一种通信对抗干扰资源分配智能决策算法

许 华 宋佰霖* 蒋 磊 饶 宁 史蕴豪
(空军工程大学信息与导航学院 西安 710077)

摘 要：针对战场通信对抗智能决策问题，该文基于整体对抗思想提出一种基于自举专家轨迹分层强化学习的干扰资源分配决策算法(BHJM)，算法针对跳频干扰决策难题，按照频点分布划分干扰频段，再基于分层强化学习模型分级决策干扰频段和干扰带宽，最后利用基于自举专家轨迹的经验回放机制采样并训练优化算法，使算法能够在现有干扰资源特别是干扰资源不足的条件下，优先干扰最具威胁目标，获得最优干扰效果同时减少总的干扰带宽。仿真结果表明，算法较现有资源分配决策算法节约25%干扰站资源，减少15%干扰带宽，具有较大实用价值。

关键词：智能干扰决策；分层强化学习；干扰资源分配；专家轨迹
中图分类号：TN975 **文献标识码**：A **文章编号**：1009-5896(2021)11-3086-10
DOI: 10.11999/JEIT210115

An Intelligent Decision-making Algorithm for Communication Countermeasure Jamming Resource Allocation

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

Abstract: Considering the intelligent decision of battlefield communication countermeasure, based on the overall confrontation, a Bootstrapped expert trajectory memory replay - Hierarchical reinforcement learning - Jamming resources distribution decision - Making algorithm(BHJM) is proposed, and the algorithm for frequency hopping jamming decision problem, according to the frequency distribution, jamming spectrum is divided, based on hierarchical reinforcement learning again decision jamming spectrum and bandwidth are divided, and finally based on the bootstrapped expert trajectory memory replay mechanism, the algorithm is optimized, the algorithm can is existing resources, especially under the condition of insufficient resources, give priority to jam the most threat target, obtain the optimal jamming effect and reduce the total jamming bandwidth. The simulation results show that, compared with the existing resource allocation decision algorithms, the proposed algorithm can save 25% of the resources of jammers and 15% of the jamming bandwidth, which is of great practical value.

Key words: Intelligent interference decision; Hierarchical Reinforcement Learning(HRL); Jamming resource allocation; Expert trajectory

1 引言

在通信对抗作战过程中，干扰决策是核心环节，选择最优的干扰策略能够节省干扰资源，提高干扰成功率。一些基于博弈论[1]、遗传算法[2]等方法的干扰决策研究相继取得成果，这些研究主要从干扰样式、目标、功率等方面入手，通过建立通信方与干扰方的对抗模型，寻找最优干扰策略。此类方法在解决小规模决策问题上理论成熟，具有一定优势，但很难用于解决战场条件下多维度、大空间、小样本决策问题。

随着人工智能技术的蓬勃发展，结合人工智能技术的认知电子战相关研究取得较大进展[3]。在认知电子战系统的智能决策环节，多采用强化学习相关方法，能够为指挥员快速、准确提供辅助决策。强化学习是一种无需先验知识，智能体通过与环境交互训练，使数值化收益值最大的一种机器学习理论，广泛应用于智能决策与控制[4]、自动驾驶[5]、组合优化[6]以及资源分配[7]等领域中。基于强化学习的干扰决策方法研究近年来取得较大突破，文献[8]建立多臂赌博机干扰模型，对物理层中信号体制、功率等级等参数进行优化，以获得功率最优分配的干扰策略；文献[9]在一种延迟信息场景下，从信息状态转移中获取奖励，针对802.11机制无线网络决策最优干扰策略；文献[10]采用双层强化学习方法，能够在未知通信协议情况下以牺牲交互时间

收稿日期：2021-02-01；改回日期：2021-05-26；网络出版：2021-06-11
*通信作者： 宋佰霖 songbail@126.com

为代价学习到最佳干扰策略；文献[11]通过学习最佳干扰信号的同相分量和正交分量，得到最优干扰参数和最佳干扰样式。然而大部分基于强化学习的干扰决策方法研究是关于干扰样式、功率、物理层参数的，而几乎没有关于干扰资源分配问题的。现如今在电磁频谱作战中，频谱管控、资源分配是关键一环，最优化分配干扰资源能够在取得最好干扰效果的同时使用较少的干扰力量，并且不过多占用电磁频谱资源，保证己方通信正常进行，所以针对资源分配的干扰决策研究是至关重要的。

文献[12]提出一种分层深度强化学习抗干扰(Hierarchical Deep Reinforcement Learning anti-jamming algorithm, HDRL)频率决策算法，该算法在分层强化学习模型下分级决策通信频率，可以在干扰样式未知的条件下有效躲避干扰并减小计算量。虽然HDRL算法应用于通信抗干扰决策场景，但其分层决策结构具有较强适用性，也能够应用于干扰资源分配决策场景。

常用的抗干扰通信手段中，跳频通信应用最为广泛。本文针对在跳频干扰中干扰资源分配决策难题，提出一种基于自举专家轨迹分层强化学习的干扰资源分配决策算法(Bootstrapped expert trajectory memory replay - Hierarchical reinforcement learning - Jamming resources distribution decision - Making algorithm, BHJM)，按照侦察到的所有跳频频点分布划分子频段，分层决策干扰频段及干扰带宽，并利用本文设计的基于自举专家轨迹的经验回放 (Bootstrapped Expert Trajectory Memory Replay, BETMR)机制采样、训练算法，使算法能够在现有干扰资源条件下，按照目标干扰优先级顺序，使用尽可能小的干扰带宽实现最优干扰效果。

2　系统模型

跳频通信电台通常使用频分方式进行组网，即在全频段内选择频点规划跳频频率集，不同的频率集之间通常无相同频点。针对跳频通信常使用跟踪式干扰、拦阻式干扰等手段，随着跳频速率不断增加，在每一跳上的驻留时间越来越短，最基本的跟踪式干扰很难完成干扰任务。拦阻式干扰通过对某一频段范围内干扰信号实施压制性干扰，只要频段内包含目标频点，且干扰功率满足干信比条件，即可使干扰奏效。忽略收发天线不同带来的极化损失，干信比计算方法可用式(1)表示

$$\mathrm{JSR}=\frac{P_{\mathrm{J}}H_{\mathrm{J}}L_{\mathrm{J}}}{P_{\mathrm{S}}H_{\mathrm{S}}L_{\mathrm{S}}} \tag{1}$$

