基于先验知识嵌入LSTM-PPO模型的智能干扰决策算法

张静克[1]，杨凯[2]，李超[1]，王洪雁[3]
（1.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 471032；2.浙江理工大学信息科学与工程学院，浙江 杭州 310018；
3.浙江理工大学计算机科学与技术学院，浙江 杭州 310018）

摘　要：针对基于传统强化学习模型的多功能雷达（MFR）干扰决策算法决策效率及有效性低、策略不稳定的问题，提出基于先验知识嵌入长短期记忆（LSTM）网络-近端策略优化（PPO）模型的智能干扰决策算法。所提算法首先将MFR干扰决策问题定义为马尔可夫决策过程（MDP）。其次，基于收益塑造理论将干扰领域先验知识嵌入PPO模型的奖励函数，利用重塑所得奖励函数引导智能体快速收敛从而提升决策效率。而后，基于LSTM优异的时序特征抽取能力，捕捉回波数据的动态特征以有效刻画雷达工作状态。最后，将所抽取动态特征输入PPO模型，经由所嵌入先验知识的引导，从而可快速获得有效干扰决策。仿真实验表明，相较于传统深度干扰决策算法，所提算法具有较高的决策效率以及有效性，且可高效稳健地达成MFR干扰决策算法。

关键词：干扰决策；多功能雷达；近端策略优化；长短期记忆网络；先验知识
中图分类号：TN92
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024270

Intelligent interference decision algorithm with prior knowledge embedded LSTM-PPO model

ZHANG Jingke[1], YANG Kai[2], LI Chao[1], WANG Hongyan[3]
1. State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, Luoyang 471032, China
2. School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China
3. School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China

Abstract: Focusing on the issues of low efficiency and effectiveness in decision-making as well as the instability of traditional reinforcement learning model-based multi-function radar (MFR) jamming decision algorithms, a prior knowledge embedded long short-term memory (LSTM) network-proximal policy optimization (PPO) model based intelligent interference decision algorithm was developed. Firstly, the MFR interference decision problem was regarded as a Markov decision process (MDP). Furthermore, by incorporating prior knowledge associated with the interference domain into the reward function of the PPO model using revenue shaping theory, a reshaped reward function was obtained to guide agent converge quickly so as to improve decision-making efficiency. Besides, leveraging LSTM's excellent temporal feature extraction ability enables capturing dynamic characteristics of echo data effectively to describe radar working states. Finally, these extracted dynamic features were inputted into the PPO model. With guidance from embedded prior knowledge, an effective interference decision can be achieved rapidly. Simulation results demonstrate that compared to traditional reinforcement learning model based interference decision algorithms, higher efficiency and effectiveness in decision-making can be attained via the proposed algorithms and the MFR interference decision can be efficiently and robustly achieved.

Keywords: interference decision, MFR, PPO, LSTM network, network prior knowledge

收稿日期：2024–07–19；修回日期：2024–12–02
通信作者：王洪雁，gglongs@163.com
基金项目：电子信息系统复杂电磁环境效应国家重点实验室基金资助项目（No.CEMEE2023K0301）
Foundation Item： The Laboratory Research Foundation of State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System (No.CEMEE2023K0301)

0 引言

多功能雷达（MFR, multifunctional radar）基于其灵活的波形变换、敏捷的波束扫描以及较强的电磁对抗特性可在现代战场上同时执行多种功能[1]，在侦察、进攻和防御等战斗任务中发挥着至关重要的作用，因此被视为电磁频谱装备体系的核心关键装备。基于此，对MFR进行有效干扰以削弱其作战效能一直是当今军事装备领域的研究热点。干扰过程主要包括电磁感知、干扰策略制定以及干扰效果评估等。干扰策略制定为该过程的关键环节，其目的在于高效制定干扰策略从而精准配置干扰样式及相应干扰资源，进而有效执行干扰任务。由以上表述可知，干扰策略制定主要包括干扰样式决策及干扰资源配置两部分，本文主要考虑干扰样式决策。

传统干扰决策算法主要依赖于“模板匹配”[2]、“博弈决策”[3]以及“推理决策”策略[4]，其需依赖于诸如雷达编码方式、功率分配等先验信息。然而MFR具有较强的自适应能力且发射波形灵活多变。由此，MFR场景下上述决策算法面临先验信息难以获取的问题，进而导致决策实时性及有效性显著降低。因此，开发一种不过分依赖于先验数据的干扰决策算法变得尤为重要。

针对此问题，一些学者开始探索将先验数据缺乏条件下可基于“试错”方式习得最佳策略的强化学习[5]（RL, reinforcement learning）技术应用于MFR干扰决策领域。文献[6]将Q-Learning算法引入MFR干扰决策，然而由于Q-Learning算法需要维护Q值表以储存各状态动作Q值，因而算法收敛较为缓慢，决策时效性较差。针对此问题，文献[7]提出基于深度Q神经网络（DQN, deep Q network）的干扰决策算法。仿真表明，相较于Q-Learning算法，基于DQN的干扰决策算法决策实时性及准确率有较大提升。然而，由于DQN模型参数更新过程易导致误差累积从而导致其陷入局部最优。基于此，文献[8]提出将基于策略梯度的异步优势行动者–评论家（A3C, asynchronous advantage actor-critic）算法引入干扰决策领域。仿真实验表明，与DQN模型相比，A3C算法可显著提升决策有效性。为进一步提升决策准确性，文献[9]将长短期记忆（LSTM, long short-term memory）网络嵌入A3C算法以处理序列数据的时间依赖性。基于迷宫环境的仿真结果表明，LSTM可协助智能体存储关于历史探索策略的记忆从而更为有效地应对新生成迷宫问题。由此表明，LSTM嵌入强化学习模型可获得序列数据更为本质的特征表达，从而显著提升决策有效性。

由于强化学习模型通常基于深度学习网络，因而具有诸如样本需求量大以及全域搜索等数据驱动模型的典型特点从而导致其学习效率较低。由此，为了提升强化学习的决策效率，须结合先验知识以引导模型寻优，从而促使其快速收敛至最优策略。为了降低样本需求量同时提升决策效率，文献[10]基于可在线训练识别模型为机器人提供先验知识从而提升控制效率。针对高维连续状态动作空间的路径规划问题，文献[11]将强化学习和逆运动先验信息相结合，以提升小样本条件下强化学习收敛速度以及路径规划效率。针对多智能体编队队形保持及协同避碰问题，文献[12]提出数据与知识联合驱动的训练模型，以基于少量训练样本提升避碰性能。由此可知，将应用相关先验知识嵌入强化学习模型可有效提升决策效率。

基于以上所述，本文提出于先验知识嵌入LSTM-近端策略优化（PPO）模型的智能干扰决策算法。所提算法首先将MFR干扰决策问题定义为马尔可夫决策过程（MDP, Markov decision process）[13]。其次，利用基于势能函数的收益塑造理论[14]，将干扰先验信息嵌入奖励函数以有效引导智能体的策略学习。而后，基于LSTM优异的时序特征抽取能力[15]，捕捉雷达数据动态特征以有效刻画其工作状态。最后，将所抽取动态特征输入PPO[16]模型，经由所嵌入先验知识的引导，从而快速获得有效干扰决策算法。仿真表明，所提算法可显著提升决策效率及有效性。

本文主要研究工作如下。

1) 基于LSTM优良的序列数据时序特征抽取能力，将其嵌入PPO模型以改善决策算法的决策有效性。

2) 利用基于势能函数的收益塑造理论，将干扰先验信息嵌入LSTM-PPO模型奖励函数以有效引导智能体学习干扰策略从而提升决策算法的决策效率。

1 强化学习理论简述

1.1 强化学习基础

强化学习中，智能体通过观察环境状态、执行

动作并接收奖励以习得最大化长期累积奖励的最优策略[17]。强化学习的核心在于智能体可通过尝试不同动作以探索环境，并基于反馈信号调整策略[18]，此方式使得智能体可高效自适应环境。

