doi：10.11809/bqzbgcxb2023.05.033

基于改进强化学习的多无人机协同对抗算法研究

张　磊[1,2]，李　姜[1,2]，侯进永[3]，高　远[1,2]，王　烨[1]

（1. 中国科学院 长春光学精密机械与物理研究所，长春　130033；
2. 中国科学院大学，北京 100049；3. 32802 部队，北京　100191）

摘要：多无人机的作战协同研究内容主要包含飞行协同、侦察协同及干扰协同，随着无人机数量及协同决策内容的增加，多智能体强化学习模型的状态空间及动作空间维度呈指数增长，多智能体强化学习算法在训练中不易收敛，协同决策水平难以得到显著提升。采用并对多智能体深度确定性策略梯度（MADDPG）算法原理进行模型构建，在此基础上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（SES-MADDPG）算法。该算法通过设置回收存储标准以及选择性因子，对进入经验池的经验进行选择性存储，以缓解奖励稀疏的问题。仿真实验表明，在保证算法时间复杂度的前提下，SES-MADDPG 算法比其他强化学习算法有了更好的收敛效果，相较于 MADDPG 算法，任务完成率提高了 25.427%。

关键词：无人机集群；强化学习；协同控制；群智能；攻防对抗

本文引用格式：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.

Citation format：ZHANG Lei，LI Jiang，HOU Jinyong，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.

中图分类号：TP18　　**文献标识码**：A　　**文章编号**：2096－2304(2023)05－0230－09

Research on multi-UAV cooperative confrontation algorithm based on improved reinforcement learning

ZHANG Lei[1,2]，LI Jiang[1,2]，HOU Jinyong[3]，GAO Yuan[1,2]，WANG Ye[1]

（1. Changchun Institute of Optics，Fine Mechanics and Physics，
Chinese Academy of Sciences，Changchun 130033，China；
2. University of Chinese Academy of Sciences，Beijing 100049，China；
3. Unit 32802 of the Chinese People's Liberation Army，Beijing 100191，China）

Abstract：The research of combat cooperation of multi-UAVs mainly includes flight cooperation，reconnaissance cooperation and interference cooperation. With the increase of both the number of UAVs and the content of cooperative decisions，state space and action space dimensions of the multi-agent reinforcement learning model grow exponentially. Multi-agent reinforcement learning algorithm is not easy to converge in training，and the level of cooperative decision-making is difficult to be significantly improved. This paper adopts and models on the principle of multi-agent deep deterministic policy gradient（MADDPG）algorithm，based on which it also proposes a multi-agent deep deterministic policy gradient algorithm of the selective experience storage policy（SES-MADDPG）. The algorithm selectively stores the experience entering the experience pool by setting the recycling storage criteria as well as selectivity factors

收稿日期：2022－07－01；**修回日期**：2022－09－11
基金项目：国家自然科学基金项目(61977059)
作者简介：张磊(1997—)，男，硕士研究生，E-mail：249812816@qq.com。
通信作者：李姜(1982—)，男，博士，副研究员，硕士生导师，E-mail：cclijiang@163.com。

to alleviate the problem of reward sparsity. The simulation experiments show that, with guaranteed time complexity of the algorithm, the SES-MADDPG algorithm has a better convergence effect than other reinforcement learning algorithms, and shows an increase of 25.427% in task completion rate compared with MADDPG algorithm.

Key words: unmanned aerial vehicle swarm; reinforcement learning; cooperative control; swarm intelligence; attack-defense countermeasure

0　引言

近年来,随着以机器学习为代表的人工智能技术的进一步突破,无人控制系统领域飞速发展[1],无人机和无人车广泛应用于物流配送[2]、航拍[3]、电力检修、工厂车间运输和军事侦察[4]。特别是在军事应用领域,各国都致力于研究控制无人机自主决策执行特定任务[5]。到目前为止,在自主决策空战算法研究领域,有3个主要研究方向:① 采用数学求解法,这个方法自从20世纪60年代就已经被提出,但是早年研究的任务较为简单,对于目前的复杂任务而言,具有很大的局限性,除此以外,这种方法需要严格的数学推导证明和复杂的数学模型。② 机器搜索方法[6],典型的方法有蒙特卡洛搜索[7]、决策树等,该类算法根据无人机所面对的不同情形进行态势评估并对威胁目标进行排序[8],最后根据评估结果和威胁目标排序进行动作决策[9],机器搜索方法的核心在于专家经验,所以要求研究人员具有很强的战场经验,模型泛化能力较弱,且难以应对复杂多变的战场情况。③ 处于研究前沿的深度强化学习方法,利用智能体的不断试错提升动作决策水平。

2013年DeepMind发表了一篇利用强化学习算法玩Atari游戏的论文,强化学习真正意义走上了大众舞台。

不同于监督学习,强化学习不需要大量已标记的数据,只需通过与环境交互进行大量的强化训练[10]。当面对不同的环境状态,智能体会根据算法选择不同的动作,环境会根据所做的动作更新下一个环境状态,同时还会根据不同的动作给予智能体一个奖励值。智能体训练的目标就是使得总奖励值最大,经过大量的训练,智能体将一步步优化决策策略。深度强化学习是强化学习的进一步发展,是与深度学习的有机结合。利用神经网络拟合策略函数或者价值函数,从而达到控制要求。相较于强化学习,深度强化学习更能胜任连续动作和复杂的任务[11]。

现如今,强化学习逐步应用于游戏、自动驾驶决策、推荐算法等领域。根据环境中智能体的数量,强化学习划分为单智能体强化学习和多智能体强化学习[12]。单智能体强化学习是指环境中只有一个智能体需要进行动作决策,AlphaGo就是典型的单智能体算法。由于环境中只有一个智能体进行决策,状态转移简单,控制相对容易。无人机群协同自主对抗属于多智能体强化学习,环境中存在多个智能体,竞争关系、合作关系以及合作竞争关系等复杂的关系存在于各个智能体之间。随着智能体数量的增加以及智能体之间的复杂关系让强化学习任务变得愈发困难。目前主流的单智能体强化学习算法包括DQN[13]、DDPG[14]、PPO[15]、A3C[16]等,主流的多智能体强化学习算法包括MADDPG[17]、QMIX[18]、VDN[19]等。

