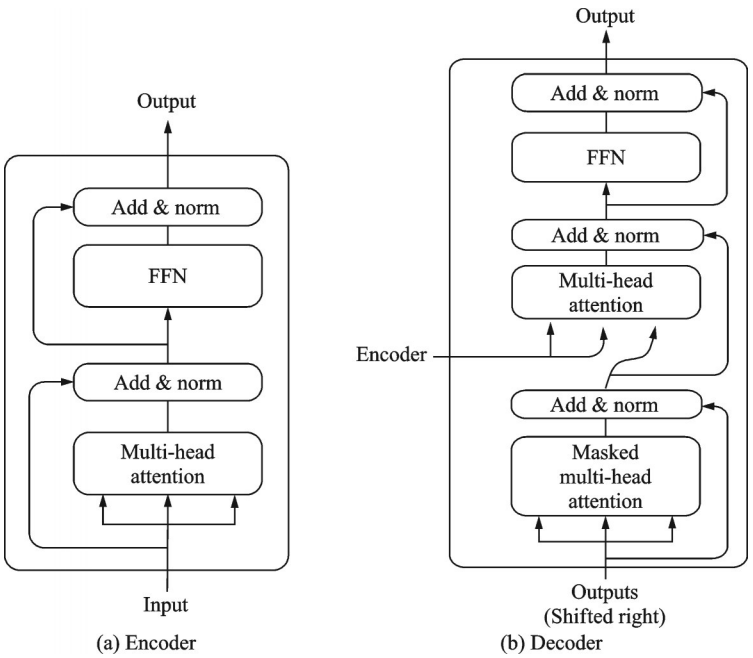


图4 Transformer结构

Fig.4 Transformer structure

$$\boldsymbol{y}_l=\boldsymbol{y}_{l-1}+\mathrm{FFN}(\mathrm{LN}(\tilde{\boldsymbol{y}}_{l-1}))\quad l=1,2,\cdots,L \tag{21}$$

这样,每个块的输出$\boldsymbol{y}_l$建立了多个连续观测值之间的时间关系,最终得到$\boldsymbol{y}_L\in\mathbf{R}^{K\times D^p}$。由于之前的输出屏蔽了未来观测,所以只有$\boldsymbol{y}_L$的最后一个元素才能表征所有观测的信息,因此把最后一个元素$\boldsymbol{y}_L[K-1]\in\mathbf{R}^{1\times D^p}$通过一个FFN和Softmax得到最终的策略$\pi(\cdot|\boldsymbol{o}_t)$。

2.3 Mask网络层

由于雷达干扰决策领域的特殊性,让干扰机以完全“试错”的方式进行学习代价高昂。在雷达对抗领域中,不同的干扰样式都是针对不同的场景和作战目标而设计的,每种干扰样式都有其特定的适用范围。例如,拖引类干扰的设计目的是为了摆脱雷达的跟踪,因此不适用于处于搜索状态的雷达。依据此类隐含在干扰样式内的先验信息,本文设计了能够根据不同雷达状态,约束TIT网络输出的Mask网络层。

TIT中最后一个FFN网络先输出一个非归一化分数向量$\boldsymbol{l}=[l_0,l_1,\cdots,l_N]$,向量中第$i$个元素对应于在当前状态下对选用干扰样式$a_i$的倾向,然后使用Softmax操作将其转化为动作概率分布。

$$\pi(\cdot|\boldsymbol{o})=\mathrm{Softmax}(\boldsymbol{l}) \tag{22}$$

为了将不适用的干扰样式(无效动作)去除,在FFN网络的输出层后添加一个Mask层。在Mask层中,将所有的无效动作对应的分数替换为一个较大的负数M(例如$M=-1\times10^8$),即

$$\mathrm{Mask}(\boldsymbol{l})_i=\begin{cases}l_i & a_i\text{为有效动作}\\ M & a_i\text{为无效动作}\end{cases} \tag{23}$$

