doi:10.11887/j.issn.1001-2486.23060016 http://journal.nudt.edu.cn

汤普森采样在线设计频率捷变雷达抗干扰策略

吴振华[1,2],钱 军[1],张 磊[3*],陈莉丽[4],朱 宁[2],杨利霞[1]

(1. 安徽大学 电子信息工程学院,安徽 合肥 230601;2. 电子信息系统复杂电磁环境效应国家重点实验室,河南 洛阳 471003;
3. 中山大学 电子与通信工程学院,广东 深圳 518107;4. 军事科学院 国防科技创新研究院,北京 100071)

摘 要:在有源干扰动态对抗背景下,基于在线学习理论中多臂赌博机模型,将雷达和干扰机工作频率作为对抗动作空间建模,通过对干扰环境状态不确定性进行多轮脉冲波形发射探索,搭建基于卷积神经网络的频率通道干扰识别器以得到频率通道干扰信念状态后验概率估计,利用汤普森采样求解算法高效求解多臂赌博机模型,实现探索与利用之间的平衡。仿真结果表明,相较于频率随机捷变及深度强化学习策略求解算法,该方法的对抗策略收敛性能更高,可适应动态快变干扰环境,充分发挥雷达波形发射主动方对抗优势。

关键词:雷达频率捷变;在线学习;多臂赌博机;汤普森采样

中图分类号:TN95 **文献标志码**:A **文章编号**:1001-2486(2025)05-206-10

论文拓展

Designing online anti-jamming strategy for agile frequency radar via Thompson sampling

WU Zhenhua[1,2], *QIAN Jun*[1], *ZHANG Lei*[3*], *CHEN Lili*[4], *ZHU Ning*[2], *YANG Lixia*[1]

(1. School of Electronic and Information Engineering, Anhui University, Hefei 230601, China;
2. State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, Luoyang 471003, China;
3. School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen 518107, China;
4. National Innovation Institute of Defense Technology, Academy of Military Sciences, Beijing 100071, China)

Abstract: In the context of dynamic countermeasures between radar and active jammer, the working frequency of radar and adversarial jammer were modeled as the combat action space based on the multi-arm bandit model in online learning theory. By exploring the uncertainty of the jamming environment state through multiple-round pulse transmission, a frequency channel jamming recognizer based on a convolutional neural network was constructed to obtain the posterior probability estimation of the belief state of each frequency channel. Then the Thompson sampling algorithm efficiently solved the built multi-arm bandit model, achieving a balance between exploration and exploitation. Simulation results show that compared with random frequency agility and deep reinforcement learning algorithms, the method had higher convergence performance and was more adaptable to dynamic fast-changing jamming environments, which can give full potential to the antagonism advantage of radar active waveform transmission.

Keywords: radar frequency agility; online learning; multi-armed bandits; Thompson sampling

现代战争形态正加速向信息化演变,电磁频谱空间成为战场信息获取、传输的重要载体和通道,频谱域作战先行已经成为现代军事活动的典型特征[1-2]。为应对复杂战场电磁干扰环境,注重自主交互式的电磁环境学习与动态智能化的对抗任务处理的认知电子战作战内涵近些年得到了极大的丰富。通过实时感知学习动态干扰对抗环境,辨识对手反馈状态,自主高效调整应对策略是认知电子战的核心属性。具体而言,在雷达与干扰认知作战过程中,包括认知发射、认知接收、智能处理以及智能调度等功能模块上的认知雷达具备闭环学习自进化能力,可以更好地对动态变化

收稿日期:2023-06-16

基金项目:国家自然科学基金资助项目(62201007);中国博士后科学基金资助项目(2020M681992);电子信息系统复杂电磁环境效应国家重点实验室开放课题资助项目(CEMEE2022Z0302B)

第一作者:吴振华(1993—),男,安徽淮北人,副教授,博士,博士生导师,E-mail:zhwu@ahu.edu.cn

***通信作者**:张磊(1984—),男,浙江金华人,教授,博士,博士生导师,E-mail:zhanglei57@mail.sysu.edu.cn

引用格式:吴振华, 钱军, 张磊, 等. 汤普森采样在线设计频率捷变雷达抗干扰策略[J]. 国防科技大学学报, 2025, 47(5): 206-215.

Citation: WU Z H, QIAN J, ZHANG L, et al. Designing online anti-jamming strategy for agile frequency radar via Thompson sampling[J]. Journal of National University of Defense Technology, 2025, 47(5): 206-215.

对抗环境进行感知、针对性进行发射和接收设计且具备智能处理和资源调度等能力[3]。

雷达频率捷变通过在工作频率捷变带宽中的多个频率点随机跳变,具有一定的波形发射主动对抗优势,可实现对压制式和转发式干扰的对抗[4]。具体来说,手动变频、去相关变频、伪随机变频、自适应捷变频和脉组变频等几种工作方式需要在技战术层面由指挥员依据干扰类型识别及战场态势结果进行人工操作[5],同时需要在反侦察与抗干扰技战术层面进行平衡,现阶段作业流式变频策略缺乏对干扰环境信息进行实时闭环利用,被动、应答式的频域资源调度较为难以应对复杂多变的干扰环境。考虑到雷达波形发射的敏捷性及频域资源的重要性,为了实现捷变雷达在应对不同干扰环境下对发射波形参数及频谱资源智能调度策略进行优选,加快“感知 - 适应 - 决策 - 行动(observe-orient-decide-act, OODA)”环路速度是本文主要研究问题,有较为重要的理论和军事意义。

针对雷达抗干扰发射策略优化问题,研究人员引入博弈论和强化学习(reinforcement learning, RL)进行最优策略求解。文献[6]利用强化学习中 Q 学习以及深度 Q 网络(deep Q-network, DQN)算法在不同类型有源主瓣干扰环境下,经过多轮与干扰环境交互,学习并躲避有源干扰策略。考虑到干扰的非稳态特性,文献[7]将监督学习与强化学习相结合,对干扰机的三种干扰策略进行对抗建模,得到了干扰时变策略下的最优抗干扰策略。考虑到在每个交互回合中,真实有源干扰状态不确定且不可直接获取,文献[8]将抗干扰波形发射频率策略问题建模成不依赖环境模型知识的部分可观测马尔可夫决策过程(partially observable Markov decision process, POMDP)并且利用双 DQN(double DQN, DDQN)算法进行发射频率策略求解。文献[9 - 10]将雷达与干扰对抗建模成马尔可夫决策过程(Markov decision process, MDP)过程,将积累效率和截获概率引入奖赏函数的设计中,并通过 Q 学习算法进行最优跳频间隔和脉冲宽度策略学习,在有源扫频压制干扰策略环境下验证了所提方法的有效性。在基于博弈论框架对雷达和干扰机建模策略方面,文献[11]将雷达与干扰机以频率为博弈要素进行非完美信息下基于博弈树的扩展式博弈建模,并利用虚拟遗憾值(counterfactual regret)算法求解纳什均衡策略。更进一步,文献[12]提出一种不需要博弈树详细结构以及干扰机对手的先验信息下使用神经虚拟自我对局(neural fictitious self play, NFSP)的算法,得到不完美信息扩展式近似纳什均衡解。

