基于强化学习的海洋移动观测网络观测路径规划方法

赵玉新， 杜登辉， 成小会， 周迪， 邓雄， 刘延龙

引用本文:
赵玉新, 杜登辉, 成小会, 等. 基于强化学习的海洋移动观测网络观测路径规划方法[J]. 智能系统学报, 2022, 17(1): 192–200.
ZHAO Yuxin, DU Denghui, CHENG Xiaohui, et al. Path planning for mobile ocean observation network based on reinforcement learning[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(1): 192–200.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202106004

DOI: 10.11992/tis.202106004
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20211215.1126.007.html

基于强化学习的海洋移动观测网络观测路径规划方法

赵玉新[1], 杜登辉[1], 成小会[1], 周迪[2], 邓雄[1], 刘延龙[1]
(1. 哈尔滨工程大学 智能科学与工程学院, 黑龙江 哈尔滨 150001; 2. 中国舰船研究设计中心, 湖北 武汉 430064)

摘 要: 合理有效地对移动海洋环境观测平台进行规划, 有利于海洋环境观测网络的设计和海洋环境信息的采集。针对庞大的海洋环境, 在有限的观测资源下, 使用深度强化学习算法对海洋环境观测网络进行规划。针对强化学习算法求解路径规划问题中的离散和连续动作设计问题, 分别使用 DQN 和 DDPG 两种算法对该问题进行单平台和多平台实验, 实验结果表明, 使用离散动作的 DQN 算法的奖赏函数优于使用连续动作的 DDPG 算法。进一步对两种算法求解的移动海洋观测平台的采样路径结果进行分析, 结果显示, 使用离散动作的 DQN 算法的采样结果也更好。实验结果证明, 使用离散动作的 DQN 算法可以最大化对海洋环境中有效资料信息采集, 说明了该方法的有效性和可行性。

关键词: 深度强化学习; 海洋环境观测; 路径规划; 无人测量船; Q 学习; 多智能体; 深度确定性策略梯度; 高斯排序
中图分类号: TP242.6 **文献标志码:** A **文章编号:** 1673−4785(2022)01−0192−09

中文引用格式: 赵玉新, 杜登辉, 成小会, 等. 基于强化学习的海洋移动观测网络观测路径规划方法 [J]. 智能系统学报, 2022, 17(1): 192–200.
英文引用格式: ZHAO Yuxin, DU Denghui, CHENG Xiaohui, et al. Path planning for mobile ocean observation network based on reinforcement learning[J]. CAAI transactions on intelligent systems, 2022, 17(1): 192–200.

Path planning for mobile ocean observation network based on reinforcement learning

ZHAO Yuxin[1], DU Denghui[1], CHENG Xiaohui[1], ZHOU Di[2], DENG Xiong[1], LIU Yanlong[1]
(1. College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001 , China; 2. China Ship Development and Design Center, Wuhan 430064, China)

Abstract: Reasonable and effective planning method of mobile vehicles for marine environmental observation is beneficial to the design of marine environmental observation network and the collection efficiency of marine environmental information. In view of the vast marine environment and limited observation resources, the deep reinforcement learning algorithm is used to plan the marine environmental observation network. In order to solve the problems in the design of discrete and continuous motion during the path planning, two algorithms, DQN and DDPG, are designed to solve the problem of single platform and multi-platform experiments. The experimental results show that the reward curve of DQN algorithm using discrete motion is better than DDPG algorithm using continuous motion. This paper further analyzes the sampling path results of the mobile vehicles for marine environmental observation, and the results show that the sampling result of DQN algorithm with discrete action is better. The experimental results show that the DQN algorithm using discrete motion can maximize the effective data information collection, which demonstrates effectiveness and feasibility of the method.