然后再使用Softmax操作将其转化为动作概率分布

$$\begin{aligned}\pi_\theta(\cdot|\boldsymbol{o}_t)&=\mathrm{Softmax}\Big(\mathrm{Mask}([l_0,l_1,\cdots,l_N])\Big)=\mathrm{Softmax}([l_0,l_1,\cdots,M,\cdots,l_N])=\\&[\pi_\theta(a_0|\boldsymbol{o}_t),\pi_\theta(a_1|\boldsymbol{o}_t),\cdots,\mu,\cdots,\pi_\theta(a_N|\boldsymbol{o}_t)]\end{aligned} \tag{24}$$

式中:μ为无效动作被选择的概率,当M足够小,此概率实际上为0。由策略梯度定理可知,无效动作对数的梯度也为0,因此Mask函数不会影响梯度的有效性。Mask-FFN网络结构如图5所示。

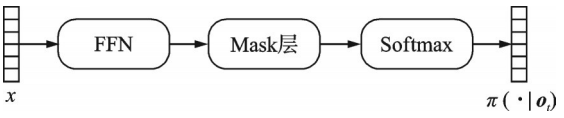


图5 Mask-FFN结构

Fig.5 Mask-FFN structure

图5的结构可表达为

$$\mathrm{Mask-FFN}(x)=\mathrm{Softmax}(\mathrm{Mask}(\mathrm{FFN}(x))) \tag{25}$$

通过Mask网络层可以将干扰样式设计中的知识融入网络,能够有效地降低干扰机每次决策的搜索空间,提高算法收敛速度。此外,还可有效地降低干扰机“试错”过程危险性。

由于Actor网络和Critic网络都以观测作为输入,因此可以共享用于特征提取的TIT网络参数以加速网络训练。使用Mask-TIT网络构建的共享网络参数的Actor-Critic结构如图6所示,其中上层为Actor网络输出,下层为Critic网络输出。

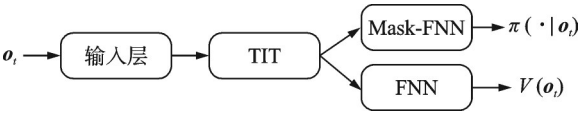


图6 Actor-Critic结构

Fig.6 Actor-Critic structure

3 PPO算法

PPO算法是一种基于策略梯度的强化学习算法,策略梯度算法的核心思想是通过最大化期望回报来优化策略。PPO算法通过限制策略更新幅度,以达到稳定、高效的训练结果。具体来说,PPO算法使用了两个损失函数:第1个损失函数为近端比率裁剪损失,用于限制策略更新幅度;第2个损失函数为价值函数损失,用于优化策略。两个损失函数的加权和是PPO算法的总损失函数。

近端比率裁剪损失通过剪裁函数限制策略更新幅度,并使用梯度上升算法优化策略网络

$$L^{\mathrm{Clip}}(\theta)=\hat{E}_t\left[\mathrm{Min}\left(r_t(\theta)A_t,\mathrm{Clip}(r_t(\theta),1-\varepsilon,1+\varepsilon)\hat{A}_t\right)\right] \tag{26}$$

式中:θ为策略函数(网络)参数;$\hat{E}_t$表示所有时间步的经验期望;$r_t(\theta)$为更新前后新旧策略的概率之比,$r_t(\theta)=\dfrac{\pi_\theta(a_t|\boldsymbol{o}_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|\boldsymbol{o}_t)}$;$\hat{A}_t$为优势函数[27],是一种衡量当前状态和动作相对于平均水平的优劣程度的函数。它表示当前状态和动作的价值与平均价值的差值,用于计算近端比率裁剪损失中的裁剪幅度。

$$\hat{A}_t=\delta_t+(\gamma\lambda)\delta_{t+1}+\cdots+(\gamma\lambda)^{T-t+1}\delta_{T-1} \tag{27}$$

