DOI:10.11918/202010082

# 融合有效方差置信上界的Q学习智能干扰决策算法

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:** 为进一步提升基于值函数强化学习的智能干扰决策算法的收敛速度,增强战场决策的有效性,设计了一种融合有效方差置信上界思想的改进Q学习智能通信干扰决策算法。该算法在Q学习算法的框架基础上,利用有效干扰动作的价值方差设置置信区间,从干扰动作空间中剔除置信度较低的干扰动作,减少干扰方在未知环境中不必要的探索成本,加快其在干扰动作空间的搜索速度,并同步更新所有干扰动作的价值,进而加速学习最优干扰策略。通过将干扰决策场景建模为马尔科夫决策过程进行仿真实验,所构造的干扰实验结果表明:当通信方使用干扰方未知的干扰躲避策略变更通信波道时,与现有基于强化学习的干扰决策算法相比,该算法在无通信方的先验信息条件下,收敛速度更快,可达到更高的干扰成功率,获得更大的干扰总收益。此外,该算法还适用于"多对多"协同对抗环境,可利用动作剔除方法降低联合干扰动作的空间维度,相同实验条件下,其干扰成功率比传统Q学习决策算法高50%以上。

**关键词:** 干扰决策;强化学习;有效方差置信上界;Q学习;干扰动作剔除;马尔科夫决策过程

**中图分类号:** TN975 **文献标志码:** A **文章编号:** 0367-6234(2022)05-0162-09

# Q-learning intelligent jamming decision algorithm based on efficient upper confidence bound variance

RAO Ning, XU Hua, SONG Bailin

(Information and Navigation College, Air Force Engineering University, Xi'an 710077, China)

**Abstract:** To further improve the convergence speed of the intelligent jamming decision-making algorithm based on value function in reinforcement learning and enhance its effectiveness, an improved Q-learning intelligent communication jamming decision algorithm was designed integrating the efficient upper confidence bound variance. Based on the framework of Q-learning algorithm, the proposed algorithm utilizes the value variance of effective jamming action to set the confidence interval. It can eliminate the jamming action with low confidence from the jamming action space, reduce the unnecessary exploration cost in the unknown environment, speed up its searching speed in the interference action space, and synchronously update the value of all actions, thus accelerating the optimal strategy learning process. The jamming decision-making scenario was modeled as the Markov decision process for simulation. Results show that when the correspondent used interference avoidance strategy against the jammer to change the communication channel, the proposed algorithm could achieve faster convergence speed, higher jamming success rate, and greater total jamming rewards, under the condition of no prior information, compared with the existing decision-making algorithms based on reinforcement learning. Besides, the algorithm could be applied to the "many-to-many" cooperative countermeasure environment. The action elimination method was used to reduce the dimension of joint jamming action, and the jamming success rate of the proposed algorithm was 50% higher than those of the traditional Q-learning decision algorithms under the same conditions.

**Keywords:** jamming decision-making; reinforcement learning; efficient upper confidence bound variance; Q-learning; jamming action elimination; Markov decision process

干扰决策是电子战中进行有效对抗的重要环节,而人工决策由于实时性与科学性较差,很难满足战场瞬息万变的态势要求。随着认知无线电[1]等技术的发展,干扰方要达到较好的干扰效果变得更加困难,但无线通信媒介自身的开放性给干扰攻击的实现保留了可行性。近年来,智能干扰已经成为认知电子战的一个重要研究领域,涌现出了利用遗传算法、粒子群算法等[2-3]实现干扰参数寻优的理论,但这些方法都需要通信方的先验通信参数。强化学习作为不需要先验信息的方法,能在未知环境中通过不断与环境交互来学习最优行为策略,目前在通信干扰与抗干扰领域已有初步应用,如文

**收稿日期:** 2020-10-26

**作者简介:** 饶 宁(1997—),男,硕士研究生;
许 华(1976—),男,教授,博士生导师
**通信作者:** 许 华,13720720010@139.com

献[4]将干扰无线网络过程建模为增广马尔科夫决策过程，并利用 Q 学习算法进行干扰决策，通过实验表明干扰方可学习到干扰成本、网络吞吐量等重要信息；文献[5]提出了基于多臂赌博机框架的干扰赌博机（jamming bandit，JB）算法，该在线学习算法通过学习最佳干扰样式对通信方进行干扰。然而，上述研究的强化学习算法学习速度较慢，这在动态变化的战场环境中十分受限。国内外学者进行了相关研究，以期提高学习算法的学习速度，文献[6]采用一种双层强化学习的干扰决策算法，以牺牲交互时间为代价来加快收敛速度；文献[7]利用正强化思想提升最佳动作的概率，并结合对信号正交分解使得算法可以在更短的时间内学习到最佳干扰样式；文献[8]提出在认知无线电网络节点中增加基于同策略 Q 学习算法的合作机制，利用宽带频谱感知和同步贪婪算法来加速学习过程并降低丢包率；文献[9]在迁移学习框架中结合 Soft-Q 算法，将从当前节点获取的抗干扰信息转移到相邻节点来加快该节点的学习过程；文献[10]在策略选择过程中通过增加一个价值附加值来平衡探索与利用，提高算法学习速度。

强化学习算法的学习速度常受探索-利用困境[11]制约，而多臂赌博机策略是平衡探索-利用的常用方法，在 $\varepsilon$-greedy、softmax、置信上界（upper confidence bound，UCB）、UCB1-Tuned[12] 等基础上，又涌现了方差置信上界（upper confidence bound variance，UCBV）及其改进算法[13-14]、有效方差置信上界（efficient upper confidence bound variance，EUCBV）[15]等。

为提高算法学习速度，本文将有效方差置信上界思想引入 Q 学习算法中，通过剔除次优干扰动作缩小动作空间，结合同策略方式贪婪利用当前最优干扰动作来提升稳定性。在认知干扰决策应用场景中，分别在单对单、多对多情形下对算法的干扰成功率和干扰总收益进行研究。

