控制与决策
Control and Decision

基于SANER-PPO算法的无人机集群干扰资源分配方法

刘旖菲, 李小帅, 杨俊安, 杨渡佳, 王健

引用本文:
刘旖菲, 李小帅, 杨俊安, 等. 基于SANER-PPO算法的无人机集群干扰资源分配方法[J]. 控制与决策, 2024, 39(12): 3937-3945.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2023.1206

基于SANER-PPO算法的无人机集群干扰资源分配方法

刘旖菲[1], 李小帅[1,2], 杨俊安[1,2†], 杨渡佳[1,2], 王 健[1,2]

(1. 国防科技大学 电子对抗学院，合肥 230037；2. 电子制约技术安徽省重点实验室，合肥 230037)

摘 要: 针对高动态通信对抗场景下无人机集群协同干扰资源分配问题,提出一种结合状态正态化、优势标准化、熵正则化机制和近端策略优化算法(state normalization, advantage normalization and entropy regularization-based proximal policy optimization, SANER-PPO)的干扰资源分配方法. 首先,以无人机集群有效干扰的目标电台数量最大化和消耗的干扰功率最小化为目标函数,建立干扰资源分配优化问题;然后,将无人机集群映射为智能体,根据干扰资源分配模型建立马尔科夫决策过程;最后,利用SANER-PPO算法求解资源分配优化问题,生成无人机集群的干扰波束和干扰功率的优化决策结果. 相比于原始PPO算法,SANER-PPO算法将状态正态化机制引入智能体的决策阶段以增强算法的有效性,将优势标准化机制和熵正则化机制引入更新阶段来提升算法的收敛速度和稳定性. 结果表明,所提出算法能有效解决协同干扰资源分配问题,相较于原始PPO和柔性演员评论家两种算法,在资源消耗量和有效干扰的成功率方面具有明显优势. 进一步,通过逐步移除所提出算法的改进机制来进行消融实验,验证了3种改进机制的有效性.

关键词: 无人机集群；干扰决策；通信对抗；资源分配；强化学习；近端策略优化

中图分类号: TN975 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2023.1206

引用格式: 刘旖菲,李小帅,杨俊安,等. 基于SANER-PPO算法的无人机集群干扰资源分配方法[J]. 控制与决策, 2024, 39(12): 3937-3945.

SANER-PPO algorithm-based jamming resource allocation for UAV swarm

LIU Yi-fei[1], *LI Xiao-shuai*[1,2], *YANG Jun-an*[1,2†], *YANG Du-jia*[1,2], *WANG Jian*[1,2]

(1. College of Electronic Engineering，National University of Defense Technology，Hefei 230037，China；2. Anhui Province Key Laboratory of Electronic Restriction，Hefei 230037，China)

Abstract: This paper proposes an approach of jamming resource allocation based on an enhanced proximal policy optimization (PPO) algorithm to handle the jamming resource allocation problem of UAV swarms in the scenario of high-dynamic communication countermeasure. The enhanced PPO algorithm combines state normalization, advantage normalization, and entropy regularization mechanisms with the PPO algorithm, which is referred to as the SANER-PPO algorithm in this paper. Firstly, we aim at maximizing the number of target radios which are jammed by a UAV swarm successfully, while minimizing the sum of jamming power consumption of the UAV swarm. Then, the UAV swarm is modeled as agents, and a Markov decision process is established based on the jamming resource allocation model. Finally, an SANER-PPO algorithm is proposed to obtain optimal decisions of jamming beamforming and power allocation. When compared to the original PPO algorithm, the SANER-PPO algorithm not only incorporates a state normalization mechanism into the decision stage of the agent to improve its effectiveness, but also introduces advantage normalization and entropy regularization mechanisms to the update stage to improve the convergence speed and stability of the algorithm. Numerical results demonstrate that the performance of the proposed algorithm outperforms the original PPO algorithm and the soft actor-critic algorithm in terms of successful interference rate and jamming power consumption. In addition, ablation experiments are conducted by gradually removing the three proposed mechanisms in the algorithm, which validate the effectiveness of these mechanisms.

Keywords: UAV swarm；jamming decision；communication countermeasures；resource allocation；reinforcement learning；proximal policy optimization

收稿日期: 2023-08-25；**录用日期:** 2024-03-05.

基金项目: 国家自然科学基金项目(62201601).

责任编委: 张文安.

†通讯作者. E-mail: yangjunan@ustc.edu.

0 引 言

近年来,无人机集群已经受到广大学者的关注,并在军事作战领域得到广泛应用[1-2]. 在通信对抗场景中,无人机集群能够在多目标干扰和干扰信号的覆盖范围、幅度等方面协作,以增强干扰效能. 同时还能够采用分散式部署策略来增加干扰机的隐蔽性和干扰效果的有效性. 干扰资源分配是无人机集群协同干扰技术的一个关键问题,在干扰资源有限的条件下,资源分配的合理性和有效性对干扰效果和资源利用率有着至关重要的影响. 传统的基于单一的干扰平台和固定的资源配置方法无法满足实际需求,难以达到阻塞目标组网电台系统通信的目的[3]. 因此,在无人机集群协同干扰组网电台的通信对抗场景下,如何对有限的干扰资源进行灵活高效的分配,使其发挥最大的干扰作用是一个亟待解决的问题.

目前的资源分配方法主要包括传统最优化算法、启发式算法和强化学习算法[3],传统最优化算法通常采用全局搜索求解空间的方式[4]. 然而,随着干扰资源类型和数量的增加,问题的解空间规模、求解难度和计算复杂度也会增大,传统最优化算法难以获得最优解. 此外,大多数干扰资源分配问题属于NP-hard类型,传统最优化算法难以解决这类问题. 针对NP-hard问题,启发式算法可以有效求解[5],但是启发式算法对干扰目标的先验信息的完备性要求较高,而且随着场景复杂度的提升,做出的假设与实际情况的差异会逐渐变大. 因此,此类算法实用性受限,不能很好地解决干扰资源分配问题.

