改进DDPG无人机航迹规划算法

高敬鹏[1,2],胡欣瑜[2],江志烨[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 471003
2.哈尔滨工程大学 信息与通信工程学院,哈尔滨 150001
3.北京航天长征飞行器研究所 试验物理与计算数学国家级重点实验室,北京 100076

摘 要:针对无人机飞行过程存在未知威胁使智能算法处理复杂度高,导致航迹实时规划困难,以及深度强化学习中调整DDPG算法参数,存在时间成本过高的问题,提出一种改进DDPG航迹规划算法。围绕无人机航迹规划问题,构建飞行场景模型,根据飞行动力学理论,搭建动作空间,依据非稀疏化思想,设计奖励函数,结合人工蜂群算法,改进DDPG算法模型参数的更新机制,训练网络模型,实现无人机航迹决策控制。仿真结果表明,所提算法整体训练时长仅为原型算法单次平均训练时长的1.98倍,大幅度提升网络训练效率,降低时间成本,且在满足飞行实时性情况下,符合无人机航迹质量需求,为推动深度强化学习在航迹规划的实际应用提供新思路。

关键词:深度确定性策略梯度算法;无人机;航迹规划;深度强化学习;人工蜂群算法
文献标志码:A **中图分类号:**TP273 doi:10.3778/j.issn.1002-8331.2106-0054

Unmanned Aerial Vehicle Track Planning Algorithm Based on Improved DDPG

GAO Jingpeng[1,2], HU Xinyu[2], JIANG Zhiye[3]
1.State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System (CEMEE), Luoyang, Henan 471003, China
2.College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China
3.National Key Laboratory of Science and Technology on Test Physics and Numerical Mathematics, Beijing Institute of Space Long March Vehicle, Beijing 100076, China

Abstract: An improved DDPG flight track planning algorithm is proposed, aiming at the problem of high processing complexity of intelligent algorithm due to unknown threats in UAV flight process which leads to the difficulty of real-time flight track planning, and long training time by adjusting the parameters of DDPG algorithm in deep reinforcement learning. The flight scene model is established under the background of UAV track planning. According to the flight dynamics theory, the action space is built. On the basis of the non-sparse idea, the reward function is designed. Combined with the artificial bee colony algorithm, the updating mechanism of the model parameters of DDPG algorithm is improved, and the network model is trained to achieve the flight track decision-making of UAV. Simulation results show that the overall training time of the proposed algorithm is only 1.98 times of the average training time of the prototype algorithm, the training efficiency is improved, and the cost of time is reduced. Besides, under the condition of satisfy real time flight, the proposed algorithm can meet the demand of UAV track quality, and provides a new idea for promoting the practical application of deep reinforcement learning in flight track planning.

Key words: deep deterministic policy gradient algorithm; unmanned aerial vehicle; track planning; deep reinforcement learning; artificial bee colony algorithm

航迹规划是无人机(unmanned aerial vehicle,UAV)完成电子对抗作战任务的有效技术手段。面对地形及敌方雷达威胁,UAV飞行时亟需合理的规划算法获取航迹以规避危险并完成任务。实际飞行过程存在未知动态威胁,更要求UAV具备实时决策能力[1],因此在未知威胁环境如何实时规划UAV航迹是亟待解决的难题。

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项目(CEMEE2021G0001)。
作者简介:高敬鹏(1980—),男,博士后,硕士生导师,研究方向为认知电子战与航迹规划,E-mail:gaojingpeng@hrbeu.edu.cn;胡欣瑜(1998—),女,硕士研究生,研究方向为多智能体与航迹规划;江志烨(1977—),男,研究员,研究方向为电子对抗。
收稿日期:2021-06-03 **修回日期:**2021-08-18 **文章编号:**1002-8331(2022)08-0264-09

群智能算法是当前规划航迹的主要手段，结合约束条件，设计目标函数，利用迭代技术解算最优航迹。文献[2]提出一种自适应遗传算法实现UAV低空三维航迹规划，可以有效适用于静态地形威胁环境，然而其忽略了未知威胁对实际飞行过程的影响。文献[3]提出一种基于改进蚁群的UAV三维航迹重规划算法，相较其他算法，减少了规划时间，然而随着威胁数目增多，算法迭代计算复杂度升高，处理速度下降，难以满足无人机飞行航迹实时控制的需求。另外，若以离散航点两两连接形成的直线段为航迹，无人机在航点切换处飞行，不符合自身飞行动力学原理，将导致飞行误差，故在航迹规划的基础上，利用航迹优化技术将离散航点优化为一条满足无人机运动约束的飞行航迹[4]。文献[5]利用改进A*算法完成离散航迹点的规划，并通过插值平均处理优化航迹，却也增大了解算航迹的时间成本。文献[6]提出一种改进RRT航迹规划算法，在得到航迹节点的基础上，采用B样条曲线平滑方法生成曲率连续的航迹，也造成整体耗时增多。虽然传统以及基于群智能优化的航迹规划算法均能够获得最优航迹，但依赖于航迹优化技术配合且解算目标函数速度慢加大了实时规划难度。因此现阶段选择高效算法对于实现UAV航迹实时规划尤为重要。

近年来，随着机器学习的发展，深度强化学习因其出色的泛化性和适配性被成功应用于规划领域[7]。2013年，DeepMind团队[8]提出基于深度Q网络(deep Q-network，DQN)的深度强化学习(deep reinforcement learning，DRL)方法，利用神经网络拟合Q值函数，能够解决高维状态空间的离散动作决策问题。文献[9]设计一种改进DQN算法，在三维空间规划移动机器人路径，控制智能体输出离散动作，但无人机实际飞行是需要连续精准控制的，故其方法无法拓展至航迹规划领域。2015年，Lillicrap等人[10]提出基于连续控制模型的深度确定性策略梯度(deep deterministic policy gradient，DDPG)算法，使智能体能在复杂环境根据自身状态决策输出连续动作。文献[11]利用DDPG算法决策无人机机动着陆的连续动作，这与航迹规划中无人机连续飞行需求不谋而合，故DDPG算法可用于无人机航迹规划。然而DDPG算法收敛性能受网络权重参数影响较大[12]，适配网络参数及优化模型将导致训练耗时长。文献[13]提出混合噪声优化DDPG算法实现无人机对机动目标的连续跟踪，DDPG算法收敛性能得以提升，但仍存在训练耗时长的弊端。因此实际应用中如何降低网络训练时间成本成为DDPG算法仍待解决的问题。