## 1　对抗场景与模型建立

### 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，文献[5-7]主要研究当通信方使用固定通信参数时干扰方如何加快学习干扰参数，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。本文主要研究侧重频域的智能干扰决策，假设通信方使用定频通信系统，且每个通信节点重要性指数相同，见图 1。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、相互独立的等带宽正交信道。每对通信用户（一对通信收发机）根据作战想定制定波道切换的策略，通过切换通信波道来躲避干扰。干扰方利用智能算法来学习其通信策略变化的内在规律，实现对通信波道的跟踪和预测并加以干扰。文中将连续时间设定为离散时隙，简化对抗过程如下：在每个时隙 $t$ 的起始时刻，每个通信用户可以随机选择 $N$ 个波道中的任意一个波道 $f_t^c(c=1,2,\cdots,N)$ 进行通信。每个干扰机根据当前的波道状态以及干扰反馈信息，利用决策算法选择一个波道 $f_t^j(j=1,2,\cdots,N)$ 进行干扰。

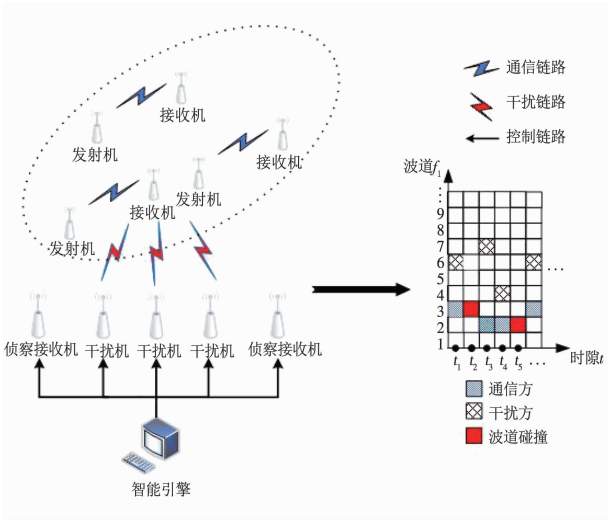


**图 1　对抗场景示意**

Fig. 1　Schematic diagram of countermeasure environment

图 1 中，智能引擎通过智能算法指引各干扰机对通信用户进行联合干扰。考虑到干扰效果实时评估一直是电子战领域的研究难题[16]，为便于分析，假设通信方使用包含确认帧/非确认帧（ACK/NACK）信息的协议机制[5]，且成功干扰需同时满足：1）通信用户和干扰方在同一个时隙内选择了同一个波道（$f_t^c=f_t^j$）；2）干扰方侦收到的单位时间 NACK 数据包数量 $\tau$ 大于预设定门限值 $\tau_0$。单位时间 NACK 平均数量可由误包率给出，传输出错时需回传 NACK 包[5]。通信接收双方亦根据 ACK/NACK 信息决定是否切换波道，如果通信用户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通信波道。只有当某时隙网内所有的通信用户都受到干扰，才可认为协同干扰有效。

### 1.2　模型建立

假设通信方传输数据采用数字调制，设通信信号低通等效表达式为

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

式中：$P_x$ 为通信接收机收到的平均信号功率；$g(t)$ 为实值脉冲波形；$T$ 为码元间隔；$x_m$ 是随机变量，为该数字调制方式的码元符号，假设其在所有可能的星座点上均匀分布，并对能量进行归一化，即有 $E(|x_m|^2)\leqslant 1$。

假设信号 $x(t)$ 在高斯白噪声（AWGN）信道中进行传输，由文献[3]可知对于数字调制信号，最佳干扰为脉冲率为 $q\in(0,1)$ 的脉冲干扰。设干扰信号的低通等效表达式为

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

式中：$P_j$ 为通信接收机收到的干扰信号功率；$j_m$ 是随机变量，表示干扰信号的码元符号。假设通信收发双方完全同步，则经过匹配滤波和抽样判决后在接收机处收到的信号表达式为

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

式中：$n_k$ 是方差为 $\sigma^2$ 的零均值加性高斯白噪声；$\varphi(f_t^c,f_t^j)$ 为波道碰撞指示函数，可由下式表示：

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

只有当 $\varphi(f_t^c,f_t^j)=1$ 时，表示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。接收机处的信噪比和干扰方的干噪比可分别表示为

$$R_{\mathrm{SN}}=\frac{P_x}{\sigma^2},R_{\mathrm{JN}}=\frac{P_j}{\sigma^2} \tag{5}$$

结合上文对抗场景可知当前环境状态只取决于上一时刻的状态，与过去所经历的无关，满足马尔科夫链属性，故将通信干扰决策问题建模为马尔科夫决策过程（MDP）[17]。MDP 可用元组 $<\boldsymbol{S},\boldsymbol{A},\boldsymbol{P},\boldsymbol{R}>$ 表示，其中 $\boldsymbol{S}$ 代表环境状态空间，$\boldsymbol{A}$ 代表干扰动作空间，$\boldsymbol{P}$ 代表状态转移概率，$\boldsymbol{R}$ 代表奖励函数。4 个元素的具体定义如下：

环境状态空间 $\boldsymbol{S}$：“一对一”（单个干扰机对抗单对通信用户）场景中，在时隙 $t$，环境的状态可表示为

$$s_t=(f_t^c,f_t^j,\tau) \tag{6}$$

式中：$f_t^c$ 为 $t$ 时隙通信方所在波道且 $c=1,2,\cdots,N$；$f_t^j$ 为 $t$ 时隙干扰机所干扰的波道且 $f_t^j\in\boldsymbol{A}$，$\boldsymbol{A}$ 为干扰机的动作空间；$\tau$ 为单位时间侦收到的 NACK 包数量。

在“多对多”场景中，环境状态可表示为

$$s_t=\{(f_t^{c_1},f_t^{j_1},\tau_1),(f_t^{c_2},f_t^{j_2},\tau_2),\cdots,(f_t^{c_i},f_t^{j_i},\tau_i)\} \tag{7}$$

式中：$i$ 为网内通信用户编号，$i=2,3$；$f_t^{c_i}$ 为 $t$ 时隙第 $i$ 个用户所在波道，$f_t^{j_i}$ 为 $t$ 时隙第 $i$ 个干扰机所干扰的第 $i$ 个波道。