在实际的通信对抗环境中,往往无法获取被干扰方的参数和工作模式等先验信息. 而强化学习算法可以通过与环境的交互,以试错的方式逐步学习环境的特征和规律,并根据即时奖励不断优化调整所学到的策略,从而在无先验信息的条件下解决序列决策问题. 因此,许多学者将强化学习算法应用于干扰问题. 针对传统的干扰决策技术的灵活性较差,且容易造成资源浪费的问题,文献[6]将强化学习技术应用于干扰决策,提出了基于DDQN的通信干扰策略生成算法. 文献[7]提出一种基于自适应启发式加速Q学习算法的干扰决策方法,该方法根据雷达威胁等级的变化自适应地调整干扰样式和干扰功率. 但是DDQN算法和Q学习算法均属于基于值函数的强化学习方法,仅适用于离散动作空间,因此可以使用的干扰资源类型较少. 文献[8]将深度强化学习算法引入到无人机集群协同侦察、干扰雷达的任务中,提出了一种选择性经验存储策略的多智能体深度确定性策略梯度算法. 但是上述方法只考虑了有限个离散动作,不适用于求解取值连续的资源分配问题. 文献[9]基于整体对抗思想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. 然而,仅针对单个干扰站进行干扰资源分配,无法解决协同干扰资源分配问题. 针对通信组网对抗场景中的干扰资源分配问题,文献[10]提出了一种基于最大策略熵深度强化学习的干扰资源分配方法,文献[11]提出一种融合噪声网络的深度强化学习通信干扰资源分配算法. 上述两种方法均适用于连续动作空间任务,但是只考虑了静态的通信对抗场景,即干扰机和目标通信链路的位置固定,难以满足实际动态对抗任务的需求.

现有方法可适用的干扰资源类型有限,且多数工作都是面向静态的通信对抗场景,很少考虑干扰机和被干扰方的通信设备动态移动的情况. 鉴于现有方法存在的缺点,本文在无人机集群协同干扰目标组网电台高动态场景下,针对无人机集群的协同干扰资源分配问题,提出一种基于SANER-PPO算法的干扰资源分配方法. 首先以有效干扰的电台数量最大化和消耗的干扰功率最小化为目标设计优化函数,并将资源有限条件下的干扰资源分配问题转化为带约束条件的组合优化问题;然后,将无人机映射为智能体,根据干扰资源分配模型建立马尔科夫决策过程(Markov decision process, MDP);最后,将状态正态化机制、 优势标准化机制和熵正则化机制与原始PPO算法结合,提出一种SANER-PPO算法,并利用SANER-PPO算法求解协同干扰的资源分配优化问题. 实验结果表明,SANER-PPO算法在实际训练时长、资源消耗量和压制干扰的成功率3种指标上明显优于原始PPO算法和柔性演员评论家(soft actor critic, SAC)算法.

1 系统模型

1.1 无人机集群协同干扰系统模型

图1给出了己方无人机集群协同干扰目标通信网络的对抗系统模型. 假设M架无人机执行协同干扰目标组网电台的通信网任务,通信网由N条通信链路组成. 在通信对抗三维场景中,无人机的位置为(x_j, y_j, z_j),目标接收机的位置为(x_i, y_i, z_i),目标发射机的位置为(x_k, y_k, z_k). 在战场环境中,敌方所采取的电磁干扰、隐身技术和电子对抗等措施增加了己方获取目标信息的难度. 由于本文的侧重点为干扰资源分配方法的研究,忽略对侦察技术的探讨,假设己方可以运用侦察手段获取目标的准确位置、发射功率和威胁系数等信息,以用于计算环境反馈的即时

奖励.

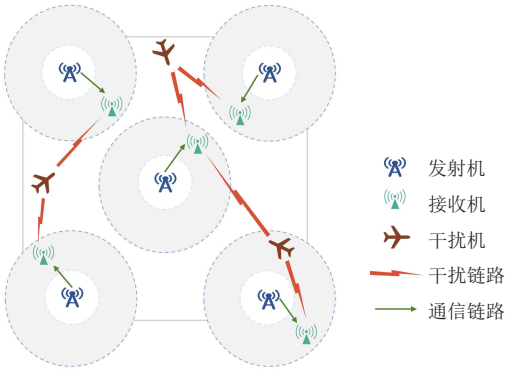

图 1　无人机集群协同干扰组网电台的通信对抗系统模型

本文设定无人机使用有源压制干扰、瞄准式干扰,且所采用的干扰系统为多波束工作模式. 此外,无人机可对干扰波束的发射功率进行分配.

由于无人机具有飞行高度高、移动性强等特点,假设无人机与目标接收机之间是视距传播信道模型[12]. 无人机 j 与目标接收机 i 间的信道增益 $g_{i,j}$ 为

$$g_{i,j} = \beta_0 d_{i,j}^{-2}. \tag{1}$$

其中: β_0 为参考距离 $d_0 = 1$ 米处的信道增益, $d_{i,j}$ 为无人机到目标接收机的距离.

由于真实地对地通信场景中,信道增益会受到大尺度衰落和小尺度衰落的影响[13],目标发射机 k 与目标接收机 i 之间的信道增益 $h_{i,k}$ 为

$$h_{i,k} = G\beta_{i,k}\zeta_{i,k}d_{i,k}^{-\alpha}. \tag{2}$$

其中: G 为由天线和放大器引入的恒定增益, $\beta_{i,k}$ 为服从指数分布的多径衰落增益,路径损耗 $\zeta_{i,k}$ 为服从对数正态分布的阴影衰落增益, α 为路径损耗指数, $d_{i,k}$ 为目标发射机到目标接收机的距离.

1.2　干扰效果评估模型

干信比为干扰信号功率与有用信号功率的比值,常被用于评估干扰效果. 因此,本文利用干信比对干扰效果进行衡量.

假设目标接收机接收到的多个干扰信号的功率符合线性叠加条件,目标接收机 i 所接收到的信号的干信比为

$$\mathrm{JSR}_i \triangleq \mathrm{JSR}_i(u_{ij}, P_{ij}) = \frac{\sum\limits_{j=1}^{M} u_{ij}P_{ij}g_{i,j}}{P_{ik}h_{i,k}}. \tag{3}$$

其中: u_{ij} 表示干扰波束分配因子,其为二元变量, $u_{ij} = 1$ 表示无人机 j 分配干扰波束干扰目标接收机 i,否则不进行干扰; P_{ij} 表示无人机 j 对目标接收机 i 的干扰功率; P_{ik} 表示目标发射机 k 与接收机 i 之间的通信功率.