强化学习过程通常采用MDP描述，由四元组$\{S,A,P,R\}$定义。其中，S为状态$s\in S$的集合，A是动作$a\in A$集合，P为由当前状态s_t采取动作a_t至下一状态s_{t+1}的转移概率$P(s_{t+1}|s_t,a_t)$，R为状态s_t、动作a_t及下一状态s_{t+1}给定条件下所得收益$R(s_t|a_t,s_{t+1})$集合。

强化学习主要采用2类算法习得最优策略：基于值函数[19]以及策略梯度的算法[20]。基于值函数的算法优化状态−动作对应值函数以选择最大长期奖励期望对应动作。常见基于值函数的算法包括Q-Learning和DQN等。基于策略梯度的算法则优化策略函数从而使得智能体选择可最大化长期奖励的动作。常见基于策略梯度的算法包括信赖域策略优化（TRPO, trust region policy optimization）[21]和PPO[22]等。

1.2　基于值函数的强化学习

如上所述，基于值函数的算法通过学习状态−动作的值函数以选择最大化长期奖励期望对应的动作。此类算法通常基于贝尔曼（Bellman）方程及其变体更新值函数估计以逐步逼近最优值函数[23]。

1.2.1　Q-Learning算法

Q-Learning算法旨在最大化值函数$V^{\pi}(s_t)$，即：状态$s_t\in S$下智能体基于策略$\pi(a_t|s_t)$选择动作a_t后所得累计奖励期望值，其可由如下Bellman方程表示[24]。

$$V^{\pi}(s_t)=\pi(a_t|s_t)\sum_{s_{t+1}\in S}P(s_{t+1}|s_t,a_t)\left[R_t+\gamma V^{\pi}(s_{t+1})\right] \tag{1}$$

其中，P为智能体由状态s_t到s_{t+1}的转移概率，R_t为状态转移之后所得奖励值，γ为折扣因子，$V^{\pi}(s_{t+1})$为状态s_{t+1}的值函数。

值函数$V^{\pi}(s_t)$的求解可视为动态规划迭代过程。各策略可对应多个具体动作，其可由如下状态−动作值函数$Q(s_t,a_t)$表征。

$$Q(s_t,a_t)=R(s_t,a_t)+\gamma\sum_{s_{t+1}\in S}P(s_{t+1}|s_t,a_t)\sum_{a_{t+1}\in A}\pi(a_{t+1}|s_{t+1})Q(s_{t+1},a_{t+1}) \tag{2}$$

其中，$R(s_t,a_t)$为状态s_t下a_t所得奖励，$Q(s_{t+1},a_{t+1})$为状态s_{t+1}下a_{t+1}所得状态−动作值函数。

由上述可知，$Q(s_t,a_t)$可视为$V^{\pi}(s_t)$上界。由此，值函数优化可转化为寻求最大化$Q(s_t,a_t)$的a_t，即

$$\pi^*(s)=\arg\max_{s_t\in S,a_t\in A}Q(s_t,a_t) \tag{3}$$

其中，$\pi^*(s)$表示最优策略。

1.2.2　DQN

DQN是Q-Learning算法与深度学习融合的产物[25]。与Q-Learning算法通过多次与环境交互以获得最优$Q(s_t,a_t)$不同，DQN通过式(4)获得目标Q值。

$$y=r+\gamma\max_{a_{t+1}}Q(s_{t+1},a_{t+1};\boldsymbol{\theta}) \tag{4}$$

其中，r为奖励值，γ为折扣因子，θ为目标网络参数，$Q(s_{t+1},a_{t+1};\theta)$为目标网络，$\max_{a_{t+1}}Q(s_{t+1},a_{t+1};\theta)$为目标最大$Q$值。

基于均方误差（MSE, mean squared error）构造如下关于$Q(s_t,a_t)$及y的损失函数$L(\boldsymbol{\theta})$[26]，并训练深度模型从而使得所估计Q值逼近真实Q值，最终获得最优策略。

$$L(\boldsymbol{\theta})=\mathbb{E}[(y-Q(s_t,a_t))^2] \tag{5}$$

其中，$\mathbb{E}$表示期望。

1.3　基于策略梯度的强化学习

基于策略梯度的算法通常基于策略梯度更新策略参数以逐步提升策略性能。

1.3.1　信赖域策略优化

TRPO旨在解决策略优化过程中策略不稳定和样本效率较低的问题。TRPO通过限制策略更新尺度以确保每次更新皆限于可信区域内，从而提升其稳定性及收敛速度。

TRPO的核心思想是当前策略条件下最大化策略函数，即

$$\max_{\theta}\mathbb{E}_{s'\sim p_{\theta_{\text{old}}},a\sim\pi_{\theta_{\text{old}}}}\left[\frac{\pi_{\theta}(a|s)}{\pi_{\theta_{\text{old}}}(a|s')}\nabla_{\theta}\log\pi_{\theta}(a|s)\right] \tag{6}$$

其中，θ是函数参数，$\pi_{\theta}(a|s)$是状态s下选择a的概率，$\pi_{\theta_{\text{old}}}(a|s')$是旧状态$s'$下动作选择概率，$\frac{\pi_{\theta}(a|s)}{\pi_{\theta_{\text{old}}}(a|s')}$为优势函数，$\nabla_{\theta}\log\pi_{\theta}(a|s)$为策略函数

梯度。

随后引入“置信域”约束条件以限制策略更新尺度。具体地，基于库尔贝克-莱布勒（KL, Kullback-Leibler）散度度量新旧策略之间差异[27]，并将其限制于可信区域内。

$$\mathbb{E}_{s'\sim p_{\theta_{\text{old}}}}[D_{\text{KL}}(\pi_{\theta_{\text{old}}}(\cdot|s')\pi_\theta(a|s)\|\pi_\theta(\cdot|s))]\leqslant\delta \quad (7)$$

其中，δ为预设阈值，D为散度，用于度量2个概率分布之间的差异。

1.3.2　近端策略优化

相较于TRPO，PPO简化了优化目标和策略更新方法，其可表示为

$$\max_\theta \mathbb{E}_{s'\sim p_{\theta_{\text{old}}},a\sim\pi_{\theta_{\text{old}}}}\left[\hat{A}(s,a)\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s')}\nabla_\theta\log\pi_\theta(a|s)\right] \quad (8)$$

其中，$\hat{A}(s,a)$为优势函数估计值。

通过添加“clipping”操作将策略更新尺度限定于可接受范围内。

$$L^{\text{CLIP}}(\theta)=\mathbb{E}_{s'\sim p_{\theta_{\text{old}}},a\sim\pi_{\theta_{\text{old}}}}\left[\min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s')}\hat{A}(s,a),\right.\right.$$
$$\left.\left.\text{clip}\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s')},1-\varepsilon,1+\varepsilon\right)\hat{A}(s,a)\right)\right] \quad (9)$$

其中，ε为裁剪系数，用于限制策略更新尺度。

随后可基于随机梯度下降（SGD, stochastic gradient descent）算法优化上述函数[28]。

综上所述，相对于Q-learning、DQN和TRPO，PPO的优势在于直接通过最大化策略梯度函数更新策略，无需学习值函数或求解复杂约束优化问题，因而较易理解、实现和调优。此外，PPO通过限制策略更新尺度及引入“clipping”操作以提升算法稳定性，因而具有较好的并行性和扩展性，进而适用于大规模策略寻优问题。基于以上所述，本文基于PPO构建MFR智能干扰决策。

2　MFR智能干扰决策建模

将干扰决策过程表述为MDP是实现MFR干扰决策的首要条件。MDP将MFR状态、复杂电磁环境以及干扰动作描述为包含如下元素的数学模型：MFR有限工作状态集$S(s\in S)$，干扰机动作集合$A(a\in A)$，依赖于状态转移的奖励集合R，收益函数$R(s_t|a_t,s_{t+1})$，由状态转移概率$P(s_{t+1}|s_t,a_t)$刻画的环境模型。干扰机发射干扰样式迫使雷达状态发生转移从而获得相应干扰收益，其尝试迭代以期最大化累计期望奖励从而获得最优干扰策略最终实现干扰目标。下面详细说明MFR智能干扰决策问题所涉及的状态集、动作集、状态转移概率以及奖励函数。