干扰动作空间 $\boldsymbol{A}$：“一对一”场景中，干扰动作为 $a_t,a_t\in\{1,2,\cdots,N\}$。“多对多”场景中，在时隙 $t$ 第 $i$ 个干扰机根据智能引擎指令选择干扰动作，将联合干扰动作表示为

$$a_t=(a_t^1,a_t^2,\cdots,a_t^i) \tag{8}$$

式中 $a_t^i$ 为第 $i$ 个干扰机的干扰动作。

状态转移概率矩阵 $\boldsymbol{P}$：在时隙 $t$，干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

$$p(s'|s,a)=\Pr\{S_{t+1}=s'|S_t=s,A_t=a\} \tag{9}$$

且满足

$$\sum_{s'\in S}\sum_{s\in S,a\in A(s)}p(s'|s,a)=1 \tag{10}$$

本文在无通信方的先验信息条件下研究干扰决策，即状态转移概率对于干扰方是未知的。

干扰收益函数 $\boldsymbol{R}$：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 $s_{t+1}$ 后干扰方获得收益 $R$。由于连续的干扰会引起通信接收端接收数据包的急剧恶化，本文将收益函数定义为某时隙侦收到的单位时间内 NACK 包数量，同时规定干扰方某时隙干扰成功获得的收益，正相关于到当前时隙为止干扰方连续干扰成功的时隙总数与当前时隙 NACK 包数之积；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非正常通信的时隙数成负相关，如式（11）所示：

$$R=\begin{cases}k*(t_2-t_1)*\tau,\varphi(f_t^c,f_t^j)=1,\tau>\tau_0\\-k*(t_4-t_3),\varphi(f_t^c,f_t^j)=0\end{cases} \tag{11}$$

式中：$k$ 为比例常数；$\tau$ 为该时隙干扰方侦收到的 NACK 包数量；$\tau_0$ 为预设定门限值；$t_1$、$t_2$ 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各个干扰机获得干扰收益之和，即

$$R_{\mathrm{sum}}=\sum_{t=0}^{T}R_t^i \tag{12}$$

式中 $t$ 为通信时隙，$R_t^i$ 为第 $i$ 个干扰机在该时隙的干扰收益，$i=1,2,3$。干扰方的最佳干扰策略 $\pi^*$ 是在一定时间内最大化干扰总收益 $R_{\mathrm{sum}}$，表示为

$$\pi^*=\mathrm{argmax}_{\pi}E_{\tau\sim\pi(\tau)}[R_{\mathrm{sum}(\tau)}] \tag{13}$$

## 2 基于改进 Q 学习的智能干扰决策算法

### 2.1 融合有效方差置信上界的改进 Q 学习算法

强化学习方法按是否已知环境状态的转移概率

可分为基于模型和无模型两类算法[18]，按学习方式可分为基于值函数和基于策略函数的强化学习算法。Q 学习[19]是一种无模型且基于值函数的算法，其算法框架见图 2。

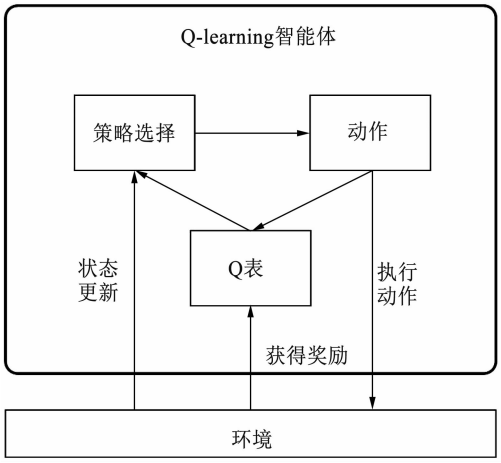


**图 2　Q 学习算法框图**

Fig. 2　Block diagram of Q-learning algorithm

智能体在当前状态 $S_t$ 通过与环境不断交互，根据学习到的知识，选择执行动作 $A_t$ 后获得反馈信息 $R_{t+1}$ 并转移到下个状态 $S_{t+1}$，按式(14)更新所采取动作的价值并存入 Q 表。

$$Q(S_t,A_t)=Q(S_t,A_t)+\alpha[R_{t+1}+\gamma\max_a Q(S_{t+1},a)-Q(S_t,A_t)] \quad (14)$$

式中：$\alpha\in(0,1)$ 为学习步长，表征算法每次迭代的步长，$\alpha$ 越大则每次更新的幅度就越大，提升了学习速度但会增加波动性，$\alpha$ 越小则学习速度较慢但更稳定；$\gamma\in(0,1)$ 为折扣因子，体现智能体对当前动作在将来产生影响的重视程度，$\gamma$ 越小表示智能体越重视眼前收益，$\gamma$ 越大表示智能体更注重长远收益。

在学习过程中，Q 学习算法在每个状态遵循 $\varepsilon$-greedy 策略选择动作，如式(15)所示：

$$\pi(a|s)=\begin{cases}\underset{a}{\operatorname{argmax}}Q(s,a), & p=1-\varepsilon\\ \text{randomly select}, & p=\varepsilon\end{cases} \quad (15)$$

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

本文为实现智能体在未知环境中能较好地从探索阶段过渡到利用阶段，将有效方差置信上界思想引入到 Q 学习算法中。有效方差置信上界核心思想是在置信区间中加入方差值，通过剔除不在置信区间的动作减小探索成本，见图 3。

在此基础上，借鉴文献[8]将使用异策略随机探索转为使用同策略贪婪利用，提出了融合有效方差置信上界的改进 Q 学习算法。在引入置信区间的同时，根据各个动作的估计价值上限，剔除不在置信区间的动作，以削减在动作空间的探索成本，使得算法能更快收敛到最佳动作，且同步更新该状态的所有动作价值来更充分地利用历史信息，使得算法在任意一个状态都能准确选择价值最高的动作。算法框架见图 4。

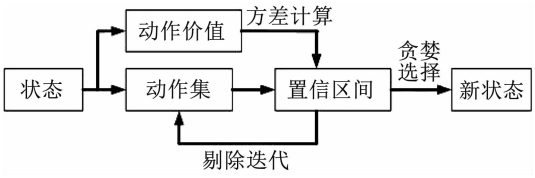


**图 3　有效方差置信上界策略**

Fig. 3　Efficient upper confidence bound variance strategy

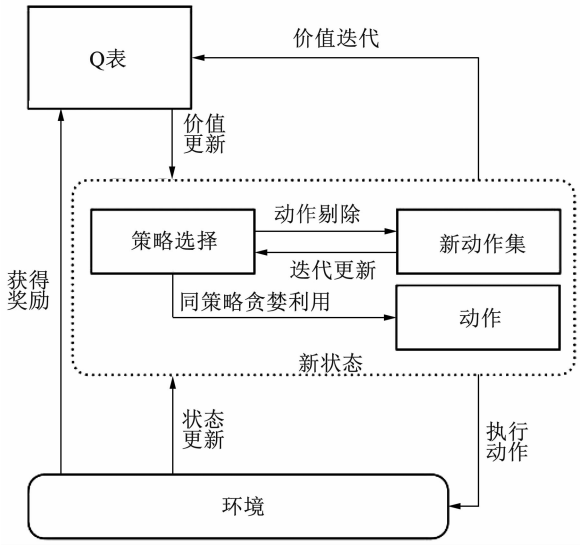


**图 4　融合有效方差置信上界的 Q 学习算法框图**

Fig. 4　Block diagram of Q-learning algorithm based on efficient upper confidence bound variance

