DOI:10.19651/j.cnki.emt.2211675

基于 DDQN 的无人机区域覆盖路径规划策略*

沈 骁[1,2] 赵彤洲[1,2]
(1.武汉工程大学计算机科学与工程学院 武汉 430205; 2.武汉工程大学智能机器人湖北省重点实验室 武汉 430205)

摘 要: 基于深度强化学习方法对未知环境的无人机区域覆盖路径规划进行研究,通过搭建栅格环境模型,在环境中随机部署无人机和禁飞区位置,利用双深度 Q 网络(DDQN)训练无人机的覆盖策略,得到了一套基于 DDQN 的无人机未知区域覆盖路径规划框架。仿真实验表明,设计的无人机未知区域覆盖路径规划框架在无禁飞区的环境下可以实现完全覆盖,在含有未知数量的禁飞区下也能比较好的完成区域覆盖任务,与 DQN 方法比较,其平均覆盖率能够在相同训练条件和训练次数下高出 2%,与 Q-Learning 方法和 Sarsa 方法对比,在无禁飞区的环境中分别高出 4%和 3%。

关键词: 未知环境;区域覆盖;深度强化学习;路径规划
中图分类号: TP391.9 **文献标识码:** A **国家标准学科分类代码:** 520.2

UAV regional coverage path planning strategy based on DDQN

Shen Xiao[1,2] Zhao Tongzhou[1,2]
(1. School of Computer Science & Engineering Artificial Intelligence, Wuhan Institute of Technology, Wuhan 430205, China;
2. Hubei Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan 430205, China)

Abstract: The path planning of UAV area coverage in unknown environment is studied based on deep reinforcement learning method. By building a grid environment model, randomly deploying UAV and no-fly zone in the environment, and using a double deep Q-network (DDQN) to train the coverage strategy of UAV, a set of UAV coverage path planning framework base on DDQN is obtained. The simulation experiment shows that the designed UAV unknown area coverage path planning framework can achieve full coverage in the environment without no fly zone, and can also better complete the area coverage task in the environment with an unknown number of no fly zones. Compared with DQN method, its average coverage rate can be 2% higher under the same training conditions and training rounds, higher than Q Learning method and Sarsa method in the environment without no fly zone.

Keywords: unknown environments; area coverage; deep reinforcement learning; path planning

0 引 言

面向区域覆盖的路径规划是指利用移动机器人或固定传感器,在物理接触或传感器感知范围内遍历目标环境区域[1],并尽量满足时间短、重复路径少和未遍历区域小的优化目标。目前区域覆盖被广泛应用于工业,军事,搜救等领域[2-5]。根据对外部环境的掌握程度,区域覆盖策略分为基于先验信息的路径规划和基于传感器探测信息的路径规划。实现方法主要包括传统方法、栅格地图法、单元分解法以及强化学习方法。传统方法主要包括经典的弓字形、回字形以及模板匹配法。传统方法的思路是对区域进行规则路径扫描,如 Avellar 等[6]使用传统的来回运动扫描路径,和车辆路径问题解决方法结合,实现无人机地面区域最小时间覆盖,Sun 等[7]则是使用贪婪算法来寻找最优覆盖路径,有效的提高了寻找最优覆盖路径速度,但这些会出现路径覆盖冗余度高、覆盖效率低下等问题。栅格法将规划空间分解为系列二值信息的网络单元并采用经典的四叉树或八叉树进行搜索,如宋宇等[8]使用优化的 A* 算法对无人机路径规划进行求解,得到了更短的搜索路径,杜健健等[9]则使用群体智能算法求解无人机的侦察路径,但这两者主要集中在从 A 点到 B 点的任务。但栅格法需要对部分栅格进行回归,计算量大且回归率较高,不适用于大场景区域覆盖。单元分解法将区域划分为若干个子区域,然后对区域进行邻域遍历完成路径规划,如 Nigam 等[10]使用了一种精确的细胞分解方法,结合单元分解法和传统方法,将目标区