其中，P_{J}为干扰机的发射功率，P_{S}为信号发射机的发射功率；H_{J}为干扰机发射天线与信号接收天线增益之积，H_{S}为信号发射机天线增益与接收天线增益之积；L_{J}和L_{S}分别为干扰机信号和通信信号传输的空间损耗，用式(2)表示，R为信号传播距离

$$\begin{aligned}L&=\left(\frac{4\pi R}{\lambda}\right)^2=20\lg\frac{4\pi R}{\lambda}\\&=32.5+20\lg f+20\lg R\end{aligned} \tag{2}$$

将式(2)代入式(1)中，可得到干信比的一般计算表示方法，如式(3)所示

$$\mathrm{JSR}=\frac{P_{\mathrm{J}}H_{\mathrm{J}}{R_{\mathrm{S}}}^2}{P_{\mathrm{S}}H_{\mathrm{S}}{R_{\mathrm{J}}}^2} \tag{3}$$

如图1所示为一个典型的干扰场景，在一个较小区域内部署了多个地面通信干扰站，其干扰空域相同，通过侦察发现干扰空域内有多个跳频通信网。在实际中需要按照某些复杂规则来划分通信网的威胁系数，本文为简便起见仅考虑距离因素，按照每个通信网与干扰方的距离不同划分威胁系数，距离越近威胁系数越高。如表1所示，由于N_1距离干扰站最近，所以其威胁系数最高为6；而N_6距离干扰站最远，其威胁系数最小为1。干扰资源分配决策一般从通信目标的威胁系数入手，威胁系数越高，对其干扰的优先级也就越高。

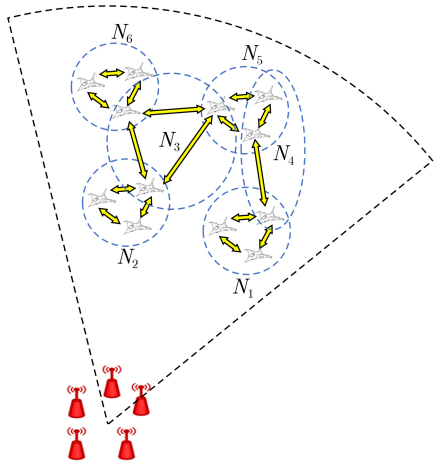


图 1 典型干扰场景

表 1　目标属性

目标	属性	威胁系数
N_1	通信网1	6
N_2	通信网2	5
N_3	通信网3	4
N_4	通信网4	3
N_5	通信网5	2
N_6	通信网6	1

假设现有通信网目标均为超短波信号，每个干扰站均采用宽带拦阻式干扰，每个频谱带宽内具有均匀相等的频谱分量，且各站干扰发射功率相同。干扰空域内共有M个通信网目标，W个干扰站；通信网内作战飞机间的信号传输距离用R_{S}表示，干扰距离用R_{J}表示。以通信网N_1为例，对其干扰的干信比可用式(4)表示，当干信比大于压制系数K_{N_1}并且干扰该目标频率集1/3以上频点时，干扰有效，通信网N_1的通信被阻断

$$\mathrm{JSR}=\frac{\sum\limits_{k=1}^{W}\dfrac{\sum\limits_{i=1}^{Z}\varphi(f_i)B_{S_i}}{B_{J_k}}P_{\mathrm{J}}H_{\mathrm{J}}{R_{\mathrm{S}}}^2}{P_{\mathrm{S}}H_{\mathrm{S}}{R_{\mathrm{J}}}^2}\geq K_{N_1} \tag{4}$$

其中，$\sum\limits_{k=1}^{W}\dfrac{\sum_{i=1}^{Z}\varphi(f_i)B_{S_i}}{B_{J_k}}$表示每个干扰站对目标$N_1$的有效干扰功率；$\sum_{i=1}^{Z}\varphi(f_i)B_{S_i}$表示目标被某个站干扰的信号带宽，$B_{S_i}$为每个跳频频点带宽，$B_{J_k}$为第$k$个干扰站的干扰带宽；$\varphi(f)$为表示干扰频段与目标频点在频率域是否对准的指示值，用式(5)表示，当频率为f的跳频频点在干扰带宽内，则指示值$\varphi(f)$为1，反之为0

$$\varphi(f)=\begin{cases}1, f\in[f_{J_1},f_{J_2}]\\0, f\notin[f_{J_1},f_{J_2}]\end{cases} \tag{5}$$

在干扰站侦收到跳频信号后，通常对其中混合的多个跳频信号进行分选。首先利用短时傅里叶变换、小波变换、谱图变换等时频分析方法分析估计跳频频率集、跳频周期等特征参数，再基于时空频信息将不同通信网的信号分开，实现对目标的精准干扰。

如图2所示为某时刻经过网台分选后跳频目标的频点分布情况，在200～400 MHz内共有6个目标，每个目标规划有一个频率集。图2中蓝色虚线方框所在频段的频点较为密集，在一个频段内有多个目标的跳频频点，并且不同目标的频点还存在交错排列的情况，此时在不同位置施放拦阻干扰带会对干扰资源分配及整体干扰效果产生不同影响。将所有目标频点合并为整体进行干扰规划，寻找包含多个不同目标的频段实施干扰，可实现对多个目标的同时干扰，进而能够降低干扰站的使用数量，减少干扰带宽，实现对干扰资源的优化分配。

3 干扰资源分配智能决策算法

3.1 基于整体对抗思想的干扰资源分配算法

针对干扰资源分配不合理、无优化算法支撑决策等问题，本文提出基于整体对抗思想的干扰资源分配算法，如表2所示，以实现在现有干扰资源下，按照干扰优先级顺序，使用尽可能小的干扰带宽实现最优干扰效果。

该算法将所有目标频点按照频率大小顺序排列，若前后两频点频率差大于拦阻干扰最大带宽$B_{\max}$，说明这两个频点不可能被同一拦阻干扰带干扰，即将两频点划入前后两个不同子频段中。按照上述方法划分频点，直至所有频点均被划入各个子频段中，图2中红色虚线方框即为划分后的子频段。