2.1　状态集

电子侦查相关研究将MFR信号构建为具有如下3层结构的层级模型：功能层（雷达句）、任务层（雷达短语）、波形层（雷达字）[29]。相较于雷达句，处于中间层的雷达短语具有较为丰富的信号信息，相较于雷达字，其与MFR工作方式相关的控制参数具有直接映射关系[30]。具体地，MFR信号层级模型如图1所示。雷达字由固定排列且数量有限的脉冲$\boldsymbol{p}_w=(p_1,p_2,\cdots,p_{D_p})^{\text{T}}\in R^{D_p}$组成[31]，其

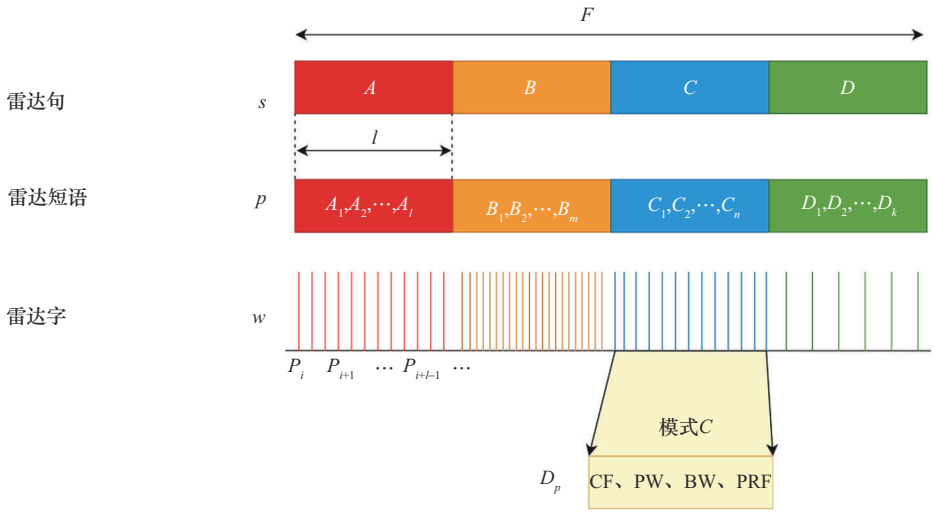


图1　MFR信号层级模型

中D_p表示雷达字脉冲参数维度。脉冲参数可由载波频率（CF）、脉冲宽度（PW）、频带宽度（BW）和脉冲重复频率（PRF）等组合表征，即$\{f_t,\mathrm{pf}_t,\mathrm{pw}_t,\mathrm{bw}_t\}$；将$l$个雷达字排列构成短语$\boldsymbol{p}_p \in R^{D_p \times l}$，其中$\boldsymbol{p}_{i,l}=(p_i,p_{i+1},\cdots,p_{i+l-1})^{\mathrm{T}}$；将$F$个短语排列成雷达句$\boldsymbol{p}_s \in R^{D_p \times F}$，其与雷达功能密切相关。

综上所述，雷达短语可以直接反映雷达工作状态，且其可与干扰样式构建较为直接的映射关系，因此可将雷达短语视为雷达工作状态，本文所构建的MDP模型的状态集即为MFR的雷达短语集。

由于MFR波形灵活多变，难以基于特定波形设定雷达威胁等级。针对此问题，基于文献[7]所提方法，本文将雷达工作状态记为$s_i^{\omega},i=0,1,2,\cdots,N_s$，威胁程度从0到$N_s$依次下降，$N_s$为工作状态数目，本文设定为11，$\omega$为雷达波形单元，$\omega=A$表示仅有一种波形，$\omega=A,B$表示存在2种波形，以此类推。

2.2　动作集

动作空间是指智能体在各时间步可选择的动作集合。可采取的行动将影响环境并产生奖励信号。MFR认知干扰情况下，干扰模式库中干扰样式可视为可选择动作，因此动作集可表示为$a_t=\{\mathrm{jam}_0,\mathrm{jam}_1,\cdots\mathrm{jam}_I\}$，$I$为动作数目，$\mathrm{jam}_i$表示$i$类干扰。

2.3　状态转移概率

状态转移概率描述了MFR工作状态之间的迁移模式，其取决于MFR信号产生机制。MFR信号生成过程中，雷达任务调度和环境适应机制密切相连，其确保了MFR状态迁移的马尔可夫性质。这种马尔可夫性质可简化表示为$P(s_{t+1}|s_t,a_t)$。

2.4　奖励函数

强化学习中，奖励函数直接影响智能体在环境中学习和执行决策的效果。智能体可基于奖励函数知晓特定状态下所执行动作的收益，从而逐步提升其决策有效性，因此设置合适的奖励函数是确保智能体可高效习得最优策略的关键。基于此，奖励函数应为非稀疏以保证奖励信号连续分布于状态动作空间从而确保学习过程中智能体可获得足够反馈，进而快速收敛至最优策略。由此，本文将奖励函数分解为状态改变奖励R_1以及状态参数改变奖励R_2两部分从而使得其可较好地适应环境变化。

$$R_1=\begin{cases}\Delta(s_{t+1}-s_t),s_t\to s_{t+1}\\ +100,s_t\to s_{\mathrm{end}}\end{cases} \tag{10}$$

$$R_2=\begin{cases}-1,L_{s_{t+1}}\geqslant L_{s_t}\\ 1,L_{s_{t+1}}<L_{s_t}\end{cases},L_{s_t}=\left\|\frac{s_t}{\|s_t\|}\right\|_2 \tag{11}$$

其中，$\Delta(s_{t+1}-s_t)$为雷达受到干扰影响迁移至新的雷达状态后威胁程度变化量，+100表示雷达状态已迁移至最优目标状态，s_{end}为威胁程度最低的目标雷达状态，R_2为受到干扰影响的状态参数与干扰前参数的威胁程度变化量，如果威胁程度增大或不变则惩罚1反之奖励1，L_{s_t}表示状态参数归一化所得范数。

3　所提智能干扰决策模型构建

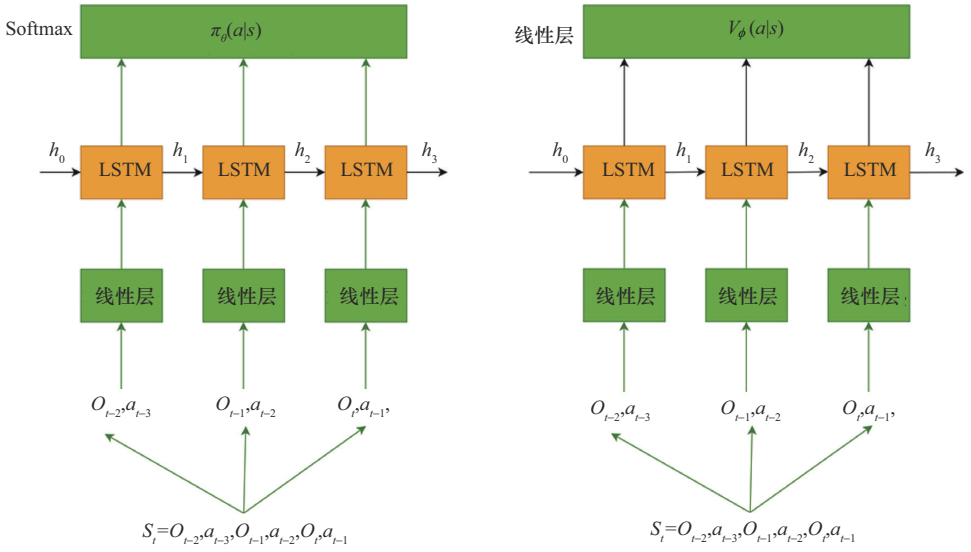
基于以上分析，本文提出基于干扰领域先验知识嵌入LSTM-PPO模型的智能干扰决策算法，以提升复杂电磁环境下干扰决策效率及有效性。通过直接最大化策略梯度函数更新策略、限制策略更新尺度及“clipping”操作，PPO具有出色的策略收敛速度与稳健性。基于LSTM强大的时序特征提取能力，所提算法将LSTM嵌入PPO模型以有效抽取回波序列数据时序本质特征，从而增强干扰决策模型的数据利用效能，进而使得干扰决策模型能够精准捕捉雷达工作状态动态变化，最终提升干扰决策算法的决策有效性。此外，基于收益塑造理论，所提算法将干扰领域先验知识嵌入LSTM-PPO模型的奖励函数，利用所得重塑奖励函数引导智能体快速收敛至最优策略路径，从而改善干扰决策算法的决策效率。下文将详述LSTM嵌入PPO模型以及干扰领域先验知识嵌入LSTM-PPO模型的具体步骤。