收稿日期:2022-10-12
* 基金项目:国家重点研发计划(2016YFC0801003)项目资助

域划分为一系列可以来回运动覆盖的凸区域,以此保证对目标区域的完全覆盖,Xu 等[11]则在此基础上进行了扩展,将要覆盖的区域编码为 Reeb 边,然后构造一个 Euler 环,该环恰好覆盖图中的每条边,但这两者都没有限制智能体的路径长度,会出现较多的重复覆盖路径。以上的方法大多都需要简单的环境或者大量的先验知识,但是这些条件在未知环境中都很难获得。因此就有学者尝试利用强化学习方法不需要先验知识的特点,让智能体与环境进行交互,通过环境给予的反馈来试错迭代不断优化自身的路径规划策略。如 Tianze 等[12]通过使用 Q 学习算法规划无人机自主到达目标区域的路径,但在其环境中不存在障碍物,无法适应有障碍物的环境。常宝娴等[13]将 Q 学习算法的机器人路径规划扩展到了有障碍物的未知环境中,通过强化学习能够自主学习的特点,让机器人能够在未知环境中完成区域覆盖路径规划任务的同时避开障碍物,但只适用于单机器人,Hung 等[14]将集群算法和 Q 学习算法结合,解决了多无人机在随机环境中的路径规划,但不能完全覆盖目标,而上述方法的缺点是在未知环境中由于复杂状态空间导致 Q 值表巨大,使得查询 Q 值的速度变慢而让算法速度变慢。随着深度神经网络(deep neural network, DNN)的发展,强化学习算法与之相结合,形成的深度强化学习(deep reinforcement learning, DRL)方法可以解决查询 Q 表速度慢的问题[15],深度强化学习方法使用神经网络来确定输入与输出之间的映射关系,由此避免了强化学习方法中巨大的 Q 值表问题,还保留了强化学习的决策性,能够适用于复杂的状态空间情况。本文基于深度强化学习的方法,设计了一个双深度 Q 网络(double deep Q network, DDQN)区域覆盖路径规划算法,利用无人机与环境的交互获得的奖励来决定动作,分别与深度 Q 网络(deep Q network, DQN),Q 学习(Q-Learning),SARSA(state action reward state action)相比,大幅提升了算法性能。

1　概念描述

无人机区域覆盖问题首先需要对环境进行建模以用于在深度强化学习方法中生成状态值和奖励值。整个覆盖环境可以用两个二维栅格地图来表示,这两个栅格地图分别描述了目标区和禁飞区。目标区域是无人机必须至少要覆盖一次的区域,禁飞区代表无人机被禁止进入的区域。

无人机区域覆盖问题可以转化为马尔可夫决策过程(Markov decision process, MDP)。MDP 由元组 $\langle S,A,R,P\rangle$ 来描述,其中 S 表示无人机状态空间的集合,通过三维数组 $[P_o,Z_{cov},Z_{no\_flv}]$ 存储,其中 P_o 表示无人机的位置,Z_{cov} 表示已覆盖的区域,Z_{no_flv} 表示已知的禁飞区;A 表示无人机可用动作的空间集合为:$A=\{$上,下,左,右$\}$;R 为奖励函数,即是无人机从某一状态到达另一状态时获得的奖励,表示为 $S\times A\to R$;P 是从某一状态转移到下一状态的概率,如无人机在 t 时刻状态 s 时选择动作 a,在 $t+1$ 时刻状态为 s' 的概率表示为 $P_{ss'}^{a}=P[S_{t+1}=s'\mid S_t=s,A_t=a]$。因此,无人机的区域覆盖问题实质上就是让无人机寻找一条路径,使得无人机根据这条路径能够覆盖所有的目标区域并且能够避开禁止飞行的区域,而无人机花费的步数和重复覆盖的区域应当都尽可能的少。

在定义无人机区域覆盖性能之前,确定如下概念:

定义 1:移动步数使用率设无人机执行区域覆盖任务中实际移动步数为 n,规定移动步数为 m,则移动步数使用率 r_{mov} 为:

$$r_{mov}=\frac{n}{m} \tag{1}$$

其中,m 通常取覆盖环境边长 l 的 10 倍,即有 $m=10l$。若移动步数到达规定步数,那么就结束这次覆盖任务,移动步数使用率为 1。

定义 2:覆盖率

设无人机已覆盖区域面积为 S_{cov},整个目标区域面积为 S_{target} 则覆盖率 r_{cov} 为:

$$r_{cov}=\frac{S_{cov}}{S_{target}} \tag{2}$$

覆盖率可以直观的看出一次覆盖任务的效果,覆盖率越高表示覆盖效果越好,当覆盖率为 1 代表目标区域全覆盖,无人机很好的完成了覆盖任务。

定义 3:重复覆盖步数

无人机在覆盖路径上对目标区域的重复覆盖所花费的步数就是重复覆盖步数 n_{recov},对于需要覆盖的目标区域,仅覆盖一次必然是最好的结果,但无人机的在移动过程中由于自身的覆盖范围必然会导致部分目标区域的重复覆盖,因此将无人机在一次移动后没有覆盖任何新的目标区域时判定为重复覆盖。重复覆盖会增加需要移动的总步数而且不会对覆盖效果造成影响,所以重复覆盖步数越少越好。

2　深度强化学习网络结构

强化学习有别于监督学习和非监督学习,它的基本原理是智能体主动对环境进行探索,对探索环境所使用的动作获得一个奖励值,如果智能体的某个行为策略导致奖励值上升,那么智能体之后产生这个行为策略的趋势就会加强,强化学习模型如图 1 所示。

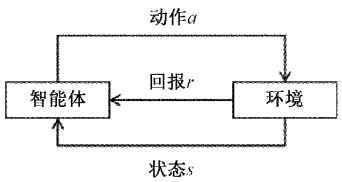


图 1　强化学习模型

Q 学习算法是一个基于值的强化学习算法[16],具有在线学习的特点,不需要太多的离线训练。它使用函数 $Q_t(s_t,a_t)$ 表示在时间 t,智能体的状态为 $s_t\in S$,执行了

