融合动作剔除的深度竞争双 Q 网络智能干扰决策算法

饶　宁，许　华，宋佰霖
(空军工程大学信息与导航学院，西安，710077)

摘要　为解决战场通信干扰决策问题，设计了一种融合动作剔除的深度竞争双 Q 网络智能干扰决策方法。该方法在深度双 Q 网络框架基础上采用竞争结构的神经网络决策最优干扰动作，并结合优势函数判断各干扰动作的相对优劣，在此基础上引入无效干扰动作剔除机制加快学习最佳干扰策略。当面对未知的通信抗干扰策略时，该方法能学习到较优的干扰策略。仿真结果表明，当敌方通信策略发生变化时，该方法能自适应调整干扰策略，稳健性较强，和已有方法相比可达到更高的干扰成功率，获得更大的干扰效能。

关键词　干扰决策；深度双 Q 网络；竞争网络；干扰动作剔除
DOI　10.3969/j.issn.1009-3516.2021.04.014
中图分类号　TN975　　**文献标志码**　A　　**文章编号**　1009-3516(2021)04-0092-07

An Intelligent Jamming Decision Algorithm Based on Action Elimination Dueling Double Deep Q Network

RAO Ning, XU Hua, SONG Bailin
(Information and Navigation College, Air Force Engineering University, Xi'an 710077, China)

Abstract　In view of the problem of intelligent jamming decision-making in battlefield communication, an action elimination dueling double deep Q network intelligent jamming decision algorithm is designed. Based on the framework of double deep Q network, this algorithm utilizes a neural network with a dueling structure for determining the optimal jamming action in combination with the advantage function to judge the relative pros and cons of each jamming action. And then on the basis of the above mentioned, an invalid jamming action elimination mechanism is introduced to speed up the learning of the best jamming strategy. The method can learn a better jamming strategy in the face of an unknown communication anti-jamming strategy. The simulation results show that when the enemy changes communication strategy, the method can adaptively adjust the jamming strategy and have stronger robustness. Compared with the existing methods, this method can achieve a higher jamming success rate with greater jamming efficacy.

Key words　jamming decision-making; double deep Q network; dueling network; jamming action elimination

电磁空间是继陆、海、空、天的第五维战场，电子对抗是在电磁空间进行军事斗争的主要手段。在感知、决策、行动、评估的闭环电磁频谱作战过程中，干扰决策是进行有效对抗的重要环节，然而目前人工

收稿日期：2020-12-29
作者简介：饶　宁(1997—)，男，江西上饶人，硕士生，研究方向：智能通信对抗。E-mail：raoningmabma@163.com
通信作者：许　华(1976—)，男，湖北宜昌人，教授，博士，研究方向：通信信号处理、盲信号处理、通信对抗。E-mail：13720720010@139.com

引用格式：饶宁，许华，宋佰霖.融合动作剔除的深度竞争双 Q 网络智能干扰决策算法[J]. 空军工程大学学报(自然科学版)，2021，22(4)：92-98. RAO Ning, XU Hua, SONG Bailin. An Intelligent Jamming Decision Algorithm Based on Action Elimination Dueling Double Deep Q Network[J]. Journal of Air Force Engineering University (Natural Science Edition), 2021, 22(4): 92-98.

决策的实时性与科学性较差，难以适应未来战场瞬息万变的态势。近年智能决策成为研究热点，出现了基于遗传算法、粒子群算法[1-2]等优化理论的干扰参数寻优方法，这些方法需要较多的先验信息，实用性不强。而随着人工智能技术的迅速发展，无需先验信息的强化学习理论在电子战领域得到初步应用。如 Amuru 等人[3]将决策干扰参数的过程建模为多臂赌博机模型，提出干扰赌博机(jamming bandit，JB)算法，该算法可自适应地优化干扰信号直至最佳；在干扰信号参数方面，颛孙少帅等人[4]提出一种双层强化学习的干扰决策算法，以牺牲交互时间来提升算法收敛速度，决策干扰参数。此外，该团队还利用正强化的思想来提高最优动作被选中的概率，以更少的交互次数获得更好的干扰效果[5]。在雷达对抗领域，邢强等人[6]针对雷达工作模式及数目未知情况，研究基于 Q 学习的智能雷达对抗方法，可实现秒量级的收敛速度。黄星源等人[7]利用双 Q 学习对战场的干扰效果进行自主学习，实现对雷达干扰资源的认知决策。Q 学习作为一种基于表格搜索型的强化学习算法，无法解决高维决策问题。而深度 Q 网络(deep Q network，DQN)[8]可利用神经网络进行 Q 学习算法中的函数拟合，能够处理高维的态势信息。基于此，张柏开等人[9]提出了对多功能雷达的 DQN 认知干扰决策方法，当可执行的任务数量增多时依然有较好的决策效率。