3.2 分层强化学习模型

分层强化学习的核心思想是将复杂的深度强化学习问题拆解为若干个子问题，通过解决各个子问题来最终解决整体问题。通过给不同层级的子问题分别设置奖励函数，能够有效解决复杂问题奖励稀疏、不容易收敛的难题[13,14]。

在干扰资源分配决策问题中，需要同时解决干扰频段的决策和干扰带宽的决策，直观上可以采用穷举法得到问题的最优解，然而在战场条件下，目标数量众多且频率分布复杂多变，解的数量呈指数级增长，计算量难以承受[15]。本文设计了一种基于分层强化学习的决策算法，将决策干扰频段和决策

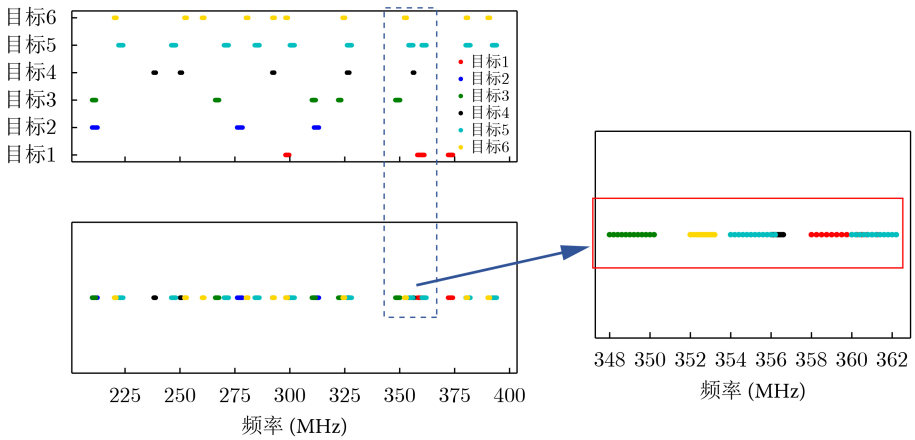


图 2 200～400 MHz频率分布

表 2　干扰资源分配算法

算法1 基于整体对抗思想的干扰资源分配算法

(1)按照威胁系数设置对目标的干扰顺序$[T_1, T_2, \cdots, T_M]$, T_1为第1目标，T_M为最末目标；

(2)按照干扰机最大干扰带宽$B_{\max}$将所有目标频点按划分为多个子频段$[J_{S_1}, J_{S_2}, \cdots, J_{S_Y}]$；

Repeat:

(3)给干扰机J_1选择干扰频段J_{S_1}；

(4)根据频段J_{S_1}设置干扰带宽B_1，找出在带宽范围内包含T_1频点数量最多的频率集，该带宽范围即为拦阻干扰带，$\boldsymbol{P}_1=[J_{S_1}, B_1]$为干扰策略$\pi$的一部分；

(5)重复步骤3和4，直至所有目标被完全阻断或干扰资源全部用完，此时共生成K个子干扰策略，干扰策略$\pi=[P_1, P_2, \cdots, P_K]$。

Break.

干扰带宽作为两个子任务来分别决策，决策网络如图3蓝色虚线方框所示。

干扰频段决策器结合环境状态S_1决策出干扰动作A_1，即干扰频段；干扰带宽决策器结合环境状态S_2和干扰动作A_1决策出干扰动作A_2，即干扰带宽。两层决策出的干扰动作组成干扰策略$P_1=[A_1, A_2]$施放干扰，改变环境状态为S'。图3所示为算法的模型结构，除各层决策器以外，模型还包括效果评估器和训练优化器部分。在效果评估器中设置奖励函数，并根据S的变化分别计算干扰动作A_1和A_2的奖励值r_1和r_2，奖励值的高低即反映了决策效果。r_1和r_2的生成无关联性，每层级决策器奖励值的设置均与当前层级解决的决策问题有关，这样可以并行训练两层决策器以提高训练效率。再由训练优化器对算法进行训练更新，在其中嵌入误差函数，通过选取一定数量包含状态S、动作A和奖励值r 3部分信息的训练样本做梯度下降计算，优化决策网络的隐藏层神经元参数，以实现对决策网络的训练更新，不断提高网络的决策水平。

3.3　基于自举专家轨迹的经验回放机制

本文设计一种基于自举专家轨迹的经验回放(Bootstrapped Expert Trajectory Memory Replay, BETMR)机制，如图4所示，在采样环节寻找专家轨迹，提高优势样本的利用率，进而提高算法的决策性能。

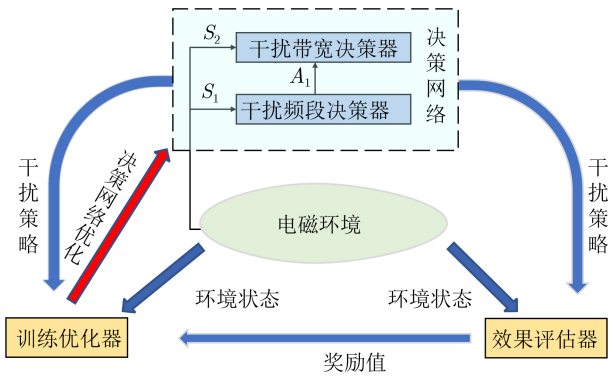


图 3 算法流程结构

通过样本训练优化决策网络是解决深度强化学习问题的核心环节，通常将每一回合的样本$\boldsymbol{e}=[S, A, r, S']$存入经验池$\boldsymbol{E}_{\text{normal}}$中，在训练神经网络时，随机抽取$\boldsymbol{E}_{\text{normal}}$中部分样本用于训练。然而在一些奖励稀疏、决策维度大的问题中，需要大量回合迭代才能采样到具有较高奖励值的优势样本，训练效率较低，使算法收敛至局部最优。

为提高算法找到全局最优策略的能力，BETMR机制将专家轨迹[16]用于算法训练中，能够“迫使”智能体学习优势样本，提高算法决策的有效性。在干扰资源分配问题中，所有的干扰目标均来自即时的通信侦察，并没有能够加以利用的专家轨迹信息，所以需要在算法训练的同时寻找专家轨迹$\boldsymbol{e}_{\text{expert}}=[S, A, r, S']$，并将其存入专家经验池$\boldsymbol{E}_{\text{expert}}$中。