目前强化学习技术在无人机自主决策领域被广泛研究,在多无人机协同搜索、路径规划和编队控制等研究中,已经获得了不俗的成果[20]。

文献[21]提出了一种基于深度强化学习的任务动态分配方法。该方法使无人机进行实时交互,对任务执行的优先级顺序和执行时间加以约束,提高了有限时间内总体的任务完成度。文献[22]提出一种基于深度确定性策略梯度算法的改进算法,提高了算法训练速度以及无人机在导航过程中对环境的适应能力。文献[23]提出了一种多机协同空战决策流程框架,该框架提高了在多架无人机协同对抗场景下智能体间的协同程度。

结合现有的成熟算法研究以及目前所遇到的工程项目难题,发现现有算法在工程应用中存在了以下的不足之处:

1) 随着实验环境中无人机数量的增加,算法适应能力下降,任务完成度低,且精度不高。

2) 状态空间和动作空间过于庞大,经验回收池中有效经验较少,有时候会出现不收敛的问题。

3) 训练时间过长,且收敛效果不理想。

针对目前算法的不足之处和实际的工程项目需求,作者在现有多智能体算法MADDPG的基础上,在经验存储过程中引入了选择性经验存储机制,设置经验回收标准以及选择性因子。并根据实际任务环境合理设定奖励函数,最后通过仿真验证,证明了改进后的算法相较其他强化学习算法,在保证算法时间复杂度的前提下,有了更好的收敛效果。

1　任务描述及模型建立

1.1　任务描述

红蓝双方展开军事对抗仿真,红方出动无人机集群,无人机具有侦察和干扰功能,蓝方阵地布设雷达、空中预警机和防空导弹发射系统。红方的任务为出动无人机集群对蓝方雷达进行协同侦察,确定蓝方雷达位置,并对雷达进行协同干扰,掩护后方轰炸机进入投弹区域。无人机群自主决策飞行路线,自主分派干扰任务,并快速完成既定任务要求。蓝方的任务为阻挡红方的进攻并保护指挥部,在指挥部周围布设地面雷达和火力打击系统,并在空中布设预警机一架,围绕蓝方阵地进行飞行预警。场景示意图如图1所示。

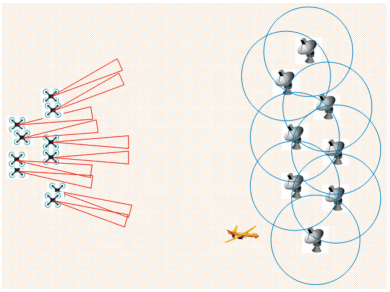


图 1 对抗过程示意图

Fig. 1 Schematic diagram of the confrontation process

1.2 模型建立

1.2.1 红方模型的建立

红方无人机群在执行任务时,受到包含风力、天气状况、地形地貌等自然因素的影响,以及蓝方防空雷达、预警机以及地面火力单元的威胁。构建智能体训练环境所需的计算模型概述如下。

红方无人机侦察到蓝方雷达信号的概率为:

$$P = 1 - \exp\left\{ - C\int_{t_0}^{t_1} [(x_t - \xi)^2 + (y_t - \zeta)^2]^{-2} \mathrm{d}t \right\} \quad (1)$$

式中:ξ 为目标的横坐标;ζ 为目标的纵坐标。该公式表示在时间(t_0, t_1)内发现目标的概率。

1.2.2 蓝方模型的建立

蓝方地面雷达侦察到红方的概率:当目标进入到雷达的探测区域后,雷达不一定发现目标,目标只是存在一定的概率会被发现。这个概率取决于雷达与目标发生直接的能量接触。一般而言,雷达的技术性能、目标的反射面积、目标的飞行高度与距离、雷达阵地(天线)高度等是影响目标被发现的主要概率。通常雷达有多种工作方式,为讨论方便,这里仅针对雷达的慢速扫描和快速扫描进行讨论[24]。

1) 雷达慢速扫描。

当雷达慢速扫描时,可将雷达对目标的探测视为离散观察,此时雷达的发现概率 P_{D} 为:

$$P_{\mathrm{D}} = 1 - \prod_{i=1}^{m} (1 - P_{\mathrm{d}_i}) \quad (2)$$

式中:m 为在持续搜索时间 t 时间段内,雷达与探测目标的接触次数,可按式(3)计算;$P_{\mathrm{d}_i}(1 \leqslant i \leqslant m)$为第 i 次与目标接触时的发现概率。

$$m = \lceil t / t_{\mathrm{search}} \rceil \quad (3)$$

式中,t_{search}为雷达的周期[24]。

在无电子干扰条件下,$P_{\mathrm{d}_i}(1 \leqslant i \leqslant m)$的计算表达式为:

$$P_{\mathrm{d}_i} = \left(\frac{n_0 S_{\mathrm{N}_i} + 1}{n_0 S_{\mathrm{N}_i}} \right)^{n-1} \exp\left\{ - \frac{Y_0}{n_0 S_{\mathrm{N}_i} + 1} \right\} \quad (4)$$

式中:n_0 为一次扫描的脉冲累积数;S_{N_i}为第 i 次与目标接触时单个脉冲的信噪比[24]。

2) 雷达快速扫描。

当雷达快速扫描时,可视为连续观察,在无干扰的情况下,雷达对点目标的发现概率为:

$$P = 1 - \exp\left\{ - C\int_0^t R(t)^{-4} \mathrm{d}t \right\} \quad (5)$$

记 $U = C\int_0^t R(t)^{-4}\mathrm{d}t$ 为发现势,则到 t 时刻,雷达到目标探测区的发现势为:

$$U = \int_0^t CR(t)^{-4}\mathrm{d}t = C\int_0^t [X_0^2 + (y_0 + V_\delta t)^2]^{-2}\mathrm{d}t \quad (6)$$