为解决在未知威胁环境无人机难以实时规划航迹且模型训练机制冗余的问题，本文提出一种改进DDPG无人机航迹规划算法。结合实际环境，搭建飞行场景模型，将DRL方法引入航迹规划领域，根据任务和飞行需求，设计状态空间、动作空间和奖励函数，利用人工蜂群改进DDPG算法，更新网络模型参数，训练并应用改进DDPG网络模型，实现无人机航迹实时规划。

1　无人机航迹规划系统模型

为完成无人机航迹实时控制，并提升DDPG算法训练效率，本文提出改进DDPG无人机航迹规划算法，其系统模型如图1所示。首先，构建环境空间，包括静态地形以及雷达探测威胁。其次，设计航迹规划问题的强化学习要素，根据无人机运动模型设计状态空间，依据飞行动力学理论设计动作空间，结合非稀疏化思想，考虑无人机与环境的交互情况设计奖励函数。随后，结合所设计的强化学习要素，构成经验数组，利用人工蜂群算法，优化DDPG算法网络参数更新机制，训练改进DDPG无人机航迹规划网络模型。最后，应用改进DDPG算法，实现无人机从实时飞行状态到实时飞行动作的决策映射，形成航迹。

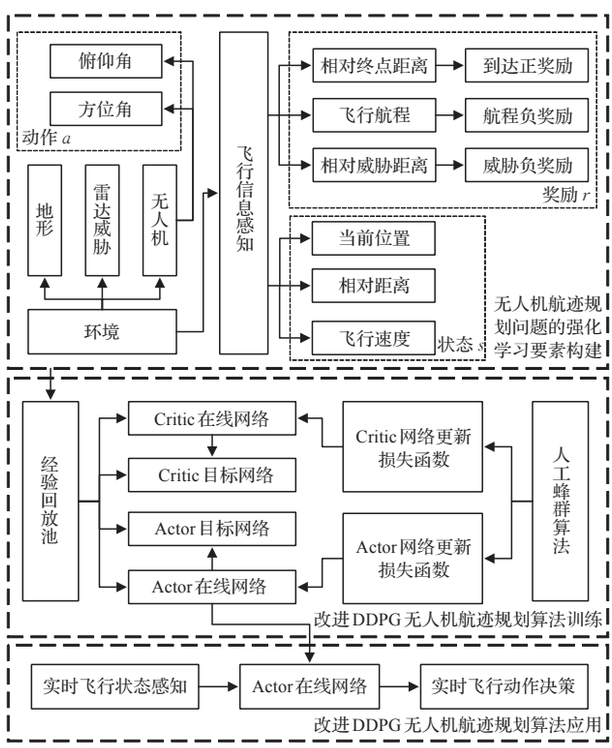


图1　无人机航迹规划系统模型

Fig.1　Model of UAV track planning system

2　强化学习与航迹规划

无人机与环境发生交互得到飞行动作的航迹规划过程可以视为序列决策过程，使用马尔科夫决策过程可以对其建模，利用强化学习算法能够对其求解。

2.1　马尔科夫决策过程模型

马尔科夫决策过程中每个 t 时刻状态的变化都只与 $t-1$ 时刻状态和动作有关，与 $t-1$ 时刻之前的状态和动作无关，其定义为一个四元组集合：

$$M=(S,A,\boldsymbol{P},R) \tag{1}$$

式中，S 表示智能体在环境中的所有状态集合，A 表示智能体在对应状态下可执行的动作集合，$\boldsymbol{P}$ 表示智能体的状态转移概率矩阵，R 表示智能体得到的奖励回报集合，$r_t(s_t,a_t,s_{t+1})\in R$ 表示智能体通过动作 a_t，从状态 s_t 转移至状态 s_{t+1} 获得奖励回报值。

2.2 无人机飞行环境设计

为更好地模拟无人机实际飞行，本节设定规划空间，搭建空间中静态地形和雷达威胁模型，将其作为无人机执行任务应考虑的威胁因素，为无人机飞行构建环境基础。

2.2.1 规划空间

在规划空间中，无人机以原点为起点，依据实时规划的航迹，避开地形威胁和雷达探测威胁，到达任务目的地。设定无人机在三维飞行空间的位置坐标 (x,y,z)，x 和 y 分别表示无人机在经纬方向的坐标点，z 表示其在空间的海拔高度，则无人机的三维规划空间数学模型 C 可表示为：

$$C=\{(x,y,z)|x\in[-x_m,x_m],y\in[-y_m,y_m],z\in[z_{\min},z_{\max}]\} \tag{2}$$

式中，x_m 和 y_m 分别为无人机在经纬方向最大飞行范围，$z_{\min}$ 和 $z_{\max}$ 分别为其在空间中最小和最大飞行高度。

2.2.2 地形和雷达威胁

考虑到无人机实际飞行环境存在地形威胁和未知位置雷达探测威胁，所以需要模拟静态地形以及不同位置的雷达威胁数学模型。静态地形模型可表示为：

$$H(x,y)=\sin(y+\upsilon)+\kappa\sin x+\chi\cos\left(\delta\sqrt{x^2+y^2}\right) \tag{3}$$

式中，$H(x,y)$ 为地形起伏高度，x 和 y 表示地面水平方向的点坐标，υ、κ、χ、δ 是模型的常系数，通过改变这些系数数值大小即能模拟起伏地貌的实际地形。