$$\delta_t=r_t+\gamma V(\boldsymbol{o}_{t+1})-V(\boldsymbol{o}_t) \tag{28}$$

式中$V(\boldsymbol{o}_t)$表示在观测$\boldsymbol{o}_t$下的平均价值。

Clip表示剪裁函数,它将$r(\theta)$约束在1附近的领域中,即$[1-\varepsilon,1+\varepsilon]$,$\varepsilon$为其超参数,通常取0.2。

$$\mathrm{Clip}(r_t(\theta),1-\varepsilon,1+\varepsilon)=\begin{cases}1-\varepsilon & r_t(\theta)<1-\varepsilon\\ 1+\varepsilon & r_t(\theta)>1+\varepsilon\\ r_t(\theta) & \text{其他}\end{cases} \tag{29}$$

PPO算法使用均方误差作为价值函数损失来拟合Critic网络。

$$L_t^{\mathrm{VF}}=E\left[\left(V_\theta(s_t)-V_t^{\mathrm{target}}\right)^2\right] \tag{30}$$

为增强稳定性,本文对价值函数也采用剪裁函数处理,即

$$L^{\mathrm{VF}}=\mathrm{Max}\left[\left(V_{\theta_t}-V_{\mathrm{targ}}\right)^2,\left(\mathrm{Clip}(V_{\theta_t},V_{\theta_{t-1}}-\varepsilon,V_{\theta_{t-1}}+\varepsilon)-V_{\mathrm{targ}}\right)^2\right] \tag{31}$$

3.1 干扰机与多功能雷达对抗过程

干扰机与多功能雷达的对抗过程如图7所示,其对抗步骤如下。

(1)干扰方的接收机在一段处理间隔内接收雷达脉冲信号并测量出脉冲信号的PDW,同时识别出多功能雷达的工作模式,并根据多功能雷达层级模型对雷达状态进行编码,将二者组合得到观测$\boldsymbol{o}_t$。

(2)将观测进行预处理后输入Mask-TIT网络,根据当前策略π_θ得到干扰样式a_i。

(3)干扰机根据决策出的干扰样式生成相应的干扰信号发射至雷达,引起雷达状态发生相应变化,并体现在其下一轮发射的脉冲信号中。

(4)干扰机重复步骤(1、2),并根据雷达工作模式的变化以及当前物理状态通过式(13)评估干扰效果,得出上一次干扰的奖励。

在无人机编队飞过雷达信号覆盖范围内时,将接收机的每一段处理间隔视作一个离散时刻。在所有离散时刻,干扰机重复上述步骤不断对雷达进行干扰并将每次对抗的数据记录成轨迹序列。每轮突防任务结束后Mask-TIT网络根据轨迹中的数据使用PPO算法优化Mask-TIT网络。