令 $y_0 + V_\delta t = X_0 \tan\phi$,则

$$U(X_0) = \frac{C}{V_\delta X_0^3}\int_{\phi_0}^{\phi} \cos^2\phi \mathrm{d}\phi =$$

$$\frac{C}{2V_\delta X_0^3}\left[(\phi - \phi_0) + \frac{1}{2}(\sin 2\phi - \sin 2\phi_0) \right] \quad (7)$$

在该段上发现目标的概率为:

$$P_{X_0} = 1 - \mathrm{e}^{-U(x_0)} \quad (8)$$

单发防空导弹打击的概率为:

$$P = 1 - \exp\left\{ \frac{\alpha W_d^\beta}{\sigma_d^\beta} \right\} \quad (9)$$

式中:W_d 为导弹的战斗部质量;σ_d 为没有干扰情况下导弹精度误差的均方差值;α、β、γ 为比例系数,在此 α 取 0.6,β 取 0.5,γ 取 0.7。

2 MADDPG 算法与 SES-ADDPG 算法

2.1 MADDPG 算法

多智能体强化学习以马尔科夫决策过程(MDP)作为算法的研究基础,可以利用一个高维元组$(S, A_1, \cdots, A_n, R_1, \cdots, R_n, P, \gamma)$进行描述。其中 S 是马尔科夫决策过程的状态集合,n 代表智能体的数量,$A_1, \cdots, A_n$ 代表各个智能体所选择的动作,$R_1, \cdots, R_n$ 代表每个智能体的收到环境给予的奖励回报,P 代表状态的转移函数,γ 代表折扣率。

多智能体深度确定性策略梯度(multi-gent deep deterministic policy gradient,MADDPG)算法是 OpenAI 团队在 2017 年提出的专门用来解决多智能体问题的算法,该算法可应用于合作、竞争以及竞争合作等多种环境场景下。它可以使多个智能体在高维度、动态化环境下通过智能体之间的通信以及智能体与环境之间的交互。能够使得多个智能体协同决策完成较为复杂的任务,是分布式计算方法在多智能体领域的优秀应用。除此之外,还能利用其他智能体的观测信息进行集中训练。训练过程采用集中训练,分散执行(centralized training with decentralized execution)的算法思想[17]。

MADDPG 是单智能体强化学习算法在多智能体领域的扩展,系统中的每个智能体都采用 DDPG 框架,每个网络学习策略函数(policy)π_{actor} 和动作价值函数(action value)Q_{critic};同时具有目标网络(target network),用 Q-learning 算法的异策略(off-policy)学习。Q 值计算公式为:

$$Q = Q(s_t, a_1, a_2, \cdots, a_n, \theta) \quad (10)$$

每个智能体都有一个 Actor 和 Critic 网络,当训练 Actor 网络时给予 Critic 更多的信息(其他智能体的观测信息以及动作信息),而在测试时去掉 Critic 部分,使智能体在得到充

分的训练之后,可以只通过 Actor 获取自己下一步的动作。这种获取全局信息的训练策略,可以避免像 Q-Learning、Policy Gradient 等单智能体算法直接迁移到多智能体环境下,由于只能获取自己的状态和动作,而产生的环境不稳定、经验回放失效等问题。MADDPG 算法能够使得每个智能体所面临的环境仍然可以视为稳定的,其原因为,系统的动力学模型可以描述为:

$$P(s' \mid s,a_1,a_2,\cdots,a_n,\pi_1,\pi_2,\cdots,\pi_n) = P(s' \mid s,a_1,\cdots,a_n) = P(s' \mid s,a_1,a_2,\cdots,a_n,\pi'_1,\pi'_2,\cdots,\pi'_n) \tag{11}$$

因此,当 $\pi_i \neq \pi'_i$ 时,整个环境仍然是稳定的。

利用 $\theta=(\theta_1,\theta_2,\cdots,\theta_n)$ 代表 n 个智能体策略函数的参数,用 $\pi=(\pi_1,\pi_2,\cdots\pi_n)$ 表示 n 个智能体的策略函数[25]。针对第 i 个智能体,我们把累计奖励期望值定义为:

$$J(\theta) = E[R_i] = E_{s\sim p^\pi,a_i\sim\pi_{\theta_i}}\left[\sum_{t=0}^{\infty}\gamma^t r_i,t\right] \tag{12}$$

式中:γ_i 为第 i 个智能体的奖励;γ 为折扣率。

针对随机策略梯度,求解策略梯度的公式为:

$$\nabla_{\theta_i} = E_{s\sim P^\pi,a_i\sim\pi_{\theta_i}}[\nabla_{\theta_i}\ln\pi_i(a_i \mid o_i)Q_i^\mu(x,a_1,a_2,\cdots,a_n)] \tag{13}$$

式中:o_i 为第 i 个智能体的观测值;$\boldsymbol{s}=[o_1,o_2,\cdots,o_n]$ 为所观测的向量,也就是状态。

$Q_i^\pi(s,a_1,\cdots,a_n)$ 表示环境场景下第 i 个智能体的集中式的状态 - 动作函数。因为环境场景下的每个智能体都在独立的学习各自的 Q_i^π 函数,所以系统中每个智能体都会得到各自的奖励函数(reward function),因此不仅可以完成合作场景的任务,还能完成竞争场景下的任务。