Keywords: deep reinforcement learning; marine environmental observation; path planning; USV; Q learning; multi-agent; DDPG; RankGauss

海洋环境观测在海洋学中有着至关重要的作用, 对海洋环境的观测是人类认识和开发海洋的基础[1]。区域海洋环境观测系统作为全球海洋观

收稿日期: 2021−06−02. **网络出版日期:** 2021−12−16.
基金项目: 国家自然科学基金项目 (41676088); 中央高校基本科研业务费项目 (3072021CFJ0401).
通信作者: 刘延龙. E-mail: yanlong_liu@hrbeu.edu.cn.

测系统中的重要组成部分, 为海洋科学研究、海洋资源探测以及海洋环境状况以及变化趋势等方面提供了有效的观测数据资料。尽管海洋环境观测对人类生活有着重要的科学意义和和社会经济价值, 但是其依然面临着巨大的挑战[2], 如何基于有限的海洋环境观测平台, 构建海洋环境移动观测网络, 实现对区域海洋环境的最优化观测, 以及如何基于海洋移动观测平台获取的实时的海洋环境观测数据, 实现海洋环境观测平台的自适应路径优化成为当前区域海洋环境观测技术发展的重要课题[3-4]。

本文将深度强化学习算法用于区域海洋环境观测网络的观测方案设计。强化学习算法是一类学习、预测、决策的方法, 通过智能体与环境的交互, 利用试错的方式学习最优策略[5]。强化学习算法已经被广泛应用到路径规划中[6-14], 以往的这些工作或将优化算法结合强化学习, 或直接采用和改进强化学习方法, 解决了传统的针对避障的路径规划问题。但是区域海洋观测网络的路径规划不只是针对避障, 其主要目的是通过获取海洋环境预报数据, 智能地选择观测价值较大的区域, 针对这个问题尚未被提出有效的方法。本文吸收了深度强化学习解决路径规划问题的经验[15-21], 考虑海洋环境预报数据, 将海洋环境自适应观测看成一类序列决策优化问题, 海洋环境移动观测平台接到指令, 通过获取当前复杂的海洋环境背景场信息做出下一步决策, 实现复杂海洋环境下的最优观测。

1　问题描述

1.1　数学模型

区域海洋环境移动观测网络由移动观测平台如无人测量船(unmanned survey vessel, USV)、水下滑翔器 (underwater glider)、自主水下航行器 (autonomous underwater vehicle, AUV) 等组成, 观测的对象是海洋中一定时间梯度下温差变化较大的区域。本文主要讨论无人测量船在海洋中的采样点观测路径规划。如图 1 所示, USV 要从选定的起始点(x_1,y_1)出发, 对海洋中的温差改变较大的区域进行测量, 并根据未知的障碍物实时对 USV 进行操控, 避免其碰撞, 目标就是在约束条件下最大化对该区域范围内温度变化梯度较大的点进行采样。

第i个USV$_i$从一个点(x_i,y_i)到另一个点(x_{i+1},y_{i+1})的路径可表示为

$$\begin{cases} x_{i+1} = x_i + v_i t\cos\theta \\ y_{i+1} = y_i + v_i t\sin\theta \end{cases} \tag{1}$$

式中: θ为 USV 在第i个路径点的航向; v_i为 USV 在第i个路径点的速度; t为时间步长。

USV 的海洋环境探测示意图如图 1 所示。USV 在一定方向范围内对周边的海洋环境进行探测, 探测角度为α_i, 探测半径为R, 在该点探测到的采样点的温度差为$(\Delta T_{i1},\Delta T_{i2},\cdots,\Delta T_{im})$, 对探测到的温度差进行比较, 选取温度差最大的$\Delta T_{ij}$为下一点的采样点。