上述研究主要解决静态环境中的干扰参数决策和资源分配问题，很少研究变化环境中的决策问题，并且有关通信电子战的智能决策研究主要针对通信方使用固定通信参数时如何学习干扰参数，而实际场景中，通信方受到干扰后通常会优先选择切换波道以躲避干扰。因此本文研究侧重频率击中的智能干扰决策，针对通信干扰决策问题提出一种融合动作剔除的深度竞争双 Q 网络智能干扰决策方法(action elimination dueling double deep Q network，AED3QN)。该方法在 Double DQN 算法基础上通过采用竞争结构的神经网络决策干扰方案，并引入干扰动作剔除机制来加快学习最佳干扰策略。当通信方采用未知且变化的通信抗干扰策略时，相对已有算法该算法能更快地学习到对应的干扰策略，实现更高的干扰成功率并获得更大的干扰收益。

1　对抗场景与马尔科夫决策过程

1.1　对抗场景

通信干扰需建立在频率击中、空域对准、能量压制等条件基础上，本文研究侧重频率击中的智能干扰决策，旨在考查决策算法对对手内在规律的学习适应能力，因此仅假设通信方使用定频通信系统。如图 1 所示，通信方在目标频谱区域内有 N 个可选择的波道，假设每个波道 f_i 对应的是互不干扰、相互独立的等带宽信道。通信方可以根据作战想定制定波道切换策略，通过在不同的波道之间切换来躲避干扰方的干扰。干扰方利用智能决策算法学习通信方的通信行为，实现对通信方选择通信波道的跟踪和预测并加以干扰。在战场条件下通信方受到干扰后会切换通信波道，为便于分析假设通信方的抗干扰策略是利用随机种子生成伪随机数列，受到干扰后根据伪随机数列切换通信波道，并且每隔一定时间变更随机种子，重新生成新的波道伪随机数列进行通信。为便于对比分析，将连续时间变为离散的时隙，并将对抗过程简化如下：在每个时隙 t 的起始时刻，通信方可以随机选择 N 个波道中的任意一个波道 $f_t^c(c=1,2,\cdots,N)$ 进行通信。同时干扰方根据当前的频谱状态信息，利用干扰算法选择 N 个波道中的一个波道 $f_t^j(j=1,2,\cdots,N)$ 进行干扰。图 1 中，只有通信方和干扰方在同一个时隙内选择了同一个波道表明双方存在波道碰撞 ($f_t^c=f_t^j$)，才可认为干扰方进行干扰。如果通信方在一个时隙内未受到干扰，则在下一个时隙将继续使用此波道进行通信；否则将按原策略在下一个时隙起始时刻重新选择一个通信波道。

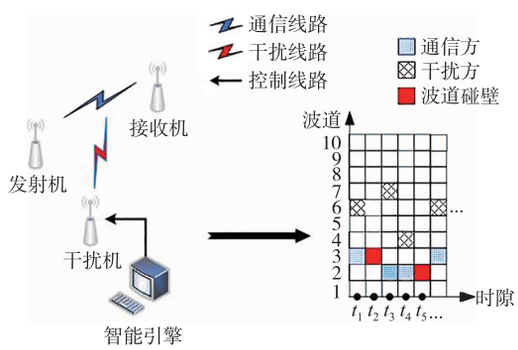


图 1　波道碰撞示意图

假设通信方传输数据采用数字调制进行通信，设通信信号的低通等效表达式为：

$$x(t)=\sum_{m=-\infty}^{\infty}\sqrt{P_x}x_m g(t-mT) \tag{1}$$

式中：P_x 表示通信接收机收到的平均信号功率；$g(t)$表示实值脉冲波形；T 是码元间隔；x_m 是随机变量表示该数字调制方式的码元符号。

设干扰信号的低通等效表达式为：

$$j(t)=\sum_{m=-\infty}^{\infty}\sqrt{P_J}j_m g(t-mT) \tag{2}$$

由于通信收发双方是完全同步，故在经过匹配滤波和抽样判决后在通信接收机处收到的信号表达

式为：

$$y_k=y(t=KT,f_t^c,f_t^j)=(\sqrt{P_x}x_k+n_k)+\sqrt{P_J}j_k\cdot\varphi(f_t^c,f_t^j),k=1,2,\cdots \tag{3}$$