本文算法具体过程如下：

首先设置步长 $\alpha\in(0,1]$、折扣因子 $\gamma>0$、迭代次数 $T$，初始状态集合为 $S^+$，状态 $s$ 对应的动作集为 $A(s)$，$s\in S^+$；初始化状态动作对价值矩阵 $\boldsymbol{Q}(s,a)=\boldsymbol{0}_{|S|\times|A|}$，状态动作对平均奖励矩阵 $\boldsymbol{P}(s,a)=\boldsymbol{0}_{|S|\times|A|}$，状态动作对执行次数矩阵 $\boldsymbol{N}(s,a)=\boldsymbol{0}_{|S|\times|A|}$，$s\in S^+$，$a\in A(s)$；探索因子 $\rho\in(0,1]$，常量 $\psi=\dfrac{T}{|A(s)|^2}$，初始化动作剔除轮次 $m=0$ 和剔除因子 $\varepsilon_m=1$，剔除轮次下限值 $M=\lfloor\dfrac{1}{2}\log_{\mathrm{e}}\dfrac{T}{\mathrm{e}}\rfloor$，e 为自然对数的底，设置动作探索访问上界值 $n_0=\lceil\dfrac{\log_{\mathrm{e}}(\psi T\varepsilon_0^2)}{2\varepsilon_0}\rceil$，所有动作总访问上界值为 $N_0=|A(s)|n_0$。

在当前状态 $s$，智能体遵循策略 $\pi(a|s)$ 即选择动作，则

$$\pi(a|s)=\begin{cases}\underset{a\in A(s)}{\operatorname{argmax}}\left\{\hat{r}_a+\sqrt{\dfrac{\rho(\hat{v}_a+2)\log_e(\psi T\varepsilon_m)}{4N(s,a)}}\right\} & \forall N(s,a)\neq 0\\ a & \exists N(s,a)=0\end{cases} \tag{16}$$

式中 $\hat{v}_a$ 为动作价值方差，$\hat{r}_a$ 为该状态动作对的平均奖励。动作价值方差计算如下：

$$\hat{v}_a=\frac{1}{N(s,a)}\sum_{l=1}^{N(s,a)}(r_{a,l}-\hat{r}_a)^2 \tag{17}$$

随后进行当前状态的动作剔除，如果动作 $i$ 满足式(18)，则从动作空间 $A(s)_m$ 剔除动作 $i$，得到新的动作集 $A(s)'_m$：

$$\hat{r}_i+\sqrt{\frac{\rho(\hat{v}_i+2)\log_e(\psi T\varepsilon_m)}{4z_i}}<\max_{j\in A(s)}\left\{\hat{r}_j-\sqrt{\frac{\rho(\hat{v}_j+2)\log_e(\psi T\varepsilon_m)}{4z_j}}\right\} \tag{18}$$

当时间步 $t\geqslant N_m$ 且 $m\leqslant M$ 时，更新剔除参数：

剔除因子

$$\varepsilon_{m+1}=\frac{\varepsilon_m}{2} \tag{19}$$

当前轮次所有状态对应的动作集合

$$A(s)_{m+1}=A(s)'_m,\ \forall s\in S^+ \tag{20}$$

动作探索访问上界值

$$n_{m+1}=\left\lceil\frac{\log_e(\psi T\varepsilon_m^2)}{2\varepsilon_m}\right\rceil \tag{21}$$

所有动作总访问上界值

$$N_{m+1}=t+|A(s)_{m+1}|n_{m+1} \tag{22}$$

动作剔除轮次

$$m=m+1 \tag{23}$$

执行动作 $a$，观察反馈奖励 $r$ 和下一状态 $s'$，按式(24)同步更新与状态 $s$ 关联的该动作集合中所有动作价值：

$$\forall a\ Q(s,a)=Q(s,a)+\alpha[r+\gamma\max_a Q(s',a)-Q(s,a)] \tag{24}$$

式中 $\alpha$ 是学习步长，为确定的非负常数，且满足下式条件：

$$\alpha\in[0,1),\ \sum_{t=1}^{\infty}\alpha_t<\infty,\ \sum_{t=1}^{\infty}(\alpha_t)^2<\infty \tag{25}$$

**算法 1** 融合有效方差置信上界的改进 Q 学习算法

**步骤 1**：设置学习步长 $\alpha\in(0,1]$，折扣因子 $\gamma>0$，迭代次数 $T$，初始化矩阵 $\boldsymbol{Q}(s,a)=\boldsymbol{0}_{|S|\times|A|}$，$\boldsymbol{P}(s,a)=\boldsymbol{0}_{|S|\times|A|}$，$\boldsymbol{N}=\boldsymbol{0}_{|S|\times|A|}$，$\rho\in(0,1]$，$\psi=\dfrac{T}{|A(s)|^2}$，$M=\left\lfloor\dfrac{1}{2}\log_e\dfrac{T}{e}\right\rfloor$，$n_0=\left\lceil\dfrac{\log_e(\psi T\varepsilon_0^2)}{2\varepsilon_0}\right\rceil$，$N_0=|A(s)|n_0$，$t=0$。