本文中专家轨迹的判定标准不是一成不变的，寻找专家轨迹是一个动态的过程，手动建立或自动生成阈值集$[\delta_0, \delta_1, \cdots, \delta_H]$。假设某一回合的目标阈值是$\delta_m$，若该回合总奖励值$R>\delta_m$，则这一回合样本为专家轨迹

$$\delta=\delta_m,\begin{cases}\delta=\delta_{m+1}, & \delta_{m+1}<R<\delta_{m+2}\\ \delta=\delta_m, & R<\delta_{m+1}\end{cases}\tag{6}$$

目标阈值δ呈阶梯式变化，从δ_0开始设置，假设某一回合$\delta=\delta_m$，若$R<\delta_{m+1}$，则下一回合目标阈值$\delta=\delta_m$保持不变；若$\delta_{m+1}<R<\delta_{m+2}$，则将下一回合目标阈值设置为$\delta=\delta_{m+1}$。当$\delta$改变时，将

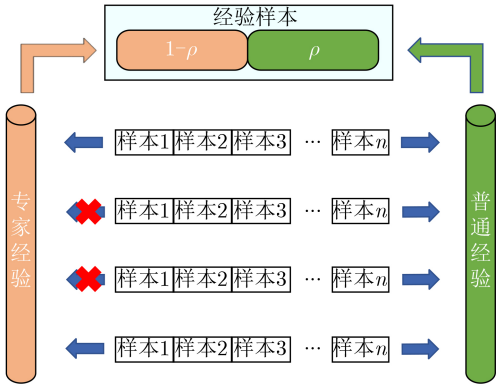


图 4 基于自举专家轨迹的经验回放机制

$\boldsymbol{E}_{\text{expert}}$重新设为空集，下一回合再次从零开始寻找更优的专家轨迹。

存储样本时，每一次决策均将样本存入$\boldsymbol{E}_{\text{normal}}$中，每一回合结束时评判当前回合样本是否满足专家轨迹条件，若满足，则将样本再存入$\boldsymbol{E}_{\text{expert}}$中。算法训练时，按照式(7)抽取样本

$$\boldsymbol{E}=\rho\boldsymbol{E}_{\text{normal}}+(1-\rho)\boldsymbol{E}_{\text{expert}} \tag{7}$$

3.4 基于BETMR的干扰资源分配决策算法

在分层强化学习框架下，结合基于整体对抗思想的干扰资源分配算法与BETMR(如表3所示)机制，提出基于自举专家轨迹分层强化学习的干扰资源分配决策算法(BHJM)，如表4所示，将算法所需基本元素定义如下：

(1) 状态空间：按照算法1步骤(2)划分子频段，按照威胁系数设置干扰目标g，分别查找各个子频段上包含g的频点个数$\boldsymbol{C}=[C_1,C_2,\cdots,C_M]$。干扰频段决策器的状态$\boldsymbol{S}_1$=[$\boldsymbol{C},g$]；干扰带宽决策器的状态$\boldsymbol{S}_2=[C_{A_1},C_{S_1},g,A_1]$，$C_{A_1}$为所选子频段内包含$g$的频点个数，$C_{S_1}$为子频段$S_1$包含所有目标的频点个数，$A_1$为干扰频段决策器的输出动作。

(2) 动作空间：两层决策器分别输出干扰动作A_1和A_2，A_1为划分子频段中的某一个，A_2用于表示干扰带宽B，$B_{\max}$为可设置带宽的最大值

$$B=B_{\max}-A_2 \tag{8}$$

(3) 奖励函数：在效果评估器中分别针对两个决策环节设置奖励函数，计算奖励值，以表征决策效果。

(a)干扰频段决策器的奖励值r_1由r_{base}和r_{appendix}组成。其中，h为成功干扰的目标，t为其威胁系数，当成功干扰某目标且为设定的目标时，获得正奖励值，否则奖励值为–2；L为所有目标的频点总数，W干扰站数量，当每回合干扰的频点总数超过均值$\dfrac{L}{3W}$时，获得正奖励值10

$$r_{\text{base}}=\begin{cases}t\times 10, & h=g\\ -2, & \text{其他}\end{cases} \tag{9}$$

$$r_{\text{appendix}}=\begin{cases}10, & l>\dfrac{L}{3W}\\ 0, & \text{其他}\end{cases} \tag{10}$$

$$r_1=r_{\text{base}}+r_{\text{appendix}} \tag{11}$$

(b)干扰带宽决策器的奖励值如式(12)表示，其中B_{tedious}表示未干扰到频点的冗余带宽

$$r_2=\begin{cases}10\times B, & B_{\text{tedious}}<0.1\\ -2, & \text{其他}\end{cases} \tag{12}$$

两层决策器在决策相应干扰动作A时，使用ε-greedy策略，该策略使得智能体有1–ε的概率选择$Q(s,a)$值最大的动作，有ε的概率随机选择动作

$$\pi(a|s)=\begin{cases}\arg\max\limits_{a}Q(s,a), & 0<x\le 1-\varepsilon\\ \text{random}, & 1-\varepsilon<x<1\end{cases} \tag{13}$$

在训练优化器中，使用3.3节提出的BETMR机制选择训练样本，按照干扰不同目标得到的不同奖励值r_1来设置δ阈值集。引入动态Q网络(Deep Q Network, DQN)算法[17]框架下的训练方法，分别设置估值神经网络和目标神经网络。两个网络的结构相同，初始参数一致，估值神经网络负责计算当前状态S的估计价值$Q(S,A;\theta_n)$，引导动作A的选择；目标神经网络负责计算目标价值$Q(S',A';\theta_n^{-})$。其中，θ_n为在n回合估值神经网络的权值参数，θ_n^{-}为在n回合目标神经网络的权值参数。

定义误差函数$L(\theta)$，由式(14)表示。对参数θ_n做梯度下降计算，以更新估值神经网络。每经过一定回合数后，将估值神经网络的权值参数赋给目标神经网络，使两个网络参数相同，不必实时更新目标价值，同时减小了目标价值选取的相关性[17]

$$L(\theta)=[r+\gamma\max_{A'}Q(S',A';\theta_n^{-})-Q(S,A;\theta_n)]^2 \tag{14}$$