式中：n_k 是方差为 σ^2 的零均值加性高斯白噪声；$\varphi(f_t^c,f_t^j)$ 表示波道碰撞指示函数，可表示为：

$$\varphi(f_t^c,f_t^j)=\begin{cases}1,f_t^c=f_t^j\\0,f_t^c\neq f_t^j\end{cases} \tag{4}$$

只有当 $\varphi(f_t^c,f_t^j)=1$ 时代表干扰机与通信方在同一波道，接收机才会受到干扰信号的影响。为便于分析干扰效果，假设通信方使用包含确认帧/非确认帧的通信协议机制[3]，根据回传的非确认帧数量决定是否切换波道。干扰方亦通过侦察通信方的非确认帧数量决定是否改变干扰波道。通信方接收机处的信噪比和干信比可分别表示为：

$$SNR=\frac{P_x}{\sigma^2},JNR=\frac{P_J}{\sigma^2} \tag{5}$$

式中：σ^2 为环境噪声方差。

1.2 马尔科夫决策过程

根据对抗场景，本文将干扰通信波道的场景建模为马尔可夫决策过程(MDP)[10]。马尔可夫决策过程可用元组$<\boldsymbol{S},\boldsymbol{A},\boldsymbol{P},\boldsymbol{R}>$表示，其中 $\boldsymbol{S}$ 代表状态空间，$\boldsymbol{A}$ 代表动作空间，$\boldsymbol{P}$ 代表状态转移概率，$\boldsymbol{R}$ 代表奖励函数。4 个元素具体定义如下：

状态空间 $\boldsymbol{S}$：在时隙 t，环境的状态可表示为：

$$s_t=(f_c,f_j) \tag{6}$$

式中：f_c 为 t 时隙通信方所在波道；f_j 为 t 时隙干扰方所在波道。其中 $f_c\in\{1,2,\cdots,N\}$，$f_j\in A$，A 为干扰方的动作空间。

动作空间 $\boldsymbol{A}$：在时隙 t，干扰方会根据当前算法选择一个波道进行干扰，干扰动作表示为 a_t，$a_t\in\{1,2,\cdots,N\}$。

状态转移概率矩阵 $\boldsymbol{P}$：在时隙 t，干扰方根据当前所处的环境状态 s_t 选择动作 a_t，环境转移到下一个时隙 $t+1$ 状态 s_{t+1}，则状态转移概率为：

$$p(s'|S,a)=\Pr\{S_{t+1}=s'|S_t=s,A_t=a\} \tag{7}$$

且满足：

$$\sum_{s'\in S}\sum_{s'\in S,a\in A(s)}p(s'\mid s,a)=1 \tag{8}$$

奖励函数 $\boldsymbol{R}$：假设在时隙 t 环境状态为 s_t，干扰方选择干扰动作 a_t，环境达到状态 s_{t+1} 后干扰方可获得奖励 r。干扰方的目标是保持持续稳定的干扰，因此在确保当前干扰成功的条件下，也需要准确预测出通信方在受到干扰后会选择的下个波道。故规定干扰方某时隙干扰成功获得的收益与到当前时隙为止干扰方连续干扰成功的时隙总数成正比，干扰方某时隙干扰失败获得的收益与到当前时隙为止通信方连续正常通信的时隙数成反比。定义干扰奖励函数为：

$$r=\begin{cases}k(t_2-t_1), & \varphi(f_t^c,f_t^j)=1\\-k(t_4-t_3), & \varphi(f_t^c,f_t^j)=0\end{cases} \tag{9}$$

式中：k 为比例常数；t_1、t_2 构成的时隙区间$[t_1,t_2]$($t_2>t_1$)，表示通信方在此区间内受到干扰方连续干扰；t_3、t_4 构成的时隙区间$[t_3,t_4]$($t_4>t_3$)，表示通信方在此区间内均正常通信。

将干扰方获得的干扰总收益定义为所有时隙内获得的奖励总和即：

$$R=\sum_{t=1}^{T}r_t \tag{10}$$

式中：t 为通信时隙；r_t 为干扰方在该时隙获得的干扰收益。

2 融合动作剔除的深度竞争双 Q 网络决策算法

2.1 动作剔除

在缺少先验信息时，干扰方对于何种干扰动作的干扰效果最好无从得知，常常需要尝试不同的干扰动作去进行探索。而在实际环境中尝试不同干扰动作成本及风险较大，故需兼顾利用目前已知效果较好的干扰动作。面对探索和利用的困境，DQN 算法[9]和 Q 学习算法[6]均采用多臂赌博机中的 ε-greedy 策略，如式(11)所示：

$$\pi(a|s)=\begin{cases}\arg\max\limits_a Q(s,a), & prob=1-\varepsilon\\ \text{random select}, & prob=\varepsilon\end{cases} \tag{11}$$