**步骤 2**：当 $t\leqslant T$ 时，对当前的状态 $s\in S^+$，遵循策略 $\pi(a|s)$ 选择动作，并更新该状态的动作空间 $A(s)_m$。

**步骤 3**：令 $t=t+1$，按式(19)～(23)更新剔除参数。

**步骤 4**：根据式(24)更新 Q 表。

**步骤 5**：重复步骤 2～4 直到 Q 表收敛或 $s$ 到达最终状态。

## 2.2 基于改进 Q 学习的智能干扰决策算法

结合本文的对抗场景与决策模型，提出基于上文改进 Q 学习的智能干扰决策算法，算法将指引多个干扰机的智能引擎作为智能体，见图 5。

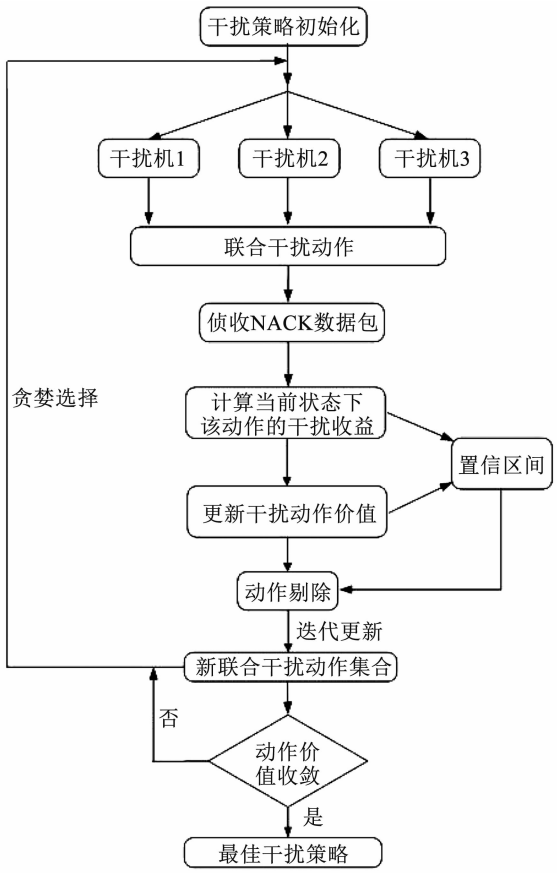


**图 5 干扰决策算法框图**

Fig. 5 Block diagram of jamming decision algorithm

详细算法如下：

**算法 2** 基于改进 Q 学习的干扰决策算法

**步骤 1**：设定初始值同算法 1，初始化干扰动作集 $A_0$。

**步骤 2**：在每个状态 $s_t=\{(f_t^{c1},f_t^{j1},\tau_1),(f_t^{c2},f_t^{j2},\tau_2),\cdots,(f_t^{ci},f_t^{ji},\tau_i)\}$，$i=2,3$；干扰方依据 $\pi(a|s)$ 即式(16)选择当前时隙要干扰的干扰动作 $a_t$，并进行当前状态下的干扰动作剔除，得到新动作集 $A_t$。

**步骤 3**：干扰方执行干扰动作 $a_t$，获得干扰反馈信息，若干扰成功，则通信方依据策略重新选择波道；否则，通信方不变更波道，环境进入下一个状态，$s_{t+1}=\{(f_t^{c1},f_t^{j1},\tau_1)',(f_t^{c2},f_t^{j2},\tau_2)',\cdots,(f_t^{ci},f_t^{ji},\tau_i)'\}$，$i=2,3$。

**步骤 4**：当 $|A_t|>1$ 时，重复步骤 2 和 3。

**步骤 5**：当 $|A_t|=1$ 时，遍历 $A_t$ 中所有的干扰动作直到所有时隙结束。

# 3 实验仿真与分析

本文在 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，取所有仿真实验数据的平均值作为最后的实验结果。实验 1 为单个干扰机对抗单对通信用户，通信方采用伪随机波道切换策略来躲避干扰；实验 2 为单个干扰机对抗单对通信用户，通信方初始阶段采用伪随机波道切换策略，当通信受阻率超过忍耐值 $P$ 后将变更随机种子，改变通信策略；实验 3 为多个干扰机对抗多个通信用户（多对通信收发机），当通信网内通信总受阻率超过忍耐值 $P$ 后各用户均变更各自伪随机波道切换序列的随机种子。3 个实验中都将本文算法和 UCBH-Q 算法[10]，以及经典强化学习算法 Q 学习、Soft-Q、动态 $\varepsilon$-Q 算法进行对比，并从干扰成功率和干扰总收益两个方面对实验结果进行分析研究。

实验及模型参数设置见表 1。

**表 1　实验及模型参数**

Tab. 1　Experiment and model parameters

| 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 $N$ | 20 |
| 波道带宽 $B_i$/kHz | 100 |
| 每时隙通信符号数 | 10 000 |
| 脉冲率 $q$ | 0.22 |
| 门限值 $\tau_0$ | 100 |
| 忍耐值 $P$/% | 30 |
| 信噪比 $R_{SN}$/dB | 20 |
| 干噪比 $R_{JN}$/dB | 15 |
| $\sigma^2$ | 1 |
| 迭代次数 $T$ | 10 000 |
| 探索因子 $\rho$ | [0,1] |
| 学习步长 $\alpha$ | 0.1 |
| 折扣因子 $\gamma$ | 0.9 |

首先考察在本文算法中设置的探索因子 $\rho$ 对算法寻优性能的影响。图 6 给出了 $\rho$ 值不同时，本文算法决策最佳干扰动作的平均概率对比。

算法 1 中探索因子 $\rho$ 不仅决定每个干扰动作的价值置信上界大小，同时也会相应改变价值置信区间，因此 $\rho$ 的大小一定程度上影响剔除次优动作与寻优的速度。由图 6 对比曲线可知，随着迭代次数的增加，算法寻优效果越明显，干扰成功率越接近于理想上界值。当总迭代数为设定的最大值 $10^5$ 时，算法的干扰成功率已趋近于 1，但实际对抗环境中大的交互次数条件常常伴随着巨大的成本代价（如暴露干扰机位置等）而很难具备。当探索因子 $\rho$ 的值介于 0.8 ~ 1.0 之间时，本文算法的寻优性能趋于饱和。在后续仿真实验中，设置本文算法的探索因子 $\rho$ 为 0.9，各仿真实验迭代次数均为 $T=10^4$。