总的来说,强化学习强调智能体在环境中通过多轮交互试错的方式学习动作,其潜在应用前提是智能体状态转移具有马尔可夫性[13],一般假设可以直接获取干扰环境状态,对于时变随机的非合作有源干扰环境,通常难以保证预先针对之前干扰环境训练好的强化学习网络泛化能力。除此之外,强化学习技术在对环境进行探索过程中,通常需要多轮交互进行样本采样,算法收敛速率低问题较为凸显,考虑到战场干扰对抗实时性高、环境变化迅速以及敌对方干扰机随时可能切换干扰策略的特点,学习能力差的智能体将难以适应瞬息万变的战场环境[14 - 15]。基于博弈论的抗干扰策略学习方法的目标是达到纳什均衡,其要求博弈玩家双方较为理智,通常在雷达与干扰激烈对抗中很难满足条件。除此之外,达到纳什均衡点追求的是损失最小,因此利用博弈论得到的抗干扰策略不一定是最优解,且需要提及的是,博弈模型的核心为构建盈利矩阵,日益复杂战场电磁环境下,构建盈利矩阵更为复杂,难以准确表征玩家双方收益。

考虑到干扰目标状态识别通常面临“小样本空间”“增量式信号数据流”以及充满“不确定性”难题,传统批量数据处理离线学习算法无法适应流式增量数据特性。在线学习技术[16],又称“增量学习”或“适应性学习”,包含一系列流式持续接收数据,动态实时地训练样本并更新模型来处理一些预测任务的机器学习技术,其目的是基于对前帧预测/学习任务的回报值以及潜在其他动作,最大限度提高决策者的预测或者动作序列的准确性/正确性。在雷达智能干扰对抗研究背景下,相比于离线强化学习对抗技术,在线学习技术可快速对感知的有源干扰信号类型进行响应,通过分析干扰样式参量并预测干扰序列工作模式,匹配生成最优的干扰对抗策略,更为适应高动态、复合干扰组态随机、干扰样式和工作参数未知的强干扰对抗作战环境,同时其具备采样探索高效、泛化能力强的特点。在雷达频率捷变与有源干扰动态对抗背景下,为了适应动态不确定干扰环境,本文基于在线学习理论中多臂赌博机(multi-armed bandits, MAB)模型,将雷达和干扰机工作频率作为动作空间建模,并将频率通道选择建模成最优臂(bandit)选取,在多轮雷达脉冲波形发射中,对不同频率通道回波信号参数进行识别,并

针对性地设计奖赏函数，利用汤普森采样（Thompson sampling，TS）算法求解最优频率发射策略以实现在探索和利用之间的平衡。

1 捷变频有源干扰对抗信号模型

1.1 捷变频雷达模型

对于脉间伪随机跳变全相参体制捷变频雷达，一个相干处理时间周期（coherent processing interval，CPI）内，捷变频信号模型可表示为：

$$s(t)=u(t)\cdot rect\left(\frac{t-nT_{\mathrm{r}}}{T_{\mathrm{p}}}\right)\mathrm{e}^{\mathrm{j}2\pi f_n(t-nT_{\mathrm{r}})} \tag{1}$$

式中：$u(t)$ 表示发射信号复包络；$rect(\cdot)$ 表示矩形窗函数，$rect(t)=\begin{cases}1 & 0\leqslant t\leqslant 1\\ 0 & \text{其他}\end{cases}$；$T_{\mathrm{p}}$ 为脉冲宽度；T_{r} 为脉冲重复间隔（pulse repetition interval，PRI）；f_n 表示第 n 个发射脉冲载频，$f_n=f_{\mathrm{c}}+d_n\Delta f$（$n=0,1,\cdots,N-1,N$）为一个 CPI 内发射的脉冲总数，$d_n$ 表示发射第 n 个脉冲的随机跳变项，且有 $d_n\in\{0,1,\cdots,M-1\}$，M 为雷达可发射频带总数，且满足 $N>M$，f_{c} 为雷达载频，$\Delta f=B/M$ 是脉冲的跳频间隔。

考虑脉冲内采用线性调频信号形式发射，复包络形式为：

$$u(t)=\mathrm{e}^{\mathrm{j}\pi K_{\mathrm{r}}(t-nT_{\mathrm{r}})^2} \tag{2}$$

式中，$K_{\mathrm{r}}=\Delta f/T_{\mathrm{p}}$ 为调斜频率，对观测场景 K 个运动目标，回波信号线性相位项可以表示为：

$$s(n)=\sum_{k=1}^{K}A_k\mathrm{e}^{-\mathrm{j}2\pi f_{\mathrm{m}}\frac{2v_kt_{\mathrm{m}}}{c}}\mathrm{e}^{-\mathrm{j}2\pi f_{\mathrm{m}}\frac{R_k}{c}} \tag{3}$$

式中，A_k、R_k、v_k 分别为第 k 个目标的散射系数、距离和速度，$\mathrm{e}^{-\mathrm{j}2\pi f_{\mathrm{m}}\frac{2v_kt_{\mathrm{m}}}{c}}$ 为目标速度、慢时间与随机载频的耦合相位，$\mathrm{e}^{-\mathrm{j}2\pi f_{\mathrm{m}}\frac{R_k}{c}}$ 为目标距离与随机载频的耦合相位。在得到目标回波信号相位项后，通过构建距离－多普勒测量字典，可恢复观测场景内目标距离和速度。

考虑到接收回波中固定存在噪声及有无有源干扰的存在，雷达接收回波信号 $e(t)$ 的广义形式为：

$$\begin{cases}\mathcal{H}_0:\ r(t)=s(t)+n(t)\\ \mathcal{H}_1:\ r(t)=s(t)+j(t)+n(t)\end{cases} \tag{4}$$