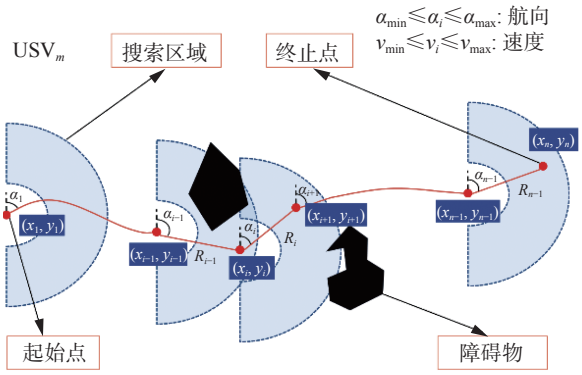


图 1　路径规划采样示意

Fig. 1　Path planning sampling diagram

所以, 对于该问题, 其目标函数为

$$\begin{aligned} &\max f = \sum \Delta T_i i = 1,2,\cdots,n \\ &\text{s.t. } d = d(t_i) \\ &0 \leqslant v_i \leqslant v_{\max} \\ &\theta_1 \leqslant \theta_i \leqslant \theta_2 \end{aligned} \tag{2}$$

式中: d为续航里程约束函数; v_i为速度约束; θ为探测方向角约束; t为时间步长。

1.2　区域耦合环境数值分析预报

在本文中, 主要是将海洋环境要素数值预报信息作为重要参考, 对海洋移动观测网络设计观测方案。因此首先需要构建一个海洋环境数值预报系统, 以获取区域的海洋环境数值预报信息。本文选择在一个中等复杂程度的耦合环流模式 (intermediate complex coupled model, ICCM) 的基础上进行优化调整, 从而获取更加符合区域海洋移动观测网络路径规划的数值预报信息。由于 ICCM 本身的水平分辨率较大, 这样大粒度的数据很难作为区域性移动观测网络路径规划的参照, 因此本文采用一种多层嵌套的方式将耦合模式系统的分辨率由 3.75°变为 0.1°, 并且采用一种最优观测时间窗口的耦合数据同化方法, 构建区域耦合环境分析预报系统。在该系统中, 本文选取经度为 124.0°~129.0°E、纬度为 16.0°~21.0°N 的范围

获取时长为 5 d 的海表温度预报信息作为海洋移动观测网络路径规划依据。

1.3　深度强化学习

深度强化学习 (deep reinforcement learning, DRL) 是近年来人工智能领域一个蓬勃发展的方向，目的是解决计算机从感知到决策控制的问题，从而实现通用人工智能[22]。图 2 展示了强化学习模型中涉及的基本思想和要素。深度强化学习本质上是使用深度神经网络来代替表格作为强化学习算法的策略函数。传统的强化学习算法 (例如 Q-learning) 采用表格作为策略函数，表格中的数值代表每个状态下采取某个动作时的价值，智能体通过比较同一个状态下价值的大小从而选择合适的动作。但是当环境越来越复杂，状态越来越多甚至是连续的时候，采用表格的方法实现决策就越来越困难，而深度神经网络正好具有强大的拟合能力，所以可以将这个决策函数用深度神经网络来代替，这样就形成了深度强化学习。目前，绝大多数强化学习算法都是使用深度神经网络代替传统强化学习算法中的表格作为策略表示。

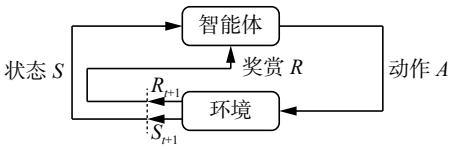


图 2　强化学习示意

Fig. 2　Schematic diagram of reinforcement learning

对于强化学习算法的研究，也在不断地发展，如图 3 所示。Mnih 等[23] 等将神经网络与传统强化学习 (RL) 中的 Q-Learning 算法相结合，提出了深度 Q 网络（deep Q-network, DQN）模型，该模型用于处理基于视觉感知的控制任务，开创了深度强化学习领域新局面[24]。

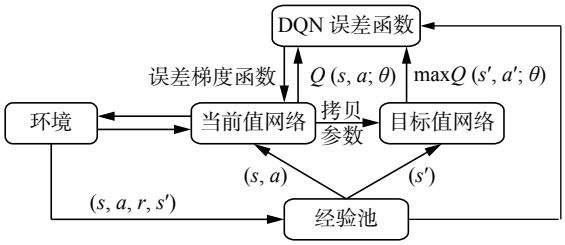


图 3　DQN 算法图解

Fig. 3　DQN algorithm diagram

DQN 算法使用均方差更新网络权重来定义目标函数：

$$L(\omega) = E[r + \gamma \max Q(s', a', \omega) - Q(s, a, \omega)]^2 \quad (3)$$