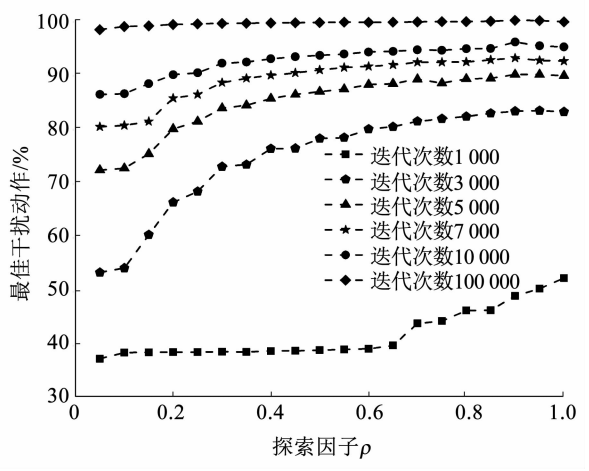


**图 6　探索因子对本文算法性能的影响**

Fig. 6　Influence of exploring factor on the performance of proposed algorithm

**实验 1**　通信方使用的固定策略为策略 1，即通过设置随机种子，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。

首先对比了干扰方直接使用有效方差置信上界等多臂赌博机（MAB）策略的干扰成功率，见图 7。

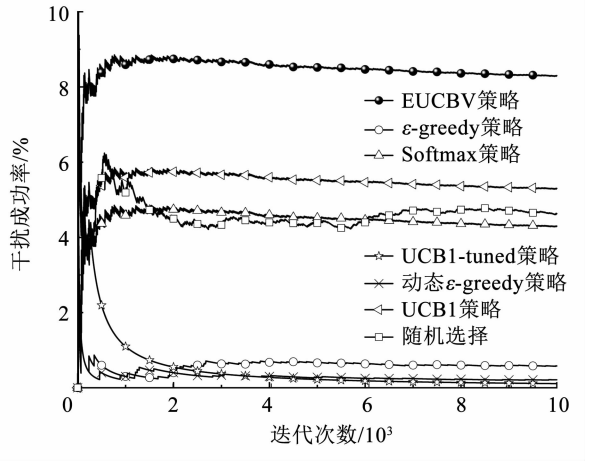


**图 7　采用不同多臂赌博机策略的干扰成功率**

Fig. 7　Jamming success rate of different MAB strategies

由图 7 可知，在 MAB 策略的干扰实验中，有效方差置信上界策略相对较优，但最高干扰成功率不超过 9%。这是由于 MAB 策略多用于收益满足平稳分布的场景，但在对抗场景下通信方受到干扰后会改变通信策略，故对于干扰方而言收益分布是非平稳的，各 MAB 策略的决策性能不佳。这同时也表明基于统计概率的干扰方式不适用于变化的对抗环境。

对抗使用伪随机策略 1 通信的单个通信用户时，各强化学习算法的干扰成功率见图 8。

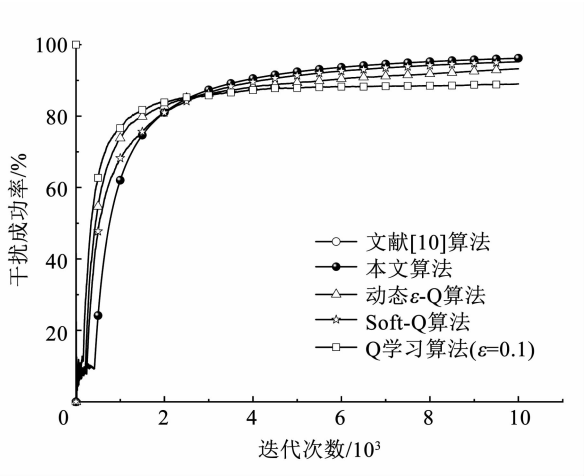


**图 8 对抗伪随机策略 1 的干扰成功率**

Fig. 8 Jamming success rate against the first pseudo-random strategy

从图 8 的干扰成功率曲线可以看到,随着迭代次数的增加,5 种算法通过与环境不断交互并学习历史经验,可使干扰成功率逐渐增加并趋于稳定。本文算法和文献[10]算法的学习轨迹大体一致,此外本文算法初始需遍历干扰动作空间因而起始阶段学习速度相对较慢,但后期可迅速收敛,在迭代次数到 3 000 时算法优势开始显现(图 8),最终干扰成功率高于其他几种算法。各算法的干扰总收益对比见图 9。

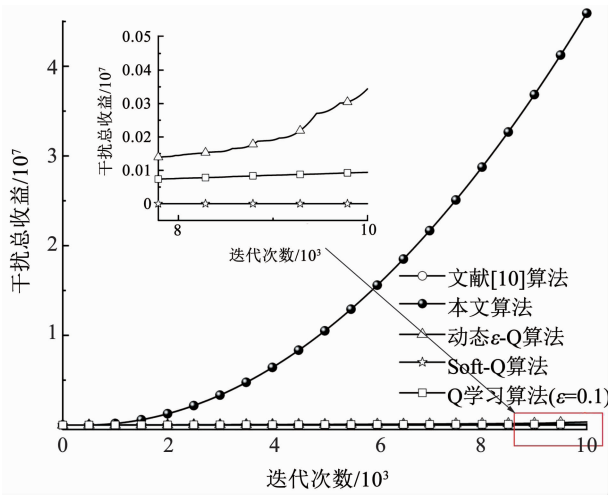


**图 9 对抗伪随机策略 1 的干扰总收益**

Fig. 9 Total jamming rewards against the first pseudo-random strategy

图 9 中,在当前设置的奖励函数条件下,与其他算法相比,本文算法和文献[10]算法所获得的干扰总收益有明显提升。

当单个通信用户使用伪随机策略 1 通信时,各算法的性能对比见表 2。

**表 2 对抗伪随机策略 1 各算法的性能对比**

Tab. 2 Performance comparison of algorithms against the first pseudo-random strategy

| 算法 | 干扰成功率/% | 干扰总收益 |
|---|---|---|
| 本文算法 | 96.2 | $4.59\times10^{7}$ |
| 文献[10]算法 | 96.0 | $4.53\times10^{7}$ |
| Soft-Q 算法 | 95.2 | $7.41\times10^{5}$ |
| 动态 ε-Q 算法 | 93.3 | $6.11\times10^{5}$ |
| Q 学习算法 | 88.6 | $1.03\times10^{5}$ |

