一种基于深度强化学习的频率捷变雷达智能频点决策方法

张嘉翔[①] 张凯翔[①] 梁振楠*[①] 陈新亮[①③] 刘泉华[①②④]

[①](北京理工大学信息与电子学院雷达技术研究所 北京 100081)

[②](北京理工大学重庆创新中心 重庆 401120)

[③](北京理工大学长三角研究院(嘉兴) 嘉兴 314000)

[④](卫星导航电子信息技术教育部重点实验室(北京理工大学) 北京 100081)

摘要：自卫式干扰机发射的瞄准干扰使多种基于信号处理的被动干扰抑制方法失效，对现代雷达产生了严重威胁，频率捷变作为一种主动对抗方式为对抗瞄准干扰提供了可能。针对传统随机跳频抗干扰性能不稳定、频点选取自由度有限、策略学习所需时间长等问题，该文面向频率捷变雷达，提出了一种快速自适应跳频策略学习方法。首先设计了一种频点可重复选取的频率捷变波形，为最优解提供了更多选择。在此基础上，通过利用雷达与干扰机持续对抗收集到的数据，基于深度强化学习的探索与反馈机制，不断优化频点选取策略。具体来说，通过将上一时刻雷达频点及当前时刻感知到的干扰频点作为强化学习输入，神经网络智能选取当前时刻各子脉冲频点，并根据目标检测结果以及信干噪比两方面评价抗干扰效能，从而优化策略直至最优。从提高最优策略收敛速度出发，设计的输入状态不依赖历史时间步、引入贪婪策略平衡搜索-利用机制、配合信干噪比提高奖励差异。多组仿真实验结果表明，所提方法能够收敛到最优策略且具备较高的收敛效率。

关键词：频率捷变雷达；抗干扰；波形设计；瞄准干扰；深度Q网络

中图分类号：TN958 **文献标识码**：A **文章编号**：2095-283X(2024)01-0227-13

DOI: 10.12000/JR23197

引用格式：张嘉翔, 张凯翔, 梁振楠, 等. 一种基于深度强化学习的频率捷变雷达智能频点决策方法[J]. 雷达学报(中英文), 2024, 13(1): 227–239. doi: 10.12000/JR23197.

Reference format: ZHANG Jiaxiang, ZHANG Kaixiang, LIANG Zhennan, *et al*. An intelligent frequency decision method for a frequency agile radar based on deep reinforcement learning[J]. *Journal of Radars*, 2024, 13(1): 227–239. doi: 10.12000/JR23197.

An Intelligent Frequency Decision Method for a Frequency Agile Radar Based on Deep Reinforcement Learning

ZHANG Jiaxiang[①] ZHANG Kaixiang[①] LIANG Zhennan*[①]
CHEN Xinliang[①③] LIU Quanhua[①②④]

[①](*Radar Research Lab*, *School of Information and Electronics*, *Beijing Institute of Technology*, *Beijing* 100081, *China*)

[②](*Chongqing Innovation Center*, *Beijing Institute of Technology*, *Chongqing* 401120, *China*)

[③](*Yangtze Delta Region Academy of Beijing Institute of Technology*, *Jiaxing* 314000, *China*)

[④](*Key Laboratory of Electronic and Information Technology in Satellite Navigation* (*Beijing Institute of Technology*), *Ministry of Education*, *Beijing* 100081, *China*)

收稿日期：2023-10-10；改回日期：2024-01-03；网络出版：2024-01-11

*通信作者： 梁振楠 liangzhennan@bit.edu.cn *Corresponding Author: LIANG Zhennan, liangzhennan@bit.edu.cn

基金项目：国家自然科学基金(62201048)

Foundation Item: The National Natural Science Foundation of China (62201048)

责任主编：全英汇 Corresponding Editor: QUAN Yinghui

Abstract: The aiming jamming emitted by self-defense jammers renders various passive anti-jamming measures based on signal processing ineffective, posing severe threats to modern radars. Frequency agility, as an active countermeasure, enables the resistance of aiming jamming. In response to issues such as the unstable anti-jamming performance of traditional random frequency hopping, limited freedom in frequency selection, and the long time required for strategic learning, the paper proposes a fast-adaptive frequency-hopping strategy for a frequency agile radar. First, a frequency agile waveform with repeatable frequency selection is designed, providing more choices for an optimal solution. Accordingly, using the data collected through continuous confrontation between a radar and a jammer, and the exploration and feedback mechanism of deep reinforcement learning, a frequency-selection strategy is continuously optimized. Specifically, considering radar frequency from the previous time and jamming frequency perceived at the current time as reinforcement learning inputs, the neural network intelligently selects each subpulse frequency at the current time and optimizes the strategy until it is optimal based on the anti-jamming effectiveness evaluated by the target detection result and Signal-to-Jamming-plus-Noise Ratio (SJNR). To improve the convergence speed of the optimal strategy, the designed input state is independent of the historical time step, the introduced greedy strategy balances the search-utilization mechanism, and the SJNR differentiates rewards more. Multiple sets of simulations show that the proposed method can converge to the optimal strategy and has high convergence efficiency.

Key words: Frequency agile radar; Anti-jamming; Waveform design; Aiming jamming; Deep Q-Network (DQN)

1 引言

在现代战争中，敌方为了获取电磁频谱优势与战场主动权，通常会发射各种有源干扰破坏雷达作战性能，从而掩护目标完成预定的作战任务[1]。雷达为了应对各种干扰，相应的抗干扰技术在对抗中不断升级[2]。一般来说，抗干扰技术按照雷达处理阶段的不同可以分为主动抗干扰和被动抗干扰[3]。在雷达发射信号阶段，主动抗干扰技术可以通过雷达波形设计降低敌方干扰机对雷达信号的截获概率或识别概率，从而降低干扰机的干扰效能[4–6]。如果雷达已经接收到了干扰信号，被动抗干扰技术可以通过空、时、频等多个处理域完成目标与干扰的分离，达到对干扰抑制的目的[7–9]。

随着雷达抗干扰研究的不断深入，被动抗干扰手段日益丰富。然而，挂载在掩护目标上的自卫式干扰机通过发射大功率瞄准干扰，使干扰与目标回波在多处理域重叠，难以分离。频率捷变雷达通过使用自主调节发射信号载频的主动抗干扰手段，使得干扰机难以截获和干扰，为对抗自卫式压制干扰提供了可能[10]。其抗干扰性能主要取决于跳频策略，传统随机跳频策略已经被证明不是最佳选择[11]。如何精准预测干扰机下一时刻将要发射的干扰频点，从而指导雷达信号的频点选择，是频率捷变雷达在与干扰机博弈中取胜的主要难点。

相比针对静态优化问题设计的启发式搜索算法，强化学习可以让智能体与环境不断交互，获得反馈，从而指导智能体在动态环境下进行决策[12]。基于深度学习模型强大的数据表征能力而衍生出的深度强化学习，能够处理高维数据并完成非线性映射，弥补了传统强化学习算法的不足[13]，在认知电子战方面已经得到了一定的研究。如果将干扰信息看作环境状态，抗干扰措施看作雷达动作，抗干扰效能看作即时回报，那么认知抗干扰决策问题可以通过强化学习技术解决。文献[14]针对干扰类型和参数固定的复合干扰场景，分别使用Q学习和SARSA (State-Action-Reward-State-Action)探索了抗干扰措施组合选取问题。文献[15]使用改进的DDPG (Deep Deterministic Policy Gradient)算法对12种抗干扰措施进行选择，以实施抗干扰措施前后干扰威胁度变化作为反馈。文献[16]使用DDPG-MADDPG (Deep Deterministic Policy Gradient and the Multi-Agent Deep Deterministic Policy Gradient)对包含复合干扰在内的12种干扰类型，以抗干扰改善因子作为反馈，进行多处理域抗干扰措施自适应选取。