为保证无人机对通信链路实施有效干扰,目标接收机接收的信号的干信比需满足条件 $\mathrm{JSR}_i \geqslant K$,其中 K 表示实施有效干扰的干信比阈值.

1.3　干扰资源分配优化模型

本文建立的干扰资源分配模型的目标是在集群干扰资源有限的条件下,通过联合优化干扰波束和干扰功率的资源分配结果,以实现无人机集群协同干扰的干扰效益最大化. 其中,干扰效益最大化是指在实现有效干扰的同时,最小化功率的消耗量. 建立干扰资源分配优化模型如下:

$$F[\mathrm{JSR}_i(u_{ij}, P_{ij}), w_i] = \sum_{i=1}^{N} w_i\left(\lambda_1 \times \mathrm{JSR}_i - \lambda_2 \times \sum_{j=1}^{M} u_{ij}P_{ij}\right). \tag{4}$$

其中: λ_1 和 λ_2 分别为干扰有效性指标和资源消耗量指标的相对重要程度, w_i 为第 i 条通信链路的威胁系数. 因此,资源有限条件下的干扰资源分配问题可转化为带约束条件的组合优化问题,具体表示为

$$\begin{aligned} &\max_{u_{ij},P_{ij}} F[\mathrm{JSR}_i(u_{ij}, P_{ij}), w_i]. \\ &\text{s.t. C1}: 0 \leqslant \left(P_j = \sum_{i=1}^{N} u_{ij} \cdot P_{ij}\right) \leqslant P_j^{\max}; \\ &\text{C2}: \sum_{i=1}^{N} u_{ij} \leqslant L; \\ &\text{C3}: u_{ij} \in \{0, 1\}. \end{aligned} \tag{5}$$

其中: C1 表示每架无人机的总干扰功率 $P_j^{\max}$ 有限; C2 表示每架无人机可分配的干扰波束个数 L 有限; C3 表示 u_{ij} 是二进制变量, $u_{ij} = 1$ 表示无人机 j 干扰目标接收机 i,即干扰功率 $P_{ij} > 0$.

2　基于SANER-PPO算法的无人机集群干扰资源分配方法

2.1　决策模型的构建

本文根据干扰资源分配模型建立的MDP如下.

1) 状态空间 S. 干扰资源分配的主要影响因素有当前时刻的通信信道增益 $\boldsymbol{H}(t)$、当前时刻的干扰信道增益 $\boldsymbol{G}(t)$ 和前一时刻的干扰效果 $\boldsymbol{E}(t-1)$. 因此,当前时刻的环境状态 $\boldsymbol{S}(t)$ 由 $\boldsymbol{H}(t)$、$\boldsymbol{G}(t)$ 和 $\boldsymbol{E}(t-1)$ 构成,其为 $(M \times 2 + 1)$ 行 N 列矩阵,即

$$\boldsymbol{S}(t) = [\boldsymbol{H}(t) \quad \boldsymbol{G}(t) \quad \boldsymbol{E}(t-1)]^{\mathrm{T}}, \tag{6}$$

其中 $\boldsymbol{H}(t)$ 和 $\boldsymbol{G}(t)$ 均为 M 行 N 列矩阵.

干扰效果 $\boldsymbol{E}(t-1)$ 为1行 N 列向量,表示为

$$\boldsymbol{E}(t-1) = [e_1(t-1) \quad e_2(t-1) \quad \ldots \quad e_N(t-1)], \tag{7}$$

其中$e_i(t-1)\in\{0,1\}$,$e_i(t-1)=1$表示目标接收机i的接收信号的干信比达到设定阈值,即干扰有效.

2) 动作空间A. 由无人机j对目标接收机i分配的干扰资源a_{ij}组成,而干扰资源分配情况a_{ij}由相互耦合的干扰波束分配因子u_{ij}和发射的干扰功率P_{ij}表征. 因此,动作空间$\boldsymbol{A}(t)$为M行N列矩阵,即

$$\boldsymbol{A}(t)=\begin{bmatrix} a_{11}(t) & a_{21}(t) & \cdots & a_{N1}(t)\\ \vdots & \vdots & & \vdots\\ a_{1M}(t) & a_{2M}(t) & \cdots & a_{NM}(t)\end{bmatrix}=$$
$$\begin{bmatrix} u_{11}(t) & u_{21}(t) & \cdots & u_{N1}(t)\\ \vdots & \vdots & & \vdots\\ u_{1M}(t) & u_{2M}(t) & \cdots & u_{NM}(t)\end{bmatrix}\circ$$
$$\begin{bmatrix} P_{11}(t) & P_{21}(t) & \cdots & P_{N1}(t)\\ \vdots & \vdots & & \vdots\\ P_{1M}(t) & P_{2M}(t) & \cdots & P_{NM}(t)\end{bmatrix}=\boldsymbol{U}(t)\circ\boldsymbol{P}(t).\quad(8)$$

其中:$u_{ij}=0$表示无人机j不对目标接收机i实施干扰,$u_{ij}\neq 0$表示无人机j对目标接收机i实施干扰且干扰功率为P_{ij}.

考虑到式(5)所示的有关干扰波束的约束条件C2和C3,本文利用干扰波束分配算法从干扰资源情况$\boldsymbol{A}(t)$中获取干扰波束和干扰功率的分配结果. 求解算法如下.

算法1 干扰波束分配算法.

输入: SANER-PPO算法输出的无人机集群的干扰资源分配情况$\boldsymbol{A}(t)$;

输出: 干扰波束分配矩阵$\boldsymbol{U}(t)$,干扰功率矩阵$\boldsymbol{P}(t)$.

① 初始化干扰波束分配矩阵$\boldsymbol{U}^{\text{ini}}(t)=\boldsymbol{0}_{M\times N}(t)$,干扰功率矩阵$\boldsymbol{P}^{\text{ini}}(t)=\boldsymbol{0}_{M\times N}(t)$

② for i in range(M)

③ 寻找$\boldsymbol{A}(t)$的第i个行向量的最大值$a_{i_1j_1}(t)$和次大值$a_{i_2j_2}(t)$

④ 更新$\boldsymbol{U}^{\text{ini}}(t)$和$\boldsymbol{P}^{\text{ini}}(t)$,$u_{i_1j_1}(t)=1$,$u_{i_2j_2}(t)=1$,$P_{i_1j_1}(t)=a_{i_1j_1}(t)$,$P_{i_2j_2}(t)=a_{i_2j_2}(t)$

⑤ end

⑥ $\boldsymbol{U}(t)=\boldsymbol{U}^{\text{ini}}(t)$,$\boldsymbol{P}(t)=\boldsymbol{P}^{\text{ini}}(t)$.

