doi: 10.7690/bgzdh.2020.09.004

未知环境下基于 PF-DQN 的无人机路径规划

何　金，丁　勇，杨　勇，黄鑫城
（南京航空航天大学自动化学院，南京 211106）

摘要：为解决无人机无模型路径规划的问题，提出一种环境信息未知情况下基于势函数(PF)奖赏的 DQN 路径规划方法。建立无人机在环境中的连续状态空间，将 360°等分成若干个角度作为航向角建立无人机的动作空间，设计目标和障碍物对无人机的势函数奖赏，刻画不同动作对无人机的影响，并进行仿真实验。实验结果表明：PF-DQN 算法能较好地实现无人机在环境信息未知下的无碰撞路径规划，且势函数奖赏能加快无人机路径规划网络的训练速度。

关键词：无人机；路径规划；势函数；深度 Q 网络
中图分类号：TP24　　**文献标志码：**A

UAV Path Planning Based on PF-DQN in Uncertain Environment

He Jin, Ding Yong, Yang Yong, Huang Xincheng
(*College of Automation Engineering*, *Nanjing University of Aeronautics & Astronautics*, *Nanjing* 211106, *China*)

Abstract: In order to solve the problem of no-model path planning for UAV, a DQN path planning method based on potential function (PF) reward in the case of unknown environmental information is proposed. Establish a continuous state space of the drone in the environment, divide the 360° into several angles as the heading angle to establish the action space of the drone, design the target and obstacles to reward the potential function of the drone, and describe the difference more carefully. Carry out the simulation experiments. The results show that the PF-DQN algorithm can better realize the collision-free path planning of the UAV under the unknown environmental information, and the potential function reward can speed up the training speed of the UAV path planning network.

Keywords: unmanned aerial vehicle (UAV); path planning; potential function; deep Q network

0　引言

近年来，随着飞行控制和自主导航技术的不断发展，无人机在军用和民用领域发挥着越来越重要的作用。无人机路径规划是指在一定的约束条件下，从起始点到目标点规划出一条最优或次优的无碰撞路径。随着无人机面临的实际环境日益复杂，规划出一条实用有效的飞行路径是无人机顺利完成各项任务的前提[1]。

从对环境的感知方面看，无人机路径规划大致分为环境信息已知和完全未知 2 种情况。环境信息已知可以理解为无人机在规划路径之前已经完全掌握环境的信息，包括环境中障碍物的位置、大小、形状或其他需要避碰的障碍信息。因为环境的信息提前确定，所以能够规划出最优的飞行路径。经典环境信息已知的路径规划算法有 A*算法、蚁群算法和扩展随机树等，或是它们与其他算法的结合。文献[2]采用双向稀疏 A*算法实现了起始点与目标点共同搜索路径，提高了搜索速度；文献[3]将 A*算法与虚拟力算法结合，先采用 A*算法进行全局预规划，然后在规划过程中构建虚拟力进行局部路径规划，有效解决了路径规划过程中陷入局部最优的问题；文献[4]将蚁群算法信息素的分泌与图形几何学结合，由图形几何构造出势场力，在路径搜索过程中迫使信息素朝着势场力的方向分泌，蚂蚁趋向于搜索更适合的子空间，使得测试模式的搜索空间变得越来越小，最终找到全局最优路径；文献[5]提出了一种称为“closed-loop RRT”的改进型快速扩展随机树算法，通过中间航点及利用可达点进行障碍冲突预测，简化了轨迹生成策略。上述这些算法都属于传统的路径规划算法，需要对特定的环境进行建模，根据实际情况制定不同的模型，所以具有局限性。

环境信息完全未知的路径规划是指在规划路径之前无人机无法预知环境的信息，信息的来源只靠机载的感知系统获取，无人机只能根据感知系统侦测范围内的环境信息进行路径规划，尽可能地规划

收稿日期：2020-05-15；**修回日期：**2020-06-07
作者简介：何　金(1995—)，男，四川人，硕士，从事智能决策研究。E-mail: 18989277346@163.com。

出次优路径或满足约束条件的任意路径，常用的算法有人工势场法、D*算法、导航向量场等。Alexander 将 D*算法与 Pareto 优化理论结合提出 D*-PO 算法，解决多机器人最优路径问题[6]；Liang 等基于 Helmholtz 理论，结合导航向量场构建出 2 维和 3 维二次可微曲线，规划出非常适合固定翼无人机的飞行路线[7]。上述算法虽然可以在环境信息未知情况下使用，但需要对具体环境建立模型，缺乏通用性。