在频点决策方面，强化学习主要围绕瞄频或扫频干扰的频率捷变波形设计展开研究[17]。文献[18]首次对雷达脉冲级跳频策略展开研究，分别对比了随机频点选择、Q学习、深度Q网络(Deep Q-Network, DQN)等3种策略，证明了DQN在决策方面具备更好的性能。并在文献[19]中继续深化研究内容，将检测概率作为奖励值，而不是之前论文中的信干噪比，同时优化了DQN模型。文献[20]在文献[18]和文献[19]工作的基础上，考虑了一种具备侦收功能的干扰机，以及子脉冲频率捷变雷达，并基于近端策略优化(Proximal Policy Optimization,

PPO)算法完成智能决策。文献[21]考虑了网络化无人机雷达工作系统，使用雷达信息表示理论作为奖励函数，基于双贪婪的改进Q学习算法优化系统抗干扰性能。文献[22]假定干扰机也具备马尔科夫性质，在预测得到干扰策略的基础上选择雷达频点与之对抗。文献[23]考虑了跳频速率会影响相干积分性能和多普勒分辨率，使用Q学习自适应调整雷达发射波形的脉宽和频点以对抗扫频干扰。

总体来说，上述研究均基于雷达不同的性能指标设计奖励函数，以此优化频点等雷达参数。虽然在对抗成功率方面超过随机频点决策方法，然而缺少对抗干扰策略收敛速度的讨论。应当指出，在现代电子战中，干扰机可能具备多种策略，并根据某种规则在不同策略间切换。因此雷达在进行抗干扰策略学习时，应当尽快收敛到最优策略，从而保持对抗先机。如果雷达还未收敛到最优策略时，干扰机改变策略，那么雷达将陷入被动地位。因此，网络收敛时间或是所需样本量是评价一个智能化算法能够应用于实际作战场景的重要衡量指标。

受上述研究启发，考虑到现代干扰机具备侦收-瞄准-干扰的基本策略，本文针对频率捷变雷达，设计了一种基于强化学习的雷达子脉冲跳频抗干扰策略。将当前时刻感知到的干扰频点以及上一时刻的雷达频点作为状态，将当前时刻的雷达频点选择策略作为动作，以目标检测结果和信干噪比作为即时奖励函数设计强化学习关键要素，基于DQN完成子脉冲频点选取策略的学习。仿真针对两种不同侦收策略的干扰机，证明了所提方法的有效性以及较高的收敛效率。

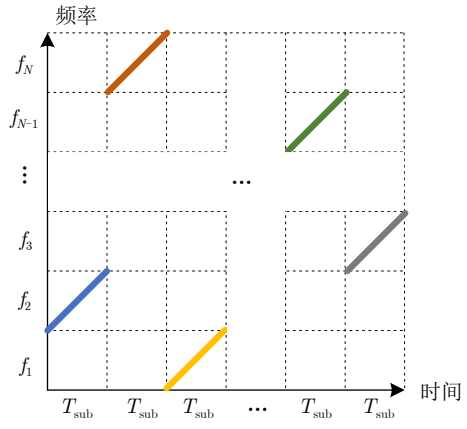
与文献[20]不同的是，本文的主要贡献在于如何通过对强化学习关键要素的设计，从而达到快速收敛到最优解的目的，而不是在于网络设计与修改。具体包括4点：(1)虽然干扰机具备侦干周期，但是我们通过状态空间的合理设计，仅使用单个时间步即可学习到干扰周期性策略，同时不需要使用长短期记忆网络(Long Short-Term Memory, LSTM)等时间记忆网络即可完成最优策略学习，显著降低了收敛时间。(2)在动作设计方面，我们设计了一种子脉冲频点可重复选取的特殊波形，增大了动作空间选取范围。(3)在动作选取方面，我们通过ε-贪婪原则，实现了搜索和利用的有效平衡。在训练初期，以随机搜索为主，减小了收敛到局部最优解的概率。随着训练过程的进行，随机搜索概率逐渐降低，选择网络输出动作的概率逐渐增加，便于收敛。(4)在奖励设计方面，围绕目标检测性能，在单次目标检测结果的基础上，引入了更具差异性的信干噪比指标，缓解了因为采样不充分可能收敛到局部最优解的情况。

2　背景

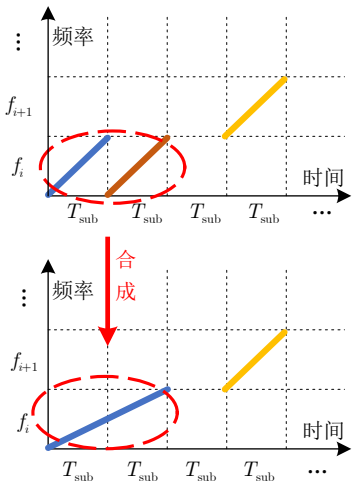
2.1　子脉冲频率捷变波形设计

由于现代干扰机可以对接收到的雷达信号进行快速测频与频率引导，对传统雷达具备较大威胁。而频率捷变雷达可以实现子脉冲级的频率调制，为与其对抗提供了可能。作为常用的雷达传输信号波形，基于线性调频(Linear Frequency Modulation, LFM)信号的子脉冲频率捷变波形如图1(a)所示，其时域表达式如下：

$$s_t(t)=\sum_{n=1}^{N}\mathrm{rect}\left[(t-\tau_n)/T_{\mathrm{sub}}\right]\exp\left[\mathrm{j}2\pi f_n(t-\tau_n)\right]\cdot\exp\left[\mathrm{j}\pi K_n(t-\tau_n)^2\right] \tag{1}$$



(a) 子脉冲频点不可重复
(a) The sub-pulse frequency is not repeatable



(b) 子脉冲频点可重复
(b) The sub-pulse frequency can be repeated

图 1 频率捷变波形示意图

Fig. 1 Schematic diagram of the frequency agility waveform

其中，rect(·)表示矩形窗函数，N表示子脉冲个数，T_{sub}表示子脉冲脉宽；τ_n表示第n个子脉冲的延时，f_n表示子脉冲频点，K_n表示第n个子脉冲的调频斜率。频率捷变雷达各可选频点应当去相关从而达到频率抗干扰的目的，即保证$s_i(\omega)s_j(\omega)=0$，其中，$s_i(\omega)$表示子脉冲i的频谱，$s_j(\omega)$表示子脉冲j的频谱。

式(1)所定义的传统频率捷变雷达在进行子脉冲频点选取时，通常会选择不同的雷达频点。为扩充频点选取自由度，增大波形复杂度，本文设计了一种子脉冲频点可重复选取的雷达发射波形，如图1(b)所示。当相邻子脉冲选取重复频点时，则将其合成一个宽脉冲，其脉宽为$T_{\mathrm{com}}=N_{\mathrm{rep}}T_{\mathrm{sub}}$，其中$N_{\mathrm{rep}}$表示选取相同频点的相邻子脉冲数量。同时保证合成后的宽脉冲带宽不变，即$B_{\mathrm{com}}=B_{\mathrm{sub}}$。合成后的脉冲数用$N_{\mathrm{com}}$表示。

2.2 强化学习与Q学习算法原理

强化学习可以由马尔科夫决策过程(Markov Decision Process, MDP)描述，满足马尔科夫性质。强化学习的优化目标为最大化累计回报，定义为

$$G_t=r_t+\gamma r_{t+1}+\gamma^2 r_{t+2}+\cdots=\sum_{k=0}^{\infty}\gamma^k r_{t+k} \tag{2}$$

其中，r_t表示智能体在状态s_t下执行动作a_t并转移到s_{t+1}后得到的回报；γ为折扣因子，是s_{t+1}及其之后的奖励权重，取值范围为0～1，表示对未来奖励的重视程度。

由于MDP是一种随机过程，其随机独立性导致累计回报G_t是一个随机变量，无法定量描述，如图2所示。因此可对累计回报取期望，获得状态值函数$V_\pi(s)$和动作状态值函数$Q_\pi(s,a)$，将优化问题变成找到一种最优策略π，使任意一个状态的$V_\pi(s)$或$Q_\pi(s,a)$为最大。而Q学习的优化目标是针对$Q_\pi(s,a)$，其贝尔曼方程及最优动作状态值函数$Q_*(s,a)$定义如下：

$$Q_\pi(s,a)=\sum_{s'\in\boldsymbol{S}}p(s'|s,a)\left[r(s,a,s')+\gamma\sum_{a'\in\boldsymbol{A}}\pi(a'|s')Q_\pi(s',a')\right] \tag{3}$$

$$Q_*(s,a)=\sum_{s'\in\boldsymbol{S}}p(s'|s,a)\left[r(s,a,s')+\gamma\max_{a'}Q_*(s',a')\right] \tag{4}$$