3) 奖励函数R. 考虑多目标优化问题,即最大化有效干扰的通信链路的数量与最小化干扰功率的消耗量,设计奖励函数的表达式为

$$R=\sum_{i=1}^{N}w_i\Big(\lambda_1\times\text{sgn}(\text{JSR}_i-K)-\lambda_2\times\sum_{j=1}^{M}u_{ij}P_{ij}\Big).\quad(9)$$

其中:w_i为第i条通信链路的威胁系数;sgn$(\cdot)$为符号函数,当$\text{JSR}_i-K\geqslant 0$时,$\text{sgn}(\text{JSR}_i-K)=1$,表示无人机集群成功协同干扰第$i$台目标接收机.

2.2 基于SANER-PPO算法的无人机集群干扰资源分配算法

针对无人机集群干扰资源分配问题,PPO方法具有以下优势: 首先,PPO方法可以通过对策略多次更新提高其性能,从而更好地适应干扰资源分配问题的动态变化特性;其次,PPO方法可以通过引入剪切参数限制策略更新的幅度,从而保证策略的稳定性. 在实际的电子对抗应用中,往往需要在一定的范围内调整干扰资源的分配结果,以适应不同的环境和需求,因此具备高稳定性和强鲁棒性的PPO方法适用于求解复杂动态场景下的无人机集群干扰资源分配问题.

本文从优化PPO算法的角度出发,引入状态正态化、优势标准化和熵正则化3种增强机制以提升原始PPO算法的性能.

1) 状态正态化机制.

强化学习中的智能体从表征环境的特征向量中学习,而该特征向量往往由多种物理量组成,例如信道增益和干扰效果评估值. 不同物理量一般具有不同的量纲,且取值范围会随着环境的变化而变化.

为了消除输入数据所具有的量纲不同特性和高度变化特性,本文引入状态正态化机制[14]. 将表征环境状态的所有物理量均缩放至同一量级,使得输入网络的特征数据具有可比较性,从而提升网络的学习效率. 通过减小表征不同环境的特征数据的差距,以避免网络对某些环境状态过度敏感,使得网络可以更好地适应不同的环境,从而增强网络的泛化性.

在实现过程中,考虑到内存和计算量的限制,通过下式对状态进行正态化处理,从而避免分配一个用于存放状态数据的无限大的内存空间,并避免多次计算所有状态的均值和方差:

$$s_t^{\text{norm}}=s_{t-1}^{\text{norm}}+(s_t-\hat{\mu}_{t-1}^s)\times(s_t-\hat{\mu}_t^s),$$
$$\hat{\mu}_t^s=\hat{\mu}_{t-1}^s+\frac{1}{t}(s_t-\hat{\mu}_{t-1}^s),$$
$$\hat{\sigma}_s=\sqrt{\frac{1}{t}s_t^{\text{norm}}}.\quad(10)$$

其中:s^{norm}为正态化处理后的状态数据,$\hat{\mu}^s$和$\hat{\sigma}^s$分别为s的均值和标准差的估计值,t为当前时刻.

2) 优势标准化机制.

在PPO算法中,通常利用广义优势估计计算优势函数的估计值$\hat{A}_t$[15],其表达式为

$$\hat{A}_t=\sum_{i=0}^{k-1}\gamma^i r(s_{t+i},a_{t+i})+\gamma^k V(s_{t+k})-V(s_t),\quad(11)$$

其中 $k \in \{1,2,\ldots,\infty\}$. 随着 k 的增大, 估计量 $\hat{A}_t$ 的偏差逐渐减小, 方差逐渐增大, 故偏差与方差之间为矛盾关系. 由于偏差和方差分别影响算法的收敛性和稳定性, 需调整 k 值以平衡优势函数估计量 $\hat{A}_t$ 的偏差和方差.

针对 $\hat{A}_t$ 的偏差和方差无法达到同步最优值的问题, 引入优势标准化机制[16], 通过自适应调整估计量的偏差和方差保证算法的收敛性和稳定性, 具体实现如下:

$$\hat{A}_t^{\text{norm}} = \frac{1}{\hat{\sigma}^A}(\hat{A}_t - \hat{\mu}^A),$$
$$\hat{\mu}^A = \frac{1}{\text{batchsize}} \sum_{i=1}^{\text{batchsize}} \hat{A}_t^i,$$
$$\hat{\sigma}^A = \sqrt{\frac{1}{\text{batchsize}} \sum_{i=1}^{\text{batchsize}} (\hat{A}_t^i - \hat{\mu}^A)^2}. \tag{12}$$

其中: $\hat{A}_t^{\text{norm}}$ 为正态化处理后的 $\hat{A}_t$, $\hat{\mu}^A$ 和 $\hat{\sigma}^A$ 分别为 $\hat{A}_t$ 的均值和标准差的批量估计值, batchsize 为批量大小. 在训练过程中, 通过将优势值标准化为具有零均值和一定标准差的形式, 使优势值的范围被控制在一个相对稳定的区间内. 此外, 该机制减小了极端值对学习过程产生的不利影响, 从而降低了方差. 因此, 优势标准化机制通过控制优势值的范围和抑制极端值, 在保证偏差为 0 的同时降低方差, 提升了算法的稳定性和收敛速度.

3) 熵正则化机制.

在训练过程中, 如果智能体未对环境中的状态-动作对进行充分采样, 则可能收敛到局部最优解, 从而导致无法学到最优策略.

针对上述问题, 引入熵正则化机制[17], 即通过引入策略熵来鼓励智能体充分探索环境. 熵是一个用于描述系统的无序程度的物理量, 因此可以用熵度量策略的不确定性. 具体地, SANER-PPO 算法最大化一个带有熵正则化项的目标函数, 即

$$L^{\text{Actor}} = \hat{\boldsymbol{E}}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t) - \beta \cdot H(\pi(a_t|s_t))]. \tag{13}$$

其中: clip$(\cdot)$ 为截断函数; $r_t(\theta)$ 为新旧策略的比值; ε 为人为设定的常数; β 为控制熵正则化项的相对权重的超参数; $H(\pi(a_t|s_t))$ 为状态 s_t 下的策略熵, 策略熵越大, 策略网络输出的动作概率分布越均匀, 智能体选择不同动作的可能性越大.