3.1　LSTM嵌入PPO模型

所构造LSTM嵌入PPO模型如图2所示。嵌入模型首先基于全连接网络对侦察所得序列数据预处理，网络权重经由式(12)正交初始化以提升网络训练稳定性，从而避免梯度消失或爆炸的问题。

$$\boldsymbol{z}_t=\boldsymbol{W}_0\boldsymbol{s}_t+\boldsymbol{b}_0 \tag{12}$$

其中，$\boldsymbol{W}_0$为权重矩阵，其可正交初始化为$\boldsymbol{W}_0=\boldsymbol{Q}_0\boldsymbol{D}_0$，$\boldsymbol{Q}_0$为正交矩阵，$\boldsymbol{D}_0$为对角矩阵，$\boldsymbol{b}_0$为偏置



(a) LSTM代理的策略网络　　(b) LSTM代理的价值网络

图2　LSTM嵌入PPO模型

向量，$\boldsymbol{s}_t$为状态向量。

而后基于非线性函数σ激活z_t，可得

$$\boldsymbol{h}_t=\sigma(z_t)=\sigma(\boldsymbol{W}_0\boldsymbol{s}_t+\boldsymbol{b}_0) \tag{13}$$

激活所得$\boldsymbol{h}_t$可作为LSTM输入，经由LSTM处理可获得如下隐藏状态。

$$\boldsymbol{h}_t'=\mathrm{LSTM}(\boldsymbol{h}_{t-1}',\boldsymbol{h}_t) \tag{14}$$

其中，$\boldsymbol{h}_t$为t时刻LSTM的隐藏状态，$\boldsymbol{h}_{t-1}'$为前一时间步隐藏状态。LSTM隐藏状态允许构成PPO框架的2个核心网络，即价值（Critic）网络和策略（Actor）网络在不同时刻之间共享信息，从而使得所构造模型能够更好地感知环境动态变化。

随后，LSTM输出特征向量被送入后续皆由全连接层构成的Actor以及Critic网络，从而为二者分别提供策略选择的有效信息以及状态长期价值估计的依据。将所得隐藏状态输入上述网络可得

$$V(s_t)=\boldsymbol{W}_v^{\mathrm{T}}\boldsymbol{h}_t'+\boldsymbol{b}_v \tag{15}$$

$$\pi(a_t|s_t)=\mathrm{Softmax}(\boldsymbol{W}_a^{\mathrm{T}}\boldsymbol{h}_t'+\boldsymbol{b}_a) \tag{16}$$

其中，$V(s_t)$为Critic网络所得状态s_t长期价值，$\boldsymbol{W}_v^{\mathrm{T}}$和$\boldsymbol{b}_v$分别为Critic网络的权重向量和偏置，$\pi(a_t|s_t)$则为Actor网络所得策略，$\boldsymbol{W}_a^{\mathrm{T}}$和$\boldsymbol{b}_a$分别是Actor网络的权重向量和偏置。

所构造LSTM嵌入PPO模型中，Actor及Critic网络共享基于LSTM所提供的时序特征。然而，两者拥有不同的输出层和损失函数，因为Actor网络生成动作的概率分布，以此选择最优动作，而Critic网络的目的则是预测状态的价值函数。由此，Actor网络损失函数由式(9)表征，Critic网络损失函数可表示为

$$L^{\mathrm{GAE}}(\boldsymbol{\theta}_v)=\frac{1}{N}\sum_{i=1}^{N}[\mathrm{GAE}_t(\gamma,\lambda)-A_{\theta_v}(s_t,a_t)]^2 \tag{17}$$

其中，$\boldsymbol{\theta}_v$为Critic网络参数，$A_{\theta_v}(s_t,a_t)$为Critic网络输出的估计优势值，N为样本数量,λ为折扣因子。

所提LSTM-PPO模型的智能干扰决策算法总体框架如图3所示，其基于当前策略$\pi(a|s)$与环境交互以获得动作a、奖励r和下一个状态s'，重复迭代直至经验回放池达到预设容量。此阶段策略网络权重参数保持不变以积累训练数据。随后，所得经验回放池中数据被用于Actor网络策略更新。数据首先通过LSTM以捕捉序列数据时序特征。基于此时序特征，Actor网络输出新策略$\pi_\theta(a|s)$。通过比较新旧策略，可计算重要性采样比例（ratio）以量化新旧策略相对变化，而后基于ratio通过“clip”操作构建损失函数$L^{\mathrm{clip}}(\theta)$以确保策略更新受限于可接受范围从而保持策略优化稳定性。最后，基于梯度下降算法更新旧策略网络参数直至策略性能达至最优。类似地，Critic网络利用GAE结合所得时序特征获得相邻时间步的值函数估计值$V_\phi(a|s)$，基于MSE构建损失函数。基于梯度下降算法优化Critic网络参数以获得精确的值函数估计。

综上所述，Actor及Critic网络共享特征提取层但拥有独立的输出层和损失函数。此协同机制可有

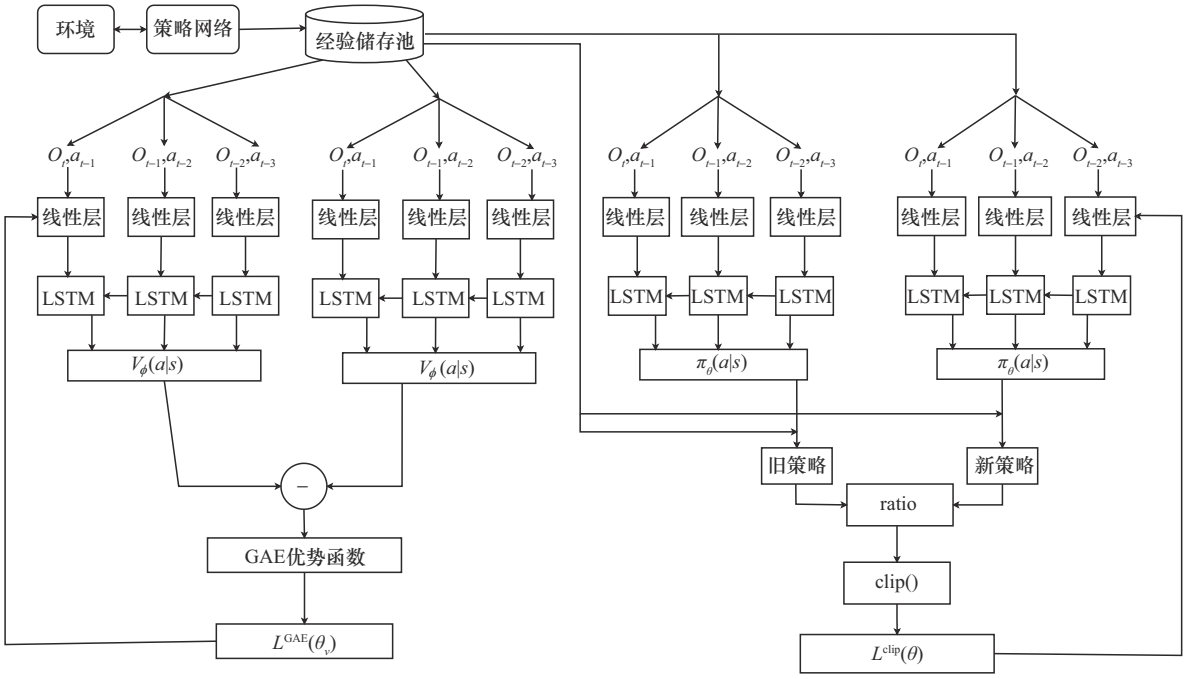


图3　LSTM-PPO模型的智能干扰决策算法总体框架

效提升智能体对环境的适应性和理解能力，优化动作决策过程，从而增强模型对环境变化的适应性，改善状态价值估计精确度，进而提升干扰决策的有效性。

3.2　基于势能函数的奖励塑造理论

基于势能函数的奖励塑造理论提供了一种将领域先验信息嵌入智能体决策过程的有效手段[14]。该理论可重塑智能体所得奖励同时不改变最优干扰策略。势能函数借鉴了物理学的势能概念，通过赋予智能体状态迁移物理意义上的“势能差”引导智能体动作。具体而言，当智能体从高势能状态（不符合领域知识的状态）迁移至低势能状态（符合领域知识的状态）时，智能体将获得额外奖励，反之则失去部分奖励。