近年来，随着人工智能技术的发展，深度学习和机器学习在无人机领域展现出巨大潜力。在路径规划方面，无人机通过一定的训练，能够自主地规划出最优或次优的路径。文献[8-10]分别采用神经网络、强化学习、DQN(deep Q network)等方法进行无人机的路径规划。这些方法无需对无人机所处的环境进行物理建模，只需通过对无人机所处环境不断地进行离线训练，就可以找出最优、次优或满足约束条件的路径。这些方法中智能体所处的环境一般为离散的栅格地图。由于这种栅格地图所能容纳的状态有限，而且是针对环境已知的情况下进行路径规划，不能满足当今无人机执行任务过程中所处的环境未知且状态连续的情况。

笔者针对无人机无环境模型路径规划问题，提出一种未知环境信息下基于势函数(potential function，PF)奖赏的 DQN(PF-DQN)路径规划方法。通过建立连续状态空间和动作空间，设计一种基于势函数的奖赏机制，而后通过 DQN 算法规划路径，实现无人机的无避碰路径规划。

1 状态空间和动作空间描述

1.1 状态空间

无人机在飞行过程中，目标和障碍物的位置随时可能发生变化，具有不确定性。如图 1 所示，为环境建立笛卡尔坐标系。设无人机在环境中的位置为(x_u, y_u)，目标的位置为(x_a, y_a)，离无人机最近的障碍物的位置为(x_o, y_o)，由图中可以很容易地计算出无人机到目标的距离 d_a、无人机到最近障碍物的距离 d_o、无人机到目标的连线与 x 轴正半轴的夹角 ϕ_a、无人机到最近障碍物的连线与 x 轴正半轴的夹角 ϕ_o。选取 $S=(d_a,\phi_a,d_o,\phi_o)$ 作为无人机在环境中的状态空间，不仅可以表达出无人机在环境中的任意状态，而且具有连续性。

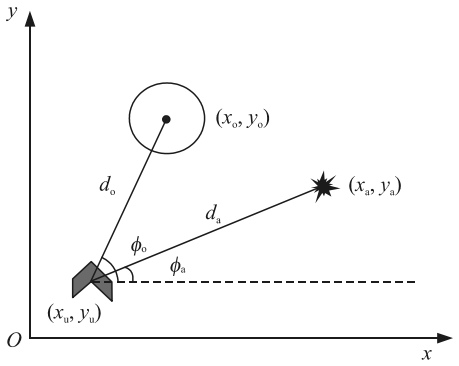


图 1 状态空间

1.2 动作空间

动作空间的定义会影响到无人机路径规划的效果。在 DQN 算法中，Q 估计网络输入为无人机的状态，输出为离散的动作空间 Q 值。理论上，无人机在环境中运动，动作空间可以是任意方向的角度。但过多的动作会导致 DQN 训练时间大大增加，而较少的动作则会使无人机的运动产生“曲折”现象，无法较好地拟合无人机的路径。如图 2 所示，笔者将 360°划分 n 等份，角度间隔 φ=360°/n，动作空间 A={0,1, … ,n-1}。假设 n=16，则 φ=360°/16=22.5°，无人机的航向精度为 22.5°。

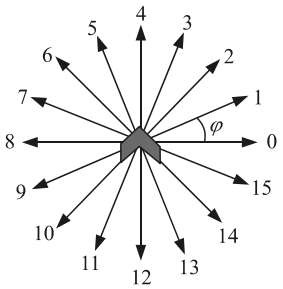


图 2 动作空间

2 基于势函数的奖赏机制

智能体在进行强化学习训练过程中，奖赏函数的设计对 DQN 训练的好坏尤为重要，决定了网络训练的效率与效果。基于势函数奖赏的无人机路径规划算法(PF-DQN)，通过构造目标与障碍物的势函数奖赏对无人机进行启发式路径规划，使得无人机更快地找到目标。

2.1 目标对无人机的势函数奖赏

传统的目标对无人机的奖赏通常定义为，当无人机执行下一动作产生的结果是接近目标时，给出固定的正奖励值；反之，给出固定的负奖励值。对于无人机来说，这种定义的缺点是无法定量地知道此动作对其产生的影响。笔者定义一种目标对无人

机的势函数奖赏 $r_a(k)$为

$$r_a(k)=\frac{d_a^k-d_a^{k+1}}{|d_a^k-d_a^{k+1}|}e^{|d_a^k-d_a^{k+1}|},\quad d_a^k-d_a^{k+1}\in[-\eta,\eta]。\quad(1)$$