2.3　基于SANER-PPO算法的无人机集群干扰资源分配方法框架

图2是基于SANER-PPO算法的无人机集群协同干扰资源分配算法框架, 输入为从通信对抗环境中采集到的状态数据 $\boldsymbol{S}(t)$, 输出为无人机集群对目标接收机分配的干扰资源 $\boldsymbol{A}(t)$.

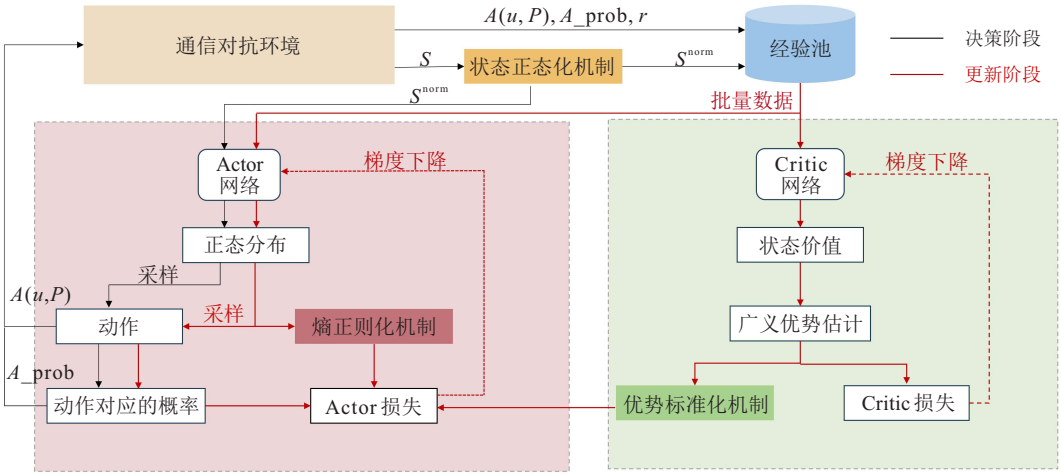

图 2　基于SANER-PPO的通信干扰资源分配算法框架

算法决策过程如下: 首先, 利用状态正态化机制对原始状态数据 S 进行处理, 处理后的状态数据 S^{norm} 作为 Actor 的输入; 然后, Actor 网络根据状态输出资源分配结果 $A(u,P)$ 和其所对应的概率值 A_prob, 其中 $A(u,P)$ 由相互耦合的干扰波束和干扰功率构成; 最后, 通过执行动作得到环境反馈的奖励值 r 和新的状态数据, 并将当前状态、动作、动作所对应的概率值、奖励和新的状态存储到经验池中.

算法更新过程如下: 从经验池中提取批量数据 buffer_S_t^{norm}、buffer_A_t、buffer_A_t_prob、buffer_r、buffer_S_{t+1}^{norm}, 计算 Critic 网络的损失函数, 先后将 buffer_S_t^{norm}、buffer_S_{t+1}^{norm} 数据作为 Critic 网络的输入, 分别得到状态价值函数 $V(S_t^{\text{norm}})$ 和 $V(S_{t+1}^{\text{norm}})$; 根据广义优势估计、buffer_r 和状态价值函数依次计

算优势函数 $\hat{A}_t$ 和Critic的损失值 L^{Critic},基于优势标准化机制处理优势函数得到 $\hat{A}_t^{\text{norm}}$,计算Actor网络的损失函数. 将buffer_S_t^{norm}数据作为Actor网络的输入,得到buffer_S_t^{norm}状态下所有动作的概率分布函数、执行动作 A_t 及其对应的概率 $\pi_\theta(A_t|S_t^{\text{norm}})$;利用批量数据buffer_$A_t$_prob(其与 $\pi_{\theta\text{old}}(A_t|S_t^{\text{norm}})$ 等效)和 $\pi_\theta(A_t|S_t^{\text{norm}})$ 计算得到新旧策略的比值 $r_t(\theta)$;根据熵正则化机制和动作的概率分布函数计算出策略熵 $H(\pi(A_t|S_t^{\text{norm}}))$,并利用上述数据和标准化后的优势函数 $\hat{A}_t^{\text{norm}}$ 计算Actor网络的损失函数;最后利用梯度下降法更新Actor的参数 θ 和Critic的参数 ϕ.

3 仿真实验及对比分析

3.1 参数设置

在本文设定的通信对抗仿真场景中,目标组网电台的通信网由 $N = 5$ 个发射机-接收机对组成,$M = 3$ 架无人机对5条通信链路实施干扰. 考虑三维通信对抗场景,目标发射机和我方无人机的水平位置在2 km × 2 km的矩形区域内随机生成,目标接收机的水平位置在以发射机位置为圆心的圆环区域内随机生成,且圆环内外半径分别为0.3 km和1 km. 此外,发射机和接收机的垂直高度均为0,无人机的垂直高度服从均值为1 km、方差为0.1 km的正态分布. 由电子侦察机获知的5条通信链路的威胁系数分别为0.222 8, 0.419 9, 0.573 4, 0.526 7, 0.797 0. 环境参数设置如表1所示[10-11,18].

表1 建立模型所用变量

参数	数值
通信发射机发射功率 P_i / dBm	23
干扰机最大发射功率 $P_j^{\max}$ / dBm	50
参考距离1米处的信道增益 β_0 / dB	−30
阴影衰落增益 $\log_{10}(\zeta_{i,k})$ / dB	$N(0,3)$
多径衰落增益 $\beta_{i,k}$ / dB	EXP(1)
干信比阈值 K / dB	5
单架无人机最大干扰波束个数 U	2
干扰有效性指标的相对重要程度 λ_1	1
资源消耗量指标的相对重要程度 λ_2	0.3

本文算法的网络架构如表2所示. 其中: 网络层参数的数值表示神经元数量,Linear表示线性网络,Relu和Tanh表示两种不同的激活函数. 仿真所用的计算机硬件参数为: Intel i7-12700KF CPU, 16 GB RAM, NVIDIA GeForce RTX 3070Ti显示适配器, Python版本为3.7, Pytorch版本为1.12.0.