关于具有塑造功能的函数项的嵌入不改变所得最优策略的论述可证明如下。旧MDP问题记作M，新MDP问题记作M'。由于强化学习策略是由状态集至动作集的映射，而状态-价值函数是策略的公式化，因此映射过程中最优状态价值函数和最优策略并非单一映射关系，即最优状态-价值函数有且只有一个，然而最优策略可具有多种表现形式。根据Bellman方程可知，最优状态-价值函数可表示为

$$Q_M^*=\mathbb{E}_{s'\sim P_{sa}(\cdot)}[R(s,a,s')+\gamma\max_{a'\in A}Q_M^*(s',a')] \quad (18)$$

其中，$P_{sa}(\cdot)$为状态s和动作a下的转移概率分布。

基于势能函数定义可得[14]

$$Q_M^*(s,a)-\phi(s)=\mathbb{E}_{s'}[R(s,a,s')+\gamma\phi(s')-\phi(s)+\gamma\max_{a'\in A}(Q_M^*(s',a')-\phi(s'))] \quad (19)$$

其中，ϕ为势能函数，用于衡量状态的相对优势。

基于势能函数差分形式$F(s,a,s')=\gamma\phi(s')-\phi(s)$可得

$$\hat{Q}_{M'}(s,a)\triangleq Q_M^*(s,a)-\phi(s)=\mathbb{E}_{s'}[R(s,a,s')+\gamma\max_{a'\in A}(\hat{Q}_M^*(s',a')] \quad (20)$$

由此可得，当M'到达最优策略时，M'的动作-价值函数$\hat{Q}_{M'}(s,a)$满足如下条件。

$$\pi_{M'}^*(s)\in\arg\max_{a\in A}Q_{M'}^*(s,a)=\arg\max_{a\in A}Q_{M'}^*(s,a)-\phi(s)=\arg\max_{a\in A}Q_M^*(s,a) \quad (21)$$

由此可得，M'最优策略与M相同，表明势能函数仅与状态相关，对同一状态下动作选择没有影响，因此不改变最优策略。受此启发，将符合领域先验信息的状态基于势能函数奖励重塑，所得重塑奖励不仅可提升智能体对先验信息的敏感性，而且加速了策略学习过程，从而使得智能体可较快收敛至最优策略，进而提升干扰决策的决策效率。

3.3　先验知识嵌入LSTM-PPO模型

基于PPO的MFR干扰智能决策场景中，智能体干扰决策依赖于优势函数[32]，而优势函数基于

动作−价值函数构建。由于智能体的试错特性，模型难以精确确定雷达状态对应的最优动作，由此限制了实际应用中优势函数的效能。基于此，可将领域先验知识基于上述奖励塑造理论嵌入奖励函数，引导智能体沿着期望目标状态方向探索，从而加速智能体策略学习过程。基于文献[7]可知，假设干扰任务的目标是通过干扰机的操作将当前雷达状态s_0转移至威胁程度最低的目标雷达状态s_{end}，此过程中雷达所经历的状态集合构成了雷达先验信息s_e。由以上所述可知，基于势能函数的奖励塑造函数的构造步骤可表述如下。

1) 设计基于环境模型的势能函数

基于环境模型的势能函数可通过求解逆向Bellman方程获得。具体地，状态s_t值函数可表示为

$$V^{\pi}(s_t)=\sum_{a\in A}\pi(a_t|s_t)\sum_{s_{t+1}\in S}P(s_{t+1}|s_t,a_t)\left[R_t+\gamma V^{\pi}(s_{t+1})\right] \tag{22}$$

量化状态s_t相对于目标状态s_{end}的“势能差”，设计如下反应状态转移能量变化的势能函数$\varphi(s_t)$。

$$\varphi(s_t)=\sum_{a\in A}\pi(a_t|s_t)\sum_{s_{t+1}\in S}P(s_{t+1}|s_t,a_t)\left[-R_t+\gamma U(s_{t+1})\right] \tag{23}$$

其中，逆转奖励函数R_t以确保奖励塑造函数$U(s_t)$随着智能体接近目标状态而单调下降，折扣因子γ的引入使得势能函数随时间推移逐渐衰减。

2) 奖励塑造函数构造

基于上述可知，智能体接近目标状态时，势能函数值逐渐变小，且所构造势能函数$\varphi(s_t)$具有非负性。因此，当智能体状态符合领域先验信息时，奖励塑造函数将智能体接近目标状态时所损失势能作为正奖励，反之则作为负奖励，从而确保智能体能够沿着先验信息方向学习策略。由此可构造式(24)所示奖励塑造函数。

$$U(s_t,s_{t+1})=\begin{cases}-\varphi(s_t), s_t\in s_e\cap s_{t+1}\neq s_t\\ \varphi(s_t), s_{t+1}\in s_e\cap s_{t+1}\neq s_t\\ 0,\text{其他}\end{cases} \tag{24}$$

3) 奖励函数塑造

结合上述所构造奖励函数R_1、R_2以及奖励塑造函数$U(s_t,s_{t+1})$以获得式(25)所示新的奖励函数从而加速智能体学习进程。

$$R'_{1,2}(s_t,s_{t+1})=R_{1,2}(s_t,s_{t+1})+\alpha\Delta U(s_t,s_{t+1}) \tag{25}$$

其中，$R_{1,2}(s_t,s_{t+1})$为原始奖励函数，$\Delta U(s_t,s_{t+1})$为势能变化，α为标量权重，用于调整势能变化对奖励塑造的影响程度。由此，所得新的奖励函数不仅考虑了即时奖励，还关注了长期潜在的“势能收益”，从而使得智能体更深入理解状态之间的相对“价值”，并综合上述信息做出最优决策。

基于以上所述，所提算法基于LSTM优异的时序特征抽取能力，将LSTM嵌入PPO模型以感知环境动态变化从而提升干扰决策模型的决策有效性。基于环境模型的势能函数表征领域先验知识并以此重塑奖励函数。而后将重塑所得奖励嵌入LSTM-PPO模型以增强智能体的环境感知能力，促进其探索开发之间平衡，从而提升复杂电磁场景下智能体的决策质量和学习效率。由此可得，所提基于先验知识嵌入LSTM-PPO模型的智能干扰决策算法具体步骤如算法1所示。

算法1 基于先验知识嵌入LSTM-PPO模型的智能干扰决策算法

输入 雷达状态集S，干扰机动作集A，折扣因子γ，学习率η，探索率ε，最大迭代次数$t_{\max}$，全局训练次数$T_{\max}$，平衡系数α，奖励函数R'_1、R'_2

输出 动作策略$\pi_\theta(a|s)$以及值函数$V_\phi(a|s)$

初始化 先验知识s_e，雷达状态s_t，策略网络参数θ_π，价值网络参数θ_v，经验回放池D;