DQN 有两个特质：1) 经验回收，DQN 有一个经验池用来存储学习数据，由于 DQN 是一种离线学习方法，所以每次 DQN 更新的时候，都可以从记忆库中随机抽取数据；2) 目标神经网络，它是一个结构与 Q 网络完全相同但参数不同的神经网络，用于产生目标值[25]。

2　设计方案

2.1　数据预处理

在本文中，数据预处理主要是指将数值模式预报出的海洋环境数据处理成可被读取、易于训练的形式。本文针对的是区域海洋环境数据，因此主要获取的是区域海洋环境要素场随时间变化的数据和该区域的障碍物信息，然后对原始数据进行数据变形、数据插值、构建时间梯度场、裁切局部环境场、构建障碍物场等。

首先对原始数据进行变形，原始数据无法体现出各点之间的相对位置关系，对于路径规划任务基本没有意义。因此本文根据数据的坐标信息将原始数据重整成均匀分布的海洋环境场数据。输入数据不仅要能表征全局的绝对数据，还要能表征海洋环境要素的时间梯度和观测平台周围的局部数据，因此分别求取海洋环境要素场的时间梯度场和对全局数据场进行裁切获得局部海洋环境要素场；另外，避障约束是路径规划的基本要求，因此输入数据也要能体现该区域障碍物的位置信息，将障碍物进行环境建模，构建区域障碍物场。其中局部环境场的裁切如图 4 所示。

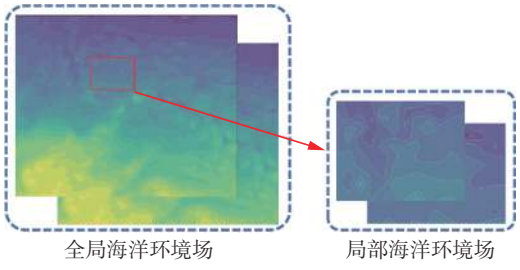


图 4　局部环境场的裁切

Fig. 4　Trimming of local environmental fields

此外，在对深度强化学习算法中的神经网络模型进行训练的过程中，通常不直接使用源数据，而是将数据进行归一化处理，由于经过上述处理后的数据为二维栅格数据，因此采用二维数据归一化。在本文中，采用了 RankGaussion 数据标准化方法。在 RankGaussion 方法中是先将数据按照大小进行排序，然后再将数据标准化为高斯分布，在训练基于梯度的模型 (如神经网络) 时使用这种方法能够加快收敛速度。

2.2　环境状态和动作设计

强化学习的环境指的是对现实环境反映模式的模拟, 或者更一般地说, 它允许对外部环境的行为进行推断。例如给定一个状态和动作, 模型就可以预测下一个状态和收益。除此之外, 环境还能模拟整个规划过程, 包括环境状态的重置, 环境数据的调度, 环境的可视化等。环境对应着我们所要解决的问题的场景, 它通过模拟现实情况进行算法的训练。总之环境就是提供给强化学习算法一个运行平台, 强化学习代理通过与环境进行交互获取状态、动作、奖赏等数据进行训练, 环境则是通过强化学习代理产生的策略根据状态得到动作, 进行完整的状态迭代过程。

强化学习算法中环境的搭建首先要明确状态和动作, 动作即为路径规划过程中的决策, 想要通过训练得到期望的动作, 那么神经网络的输入即状态必须包含足够且精准的环境信息。考虑到海洋环境观测路径规划的目的, 于是取状态为全局海洋环境场、局部海洋环境场和观测平台的位置, 如图 5 所示。