针对每个智能体建立价值函数,解决了传统强化学习算法在多智能体领域的不足[26]。其中 D 为一个经验存储池(experience replay buffer),它由一条条经验组成,经验表示为 $(s,s',a_1,\cdots,a_n,r_1,\cdots,r_n)$。式中,$S_t=(o_1^t,o_2^t,\cdots,o_n^t)$ 表示在 t 时刻环境内全部智能体的观测值集合,$a_1^t,a_2^t,\cdots,a_n^t$ 表示 t 时刻环境内全部智能体的动作集合,$r_1^t,r_2^t,\cdots,r_n^t$ 表示 t 时刻所有智能体在执行完相应动作后获得的奖励,$S_{t+1}=(o_1^{t+1},o_2^{t+1},\cdots,o_n^{t+1})$ 表示 $t+1$ 时刻所有智能体的观测值的集合。集中式的 critic 的更新方法借鉴了 DQN 中 TD 算法。

系统的损失函数定义为:

$$L(\theta_i) = E_{s,a,r,s'}[(Q_i^\mu(s,a_1,\cdots,a_n) - y)^2] \tag{14}$$

式中:$y=r_i+\gamma Q_i^{\mu'}(s',a'_1,\cdots,a'_n)\big|_{a_j=\mu_j(o_j)}$;$Q_i^\mu$ 为目标网络;$Q_i^\mu\mu'=[\mu'_1,\mu'_2,\cdots,\mu'_n]$ 为目标策略具有滞后更新的参数 θ'_j。

对于 actor 网络,参数的更新公式为:

$$\theta^u = \theta^u + \alpha_u\nabla_{\theta\pi}J \tag{15}$$

$$\nabla_{\theta x}J = \frac{1}{K}\sum_{t=1}^{K}\nabla_{\theta\pi}\pi(o,\theta^\pi)\nabla Q(s,a_1,a_2,\cdots,a_n,\theta^Q) \tag{16}$$

其中:θ^μ 为 actor 网络的参数值;α^μ 为 actor 网络的学习率。

2.2　SES-MADDPG 算法

选择性经验存储策略的多智能体深度确定性策略梯度(selective experience storage multi-agent deep deterministicpolicy gradient,SES-MADDPG)算法是 MADDPG 算法的改进提升。经过前期仿真实验可知,随着环境系统内智能体的数量增加,状态空间爆炸式扩张,导致算法训练时间延长,算法的奖励值收敛缓慢或者收敛值不理想。MADDPG 算法流程中存在经验池机制,智能体与环境交互产生的经验被存入经验池中,经验池里的经验将会被二次抽取,重新用于训练。经验池无保留地存储了所有的经验,其中高质量的经验便于算法的快速收敛,低质量的经验将不利于算法训练。其中低质量的经验占大多数,采用随机抽取将会抽取大量的低质量经验,因此将会消耗了大量的训练时间。前人研究者们为了改善这种问题,提出了一种优先经验抽取的机制[27],该机制为了抽出更好的经验,不再采用随机抽取,而是将进入经验池的经验根据损失进行排序,损失越大,排序越靠前。这种改进可以优先抽取高质量经验,加快算法的收敛速度,但是该机制存在时间复杂度较高的问题。每当一条新的经验进入经验池,该经验将会与经验池里的其他经验进行排序,排序的时间复杂度较高,大大增加了系统开销。

一方面为了改善经验优先回放算法时间复杂度过高的问题,另一方面需要控制经验池中经验的抽取。除了控制抽取的过程,还可以控制经验存储经验的过程。在经验回收存储时,并非无选择性地将交互产生的经验逐条存储至经验池内,而是设立经验回收标准,回收标准的具体数值应该根据奖励函数和实际问题进行设定。对于每条经验里的奖励值参量,对其求累积均值,当均值大于回收标准时,该条经验将会被存入经验池中 ,当小于回收标准时,系统产生 0 ~ 1 的随机数,当随机数小于选择性因子时,该条经验将会被存入经验池。该经验选择机制,既保证了对低质量经验的过滤,又避免了训练初期经验池内缺乏经验数据。除此之外,该算法实现简单,算法的时间复杂度为常数级别,有效地减轻了系统的开销。算法基本框架示意图如图 2 所示,SES-MADDPG 算法示意图如图 3 所示。

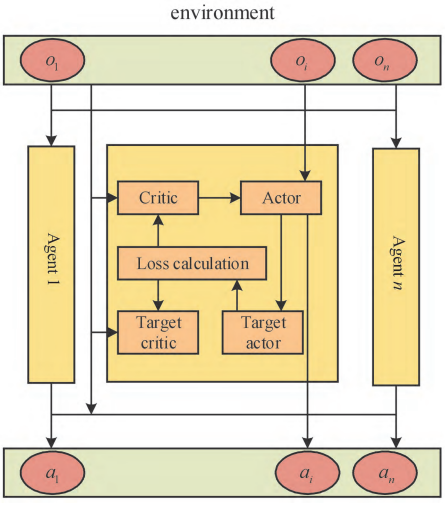


图 2　基本算法框架示意图

Fig. 2 Schematic diagram of the basic algorithm framework

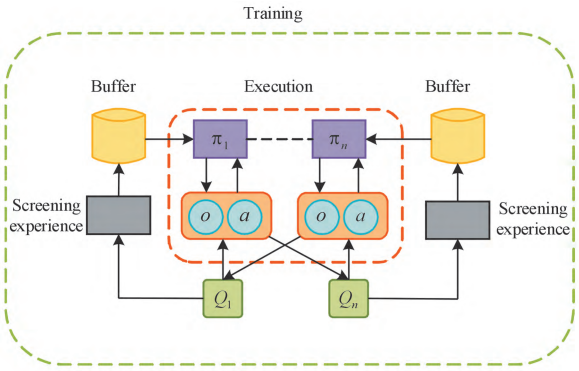


图 3 SES-MADDPG 算法示意图

Fig. 3 SES-MADDPG algorithm diagram

算法的伪代码实现如下。

对超参数进行随机初始化

对价值网络和策略网络的参数进行随机初始化

对目标价值网络和目标策略网络的参数进行随机初始化

初始化经验池 D 和动作噪声 N_t

for episode from 1 to num_episode do:

对环境和所有智能体的状态集合进行随机初始化

for step from 1 to max_episode_length do:

对于每个智能体进行动作选择,其中 $a_i=\mu_{\theta_i}(o_i)+N_t$

执行动作 $a=(a_1,\cdots,a_n)$,环境给与奖励 r,进入下一个环境 s'

获得到一条经验(s,a,r,s')

if ($r>W$):

存储经验进入经验池 D

else:

if(random(0,1) <β):

存储经验进入经验池 D(其中 W 是回收标准,β 是选择因子)

结束 if 判断语句

结束 else 判断语句

更新环境 $s\leftarrow s'$

for agent from 1 to n do:

从回收池随机抽取一条经验(s,a,r,s')

根据目标评估网络计算每个动作的期望回报 y

通过最小化损失更新 critic 网络的参数

$L(\theta_i)=E_{s,a,r,s'}[(Q_i^{\mu}(s,a_1,\cdots,a_n)-y)^2]$

使用随机梯度下降更新 actor

结束(agent) 循环

对于每个智能体更新目标网络参数

$\theta^{Q'}=\tau\theta^{Q}+(1-\tau)\theta^{Q'}$

$\theta^{\pi'}=\tau\theta_i^{\pi}+(1-\tau)\theta^{\pi'}$

结束(step)循环

结束(episode)循环

3 基于 SES-MADDPG 的协同对抗算法

3.1 状态空间的设计

本文中将无人机集群侦察、干扰敌方雷达的问题求解过程抽象为序列化决策过程,将作战环境中每个无人机视为一个智能体。强化学习的训练目标是构造一个智能网络模型,在每个状态都能做出决策,在避免被敌方发现的情况下,实现对敌方雷达的侦察与干扰。为了减少维度,训练环境在二维空间内进行。

无人机群的状态空间分为 2 个部分:第一部分为环境状态空间 S,代表了总体的环境状态;第二部分是智能体的观测状态 O,代表了无人机自身的状态以及对环境的捕获数据。分别如表 1 和表 2 所示。

表 1 环境状态空间

Table 1 Environmental state space

字段	说明
雷达位置	地面雷达的坐标
雷达频率	地面雷达信号发射频率
雷达探测范围	雷达实时的探测距离
预警机位置	预警机的实时坐标
预警机探测范围	预警机实时探测距离
目标位置	打击目标的坐标

表 2 智能体观测状态空间

Table 2 Agent observation state space

字段	说明
位置	无人机的坐标
朝向	无人机飞行方向
速度	无人机的飞行速度
方向	无人机的定向侦察方向
强度	无人机的干扰强度
频段	无人机的干扰频段
续航	无人机剩余生命值
标志位	是否开启干扰
目标	已经定位到的雷达位置

3.2 动作空间的设计

为了减小动作空间的维度,对部分动作做了离散和简化处理,具体动作可分为以下 6 个方面。① 飞行动作:无人机的飞行动作可以选择前、后、左、右和悬停等 5 个飞行动作。② 飞行速度:无人机的飞行速度可以选择低速、中速和高速等 3 个飞行速度。③ 定向侦察方向:无人机的定向侦察方向可以选择左前方、正前方和右前方等3 个方向。④ 定向干扰

强度:无人机的定向干扰强度可以选择不开干扰、低强度、中强度和高强度等4个强度。⑤ 干扰频段:无人机的干扰频段可以选择低频段(0.03 ~ 1 GHz)、中频段(1 ~ 15 GHz)和高频段(15 ~ 30 GHz)等3个频段。⑥ 干扰目标:无人机可以选择7个雷达的任意一个,共有7个选择目标。

根据以上6个方面进行动作组合选择,可产生3 780种不同的动作,即为动作空间,所有的动作选择采用独热编码格式。

3.3　奖励函数的设计

强化学习的目标是要获取最大的奖励值,根据任务场景设定奖励值,将有利于完成任务的状态设置正奖励值,将不利于完成任务的状态设置负奖励值。

由于无人机群之间需要协同完成任务,如果距离太远,将无法完成通信,因此需要设置无人机之间的距离奖励。

$$D(i,j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{17}$$

式中:$D(i,j)$为无人机i和无人机j之间的距离;(x_i,y_i)为无人机i的坐标;(x_j,y_j)为无人机j的坐标。

无人机之间的距离奖励为:

$$R = \begin{cases} 100 \times (\dfrac{C}{D} - 1), & D > C \\ 0, & D \leqslant C \end{cases} \tag{18}$$

式中,C为无人机之间的通信距离。

接近目标区域的奖励为:

$$R = \begin{cases} 500 \times \dfrac{D_{蓝} - d_{\text{now}}}{d_{\text{last}} - D_{红}}, & d_{\text{now}} > D_{蓝} - 2 \\ 100 \times (1 - \dfrac{d_{\text{now}}}{D_{蓝}}), & \text{其他} \end{cases} \tag{19}$$

式中:$D_{蓝}$代表蓝方雷达的探测距离;d_{now}代表此时无人机与蓝方雷达中心的距离;d_{last}代表上一时刻无人机与蓝方雷达中心的距离;$D_{红}$代表无人机的探测距离。