其中，$r_t=r(s,a)=\sum_a p(s'|s,a)r(s,a,s')$。$p(s'|s,a)$为某状态$s$执行动作$a$后，转移到下一状态$s'$的概率。

由于在实际场景中，我们可能不知道环境先验信息$p(s'|s,a)$，因此无法获得值函数的解析表示。而Q学习可以通过多次取平均的方式，近似估计得到Q。具体来说，从任意状态开始与环境1个时间步长，利用t时刻的即时回报r_t和下一时刻最大的状态动作值函数$Q(s_{t+1},a'_{t+1})$对当前时刻动作状态值函数$Q(s_t,a_t)$进行估计，最后重复上述动作多次取平均。值函数的更新公式为

$$Q(s_t,a_t)=Q(s_t,a_t)+\alpha\left[r_t+\gamma\max_{a'}Q(s_{t+1},a'_{t+1})-Q(s_t,a_t)\right] \tag{5}$$

其中，α为学习率，表示更新的步长。

Q学习通过不断与环境进行交互来获取并更新Q值，并将Q值存入到由状态和动作组成的Q表中。待智能体学习完成后，根据当前状态的Q值来选取能够获取最大收益的动作。

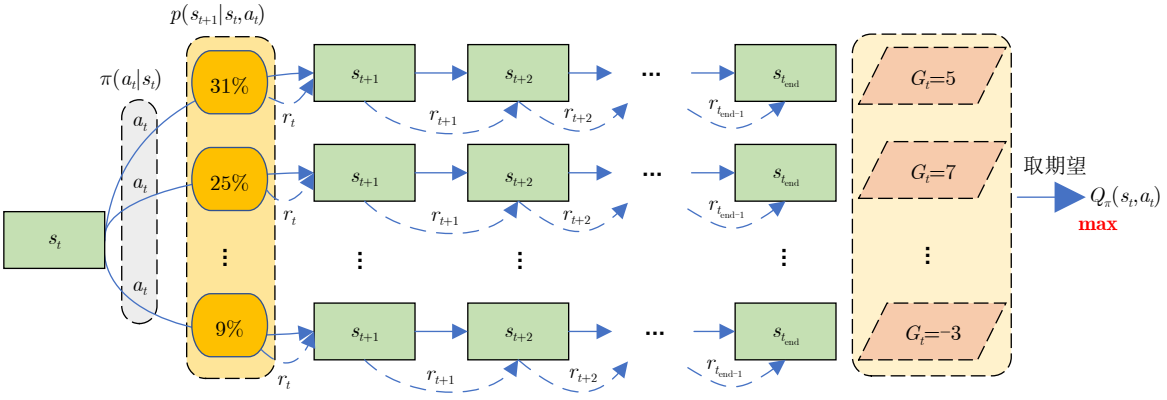


图 2 MDP的随机独立性与强化学习的优化目标

Fig. 2 The random independence of MDP and the optimization objectives of reinforcement learning

3　基于深度Q网络的自适应频点决策

3.1　基于深度Q网络的子脉冲频点决策模型

雷达子脉冲级频点决策往往对应于指数级增长的动作空间，而传统Q学习基于Q表存储和查找Q值，维护难度巨大。而DQN利用神经网络拟合值函数，替换了传统Q表的存储方式，有效解决了高维状态和动作空间的寻优问题。

DQN与Q学习的主要区别在于网络部分，其采用目标值网络和估计值网络组成的双网络。估计值Q网络输出$Q\left(s_t,a_t;\theta\right)$，用来评估当前状态动作对的未来累计回报期望。目标值$\hat{\mathrm{Q}}$网络输出$\hat{Q}\left(s_{t+1},a'_{t+1};\theta^-\right)$，并根据贝尔曼最优方程，使用$y=r_t+\gamma\max\hat{Q}\left(s_{t+1},a'_{t+1};\theta^-\right)$表示$Q$函数的优化目标。其网络训练过程如图3所示。

输入当前状态s_t，通过估计值网络预测得到当前状态s_t对应的不同动作a_t的Q值，然后通过ε-贪婪原则选择a_t并转至下一状态s_{t+1}，同时获得r_t。通过目标值网络计算下一状态s_{t+1}的最大$\hat{Q}$值，将其与估计值作差更新估计值网络参数θ，表示为

$$L=\left[r_t+\gamma\max_{a'}\hat{Q}\left(s_{t+1},a'_{t+1};\theta^-\right)-Q\left(s_t,a_t;\theta\right)\right] \tag{6}$$

其中，ε-贪婪原则以概率$1-\varepsilon$选择估计值网络输出的具有最大Q值的频点，以概率ε随机选择频点，并随着训练步数的增加减小ε，从而达到搜索和利用的充分结合。

上述流程经过一定次数后，基于软更新来更新目标值网络参数θ^-：

$$\theta^-=\tau\theta+\left(1-\tau\right)\theta^- \tag{7}$$

其中，$0<\tau\ll1$表示软间隔更新系数。由于在一段时间内目标值具有一定稳定性，这能在一定程度上降低估计值Q网络和目标值$\hat{\mathrm{Q}}$网络之间的耦合性，提升了网络的收敛性和稳定性。

训练完成后，测试时直接输入当前时刻状态至训练好的模型中，即可获取最优动作。

3.2　强化学习关键要素设计

上述提及的状态、动作和奖励是强化学习的关键要素，其中状态和奖励是算法的输入，动作是算法的输出。设置如下：

(1) 状态空间：假设雷达能够通过干扰感知等手段获取干扰频点信息，则状态空间由雷达子脉冲频点和干扰频点组成。

$$\begin{aligned}\boldsymbol{S}&=\left[f_{\mathrm{R},t-1},f_{\mathrm{J},t}\right]\\&=\left[f_{\mathrm{sub}1,t-1},f_{\mathrm{sub}2,t-1},\cdots,f_{\mathrm{sub}N,t-1},f_{\mathrm{J},t}\right]\end{aligned} \tag{8}$$

其中，$f_{\mathrm{R},t-1}=\left[f_{\mathrm{sub}1,t-1},f_{\mathrm{sub}2,t-1},\cdots,f_{\mathrm{sub}N,t-1}\right]$和$f_{\mathrm{J},t}$分别表示$t-1$时刻雷达$N$个子脉冲的频点选择以及$t$时刻干扰瞄准频点。$f_{\mathrm{J},t}$取值范围为1～$(N+1)$，1～$N$表示干扰机发射窄带瞄频干扰的瞄准频点，$(N+1)$表示干扰机发射宽带阻塞干扰。$f_{\mathrm{sub}n,t}\,(1\le n\le N)$的取值范围为1～$N$，表示第$n$个子脉冲的频点。

(2) 动作空间：t时刻雷达N个子脉冲频点选择：

$$\boldsymbol{A}=f_{\mathrm{R},t}=\left[f_{\mathrm{sub}1,t},f_{\mathrm{sub}2,t},\cdots,f_{\mathrm{sub}N,t}\right] \tag{9}$$

(3) 奖励函数：奖励函数应当围绕雷达作战任务设置，本文以预警雷达为例，采用目标检测结果F_{d}和信干噪比(Signal-to-Jamming-plus-Noise Ratio, SJNR)作为评价指标。前者直接反映了目标检测能力，而后者的存在加快了最优解的收敛速度，降低收敛到局部最优解的可能，从而最大化目标检测性能。定义如下：