动作 $a_t \in A$ 之后累计获得的回报值。该算法的最终目的是最大化 Q 函数的值，Q 值的更新公式为：

$$Q(s,a) \leftarrow Q(s,a)+\alpha[r+\gamma max_{a'}Q(s',a')-Q(s,a)] \tag{3}$$

其中，$Q(s,a)$ 表示智能体在当前状态 s 下选择动作 a 的 Q 值，α 表示学习率（$\alpha>0$），γ 表示折扣因子（$0<\gamma<1$），r 表示执行动作 a 后获得的回报值，$max_{a'}Q(s',a')$ 表示下一状态的最大 Q 值。Q 学习算法的动作选择策略一般都是使用 ε-$greedy$ 策略：

$$\pi(s)=\begin{cases} argmaxQ(s,a), & q<1-\varepsilon \\ a_{random}, & \text{其他} \end{cases} \tag{4}$$

其中，q 表示在范围 [0,1] 中的随机数，a_{random} 表示在可选的动作集合中选择随机的动作，ε 表示探索率，ε 值越大选择随机动作的概率越大。

Q 学习算法中的 Q 值表在处于较大状态或动作空间时容易发生维度爆炸，这使得在 Q 值表中搜索对应的状态会变得更耗时。深度 Q 网络（DQN）则利用神经网络替代 Q 值表，输入状态即可通过 DQN 网络获得 Q 值。使用参数向量 θ 参数化 DQN 网络，其最小化损失函数为：

$$L(\theta)=E[(Q(s,a;\theta)-Y(s,a,s'))^2] \tag{5}$$

其中目标值函数为：

$$Y(s,a,s')=r(s,a)+\gamma max_{a'}Q(s',a';\theta) \tag{6}$$

与 Q 值表相比，DQN 采用神经网络建立输入参数和输出参数之间映射关系的方法使得数据效率高了很多，但是使用神经网络逼近值函数的是不稳定的，并且可能由于相关样本的偏差而导致发散。2015 年，Mnih 等[17]提出了一种稳定 DQN 学习的方法，该方法由 θ 参数化的目标网络，以使样本不相关。此外，生成的样本存储在经验回放内存中，并在训练智能体时从中随机检索。因此目标值函数可以更新为：

$$Y^{DQN}(s,a,s')=r(s,a)+\gamma max_{a'}Q(s',a';\theta') \tag{7}$$

目标网络的参数 θ' 的更新方式分为硬更新和软更新，硬更新是在规定的步数之后直接将当前网络参数 θ 复制到目标网络中，软更新则使用：

$$\theta' \leftarrow (1-\varepsilon)\theta'+\varepsilon\theta \tag{8}$$

其中，$0<\epsilon<1$，软间隔更新系数 ϵ 越小，算法会越稳定，目标网络参数变化越小，算法收敛速度会越慢。

在 DQN 的基础上，Hasselt 等[18]提出了一项新的改进，发现在某些条件下，DQN 的 Q 值会被高估。为了解决这个问题，将双 Q 学习应用到了 DQN 中提出了双 Q 深度网络（DDQN）。其目标值函数改为式(7)：

$$Y^{DDQN}(s,a,s')=r(s,a)+\gamma Q(s',argmax_{a'}Q(s',a';\theta);\theta') \tag{9}$$

DDQN 使用当前正在更新的 Q 网络的参数 θ 来选择最优动作，然后使用目标网络的参数 θ' 来估计这个动作的 Q 值，这样在一定程度上降低了对动作值的高估，使得 Q 值更加接近真实值。

3　基于 DDQN 的区域覆盖路径规划算法

为了将 DDQN 方法应用到无人机区域覆盖路径规划中，需要提出一种适用于区域覆盖路径规划的 DDQN 算法，如算法 1 所示。无人机区域覆盖的 DDQN 网络结构由卷积层和完全连接层组成，网络的输入由目标区、禁飞区和覆盖区的二维栅格地图和无人机自身位置叠加在一起组成。然后，卷积层的内核能够在当前位置和附近的单元之间形成直接的空间连接。为了使其输出形状与输入形状保持一致，对卷积层进行填充。除禁飞区为一填充外，其他通道全部是零填充。选择整流线性单元（rectified linear unit，ReLU）作为卷积层的激活函数。卷积网络最后一层展平并连接具有 ReLU 激活的完全连接层，完全连接层的最后一层作为智能体网络对应的 Q 值输出，其大小为无人机可采取的动作空间 $|A|$ 的大小，无人机区域覆盖神经网络结构图如图 2 所示。

算法 1：基于 DDQN 的覆盖路径网络训练

输入：迭代次数 N_{max}，状态集合 S，动作集合 A，网络参数 θ 和 θ'，更新系数 ϵ，批量梯度下降样本 m，规定步数 l，剩余覆盖栅格数 n_{cov}，折扣因子 γ