被雷达发现的奖励为:

$$R = -10 \tag{20}$$

发现雷达的奖励为:

$$R = 20 \tag{21}$$

对雷达的干扰奖励为:

$$R = 1000 \times (1 - \frac{D_{蓝now}}{D_{蓝}}) \tag{22}$$

式(22)中:$D_{蓝\text{now}}$代表被干扰后雷达的探测距离;$D_{蓝}$代表雷达最大的探测距离。

无人机被火力击落的奖励为:

$$R = -100 \tag{23}$$

开辟投弹区域的奖励为:

$$R = 200 \tag{24}$$

4　仿真实验

4.1　仿真环境介绍

为了验证由SES-MADDPG算法控制的无人机集群在战场对抗的有效性,在自建的多无人机智能对抗仿真推演平台进行对比实验验证。该仿真平台以海上登陆战为作战背景,以固定空域为作战环境,红方无人机集群在前方负责对蓝方的地面雷达和空中预警机进行侦察、干扰,为后方的轰炸机开辟投弹通道,使之顺利进入投弹范围进行投弹,对蓝方阵地进行火力打击。该仿真模拟环境选取了1 000 km × 850 km的空域范围作为作战区域,以1 km为一单位进行划分,将整个作战区域划分为1 000 km × 850 km的网格区域,便于多无人机集群在此区域进行飞行动作模拟和侦察动作模拟等。红方配备由10架侦干一体机构成的无人机集群,通过强化算法进行自主决策。蓝方配备7台地面雷达,1架空中预警机和数发航空导弹,其中地面雷达位置固定,预警机绕蓝方阵地作"8"字形或者沿跑道飞行。

红方无人机集群需要自主决策飞行路线,自主选择侦察方向等,对雷达进行侦察,同时锁定雷达位置,并对其进行干扰,为后方的轰炸机开辟投弹通道(即通道内无雷达探测信号覆盖)。

4.2　仿真实验设置

在自建的多无人机智能对抗仿真推演平台分别采用DQN算法、DDPG算法、MADDPG算法和SES-MADDPG算法进行20 000个实验周期的训练。每个周期的最大时间步为1 900步,当环境内的无人机个数不满足完成任务的最低个数或者任务提前完成时,该实验周期将会提前结束。通过对4种不同的算法进行对比,对算法进行评价比较。

以每个实验周期内的累积奖励以及任务的完成率作为评价指标。由于长机的设置与其他无人机稍有区别,因此当对比每个实验周期的奖励时,不仅比较10架无人机的平均奖励,也对长机获得的奖励进行单独比较。实验代码中部分超参数如表3所示。

表3　超参数

Table 3 Hyper-parameters

超参数	参数值
Actor 学习率	1e-4
Critic 学习率	1e-3
折扣率	0.9
噪声率	0.1
目标网络参数的更新率	0.01
buffer size	5e5
batch size	256

5　仿真实验分析

5.1　训练奖励值分析

在自建的多无人智能仿真推演平台分别使用了DQN算法、DDPG算法、MADDPG算法和SES-MADDPG算法进行20 000个周期的训练。其中图4为集群内所有无人机平均奖

励的对比图片,图5为长机平均奖励的对比图片。由图4、图5中可以看出,大约5 000个周期后,训练过程进入了较为平稳的收敛状态。MADDPG算法和SES-MADDPG算法的奖励收敛值明显高于DQN算法和DDPG算法。其中SES-MADDPG算法的收敛效果最好,相较于没有选择性回收机制的MADDPG算法,收敛值有了一定的提升。

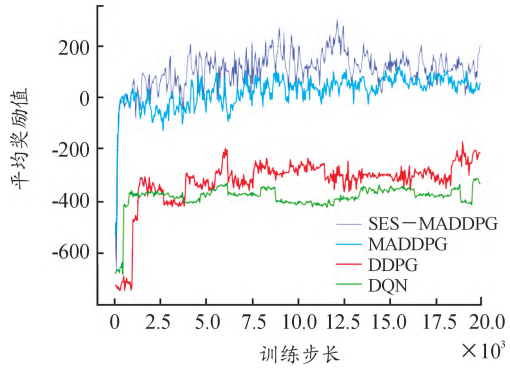


图4 所有无人机平均奖励对比图

Fig. 4 Comparison of average rewards for all UAVs

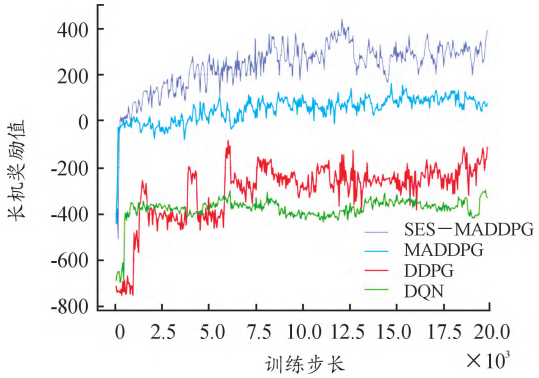


图5 长机的奖励对比图

Fig. 5 The reward comparison chart of the leader of UAV

表4和表5分别展示了不同强化学习算法在20 000个实验周期内的每架无人机的平均奖励和长机的奖励,由实验数据可看出,DQN和DDPG算法的平均奖励值均为负数,而SES-MADDPG算法的奖励值在100左右,远远高于其他的算法,充分证明了该算法的优越性。

表4 每架无人机前20 000轮的平均奖励对比

Table 4 Average reward comparison for the first 20 000 rounds per UAV

算法	平均奖励
DQN	-384.42
DDPG	-326.77
MADDPG	23.50
SES-MADDPG	106.42

表5 长机前20 000轮的平均奖励对比

Table 5 Comparison of the average rewards for the first 20 000 rounds of the leader UAV

算法	平均奖励
DQN	-381.49
DDPG	-303.59
MADDPG	48.36
SES-MADDPG	236.51

评估算法好坏的另一种方式是任务的完成度,为了避免训练前期收敛值不稳定对实验产生的影响,分析了后10 000个实验周期的任务完成情况,如表6所示,展示了在10 000次的训练过程中,成功完成任务的次数。虽然任务的完成率不高,但是相较于MADDPG算法,任务完成率提高了25.427%。