威胁辐射源的探测范围决定了其对无人机的威胁程度，常用的方法通过计算威胁高度数据，将其等效为地形模型[14]。雷达对不同距离的目标有不同的探测能力，因而在建立雷达威胁模型时，应将雷达与目标间距离 D 和检测概率 P_d 纳入考虑范围。基于此，本文结合雷达原理，依据文献[15]推导目标和雷达间任意距离与检测概率的关系 $P_d(D)$ 为：

$$P_d(D)=\exp\left(\frac{-(D/D_{\max})^4\ln P_f}{\ln P_f-(D/D_{\max})^4+1}\right) \tag{4}$$

式中，$D_{\max}$ 表示雷达最大探测距离，P_f 表示虚警概率。

利用上述将威胁源等效为地形模型的方法，把雷达威胁范围处理为地形高程数据后数学表达式为：

$$H_{\text{radar}}(x,y)=K_r\left(D_{\max}^2-(x-x_0)^2-(y-y_0)^2\right) \tag{5}$$

式中，$H_{\text{radar}}(x,y)$ 为整合后的雷达威胁高程，K_r 表示与雷达相关的性能系数，$D_{\max}$ 为雷达的最大作用半径，(x_0,y_0) 为雷达中心坐标。最后，将静态地形和雷达威胁模型叠加后得：

$$H'(x,y)=P_d(D)\cdot H_{\text{radar}}(x,y)+H(x,y) \tag{6}$$

式中，$H'(x,y)$ 表示整体高程数据。

2.3 航迹规划问题的强化学习要素设计

无人机航迹规划问题的强化学习基本要素主要体现为其在飞行空间的状态，由一个状态转换到下一状态对应的动作以及执行动作后与环境交互所得奖励。

2.3.1 状态空间

无人机在飞行时，应具有实时感知环境信息并决策航迹的能力，从而避开地形和未知雷达威胁。考虑到以上需求，利用无人机能够根据传感器和情报等途径获取飞行信息的特点，本文设计无人机当前位置、相对威胁距离和飞行速度方向三方面信息为状态，将其在任意时刻状态信息联合，用公式表示为：

$$\begin{aligned}s_t&=(p_{u,t},p_{t,t}-p_{u,t},v_{u,t})=\\&([x_{u,t},y_{u,t},h_{u,t}],[dx_t,dy_t,dh_t],[v_{x,t},v_{y,t},v_{z,t}])\end{aligned} \tag{7}$$

式中，$p_{u,t}$ 和 $p_{t,t}$ 分别为终点和无人机位置，$v_{u,t}$ 为无人机速度，$[x_{u,t},y_{u,t},z_{u,t}]$ 为 t 时刻无人机在飞行空间的坐标位置，$[dx_t,dy_t,dz_t]$ 为无人机和终点的相对距离，$[v_{x,t},v_{y,t},v_{z,t}]$ 为无人机飞行时三个方向的分速度。

2.3.2 动作空间

从无人机飞行动力学角度出发，为避开地形和雷达威胁并安全到达终点，其需要在飞行时改变速度方向。本文设定无人机按照恒定速率飞行，因而调整其飞行角度即可改变速度方向，并规定飞行角度精度，以期形成平滑的航迹，满足飞行动力学要求。所以将其在任意时刻的动作信息联合，用公式表示为：

$$a_t=(\varphi_t,\vartheta_t) \tag{8}$$

式中，φ_t 和 ϑ_t 分别表示无人机飞行的方向角和俯仰角。

2.3.3 奖励函数

强化学习算法的收敛性依赖于合理的奖励设置，本文结合非稀疏思想设计奖励函数，使无人机执行每一步到达终点的趋势更加明显。无人机在规划空间内飞行的首要目的是到达任务终点，其航程受到自身携带燃料限制，同时飞行过程要避免被雷达探测，因此本文奖励函数的设计主要考虑以下3个方面。

(1)到达正奖励 r_{appr}。无人机航迹规划的首要任务是成功到达任务目的地，因而当任务终点在无人机的探测范围内时，系统反馈正奖励以使到达趋势更加明显，具体表示为：

$$r_{\text{appr}}=\begin{cases}2-\left\|\hat{N}(p_{t,t}-p_{u,t})\right\|, \left\|p_{t,t}-p_{u,t}\right\|<\rho_{\max}\\-\left\|\hat{N}(p_{t,t}-p_{u,t})\right\|, \left\|p_{t,t}-p_{u,t}\right\|\geqslant\rho_{\max}\end{cases} \tag{9}$$

式中，$\hat{N}(\cdot)$ 表示归一化，$\|\cdot\|$ 表示取模长，$\rho_{\max}$ 为无人机最大探测距离。

(2)航程负奖励 r_{path}。实际飞行时，无人机飞行航程受到燃料等能源限制，所以设置航程负奖励 r_{path}，使无人机经历越短的航程便能到达终点，具体表示为：

$$r_{\text{path}}=\begin{cases}-\left\|\hat{N}(d)\right\|, d\geqslant L_{\max}\\ -2, d<L_{\max}\end{cases} \tag{10}$$

式中，d 表示无人机已经飞过的航程，$L_{\max}$ 表示无人机携带燃料对应的最大飞行航程。

(3)威胁负奖励 r_{threat}。依据前文建立的威胁模型，若无人机进入雷达威胁区域则视为被敌方雷达发现，因此设置威胁负奖励 r_{threat}，以降低无人机进入雷达探测区域的概率，具体表示为：

$$r_{\text{threat}}=\begin{cases}\left\|\hat{N}(p_{u,t}-p_{r,t})\right\|-1, \left\|p_{u,t}-p_{r,t}\right\|\geqslant D_{r,\max}\\ -1, \left\|p_{u,t}-p_{r,t}\right\|<D_{r,\max}\end{cases} \tag{11}$$