1) while $T<T_{\max}$,do

2) 重置策略网络和价值网络参数梯度$\mathrm{d}\boldsymbol{\theta}_\pi\leftarrow 0,\mathrm{d}\boldsymbol{\theta}_v\leftarrow 0$，正交初始化网络权重；

3) 基于策略网络获得干扰样式概率分布π以选取干扰样式a_t，并获取下一个雷达状态s_{t+1}；

4) $t\leftarrow t+1,T\leftarrow T+1$；

5) if $t=t_{\max}$或$s_t=s_{\text{end}}$；

6) 获取最终状态$s_t\rightarrow s_{\text{end}}$的奖励值$R'_{1,2}$；

7) else

8) 返回步骤3)；

9) for $i\in\{0,1,2,\cdots,t-1\}$:

10) 计算R:$R\leftarrow r_i+\gamma R'_{1,2}(s_t,s_{t+1})$；

11) 策略网络梯度更新

$\mathrm{d}\boldsymbol{\theta}_\pi\leftarrow\mathrm{d}\boldsymbol{\theta}_\pi+\nabla_{\theta_\pi}\ln\pi(a|s)(R-V(s;\boldsymbol{\theta}_v))+\nabla_{\theta_\pi}H(\pi(s,\boldsymbol{\theta}_\pi))$；

12) 价值网络梯度更新

$$\mathrm{d}\boldsymbol{\theta}_v \leftarrow \mathrm{d}\boldsymbol{\theta}_v + \frac{\partial (R - V(s|\boldsymbol{\theta}_v))^2}{\partial \boldsymbol{\theta}_v};$$

13) end for

14) 更新策略网络和价值网络的梯度参数 $\boldsymbol{\theta}_\pi = \boldsymbol{\theta}_\pi - \eta \mathrm{d}\boldsymbol{\theta}_\pi, \boldsymbol{\theta}_v = \boldsymbol{\theta}_v - \eta \mathrm{d}\boldsymbol{\theta}_v$

15) 返回步骤1)；

16) end if

17) end while

4　实验仿真和结果分析

本节通过与DQN、Q-Learning、TRPO、PPO等主流强化学习算法对比以验证所提算法的有效性。通过基于不同神经网络架构的PPO模型及先验信息嵌入等算法，对各算法在决策有效性、决策效率和决策路径方面的性能进行验证与分析，以全面展示所提算法在复杂电磁环境中的适用性与优越性。

所构建对抗场景包括一部智能干扰机及其干扰目标，即单部机载MFR，其中干扰机固定部署，机载MFR则于干扰机作用范围内移动。基于侦查设备，智能干扰机可实时获取MFR定位、工作状态及电磁信息，从而为干扰决策提供精准的数据支撑。其次，智能干扰机可设置多种干扰模式，包括噪声干扰、欺骗干扰和频率调制干扰等。再者，智能干扰机可根据其与MFR的交互经验动态调整干扰样式，从而实现最佳干扰效果。该对抗场景中，MFR具备4种工作状态：搜索、跟踪、成像和制导。基于表1所示的MFR状态参数和具体任务需求，本文将上述4种工作状态细分为11种具体工作模式，由此，不同作战条件下MFR可基于威胁等级灵活切换工作模式。

表1　MFR状态参数

工作状态	CF/GHz	PW/μs	BW/MHz	PRF/kHz
搜索	[1,4]	[0.5,20]	[1,10]	[0.5,5]
跟踪	[1,4]	[10,30]	[1,50]	[20,50]
成像	[2,8]	[30,50]	[50,100]	[0.5,100]
制导	[2,8]	[10,30]	[1,50]	[50,100]

4.1　先验知识嵌入LSTM-PPO模型仿真环境构建

仿真实验环境设置如下：CPU为AMD Ryzen 7 5800H with Radeon Graphics 3.20 GHz；GPU为NVIDIA GeForce RTX 3050Ti，算法基于PyTorch深度学习框架采用Python编写。所提算法参数设置如下：学习率 $\eta = 0.01$、奖励折扣因子 $\gamma = 0.99$、衰减权重 $\beta = 0.9$、GAE优势函数权衡值 $\lambda = 0.95$、初始探索率 $\varepsilon = 0.1$、终止探索率 $\varepsilon = 0.9$、PPO裁剪系数 $\delta = 0.2$、PPO熵正则化项 $H = 0.01$、先验信息平衡系数 $\alpha = 32$，经验回放池 D 总量为1 000，采用每15步或者当前网络回合结束更新网络参数。对比强化学习算法参数设置如下：TRPO共轭梯度算法优化迭代次数为15，阻尼项为0.1，控制策略更新幅度值为0.01，其余参数设置与PPO相同。DQN学习率 $\eta = 0.01$，奖励折扣因子 $\gamma = 0.99$，经验回放池 D 总量为1 000，每隔15步更新网络参数。Q-Learning学习率 $\eta = 0.01$，奖励折扣因子 $\gamma = 0.99$。实验中，每轮迭代结束条件为雷达状态进入终止状态，或者试错次数达到300次。基于上述MFR雷达环境设置可得表2所示的雷达任务转换关系，其中JM表示干扰类型。

表2　雷达任务转换关系

雷达任务状态	JM											
	1	2	3	4	5	6	7	8	9	10	11	12
s_1^A		s_2^B		s_2^A		s_1^A				s_1^A		
s_2^A	s_1^A				s_2^A	s_3^A		s_3^B				
s_2^B	s_1^A				s_2^B		s_3^C		s_3^A			
$s_3^A \sim s_9^C$	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
s_{10}^A		s_{11}^A			s_9^B			s_{10}^A				
s_{10}^B				s_9^C		s_{10}^B			s_9^B			s_{11}^A
s_{11}^A												

4.2　干扰有效性验证

4.2.1　不同强化学习算法所得奖励对比

相同电磁环境下，基于Q-Learning、DQN、TRPO和PPO的干扰决策算法所得平均奖励值随回合数变化如图4所示。由图4可知，上述4种算法所得平均奖励值均随回合数增加而增加，表明随着试错次数增加，所得策略有效性随之增加。其次，Q-Learning和DQN权衡探索利用时，将导致短期内奖励明显波动。此波动是不断尝试新策略过程中不可避免的现象。另外，随着回合数增加，DQN可能陷入局部最优，其不仅使得所得奖励持续波动，而且限制了智能体策略的有效提升。再者，离散至连续空间映射时，由于估计及近似误差的存在，TRPO并不能始终保证新策略所得奖励高于旧策略。由此，随着回合数增加，TRPO平均奖励值可能会下降，这是由于策略迭代过程中上述误差所导致。此外，相较于对比算法，PPO不仅可快速收敛至较高奖励值，而且随着回合数增加所得平均奖励波动逐渐变小，表明所提算法具有较高的决策有效性、有效率以及稳定性。

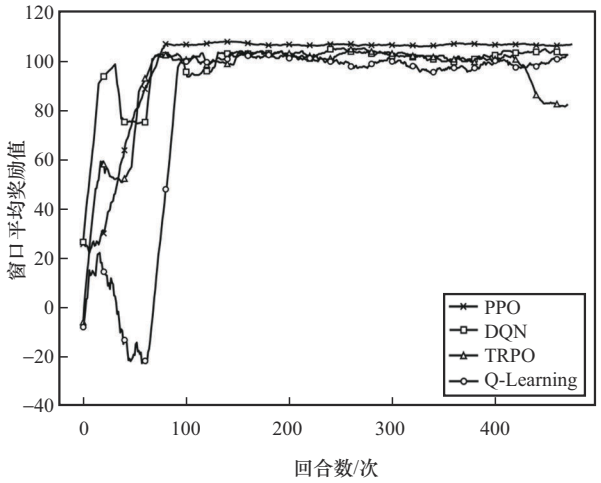


图4　基于不同强化学习模型的干扰决策算法所得平均奖励值随回合数变化

4.2.2　基于不同深度模型构建的PPO干扰决策模型所得奖励对比

电磁环境不变条件下，基于线性层、CNN以及LSTM构建的PPO干扰决策模型所得平均奖励值对比如图5所示。由图5可知，CNN作为基础架构所构造的PPO干扰决策模型所得平均奖励值明显低于LSTM。此可归因于如下事实：尽管CNN在空间特征提取方面具有独特优势，然而干扰决策模型并不涉及数据空间特性，因而其在策略优化过程中决策效率及有效性皆较为有限。相较之下，LSTM嵌入PPO模型具有较高的决策效率以及有效性。这是由于干扰决策任务具有序贯特性，而LSTM因其固有长时记忆机制，在处理时序依赖相关问题时具有显著优势，其所具有的门控结构使得网络可有效捕捉并充分利用历史决策信息，从而提升干扰决策的效率以及有效性。