其中，$s(t)$、$j(t)$、$n(t)$ 分别表示目标回波信号、有源干扰信号及噪声项。$\mathcal{H}_0$ 及 $\mathcal{H}_1$ 分别表示有无干扰信号的假设，若 $\mathcal{H}_1$ 假设成立，则由式（3）做距离速度重建得到的目标参数准确性极大降低。鉴于此，雷达需在多轮脉冲频率捷变波形发射下，通过对干扰环境的感知和学习，对干扰信号所在频率通道进行自主学习来避让发射。

1.2 MAB 模型

多摇臂赌博机模型是在线学习理论中较为经典的一类模型，它最初起源于如何在多个赌博机中连续摇动 N 次，赢得尽可能多的钱。由于不同赌博机可能有不同的赔率，因此观测到的经验赔率对于之后的选择将有所帮助。如何利用实时信息尽快学得哪个赌博机赔率最高，以获得更高的收益，就是多摇臂赌博机问题的原型。赌博机共有 M 个臂，M 个臂的集合即玩家的动作集合 A，记为 $A=\{1,2,\cdots,M\}$，每个臂 $m\in A$ 都有各自未知的奖励分布，该分布的期望记作 μ_m。玩家在每一轮游戏 $n\in\{1,2,\cdots,N\}$ 拉动其中一个臂 $a_n\in A$，环境将从该臂的奖励分布中采样一个随机变量 $X_{A,n}$ 作为玩家拉动该臂的奖励值。该奖励值将帮助玩家更新对臂的奖励分布的了解，进而更新其后续选择臂的策略。玩家的目标是最大化 N 轮的累积期望收益，即最小化与最优臂之间的累积期望收益的差距。令 $\mu_n^*=\max_{m\in A}\mu_m$，表示 n 时刻最高期望收益，则玩家的目标为最小化累积期望悔恨（regret）值：

$$Reg(N)=E\left[\sum_{n=1}^{N}(\mu_n^*-X_{A,n})\right] \tag{5}$$

式中，$Reg(N)$ 表示与最优臂收益之间的平均累计收益损失之差，E 表示期望。通过与环境的多轮学习，玩家将收集到对各个臂的奖励分布的观察。玩家一方面希望选取历史观察中表现最好的臂，以获取相对较高的收益（利用）；另一方面希望尝试一些尚未受到足够观察的臂，以获取潜在较高的收益（探索）。过度利用可能导致玩家错过最优臂，过度探索则将导致玩家付出过多的学习代价，如何平衡利用与探索是多臂赌博机模型需要考虑的核心问题。考虑到有源干扰环境下难以获取完整状态信息，需在不完全干扰环境信息下利用有限次脉冲发射生成最优对抗策略，本文将捷变频雷达发射频率建模成赌博机选择臂，己方雷达基础波形发射动作集为 $a_{\mathrm{R}}(n)\in A_{\mathrm{R}}$，每个脉冲发射时刻敌方干扰机攻击的子频率通道表示为 $a_{\mathrm{J}}(n)\in A_{\mathrm{J}}$，不同回波假设模型下，$\mathcal{H}_0$、$\mathcal{H}_1$ 对应不同的奖赏，通过多轮的干扰环境主动学习，求解生成最优频率发射策略。

对不同臂的选取，奖赏函数设计为：

$$R(n)=\begin{cases}0 & a_{\mathrm{R}}(n)=a_{\mathrm{J}}(n)\\ 1 & |a_{\mathrm{R}}(n)-a_{\mathrm{J}}(n)|=1\\ 2 & |a_{\mathrm{R}}(n)-a_{\mathrm{J}}(n)|\geqslant 2\end{cases} \tag{6}$$

式(6)设计的奖赏函数与己方雷达和敌方干扰机发射脉冲频率通道间隔距离直接相关,充分考虑后续频率捷变雷达进行距离、速度测量信号处理需求,$a_R(n)\in A_R$ 和 $a_J(n)\in A_J$ 分别表示 n 脉冲时刻己方雷达选择频率通道和敌方干扰机攻击的频率通道索引,$|a_R(n)-a_J(n)|$为己方雷达与敌方干扰机攻击频率通道之间的绝对间隔。$R(n)=2(|a_R(n)-a_J(n)|\geqslant 2)$为给予和敌方干扰机频率通道间隔不小于 2 的己方雷达发射频率通道同等大小的奖赏 2,表示同等看待所有远离敌方干扰机频率通道的选取动作,以保持发射频率的随机性,$R(n)=1(|a_R(n)-a_J(n)|=1)$表示给予和敌方干扰机频率通道间隔为 1 的己方雷达发射频率通道更低的奖赏 1。$R(n)=0(a_R(n)=a_J(n))$为若己方雷达发射脉冲与敌方干扰机脉冲占据相同的频率通道,此时奖赏为 0。

1.3　基于卷积神经网络有源干扰鉴别器

对雷达而言,脉冲波形发射面临的动态快变干扰状态未知,直接利用第 n 个脉冲之前的全历史信息 $h_n=[(s_1,R(1)),(s_2,R(2)),\cdots,(s_{n-1},R(n-1))]s_1=\{a_R(1),a_J(1)\}$为己方雷达动作与干扰攻击动作集,估计当前发射脉冲下干扰环境状态信息较为困难。因此引入信念状态(belief state)[17]对所在频率通道干扰状态 $b(s)$进行概率表示。为了获得不同频率通道下对干扰环境的观察结果来形成对环境状态的信念,进而计算干扰信号所在每个频道的后验概率,搭建基于卷积神经网络(convolution neural network,CNN)的有源干扰识别器,如图 1 所示,输出干扰环境状态的后验概率分布。设计的第一层为网络的输入层,将接收回波信号每一点的实部和虚部分开,分开的实部依次连接、分开的虚部依次连接,再将连接后的实部和虚部拼接成新的一维数据;接下来是三层卷积层,卷积层输入分别为 32,64,128,翻倍递增,每一卷积层的卷积核大小均为 3,卷积步长为 2;数据经过每一卷积层处理后,都会经过激活函数层(ReLU)、最大池化层(softmax)和 Dropout 层处理;经过三层卷积层处理后数据输入全局平均池化层(global average pooling, GAP);最后,全连接层将数据分类,通过 softmax 层进行归一化。在对网络进行训练后,将不同频率通道回波数据进行干扰信号识别,得到第 n 时刻观测数据中任意 M 个频率通道存在有源干扰的概率 $\boldsymbol{p}_n=[p_n^1,p_n^2,\cdots,p_n^M]$。