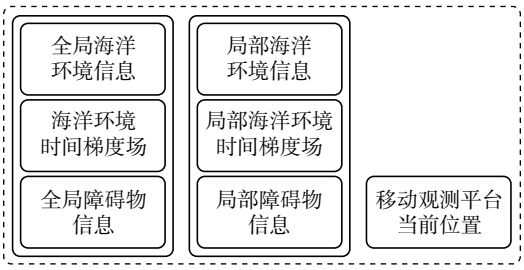


图 5　环境状态设计

Fig. 5　Environmental state design

环境动作空间的设计指定了智能体所能采取的动作的范围, 也决定了其所能探索的状态空间的最大范围。一个好的动作空间的设计是在探索范围和训练效率之间的权衡, 既不能将动作空间设计过于保守, 压缩探索空间范围造成局部最优; 同时又不能将动作空间设计得过于繁琐, 导致训练过程难以收敛; 另外, 动作空间的设计还要考虑动作的“合法性”, 即需要考虑设计的动作是否能够达到或者会不会造成严重的后果, 在设计动作空间的过程中要抛弃不合法的动作。

本文中的动作空间主要指能够对移动观测平台的移动造成影响的变量, 对于宏观的路径点规划来说, 将运动变量归纳为航向和航速。如图 6 所示, 航向和航速两个变量都对移动观测平台的空间探索范围有所影响, 因此为了权衡探索范围和训练效率, 分别将两个变量限制在一定的范围内。

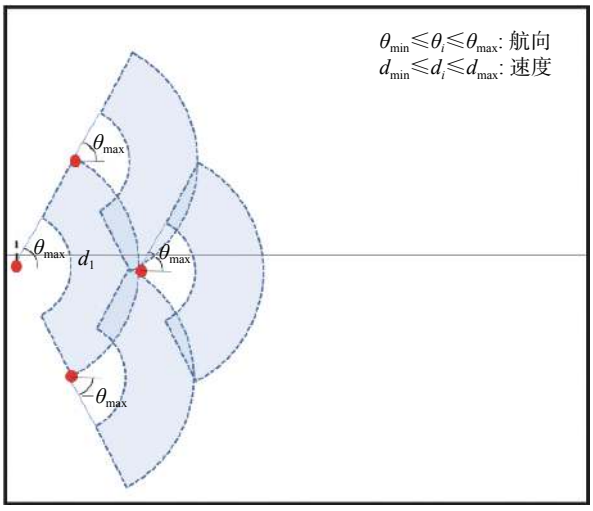


图 6　动作空间设计

Fig. 6　Action space design

2.3　奖赏函数设计

奖赏函数的设计对强化学习算法来说至关重要。强化学习的最终目标就是使得累计期望奖赏最大化, 因此奖赏函数的设计决定了训练的方向, 奖赏函数的设计在一定程度上也就决定了训练效果的上限。本文奖赏函数涉及多个目标, 对多目标优化的处理是通过线性加权的方式转化为单目标优化。奖赏函数应该体现所规划路径的目标以及约束, 即应包含海洋环境待测要素的信息梯度、移动观测平台的测量属性约束、多个移动观测平台之间的避障和重叠约束等。

1) 海洋环境待测要素的信息梯度

在本文中, 移动观测平台执行海洋观测任务主要的目的就是捕捉海洋环境要素的变化特性, 所以当观测资源有限时, 观测应该集中在变化剧烈的区域。待观测要素分析预报场的标准差和水平梯度能有效表征待测要素在时间和空间上的变化特性, 所以分别采用基于待测海域海洋要素的时间梯度和空间梯度作为奖赏函数:

$$F(\mathrm{std}(f(x))) = \left(\sum_{f(x)} \mathrm{std}(V_x)\right) \tag{4}$$

$$F(\mathrm{grad}(f(x))) = \left(\sum_{f(x)} \mathrm{grad}(V_x)\right) \tag{5}$$

2) 移动观测平台的测量属性约束

本文针对观测平台自身的测量属性, 包括时间间隔、测量范围、续航里程, 构建了相应的约束。移动观测平台续航里程则对应整个观测平台的观测轨迹总长度。

3) 观测平台的避障约束

对移动观测平台进行路径规划, 避障是一个

不可能回避的问题，任何观测任务如果不能保证其安全性那么将失去意义。本文针对的是相对全局的路径规划，因此只需考虑海面存在的岛屿等固定障碍元素，这些障碍信息也是执行路径规划的重要信息考量。在本文中，为了完成避障任务需要对智能体施加一个避障约束，当智能体遇到障碍时对其施加一个负的奖赏值，训练智能体避免再碰到障碍。