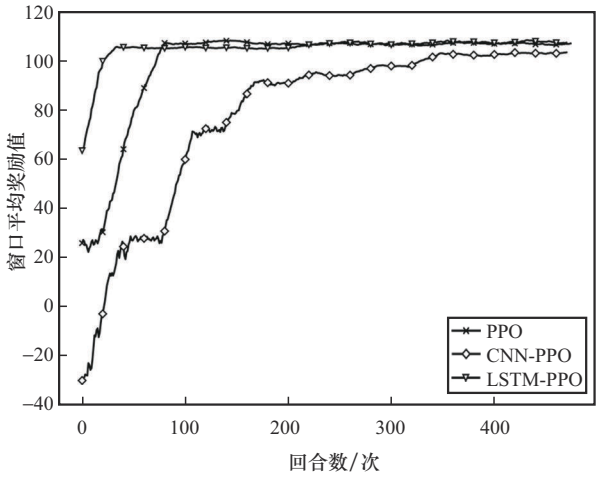


图5　基于不同深度网络构建的PPO干扰决策模型所得平均奖励值对比

4.2.3　嵌入先验知识所得奖励对比

基于奖励塑造理论将干扰领域先验知识嵌入上述基于不同深度模型构建的PPO干扰决策模型，所得平均奖励值如图6所示。由图6可知，相较于先验信息缺失的干扰决策模型，嵌入先验信息的干扰决策模型具有较高的决策效率及有效性，表明先验信息可引导决策模型快速收敛至较优策略并获得较高的策略收益。再者，由于LSTM优良的时序特征抽取能力，相较于CNN及传统的线性模型，LSTM-PPO在策略优化初期阶段便展现出较好的决策有效性。此外，先验知识嵌入的LSTM-PPO（PK-LSTM-PPO, prior knowledge-LSTM-PPO）在策略优化中后期所得平均奖励值始终优于先验知识嵌入的CNN-PPO（PK-CNN-PPO, prior knowledge-CNN-PPO）以及先验知识嵌入的PPO（PK-PPO, prior knowledge-PPO）。由此可知，相较于对比算法，所提算法可较快收敛至较优且稳定的干扰样式决策。

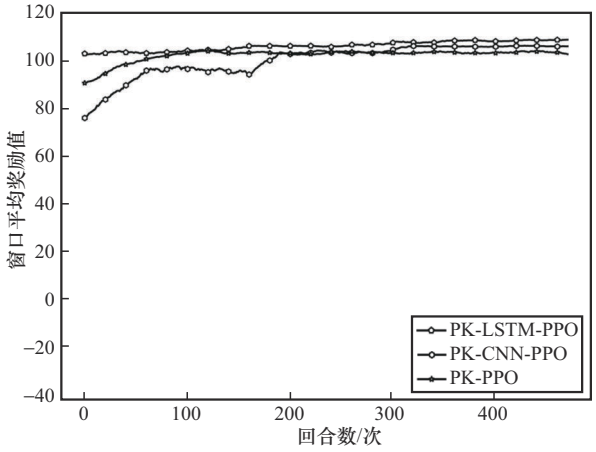


图6　先验知识嵌入所得平均奖励值对比

4.3　收敛性能对比

为了获得基于不同深度强化模型的干扰决策算法收敛性能的有效对比，此处将训练回合数从500次增加至1 000次以较为全面评估长时跨度上不同算法收敛性能，其他因素，如计算硬件配置、网络架构等与上述实验保持一致。由此，不同强化学习算法各回合收敛平均所用时间如表3所示，其中算法可分为先验知识嵌入前后两部分。由表3可知，先验知识未嵌入算法中，LSTM-PPO具有较快的收敛速度，其收敛时间相对于较慢的TRPO提升38.1%。再者，相较于先验知识未嵌入模型，先验知识嵌入算法干扰决策效率显著提升。此外，相较于其他对比算法，PK-LSTM-PPO具有较好的收敛性能，其收敛时间相对于先验知识未嵌入的LSTM-PPO提升约7.64倍，相较于PK-PPO以及PK-CNN-PPO分别提升约34.6%和47.0%。由此可得，基于LSTM优良的时序特征提取能力以及领域先验知识较强的收敛引导能力，相较于对比算法，所提算法具有较好的收敛性能，因而可较为有效地对抗复杂电磁环境。

表3　不同强化学习算法各回合收敛平均所用时间

干扰决策算法	平均每回合所用时间/ms
Q-Learning	129.712
DQN	142.435
TRPO	188.342
PPO	141.208
CNN-PPO	156.365
LSTM-PPO	116.423
PK-PPO	23.345
PK-CNN-PPO	28.742
PK-LSTM-PPO	**15.247**

4.4　决策路径对比

基于文献[7]所得先验信息，可以获得如下2个最短干扰路径。

$$\begin{aligned} & s_1^A \xrightarrow{2} s_2^B \xrightarrow{7} s_3^C \xrightarrow{8} s_4^B \xrightarrow{8} s_4^D \xrightarrow{4} s_6^B \xrightarrow{8} s_9^B \xrightarrow{} \\ & \begin{cases} \xrightarrow{3} s_{10}^B \xrightarrow{12} s_{11}^A \\ \xrightarrow{7} s_{10}^A \xrightarrow{2} s_{11}^A \end{cases} \end{aligned} \tag{26}$$

其中，$s_i^X, i = 1,2,\cdots,11$；$X = A,B,C,D$为雷达状态。

基于干扰设备免于陷入危险状态的原则，定义使得雷达状态沿着威胁等级递减方向并尽快转移至威胁最低状态的决策为正确的干扰样式决策。表4列出500回合后各算法所得最短干扰路径的正确干扰决策、未获得最短路径的干扰决策和干扰决策失败的比例。由表4可知，相较于PPO的各种变体算

表4　决策路径对比

强化学习算法	最短干扰路径	其他干扰路径	干扰失败
Q-Learning	45.5%	37.7%	16.8%
DQN	46.4%	44.6%	9%
TRPO	39.4%	40.2%	20.4%
PPO	51.8%	34.1%	14.1%
CNN-PPO	65.8%	33.0%	1.2%
LSTM-PPO	71.1%	25.4%	3.6%
PK-PPO	89.2%	9.1%	1.7%
PK-CNN-PPO	90.8%	9.2%	0
PK-LSTM-PPO	**93.4%**	**6.6%**	**0**

法，Q-Learning、DQN、TRPO和PPO等传统强化学习算法所得干扰失败率较高，表明传统强化学习算法选择干扰路径时鲁棒性较低。再者，基于LSTM和CNN构建的PPO所得干扰失败率较低，表明深度网络模型可抽取回波数据有效特征从而可显著提高决策效率及有效性。此外，先验知识的嵌入可给予PPO及其变体算法最短路径引导，使其倾向于最短且有效的干扰路径，从而缩短干扰路径并降低干扰失败概率。

综上所述，所提算法在收敛速度、稳定性以及干扰决策路径选择等方面，均表现出相对于传统强化学习算法的显著优势。由此表明，嵌入领域先验知识以及LSTM可有效提升复杂电磁环境下基于强化学习的干扰决策算法的干扰决策效率及有效性。

5 结束语

本文提出一种基于先验知识嵌入LSTM-PPO模型的智能干扰决策算法。所提算法将具有良好时序特征抽取能力的LSTM嵌入适宜于大规模策略寻优的PPO模型以感知MFR状态动态变化从而提升干扰决策有效性。此外，所提算法基于势能函数表征干扰领域先验知识并将其以重塑奖励函数形式嵌入LSTM-PPO模型以增强智能体对环境的感知精准度，从而提升复杂电磁场景下干扰决策算法的决策质量及效率。实验表明，相较于Q-Learning、DQN、TRPO和PPO等传统的深度强化学习模型，LSTM-PPO干扰决策模型可显著提升干扰决策模型的决策有效性。相较于先验知识缺失的干扰决策算法，所提算法可基于先验知识将策略优化引导至具有较高决策质量的较短决策路径，因而具有较好的决策效率及有效性。需要注意的是，干扰决策包含干扰样式选取及干扰资源配置，本文仅考虑干扰样式的智能选取，如何基于选定干扰样式智能配置干扰资源以提升干扰有效性是后续工作的重点。

参考文献：

[1] FENG L W, LIU S T, XU H Z. Multifunctional radar cognitive jamming decision based on dueling double deep Q-network[J]. IEEE Access, 2022, 10: 112150-112157.