表2 网络架构

参数	Actor网络	Critic网络
输入层	Linear, 35	Linear, 35
隐藏层1	64, Relu	64, Relu
隐藏层2	64, Relu	64, Relu
输出层	15, Tanh	Linear, 1

3.2 实验仿真分析

为了验证所提出SANER-PPO算法的有效性,将其与4种基准方法的学习性能进行比较,即原始PPO算法[19]、SAC 算法[20]、基于最大策略熵的深度强化学习(MPEDRL)算法[10]、融合噪声网络的深度强化学习(FNNDRL)算法[11]. 在具体实现中,原始 PPO算法没有采用2.2节表述的3个机制,其策略熵系数为0,其余参数设置与SANER-PPO算法相同. SAC、MPEDRL和FNNDRL算法的网络层数、 神经元数量与SANER-PPO算法相同,其余参数设置如表3所示[10-11,18-19].

表3 参数设置

参数	取值	参数	取值
训练回合数	1 000	批次样本大小	256
回合交互数	700	软更新系数	0.005
优化器	Adam	熵系数初始值	1
学习率	0.000 3	批量大小	2 048
折扣因子	0.1	策略熵系数	0.01
经验池大小	10^6		

3.2.1 模型训练

基于上述设计方法和参数设置进行实验,分别记录5种算法在训练阶段的奖励函数曲线,如图3所示. 随着训练的迭代步长数的增加,奖励值逐渐上升且趋于平稳. 由于算法始终需要权衡探索与利用之间的关系,有一定的概率输出奖励值较低的次优动作,收敛阶段的曲线存在小幅度的波动. 相较于原始PPO算法, SANER-PPO算法的波动幅度更小,即算法的稳定性更好. 此外,原始PPO算法在50 000个迭代步长后收敛到局部最优值,为165左右. SAC算法在70 000个迭代步长左右稳定在217, MPEDRL算法和FNNDRL算法均在200 000个迭代步长左右稳定在220. SANER-PPO算法在300 000个迭代步长后能稳定输出最优决策,奖励值收敛在225左右,优于其他基准算法. MPEDRL算法和FNNDRL算法均是SAC算法的改进算法,前者是目标函数方面的改进,后者是网络架构方面的改进,两者的有效性均优于SAC算法,但是收敛速度均比SAC算法慢. 由于SAC算法是离线策略方法,相较于在线策略PPO方法, SAC算法对训练阶段获取的样本数据的利用率更高,收敛速度更快. 但是训练SAC算法所耗费的实际时间远多于SANER-PPO算法,两种算法的训练迭代步长和耗费时间的对比如表4所示.

图4显示了SANER-PPO算法下,各接收机处的

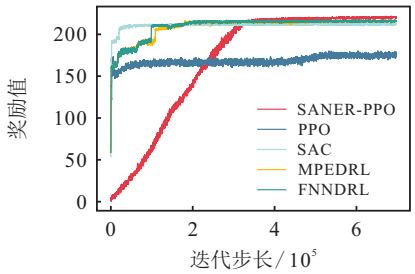

图 3　训练过程中不同算法的奖励值比较

表 4　不同算法的训练迭代步长和耗费时间对比

	SANER-PPO 算法		SAC 算法	
	步长	时间/s	步长	时间/s
收敛点	300 000	384.72	70 000	699.8
结束点	700 000	763.6	700 000	12 593.5

干信比随训练迭代步长的变化情况,为了评估干扰效果的有效性,将有效干扰的干信比阈值设定为 5 dB,即当接收机处的干信比值大于 5 dB 时,可以认为干扰机成功干扰该台接收机. 随着训练的进行,5 台接收机处的干信比稳定值均大于 5 dB,即 SANER-PPO 算法实现了对接收机的完全压制干扰,验证了基于 SANER-PPO 算法的无人机集群干扰资源分配方法的有效性和收敛性.

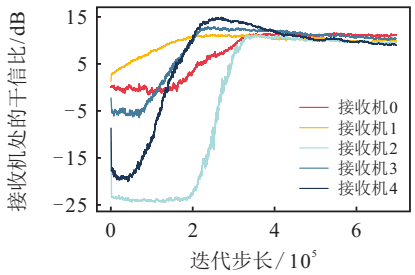

图 4　SANER-PPO算法下各接收机处的干信比

表 5 显示了训练过程中 5 种算法收敛时各接收机接收信号的干信比(单位: dB). 由表 5 可知,随着训练的进行,SAC 算法、MPEDRL 算法、FNNDRL 算法和 SANER-PPO 算法均可以实现对 5 台目标接收机的有效干扰,而原始 PPO 算法只能实现对 4 台接收机的有效干扰. 在被有效干扰的接收机处,原始 PPO 算法的干信比值最大,均大于 19 dB; SANER-PPO 算法的干信比值最小,均小于 12 dB; SAC 算法、MPEDRL 算法和 FNNDRL 算法的干信比值稳定在 13 dB 左右. 干信比值越大代表功率的消耗量越多,因此 SANER-PPO 算法的干扰效能最大,以最少的功率实现了对所有接收机的有效干扰.

表 5　算法收敛时各接收机处的接收信号的干信比

	接收机 0	接收机 1	接收机 2	接收机 3	接收机 4
SANER-PPO	11.230	10.034	9.713	10.467	9.202
PPO	21.269	19.987	19.873	21.545	3.349
SAC	13.816	13.744	13.819	13.839	13.838
MPEDRL	13.808	13.693	13.793	13.675	13.730
FNNDRL	13.841	13.672	13.825	13.691	13.733

3.2.2　模型测试

为了验证 SANER-PPO 算法的泛化性,在仿真环境中进行 70 000 个时间步长的测试实验,从实现完全压制干扰的成功率和干扰功率的消耗量两方面评估算法的性能. 表 6 显示了测试过程中无人机集群对所有目标接收机发射的总干扰功率(单位: dBm), SANER-PPO 算法消耗的总干扰功率最少,进一步验证了 SANER-PPO 算法可以在实现有效干扰的同时,尽可能节约干扰资源,因此其分配结果更符合本文的目标想定.

表 6　测试过程中干扰机消耗的总功率

算法	SANER-PPO	PPO	SAC	MPEDRL	FNNDRL
功率	66.777 1	77.211 1	70.761 7	70.913 8	70.867 6

表 7 为测试阶段不同算法实现完全压制干扰的成功率,原始 PPO 算法为 41.54 %. SANER-PPO 算法为 96.30 %,高于 SAC、MPEDR 和 FNNDRL 算法,表明其在长时隙动态的通信对抗环境下具有优异的泛化性能.