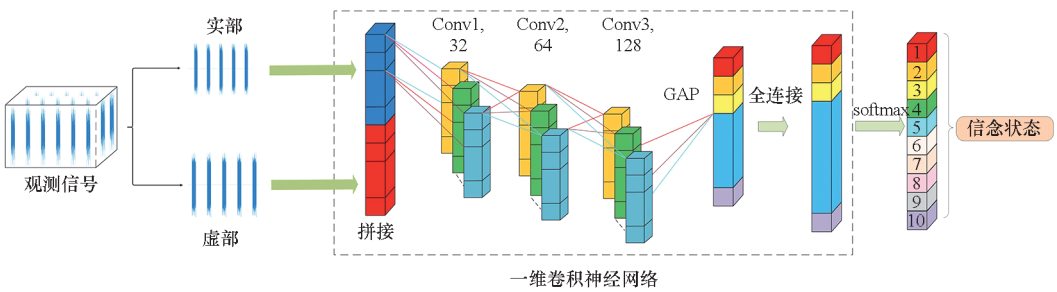


图 1　有源干扰识别器网络结构

Fig. 1　Network structure of active jamming recognizer

2　基于汤普森采样算法的有源干扰对抗策略

对于多臂赌博机经典的开发和利用最优策略求解问题,TS 算法根据每个臂的概率分布进行采样,然后选择具有最大样本平均值的臂进行开发。假设每个臂 m 具有先验分布 μ_m,在每一轮决策开始,算法从每个臂 m 的先验分布 μ_m 中执行动作,并且得到对应不同的奖赏,随着探索过程的不断继续,将动作 m 的历史数据记录为 L_m,同时记录每一选择动作 m 的时间节点和返回的奖励。基于贝叶斯估计定理,可以得到 μ_m 的后验分布计算公式为:

$$p(\mu_m|D_m)=\frac{p(\mu_m)p(L_m|\mu_m)}{p(L_m)}\tag{7}$$

其中,

$$p(L_m|\mu_m)=\mu_m^{S_m}(1-\mu_m)^{F_m}\tag{8}$$

为 L_m 的似然函数,S_m、F_m 分别是使用动作 m 成功和失败的次数,将 Beta 分布(beta distribution)函数作为先验分布 μ_m,且有

$$p_m(\mu_m)=\frac{\Gamma(\alpha_m+\beta_m)}{\Gamma(\alpha_m)\Gamma(\beta_m)}\mu_m^{\alpha_m-1}(1-\mu_m)^{\beta_m-1}\tag{9}$$

参数 α_m 和 β_m 决定了 Beta 分布的均值和方差,与第 m 个臂相关联,$\Gamma(\cdot)$为 Gamma 函数。

将式(7)与式(8)代入式(9),可以得到后验分布的化简形式为:

$$p(\theta_m|L_m)=\frac{\mu_m^{\alpha_m+S_m-1}(1-\mu_m)^{\beta_m+F_m-1}}{\int_0^1\mu_m^{\alpha_m+S_m-1}(1-\mu_m)^{\beta_m+F_m-1}\mathrm{d}\mu_m}$$

$$=\frac{\Gamma(\alpha_m+\beta_m+S_m+F_m)}{\Gamma(\alpha_m+S_m)\Gamma(\beta_m+F_m)}\mu_m^{\alpha_m+S_m-1}(1-\mu_m)^{\beta_m+F_m-1} \tag{10}$$

由式(10)可以看出,μ_m 的后验分布依然为 Beta 分布,其参数分别为 α_m+S_m 及 β_m+F_m。因此对于任意时刻 n 的雷达脉冲发射频率动作 $a_{\mathrm{R}}^n\in A=\{1,2,\cdots,M\}$ 及动作选取奖励 $R(n)$,Beta 分布的参数更新规则为:

$$(\alpha_{a_{\mathrm{R}}(n)},\beta_{a_{\mathrm{R}}(n)})\leftarrow(\alpha_{a_{\mathrm{R}}(n)}+R(n),\beta_{a_{\mathrm{R}}(n)}+2-R(n)) \tag{11}$$

式中,$\alpha_{a_{\mathrm{R}}(n)}$ 和 $\beta_{a_{\mathrm{R}}(n)}$ 是 MAB 模型的重要参数,分别表示选择频带 $a_{\mathrm{R}}(n)$ 获得的累积奖励和累积悔恨值。总的来说,汤普森采样算法利用对每个臂后验分布的高效采样可以较好地平衡开发与利用矛盾难题。

图 2 是雷达频率捷变在线学习抗干扰策略的 OODA 闭合环路框图,图中 MAB 模型和 TS 算法以伪代码形式详细表示,见算法 1。在第 4 ~ 6 行,该算法从后验数据中对估计值进行采样,然后通过将概率分布参数稍微“贴现未来”“忘记过去”,然后根据提升系数 γ 更新 α_m 和 β_m;在第 8 行,算法取抽样中最大值决定在这一轮中选择哪个超级臂,即发射哪个雷达动作 $a_{\mathrm{R}}(n)$。

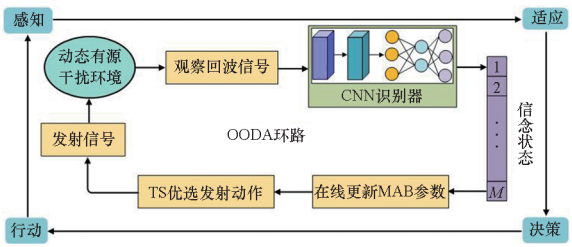


图 2 雷达频率捷变在线学习抗干扰策略框图

Fig. 2 Diagram of online learning anti-jamming strategy for frequency agile radar

算法 1 基于汤普森采样的捷变频雷达在线干扰对抗策略

Alg. 1 Research on online anti-jamming strategy for radar frequency agility based on Thompson sampling