由本文设置的干扰奖励函数可知,连续成功的干扰可获得更大的收益,即使在干扰成功率差别不大的情况下,干扰总收益仍可能存在较大的差距。例如表 2 中 Soft-Q 算法虽然干扰成功率只比前两种算法低了一个百分点,但在干扰总收益上却相差了两个数量级,原因主要在于其他算法是根据最佳干扰动作的概率分布选择干扰动作,具有一定随机性,很难实现连续稳定的成功干扰,而本文算法是依据价值高低直接选择最佳动作,消除了决策过程的偶然性。

**实验 2** 实验 2 为单个干扰机对抗使用可变通信策略的单用户,进一步考察各算法在变化环境中的学习速度。通信方使用的可变通信策略为策略 2,即初始阶段采用伪随机的波道切换策略,当通信受阻率超过规定忍耐值后,则变更随机种子,使用新波道序列进行通信。

对抗使用伪随机策略 2 通信的单通信用户时,各算法的干扰成功率见图 10。

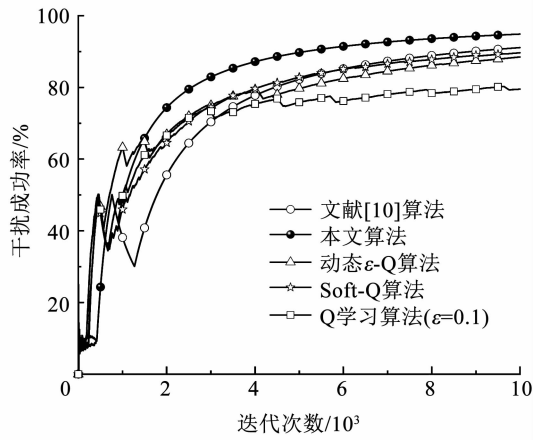


**图 10 对抗伪随机策略 2 的干扰成功率**

Fig. 10 Jamming success rate against the second pseudo-random strategy

从图 10 的干扰成功率曲线可以看出,初始阶段上述几种算法都可以实现干扰成功率的缓慢上升。当通信受阻率超过忍耐值后,由于通信方改变波道切换的策略,干扰成功率急剧下降,而后随着时间推移,干扰成功率再次上升。说明在通信方改变通信

策略后，强化学习算法仍可以学习到新的干扰策略来应对通信方策略的变化。且从曲线对比可知，本文算法的学习轨迹更稳定，波动性更小。

对抗伪随机策略 2 时，各算法获得的干扰总收益见图 11。

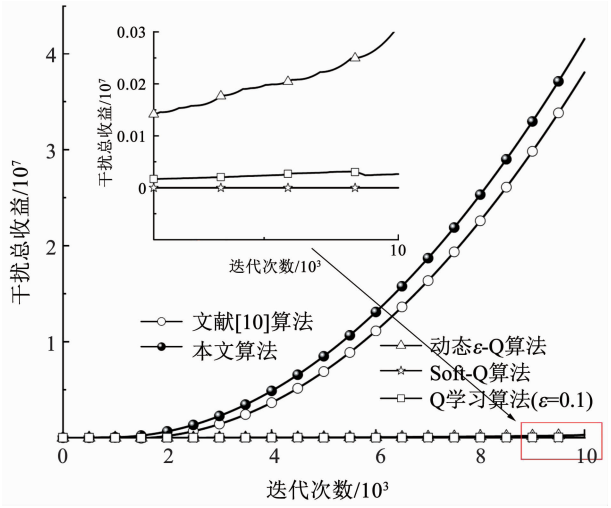


**图 11　对抗伪随机策略 2 的干扰总收益**

Fig. 11　Total jamming rewards against the second pseudo-random strategy

图 11 中，由于初始阶段各学习算法需要大量的交互经验，前期干扰获得的总收益一直处于较低水平，而后获得足够的经验样本后，干扰总收益回升。从曲线对比可看到当通信方的策略发生改变时，本文算法可以获得更高的干扰总收益，表现出比其他学习算法在变化环境中更强的学习和适应能力。表 3 给出了对抗伪随机策略 2 时各算法的性能对比。

**表 3　对抗伪随机策略 2 的各算法性能对比**

Tab. 3　Performance comparison of algorithms against the second pseudo-random strategy

| 算法 | 干扰成功率/% | 干扰总收益 |
|---|---|---|
| 本文算法 | 95.0 | $4.16\times10^{7}$ |
| 文献[10]算法 | 91.1 | $3.80\times10^{7}$ |
| Soft-Q 算法 | 88.9 | $5.04\times10^{5}$ |
| 动态 $\varepsilon$-Q 算法 | 88.6 | $4.03\times10^{5}$ |
| Q 学习算法 | 78.5 | $2.43\times10^{4}$ |

**实验 3**　构造“多对多”对抗场景，对比多个干扰机对抗使用可变通信策略的多用户时各算法性能，进一步考察算法在变化环境中的协同干扰决策能力。实验设置 3 个干扰机（或有 3 个干扰通道的单干扰机）和 3 个通信用户，每个通信用户均使用伪随机策略 2 进行通信，当某时隙所有用户都受到干扰时才视为协同干扰成功。

此时各算法的协同干扰效果见图 12。

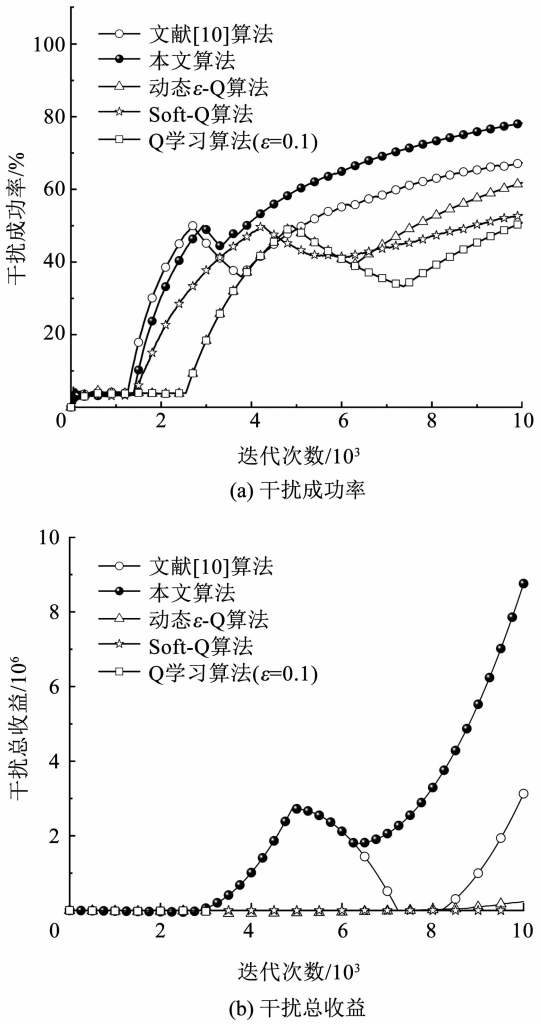


**图 12　“多对多”场景下各算法的协同对抗性能对比**

Fig. 12　Performance comparison of “many-to-many” cooperative countermeasure of different algorithms