其中：d_a^k 为第 k 时刻无人机到目标的距离；d_a^{k+1} 为第 k+1 时刻无人机到目标的距离；η 为无人机的飞行步长。

由式(1)可知：当 $d_a^k>d_a^{k+1}$ 时，无人机执行当前动作后距离目标更近，且 $r_a(k)$为正值，表明环境给予无人机一个正的奖励值，正奖励值的大小与无人机前后时刻距离目标的差值成指数变化；反之，无人机则会得到一个负的奖励值。

2.2　障碍物对无人机的势函数奖赏

无人机在到达目标点的过程中，需要规划出无碰撞的路径，无论是静态障碍物或动态障碍物，无人机都必须与障碍物处在一定距离外。这个距离最大为无人机机载感知系统中用于观测障碍物的传感器观测距离 d_{obs}。无人机在未知环境下进行路径规划时，事先无法预知环境信息，只有当无人机距离障碍物在 d_{obs} 内，才能获得障碍物的位置，这就给无人机避障带来一定的困难。

针对障碍物对无人机的奖励，从无人机传感器的探测范围方面可划分为以下 4 种情况：

1) 第 k 时刻无人机未探测到障碍物，第 k+1 时刻无人机也未探测到障碍物；

2) 第 k 时刻无人机未探测到障碍物，第 k+1 时刻无人机探测到障碍物；

3) 第 k 时刻无人机探测到障碍物，第 k+1 时刻无人机未探测到障碍物；

4) 第 k 时刻无人机探测到障碍物，第 k+1 时刻无人机也探测到障碍物。

图 3 为无人机相对于障碍物的 4 种飞行情况，图中标号 1, 2, 3, 4 对应上述所列的 4 种情况。对第 1 种情况来说，无人机在当前状态下执行某个动作后没有探测到障碍物，设定障碍物对无人机的奖励为无人机的飞行步长值 η；对第 2 种情况来说，由于无人机缺乏环境的预知信息，无法判断出下一时刻是否会靠近障碍物，与无人机执行哪个动作无关，设定障碍物对无人机的奖励为 0；对于第 3 种情况，无人机在当前状态下已经探测到障碍物，在下一时刻却探测不到障碍物，说明无人机远离了障碍物，设定障碍物对无人机的奖励也为步长值 η；对于第 4 种情况，就要定量地给出障碍物对无人机的奖励函数，这里采用和目标对无人机的奖励函数相似的势函数奖赏表示。

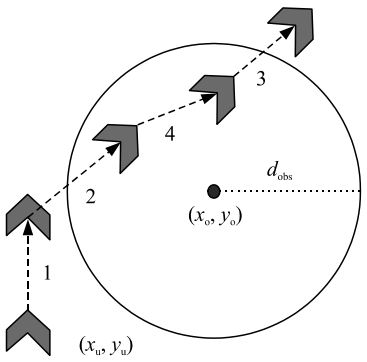


图 3　无人机相对障碍物的飞行

笔者定义此时障碍物对无人机的势函数奖赏

$$r_o(k)=\begin{cases}\eta & d_o^k>d_{obs},d_o^{k+1}>d_{obs}\\ 0 & d_o^k>d_{obs},d_o^{k+1}\leqslant d_{obs}\\ \eta & d_o^k\leqslant d_{obs},d_o^{k+1}>d_{obs}\\ \dfrac{d_o^{k+1}-d_o^k}{|d_o^{k+1}-d_o^k|}e^{|d_o^{k+1}-d_o^k|} & d_o^k\leqslant d_{obs},d_o^{k+1}\leqslant d_{obs}\\ & d_o^{k+1}-d_o^k\in[-\eta,\eta]\end{cases}。\quad(2)$$

其中：d_o^k 为第 k 时刻无人机到障碍物的距离；d_o^{k+1} 为第 k+1 时刻无人机到障碍物的距离；η 为无人机的飞行步长。

2.3　无人机总的势函数奖赏

无人机总的势函数奖赏 r_k 为目标对无人机的势函数奖赏 $r_a(k)$和障碍物对无人机势函数奖赏 $r_o(k)$ 的叠加，即

$$r_k=r_a(k)+r_o(k)。\quad(3)$$

3　基于 PF-DQN 的无人机路径规划

深度强化学习由 DeepMind 团队于 2013 年提出[11]，2015 年该团队提出改进版深度强化学习算法即深度 Q 网络(deep Q network，DQN)模型[12]，在机器学习领域取得了极大的突破。

3.1　Q-learning

Q-learning(Watkins，1989)是强化学习中最早具有突破性进展的时序差分(temporal difference，TD)控制算法，是一种异策略强化学习算法[13]。所谓异策略，即动作策略采用 ε-greedy 策略，而目标策略采用贪婪策略。