输入:己方雷达发射频率动作集 A_{R},敌方干扰机攻击动作集 A_{J},脉冲频带宽度 B/M,提升系数 γ

输出:参数 α_{a_n} 和 β_{a_n}

1. 初始化每个基础臂参数 $\alpha_{f_m}=\beta_{f_m}=1$,其中 $m\in\{1,2,\cdots,M\}$
2. **for** $n=1,2,\cdots,N$ **do**
3. **for** $m=1,2,\cdots,M$ **do**
4. 采样 $\theta_m\sim Beta(\alpha_m,\beta_m)$
5. 更新 $\alpha_m=\gamma\cdot\alpha_m+(1-\gamma)$
6. 更新 $\beta_m=\gamma\cdot\beta_m+(1-\gamma)$
7. **end**
8. 计算 $a_{\mathrm{R}}(n)=\arg\max_{m\in A_{\mathrm{R}}}\theta_m$,选择频率通道 $f_{a_{\mathrm{R}}(n)}$ 发射信号
9. 观察奖励值 $R(n)$,计算单步悔恨值 $2-R(n)$
10. 更新 MAB 选择臂 $a_{\mathrm{R}}(n)$ 对应的参数 $\alpha_{a_{\mathrm{R}}(n)}$ 和 $\beta_{a_{\mathrm{R}}(n)}$
11. $\alpha_{a_{\mathrm{R}}(n)}=\alpha_{a_{\mathrm{R}}(n)}+R(n)$
12. $\beta_{a_{\mathrm{R}}(n)}=\beta_{a_{\mathrm{R}}(n)}+2-R(n)$
13. **end**

3 实验结果及分析

3.1 雷达有源干扰对抗场景系统参数设置

在仿真对抗场景中,重点考虑地基雷达在复杂区域防空作战任务下,对敌方机载自卫式主瓣有源压制干扰进行自主学习策略对抗。在频率发射动作建模中,考虑脉冲级频率捷变对抗敌方干扰机攻击,对敌方干扰机动态环境建模,假定敌方干扰机具备己方雷达脉冲截获分析辨识能力,可通过电子情报系统获取己方雷达可用工作带宽、载波频率、脉冲带宽、PRI 系统参数,即己方雷达面临单方面作战系统参数严峻暴露问题。同时对于敌方干扰机,在截获己方雷达脉冲后,发射攻击脉冲时宽和频宽均与所截获的雷达脉冲保持一致,所选动作空间与己方雷达动作空间保持一致,即有 $A_{\mathrm{R}}=A_{\mathrm{J}}=\{1,2,\cdots,M\}$,且采取典型的的上/下、锯齿、瞄准频率在己方雷达工作带宽内进行循环压制频率攻击。为了定量对比分析在线学习算法的雷达频率捷变有源压制干扰对抗能力,仿真实验设置参数如表 1 所示。

在强对抗作战环境下,己方雷达需以最少的探索代价即发射脉冲数进行有源压制干扰环境动态感知学习,针对性地发射未被敌方攻击的频率通道,进行稳健目标距离、速度参数估计。每个臂在未知干扰环境的奖赏函数需结合有源干扰识别器进行信念状态后验概率估计。为对比策略求解性能,使用随机(Random)策略、贪婪(ε-greedy)策略、最大置信度上界(upper confidence bound, UCB)算法和 DQN 算法进行对比。

表 1 为己方雷达及敌方干扰机系统参数,表 2为敌方干扰机采用的有源压制干扰频率扫描策略,表 3 为有源干扰识别器卷积神经网络具体参数。

表 1　己方雷达/敌方干扰机系统及 MAB 参数

Tab. 1　Parameters of own radar/adverse jammer system and MAB

参数	数值
己方雷达/敌方干扰机载频/GHz	10
己方雷达/敌方干扰机带宽/MHz	500
载频跳变带宽/MHz	50
捷变通道(选择臂)总数	10
脉冲重复周期/μs	500
发射脉冲时宽/μs	50
提升系数	1.05
发射脉冲总数(总轮数)	5 000
一个 CPI 包含脉冲数	32

表 2　有源压制干扰频率扫描策略

Tab. 2　Frequency scanning strategy for active suppression jamming

干扰	切换策略
上三角扫频	循环递增方式
下三角扫频	循环递减方式
锯齿式	循环递增递减方式
瞄准式	干扰机与上一时刻所截获的雷达脉冲频率通道保持一致

表 3　识别器卷积神经网络参数

Tab. 3　CNN parameters for the recognizer

网络结构	Conv1	Conv2	Conv3	Output
卷积核尺寸	3×1	3×1	3×1	
激活函数	ReLU	ReLU	ReLU	Softmax
通道数	32	64	128	1
训练方法	Adam(自适应运动估计)			

3.2　基于卷积神经网络的有源干扰识别结果

有源干扰识别器数据集构建具体过程为:根据式(4)生成回波数据集,雷达发射信号和环境交互所得回波信号 $e(t)$ 的信噪比、干噪比和时延分别在 20 ~ 30 dB、5 ~ 10 dB 和 20 ~ 30 μs(对应 3 000 ~ 4 500 m 雷达观测场景范围)中随机选取,回波信号的标签为干扰信号所在频道。其中,在雷达同一发射动作下,随机选取干扰机不同攻击动作下的回波信号和无攻击动作的回波信号共 550 个,回波信号数据集共包含 5 500 个回波信号。将回波信号数据按照 7∶2∶1 分为训练集、验证集、测试集三部分。依据表 3 设置有源干扰识别器网络的训练参数,最小学习率 $\eta_{\min}=0.000\ 8$ 和最大学习率 $\eta_{\max}=0.01$,学习率根据式(12)不断衰减得到。

$$\eta=\eta_{\min}+0.5\times(\eta_{\max}-\eta_{\min})\times\left[1+\cos\left(\frac{e}{T}\pi\right)\right] \tag{12}$$

其中,e 是当前训练次数,T 是训练总次数。上述有源干扰识别器网络的训练参数的设置使得有源干扰识别器网络前期学习率大、收敛快、后期学习率小,避免在极值附近振荡,使用交叉熵损失函数(cross entropy loss)作为网络的损失函数。

将接收脉冲信号输入有源干扰识别器,识别器输出的信念状态 p_n 中最大后验概率的子频带为干扰机的干扰频带,己方雷达在脉冲波形发射时可选择 10 个子频率通道,识别器在任意时刻接收回波脉冲进行被攻击的子频率通道鉴别,图 3 为 $M=10$ 时,对任意己方雷达脉冲发射频率通道进行敌方干扰机攻击子频率通道鉴别结果,颜色越深表示输出概率越大,该识别器的识别准确率在 98% 以上,依据频率通道识别器,己方雷达对任意时刻发射脉冲回波进行信号分析,可对敌方干扰机所发射的攻击脉冲的子频率通道进行准确估计。