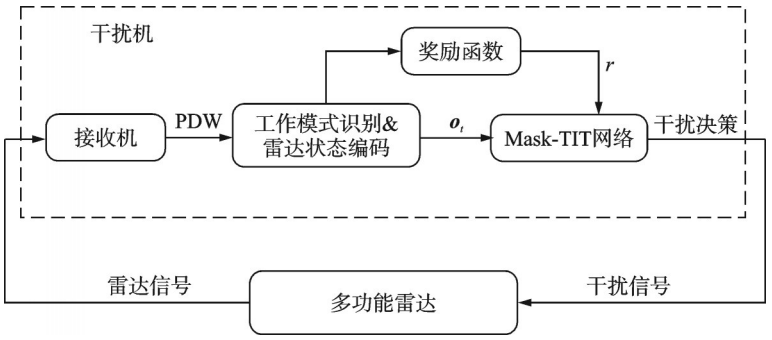


图7 干扰机与多功能雷达对抗过程

Fig.7 Jammer against multifunctional radar process

3.2 算法伪代码

本文所提算法的伪代码如下。

随机初始化Actor网络参数θ和Critic网络参数ϕ。

for $k=1:K$ do

无人机编队按照既定路线飞行,干扰机根据干扰策略π_{θ_k}与目标雷达进行交互直至任务结束,得到轨迹$\tau_i=(\boldsymbol{o}_0,a_0,r_0),(\boldsymbol{o}_1,a_1,r_1),\cdots,(\boldsymbol{o}_T,a_T,r_T)$。并将交互轨迹数据和干扰策略$\pi_{\theta_k}$保存至经验回放集$D_k=\left\{(\pi_{\theta_k},\tau_i)\right\}$;

计算轨迹中每个时间步后,将得到的折扣奖励$\hat{G}_t=\sum_{i=0}^{T-t}\gamma^i r_{t+i}$保存至经验回放集$D_k=\left\{\left(\pi_{\theta_k},\tau_i,\hat{G}\right)\right\}$;

根据式(27)基于当前的价值函数V_{ϕ_k}计算优势函数$\hat{A}_t$,并保存至经验回放集$D_k=\left\{\left(\pi_{\theta_k},\tau_i,\hat{G},\hat{A}\right)\right\}$;

for epoch $=1:M$ do

从经验回放集中按照优先级提取一段小批次轨迹数据$\{(s_1,a_1,r_1),(s_2,a_2,r_2),\cdots,(s_N,a_N,r_N)\}$,并对该批次数据的优势进行标准化$A_t'=\dfrac{A_t-\mu}{\sigma}$,其中$\mu$和$\sigma$分别为该批数据的均值和标准差;

根据式(26)计算目标损失L和策略函数梯度$g(\theta)$;

根据式(31)计算均方误差损失和价值函数的梯度$g(\phi)$;

使用梯度上升算法根据策略梯度$g(\theta)$更新策略参数θ;

使用梯度下降算法根据价值函数梯度$g(\phi)$更新价值网络参数ϕ;

end for

end for

4 仿真实验与分析

4.1 实验参数设置

设定无人机编队(共10架无人机)从进入雷达覆盖范围到离开共需要100个离散时刻,即$T=100$。将无人机编队全部被摧毁或成功突防视为最终状态,即任务结束。在强化学习中一次任务被称为一幕(Episode)。参考文献[28],设置多功能雷达各工作模式下脉冲信号参数范围如表3所示。PPO算法和

表3 雷达信号参数

Table 3 Radar signal parameters

工作模式	CF/MHz	PRI/μs	PW/μs	PA/dB
搜索	2 800～3 000	500～700	2～10	－5
边搜索边跟踪	2 300～2 600	20～400	8,10	－5～－20
搜索加跟踪	2 600～2 800	30～500	1,6	－2～－22
单目标跟踪	3 000～4 000	30～50	1～5	－22
多目标跟踪	3 600～4 400	50～250	5～20	－20
导弹制导	4 000～4 600	20～40	2～6	－10

Mask-TIT网络超参数设置分别如表4和表5所示。为验证算法的优越性,选取近期研究文献[9-12]的方法作性能对比,这些方法均为智能干扰决策研究领域的代表性方法,并且在比较时将算法参数设置为与原文献一致。

表4 PPO算法超参数

Table 4 Hyperparameters of PPO algorithm

超参数	参数值
初始学习率α	0.01
折扣系数γ	0.9
剪裁函数参数ε	0.2
GAE权衡因子gae-λ	0.95
Adam参数ε	10～5

表5 Mask-TIT网络超参数

Table 5 Hyperparameters of Mask-TIT network

超参数	参数值
Patch大小	1
Embedding维度	32,64
Inner Transformer中head数	4
Outer Transformer中head数	8
激活函数	高斯误差线性单元(Gaussian error linear unit,GELU)