Q-learning 的值函数更新公式[14]为：

$$Q(s_k,a_k)\leftarrow Q(s_k,a_k)+\alpha\left[r_k+\gamma\max_{a_{k+1}}Q(s_{k+1},a_{k+1})-Q(s_k,a_k)\right]。\quad(4)$$

式中：$Q(s_k,a_k)$为在状态 s_k 下采取动作 a_k 的 Q 值；α

为学习速率，α 越大，表示 Q-learning 越注重当前的及时回报，反之，之前训练的效果保留得就越多；r_{k+1} 为智能体达到状态 s_{k+1} 所获得的回报；γ 为回报折扣因子；$r_k+\gamma\max_{a_{k+1}}Q(s_{k+1},a_{k+1})$ 为 Q 学习的学习目标。

Q-learning 给定起始状态 s_k 后，根据 ε-greedy 策略在状态 s_k 选择动作 a_k，得到回报 r_k 和下一步状态 s_{k+1}，按照式(4)进行值函数更新，依次循环，智能体不断与环境交互，最终达到所有的 $Q(s,a)$收敛。通过学习训练，最优的策略[15]为：

$$\pi(s)=\arg\max_a Q(s,a)。\tag{5}$$

强化学习的基本思想是使用贝尔曼方程作为迭代更新来评估动作值函数。经证明，Q-learning 的动作值最终会收敛到最优的动作值函数 Q^*，即当迭代个数 $i\to\infty$，$Q_i\to Q^*$。实际上，该方法因为对于每个序列动作值函数都分开估计，没有任何泛化能力。Q 值函数早期通过表格形式表达，对最优策略的求解很有帮助，但也存在先天缺陷。如果所求问题的状态和动作空间非常大或者无穷多，就会产生数学意义上的“维数灾难”，导致无法建立模型或模型很难求解，而深度强化学习正是有效解决该难题的方法[16]。

3.2 Deep Q-network（DQN）

DQN 模型作为一种典型的深度强化学习算法，其算法框图如图 4 所示。由图中可以看出：DQN 算法有 2 个卷积神经网络，即 Q 估计网络和 Q 目标网络。Q 估计网络以具有权重 θ 的卷积神经网络来评估动作值函数 $Q(s,a;\theta_i)\approx Q^*(s,a)$，以解决 Q-learning 中状态-动作表格的“维数灾难”问题[17]。在 DQN 训练过程中，Q 估计网络通过训练迭代 i 次来调整网络权值 θ_i，目的是降低贝尔曼方程的均方误差，其中拟合目标值 $y=r+\gamma\max_{a'}Q(s',a';\theta_i^-)$ 替代最优目标值 $r+\gamma\max_{a'}Q^*(s',a')$，参数 θ_i^- 从先前的训练迭代中获得。在第 i 次迭代中，每一序列的损失函数 $L_i(\theta_i)$都会改变，具体为

$$L_i(\theta_i)=E_{s,a,r,s'}\left[(y-Q(s,a;\theta_i))^2\right]。\tag{6}$$

式中：s 为当前时刻的状态；s'为下一时刻的状态；a 为当前时刻的动作；a'为下一时刻采取的动作。

Q 估计网络的参数采用梯度下降法进行更新，具体为

$$\nabla_{\theta_i}L(\theta_i)=E_{s,a,r,s'}[(r+\gamma\max_{a'}Q(s',a';\theta_i^-)-Q(s,a;\theta_i))\nabla_{\theta_i}Q(s,a;\theta_i)]。\tag{7}$$

DQN 的目标值来自于和 Q 估计网络结构完全相同的 Q 目标网络。Q 目标网络每隔一定步数 C 后，将 Q 估计网络的参数 θ_i 赋给 Q 目标网络的参数 θ_i^-，即 $\theta_i^-=\theta_i$。

DQN 采用 ε-greedy 策略进行动作的选择。在网络训练时，该策略遵循以 ε 的概率随机选择动作，以 1-ε 的概率选择最优动作，最优动作为 Q 估计网络输出的动作空间中动作的最大值参数 $\arg\max_a Q(s,a)$。这正是 DQN 学习中“探索”与“利用”的体现。

此外，DQN 还采用“经验回放”技术，把智能体每一个时间步的经验 $e_t=(s_t,a_t,r_t,s_{t+1})$ 以数据集合 $D=\{e_1,\cdots,e_t\}$ 存储在“经验池”中。在网络训练过程中，从经验池中随机选择最小数据块 $D_{\min}$ 作为样本进行网络训练，减少样本的相关性。通过采用经验回放技术，智能体先前状态的行为分布变得均匀，平滑了学习过程，避免参数波动或产生分歧。