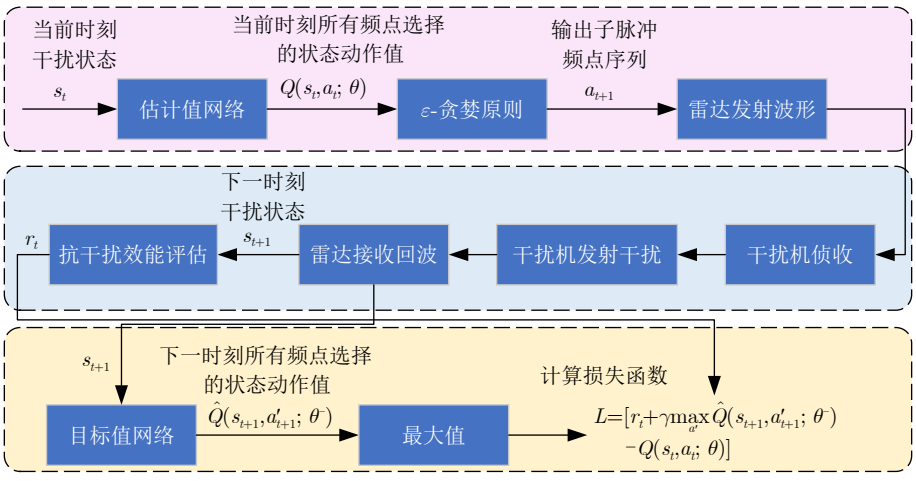


图 3 DQN网络参数的更新过程

Fig. 3 The network parameter update process of DQN

$$R=\sum_{n=1}^{N_{\mathrm{com}}}\left(N_{\mathrm{rep},n}F_{\mathrm{d},n}-\mathrm{SJNR}_n/N_{\mathrm{com}}\right) \tag{10}$$

$$\mathrm{SJNR}_n=\begin{cases}\left(P_{\mathrm{T},n}-\bar{P}_{\mathrm{JN},n}\right)/\eta, & F_{\mathrm{d},n}=1\\ 0, & F_{\mathrm{d},n}=-1\end{cases} \tag{11}$$

其中，对于目标检测结果F_{d}，我们可以根据提前获取的战场态势信息预估目标距离波门，在子脉冲脉压后基于单元平均恒虚警率(Cell Average-Constant False Alarm Rate, CA-CFAR)检测判断目标能否被检测到[24]。如果第n个子脉冲检测到目标则$F_{\mathrm{d},n}=1$，反之则$F_{\mathrm{d},n}=-1$。同时可以获取目标平均功率$P_{\mathrm{T},n}$和干扰噪声平均功率$\bar{P}_{\mathrm{JN},n}$。η为归一化系数，用来将信干噪比限制在0～1之间，从而提高训练稳定性。

结合状态、动作和奖励的定义，基于深度Q网络的雷达子脉冲频点决策流程如算法1所示。

算法 1 基于深度Q网络的雷达子脉冲频点决策

Alg. 1 Radar sub-pulse frequency decision based on Deep Q-Network (DQN)

Step 1：初始化：

Step 1-1：使用随机参数θ初始化估计值Q网络

Step 1-2：使用参数$\theta^{-}=\theta$初始化目标值$\hat{\mathrm{Q}}$网络

Step 1-3：初始化经验池D

Step 1-4：初始化干扰策略，雷达子脉冲数量及频点，折扣因子γ，学习率α，贪婪因子ε，软间隔更新系数τ等参数

Step 2：每幕：

Step 2-1：设置初始状态$s_1=\left[f_{\mathrm{R},0},f_{\mathrm{J},1}\right]$

Step 2-2：每个时间步：

Step 2-2-1：使用ε-贪婪原则依据估计值网络的输出结果选择各子脉冲频点$a_t=f_{\mathrm{R},t}=\left[f_{\mathrm{sub}1,t},f_{\mathrm{sub}2,t},\cdots,f_{\mathrm{sub}N,t}\right]$，即以$1-\varepsilon$概率选择估计值网络输出的最佳的频点或者以$\varepsilon$概率随机选择频点

Step 2-2-2：雷达发射子脉冲频率捷变波形，接收到回波后，感知得到下一时刻状态s_{t+1}并根据目标检测结果和脉压后的信干噪比评估当前时刻奖励r_t

Step 2-2-3：将(s_t,a_t,r_t,s_{t+1})存储到经验池D中，如果经验池中的样本数超出预定数量，则删除早期训练样本数据，以便存储并使用最新样本数据

Step 2-2-4：如果经验池D中保存数量超过起始值，则从D中选择批大小(batchsize)个样本作为训练集输入到估计值和目标值网络中，分别计算得到$Q\left(s_t,a_t;\theta\right)$和$y=r_t+\gamma\max\hat{Q}(s_{t+1},a'_{t+1};\theta^{-})$，并反向梯度求导使误差函数$L\left(\theta\right)=[y-Q\left(s_t,a_t;\theta\right)]^2$趋近0，更新估计值网络参数$\theta$

Step 2-2-5：每隔一定的时间步软更新目标值网络参数θ^{-}

Step 2-3：结束该时间步

Step 2-4：降低贪婪概率ε

Step 3：结束该幕

4 仿真与分析

4.1 场景设置

4.1.1 仿真参数设置

本文以3个子脉冲和3个可选频点为例，讨论DQN应用于子脉冲频点自适应选取的可行性。为避免子脉冲脉压后出现虚假目标，非相邻子脉冲不能选取重复频点，因此动作总数为$3^3-6=21$。频率捷变信号、干扰、DQN的仿真参数分别如表1—表3所示。其中，每幕表示1个相参处理间隔(Coherent Processing Interval, CPI)，时间步t表示某个CPI中的第t个脉冲重复周期。

表 1 频率捷变信号参数设置

Tab. 1 The parameter settings of frequency agile signal

参数	数值
子脉冲调制类型	LFM
子脉冲个数	3
子脉冲频点	[10 MHz, 30 MHz, 50 MHz]
子脉冲脉宽	5 μs
子脉冲带宽	5 MHz
信噪比	0 dB

表 2 干扰参数设置

Tab. 2 The parameter settings of jamming

干扰类型	参数	数值
窄带瞄频	瞄准频点	[10 MHz, 30 MHz, 50 MHz]
	带宽	10 MHz
	干噪比	35 dB
宽带阻塞	带宽	120 MHz
	干噪比	30 dB

很重要的一个技巧是，本文在基于贪婪原则随机选取动作时，只考虑所有子脉冲选择相同频点的情况，即脉内不跳频。该处理旨在尽可能提高相参处理增益以及使干扰机侦收到单频信号并诱导其发射窄带瞄频干扰，从而加快最优策略学习。同样出于加速收敛的目的，输入到神经网络的奖励按照子脉冲个数进行了归一化。

估计值网络和目标值网络的结构相同，均使用4层全连接神经网络，分别为输入层、2个隐藏层和输出层。其中，隐藏层的神经元个数均为64，并使用ReLU作为激活函数，如图4所示。

4.1.2 干扰策略设置

考虑一个具备侦收功能的干扰机，并根据侦-干时间长短分别设置了脉内侦干和脉间侦干等两种固定干扰策略，分别如图5、图6所示。由于切片转发干扰的对抗效果受限于切片宽度、转发次数等参数，灵活的参数变化可能会导致对抗失效，因此本文考虑的干扰类型为压制干扰，包括窄带瞄频和宽

表 3　DQN参数设置

Tab. 3　The parameter settings of DQN

参数	数值
批大小	64
学习率	0.001
折扣因子	0.99
缓冲区大小	10000
起始训练样本量	64
贪婪因子衰减系数	0.2
幕	32个时间步
目标值网络更新周期	4个时间步
目标值网络软间隔更新系数	0.01
隐藏层数量	2
隐藏层神经元个数	64
归一化系数	80