式中，$p_{r,t}$ 表示雷达位置坐标，$D_{r,\max}$ 表示雷达最大探测距离。

将任意时刻奖励综合表示为：

$$r_t=r_{\text{appr}}+r_{\text{path}}+r_{\text{threat}} \tag{12}$$

综上所述，本文结合无人机实际飞行需求，设计基于航迹规划问题的强化学习基本要素，为构建网络训练经验集奠定基础。

2.4　DDPG与航迹规划

在众多强化学习算法中，DDPG算法因其能在连续动作空间确定性选择唯一动作的优点受到青睐。又由前文设计的强化学习基本要素可知，航迹规划问题是基于高维状态空间以及连续动作决策的，因此采用DDPG算法可以很好地完成无人机航迹决策。

DDPG网络中包含Actor策略网络和Critic值函数网络。Actor网络用来拟合策略函数，进而提取可执行的动作，其网络权重参数为 θ，输入为状态 s_t，输出为动作 a_t；Critic网络通过内部的值函数信息估计Actor策略网络中对应梯度更新的方向，其网络权重参数为 ω，输入为状态 s_t 和动作 a_t，输出为评估值 Q。

Actor网络更新采用策略梯度下降法，具体表示为：

$$\nabla_{\theta}J(\theta)=\frac{1}{m}\sum_{i}\nabla_{a_i}Q(s_i,a_i|\omega)\nabla_{\theta}\mu(s_i|\theta) \tag{13}$$

式中，m 为经验数据 (s,a,r,s') 的采样个数。Critic网络采用均方误差损失函数进行参数更新：

$$Loss=\frac{1}{m}\sum_{i}(r_i+\gamma Q'(s_{i+1},a_{i+1}|\omega')-Q(s_i,a_i|\omega))^2 \tag{14}$$

式中，γ 为奖励折扣因子。

另外，DDPG算法分别复制Actor策略网络和Critic值函数网络作为目标网络，使智能体对任务策略进行稳定学习，其网络权重参数分别表示为 θ' 和 ω'。结合软迭代思想，缓慢更新目标网络，使智能体在训练时，学习过程稳定性大幅度增强。Actor目标网络具体更新方式为：

$$\theta'\leftarrow\tau\theta+(1-\tau)\theta' \tag{15}$$

式中，τ 用来控制Actor目标网络权重 θ' 的更新速度。同样，利用式(15)的方式更新Critic目标网络参数 ω'。

此外，DDPG算法利用随机噪声，增加Actor策略网络在连续动作空间的探索能力，形成策略映射 μ'：

$$\mu'(s_t)=\mu(s_t|\theta)+N \tag{16}$$

式中，N 为该噪声随机过程。

本文设计Actor策略网络和Critic值函数网络均由两个全连接层FC构成，网络结构简单且运算方便，时间复杂度低。故结合Actor网络输入状态，输出动作，Critic网络输入状态和动作，输出 Q 值的特点，根据上文选定的9维状态和2维动作，设计DDPG网络结构如表1所示。表中ReLu和tanh为神经网络常用的两种非线性激活函数。

表1　DDPG网络结构

Table 1　Network structure of DDPG

Actor策略网络	Critic值函数网络
2*(FC(9,300,ReLu)→	2*(FC(11,300,ReLu)→
FC(300,2,tanh))→输出	FC(300,1,None))→输出

依据DDPG网络训练原理，采用表1设计的网络结构，根据式(13)至式(16)，训练DDPG网络。训练完成后，获取从飞行状态到飞行动作端到端的决策映射，其Actor在线网络策略映射公式如下：

$$a_t=\boldsymbol{\mu}_{\theta}(s_t) \tag{17}$$

式中，$\boldsymbol{\mu}_{\theta}(\cdot)$ 为已训练Actor在线网络的策略映射关系，θ 是其网络权重参数，s_t 为无人机实时飞行状态，a_t 即为由映射关系 $\boldsymbol{\mu}_{\theta}(\cdot)$ 得到的实时飞行动作。

在实际应用中，无人机实时采集飞行状态，迁移已训练Actor在线网络，即可得到实时飞行动作，实现航迹规划。

3　基于改进DDPG的无人机航迹规划算法

DDPG网络训练过程中，学习率的改变会直接影响网络收敛性能，传统方法通过调试学习率，直至网络具有较好的收敛效果，但调整至合适的学习率将会耗费大量时间成本。群智能算法通过不断迭代更新求解适应度函数最优值的思想，与神经网络优化权重参数的思想异曲同工，因此结合群智能算法寻优DDPG网络权重参数能够避免学习率对网络收敛性能的影响，最终解决网络训练时间长的问题。

3.1　改进人工蜂群算法

人工蜂群(artificial bee colony，ABC)算法具备寻优能力强以及收敛速度快等优点，故本文采用ABC算法优化DDPG网络更新机制。但直接采用ABC算法需在一次完整DDPG网络训练中，利用不同的蜂群寻优策

略和值函数两类网络的最佳更新方式,必然导致计算冗余。为弥补该缺陷,本文设计一种二维人工蜂群(two dimensional artificial bee colony,2D-ABC)算法,改进初始解和位置更新公式,共享种群行为机制,减少计算复杂度,提升训练效率。

2D-ABC算法将蜂群分为二维开采蜂、二维随从蜂和二维侦察蜂,二维蜜源每一维位置分别对应两个优化问题可能解,每一维蜜源花粉量分别对应两个解的适应度。二维蜂群采蜜的行为机制有以下三种,

(1)初始化种群。蜜蜂群体派出 SN 个二维开采蜂,开采蜂和随从蜂各占蜂群总数的一半,蜜源数与开采蜂相同,依据式(18)随机产生 SN 个二维初始解:

$$\boldsymbol{x}_{i,k}^{j}=\boldsymbol{x}_{\min,k}^{j}+\varepsilon\times(\boldsymbol{x}_{\max,k}^{j}-\boldsymbol{x}_{\min,k}^{j}) \tag{18}$$

式中,$\boldsymbol{x}_{i,k}^{j}$ 是初始化时第 k 维蜜源 i 的第 j 维向量,$i=1,2,\cdots,SN$, $j=1,2,\cdots,M$, $k=1,2,M$ 是解的个数,$\boldsymbol{x}_{\max,k}^{j}$ 和 $\boldsymbol{x}_{\min,k}^{j}$ 分别为第 k 维蜜源中第 j 维向量的极大值和极小值,ε 为区间 $[0,1]$ 的随机数。