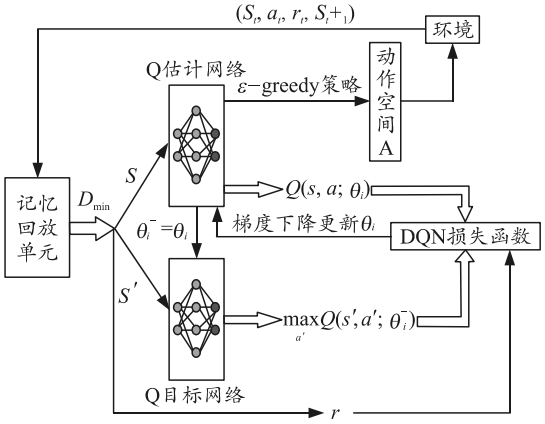


图 4 DQN 算法

3.3 PF-DQN 的无人机路径规划算法流程

状态空间 S、动作空间 A 和势函数奖赏 r_k 是 PF-DQN 算法进行无人机路径规划的 3 大基本要素，无人机将当前的状态、动作和回报以及下一时刻状态存储在经验池中，每次随机抽取最小经验块 $D_{\min}$ 对 Q 估计网络进行训练，使得训练出的 Q 估计网络可以在当前状态下拟合出最优的动作值。当对 Q 估计网络进行足够多的训练，其权值向最佳参数逼近。具体地，PF-DQN 无人机路径规划算法的训练步骤描述如下：

算法 1：PF-DQN 无人机路径规划训练过程。

初始化经验池 D；

建立动作值 Q 估计网络，随机初始化网络的权重 θ；

建立动作值目标网络 $\hat{Q}$，初始化网络权重 $\theta^{-}=\theta$。

Repeat(对每一个情节)

初始化无人机状态序列 $s_1=(d_{\mathrm{t}}^{1},\phi_{\mathrm{t}}^{1},d_{\mathrm{o}}^{1},\phi_{\mathrm{o}}^{1})$

Repeat(对该情节中的每一步)

在概率 ε 内选择一个随机动作 a_k，否则选择 $a_k=\arg\max_a Q(s,a;\theta)$；

无人机执行动作 a_k，计算 k 时刻势函数奖赏 r_k；

计算无人机 k+1 时刻的状态 $s_{k+1}=(d_{\mathrm{t}}^{k+1},\phi_{\mathrm{t}}^{k+1},d_{\mathrm{o}}^{k+1},\phi_{\mathrm{o}}^{k+1})$；

将当前经验 (s_k,a_k,r_k,s_{k+1}) 存储在经验池 D 中；

从经验池 D 中随机抽取最小经验块 $D_{\min}$；

计算目标值函数 y_k 为

$$y_k=\begin{cases} r_k & \text{情节终止} \\ r_k+\gamma\max_{a'}\hat{Q}(s_{k+1},a';\theta^{-}) & \text{其他} \end{cases};$$

对 $(y_k-Q(s_k,a_k;\theta))^2$ 执行梯度下降法，更新 Q 估计网络权值 θ；

每隔 C 步设置 $\theta^{-}=\theta$。

Until 该情节结束

Until 训练结束

在执行阶段，无人机每飞行一个时间步长，通过 Q 估计网络计算出当前状态下的最优动作 $a^{*}\in\{0,1,\cdots,n-1\}$。假设无人机当前的位置坐标为 $(x_{\mathrm{u}},y_{\mathrm{u}})$，飞行步长为 η，无人机在下一时刻的位置坐标 $(x'_{\mathrm{u}},y'_{\mathrm{u}})$ 由下式计算得到：

$$\left.\begin{aligned} x'_{\mathrm{u}}&=(1-\lambda)x_{\mathrm{u}}+\lambda\bullet\eta\bullet\cos(a^{*}\bullet\varphi)\\ y'_{\mathrm{u}}&=(1-\lambda)y_{\mathrm{u}}+\lambda\bullet\eta\bullet\sin(a^{*}\bullet\varphi)\end{aligned}\right\}。\tag{8}$$

式中 λ 为路径平滑系数，起细分动作空间的作用，在一定意义上增加了动作数目，使得规划出的路径更加平滑。

依次类推，即可规划出到达目标点的最优无碰撞路径，至此路径规划任务结束。

4　仿真验证与结果分析

为了验证 PF-DQN 算法在未知环境下无人机路径规划的效果，在 Ubuntu 操作系统上，使用 python 语言在 pycharm 上搭建仿真环境，DQN 神经网络框架采用著名的谷歌开源的基于数据流编程的网络框架 TensorFlow 进行仿真实验。