4.2 结果与分析

(1) 干扰对抗模型有效性验证

为验证本文提出的干扰对抗场景和POMDP模型的合理性和真实性,在干扰机不进行任何操作时($a=a_0$),对多功能雷达工作情况进行实验仿真,同时验证两个超参数ω_1和ω_2对雷达探测性能的影响,对ω_1和ω_2分别取1, 1/2, 1/3时进行20次仿真实验,并对结果取均值,实验结果如图8所示。从图8可以看出,在干扰机不进行任何干扰的情况下,所有无人机在35个时间步内被全部制导摧毁。由于多功能雷达会对已搜索到的目标附近加强搜索,因此一旦有一个目标暴露,整个飞行编队的存活率将快速下降。并且随着ω的减小,存活时间变长。ω过大会导致雷达检测概率过大,干扰机来不及反应,ω设置过小导致雷达性能过差,无法体现算法的优劣,综合后续实验考虑,选择$\omega_1=1/2, \omega_2=1/3$进行后续实验。

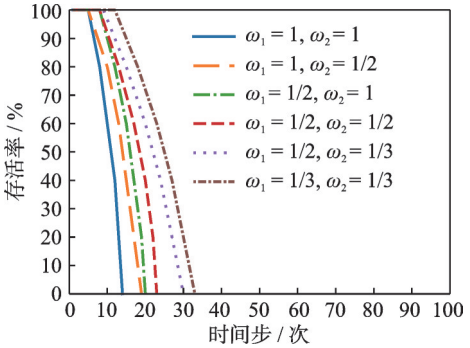


图8 POMDP模型验证

Fig.8 POMDP model validation

(2) 算法性能对比

为验证本文提出的算法有效性和优越性,将该算法应用在图1中的干扰对抗场景下,并与基于DQN[9]、DDQN[10]、D3QN[11]或A3C[12]的干扰决策方法进行对比,实验结果如图9所示。图9(a)为各种算法单次训练500个Episode的结果,从中可以看出强化学习算法的训练过程极不稳定,因此将每种算

法分别进行50轮蒙特卡洛实验，实验结果如图9(b)所示。图9中不同的颜色带表示每种算法在训练过程中的波动范围，颜色带中心的线代表50次训练的平均水平。从图9(b)可以看出，本文提出的算法显著优于其他算法，并且与原始PPO算法相比，有着更快的收敛速度，训练过程中算法的稳定性也有所提升。

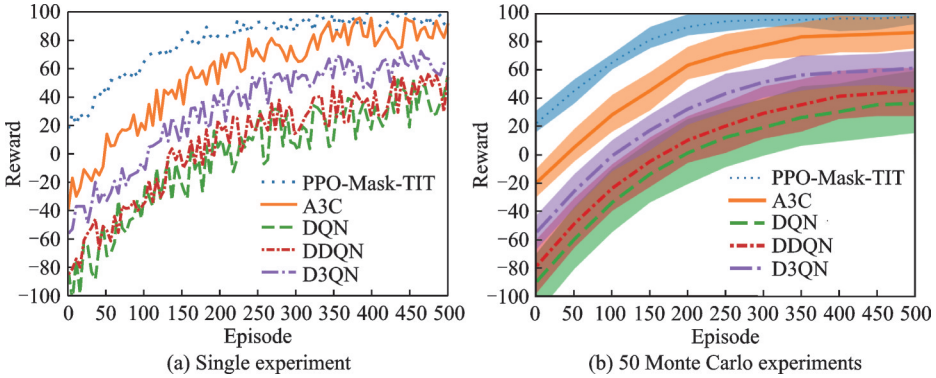


图9 5种算法性能对比

Fig.9 Performance comparison of five algorithms

无人机编队的存活率与Episode之间的仿真结果如图10所示。由于单次实验目标数量较少，因此统计了50次蒙特卡洛实验的总存活率。从图10中可以看出本文所提方法训练240个Episode左右基本收敛，存活率接近100%，显著优于其他算法。最终算法收敛结果对比如表6所示。从表6中可以看出本文提出的PPO-Mask-TIT干扰决策方法收敛所需Episode数相较于各对比算法平均减少25.6%，并且收敛后的方差显著降低。