输出：网络参数 θ

1. 初始化经验回放 D，随机初始化参数 θ，$\theta' \leftarrow \theta$，初始化覆盖环境
2. **For** $n=0$ to N_{max} **do**
3.　初始化状态 s，规定步数 l，剩余覆盖栅格数 n_{cov}
4.　**While** $l>0$ and $n_{cov}>0$ **do**
5.　采用 soft-max 策略选取动作 a，执行动作 a 得到收益 r 和下一状态 s'，将 (s,a,r,s') 放入 D
6.　**For** $i=1$ to m **do**
7.　从 D 中随机采样样本 (s_i,a_i,r_i,s_i')
8.　$Y_i=\begin{cases} r_i \\ r_i+\gamma Q(s_i',argmax_{a'}Q(s_i',a_i;\theta);\theta') \end{cases}$
9.　计算 $L_i(\theta)$
10.　**end for**
11.　根据梯度损失函数 $\frac{1}{m}\sum_{i=1}^{m}L_i(\theta)$ 更新参数 θ
12.　根据式子 $\theta' \leftarrow (1-\epsilon)\theta'+\epsilon\theta$ 更新参数 θ'
13.　$l=l-1$
14.　**end while**
15. **end for**

选择 Q 值的贪婪策略由式(10)表示：

$$\pi(s)=argmaxQ(s,a;\theta) \tag{10}$$

在训练过程中，使用经验重放机制和随机抽样来降低样本之间的相关性，采用 soft-max 策略来探索状态和动作

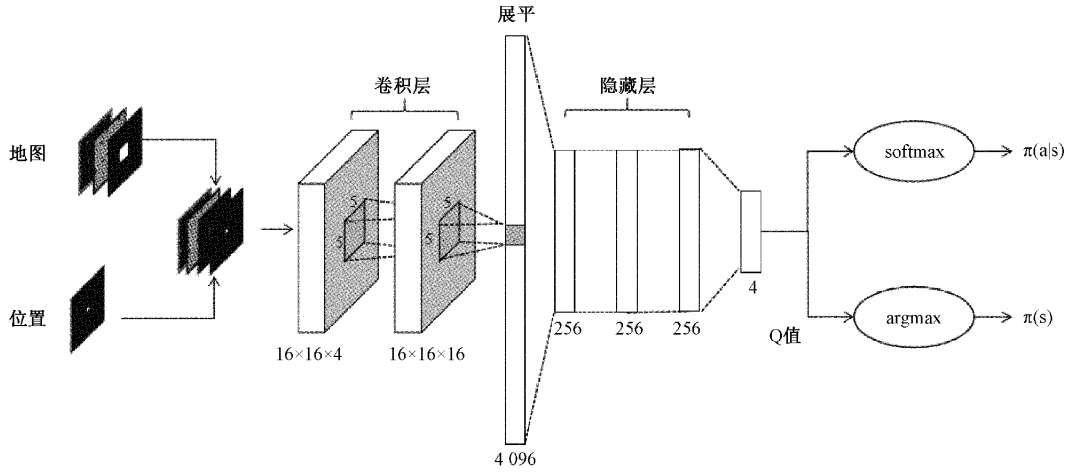


图 2　无人机区域覆盖神经网络结构

空间。其表达式为:

$$\pi(a_i \mid s)=\frac{e^{Q_\theta(s,a_i)/\tau}}{\sum_{a_j\in\Lambda} e^{Q_\theta(s,a_j)/\tau}} \tag{11}$$

其中,参数 τ 是一个超参数,当 τ 增加时,模型会更加偏向于执行探索动作而非总是执行获胜的动作。当 τ 趋向于 0 的时候式(9)就是贪婪策略。

无人机获得奖励的规则为:设无人机的覆盖半径为 μ,奖励值为 r,无人机当前的移动步数为 n,规定步数为 m,当无人机当前状态 s_i 的覆盖范围与前一状态 s_{i-1} 相比,使得未被覆盖的区域被覆盖,其新增的覆盖单元格数为 n_{cov},那么奖励值就为 $r=n_{cov}$。若 $n_{cov}=0$,这说明无人机没有新增的覆盖单元格,那么这一步是无效覆盖,应当给予负面收益,奖励值就为 $r=-(2\mu-1)$。若无人机的移动步数 $n>m$ 或是移动到了禁止区域 Z_{no_flv},判定无人机这次覆盖任务失败,奖励值为 $r=-150$。若无人机执行了当前步的动作之后,所有待覆盖区域都已被覆盖,判定无人机完成了覆盖任务,奖励值为 $r=150$。