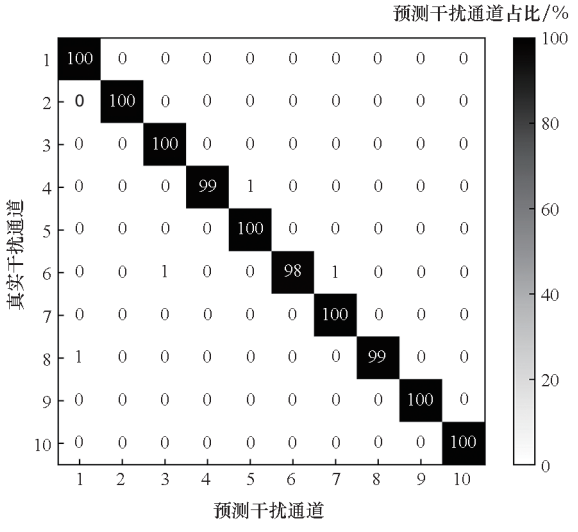


图 3　子频率通道干扰识别结果

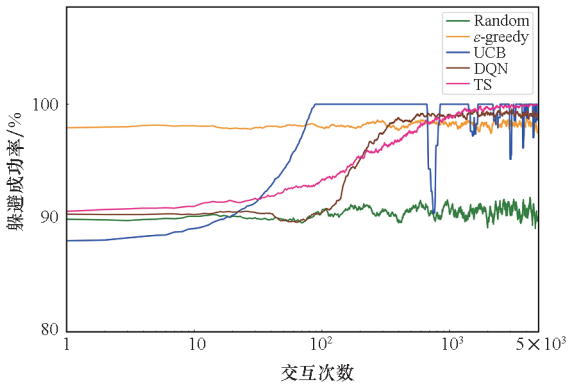
Fig. 3　Results of jamming recognition of each frequency channel

3.3　基于 TS 算法的有源压制干扰主动对抗策略生成结果

经过一段时间的捷变波形发射脉冲探索,对 1 个 CPI 内的 32 个脉冲进行分析,分别从抗干扰策略的躲避干扰成功率、在线选择悔恨值、动作选择 3 个指标定量分析不同策略的自主学习对抗

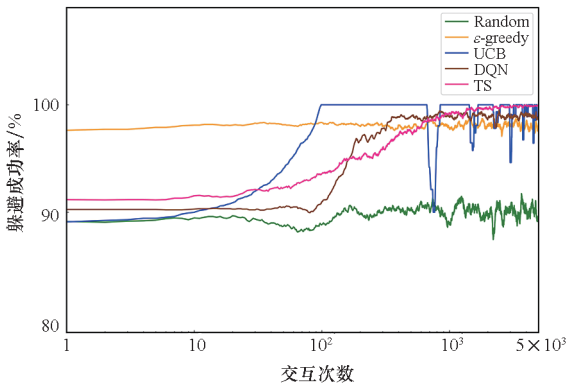
性能。

图 4(a) ~ (d)为不同抗干扰策略分别在下三角扫频干扰、上三角扫频干扰、锯齿式扫频干扰和瞄准式干扰场景下的躲避干扰成功率曲线图。Random 策略由于缺乏对环境知识的利用，在全部有源压制干扰场景中表现较为稳定。而采用对干扰环境进行学习得到的抗干扰策略的躲避干扰成功率均明显高于随机抗干扰策略。ε-greedy 策略利用环境知识进行学习，躲避成功率最终趋近于 100%，但是该算法对探索因子 ε 最优参数设置要求较高，不合适的 ε 容易使策略陷入局部最优。UCB 算法在计算出动作的置信度后采取贪婪策略，提高躲避干扰成功率，收敛速度快于 ε-greedy 策略，但是探索后期仍兼顾选择次数少的动作，导致振荡剧烈，最终收敛困难。DQN 算法利用深度网络进行训练，在训练过程中把 DQN 模型部署在 3070Ti 显卡上运行，在交互 5 000 次的条件下共用时 180 s。随机 ε-greedy、UCB、TS 算法均没有使用显卡，在 Intel i7 - 11700、32 GB 运存下的运行时间为 90 ~ 100 s。在算法收敛时间方面，DQN 算法收敛速度显著慢于 TS 方法，且较为依赖数据量及数据多样性和训练时间，DQN 算法对于样本的使用更为低效，训练深度神经网络时，可能会出现发散或振荡不稳定的情况，通常需要较为精细的超参数调整来解决，这点在图 4(d)中敌方干扰机使用更为灵活的瞄准式策略进行攻击时更为明显，在瞄准式干扰场景中，干扰机发射的频率始终跟随雷达，导致 DQN 收集到的数据高度重合，出现网络过拟合，初始性能反而劣于随机策略，但因其具有智能学习能力，躲避成功率最终提升到 99% 左右。总的来说，与其他抗干扰策略相比，TS 策略前期积极探索全部通道，后期保持稳定在最优动作上，在 4 种干扰策略下收敛速度较为稳定且收敛结果准确率保持最高且波动最小。