4.1　仿真环境与参数设置

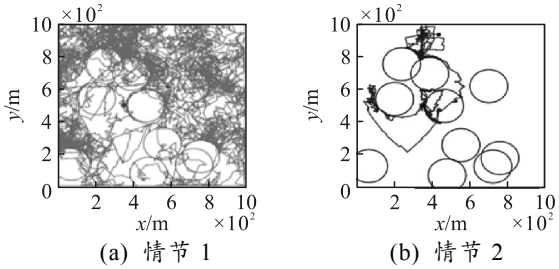
假设在一个 1 000 m×1 000 m 的连续区域，在该区域内无人机的起始位置和目标位置随机产生，其中随机产生 10 个障碍物，无人机感知系统对障碍物的感知距离设为 100 m。PF-DQN 路径规划算法在训练过程中的各项参数设置如表 1 所示。

表 1　PF-DQN 路径规划算法训练参数

参数名称	参数大小	参数名称	参数大小
隐含层个数	3	路径平滑系数 λ	0.75
隐含层神经元个数	100	累积回报折扣因子 γ	0.9
输出层神经元个数	100	ε-greedy 探索率 ε	0.1
神经元激活函数	ReLu	Q 目标网络更新间隔步数 C	200
神经网络学习率 α	0.01	经验池容量 D	20 000
动作划分个数 n	36	经验最小块大小 $D_{\min}$	500

4.2　仿真结果及分析

为了验证笔者所提的 PF-DQN 算法在无人机路径规划的有效性，首先给出 DQN 训练过程中，无人机探索路径的 4 种情况如图 5 所示。图中，圆点表示无人机的起始位置，六角星表示目标位置，圆圈表示无人机对障碍物的探测距离，即无人机一旦进入圆圈内就可感知到障碍物的位置。情节 1、2、1 000、2 071 中障碍物位置相同，无人机的起始位置不同，目标的位置只有情节 1 和情节 2 相同。在图 5 情节 1 中，为神经网络训练的开始阶段，无人机没有任何有价值的经验可利用，所选择的动作值最大的动作并非最优。按照 ε-greedy 策略，无人机有 ε=0.1 的概率进行探索，随机选择动作，所以情节 1 中无人机的路径看起来相当漫长且复杂，但最终经过不停地“摸索”，找到了目标的位置，并为自己积累了许多经验。在图 5 情节 2 中，障碍物和目标位置不变，无人机的起始位置随机产生，由于情节 1 无人机在遍历很多状态空间后并最终找到目标，Q 估计网络已经存储了很多正样本，所以情节 2 中，无人机能够较快地找到目标并在一定程度上避开障碍物。图 5 情节 1 000 情况为更换目标位置后训练的初期，和图 5 情节 1 相似也需要遍历大量状态才能找到目标。图 5 情节 2 071 为对当前目标位置进行大量训练后的情况，由图中可明显看出：无人机已经能很快地避开障碍物找到目标，由于无人机仍有 ε=0.1 的概率对环境进行探索，所以该路径还存在一定概率的“曲折”。

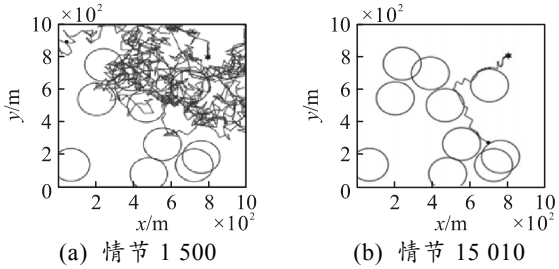


(a) 情节 1 (b) 情节 2

(c) 情节 1 000 (d) 情节 2 071

图 5 PF-DQN 部分训练过程

图 6 列出了在不同障碍物、不同目标位置、不同无人机起始点下 PF-DQN 训练的其他 4 种情况。从图中可明显看出，随着训练的深入，无人机能更快地搜索到目标且实现避障。