[2] ZHANG C D, WANG L, JIANG R D, et al. Radar jamming decision-making in cognitive electronic warfare: a review[J]. IEEE Sensors Journal, 2023, 23(11): 11383-11403.

[3] GENG J, JIU B, LI K, et al. Radar and jammer intelligent game under jamming power dynamic allocation[J]. Remote Sensing, 2023, 15 (3): 581.

[4] SUN H, TONG N, SUN F. Electronic jamming mode selection based on D-S evidence theory[J]. Journal of Projectiles, Arrows and Guidance, 2003, 23(2): 218-220.

[5] LADOSZ P, WENG L L, KIM M, et al. Exploration in deep reinforcement learning: a survey[J]. Information Fusion, 2022, 85: 1-22.

[6] ZHENG S J, ZHANG C D, HU J, et al. Radar-jamming decision-making based on improved Q-learning and FPGA hardware implementation[J]. Remote Sensing, 2024, 16(7): 1190.

[7] XIA L Q, WANG L L, XIE Z D, et al. GA-dueling DQN jamming decision-making method for intra-pulse frequency agile radar[J]. Sensors, 2024, 24(4): 1325.

[8] 邹玮琦, 牛朝阳, 刘伟, 等. 基于A3C的多功能雷达认知干扰决策方法[J]. 系统工程与电子技术, 2023, 45(1): 86-92.
ZOU W Q, NIU C/Z)Y, LIU W, et al. Cognitive jamming decision-making method against multifunctional radar based on A3C[J]. Systems Engineering and Electronics, 2023, 45(1): 86-92.

[9] RAO N, XU H, WANG D, et al. Efficient jamming resource allocation against frequency-hopping spread spectrum in WSNs with asynchronous deep reinforcement learning[J]. IEEE Sensors Journal, 2024, 24 (8): 13560-13577.

[10] SHI Q, YING W D, LYU L, et al. Deep reinforcement learning-based attitude motion control for humanoid robots with stability constraints [J]. Industrial Robot: the International Journal of Robotics Research and Application, 2020, 47(3): 335-347.

[11] ZHONG J, WANG T, CHENG L L. Collision-free path planning for welding manipulator via hybrid algorithm of deep reinforcement learning and inverse kinematics[J]. Complex & Intelligent Systems, 2022, 8 (3): 1899-1912.

[12] SUI Z Z, PU Z Q, YI J Q, et al. Formation control with collision avoidance through deep reinforcement learning using model-guided demonstration[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(6): 2358-2372.

[13] LAURI M, HSU D, PAJARINEN J. Partially observable Markov decision processes in robotics: a survey[J]. IEEE Transactions on Robotics, 2023, 39(1): 21-40.

[14] LU R Z, JIANG Z Y, WU H M, et al. Reward shaping-based actor–critic deep reinforcement learning for residential energy management[J]. IEEE Transactions on Industrial Informatics, 2023, 19(3): 2662-2673.

[15] VAN HOUDT G, MOSQUERA C, NÁPOLES G. A review on the long short-term memory model[J]. Artificial Intelligence Review, 2020, 53 (8): 5929-5955.

[16] GU Y, CHENG Y H, CHEN C L P, et al. Proximal policy optimization with policy feedback[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52(7): 4600-4610.

[17] CANESE L, CARDARILLI G C, NUNZIO L D, et al. Multi-agent reinforcement learning: a review of challenges and applications[J]. Applied Sciences, 2021, 11(11): 4948.

[18] HICKLING T, ZENATI A, AOUF N, et al. Explainability in deep reinforcement learning: a review into current methods and applications[J]. ACM Computing Surveys, 2024, 56(5): 1-35.

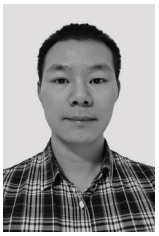
[19] RASHID T, SAMVELYAN M, WITT C S D, et al. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning[J]. Advances in Neural Information Processing Systems, 2020, 33:

10199-10210.
[20] ZHANG J Y, KOPPEL A, BEDI A S, et al. Variational policy gradient method for reinforcement learning with general utilities[J]. Advances in Neural Information Processing Systems, 2020, 33: 4572-4583.
[21] LI H P, HE H B. Multiagent trust region policy optimization[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(9): 12873-12887.
[22] ZHANG J W, ZHANG Z H, HAN S, et al. Proximal policy optimization via enhanced exploration efficiency[J]. Information Sciences, 2022, 609: 750-765.
[23] MOON J. Generalized risk-sensitive optimal control and Hamilton–jacobi–bellman equation[J]. IEEE Transactions on Automatic Control, 2021, 66(5): 2319-2325.
[24] MEYN S. The projected Bellman equation in reinforcement learning [J]. IEEE Transactions on Automatic Control, 2024, 69(12): 8323-8337.
[25] CLIFTON J, LABER E. Q-learning: theory and applications[J]. Annual Review of Statistics and Its Application, 2020, 7: 279-301.
[26] WENG W, GUPTA H, HE N, et al. The mean-squared error of double q-learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 6815-6826.
[27] CAMAGLIA F, NEMENMAN I, MORA T, et al. Bayesian estimation of the Kullback-Leibler divergence for categorical systems using mixtures of dirichlet priors[J]. Physical Review E, 2024, 109(2): 024305.
[28] TIAN Y J, ZHANG Y Q, ZHANG H B. Recent advances in stochastic gradient descent in deep learning[J]. Mathematics, 2023, 11(3): 682.
[29] ZHAO Y R, WANG X, HUANG Z T. Multi-function radar modeling: a review[J]. IEEE Sensors Journal, 2024, 24(20): 31658-31680.
[30] FENG H C, JIANG K L, ZHOU Z X, et al. Syntactic modeling and neural-based parsing for multifunction radar signal interpretation[J]. IEEE Transactions on Aerospace and Electronic Systems, 2024, 60(4): 5060-5072.
[31] ZHU M T, LI Y J, WANG S F. Model-based time series clustering and interpulse modulation parameter estimation of multifunction radar pulse sequences[J]. IEEE Transactions on Aerospace and Electronic Systems, 2021, 57(6): 3673-3690.
[32] MITCHELL E, RAFAILOV R, PENG X B, et al. Offline meta-reinforcement learning with advantage weighting[J]. arXiv Preprint, arXiv: 2008.06043, 2020.

[作者简介]

张静克（1988– ），男，河南洛阳人，博士，电子信息系统复杂电磁环境效应国家重点实验室助理研究员，主要研究方向为雷达对抗。

杨凯（1999– ），男，浙江绍兴人，浙江理工大学硕士生，主要研究方向为电子对抗。



李超（1986– ），男，河南洛阳人，博士，电子信息系统复杂电磁环境效应国家重点实验室助理研究员，主要研究方向为雷达对抗。

王洪雁（1979– ），男，河南南阳人，博士，浙江理工大学特聘教授，主要研究方向为雷达对抗、MIMO雷达信号处理、机器视觉等。