即以 $1-\varepsilon$ 的概率选择当前状态下收益最高的动作，以 ε 的概率进行随机选择。

本文借鉴文献[11]提出的多臂赌博机策略 EUCBV，利用干扰动作的干扰效能设置置信上界值，从干扰动作集合中剔除干扰效能低于该上界值得干扰动作，减少对无效干扰动作不必要的探索，如图 2 所示。

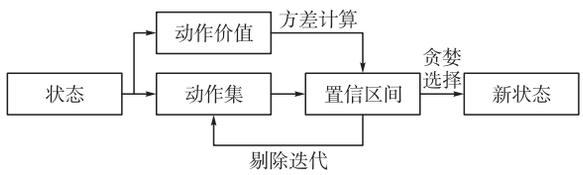


图 2 EUCBV 策略

EUCBV 策略为：

$$\pi(a|s)=\begin{cases}\arg\max\limits_{a\in A(s)}\left\{\hat{r}_a+\sqrt{\dfrac{\rho(\nu_a+2)\ln(\psi T\varepsilon_m)}{4N(s,a)}}\right\}, & \forall N(s,a)\neq 0\\ a, & \forall N(s,a)=0\end{cases} \tag{12}$$

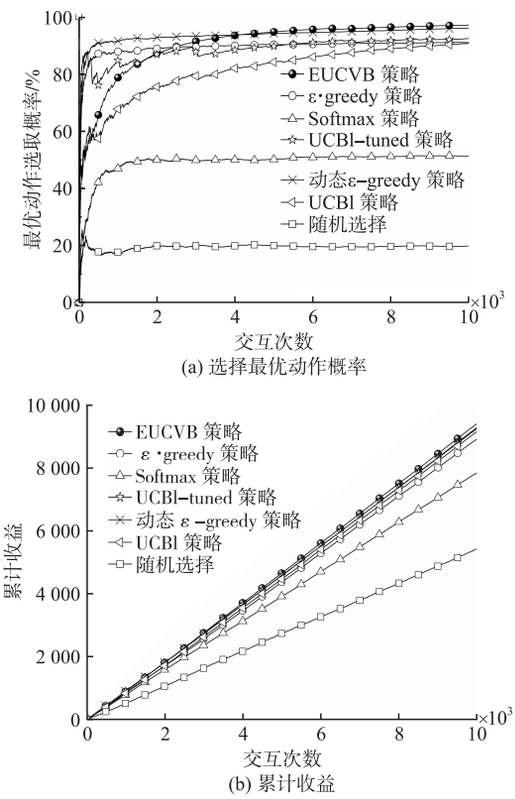
式中：ρ、ψ、ε_m 均为常数；T 为总执行次数；$\hat{r}_a$ 为在状态 s 执行动作 a 后的即时奖励；$\hat{\nu}_a$ 为动作价值的方差；$N(s,a)$为在状态 s 执行动作的总次数。

依据各动作的效能设置置信上界值，剔除无效动作，即若动作 i 满足式(13)，则剔除动作 i。

$$\hat{r}_i+\sqrt{\frac{\rho(\hat{\nu}_i+2)\ln(\psi T\varepsilon_m)}{4N(s,i)}}<\max_{j\in A(s)}\left\{\hat{r}_j-\sqrt{\frac{\rho(\hat{\nu}_j+2)\ln(\psi T\varepsilon_m)}{4N(s,j)}}\right\} \tag{13}$$