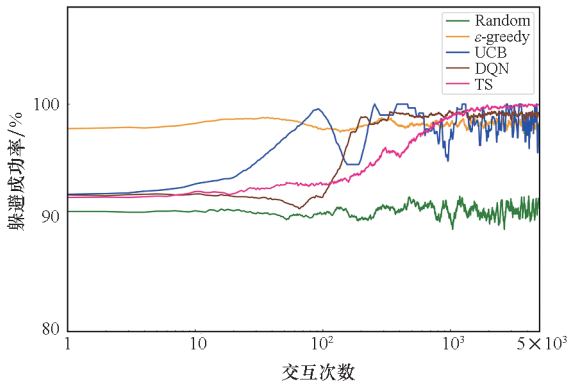
(a) 下三角扫频干扰

(a) Lower triangle sweep jamming



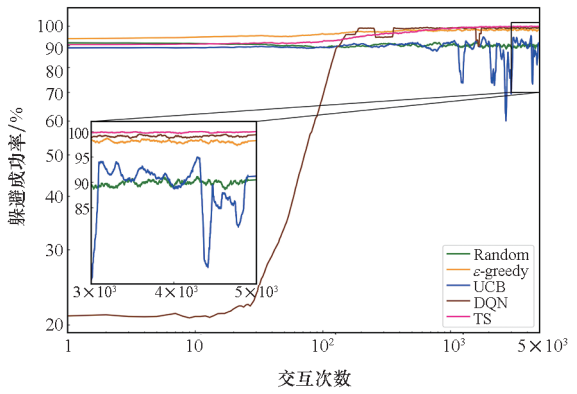
(b) 上三角扫频干扰

(b) Upper triangle sweep jamming



(c) 锯齿式扫频干扰

(c) Triangular sweep jamming



(d) 瞄准式干扰

(d) Spot jamming

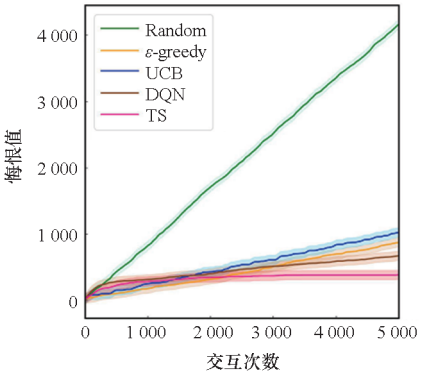
图 4　不同有源压制干扰策略下雷达躲避干扰成功率

Fig. 4　Radar jamming avoidance success rate under diverse active suppression jamming strategies

图 5(a) ~ (d)是不同算法在下三角扫频干扰、上三角扫频干扰、锯齿式扫频干扰和瞄准式干扰场景下的悔恨值曲线图。DQN 算法较为强大的深度网络学习能力使其在灵活切换的瞄准式干扰策略下，悔恨值低于 ε-greedy 算法和 UCB 算

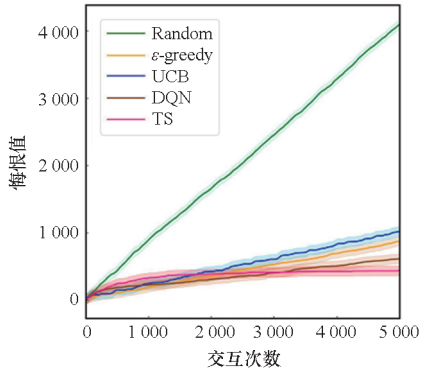
法，但是需要近千次的训练，才能达到最终收敛的状态，而 TS 算法不需要经过深度网络提前训练模型，悔恨值曲线始终低于其他算法。从悔恨值曲线分析，TS 算法在前期平衡了“利用”和“探索”，没有惰于过去的正确选择，在学习回合数足够之后，其悔恨值增长变缓，最终达到全局最优的稳定状态。

图 6(a) ~ (d)是 TS 算法有源压制动态干扰环境下的 1 个 CPI 内的动作选择情况。TS 算法通过学习敌方干扰机的攻击切换策略，不断调整自身发射频率通道，远离所预测的敌方干扰机占据的频率通道，同时同等看待其他可选频率通道，在避免被干扰的前提下最大限度地保持了宽带频率发射的完整性和随机性，保持观测场景目标距离和速度测量能力。



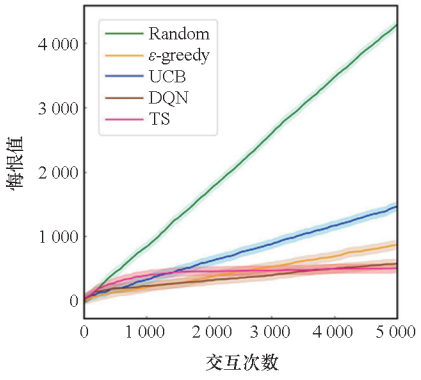
(a) 下三角扫频干扰

(a) Lower triangular sweep jamming



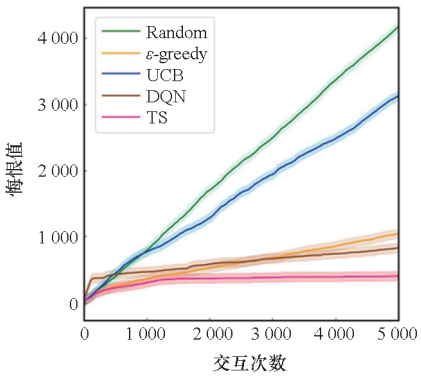
(b) 上三角扫频干扰

(b) Upper triangular sweep jamming



(c) 锯齿式扫频干扰

(c) Triangular sweep jamming

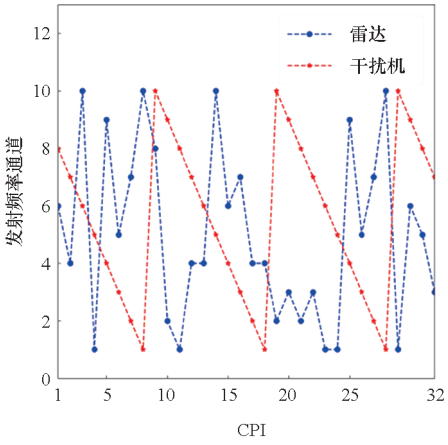


(d) 瞄准式干扰

(d) Spot jamming

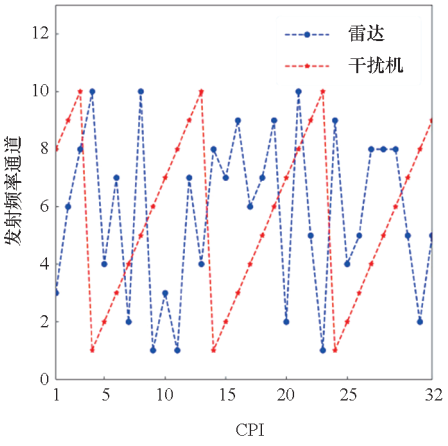
图 5　不同有源压制干扰策略抗干扰算法的悔恨值曲线图

Fig. 5　Regret value curves of anti-jamming algorithms under diverse active suppression jamming strategies