此算法详细的描述了双深度 Q 网络的训练过程。首先初始化学习率、训练次数、记忆存储容量、奖励的折扣因子等参数,新的训练开始时重置无人机状态和地图禁飞区域,随机选择无人机起始位置,当无人机尚未完全覆盖区域且无人机的移动步数没有达到规定步数,训练就会继续。从经验回放内存 D 中采样一小批大小为 m 的样本,通过使用 Adam 优化器执行梯度步骤来更新主要网络参数 θ。随后,使用软更新方法更新目标网络参数 θ',并减少可移动步数。当无人机完成覆盖或达到了规定移动步数,本轮结束。然后开始新的训练轮次,直到达到最大训练次数。

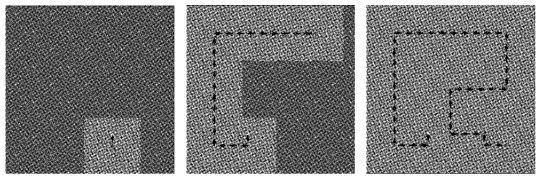
4　仿真实验与结果分析

无人机可以在一个 $n\times n$ 网格中通过 A 中的动作命令进行移动,每一次移动都会增加一次移动步数,无人机的初始状态 $s_0\in S$。设单元网格长度为 1,则地图长度为 n,初始的可移动步数为 $10n$,无人机的覆盖半径设置为 $\mu=3$。为验证在覆盖环境含有不同的禁飞区数量中此算法的有效性,分别设置覆盖区域含禁飞区比例 $\omega=\{0\%,5\%,10\%\}$,其余实验参数如表 1 所示,分别使用 DDQN 算法和 DQN 算法进行覆盖实验。

表 1　仿真实验参数设置

参数	值	含义
$\mid D \mid$	50 000	回放经验内存大小
N_{max}	1 500 000	最大训练次数
τ	0.1	式(11)的参数
m	128	取样大小
γ	0.95	折扣因子
ϵ	0.005	参数 θ' 的更新系数

实验 1:取 $\omega=0$,即无人机巡逻区域无禁飞区,其中无人机出生位置随机。该实验设计为无人机在无禁飞区条件下执行区域覆盖任务的路径规划执行情况。图 3(a)为无人机完成一次覆盖任务在 $t=1$ 时无人机对地图的覆盖情况,图 3(b)为 $t=22$ 时无人机完成了接近一半的地图覆盖,图 3(c)为 $t=43$ 时无人机完成了地图的覆盖。



(a) t=1　(b) t=22　(c) t=43

图 3　无禁飞区时不同时刻区域覆盖路径规划状况

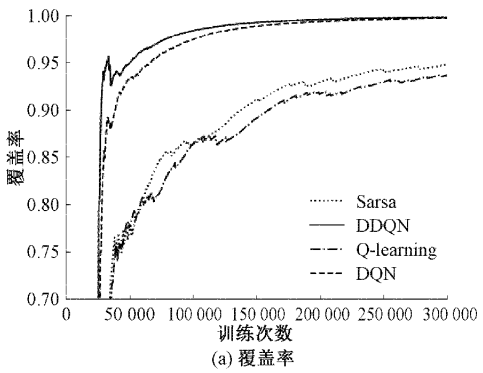
表 2 对比 4 种方法的实验数据发现,在训练次数为 100 000 时,DDQN 方法的平均覆盖率为 0.94 最高,DQN 方法为 0.90,Q-Learning 方法为 0.87,Sarsa 方法为 0.88,

随着训练次数的增加，4 种方法的平均覆盖率都在增长，DDQN 方法在训练次数为 300 000 时平均覆盖率仍然是最高的 0.99，DQN 方法则为 0.97，两种强化学习方法 Q-Learning 和 Sarsa 分别为 0.95 和 0.96，这说明在相同训练次数时，深度强化学习方法覆盖率高于强化学习方法，其中 DDQN 方法的覆盖率最高。

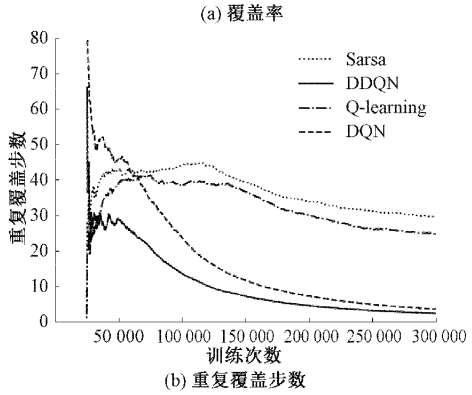
表 2　无禁飞区两种算法平均覆盖率　%

方法	次数		
	100 000	200 000	300 000
DDQN	0.94	0.97	0.99
DQN	0.90	0.95	0.97
Q-Learning	0.87	0.93	0.95
Sarsa	0.88	0.94	0.96