表 7 测试过程中实现完全压制干扰的成功率　　%

算法	SANER-PPO	PPO	SAC	MPEDRL	FNNDRL
成功率	96.30	41.54	92.78	93.40	93.82

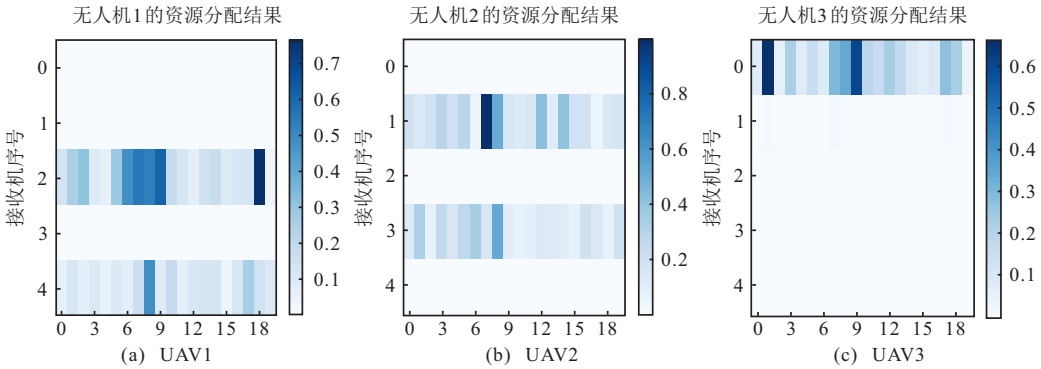

图 5　无人机的资源分配结果

图 5 给出了测试阶段 20 个连续时间步长中, 每台无人机的干扰资源分配结果. 网格颜色表示无人机集群分配的波束干扰对象和发射的干扰功率, 若功率值为 0, 则表示无人机未分配干扰波束指向目标接收机. 为了增强可比较性, 归一化处理分配的功率值. 从图中可以看出, 无人机 1 稳定干扰接收机 2 和接收机 4, 无人机 2 稳定干扰接收机 1 和接收机 3, 无人机 3 稳定干扰接收机 0. 此外, 无人机 3 的另一个干扰波束指向接收机 1, 且分配的干扰功率极少, 从而避免浪费干扰资源.

3.2.3 消融实验

为了探究各种增强机制对模型效果的影响, 在 SANER-PPO 算法的基础上删除不同的增强机制以产生 3 种算法, 通过比较不同算法性能评估各种机制的作用. 消融实验中 4 种算法的具体设置如表 8 所示. 其中: "●" 代表包含该机制, "○" 代表不包含该机制.

表 8 消融实验

算法	状态正态化	优势标准化	熵正则化
SANER-PPO	●	●	●
SANER-PPO-S	○	●	●
SANER-PPO-A	●	○	●
SANER-PPO-H	●	●	○

图 6 是消融实验中不同算法的性能对比曲线. 相较于原始 PPO 算法, 消融实验中 4 种算法的学习性能均有一定程度的提高, 且不同机制对算法的影响层面不同. SANER-PPO-S 算法获取的奖励值最小, 即陷入局部最优, 因此状态正态化机制影响了算法的有效性; SANER-PPO-A 算法和 SANER-PPO-H 算法均可以收敛到全局最优, 但收敛速度和稳定性不如 SANER-PPO 算法, 且 SANER-PPO-H 算法的收敛速度优于 SANER-PPO-A 算法, 因此优势标准化机制和熵正则化机制均提升了算法的收敛速度和稳定性, 优势标准化机制对收敛速度的影响更明显. 消融实验验证了本文引入的 3 种增强机制可以提升原始 PPO 算法的有效性、稳定性和收敛速度, 基于 SANER-PPO 算法的无人机集群干扰资源分配方法可以高效稳定地输出最优的资源分配结果.

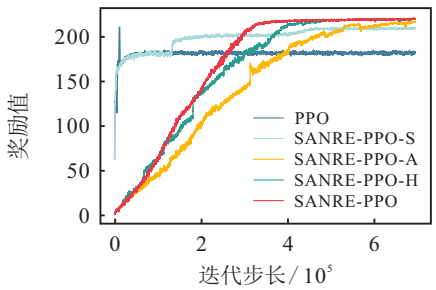

图 6 消融实验算法性能对比

表 9 为消融实验算法在测试过程中实现完全压制干扰的成功率. 结果表明, 相较于原始 PPO 算法, 消融实验中 4 种算法的压制干扰成功率均有一定程度的提升, 即 3 种增强机制均提升了原始 PPO 算法的有效性. SANER-PPO-S 算法压制干扰成功率的提升程度最小, 即状态正态化机制对算法的有效性影响最大. 结合理论分析, 状态正态化机制通过将信道增益的数值和干信比的数值缩放至同一量级, 减小了表征不同环境特征数据的差距, 从而提升了算法的泛化性和有效性.

表 9 消融实验算法实现完全压制干扰的成功率 %

算法	SANER-PPO	SANER-PPO-H	SANER-PPO-A	SANER-PPO-S	PPO
成功率	96.30	95.32	89.09	85.79	41.54

4 结 论

本文针对高动态的无人机集群协同干扰目标组网电台的通信对抗场景, 提出了一种基于 SANER-PPO 算法的干扰资源分配方法. 将无人机集群作为智能体, 利用 SANER-PPO 算法求解有限资源条件下的无人机集群协同干扰资源分配问题, 生成无人机集群干扰波束和干扰功率优化决策结果. 此外, SANER-PPO 算法通过状态正态化机制提升算法的有效性, 利用优势标准化机制和熵正则化机制提升算法的收敛速度和稳定性. 仿真结果表明, 所提出方法可以在最大化有效干扰电台数量的同时, 最小化无人机集群消耗的干扰功率资源, 与原始 PPO 算法和 SAC 算法相比, SANER-PPO 算法具有更优异的干扰效能. 消融实验进一步验证了引入的 3 种增强机制可以在有效性、稳定性和收敛速度 3 个层面提升算法性能.