(2)开采蜂工作。每一维开采蜂从对应维原蜜源的位置 $\boldsymbol{x}_{i,k}^{j}$ 产生一个新的蜜源位置 $\boldsymbol{v}_{i,k}^{j}$,并与该维蜜源初始花蜜量比较,若更优则将该维新蜜源作为该维的标记蜜源,否则以初始蜜源为标记蜜源,同时与对应维度随从蜂共享蜜源信息。二维蜜源的位置更新公式为:

$$\boldsymbol{v}_{i,k}^{j}=\boldsymbol{x}_{i,k}^{j}+\phi_k\times(\boldsymbol{x}_{i,k}^{j}-\boldsymbol{x}_{l,k}^{j}) \tag{19}$$

式中,ϕ_k 为第 k 维中区间 $[0,1]$ 的随机数。

(3)随从蜂采蜜。二维随从蜂根据对应维蜜源花蜜量的多少选择每一维较优的标记蜜源,并在其附近按照式(20)探索第 k 维新蜜源,选择概率 P_i^k 表达式为:

$$P_i^k=\frac{fit_i^k}{\sum_{n=1}^{SN}fit_n^k} \tag{20}$$

式中,fit_i^k 为第 k 维解 $\boldsymbol{x}_i^k$ 的适应度值,即花蜜量。

本文提出2D-ABC算法流程如图2所示,其改进DDPG算法具体步骤描述如下:

步骤1 初始化二维蜜源和二维蜂群。根据式(18),在 M 维空间随机初始化 SN 个二维蜜源位置,第一维和第二维蜜源位置分别代表Actor在线网络和Critic在线网络权重参数。同时,设置开采蜂和随从蜂数目均为 $2\times SN$,第一维和第二维蜂群的工作对象分别为第一维和第二维蜜源。

步骤2 计算二维适应度。将Critic在线网络权重更新的均方误差损失函数,即式(14)作为第一维适应度函数,得到第一维蜜源评价值;将Actor在线网络权重更新的策略梯度下降函数,即式(13)作为第二维适应度函数,得到第二维蜜源评价值。

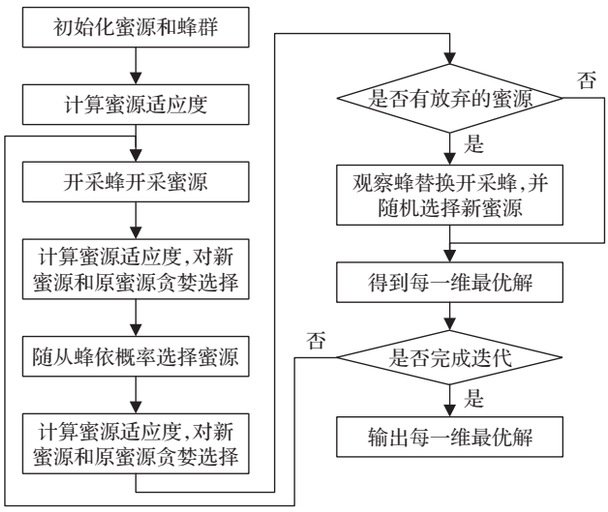


图2　2D-ABC算法流程图

Fig.2　Flow chart of 2D-ABC algorithm

步骤3 二维开采蜂开采蜜源。根据式(19),开采蜂分别在每一维蜜源位置附近开采,获得新蜜源位置。

步骤4 根据式(13)和式(14),再次分别计算每一维新位置蜜源评价值,并与原位置蜜源评价值相比较,进行贪婪选择,保留更优的二维蜜源。

步骤5 随从蜂选择蜜源。二维随从蜂依据式(20)得到的概率,选择每一维新蜜源。

步骤6 再次执行步骤4。

步骤7 在 $Limit$ 次蜜源位置更新后,若每一维有放弃的蜜源则利用观察蜂替换开采蜂,并随机选择新蜜源,若无则从已保留的优质蜜源得到每一维最优蜜源位置,即最优的Actor网络和Critic网络权重参数。

3.2　改进DDPG算法模型训练及应用

本文融合2D-ABC算法寻优与DDPG算法模型更新机制,将Actor在线网络权重更新的策略梯度下降函数和Critic在线网络权重更新的均方误差损失函数作为适应度函数,利用2D-ABC算法分别寻优每一回合DDPG算法Actor和Critic在线网络权重参数,完成改进DDPG算法模型的训练,从而提升网络训练效率,降低总体的训练时间成本。改进DDPG算法模型训练及应用结构框图如图3所示,具体训练步骤如下:

步骤1 结合式(7)至式(12),设计航迹规划问题的强化学习要素。

步骤2 初始化状态 s,清空经验回放池。

步骤3 根据状态 s,Actor在线网络得到对应动作 a,智能体执行动作 a,并得到新状态 s' 以及与环境交互后的奖励 r。

步骤4 将经验数组存入经验回放池,并从经验回放池中采样 m 个经验数组,送入Critic值函数网络,计算得在线 Q 值 $Q_\omega(s,a)$ 和目标 Q 值 $Q_{\omega'}(s',a')$。