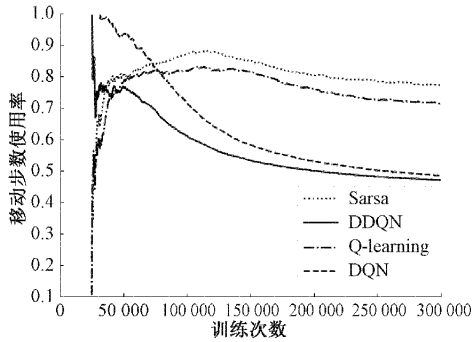
图 4(a)～(c)分别为 4 种方法在无人机区域覆盖训练过程中的覆盖率，重复覆盖次数，移动步数使用率，对比发现 DDQN 方法在前半段有不稳定的现象，这是由于 DDQN 方法在前半段训练次数中由于次优选择陷入了局部最优，经过足够次数的试错之后跳出了局部最优，而在训练过程中使用 DDQN 方法的无人机的重复覆盖次数和移动步数使用率较 DQN 方法低，且两种深度强化学习方法的重复覆盖次数和移动步数使用率在训练开始时高于两种强化学习方法，随着训练时间的增加而逐渐变低，这说明在无禁飞区环境下 DDQN 方法优于 DQN 方法，Q-Learning 方法和 Sarsa 方法。



(a) 覆盖率



(b) 重复覆盖步数



(c) 移动步数使用率

图 4　无禁飞区时 DQN 与 DDQN 训练数据对比

实验 2 取 $\omega = 5\%$，即禁飞区占整个禁飞区面积的 5%，其中无人机出生位置随机，禁飞区位置随机分布，且开始时对无人机是未知的。对覆盖过程进行多次截图，图 5(a)为无人机完成一次覆盖任务在 $t = 1$ 时无人机对地图的覆盖情况，图 5(b)为 $t = 23$ 时无人机避开禁飞区完成了接近一半的地图覆盖，图 5(c)为 $t = 46$ 时无人机对地图的覆盖情况。

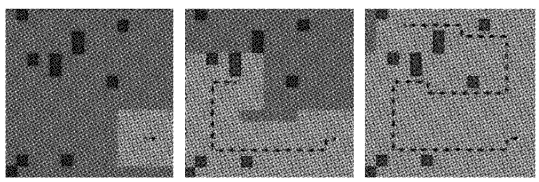

(a) t=1　(b) t=23　(c) t=46

图 5　5%禁飞区时不同时刻区域覆盖路径规划状况

表 3 实验数据发现当环境中在加入了一定数量的随机位置禁飞区之后，4 种方法的覆盖效果与无禁飞区相比都变差了些，在训练次数为 200 000 次时，DDQN 方法的平均覆盖率只达到了 0.92，而 DQN 方法是 0.86，两种强化学习方法的覆盖效率最差只有 0.32，随着训练次数增加，4 种方法的平均覆盖率都在增加，但增长速率在放缓，在训练次数为 800 000 次时，Q-Learning 方法平均覆盖率到达 0.47，Sarsa 方法的平均覆盖率到 0.45，DDQN 方法的平均覆盖率到达 0.98，高于 DQN 方法的 0.96，两种强化学习方法在含有随机数量的禁飞区环境中覆盖效果比无禁飞区环境差，而两种深度强化学习方法更能适应随机环境，其中 DDQN 方法平均覆盖率最高。

表 3　5%禁飞区两种算法平均覆盖率　%

方法	次数		
	200 000	500 000	800 000
DDQN	0.92	0.97	0.98
DQN	0.86	0.95	0.96
Q-Learning	0.32	0.41	0.47
Sarsa	0.32	0.40	0.45

图 6(a)～(c)分别为 4 种方法在无人机区域覆盖训练过程中的覆盖率,重复覆盖次数,移动步数使用率,对比发现在训练过程中使用 DDQN 方法的无人机的覆盖率最高,DQN 和 DDQN 两种方法的移动步数使用率和重复覆盖次数相似,Q-Learning 方法和 Sarsa 方法则因为覆盖率低,其重复覆盖次数和移动步数使用率随着覆盖率增长,这说明在 5%禁飞区环境下 DDQN 方法仍然优于 DQN 方法,而两种强化学习方法 Q-Learning 和 Sarsa 在这种环境下覆盖效果大幅下降。