表4为本文提出的BHJM算法，每个决策器的神经网络均设置输入层、2个隐藏层以及输出层，干扰频段决策器网络的隐藏层神经元数量用式(15)表示，x为输入层神经元数量；干扰带宽决策器的隐藏层神经元数量为16，网络参数的更新过程可分别用式(16)、式(17)表示

$$y=\left\lfloor\frac{\left\lfloor x\times\frac{3}{2}\right\rfloor}{2}\right\rfloor\times 2 \tag{15}$$

表 3 BETMR算法

算法2 BETMR算法

(1)建立经验池$\boldsymbol{E}_{\text{normal}}$和$\boldsymbol{E}_{\text{expert}}$，初始化为空集；

(2)设置初始阈值δ，δ=δ_0；

Repeat：

(3)将样本e存入$\boldsymbol{E}_{\text{normal}}$中；

(4)当回合结束：

Break；

(5)判断该回合样本是否满足专家轨迹条件，

若满足：将样本e存入$\boldsymbol{E}_{\text{expert}}$中；

若不满足：**pass**

(6)按式(6)计算下一回合δ，若δ改变，重置$\boldsymbol{E}_{\text{expert}}$为空集；

(7)按式(7)抽取样本$\boldsymbol{E}$。

$$\theta \leftarrow \theta - [r(S_t, A_t) + \gamma \max_{A_{t+1}} Q(S_{t+1}, A_{t+1}; \theta^-) - Q(S_t, A_t; \theta)]\nabla Q(S_t, A_t; \theta) \tag{16}$$

$$\theta^- \leftarrow \theta, t = nL \ (n = 1, 2, \cdots) \tag{17}$$

4　实验与仿真

4.1　场景及参数设置

经过通信侦察获取当前干扰空域内的6个跳频目标，频率范围均在200～400 MHz内，其各类信息如表5所示。其中，根据长期情报或侦察情报，可知干扰方已知每个通信网目标的通信距离、信号发射机功率等参数，假设每个目标的压制系数均为2。各个目标的频率集分布情况如图5所示。

为确保干信比能够大于压制系数，保证在功率域满足干扰条件，设置干扰功率为30 kW；干扰带宽最小为1 MHz，最大为3 MHz，其中每隔0.2 MHz设置一个可选带宽，共有11种选择。

4.2　不同数量干扰资源的干扰效果分析

将干扰站数量设置为6～12个共7种情况进行仿真实验，分析在不同干扰资源条件下算法的干扰效果。首先对干扰带宽决策器进行6000回合的预训练，降低其对干扰频段决策器及整体决策效果的影响。

图6所示为不同数量干扰站的干扰效果，可见当干扰站数量超过9个时，决策出的干扰策略均能够将目标全部干扰，即干扰这6个目标最少需要9个干扰站。同时可以看出，算法训练中各目标被成功干扰的收敛顺序是与目标威胁系数顺序相符的，威胁系数越高的最先保证干扰。

当干扰站数量为6,7,8时，干扰资源不足，无法将所有的目标全部干扰。当干扰站数量为8时无法将目标3干扰成功，干扰站数量为7时无法干扰2和3，而都能干扰目标1，原因与目标各频率集的频点分布有关，目标1规划的频点与目标5和6的频点存在交错情况，处于同一个小区域内，所以在干扰目标5和6时能够将目标1一起干扰。当干扰站数量为6时，能够干扰目标6, 4和2，而无法干扰前面都能干扰的目标5和1，原因是目标5的频率集有10个，频点数量有128个，现有干扰资源不足，但在尝试干扰目标6和5时能够将频率集数量相对较少并且存在频点交错现象的目标4同时干扰。

表 4　BHJM算法

算法3 BHJM算法

(1)初始化干扰频段、带宽决策器，分别建立两个神经网络：权值参数为θ的估值神经网络和权值参数为θ的目标神经网络；

Repeat:

While $j < W$：

(2)目标生成器获取环境状态S_0，给出干扰目标g；

(3)干扰频段决策器获取环境状态S_1，决策干扰频段即干扰动作A_1；

(4)干扰频段决策器获取环境状态S_2，决策干扰带宽即干扰动作A_2；

(5)效果评估器计算奖励值r_1和r_2；

(6)两层决策器分别按照BETMR机制存储样本$\boldsymbol{E}_1$和$\boldsymbol{E}_2$；

(7)获取下一步环境状态S_1'和S_2'；

(8)当完成干扰任务或干扰资源用尽：

Break；

(9)当前回合结束后，两层决策器分别按式(14)更新各自的估值神经网络；

(10)每L个回合后，按式(16)、式(17)分别更新两层决策器的目标神经网络；

(11)当算法训练至最优后，循环结束。

表 5　侦察目标信息

目标	威胁系数	通信距离(km)	通信发射功率(W)	干扰距离(km)
N_1	6	20	100	30
N_2	5	20	100	50
N_3	4	50	100	70
N_4	3	50	100	90
N_5	2	20	100	110
N_6	1	20	100	130