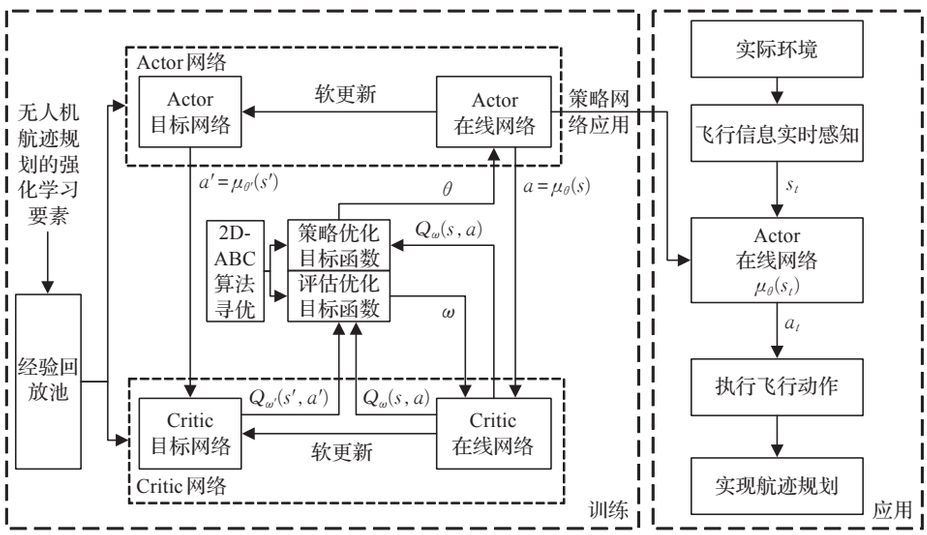


图3　改进DDPG算法模型训练及应用结构框图

Fig.3　Training and application structure diagram of improved DDPG algorithm model

步骤5　根据式(13)和式(14)，结合Critic值函数网络的在线 Q 值和目标 Q 值，利用2D-ABC算法求得最优Actor网络权重参数和最优Critic网络权重参数。

步骤6　根据式(15)，通过软迭代更新Actor网络以及Critic网络权重参数。

步骤7　判断是否满足DDPG网络训练结束条件，结束训练。

最后，与改进前方法相同，无人机实时采集飞行状态，根据式(17)，获取该状态下的决策映射，执行飞行动作，实现航迹规划。

4　仿真与分析

对本文提出的改进DDPG无人机航迹规划算法进行仿真分析，无人机飞行约束参数、相关威胁仿真参数和改进DDPG算法参数分别如表2、表3和表4所示。本文设定无人机航迹规划空间大小为15 km×15 km×7.5 km，且假设无人机飞行恒定速率，同时设置算法测试500次，另外忽略自然环境干扰因素影响。本文涉及仿真的实验设备及环境满足：Intel® Core™ i7-9700k CPU，32 GB双通道内存，Windows 10 64位操作系统，Python 3.5，TensorFlow 1.7.0。

表2　无人机飞行约束参数

Table 2　Fight constraint parameters of UAV

参数	值
飞行速度/(m/s)	300
方向角范围/(°)	[0，180]
俯仰角范围/(°)	[−90，90]
飞行高度范围/km	[0.5，8]
东西方向飞行范围/km	[−7.5，7.5]
南北方向飞行范围/km	[−7.5，7.5]
最大飞行航程/km	25

表3　相关威胁仿真参数

Table 3　Simulation parameters of related threat

模型	参数	值
静态地形	常系数 υ	0.5
	常系数 κ	0.1
	常系数 χ	1
	常系数 δ	0.005
雷达威胁	虚警概率 P_{fa}	10^{-6}
	性能系数 K_r	0.25
	最大作用距离 $D_{\max}$ /km	15

表4　改进DDPG算法参数

Table 4　Parameters of improved DDPG

参数	DDPG	参数	ABC
经验回放池大小	3×10^5	迭代次数	500
奖励折扣因子	0.99	蜂群数量	100
替换因子	5×10^{-3}	蜜源个数	50
批量大小	128	单个蜜源最大搜索次数	100
训练回合数	10 000		

为验证改进DDPG算法有效性和在未知环境的适应性，本文选取网络训练时长、测试成功率和航迹偏差率为评估指标，评估算法的训练和测试结果。其中，网络训练时长用于评估算法训练效率，测试成功率用于评估无人机满足航程约束情况下依照航迹决策顺利达到终点的能力，其计算公式为：