式中：$A(s)$表示在状态 s 的可选动作集合。

EUCBV 策略和经典多臂赌博机策略如 UCB1 等策略[12-14]的性能对比见图 3。



(a) 选择最优动作概率

(b) 累计收益

图 3　策略对比

从图 3 可看出 EUCBV 策略在解决探索-利用困境中表现最佳，证明了在未知环境中通过估计价值方差来剔除无效干扰动作的可行性。

2.2　深度竞争双 Q 网络

Q 学习和 DQN 算法在估计动作价值时均采用选取最大估计值，如式(14)，这在学习过程中会导致过估计，最终使得学到的策略偏离最佳策略。

$$y^{Q}=r+\gamma Q[s',\arg\max_a Q(s,a;\theta);\theta] \tag{14}$$

式中：s 表示状态；a 表示动作；r 为执行动作后获得的即时奖励；γ 为折扣因子；θ 为网络参数；$Q(s,a;\theta)$表示 Q 函数。

针对过估计问题，借鉴文献[15]利用在线网络进行动作选择，本文利用目标网络估算其价值降低过估计对算法学习过程的影响，见式(15)：

$$y^{DoubleQ}=r+\gamma Q[s',\arg\max_a Q(s,a;\theta);\theta^-] \tag{15}$$

式中：θ 为在线网络的网络参数，θ^- 为目标网络的网络参数。

此外，为了进一步比较相同环境状态下不同干扰动作的优劣，更准确地剔除无效动作，借鉴文献[16]，采用竞争结构的神经网络，引入优势函数来评估某个干扰动作在当前状态相对其他动作的好坏程度。如图 4 所示，将全连接神经网络的单个输出改为两个输出，一个输出当前状态的价值，另一个输出干扰动作的优势函数，最终合并为干扰动作的 Q 函数。

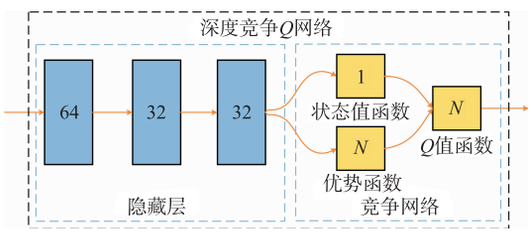


图 4　竞争网络结构

干扰动作的 Q 函数表示为：

$$Q(s,a)=V(s;\theta,\alpha)+A(s,a;\theta,\beta) \tag{16}$$

式中：$V(s;\theta,\alpha)$表示状态值函数，$A(s,a;\theta,\beta)$表示优势函数，α 和 β 分别表示各函数对应的参数变量。实际工程应用中，常用优势函数中各个动作的平均值代替优势函数，可将式(13)改写为：

$$Q(s,a)=V(s;\theta,\alpha)+\left[A(s,a;\theta,\beta)-\frac{1}{|A|}\sum_{a'}A(s,a';\theta,\beta)\right] \tag{17}$$

引入竞争网络结构后，可将每个动作的 Q 值拆分为状态值函数加上每个动作的优势函数。而优势函数恰恰体现了该动作的相对优劣，故将优势函数替换式中的即时奖励部分，利用优势函数表征的动作相对优劣情况可得到更准确的无效动作剔除方法，如式(18)所示。

$$A(s,i;\theta,\beta)+\sqrt{\frac{\rho(\hat{\nu}_i+2)\ln(\psi T\varepsilon_m)}{4N(s,i)}}<\max_{j\in A(s)}\left\{A(s,j;\theta,\beta)-\sqrt{\frac{\rho(\hat{\nu}_j+2)\ln(\psi T\varepsilon_m)}{4N(s,j)}}\right\} \tag{18}$$

2.3　融合动作剔除的深度竞争双 Q 网络智能干扰决策算法

本文在深度竞争双 Q 网络基础上，引入无效干扰动作剔除机制，结合对抗场景提出了融合动作剔除的深度竞争双 Q 网络智能干扰决策算法(AED3QN)。

算法框架如图 5 所示。AED3QN 算法包含两个神经网络，分别是在线决策网络和价值评估网络，每个网络均采用竞争网络结构。在线决策网络根据当前环境状态 s_t 给出所有干扰动作的干扰效能，根

据式(18)进行无效干扰动作的剔除,在新干扰动作集合中依据贪婪策略选择干扰动作 a_t 并执行。价值评估网络根据该干扰动作并结合环境状态给出该干扰动作的干扰效能 r_t,得到下一个环境状态 s_{t+1},将交互经验(s_t,a_t,s_{t+1},r_t)存入经验回放池。训练时,在经验回放池中随机采样 S 个经验样本,根据式(19)进行梯度下降来训练在线决策神经网络。

$$\begin{aligned} y_t &= r_t + \gamma Q[s_{t+1}, \arg\max_Q(s_{t+1}, a'; \theta); \theta^-] \\ \theta &= \theta + \alpha[y_t - Q(s_t, a_t; \theta)]2\nabla Q(s_t, a_t; \theta) \end{aligned} \tag{19}$$