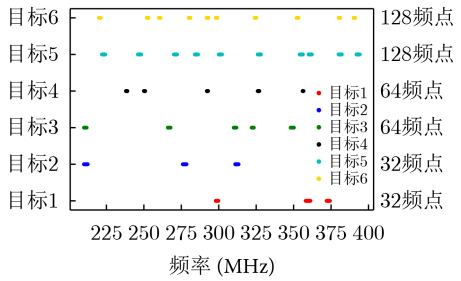
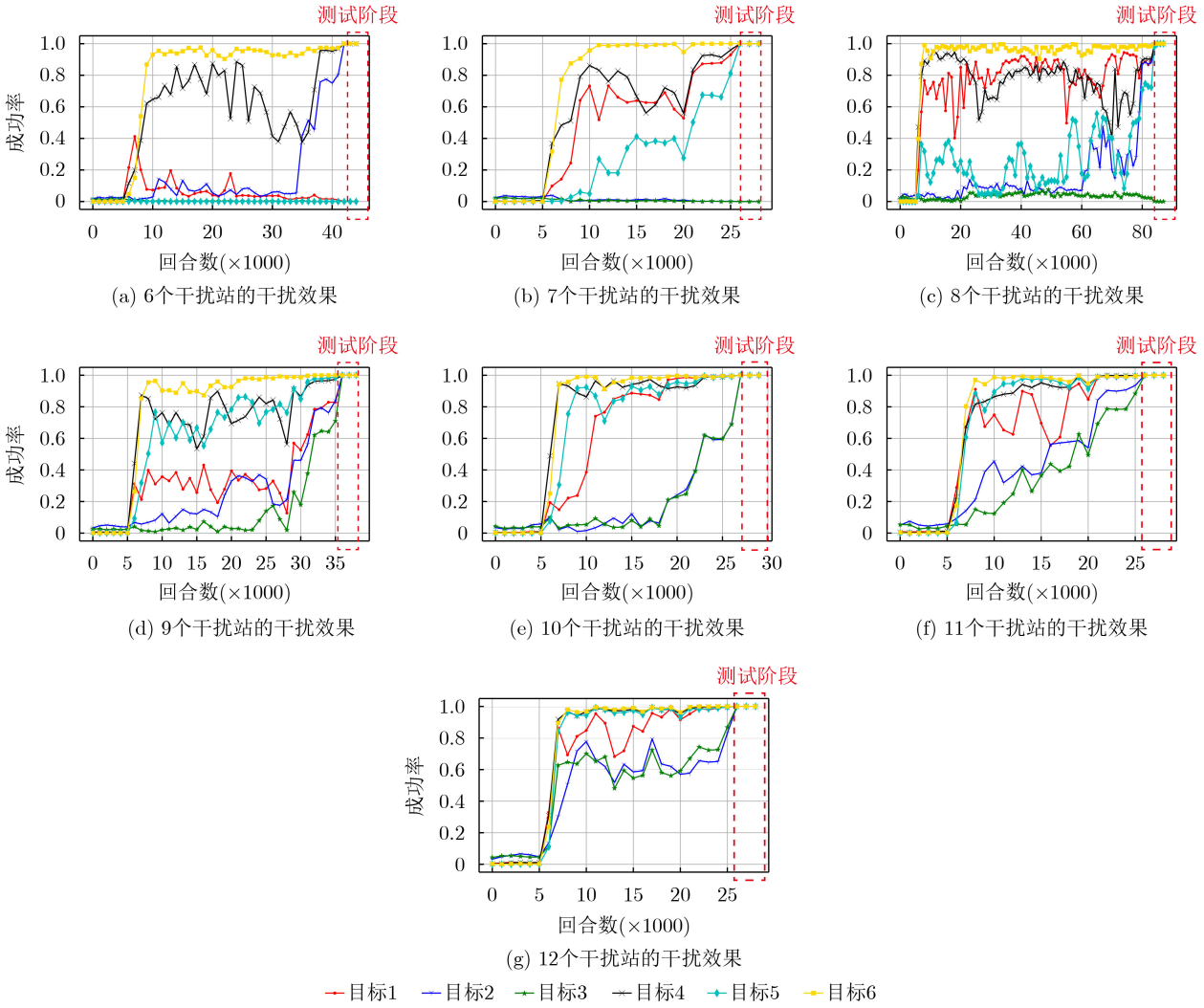


图 5 目标频率集分布情况



(a) 6个干扰站的干扰效果 (b) 7个干扰站的干扰效果 (c) 8个干扰站的干扰效果

(d) 9个干扰站的干扰效果 (e) 10个干扰站的干扰效果 (f) 11个干扰站的干扰效果

(g) 12个干扰站的干扰效果

图 6 不同数量干扰站的干扰效果

当干扰资源不足时，各目标干扰成功的收敛顺序仍然是与目标威胁系数顺序相符的，算法能够保证威胁系数越高的先被干扰。同时可以分析得出，在干扰同样目标时，干扰资源越充足，算法训练收敛更快，训练过程更稳定。

4.3 算法探索性对仿真效果的影响分析

基于强化学习的算法需要平衡探索与利用之间的关系，也就是使算法既要有一定探索性，一部分时间选择最好的动作，剩下时间随机选择动作，避免算法收敛到局部最优状态；又要把握好探索性的大小，以免算法长时间处于不收敛状态。

从图6中可以看出本实验分成了训练和测试两个阶段，当实验进入到测试阶段时，决策网络停止训练更新，同时将选择干扰动作的ε-greedy策略中ε值置为0，即每次均选择$Q(s,a)$最大值对应的动作。这样做的目的是消除决策算法的探索性，用训练好的网络来测试算法性能。

本文算法中的ε-greedy策略就是一种兼顾探索与利用的好方法，但由于实验中每一回合均有6～12次使用该策略选择干扰动作的环节，每一回合能够顺利决策出最优干扰策略的概率最多只有$(0.9)^6=0.53(\varepsilon{=}0.1)$，所以很难通过训练阶段的结果来判断算法是否已经训练收敛。为了避免长时间训练算法使模型过度训练导致过拟合，需要使算法在训练出最优策略后即停止训练。

本文设置阈值$\sigma=(0.9)^{N_J}$，当专家轨迹样本在之前1500回合内出现的概率超过σ，即可认为样本对应策略就是算法能决策出的最优策略，算法也已训练到最优状态，此时停止算法的训练更新，转入测试阶段。

分析图6各子图可以看出，算法按干扰优先级顺序决策干扰策略，探索性导致优先级较低的目标在训练阶段干扰成功率较低，但按照本文方法判定算法训练收敛转入测试阶段后，之前成功率处于上升阶段的目标均能够被成功干扰，证明了本文的算法收敛判断方式是有效的。

4.4 BHJM算法与现有算法的决策对比

本文引用文献[12]中的HDRL算法与BHJM算法对比决策效果。图7展示了两个算法的干扰效果对比情况，当有9个干扰站时BHJM算法即可干扰全部目标，而此时HDRL算法只能干扰4个目标。当干扰站数量为9个以下时，BHJM算法至少能干扰3个目标，而HDRL算法最多只能干扰3个目标。当干扰站数量为12时，HDRL算法才能够将所有目标全部干扰，此时较BHJM算法多用了3个干扰站，BHJM算法节省干扰站资源比例达到了25%。