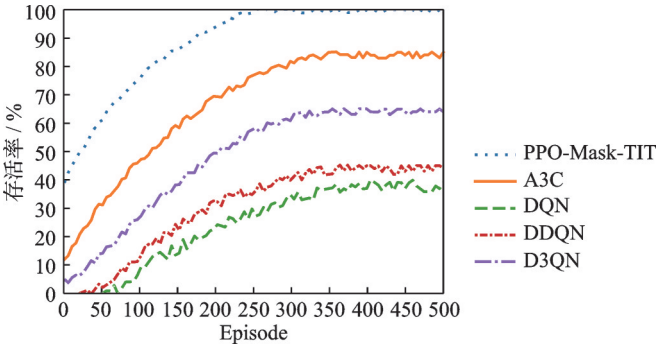


图10 5种算法的总存活率

Fig.10 Total survival rates of five algorithms

离散时刻数量对PPO-Mask-TIT干扰决策方法性能的影响如表7所示。从表7中可以看出，离散时刻数量越大代表无人机编队飞越雷达信号覆盖区域所需的时间越久，整个过程的危险性就越大，对

表6 算法收敛结果对比

Table 6 Comparison of convergence results for five algorithms

算法	收敛奖励		收敛所需Episode		收敛存活率/%
	均值	方差	均值	方差	
DQN[9]	36.6	30.2	355	159	39
DDQN[10]	45.3	26.7	323	122	44
D3QN[11]	61.8	22.3	311	91	67
A3C[12]	86.3	13.3	304	74	86
PPO-Mask-TIT	98.5	3.1	238	32	99

算法的性能要求就越高。

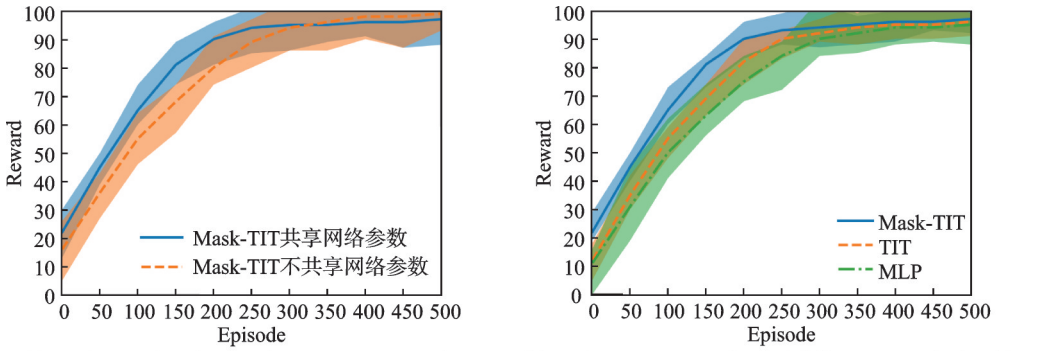
(3) 消融实验

通过消融实验验证Mask-TIT网络的有效性并分析网络参数共享对算法性能的影响,对比共享网络参数和不共享网络参数的训练情况,同时,将Mask-TIT网络与TIT网络和常见的多层感知机网络(Multilayer perceptron,MLP)进行对比。所有实验均使用PPO算法进行训练,并同样进行50轮蒙特卡洛实验,实验结果如图11所示。从图11(a)可以看出,共享网络参数能够加速网络收敛,但最终算法性能会略低于Actor和Critic网络独立的形式。从图11(b)可以得到Mask-TIT网络相较于MLP在算法收敛速度上平均提升11.2%,并且Mask网络层对算法的收敛起到了较大的促进作用,算法稳定性也更强。