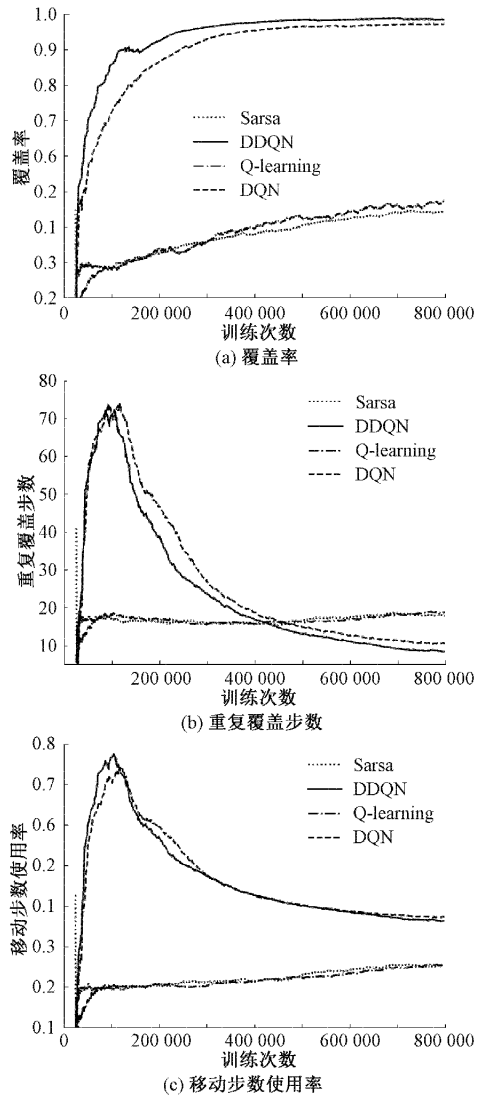


图 6　5%禁飞区时 DQN 与 DDQN 训练数据对比

实验 3 取 $\omega = 10\%$,即禁飞区占整个禁飞区面积的 5%,其中无人机出生位置随机,禁飞区位置随机分布,且开始时对无人机是未知的。图 7(a)为无人机完成一次覆盖任务在 $t = 1$ 时无人机对地图的覆盖情况,图 7(b)为 $t = 30$ 时无人机完成了接近一半的地图覆盖,图 7(c)为 $t = 55$ 时无人机对地图的覆盖情况。

表 4 实验数据中发现,在环境中加入更多的随机位置

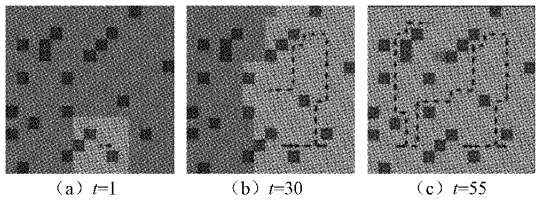


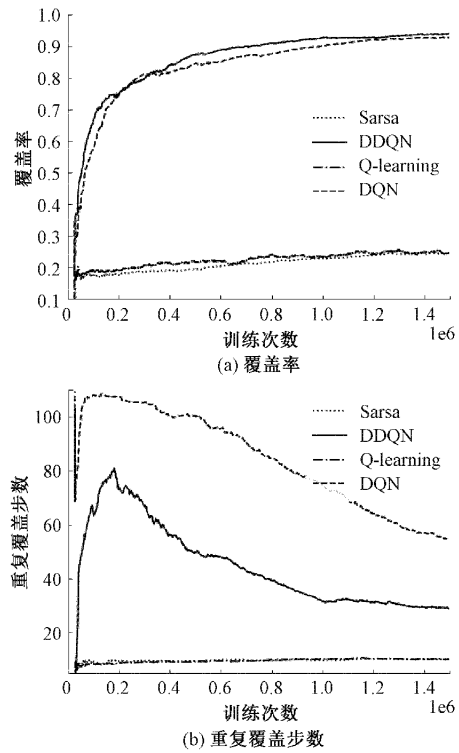
图 7　10%禁飞区时不同时刻区域覆盖路径规划状况

禁飞区后,4 种方法在需要更多的训练次数来跳高覆盖率,在 500 000 次数时,DDQN 方法的平均覆盖率达到 0.95,DQN 方法则是 0.90,Q-Learning 为 0.21,Sarsa 为 0.19,随着训练次数增加,4 种方法的平均覆盖率都在增长,但在相同的训练次数中,DDQN 方法的平均覆盖率一直都高于 DQN 方法和两种强化学习方法,训练次数为 1 500 000 时,DDQN 平均覆盖率在所有方法中最高,为 0.98,其次是 DQN 为 0.96,最差的是两种强化学习方法 Q-Learning 和 Sarsa 为 0.26。

表 4　10%禁飞区两种算法平均覆盖率　　%

方法	次数		
	500 000	1 000 000	1 500 000
DDQN	0.95	0.97	0.98
DQN	0.90	0.95	0.96
Q-Learning	0.21	0.25	0.26
Sarsa	0.19	0.24	0.26

图 8(a)发现 DDQN 方法覆盖率与 DQN 方法覆盖率相似且都大幅高于 Q-Learning 方法和 Sarsa 方法,而在



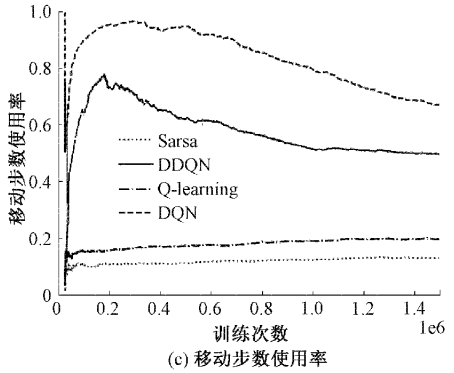


图 8　10%禁飞区时 DQN 与 DDQN 训练数据对比

图 8(b)和(c)显示 DDQN 方法重复覆盖次数与移动步数使用率低于 DQN 方法，而 Q-Learning 方法和 Sarsa 方法因为覆盖效果差导致其重复覆盖步数和移动步数使用率都很低，这说明在 10%禁飞区环境下 DDQN 方法在四种方法中有最好的覆盖效果，而强化学习方法 Q-Learning 和 Sarsa 因为更多的禁飞区而导致其覆盖效果变得更差。