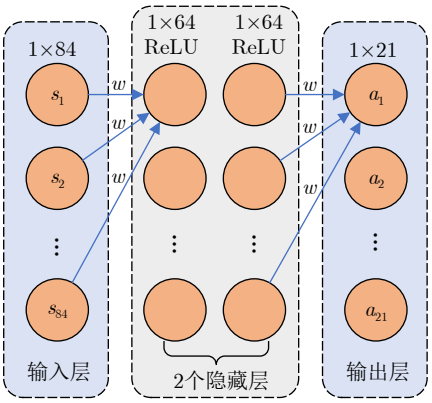


图 4 全连接神经网络结构示意图

Fig. 4 The schematic diagram of fully connected neural network structure

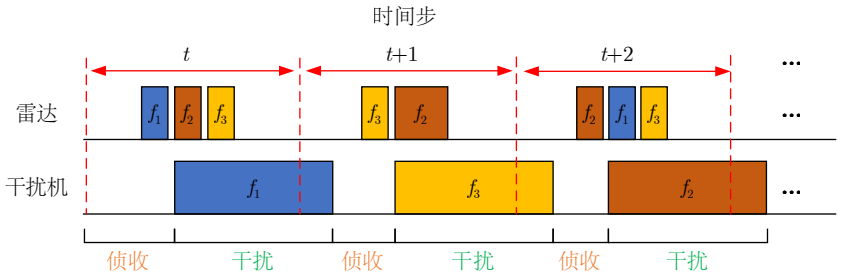


图 5 脉内侦干策略

Fig. 5 The intra-pulse interception-jamming strategy

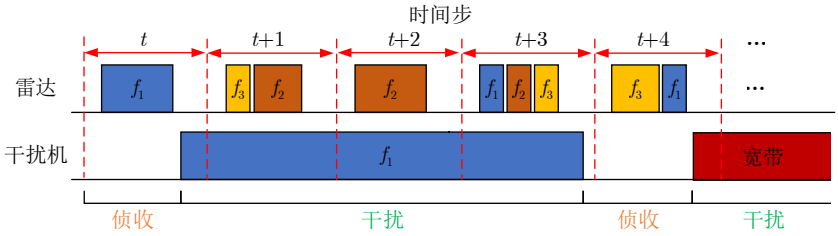


图 6 脉间侦干策略

Fig. 6 The pulse-to-pulse interception-jamming strategy

带阻塞。其中，窄带瞄频干扰的带宽为雷达子脉冲带宽的2倍，更宽的带宽会使得全部状态的奖励值发生整体偏移，但在归一化后会消除该影响。

对于脉内侦干策略，假设干扰机侦收到雷达脉冲上升沿及下降沿，立即对其测频，转发对应频点的窄带瞄频干扰。值得注意的是，干扰时长设置略小于1个脉冲重复周期(Pulse Repetition Time, PRT)，从而使得在当前PRT会同时受到上一时刻以及当前时刻的干扰。因此，雷达在该干扰策略下的一种较为合适的选择为后续子脉冲发射不同于子脉冲1的雷达频点，并且每个PRT均保持相同的发射策略。由于干扰所在频点在滤波后可能会在邻近频点上存在干扰功率残留，因此最优策略为雷达后续子脉冲跳频到距离子脉冲1所选频点的最远频点上。即雷达最优频点选择为$[1,N,N]$或$[N,1,1]$。

对于脉间侦干策略，假设干扰机从侦收到第1个子脉冲开始持续侦收一段时间，直至没有检测到子脉冲时侦收结束。根据侦收结果发射一段时间长度的干扰，干扰时长在3～4个PRT之间。相比脉内侦干策略，后者不会在某个PRT同时受到两部分干扰。在侦收阶段若只侦收到1个频点，则发射对应频点的窄带瞄频干扰，反之则发射宽带阻塞干扰。雷达需要尽量避免干扰机发射宽带阻塞干扰，为此雷达需要在干扰机侦收阶段时只发射单频信号，而在干扰阶段时选择其余频点。类似地，考虑到滤波引起的干扰功率残留，在干扰机侦收时雷达最优策略为$[1,1,1]$或$[N,N,N]$，对应的干扰时雷达最优策略为$[N,N,N]$或$[1,1,1]$。

值得注意的是，脉间侦干策略虽然具备周期性，但当前时刻的干扰动作不完全取决于上一时刻

的状态，而是按照固定的时序执行侦收和干扰，因此不具备马尔科夫性。脉间侦干策略寻求的是由4个PRT组成的侦干周期的最大奖励，满足式(5)所示的贝尔曼最优方程的价值迭代原理，因此可以使用强化学习解决。

4.2 脉内侦干策略

此时干扰机侦收到1个子脉冲的上升沿与下降沿后，完成测频并立刻发射干扰，雷达频点对抗的训练结果如图7所示。得分曲线在第4个CPI左右即可收敛，在36分附近波动，如图7(a)所示。图7(b)展示了文献[20]提出的基于PPO与LSTM相结合的频点决策算法，其至少需要30幕的时间才能提升到32分附近震荡，因此策略学习耗时且鲁棒性较差。其本质原因在于PPO为on-policy算法，只能利用神经网络进行动作搜索，导致探索性不足，所以存在收敛速度慢、可能会收敛到局部最优解、得分无法保持等诸多问题。

根据图7(a)的收敛情况，保存前10个CPI的训练模型，每个模型对抗100幕，对抗成功率如图8所示。根据4.1.2节对脉内侦干策略的分析，雷达应将未被侦收到的子脉冲频点设置为距离侦收频点的最远频点。因此，PRT对抗成功定义为$\{f_{\mathrm{R}}=[1,3,3]\,\&\,f_{\mathrm{J}}=1\}$或$\{f_{\mathrm{R}}=[3,1,1]\,\&\,f_{\mathrm{J}}=3\}$，即21个动作中只有2个动作为最优，占比9.5%。CPI对抗成功的判决依据是当前CPI内所有PRT均对抗成功。

发现训练所用CPI数量对对抗成功率的影响与收敛情况基本对应，从第3个CPI开始，对抗成功率即可达到100%。

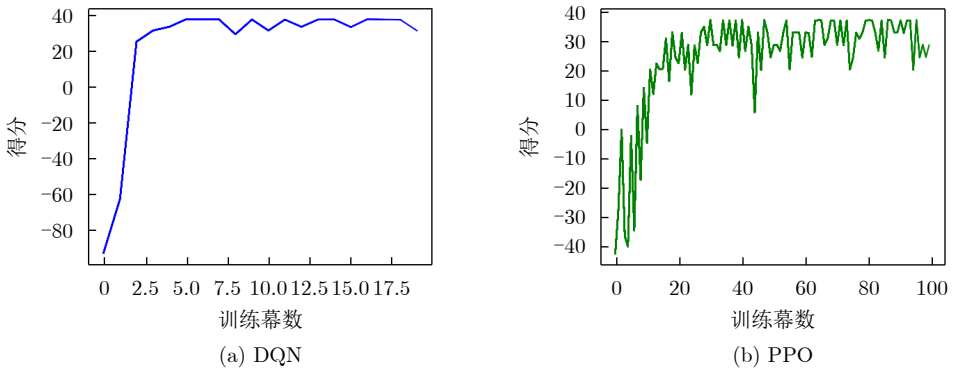
表4展示了随机频点、PPO-LSTM和DQN的单次对抗(PRT)成功率，单幕(CPI)对抗成功率。随机频点决策的成功率与最优动作占比，即理论值大致相同。基于PPO的频点决策虽然在第2个和第3个子脉冲避开了干扰频点，但是由于其搜索力度不够，有一定概率选取到次优策略。而基于DQN的频点决策算法由于使用了ε-贪婪算法，大大扩展了动作搜索空间，更容易收敛到最优策略。

PPO算法由于可以处理连续动作空间问题，并且可以学习到随机策略，因此是强化学习中受众面最广的基线方法。然而在本文研究的频点决策场景中，不涉及连续动作空间，最优策略也可以由随机策略退化到确定性策略，因此PPO算法优势没有得到充分利用。更为重要的是，由于每幕对抗中次优策略不低于最优策略得分的10%，大大提高了仅依靠神经网络参数进行动作搜索的最优策略收敛难度。

图9(a)展示了雷达和干扰在4个PRT下的频点选取情况。对于第1个PRT，由于初始状态的随机性，雷达选取频点[1,2,3]，干扰瞄准频点1。由于单个子脉冲的信噪比增益有限，因此除被干扰的子脉冲外，另有1个子脉冲未能检测到目标，奖励为负值，如图9(b)所示。在第2, 3, 4个PRT，基于训练好的模型，雷达的第2个和第3个子脉冲均选择离干扰频点1最远的频点3，降低了干扰剩余能量的同时，合成了宽脉冲，提高了信噪比增益。