表7 离散时刻数量对PPO-Mask-TIT算法性能的影响

Table 7 Effect of number of discrete moments on performance of PPO-Mask-TIT algorithm

离散时刻数量	收敛奖励		收敛所需Episode		收敛存活率/%
	均值	方差	均值	方差	
50	98.8	2.8	189	27	99
100	98.5	3.1	238	32	99
150	96.4	3.9	287	43	95
200	90.1	5.2	346	59	91
250	85.3	9.6	398	71	84



(a) Mask-TIT shared network parameters comparison (b) Performance comparison of different network algorithms

图11 不同网络结构消融实验

Fig.11 Ablation experiment of different network architectures

5 结束语

针对多功能雷达干扰决策问题,本文考虑一种典型的雷达干扰对抗场景——单架电子战飞机实施支援干扰,掩护无人机编队突防雷达警戒区域,将该场景下干扰机与雷达的对抗过程建模为POMDP。在此基础上,提出一种结合PPO算法和Mask-TIT网络的干扰决策方法。实验表明,Mask-TIT-PPO算法较其他对比方法,收敛所需Episode平均减少25.6%,并且收敛后的方差显著降低。Mask-TIT网络相较于MLP在算法收敛速度上平均提升11.2%。

本文主要考虑单架干扰机对抗单部雷达的场景,目前由多部雷达构成雷达组网系统已成为新的趋势,如何应用强化学习理论控制多架干扰机协同应对雷达组网系统是下一步值得深入研究的问题。

参考文献:

[1] HAYKIN S. Cognitive radar: A way of the future[J]. IEEE Signal Processing Magazine, 2006, 23(1): 30-40.

[2] MIRANDA S, BAKER C, WOODBRIDGE K, et al. Knowledge-based resource management for multifunction radar: A look at scheduling and task prioritization[J]. IEEE Signal Processing Magazine, 2006, 23(1): 66-76.

[3] ZHANG Bokai, ZHU Weigang. Research on decision-making system of cognitive jamming against multifunctional radar[C]// Proceedings of 2019 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC).[S.l.]:

IEEE, 2019: 1-6.

[4] XING Qiang, ZHU Weigang, CHI Zhou, et al. Jamming decision under condition of incomplete jamming rule library[J]. Journal of Engineering, 2019, 2019(21): 7449-7454.

[5] 李云杰, 朱云鹏, 高梅国. 基于Q-学习算法的认知雷达对抗过程设计[J]. 北京理工大学学报, 2015, 35(11): 1194-1199.
LI Yunjie, ZHU Yunpeng, GAO Meiguo. Design of cognitive radar countermeasure process based on Q-learning algorithm[J]. Journal of Beijing Institute of Technology, 2015, 35(11): 1194-1199.

[6] XING Qiang, ZHU Weigang, JIA Xin. Research on method of intelligent radar confrontation based on reinforcement learning [C]//Proceedings of 2017 the 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA). [S.l.]: IEEE, 2017: 471-475.

[7] 邢强, 贾鑫. 基于Q-学习的智能雷达对抗[J]. 系统工程与电子技术, 2018, 40(5): 1031-1035.
XING Qiang, JIA Xin. Intelligent radar countermeasure based on Q-learning[J]. Systems Engineering and Electronics, 2018, 40(5): 1031-1035.

[8] LI Huiqin, LI Yanling, HE Chuan, et al. Cognitive electronic jamming decision-making method based on improved Q-learning algorithm[J]. International Journal of Aerospace Engineering, 2021, 2021: e8647386.

[9] 张柏开, 朱卫纲. 对多功能雷达的DQN认知干扰决策方法[J]. 系统工程与电子技术, 2020, 42(4): 819-825.
ZHANG Baikai, ZHU Weigang. DQN-based cognitive jamming decision method for multi-function radars[J]. Systems Engineering and Electronics, 2020, 42(4): 819-825.