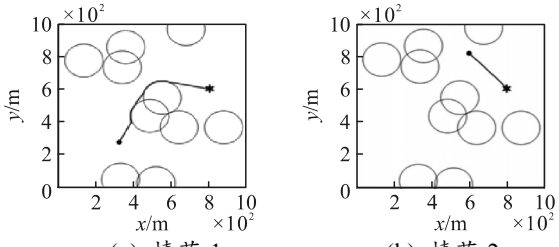


(a) 情节 1 500 (b) 情节 15 010

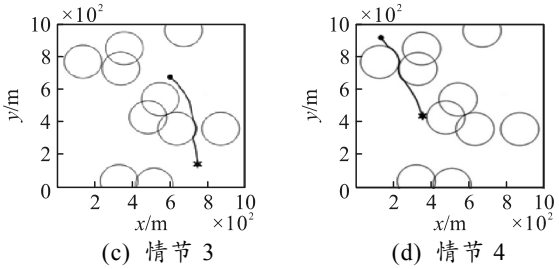
(c) 情节 1 000 (d) 情节 2 071

图 6 PF-DQN 训练过程的其他 4 种情况

如图 7 所示，在测试阶段，随机生成 10 个障碍物，此后障碍物位置固定不动，测试在目标和无人机起始位置随机产生的情况下，PF-DQN 进行路径规划的效果。从图中可以看出，无人机能够有效避开障碍到达目标位置。



(a) 情节 1 (b) 情节 2



(c) 情节 3 (d) 情节 4

图 7 PF-DQN 路径规划效果测试

图 8 为不同动作个数的 DQN 训练耗时情况。柱形图表示不同动作个数在前 200 个情节训练过程的耗时。从图中可看出：动作的个数越多，训练消耗的时间越长。

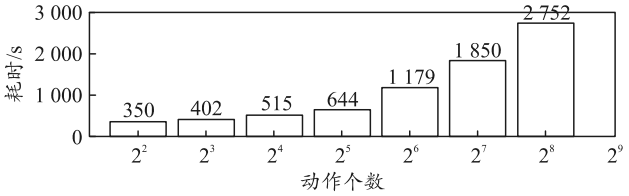
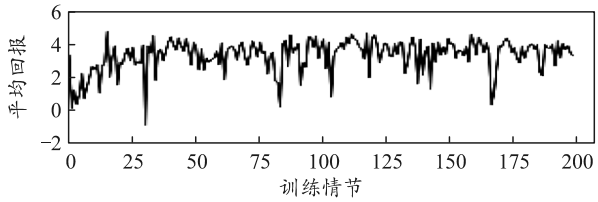
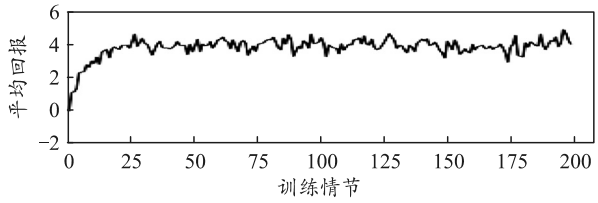


图 8 不同动作个数的 DQN 训练耗时情况

图 9 为基于简单的奖赏规则和基于 PF 奖赏函数 2 种定义下，DQN 训练过程中前 200 个情节的平均回报曲线比较。所谓基于简单的奖赏规则是指目标对无人机的奖赏定义为当无人机执行下一动作产生的结果是接近目标，则给出固定的正奖励值，反之，给出固定的负奖励值。这里的固定奖励值设定为无人机的飞行步长 η。同样，障碍物对无人机的奖赏也按此规则定义。由图 9(a)可以看出：无人机获得的回报趋势变化缓慢，且整体趋势波动较大，平均回报收敛慢。由图 9(b)可以看出：PF-DQN 平均回报虽然伴随着一定的振荡，但获得的回报整体增加较快。由此可见，PF-DQN 比基于简单的奖赏规则 DQN 能够更快地将回报最大化，具有更快的回报收敛速度。



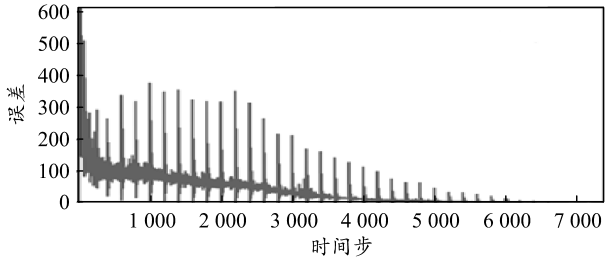
(a) 基于简单的奖赏规则的平均回报



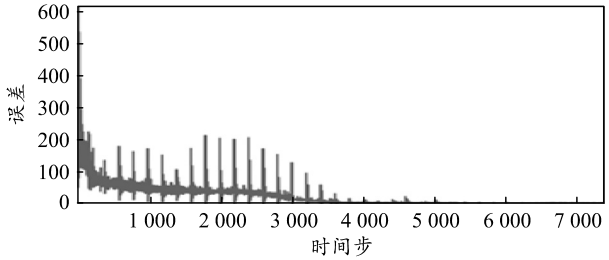
(b) 基于 PF 的奖赏函数的平均回报

图 9 2 种奖赏规则下的平均回报比较

图 10 为基于简单的奖赏规则 DQN 和 PF-DQN 在 7 000 步训练过程中动作估计值和目标值的误差比较。由图中可以看出：无人机在路径探索过程中，存在 ε=0.1 的概率随机选择动作，所以 2 个曲线都会出现尖峰误差的现象。但图 10(a)是基于简单的奖赏规则 DQN 算法，在训练 5 000 步后，网络权值才达到收敛状态，而图 10(b)是 PF-DQN 在训练到 3 500 步时已经接近收敛，所以相比基于简单的奖赏规则 DQN，在路径规划训练过程中网络权值收敛更快。