由图 12 可知，由于协同干扰条件下环境更复杂且干扰方的联合干扰动作空间变大，加之学习过程中各通信用户的策略变化，使得决策空间维度增加，决策难度增大，当通信用户改变策略后各算法的干扰成功率和干扰总收益出现了大幅度下降。本文算法最终干扰成功率接近 80%，较传统 Q 学习算法高 50% 以上。

## 4　结　论

针对当前智能干扰决策算法收敛速度较慢、决策效率不高的问题，提出了一种基于改进 Q 学习的智能干扰决策算法。在通信干扰决策场景下，仿真验证了该算法提升决策的有效性。本文算法有较强扩展性，可将干扰动作空间延拓至多个干扰机的联合干扰动作，利用动作剔除方法降低联合干扰动作的空间维度，实现干扰资源受限情况下的协同干扰决策，实验表明该算法可实现相对较好的协同对抗

效果。本文也存在一些不足,例如:由于在初始阶段需遍历动作空间而当决策空间过大或决策空间连续时,算法学习速度较慢、效率不高。下一步考虑更充分地利用与环境交互得到的历史经验,在进一步缩减迭代次数条件下保证算法决策的可靠性。

## 参考文献


[1] SUN Hongjian, NALLANATHAN A, WANG Chengxiang, et al. Wideband spectrum sensing for cognitive radio networks: A survey [J]. IEEE Wireless Communications, 2013, 20(2): 74. DOI: 10.1049/iet-net.2020.0122

[2] 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, 16(8): 1154. DOI: 10.1109/LCOMM.2012.052112.120098

[3] AMURU S, BUEHRER R M. Optimal jamming against digital modulation [J]. IEEE Transactions on Information Forensics & Security, 2015, 10(10): 2212. DOI: 10.1109/TIFS.2015.2451081

[4] AMURU S, BUEHRER R M. Optimal jamming using delayed learning[C]//Proceedings of 2014 IEEE Military Communications Conference. Baltimore: IEEE, 2014: 1530. DOI: 10.1109/MILCOM.2014.252

[5] AMURU S, TEKIN C, van der SCHAAR M, et al. Jamming bandits—A novel learning method for optimal jamming[J]. IEEE Transactions on Wireless Communications, 2016, 15(4): 2794. DOI: 10.1109/TWC.2015.2510643

[6] 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): 65
ZHUANSUN Shaoshuai, YANG Junan, LIU Hui, et al. An algorithm for jamming decision using dual reinforcement learning[J]. Journal of Xi'an Jiaotong University, 2018, 52(2): 65. DOI: 10.7652/xjtuxb201802010

[7] 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): 37
ZHUANSUN Shaoshuai, YANG Junan, LIU Hui, et al. Jamming strategy learning based on positive reinforcement learning and orthogonal decomposition[J]. Systems Engineering and Electronics, 2018, 40(3): 37. DOI: 10.3969/j.issn.1001-506X.2018.03.05

[8] SLIMENI F, CHTOUROU Z, SCHEERS B, et al. Cooperative Q-learning based channel selection for cognitive radio networks[J]. Wireless Networks, 2019, 25(7): 4162. DOI: 10.1007/s11276-018-1737-9

[9] HAN Chen, NIU Yingtao. Multi-regional anti-jamming communication scheme based on transfer learning and Q learning[J]. KSII Transactions on Internet and Information Systems, 2019, 13(7): 3340. DOI: 10.3837/tiis.2019.07.001

[10] JIN Chi, ZHU Z A, BUBECK S, et al. Is Q-learning provably efficient? [C]// Proceedings of 32nd Conference on Neural Information Processing Systems. Montréal: [s.n.], 2018: 4867

[11] WANG Wenbo, KWASINSKI A, NIYATO D, et al. A survey on applications of model-free strategy learning in cognitive wireless networks[J]. IEEE Communications Surveys & Tutorials, 2016, 18(3): 1740. DOI: 10.1109/COMST.2016.2539923

[12] MAILLARD O A, MUNOS R, STOLTZ G. A finite-time analysis of multi-armed bandits problems with Kullback-Leibler divergences [J]. Journal of Machine Learning Research, 2011, 19(1): 18

[13] AUDIBERT J Y, MUNOS R, SZEPESVÁRI C. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits [J]. Theoretical Computer Science, 2009, 410(19): 1885. DOI: 10.1016/j.tcs.2009.01.016

[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): 58. DOI: 10.1007/s10998-010-3055-6

[15] MUKHERJEE S, NAVEEN K P, SUDARSANAM N, et al. Efficient-UCBV: An almost optimal algorithm using variance estimates[C]//Proceedings of 32nd AAAI Conference on Artificial Intelligent. San Francisco: AAAI, 2018: 6419

[16] 张春磊, 杨小牛. 认知电子战与认知电子战系统研究[J]. 中国电子科学研究院学报, 2014, 9(6): 551
ZHANG Chunlei, YANG Xiaoniu. Research on the cognitive electronic warfare and cognitive electronic warfare system [J]. Journal of CAEIT, 2014, 9(6): 551. DOI: 10.3969/j.issn.1673-5692.2014.06.001

[17] NIE Junhong, HAYKIN S. A Q-learning-based dynamic channel assignment technique for mobile communication systems[J]. IEEE Transactions on Vehicular Technology, 1999, 48(5): 1677. DOI: 10.1109/25.790549

[18] SUTTON R S, BARTO A G. Reinforcement learning: An introduction[M]. Boston: MIT Press, 1998: 25

[19] WATKINS C J C H. Learning from delayed rewards [D]. Cambridge: University of Cambridge, 1989: 27


(编辑 苗秀芝)