$$\text{成功率}=\frac{\text{不碰撞到威胁且成功到达的次数}}{\text{总测试次数}} \tag{21}$$

航迹偏差率 TE 用于评估无人机在成功到达终点前提下的航迹质量，其计算公式为：

$$TE=\frac{1}{F}\sum_{i=1}^{F}\frac{|\alpha_i-\beta_i|}{\alpha_i} \tag{22}$$

式中，F 为测试次数，α_i 和 β_i 分别为设定相同条件下

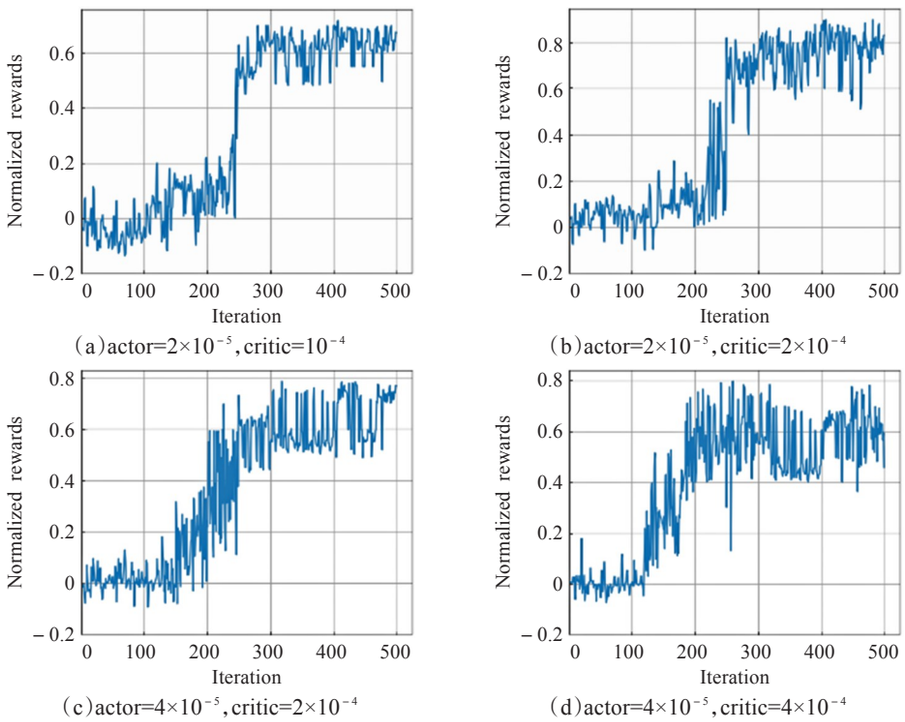


图4　四种不同学习率情况下DDPG算法的奖励收敛曲线

Fig.4　Reward convergence curve of DDPG under four different learning rates

用智能算法解算得第 i 条航迹长度和改进DDPG算法决策得第 i 条航迹长度，航迹偏差率越低航迹质量越高，本文设定航迹偏差率低于7.5%时航迹质量达标。

由于训练次数多，且算法随机波动较大，直接显示所有训练回合奖励收敛曲线效果不佳，为更好展示算法训练效果，本文将每20个训练回合所得奖励和取平均并作归一化处理，将10 000次迭代收敛曲线等效处理为500次迭代收敛曲线。图4和表5分别给出了在网络结构设置如表1，超参数设置如表4，设定4组不同Actor网络和Critic网络学习率情况下，DDPG算法的奖励收敛曲线和训练时长表。

表5　四种不同学习率情况下DDPG网络训练时长

Table 5　Network training duration under four different learning rates

Actor网络学习率/10^{-5}	Critic网络学习率/10^{-4}	训练时长	平均训练时长
2	1	36 h 15 min 54 s	35 h 8 min 41 s
2	2	35 h 22 min 17 s	
4	2	34 h 46 min 29 s	
4	4	34 h 10 min 4 s	

由图4可知，随着学习率的增大，DDPG算法收敛速度明显加快，当Actor网络和Critic网络学习率分别为2×10^{-5}和10^{-4}时，归一化奖励值在5 600次训练回合左右才趋于稳定，而当Actor网络和Critic网络学习率分别为4×10^{-5}和4×10^{-4}时，归一化奖励值在3 800次训练回合左右即逐渐收敛。另外，不同学习率情况下，归一化奖励最终收敛值也不同，当Actor网络和Critic网络学习率分别为2×10^{-5}和2×10^{-4}时，归一化奖励值在0.8上下波动，而当Actor网络和Critic网络学习率分别为4×10^{-5}和4×10^{-4}时，归一化奖励值在0.6上下浮动，且浮动幅度较大。这是因为学习率是强化学习算法学习能力的数值体现，过高会导致算法早期样本过拟合，过低会导致样本利用率低使算法收敛慢，因此降低学习率对网络性能的影响尤为重要。

由表5可知，仅调试4组学习率情况下网络总训练时间累计140 h 34 min 44 s，训练耗时长，而调整至合适的学习率需要大量的训练时间，本文提出改进DDPG算法优化网络更新机制，提升算法训练效率。

图5给出了网络结构和参数设置分别如表1和表4情况下改进DDPG算法归一化奖励收敛曲线。

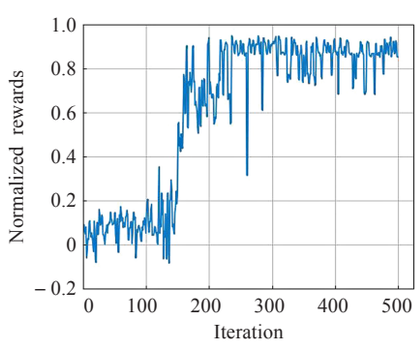
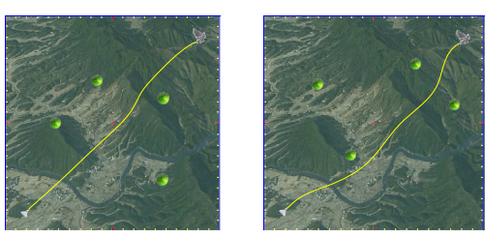


图5　改进DDPG算法奖励收敛曲线

Fig.5　Reward convergence curve of improved DDPG

由图5可知，改进DDPG算法归一化奖励值在4 400次训练回合左右即趋于收敛，且稳定在0.9左右。另外，记录其经历10 000次训练回合耗费时长为69 h 40 min 34 s，对比表5结果可知，改进DDPG算法整体训练时长仅为原算法在表5所设4组学习率情况下平均训练时长的1.98倍。这是因为所提算法每个训练回合内利用ABC算法迭代更新寻优网络参数，导致网络训练复杂度增加，引起单个训练回合耗时增长的代价。得益于改进DDPG算法网络训练不依赖于学习率的优势，仅一次训练就能完成对模型权重参数的寻优，因此总体上网络训练时长大幅度减少，所提算法具有一定的有效性。

图6给出了在无人机仿真参数设置如表2，威胁模型仿真参数设置如表3的情况下，在两种随机位置多雷达环境中，无人机利用改进DDPG算法航迹规划测试效果图。



(a)随机位置雷达环境1　(b)随机位置雷达环境2

图6　改进DDPG算法航迹规划效果图

Fig.6　Track planning effect chart by using improved DDPG

由图6可知，无人机能以连续平滑的航迹飞行，有效避开实际环境地形和不同位置未知雷达探测威胁，成功到达任务终点，验证了所提算法应用的可行性。

尽管智能算法解算航迹速率慢导致测试成功率不尽如人意，但迭代计算的特点决定了其能在不限时间内得到更优航迹。本文以智能算法在测试回合内解得航迹为参照，用航迹偏差率评估改进DDPG算法每次测试形成航迹的质量。蚁群算法具有启发式概率搜索特点，易于找到全局最优解，在规划领域广泛应用，因此选择蚁群算法作为对比算法。表6给出在相同飞行环境内无人机利用改进DDPG算法进行航迹决策和用蚁群算法解算航迹的测试结果对比。其中蚁群算法种群数量为40，全局信息素浓度更新率为0.5，局部信息素浓度更新率为0.4，信息素浓度重要程度因子为1.5，启发值重要程度因子为5。