(a) 基于简单的奖赏规则误差



(b) 基于 PF 的奖赏函数误差

图 10　2 种奖赏规则下的 DQN 训练误差比较

5　结束语

针对无人机无环境模型路径规划问题，笔者提出基于势函数奖赏的 DQN 无人机路径规划方法。通过建立的状态空间，可以表达无人机在环境中的任意连续状态，更符合实际作战环境需要。改进常规的固定奖赏设置规则，设计了目标和障碍物对无人机分段的势函数奖赏规则，刻画了不同动作对无人机的影响。仿真实验结果表明：PF-DQN 算法能够实现无人机在环境信息未知下有效避障的路径规划，且势函数奖赏加快了网络的收敛速度。

参考文献：

[1] 张智, 翁宗南, 苏丽, 等. 室内机器人避碰路径规划[J]. 小型微型计算机系统, 2019, 40(10): 2077-2081.

[2] MENG B B, GAO X. UAV Path Planning Based on Bidirectional Sparse A* Search Algorithm[C]. International Conference on Intelligent Computation Technology and Automation. IEEE, 2010: 1106-1109.

[3] DONG Z N, CHEN Z J, ZHOU R, et al. A hybrid approach of virtual force and A* search algorithm for UAV path re-planning[P]. Industrial Electronics and Applications (ICIEA), 2011 6th IEEE Conference on, 2011.

[4] LIU J, YANG J, LIU H, et al. An improved ant colony algorithm for robot path planning[J]. Soft Computing, 2016, 1(11): 1-11.

[5] Jae Byung Park. Multiple mobile robot path planning for rollover prevention and collision avoidance[P]. Control, Automation and Systems (ICCAS), 2011 11th International Conference on, 2011.

[6] 关震宇, 杨东晓, 李杰, 等. 基于 Dubins 路径的无人机避障规划算法[J]. 北京理工大学学报, 2014, 34(6): 570-575.

[7] LIANG Y, JIA Y. Combined Vector Field Approach for 2D and 3D Arbitrary Twice Differentiable Curved Path Following with Constrained UAVs[J]. Journal of Intelligent & Robotic Systems, 2016, 83(1): 133-160.

[8] ZHANG Y, LI S, GUO H. A type of biased consensus-based distributed neural network for path planning[J]. Nonlinear Dynamics, 2017, 89(3): 1-13.

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

[10] WU J, SHIN S, KIM C G, et al. Effective lazy training method for deep q-network in obstacle avoidance and path planning[C]. IEEE International Conference on Systems, Man and Cybernetics. IEEE, 2017: 1799-1804.

[11] 白辰甲, 刘鹏, 赵巍, 等. 基于 TD-error 自适应校正的深度 Q 学习主动采样方法[J].计算机研究与发展, 2019, 56(2): 262-280.

[12] 陈雪超, 开超, 卢飞宇. 室内环境下移动机器人地图构建与路径规划技术[J]. 电子技术与软件工程, 2019(20): 83-84.

[13] 刘帆, 刘鹏远, 张峻宁, 等. 一种改进的深度学习模型自适应学习率策略[J]. 兵工自动化, 2019, 38(1): 78-83.

[14] 徐安, 寇英信, 于雷, 等. 基于 RBF 神经网络的 Q 学习飞行器隐蔽接敌策略[J]. 系统工程与电子技术, 2012, 34(1): 97-101.

[15] ARULKUMARAN K, DEISENROTH M P, BRUNDAGE M, et al. Deep Reinforcement Learning: A Brief Survey[J]. IEEE Signal Processing Magazine, 2017, 34(6): 26-38.

[16] 赵冬斌. 深度强化学习综述: 兼论计算机围棋的发展[J]. 控制理论与应用, 2016, 33(6): 701-717.

[17] 李晨溪, 曹雷, 张永亮, 等. 基于知识的深度强化学习研究综述[J]. 系统工程与电子技术, 2017, 39(11): 2603-2613.