表6 任务成功完成的次数

Table 6 Number of times of the task successfully completed

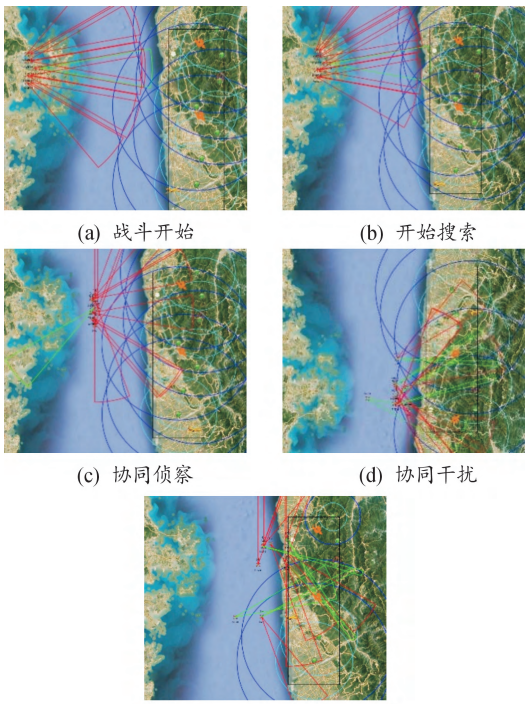
算法	完成次数
DQN	138
DDPG	196
MADDPG	2 167
SES-MADDPG	2 718

5.2 单次仿真可视化结果分析

利用SES-MADDPG算法经过20 000次训练后得到的模型,执行单次仿真推演进行效果的可视化评估分析。

可视化演示如图6所示。图6(a)为仿真开始,10架侦干一体机构成的集群做好出发准备,设置0号为长机。图6(b)集群内的无人机试探前进,对范围内的信号进行扫描探测。图6(c)无人机集群进入蓝方阵地,开始对雷达进行分散探测定位。图6(d)无人机确定雷达方位,对雷达进行持续干扰,被干扰后的雷达的探测范围大大降低。图6(e)集群内无人机团结协作,在蓝方阵地开辟出投弹通道,任务成功结束。

图7为DQN算法模型经过20 000次训练后得到的无人机运动轨迹图,图7中蓝色点代表雷达的位置。有4架无人机被蓝方雷达发现并被击毁。其余无人机没有进行有效的侦察和干扰,运动无规律性,仅仅在某个区域进行徘徊。图8为SES-MADDPG算法模型经过20 000次训练后得到的无人机运动轨迹图,从图8中可以看出,在未发现雷达前,无人机集群试探性前进,当发现蓝方雷达后,0号无人机绕着蓝方阵地进行往复移动,其目的是对运动的侦察机进行持续性干扰。其余的无人机各自进行任务分配,对蓝方的地面雷达进行持续性干扰,最后成功压制了雷达的探测范围,为轰炸机开辟了投弹通道。



(a) 战斗开始　(b) 开始搜索

(c) 协同侦察　(d) 协同干扰

(e) 投弹区域开辟，任务完成

图 6　对抗的仿真结果

Fig. 6 Simulation results of the confrontation

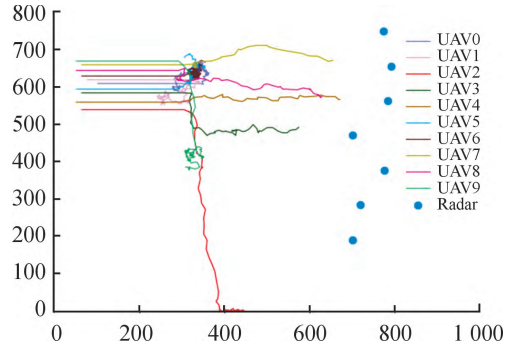


图 7　基于 DQN 算法的无人机飞行轨迹

Fig. 7 UAVs flight path based on DQN algorithm

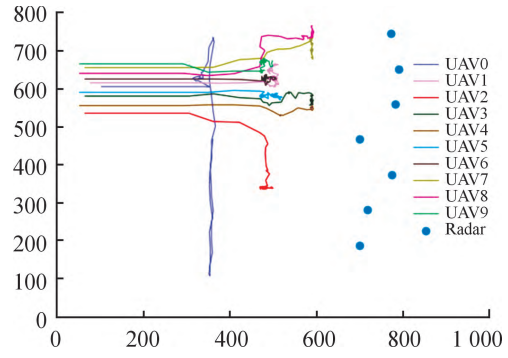


图 8　基于 SES-MADDPG 算法的无人机飞行轨迹

Fig. 8 UAVs flight path based on SES-MADDPG algorithm

6　结论

针对红蓝对抗问题,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。为了解决收敛效果差、任务完成率低的问题,在 MADDPG 算法的基础上,加入选择性经验回收机制,提出了 SES-MADDPG 算法。仿真实验结果表明:SES-MADDPG 算法比其他几种强化学习算法具有更好的收敛效果,同时任务完成率相较于 MADDPG 算法提高了 25.427%。

该算法虽然提高了收敛效果和任务完成率,但是会存在一定概率陷入局部最优的情况。下一步研究方向:一方面要克服陷入局部最优的缺陷,另一方面将该算法的环境推广至三维空间环境中。

参考文献:

[1] 杜威,丁世飞. 多智能体强化学习综述[J]. 计算机科学,2019,46(8):1 - 8.
DU Wei, DING Shifei. Overview on multi-agent reinforcement learning[J]. Computer Science,2019,46(8):1 - 8.

[2] YU H,MEIER K,ARGYLE M,et al. Cooperative path planning for target tracking in urban environments using unmanned air and ground vehicles[J]. IEEE/ASME Transactions on Mechatronics,2014,20(2):541 - 552.

[3] TODA K,TAMAMITSU M,IDEGUCHI T. Adaptive dynamic range shift (ADRIFT) quantitative phase imaging[J]. Light:Science & Applications,2021,10(1):1 - 10.

[4] ZHAO W,CHU H,MIAO X,et al. Research on the multiagent joint proximal policy optimization algorithm controlling cooperative fixed-wing UAV obstacle avoidance[J]. Sensors,2020,20(16):4546.

[5] BAYERLEIN H,THEILE M,CACCAMO M,et al. Multi-UAV path planning for wireless data harvesting with deep reinforcement learning[J]. IEEE Open Journal of the Communications Society,2021,2:1171 - 1187.