最优动作的时频图及一维距离像如图10所示。当前PRT会同时收到瞄准上一时刻第1个子脉冲以及瞄准当前时刻第1个子脉冲的窄带瞄频干扰，后者会在瞄准后立即发射。因此，第1个子脉冲脉压后，目标尖峰出现在当前时刻产生的大功率噪声干扰边缘，导致漏检。第2个子脉冲由于跳频策略与干扰频域正交，因此脉压后能够检测到目标尖峰，具有较高的信干噪比。

本文围绕目标检测性能，基于单个PRT能否检测到目标以及脉压后的信干噪比两方面评价跳频抗干扰效能。表5展示了蒙特卡洛1000次下，雷达的几个典型频点选取策略的目标检测率、脉压后的信干噪比以及平均得分。为便于分析，假设当前时刻



(a) DQN (b) PPO

图 7 脉内侦干策略的子脉冲频点决策训练结果

Fig. 7 The training results of sub-pulse frequency decision for the intra-pulse interception-jamming strategy

和上一时刻均干扰相同的频点，频点[3,1,1]和[1,3,3]为本文所提模型的策略。可以看出：

(1) 由于在当前PRT能同时受到上一时刻和当前时刻的干扰，因此至少有一个雷达频点会被干扰到。根据式(10)所示的奖励函数计算方式，最大得分始终小于2；

(2) 当子脉冲2和子脉冲3跳频成功时，两个子脉冲均选择离干扰频点的最远频点时，平均得分最高，为最优策略，即[1,3,3]和[3,1,1]；

(3) 诸如[1,2,3]和[2,1,3]等传统频点选取策略，由于脉压增益有限，导致目标检测率较低；而[1,2,2]和[2,1,1]等选择了干扰频点相邻频点的动作，由于滤波后的干扰能量残余，从而降低了信干噪比，非最优策略；

(4) 次优策略和最优策略的单次对抗得分仅差0.06，网络能够捕获到细微差异，收敛到最优解。

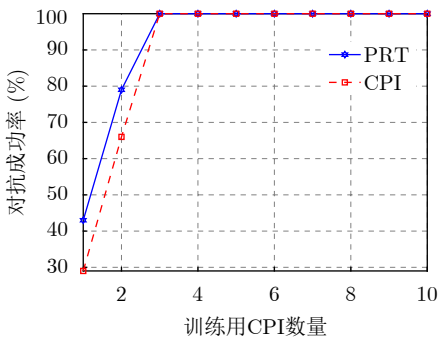


图 8 训练用CPI数量对脉内侦干策略下对抗成功率的影响

Fig. 8 The impact of the number of CPI used for training on the success rate of confrontation for the intra-pulse interception-jamming strategy

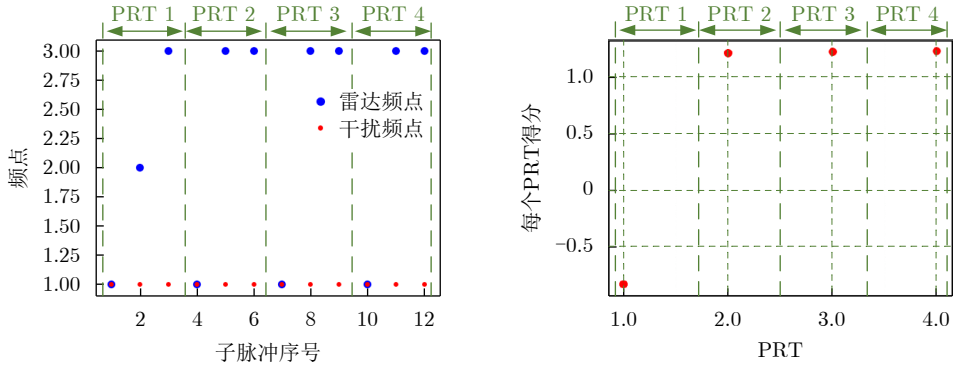
4.3 脉间侦干策略

针对脉间侦干策略，DQN和PPO的训练曲线

表 4 脉内侦干策略的对抗成功率(%)

Tab. 4 The success rate of confrontation for the intra-pulse interception-jamming strategy (%)

策略	PRT对抗成功率	CPI对抗成功率
随机频点	9.7	0
PPO	94	9
DQN	100	100



(a) 雷达子脉冲频点和干扰频点选取

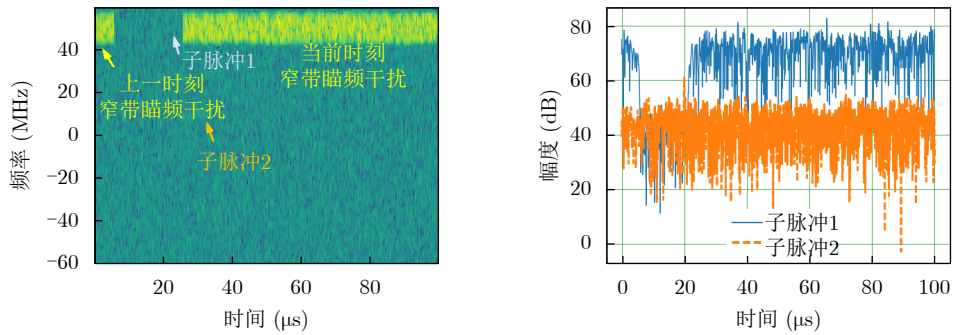
(a) The selection of radar sub-pulse frequencies and jamming frequencies

(b) 每次对抗的奖励

(b) Rewards for each confrontation

图 9 雷达与干扰对抗4个PRT的策略及对抗奖励

Fig. 9 The strategies and rewards for radar anti-jamming during four PRT periods



(a) 回波时频图

(a) The time-frequency map of echo

(b) 一维距离像

(b) The high range resolution profile

图 10 雷达执行最优策略的时频图及一维距离像

Fig. 10 The time-frequency map and the one-dimensional High-Resolution Range Profile (HRRP) for radar executing optimal strategy

表 5 脉内侦干策略下各种雷达策略对抗1000次结果(f_J=f_{sub1})

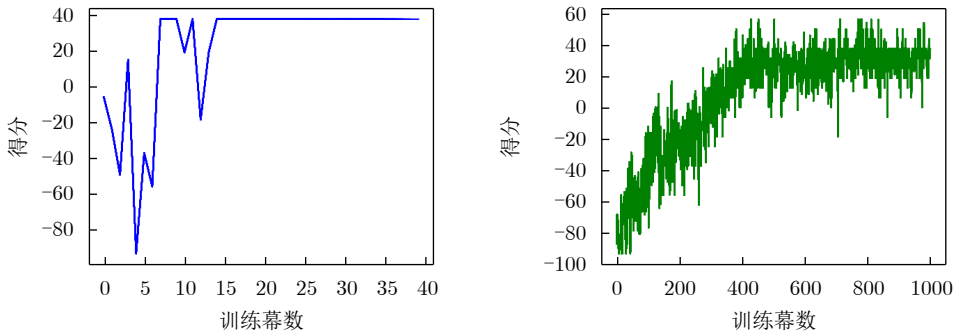
Tab. 5 The results of 1000 confrontations with various radar strategies for the intra-pulse interception-jamming strategy (f_J=f_{sub1})

雷达频点选择	目标检测率(%)	信干噪比(dB)	平均得分
[1,1,1]	0	—	–3.00
[1,1,2]	0	11.09	–1.12
[1,1,3]	0	12.25	–0.96
[1,2,2]	97.6	15.20	1.09
[1,2,3]	81.7	12.78	0.78
[1,3,3]	**99.7**	**16.06**	**1.19**
[2,1,1]	98.3	15.35	1.12
[2,1,3]	75.6	12.47	0.64
[2,3,3]	97.7	15.19	1.10
[3,1,1]	**99.6**	**16.07**	**1.18**

注：综合考虑噪声随机性引起的得分波动情况，加粗项为最优策略

如图11所示。DQN在第15幕(CPI)左右即可收敛，得分在37分附近。而PPO的训练过程虽然整体呈现上升-平稳，但是其波动始终较为剧烈，且至少需要400幕左右才能趋于平稳。