未来将开展分布式方法的研究, 以解决本文所采用的集中式方法存在的信息交换开销大的问题, 提高集群系统的可扩展性和效率. 此外, 为了凸显算法性能, 本文对环境反馈的数据进行了理想化假设, 未考虑无法精确获得目标的参数信息的问题. 针对上述问题, 未来将开展概率模型和鲁棒性算法的研究, 以提高算法的实用性和稳健性.

参考文献 (References)

[1] 高程, 都延丽, 步雨浓, 等. 基于顺序扩展一致性包算法的多无人机分布式任务分配 [J]. 控制与决策, 2023, 38(11): 3242-3250.
(Gao C, Du Y L, Bu Y N, et al. Distributed task allocation of multiple UAVs based on sequential extended consensus based bundle algorithm[J]. Control and Decision, 2023, 38(11): 3242-3250.)

[2] 刘学达, 何明, 禹明刚, 等. 基于公共物品博弈的无人机集群弹药分配方法[J]. 控制与决策, 2022, 37(10): 2696-2704.
(Liu X D, He M, Yu M G, et al. UAV swarm ammunition distribution method based on public goods game[J]. Control and Decision, 2022, 37(10): 2696-2704.)

[3] 张大琳. 针对组网雷达的协同干扰资源调度方法研究[D]. 成都: 电子科技大学, 2022: 1-2.
(Zhang D L. Cooperative jamming resource scheduling for multi-jammer jamming netted radar system[D]. Chengdu: University of Electronic Science and Technology of China, 2022: 1-2.)

[4] 王瀚. 信息不确定条件下的IRS通信对抗干扰策略研究[D]. 南京: 南京邮电大学, 2022: 23-43.
(Wang H. Research on anti-jamming strategy of IRS communication pair under information uncertainty[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2022: 23-43.)

[5] Yao Z K, Liu T H, Wang C. Cooperative jamming resource allocation model based on the improved firefly algorithm[C]. Proceedings of the 2022 6th International Conference on Electronic Information Technology and Computer Engineering. Xiamen, 2022: 1719-1724.

[6] 曲正. 基于强化学习的通信干扰策略生成技术研究[D]. 石家庄: 中国电子科技集团公司电子科学研究院, 2022: 23-39.
(Qu Z. Research on communication interference strategy eneration based on rein for cement learning[D]. Shijiazhuang: China Academic of Electronics and Information Technology, 2022: 23-39.)

[7] Zhang W, Ma D, Zhao Z, et al. Design of cognitive jamming decision-making system against MFR based on reinforcement learning[J]. IEEE Transactions on Vehicular Technology. Florence, 2023: 1-15.

[8] 张磊, 李姜, 侯进永, 等. 基于改进强化学习的多无人机协同对抗算法研究[J]. 兵器装备工程学报, 2023, 44(5): 230-238.
(Zhang L, Li J, Hou J Y, et al. Research on multi-UAV cooperative confrontation algorithm based on improved reinforcement learning[J]. Journal of Ordnance Equipment Engineering, 2023, 44(5): 230-238.)

[9] 许华, 宋佰霖, 蒋磊, 等. 一种通信对抗干扰资源分配智能决策算法[J]. 电子与信息学报, 2021, 43(11): 3086-3095.
(Xu H, Song B L, Jiang L, et al. An intelligent decision-making algorithm for communication countermeasure jamming resource allocation[J]. Journal of Electronics & Information Technology, 2021, 43(11): 3086-3095.)

[10] 饶宁, 许华, 齐子森, 等. 基于最大策略熵深度强化学习的通信干扰资源分配方法[J]. 西北工业大学学报, 2021, 39(5): 1077-1086.
(Rao N, Xu H, Qi Z S, et al. Allocation method of communication interference resource based on deep reinforcement learning of maximum policy entropy[J]. Journal of Northwestern Polytechnical University, 2021, 39(5): 1077-1086.)

[11] 彭翔, 许华, 蒋磊, 等. 一种融合噪声网络的深度强化学习通信干扰资源分配算法[J]. 电子与信息学报, 2023, 45(3): 1043-1054.
(Peng X, Xu H, Jiang L, et al. A deep reinforcement learning communication jamming resource allocation algorithm fused with noise network[J]. Journal of Electronics & Information Technology, 2023, 45(3): 1043-1054.)

[12] Wang L. Research on link performance analysis and resource allocation technology of UAV relaying communications[D]. Beijing: Beijing University of Posts and Telecommunications, 2021: 19-30.

[13] Li X, Ma L, Shankaran R, et al. Joint power control and resource allocation modes election for safety-related V2X communication[J]. IEEE Transactions on Vehicular Technology, 2019, 68(8): 7970-7986.

[14] Passalis N, Tefas A, Kanniainen J, et al. Deep adaptive input normalization for time series forecasting[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(9): 3760-3765.

[15] Schulman J, Moritz P, Levine S, et al. High-dimensional continuous control using generalized advantage estimation[J/OL]. 2015, arXiv: 1506.02438.

[16] Tucker G, Bhupatiraju S, Gu S, et al. The mirage of action-dependent baselines in reinforcement learning[C]. International Conference on Machine Learning. Stockholm: PMLR, 2018: 5015-5024.

[17] Grandvalet Y, Bengio Y. Semi-supervised learning[M]. Cambridge: The MIT Press, 2006: 151-168.

[18] Zhao F. Research on multi-UAV cooperation and location planning method for wireless relay[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2022: 25-26.

[19] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J/OL]. 2017, arXiv: 1707.06347.

[20] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]. International Conference on Machine Learning. Piscataway: IEEE, 2018: 1861-1870.

作者简介

刘旖菲 (2001−), 女, 硕士生, 主要研究方向为强化学习、资源调度, E-mail: yifeiliu0616@163.com;

李小帅 (1989−), 女, 副教授, 博士, 主要研究方向为认知电子战、无人机集群任务决策、车联网资源分配, E-mail: xiaoshuai.li@nudt.edu.cn;

杨俊安 (1965−), 男, 教授, 博士生导师, 主要研究方向为信号处理、智能计算, E-mail: yangjunan@ustc.edu;

杨渡佳 (1991−), 男, 讲师, 博士, 主要研究方向为认知电子战、群体智能, E-mail: yangdj@nudt.edu.cn;

王健 (1991−), 男, 讲师, 博士, 主要研究方向为认知电子战、群体智能, E-mail: wangjiannudt@nudt.edu.cn.