[6] HE W,QI X,LIU L. A novel hybrid particle swarm optimization for multi-UAV cooperate path planning[J]. Applied Intelligence,2021,51(10):7350 - 7364.

[7] 何旭,景小宁,冯超. 基于蒙特卡洛树搜索方法的空战机动决策[J]. 空军工程大学学报(自然科学版),2017,18(5):36 - 41.
HE Xu,JING Xiaoning,FENG Chao. Air combat maneuver decision based on MCTS methods[J]. Journal of Air Force Engineering University(Natural Sciences Edition),2017,18(5):36 - 41.

[8] RASHKI M. The soft Monte Carlo method[J]. Applied Mathematical Modelling,2021,94:558 - 575.

[9] 嵇慧明,余敏建,乔新航,等.改进BAS-TIMS算法在空战机动决策中的应用[J].国防科技大学学报,2020,42(4):123-133.
JI Huiming,YU Minjian,QIAO Xinhang,et al. Application of improved BAS-TIMES algorithm in air combat maneuvaer desion[J]. Journal of National University of Defense Technology,2020,42(4):123-133.

[10] 刘全,翟建伟,章宗长,等.深度强化学习综述[J].计算机学报,2018,41(1):1-27.
LIU Quan,ZHAI Jianwei,ZHANG Zongchang,et al. A survey on deep reinforcement learning[j]. Chinese Journal of Computers,2018,41(1):1-27.

[11] 王毅然,经小川,贾福凯,等.基于多智能体协同强化学习的多目标追踪方法[J].计算机工程,2020,46(11):90-96.
WANG Yiran,JING Xiaochuan,JIA Fukai,et al. Multi-target tracking method based on multi-agentcollaborative reinforcement learning[J]. Computer Engineering,2020,46(11):90-96.

[12] 韩统,崔明朗,张伟,等.多无人机协同空战机动决策[J].兵器装备工程学报,2020,41(4):117-123.
HAN Tong,CUI Minglang,ZHANG Wei,et al. Multi-UCAV cooperative air combat maneuvering decision[J]. Journalof Ordnance Equipment Engineering,2020,41(4):117-123.

[13] MNIH V,KAVUKCUOGLU K,SILVER D,et al. Human-level control through deep reinforcement learning[J]. Nature,2015,518(7540):529-533.

[14] LILLICRAP T P,HUNT J J,PRITZEL A,et al. Continuous control with deep reinforcement learning[C]//The International Conference on Learning Representations,San Juan,Puerto Rico,2016.

[15] SCHULMAN J,WOLSKI F,DHARIWAL P,et al. Proximal policy optimization algorithms[EB/OL]. 2017-08-28[2022-09-10]. https://arxiv.53yu.com/abs/1707.06347.

[16] MNIH V,BADIA A P,MIRZA M,et al. Asynchronous methods for deep reinforcement learning[C]//International Conference on Machine Learning,2016:1928-1937.

[17] LOWE R,WU Y,TAMAR A,et al. Multi-agent actor-critic for mixed cooperative-competitive environments[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach,USA:MIT Press,2017:6379-6390.

[18] RASHID T,SAMVELYAN M,SCHROEDER C,et al. Qmix:Monotonic value function factorisation for deep multi-agent reinforcement learning[C]//Proceedings of theInternational Conference on Machine Learning,2018:4295-4304.

[19] SUNEHAG P,LEVER G,GRUSLYS A,et al. Value-decomposition networks for cooperative multi-agent learningbased on team reward[C]//Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. Richland,2018:2085-2087.

[20] SUTTON R S,BARTO A G. Reinforcement learning:An introduction[M]. MIT press,2018.

[21] 唐峯竹,唐欣,李春海,等.基于深度强化学习的多无人机任务动态分配[J].广西师范大学学报(自然科学版):2021,39(6):63-71.
TANG Fengzhu,TANG Xin,LI Chunhai,et al. Multi-layer interactive color image encryption algorithm based on chaotic map[J]. Journal of Guangxi NormalUniversity (Natural Science Edition),2021,39(6):63-71.

[22] 薛喜地.基于深度强化学习的室内无人机避障[D].哈尔滨:哈尔滨工业大学,2020.
XUE Xidi. Collision avoidance for indoor UAVbased on deep reinforcement learning[D]. Harbin:Harbin Institute of Technology,2020.

[23] 施伟,冯旸赫,程光权,等.基于深度强化学习的多机协同空战方法研究[J].自动化学报,2021,47(7):1610-1623.
SHI Wei,FENG Yanghe,CHENG Guangquan,et al,Research on multi-aircraft cooperative air combat method based on deep reinforcement learning[J]. Acta Automatica Sinica,2021,47(7):1610-1623.

[24] 方洋旺,伍友利,魏贤智,等.航空装备作战建模与仿真[M].北京:国防工业出版社,2012.
FANG Yangwang,WU Youli,WEI Xianzhiet al. Operational modelling and simulation of aviation equipment[M]. National Defense Industry Press,2012.

[25] SUN R,SILVER D,TESAURO G,et al. Introduction to the special issue on deep reinforcement learning:An editorial[J]. Neural Networks:the Official Journal of the International Neural Network Society,2018,107:1-2.

[26] LV L,CHEN Z,LU Z. A novel neural-network gradient optimization algorithm based on reinforcement learning[C]//Proceedings of the 2019 International Conference on Security,Pattern Analysis,and Cybernetics (SPAC),2019.

[27] 何明,张斌,柳强,等.MADDPG算法经验优先抽取机制[J].控制与决策,2021,36(1):68-74.
HE Ming,ZHANG Bin,LIU Qiang,et al. Multi-agent deep deterministic policy gradient algorithm VIA prioritizedexperience selected method[J]. Control and Desion,2021,36(1):68-74.

科学编辑 甄子洋 博士(南京航空航天大学教授、博导)
责任编辑 贺 柳