图8展示了两个算法干扰带宽的对比情况，当干扰站数量超过10个时，BHJM算法在干扰更多目标的同时仍能够节约1 MHz以上的干扰带宽。当干扰站数量不足10个时，BHJM算法使用的干扰带宽比HDRL算法更大，但BHJM算法能干扰的目标更多，而HDRL算法虽然能够节省干扰带宽，但其无法决策出具有更好干扰效果的策略。当干扰全部目标相同时，BHJM算法能够节约4 MHz干扰带宽，比例达到15%。

以12个干扰站为例，若不使用任何智能算法，干扰全部目标所需带宽可达到3×12=36 MHz带宽，BHJM算法可减少使用12 MHz带宽，比例超过30%，能够节省大量频谱资源。

通过上述两个对比可以看出，BHJM算法能够在取得较好干扰效果的同时，还能节约大量干扰站资源及频谱资源，实现了对干扰资源的更优分配。

4.5 分层强化学习模型及BETMR机制对算法决策结果的影响分析

实验以10个干扰站为例，分别对比BHJM算法、文献[12]中HDRL算法以及文献中基于DQN的资源分配决策算法的决策效果。由于奖励值r_1能够反映出干扰目标的相关信息，而r_2只表征了干扰带宽的选择情况，不能直接反映决策效果，所以本实验使用每回合r_1的总和$\sum_{i=1}^{J} r_i$来对比3种算法的决策效果。如图9所示，3种算法每500回合取$\sum_{i=1}^{J} r_i$的平均值绘制曲线，经过30000回合实验后，对比决策效果。

从图9中可以看出，BHJM算法收敛后的平均奖励值最高，HDRL算法次之，基于DQN的算法几乎未学习到任何有用信息，算法基本不具有决策

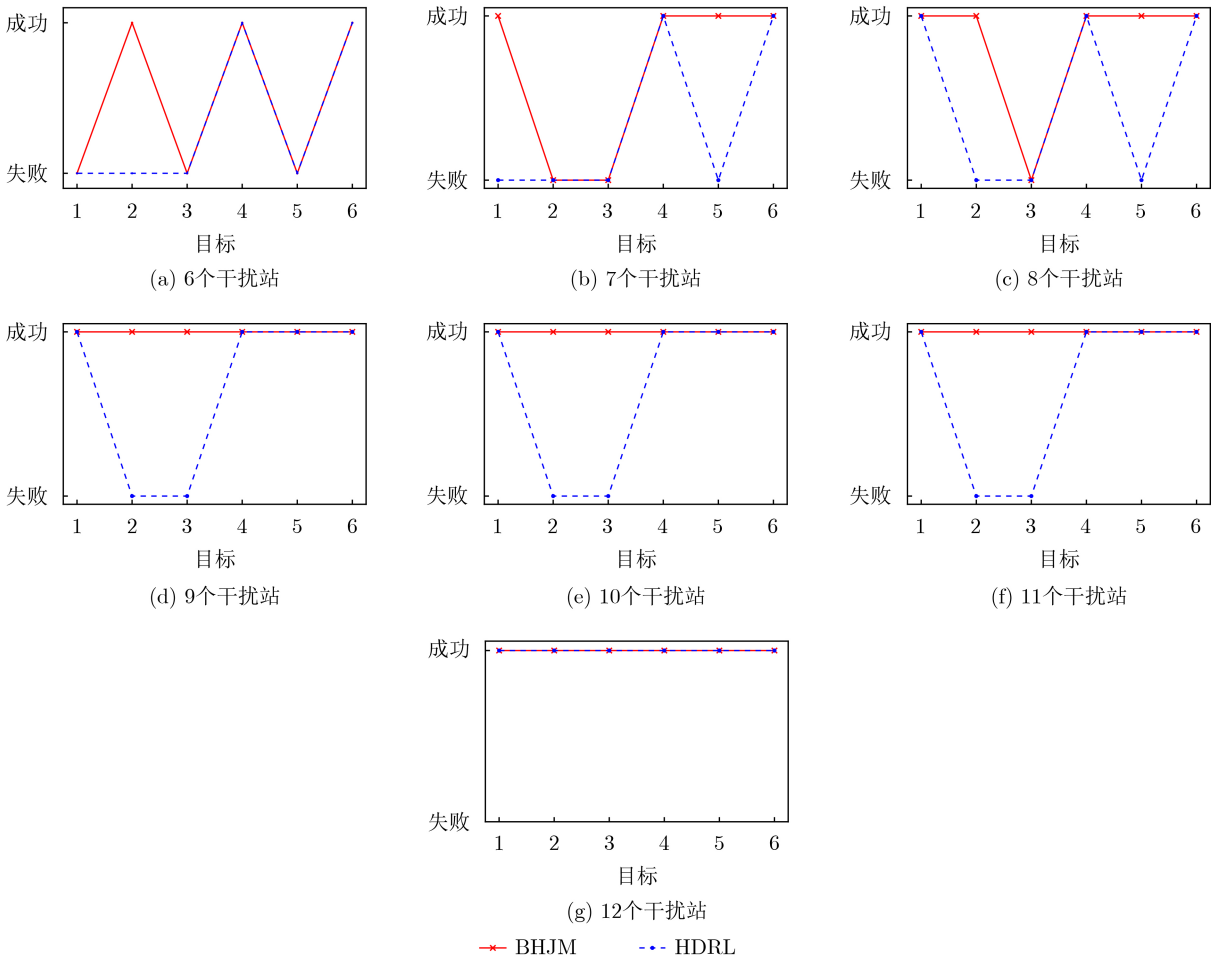


图 7 干扰效果对比

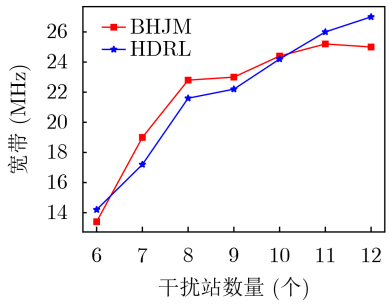


图 8 干扰带宽对比

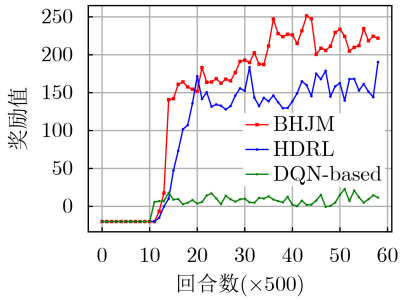


图 9 决策效果对比

能力，奖励值保持在0～25内未有明显变化。可见对于模型复杂、决策维度高的各类问题，将其拆解成各个子任务，采用分层强化学习模型就能够较好解决。而DQN等传统1维深度强化学习方法需要提前将不同的频段与不同的带宽组合成不同的干扰策略，每次决策出一个策略，但这样会使得决策空间成倍增加，算法决策效率较低，无法解决此类问题。