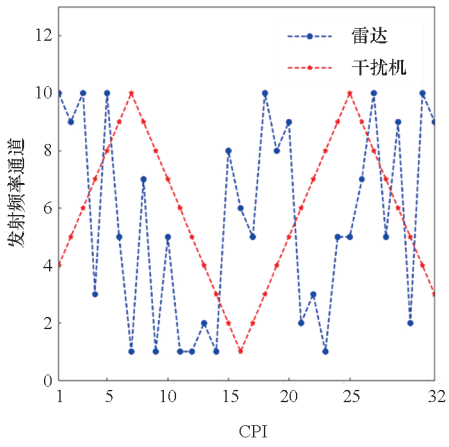
(a) 下三角扫频干扰

(a) Lower triangular sweep jamming



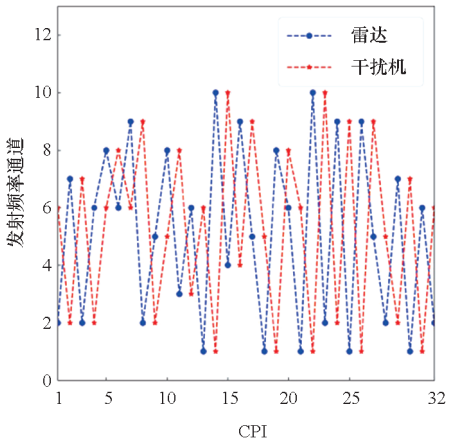
(b) 上三角扫频干扰

(b) Upper triangular sweep jamming



（c）锯齿式扫频干扰

（c）Triangular sweep jamming



（d）瞄准式干扰

（d）Spot jamming

图 6 TS 算法在不同有源压制干扰下的动作选择序列

Fig. 6 Action selection sequence of TS algorithm under different active suppression jamming scenarios

4 结论

本文设计己方雷达与敌方干扰机有源压制动态干扰下的发射频率 OODA 自主学习对抗框架，并提出 TS 算法求解在线干扰对抗策略，比较躲避干扰成功率和累积悔恨值，所学习的发射策略在躲避预测敌方干扰频率通道的同时兼顾己方发射频带的完整性和子频带的随机性。

1）考虑到面对未知动态快变环境状态难以利用全历史信息，引入信念状态对回波信号中的频率通道进行概率表示，利用搭建的 CNN 对通道干扰信念状态进行估计，正确识别通道存在干扰概率达到 98% 以上，为奖赏函数及策略优化算法提供基础。

2）利用 MAB 模型对捷变频雷达在线对抗干扰进行建模，应用 TS 算法高效求解探索和利用难题。该算法前期探索优先，后期稳定于最优动作选择，累积悔恨值保持最低，最终达到全局最优的状态。

3）TS 算法采样高效，同时泛化能力较强，不需要多次的离线预训练和高要求的算力资源，可更好地适应动态快变的有源压制干扰场景。

参考文献（References）

[1] 李凯，朱璇，张宝良，等. 联合电磁频谱作战的发展特点与技术分析[J]. 战术导弹技术，2022(6)：138－144.
LI K，ZHU X，ZHANG B L，et al. Analysis on development characteristics and technology of joint electromagnetic spectrum operations [J]. Tactical Missile Technology，2022(6)：138－144.（in Chinese）

[2] 王沙飞，鲍雁飞，李岩. 认知电子战体系结构与技术[J]. 中国科学：信息科学，2018，48(12)：1603－1613.
WANG S F，BAO Y F，LI Y. The architecture and technology of cognitive electronic warfare [J]. Science in China（Informationis Sciences），2018，48(12)：1603－1613.（in Chinese）

[3] 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.
CUI G L，YU X X，WEI W Q，et al. An overview of antijamming methods and future works on cognitive intelligent radar[J]. Journal of Radars，2022，11(6)：974－1002.（in Chinese）

[4] 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.
QUAN Y H，FANG W，SHA M H，et al. Present situation and prospects of frequency agility radar waveform countermeasures[J]. Systems Engineering and Electronics，2021，43(11)：3126－3136.（in Chinese）

[5] 刘冬利，兰慧，侯建强. 舰载雷达变频方式及变频使用策略研究[J]. 兵器装备工程学报，2021，42(4)：123－127.
LIU D L，LAN H，HOU J Q. Research on frequency conversion strategy and mode of shipborne radar[J]. Journal of Ordnance Equipment Engineering，2021，42(4)：123－127.（in Chinese）

[6] LI K，JIU B，WANG P H，et al. Radar active antagonism through deep reinforcement learning：a way to address the challenge of mainlobe jamming[J]. Signal Processing，2021，186：108130.

[7] GENG J，JIU B，LI K，et al. Reinforcement learning based radar anti-jamming strategy design against a non-stationary jammer [C]//Proceedings of the IEEE International Conference on Signal Processing，Communications and Computing（ICSPCC），2022.

[8] WANG S S，LIU Z，XIE R，et al. Reinforcement learning for compressed-sensing based frequency agile radar in the presence of active interference[J]. Remote Sensing，2022，14(4)：968.

[9] AILIYA，YI W，YUAN Y. Reinforcement learning-based joint adaptive frequency hopping and pulse-width allocation for radar anti-jamming [C]//Proceedings of the IEEE Radar Conference，2020.

[10] AILIYA，YI W，VARSHNEY P K. Adaptation of frequency hopping interval for radar anti-jamming based on reinforcement

learning[J]. IEEE Transactions on Vehicular Technology, 2022, 71(12): 12434 – 12449.

[11] LI H Y, HAN Z W, PU W Q, et al. Counterfactual regret minimization for anti-jamming game of frequency agile radar[C]//Proceedings of the IEEE 12th Sensor Array and Multichannel Signal Processing Workshop (SAM), 2022.

[12] LI K, JIU B, PU W Q, et al. Neural fictitious self-play for radar antijamming dynamic game with imperfect information[J]. IEEE Transactions on Aerospace and Electronic Systems, 2022, 58(6): 5533 – 5547.

[13] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. Cambridge: the MIT Press, 1998.

[14] FANG Y Y, ZHANG L, WEI S P, et al. Online frequency-agile strategy for radar detection based on constrained combinatorial nonstationary bandit[J]. IEEE Transactions on Aerospace and Electronic Systems, 2023, 59(2): 1693 – 1706.

[15] THORNTON C E, BUEHRER R M, MARTONE A F. Constrained contextual bandit learning for adaptive radar waveform selection[J]. IEEE Transactions on Aerospace and Electronic Systems, 2022, 58(2): 1133 – 1148.

[16] HOI S C H, SAHOO D, LU J, et al. Online learning: a comprehensive survey[J]. Neurocomputing, 2021, 459: 249 – 289.

[17] AK S, BRUGGENWIRTH S. Avoiding jammers: a reinforcement learning approach[C]//Proceedings of the IEEE International Radar Conference (RADAR), 2020.