图12展示了训练所用CPI数量对对抗成功率的影响，蒙特卡洛次数为100幕。由于雷达初始频点随机选取，不参与决策，因此去除包含初始状态在内的第1个干扰侦干周期。从第2个周期开始统计，即每幕(CPI)对抗28次。根据4.1.2节对脉间侦干策略的分析，雷达应始终发射单频信号，并在干扰机对当前脉冲侦收干扰后的下个脉冲跳到另一频点，从而诱导干扰机在后续干扰周期内发射窄带瞄频干扰，避免发射宽带阻塞干扰导致跳频手段失效。由于干扰机可以在侦收后立即发射对应频点的干扰，所以每个侦干周期内，无论采取何种手段，至少会存在1个PRT抗干扰失败。因此可以仅针对剩余PRT计算抗干扰成功率，将PRT对抗成功定义为干扰机处于发射干扰阶段时雷达选取到最优策略，即$\{f_J=3\ \&\ f_R=[1,1,1]\}$或$\{f_J=1\ \&\ f_R=[3,3,3]\}$；CPI对抗成功的判决依据是当前CPI内所有PRT均对抗成功。



(a) DQN (b) PPO

图 11 脉间侦干策略的子脉冲频点决策训练结果

Fig. 11 The training results of sub-pulse frequency decision for the pulse-to-pulse interception-jamming strategy

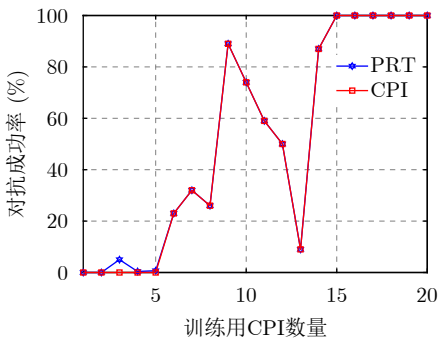


图 12 训练用CPI数量对脉间侦干策略对抗成功率的影响

Fig. 12 The impact of the number of CPI used for training on the success rate of confrontation for the pulse-to-pulse interception-jamming strategy

可以发现，在前20个CPI的训练过程中模型学习到的策略不是一直向好，而是波动变化。在第13个PRT策略出现了明显恶化，这与图11(a)的训练结果相一致。此时模型尚未稳定学习到干扰机的侦干策略，因此仍主要处于试错探索阶段。从第15～20个CPI，模型探索到干扰机策略，并学习到有效对抗策略，保持稳定。

100次蒙特卡洛仿真下的随机频点、PPO和DQN决策的单次对抗(PRT)成功率，单幕(CPI)对抗成功率如表6所示。由于对抗成功率隐含雷达在干扰机侦-干PRT和干扰PRT均发射不同的单频信号，因此随机频点选择的成功概率极低，仅有0.7%。相比PPO，DQN动作搜索更加充分，使对抗成功率得到有效提高，达到100%。

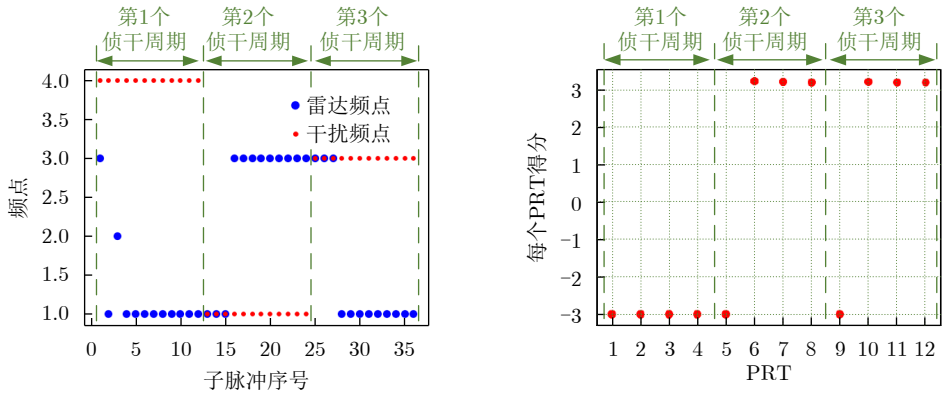
图13(a)展示了干扰机的3个侦干周期下的雷达子脉冲频点选取和干扰瞄准频点。在第1个侦干周期中，由于雷达初始状态的随机性，3个子脉冲分

表 6　脉间侦干策略的对抗成功率(%)

Tab. 6　The success rate of confrontation for the pulse-to-pulse interception-jamming strategy (%)

策略	PRT对抗成功率	CPI对抗成功率
随机频点	0.7	0
PPO	93.6	31
DQN	100	100

别选取不同频点，导致干扰机在接下来的3个PRT中发射宽带阻塞干扰，此时无论雷达如何跳频，目标均未被检测到，奖励为负值，如图13(b)所示。在第2个侦干周期的第1次对抗中，雷达3个子脉冲均选择频点1，干扰机侦收到并立刻发射对应频点的干扰，因此第1个PRT的奖励为负值。接下来的3个PRT，干扰机继续发射频点1，而雷达选择离频点1最远的频点3。至此第2个侦干周期结束，雷达频点选取成功。在第3个侦干周期中，雷达和干扰的频点选取对调，雷达仍然能够通过频点决策选择受到干扰最小的频点。



(a) 雷达子脉冲频点和干扰频点选取
(a) The selection of radar sub-pulse frequencies and jamming freuencies

(b) 每次对抗的奖励
(b) Rewards for each confrontation

图 13 对抗3个侦干周期的雷达策略及对抗奖励

Fig. 13 The strategies and rewards for radar anti-jamming during three interception-jamming periods

以干扰瞄准频点1为例，蒙特卡洛1000次，统计各种策略对抗的目标检测率、脉压后的信干噪比以及平均得分，如表7所示，其中频点[3,3,3]为本文所提模型的策略。可以看出：

(1) 对于传统雷达跳频策略[1,2,3]，有1个子脉冲会被干扰到，此时奖励虽然为正值，但是较低；

(2) 对于[2,2,2]，虽然从频点数值上看确实跳频成功，但此时瞄准频点1的干扰功率可能未被全部滤掉，有很少一部分的功率会溢出到频点2，使得其信干噪比略低于频点3；

(3) 当雷达所有子脉冲均选择频点3时，接收到的干扰平均功率达到最小值，平均得分最高，为最优策略。

表 7　脉间侦干策略下各种雷达策略对抗1000次的结果(f_J=1)

Tab. 7　The results of 1000 confrontations with various radar strategies for the pulse-to-pulse interception-jamming strategy (f_J=1)

雷达频点选择	目标检测率(%)	信干噪比(dB)	平均得分
[1,1,1]	0	—	–3.00
[1,2,3]	81.3	12.74	0.76
[2,2,2]	99.7	17.08	3.17
[3,3,3]	**100**	**17.58**	**3.22**

注：加粗项表示最优策略

5　结语

针对瞄准式压制干扰，本文面向频率捷变雷达，提出了一种基于深度强化学习的频点自适应快速选取方法。根据当前时刻干扰状态，以及上一时刻雷达动作，依靠神经网络自适应选取当前时刻最优雷达频点，并基于目标检测结果以及脉压后的信干噪比作为奖励反馈，迭代改进策略。仿真部分考虑了具备侦收-瞄准-干扰功能的干扰机，证明了通过关键要素设计可以以单个时间步长作为输入学习到干扰策略的时序性。同时，所用DQN算法配合贪婪准则实现了搜索-利用的平衡，配合信干噪比的反馈加速最优抗干扰策略收敛，相比PPO算法收敛速度提升至少10倍。考虑到实际场景中，干扰频点在滤波后可能在邻近频点存在能量残余的情况，所提频率捷变波形设计方法允许子脉冲多次重复选取距离干扰频点最远的雷达频点，有效降低了回波中的干扰剩余能量，提高了信干噪比。同时扩展了动作空间，提供了最优动作选取的基础。

通过本文研究发现，当子脉冲数或脉冲数较多时，增大了网络的搜索和决策空间，使得收敛时间进一步增加，并且提高了最优策略的收敛难度。但这不会影响强化学习的关键要素设计，因此所提方

法仍能根据交互数据的反馈结果进行策略优化。另外，考虑到子脉冲间、脉冲间的相位不一致，在积累时会带来一定程度上的增益损失。因此在未来的研究中，考虑将子脉冲以及脉冲间的积累情况纳入到奖励函数中，从而指导策略选取。

利益冲突 所有作者均声明不存在利益冲突

Conflict of Interests The authors declare that there is no conflict of interests

参 考 文 献

[1] 李永祯, 黄大通, 邢世其, 等. 合成孔径雷达干扰技术研究综述[J]. 雷达学报, 2020, 9(5): 753–764. doi: 10.12000/JR20087.
LI Yongzhen, HUANG Datong, XING Shiqi, *et al*. A review of synthetic aperture radar jamming technique[J]. *Journal of Radars*, 2020, 9(5): 753–764. doi: 10.12000/JR20087.