式中:α 为学习步长。

每隔一定时间将在线决策网络参数赋值给价值评估网络。

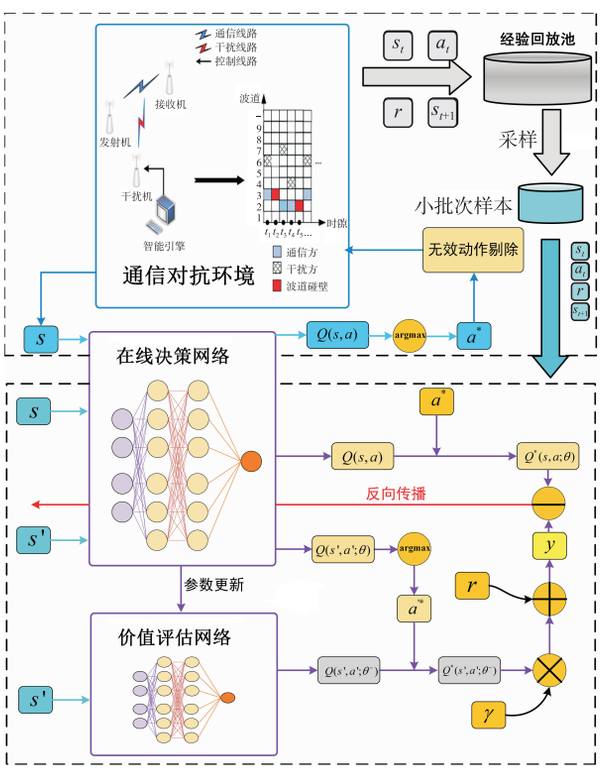


图 5　AED3QN 智能决策算法框图

3　实验仿真与分析

本文在干扰方先验信息较少的条件下,研究当敌方采用切换通信波道的抗干扰手段时,在未知敌抗干扰策略时干扰方如何决策干扰方案才能获得更好的干扰效果。

仿真实验中,通信方有一对信号发射接收机,干扰方只对通信接收机进行干扰,通信接收机可更换通信波道躲避干扰。通信方为达到通信安全目的,采用伪随机波道切换策略,并且每隔一段时间改变通信波道切换的策略。预设的通信波道及通信波道切换策略对于干扰方而言未知,且干扰方为确保功率集中每次只能选择一个波道释放干扰信号。实验从干扰成功率和干扰总收益 2 个方面对比本文算法(AED3QN)、Q 学习算法[6]和 DQN[9]的性能。0～5 MHz 频率范围内划置了 N 个正交波道,设每个波道的带宽均为 B_i,为了减少仿真环境存在的随机性与偶然性,每组仿真实验重复 1 000 次,取 1 000 次仿真实验数据的平均值作为最后的实验结果。实验及模型参数设置见表 1。

表 1　实验及模型参数

参数	初始值
通信调制样式	16-QAM
干扰调制样式	QPSK
正交波道数目 N	50
波道带宽 B_i/kHz	10
每时隙通信符号数	10 000
批次样本数量 S	64
折扣因子 γ	0.9
SNR/dB	20
JNR/dB	15
σ^2	1
训练回合 T	10 000
学习步长 α	0.1

通信方使用伪随机波道切换策略进行通信,通过设置随机种子,产生伪随机数列,数列中的元素代表波道序号,根据伪随机波道序列切换波道。并且每隔一定时间变更随机种子重新生成伪随机波道序列,策略时频图样如图 6 所示,黄色频点表示所在波道的中心频率。

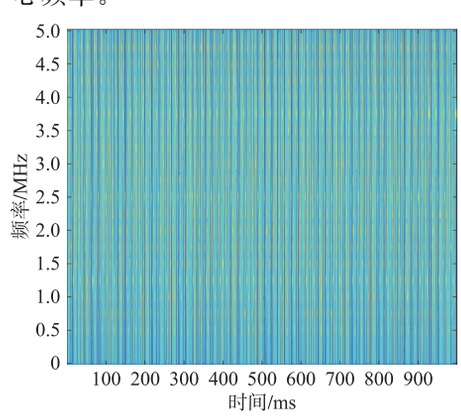
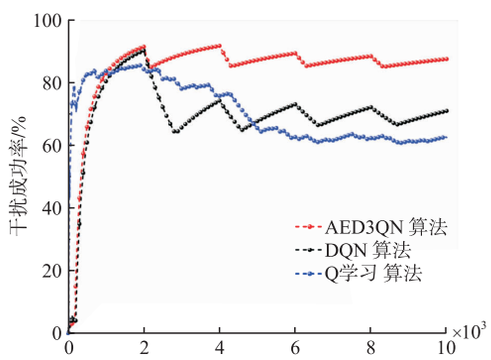