表6　不同算法航迹规划测试结果

Table 6　Test results of different algorithms for track planning　%

算法	测试成功率	航迹偏差率
改进DDPG算法	97.2	3.78
蚁群算法	48.2	无

由表6可知，500次测试中，用蚁群算法解算航迹无人机测试成功率仅48.2%，而改进DDPG算法成功率高达97.2%。这是由于大量的训练增强了改进DDPG算法学习能力，能够实时决策无人机飞行航迹，获得较高飞行成功率。同时，以蚁群算法获得最优航迹为参照，改进DDPG算法所得航迹偏差率仅为3.78%，其原因是所提算法采取的航迹决策使无人机飞行航迹有效且平滑，形成的航迹满足航迹质量需求，进一步验证了所提算法在工程应用的可行性。

5　结语

本文提出一种改进DDPG无人机航迹规划算法，解决了用传统算法解算航迹速度慢的问题，同时优化了DDPG网络权重参数更新过程。所提算法将深度强化学习应用于航迹规划领域，为无人机飞行提供连续确定性动作决策，并设计2D-ABC算法，改进DDPG算法模型更新机制。仿真结果表明，所提算法无需调整学习率的过程，提升了无人机在未知威胁环境飞行的实时反应能力，降低了训练的时间成本，且在达到97.2%飞行成功率前提下，保证了航迹质量。忽略自然干扰因素影响，所提算法相比典型智能算法，凭借连续飞行动作输出和实时航迹决策的优势，在无人机航迹规划领域更具可行性。面对实际环境天气、风力和气流等变化影响，可联合卡尔曼滤波等技术完善飞行动作，使得所提算法在自然环境应用可行。下一步工作，本团队将研究所提算法的优化技术，同时探讨超参数对于深度强化学习网络模型性能的影响。

参考文献：

[1] 朱杰，鲁艺，张辉明.突发威胁情况下的无人机航迹重规划[J].计算机工程与应用，2018，54(8)：255-259.
ZHU J，LU Y，ZHANG H M.Path replanning for UAV in emergent threats[J].Computer Engineering and Applications，2018，54(8)：255-259.

[2] 任鹏，高晓光.基于NAPPGA算法的无人机低空突防航迹规划[J].计算机仿真，2014，31(4)：102-105.
REN P，GAO X G.Flight path planning for UAV low-altitude penetration based on niche adaptive pseudo parallel genetic algorithm[J].Computer Simulation，2014，31(4)：102-105.

[3] 唐必伟，朱战霞，方群，等.基于改进蚁群算法的无人驾驶飞行器三维航迹规划与重规划[J].西北工业大学学报，2013，31(6)：901-907.
TANG B W，ZHU Z X，FANG Q，et al.Planning and replanning 3D route of UAV using improved ant colony algorithm[J].Journal of Northwestern Polytechnical University，2013，31(6)：901-907.

[4] 贾文涛，李春涛.无人机航迹优化与跟踪技术研究[J].机械

制造与自动化,2020,49(6):156-161.
JIA W T,LI C T.Trajectory optimization of unmanned aerial vehicle and research on its following technology[J].Machine Building & Automation,2020,49(6):156-161.

[5] 李海,郭水林,周晔.融合动态风险图和改进A*算法的动态改航规划[J].航空科学技术,2021,32(5):61-71.
LI H,GUO S L,ZHOU Y.Dynamic diversion planning combining dynamic risk map and improved A* algorithm[J].Aeronautical Science & Technology,2021,32(5):61-71.

[6] 高升,艾剑良,王之豪.混合种群RRT无人机航迹规划方法[J].系统工程与电子技术,2020,42(1):101-107.
GAO S,AI J L,WANG Z H.Mixed population RRT algorithm for UAV path planning[J].Systems Engineering and Electronics,2020,42(1):101-107.

[7] ZHANG W,SONG K,RONG X.Coarse-to-fine UAV target tracking with deep reinforcement learning[J].IEEE Transactions on Automation Science and Engineering,2019,16(4):1522-1530.

[8] VOLODYMYR M,KORAY K,DAVID S,et al.Human-level control through deep reinforcement learning[J].Nature,2015,518(7540):529-533.

[9] 封硕,舒红,谢步庆.基于改进深度强化学习的三维环境路径规划[J].计算机应用与软件,2021,38(1):250-255.
FENG S,SHU H,XIE B Q.3D environment path planning based on improved deep reinforcement learning[J].Computer Applications and Software,2021,38(1):250-255.

[10] LILLICRAP T P,HUNT J J,PRITZEL A.Continuous control with deep reinforcement learning[J].arXiv:1509.02971,2015.

[11] RODRIGUEZ R,ALEJANDRO S,CARLOS B,et al.A deep reinforcement learning strategy for UAV autonomous landing on a moving platform[J].Journal of Intelligent & Robotic Systems,2019,93(1):351-366.

[12] 张耀中,许佳林,姚康佳,等.基于DDPG算法的无人机集群追击任务[J].航空学报,2020,41(10):314-326.
ZHANG Y Z,XU J L,YAO K J,et al.Pursuit missions for UAV swarms based on DDPG algorithm[J].Acta Aeronautica et Astronautica Sinica,2020,41(10):314-326.

[13] LI B,YANG Z P,CHEN D Q,et al.Maneuvering target tracking of UAV based on MN-DDPG and transfer learning[J].Defence Technology,2021,17(2):457-466.

[14] 熊礼阳,汤国安,杨昕,等.面向地貌学本源的数字地形分析研究进展与展望[J].地理学报,2021,76(3):595-611.
XIONG L Y,TANG G A,YANG X,et al.Geomorphology-oriented digital terrain analysis:progress and perspectives[J].Acta Geographica Sinica,2021,76(3):595-611.

[15] TONG X R,Modeling and realization of real time electronic countermeasure simulation system based on SystemVue[J].Defence Technology,2020,16(2):470-486.