5　结　　论

深度强化学习方法可以解决在未知环境下无人机执行区域覆盖任务的路径规划问题。当前基于强化学习的路径规划方法普遍存在 Q 值表难以查询的问题，本文提出通过将空间信息转化成二维网格进而构建双深度 Q 网络作为无人机区域覆盖的路径规划训练网络结构。该网络通过评价及奖惩函数，计算出无人机的动作选择，在不限定无人机位置的条件下完成区域覆盖任务，与深度 Q 网络方法对比，该方法在相同的训练条件和训练次数下，其平均覆盖率至少高出 2%，在无禁飞区的环境中与强化学习方法 Q-Learning 和 Sarsa 比较，其平均覆盖率分别高出 4%和 3%，在含有 5%和 10%禁飞区的环境中分别高出 51%和 72%。但此算法只能够解决单个无人机的区域覆盖问题，无法实现多个无人机的区域覆盖，因此后续工作将解决多无人机协同完成区域覆盖任务的问题。

参考文献

[1]　付梦家，游晓明. 多机器人系统及其路径规划方法综述[J]. 软件导刊，2017,16(1):177-179.

[2]　KONDA R, LA H M, ZHANG J. Decentralized function approximated Q-learning in multi-robot systems for predator avoidance[J]. IEEE Robotics and Automation Letters, 2020,5(4):6342-6349.

[3]　GALCERAN E, CARRERAS M. A survey on coverage path planning for robotics[J]. Robotics and Autonomous Systems, 2013,61(12):1258-1276.

[4]　CABREIRA T, BRISOLARA L, FERREIRA JR. P R. Survey on coverage path planning with unmanned aerial vehicles[J]. Drones(Basel), 2019,3(1):4.

[5]　HONG Y, JUNG S, KIM S, et al. Autonomous mission of multi-UAV for optimal area coverage[J]. Sensors, 2021,21(7):2482.

[6]　AVELLAR G S C, PEREIRA G A S, PIMENTA L C A, et al. Multi-UAV routing for area coverage and remote sensing with minimum time [J]. Sensors (Basel, Switzerland), 2015,15(11):27783-27803.

[7]　SUN X, CASSANDRAS C G, MENG X. Exploiting submodularity to quantify near-optimality in multi-agent coverage problems[J]. Automatica, 2019,100:349-359.

[8]　宋宇，顾海蛟. 基于改进 A* 算法的无人机航路规划[J]. 长春工业大学学报，2020,41(6):597-601.

[9]　杜健健，万晓冬. 基于蝙蝠算法的多无人机协同侦察任务规划[J]. 电子测量技术，2019,42(7):40-43.

[10]　NIGAM N, BIENIAWSKI S, KROO I, et al. Control of multiple UAVs for persistent surveillance: algorithm and flight test results [J]. IEEE Transactions on Control Systems Technology, 2012, 20(5):1236-1251.

[11]　XU A, VIRIYASUTHEE C, REKLEITIS I. Optimal complete terrain coverage using an unmanned aerial vehicle[C]. 2011 IEEE International Conference on Robotics and Automation, 2011:2513-2519.

[12]　TIANZE Z, XIN H, SONGLIN C, et al. Hybrid path planning of a quadrotor UAV based on Q-learning algorithm [C]. Technical Committee on Control Theory, Chinese Association of Automation, 2018: 301-305.

[13]　常宝娴，丁洁，朱俊武，等. 未知环境下机器人 Q 学习覆盖算法[J]. 南京理工大学学报，2013,37(6):792-798.

[14]　HUNG S M, GIVIGI S N. A Q-learning approach to flocking with UAVs in a stochastic environment[J]. IEEE Transactions on Cybernetics, 2017, 47 (1): 186-197.

[15]　DU W, DING S. A survey on multi-agent deep reinforcement learning: from the perspective of challenges and applications[J]. Artificial Intelligence Review, 2021,54(5):3215-3238.

[16]　LI R, WANG R, TIAN T, et al. Multi-agent reinforcement learning based on value distribution[J]. Journal of Physics Conference Series, 2020,1651:12017.

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

[18]　HASSELT H V, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning[C]. National Conference on Artificial Intelligence, AAAI Press, 2016.

作者简介

沈骁，硕士研究生，主要研究方向为无人机区域覆盖。
E-mail:895564945@qq.com

赵彤洲(通信作者)，博士，副教授，主要研究方向为模式识别、智能计算、模拟仿真等。
E-mail:tongzhouzhao@wit.edu.cn