图 6　时频图样

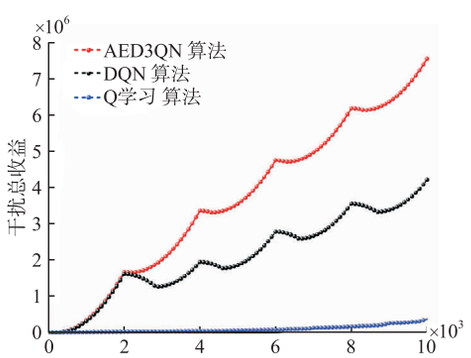
当通信方采用伪随机波道切换策略且每 2 000 回合改变一次随机数种子时,AED3QN 算法、DQN 算法和 Q 学习算法的干扰效果见图 7。

从图 7(a)的干扰成功率曲线可以看到,初始阶段随着训练回合的增加,3 种学习算法通过与环境不断交互,并学习利用交互得到的历史经验,干扰成功率迅速上升。其中 Q 学习算法最先达到 80%的干扰成功率,而 DQN 和 AED3QN 算法曲线轨迹相仿,初始阶段学习速率不及 Q 学习,但在 1 000 回合

后干扰成功率逐渐超过 Q 学习。而当每 2 000 回合通信方改变策略时，在每次策略改变后 3 种算法的干扰成功率均有显著下降，其中 DQN 算法下降幅度最大，原因在于环境的快速改变使得神经网络学习到的拟合函数需重新拟合，而神经网络相比于 Q 表需要更多的数据进行训练。AED3QN 算法由于在训练网络的同时进行无效干扰动作的剔除，降低了决策空间的维度，在敌方策略发生改变后，能更快地学习到对应的干扰方案。图 7(b)中 AED3QN 和 DQN 算法获得的干扰总收益明显高于 Q 学习，从曲线变化趋势可以看到当通信方的策略发生改变时，AED3QN 算法可以更快地学习到应对的干扰方案，表现出比其他 2 种学习算法在变化环境中更强的学习和适应能力。



(a) 干扰成功率

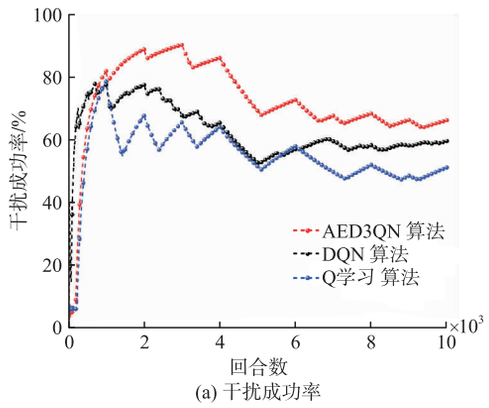


(b) 干扰总收益

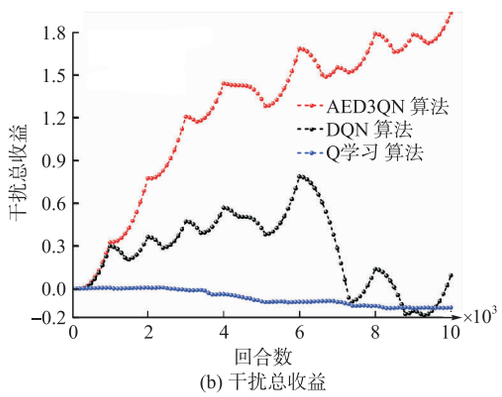
图 7　通信方第 2 000 回合改变策略干扰效果对比

当通信方加快改变波道切换策略的速度时(每过 1 000 回合改变一次通信策略)，此时算法干扰效果见图 8。

由图 8 可知，当环境变化加快时 3 种算法最终的干扰成功率都有相对较大幅度的下降。而从图 8(b)的干扰收益曲线可以看到，在每次环境变化后 AED3QN 算法的干扰收益出现短暂下降后能更快地回升，稳健性更强。表 2 给出了 3 种算法的干扰效果对比。



(a) 干扰成功率



(b) 干扰总收益

图 8　通信方每 1 000 回合改变策略效果对比

表 2　干扰效果对比

策略变化回合	算法	干扰成功率/%	干扰总收益
2 000	AED3QN	87.5	7.51×10^{6}
	DQN	71.0	4.19×10^{6}
	Q 学习	62.5	3.72×10^{5}
1 000	AED3QN	66.1	1.93×10^{6}
	DQN	59.4	9.19×10^{4}
	Q 学习	51.0	-1.34×10^{5}