[2] 崔国龙, 余显祥, 魏文强, 等. 认知智能雷达抗干扰技术综述与展望[J]. 雷达学报, 2022, 11(6): 974–1002. doi: 10.12000/JR22191.
CUI Guolong, YU Xianxiang, WEI Wenqiang, *et al*. An overview of antijamming methods and future works on cognitive intelligent radar[J]. *Journal of Radars*, 2022, 11(6): 974–1002. doi: 10.12000/JR22191.

[3] 李康. 雷达智能抗干扰策略学习方法研究[D]. [博士论文], 西安电子科技大学, 2021. doi: 10.27389/d.cnki.gxadu.2021.003098.
LI Kang. Research on radar intelligent antijamming strategy learning method[D]. [Ph.D. dissertation], Xidian University, 2021. doi: 10.27389/d.cnki.gxadu.2021.003098.

[4] JIANG Wangkui, LI Yan, LIAO Mengmeng, *et al*. An improved LPI radar waveform recognition framework with LDC-Unet and SSR-Loss[J]. *IEEE Signal Processing Letters*, 2022, 29: 149–153. doi: 10.1109/LSP.2021.3130797.

[5] GARMATYUK D S and NARAYANAN R M. ECCM capabilities of an ultrawideband bandlimited random noise imaging radar[J]. *IEEE Transactions on Aerospace and Electronic Systems*, 2002, 38(4): 1243–1255. doi: 10.1109/TAES.2002.1145747.

[6] GOVONI M A, LI Hongbin, and KOSINSKI J A. Low probability of interception of an advanced noise radar waveform with linear-FM[J]. *IEEE Transactions on Aerospace and Electronic Systems*, 2013, 49(2): 1351–1356. doi: 10.1109/TAES.2013.6494419.

[7] CUI Guolong, JI Hongmin, CAROTENUTO V, *et al*. An adaptive sequential estimation algorithm for velocity jamming suppression[J]. *Signal Processing*, 2017, 134: 70–75. doi: 10.1016/j.sigpro.2016.11.012.

[8] YU K B and MURROW D J. Adaptive digital beamforming for angle estimation in jamming[J]. *IEEE Transactions on Aerospace and Electronic Systems*, 2001, 37(2): 508–523. doi: 10.1109/7.937465.

[9] DAI Huanyao, WANG Xuesong, LI Yongzhen, *et al*. Mainlobe jamming suppression method of using spatial polarization characteristics of antennas[J]. *IEEE Transactions on Aerospace and Electronic Systems*, 2012, 48(3): 2167–2179. doi: 10.1109/TAES.2012.6237586.

[10] 鲍秋香. 频率随机捷变雷达抗扫频干扰性能仿真[J]. 舰船电子对抗, 2021, 44(5): 78–81. doi: 10.16426/j.cnki.jcdzdk.2021.05.017.
BAO Qiuxiang. Simulation of anti-sweep jamming performance of frequency random agility radar[J]. *Shipboard Electronic Countermeasure*, 2021, 44(5): 78–81. doi: 10.16426/j.cnki.jcdzdk.2021.05.017.

[11] 全英汇, 方文, 沙明辉, 等. 频率捷变雷达波形对抗技术现状与展望[J]. 系统工程与电子技术, 2021, 43(11): 3126–3136. doi: 10.12305/j.issn.1001-506X.2021.11.11.
QUAN Yinghui, FANG Wen, SHA Minghui, *et al*. Present situation and prospects of frequency agility radar wave form countermeasures[J]. *Systems Engineering and Electronics*, 2021, 43(11): 3126–3136. doi: 10.12305/j.issn.1001-506X.2021.11.11.

[12] MINSKY M. Steps toward artificial intelligence[J]. *Proceedings of the IRE*, 1961, 49(1): 8–30. doi: 10.1109/JRPROC.1961.287775.

[13] 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. doi: 10.1109/MSP.2017.2743240.

[14] JIANG Wen, REN Yihui, and WANG Yanping. Improving anti-jamming decision-making strategies for cognitive radar via multi-agent deep reinforcement learning[J]. *Digital Signal Processing*, 2023, 135: 103952. doi: 10.1016/j.dsp.2023.103952.

[15] JIANG Wen, WANG Yanping, LI Yang, *et al*. An intelligent anti-jamming decision-making method based on deep reinforcement learning for cognitive radar[C]. 2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Rio de Janeiro, Brazil, 2023: 1662–1666. doi: 10.1109/CSCWD57460.2023.10152833.

[16] WEI Jingjing, WEI Yinsheng, YU Lei, *et al*. Radar anti-jamming decision-making method based on DDPG-MADDPG algorithm[J]. *Remote Sensing*, 2023, 15(16): 4046. doi: 10.3390/rs15164046.

[17] AZIZ M M, MAUD A, and HABIB A. Reinforcement learning based techniques for radar anti-jamming[C]. 2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST), Islamabad, Pakistan, 2021:

1021–1025. doi: 10.1109/IBCAST51254.2021.9393209.

[18] LI Kang, JIU Bo, LIU Hongwei, *et al*. Reinforcement learning based anti-jamming frequency hopping strategies design for cognitive radar[C]. 2018 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Qingdao, China, 2018: 1–5. doi: 10.1109/ICSPCC.2018.8567751.

[19] LI Kang, JIU Bo, and LIU Hongwei. Deep Q-network based anti-jamming strategy design for frequency agile radar[C]. 2019 International Radar Conference (RADAR), Toulon, France, 2019: 1–5. doi: 10.1109/RADAR41533.2019.171227.

[20] LI Kang, JIU Bo, WANG Penghui, *et al*. Radar active antagonism through deep reinforcement learning: A way to address the challenge of mainlobe jamming[J]. *Signal Processing*, 2021, 186: 108130. doi: 10.1016/j.sigpro.2021.108130.

[21] WU Qinhao, WANG Hongqiang, LI Xiang, *et al*. Reinforcement learning-based anti-jamming in networked UAV radar systems[J]. *Applied Sciences*, 2019, 9(23): 5173. doi: 10.3390/app9235173.

[22] AK S and BRÜGGENWIRTH S. Avoiding jammers: A reinforcement learning approach[C]. 2020 IEEE International Radar Conference (RADAR), Washington, USA, 2020: 321–326. doi: 10.1109/RADAR42522.2020.9114797.

[23] AILIYA, YI Wei, and YUAN Ye. Reinforcement learning-based joint adaptive frequency hopping and pulse-width allocation for radar anti-jamming[C]. 2020 IEEE Radar Conference (RadarConf20), Florence, Italy, 2020: 1–6. doi: 10.1109/RadarConf2043947.2020.9266402.

[24] ZHANG Jiaxiang and ZHOU Chao. Interrupted sampling repeater jamming suppression method based on hybrid modulated radar signal[C]. 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP), Chongqing, China, 2019: 1–4. doi: 10.1109/ICSIDP47821.2019.9173093.

作 者 简 介

张嘉翔，博士生，主要研究方向为智能干扰感知与抗干扰决策。

张凯翔，博士生，主要研究方向为分布式雷达和抗干扰。

梁振楠，博士，副研究员，硕士生导师，主要研究方向为数字阵列雷达系统和宽带雷达信号处理。

陈新亮，博士，讲师，硕士生导师，主要研究方向为目标检测跟踪和软件化雷达。

刘泉华，博士，教授，博士生导师，主要研究方向为高分辨雷达系统及信号处理。

(责任编辑：高山流水)