[10] 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): 1801-1808.
HUANG Xingyuan, LI Yanyi. Jamming resource allocation strategy based on double Q-learning algorithm[J]. Journal of System Simulation, 2021, 33(8): 1801-1808.

[11] FENG Luwei, LIU Songtao, XU Huazhi. Multifunctional radar cognitive jamming decision based on dueling double deep Q-network[J]. IEEE Access, 2022, 10: 112150-112157.

[12] 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. A3C-based cognitive jamming decision method for multi-function radars[J]. Systems Engineering and Electronics, 2023, 45(1): 86-92.

[13] PUTERMAN M L. Markov decision processes: Discrete stochastic dynamic programming[M]. [S.l.]: John Wiley & Sons, 2014.

[14] KAELBLING L P, LITTMAN M L, CASSANDRA A R. Planning and acting in partially observable stochastic domains[J]. Artificial Intelligence, 1998, 101(1): 99-134.

[15] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[EB/OL]. (2024-11-29). https://arxiv.org/abs/1707.06347.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2024-11-29). https://user.phil.hhu.de/~cwurm/wp-content/uploads/2020/01/7181-attention-is-all-you-need.pdf.

[17] RAM S S, SINGH G, GHATAK G. Estimating radar detection coverage probability of targets in a cluttered environment using stochastic geometry[C]//Proceedings of 2020 IEEE International Radar Conference (RADAR). [S. l.]: IEEE, 2020: 665-670.

[18] ZHANG Zilin, LI Yan, ZHAI Qihang, et al. Mode recognition of multifunction radars for few-shot learning based on compound alignments[J]. IEEE Transactions on Aerospace and Electronic Systems, 2022, 58(6): 5860-5874.

[19] 刘章孟, 袁硕, 康仕乾. 多功能雷达脉冲列的语义编码与模型重建[J]. 雷达学报, 2021, 10(4): 559-570.
LIU Zhangmeng, YUAN Shuo, KANG Shiqian. Semantic coding and model reconstruction of multi-function radar pulse sequences [J]. Journal of Radars, 2021, 10(4): 559-570.

[20] VISNEVSKI N, KRISHNAMURTHY V, WANG A, et al. Syntactic modeling and signal processing of multifunction radars: A stochastic context-free grammar approach[J]. Proceedings of the IEEE, 2007, 95(5): 1000-1025.

[21] SUTTON R S, BARTO A G. Reinforcement learning: An introduction[M]. [S.l.]: MIT Press, 2018.

[22] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:

Human Language Technologies. Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171-4186.

[23] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020,33: 1877-1901.

[24] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[EB/OL]. (2024-11-29). https://arxiv.org/abs/2010.11929.

[25] LIU Ze, LIN Yutong, CAO Yue, et al. Swin Transformer: Hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE,2021: 10012-10022.

[26] HAN Kai, XIAO An, WU Enhua, et al. Transformer in transformer[EB/OL]. (2024-11-29).https://proceedings.neurips.cc/paper/2021/hash/854d9fca60b4bd07f9bb215d59ef5561-Abstract.html.

[27] SCHULMAN J, MORITZ P, LEVINE S, et al. High-dimensional continuous control using generalized advantage estimation [EB/OL]. (2024-11-29).https://arxiv.org/abs/1506.02438.

[28] 廖桂悦. 基于深度学习的雷达辐射源信号精准识别方法研究[D]. 西安: 西安电子科技大学, 2021.
LIAO Guiyue. Research on precise recognition method of radar emitter signals based on deep learning[D]. Xi'an: Xidian University, 2021.

作者简介:

娄雨璇(1999-),男,硕士研究生,研究方向:干扰智能自主决策,E-mail: 572739554@qq.com。

孙闽红(1974-),**通信作者**,男,博士,教授,研究方向:雷达通信信号处理及抗干扰,E-mail: cougar@hdu.edu.cn。

尹帅(2000-),男,硕士生,研究方向:干扰智能自主决策。

(编辑:陈珺)