表 2 中，当通信方改变通信策略时，本文算法无论是干扰成功率还是最终获得的干扰总收益均高于 DQN 和 Q 学习算法。当敌方策略改变后，本文算法能更快地学习到新的对抗方案，在变化环境中表现出更强的稳健性。

4　结语

本文设计了一种通信干扰智能决策方法，在深度双 Q 网络基础上，采用竞争结构的神经网络来输出干扰方案，利用竞争结构中的优势函数进一步对比各干扰动作的优劣，剔除无效的干扰动作，加快算法学习速度。仿真结果表明，当环境发生改变时本文所提出的方法能达到更高的干扰成功率，稳健性更强，与已有方法相比性能更优。但本文也存在一些不足，例如环境发生改变后本文算法仍需要一定

的时间重新学习适应环境,这在连续动态变化的环境中效率不高。今后的工作主要围绕如何更充分地利用与环境交互得到的历史经验,加快重新学习环境模型的速度。

参考文献

[1] BAYRAM S, VANLI N D, DULEK B, et al. Optimum Power Allocation for Average Power Constrained Jammers in the Presence of Non-Gaussian Noise[J]. IEEE Communications Letters, 2012, 8(16): 1153-1156.

[2] AMURU S, BUEHRER R. Optimal Jamming against Digital Modulation[J]. IEEE Transactions on Information Forensics & Security, 2015, 10(10): 2212-2224.

[3] AMURU S, SAIDHIRAJ, TEKIN, et al. Jamming Bandits-A Novel Learning Method for Optimal Jamming[J]. IEEE Transactions on Wireless Communications, 2016, 15(4): 2792-2808.

[4] 颛孙少帅, 杨俊安, 刘辉,等. 采用双层强化学习的干扰决策算法[J]. 西安交通大学学报, 2018, 52(2):63-69.

[5] 颛孙少帅, 杨俊安, 刘辉,等. 基于正强化学习和正交分解的干扰策略选择算法[J]. 系统工程与电子技术, 2018, 40(3): 35-42.

[6] 邢强, 贾鑫, 朱卫纲. 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):76-80.

[7] 黄星源,李岩屹. 基于双 Q 学习算法的干扰资源分配策略[EB/J]. 系统仿真学报. (2020-10-27)[2020-12-29]https://doi.org/10.16182/j.issn1004731x.joss.20-0253.

[8] MNIHL V, KAVUKCUOGLU1 K, SLIVER D, et al. Human-Level Control Through Deep Reinforcement Learning[J]. Nature, 2015, 518(7540):529-540.

[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): 93-99.

[10] NIE J, HAYKIN S. A Q-Learning-Based Dynamic Channel Assignment Technique for Mobile Communication Systems[J]. IEEE Transactions on Vehicular Technology, 1999, 48(5): 1676-1687.

[11] MUKHERJEE S, NAVEEN K P, SUDARSANAM N, et al. Efficient-UCBV: An Almost Optimal Algorithm using Variance Estimates[C]//Thirty-Second AAAI Conference on Artificial Intelligent. San Francisco, CA, USA: AAAI, 2018: 6417-6424.

[12] MAILLARD O A, RÉMI M, Stoltz G. A Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences[J]. Journal of Machine Learning Research, 2011, 19(1):14-25.

[13] AUDIBERT J Y, RÉMI M, CSABA S. Exploration-Exploitation Tradeoff Using Variance Estimates in Multi-Armed Bandits[J]. Theoretical Computer Science, 2009, 410(19): 1876-1902.

[14] AUER P, ORTNER R. UCB Revisited: Improved Regret Bounds for the Stochastic Multi-Armed Bandit Problem[J]. Periodica Mathematica Hungarica, 2010, 61(1-2):55-65.

[15] VAN H, HADO, ARTHUR G, et al. Deep Reinforcement Learning with Double Q-Learning [C]// Thirtieth AAAI Conference on Artificial Intelligent. Phoenix, Arizona, USA: AAAI, 2016: 3878-3884.

[16] WANG Z, SCHAUL T, HESSEL M,et al. Dueling Network Architectures for Deep Reinforcement learning[C]// Thirty-Third International Conference on Machine Learning. New York, USA: ICML, 2016: 1995-2003.

(编辑:徐楠楠)