通过对比BHJM算法和HDRL算法的奖励值可以看出，前者的平均值相较于后者高出40%以上，具有更好的决策效果。结合上一小节干扰效果对比情况可以分析得出，在分层强化学习模型的基础上引入BETMR机制能够让算法学习专家轨迹，具有更强的决策能力。

5 结论

本文针对战场环境下跳频信号的干扰难题，提出一种干扰资源分配智能决策算法。该算法融合分层强化学习与专家轨迹等相关知识，分级决策干扰频段和干扰带宽，设计BETMR机制来采样并训练优化算法，使算法能够在现有干扰资源特别是干扰资源不足的条件下，优先干扰最具威胁目标，最优分配干扰资源，具有首创性意义。仿真结果表明，基于分层强化学习模型能够解决复杂的干扰问题，设计的BETMR机制能够使算法具有更强的决策能力，算法整体较现有资源分配决策算法节约25%干扰站资源，减少15%干扰带宽，具有较大实用价值。

参 考 文 献

[1] XIAO Liang, LIU Jinliang, LI Qiangda, *et al*. User-centric view of jamming games in cognitive radio networks[J]. *IEEE Transactions on Information Forensics and Security*, 2015, 10(12): 2578–2590. doi: 10.1109/TIFS.2015.2467593.

[2] AMURU S D and BUEHRER R M. Optimal jamming against digital modulation[J]. *IEEE Transactions on Information Forensics and Security*, 2015, 10(10): 2212–2224. doi: 10.1109/TIFS.2015.2451081.

[3] 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. doi: 10.1360/N112018-00153.

WANG Shafei, BAO Yanfei, and LI Yan. The architecture and technology of cognitive electronic warfare[J]. *Science in China*: *Information Sciences*, 2018, 48(12): 1603–1613. doi: 10.1360/N112018-00153.

[4] VINYALS O, BABUSCHKIN I, CZARNECKI W M, *et al*. Grandmaster level in StarCraft II using multi-agent reinforcement learning[J]. *Nature*, 2019, 575(7782): 350–354. doi: 10.1038/s41586-019-1724-z.

[5] QIAO Zhiqian, TYREE Z, MUDALIGE P, *et al*. Hierarchical reinforcement learning method for autonomous vehicle behavior planning[J]. arXiv preprint arXiv: 1911.03799, 2019.

[6] BELLO I, PHAM H, LE Q V, *et al*. Neural combinatorial optimization with reinforcement learning[C]. The International Conference on Learning Representations, Toulon, France, 2017.

[7] NAPARSTEK O and COHEN K. Deep multi-user reinforcement learning for distributed dynamic spectrum access[J]. *IEEE Transactions on Wireless Communications*, 2019, 18(1): 310–323. doi: 10.1109/TWC.2018.2879433.

[8] AMURU S D, 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): 2792–2808. doi: 10.1109/TWC.2015.2510643.

[9] AMURU S and BUEHRER R M. Optimal jamming using delayed learning[C]. 2014 IEEE Military Communications Conference, Baltimore, USA, 2014: 1528–1533.

[10] 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. doi: 10.7652/xjtuxb201802010.

ZHUANSUN Shaoshuai, YANG Jun'an, LIU Hui, *et al*. An algorithm for jamming decision using dual reinforcement learning[J]. *Journal of Xi'an Jiaotong University*, 2018, 52(2): 63–69. doi: 10.7652/xjtuxb201802010.

[11] 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):

518–525. doi: 10.3969/j.issn.1001-506X.2018.03.05.

ZHUANSUN Shaoshuai, YANG Jun'an, LIU Hui, *et al*. Jamming strategy learning based on positive reinforcement learning and orthogonal decomposition[J]. *Systems Engineering and Electronics*, 2018, 40(3): 518–525. doi: 10.3969/j.issn.1001-506X.2018.03.05.

[12] LI Yangyang, XU Yuhua, XU Yitao, *et al*. Dynamic spectrum anti-jamming in broadband communications: A hierarchical deep reinforcement learning approach[J]. *IEEE Wireless Communications Letters*, 2020, 9(10): 1616–1619. doi: 10.1109/LWC.2020.2999333.

[13] KULKARNI T D, NARASIMHAN K R, SAEEDI A, *et al*. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation[C]. The 30th Conference on Neural Information Processing Systems, Barcelona, Spain, 2016: 3675–3683.

[14] RAFATI J and NOELLE D C. Learning representations in model-free hierarchical reinforcement learning[C]. The AAAI Conference on Artificial Intelligence, Palo Alto, USA, 2019: 10009–10010.

[15] FESTA P. A brief introduction to exact, approximation, and heuristic algorithms for solving hard combinatorial optimization problems[C]. 2014 16th International Conference on Transparent Optical Networks, Graz, Austria, 2014: 1–20.

[16] GULCEHRE C, LE PAINE T, SHAHRIARI B, *et al*. Making efficient use of demonstrations to solve hard exploration problems[C]. The International Conference on Learning Representations, 2020.

[17] MNIH V, KAVUKCUOGLU K, SILVER D, *et al*. Human-level control through deep reinforcement learning[J]. *Nature*, 2015, 518(7540): 529–533. doi: 10.1038/nature14236.

许　华：男，1976年生，教授，博士生导师，研究方向为通信信号处理、智能通信对抗.

宋佰霖：男，1997年生，硕士生，研究方向为通信对抗智能决策.

蒋　磊：男，1974年生，副教授，研究方向为通信抗干扰、智能通信对抗.

饶　宁：男，1997年生，硕士生，研究方向为通信对抗智能决策.

史蕴豪：男，1996年生，博士生，研究方向为通信信号识别.

责任编辑：余　蓉