2.4　智能体设计

对智能体的设计首先是选择深度强化学习算法，它决定了智能体的结构以及参数更新方式，本文选择了 DQN 和深度确定性策略梯度 (deep deterministic policy gradient, DDPG) 算法。另外就是神经网络的搭建。

神经网络是强化学习算法中策略的表征，它是状态空间到动作空间的映射。深度神经网络实现对环境的精确感知，以及强化学习算法从环境状态到决策动作映射的决策能力，实现海洋环境观测路径规划结果最优。神经网络的架构应与状态以及动作相符合，如图 7 所示。由于本文的状态包括全局海洋环境场、局部海洋环境场，以及移动海洋环境观测平台的 X、Y 坐标，因此神经网络的输入为混合输入，采用卷积神经网络对海洋环境场数据进行处理，再与观测平台坐标进行融合作为整个神经网络的输入。神经网络的输出为各离散动作 Q 值，输出的维度为离散动作的个数。

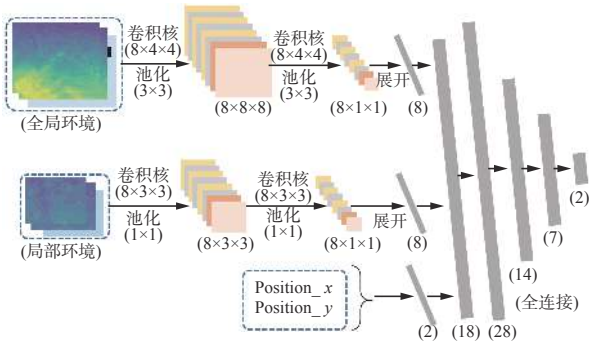


图 7　神经网络架构

Fig. 7　Neural network architecture

以上设计主要是针对单个智能体的情况，对于多智能体的设计主要是对多个单智能体进行组合，以达到整体最优结果。本文所采用的多智能体是完全合作的关系，所有智能体的目标一致，均是改善观测效果，只需要调整智能体训练时的奖赏即可。因此将单个智能体奖赏函数中海洋环境待测要素的信息梯度部分进行求和，作为整体奖赏函数替换单个智能体的梯度奖赏。

3　实验结果与分析

3.1　实验参数设置

在第 2 节中，搭建了采用强化学习训练移动观测平台进行路径规划的框架，分别设计了环境的状态、动作、奖励函数以及智能体的神经网络架构，本节主要是进行实验以及对实验结果进行分析。采用 DQN 及 DDPG 算法进行训练的伪代码分别算法 1 和算法 2 所示。

算法 1　使用 DQN 算法生成路径

1) 创建环境，生成并初始化智能体

2) **for** episode=1, M **do**

3) 初始化环境状态 s_1

4) **for** t=1, T **do**

5) 以 ϵ 的概率随机选择一个动作 a_t

否则选择 $a_t = \max_a Q^*(s_t, a; \theta)$

6) 在环境中执行 a_t 得到奖赏 r_t 和 s_{t+1}

7) 在记忆池中存储样本 (s_t, a_t, r_t, s_{t+1})

8) 从记忆池抽取样本 (s_j, a_j, r_j, s_{j+1})

9) 当 s_{j+1} 为回合终止状态时，$y_j = r_j$，否则

$y_j = r_j + \gamma Q(s_{j+1}, \mathrm{argmax}_a Q(s_{j+1}, a_j, \theta_j); \theta_j^-)$

10) 根据式 $(y_j - Q(s_j, a_j; \theta))^2$ 执行梯度下降；

11) **end for**

12) **end for**

算法 2　使用 DDPG 算法生成路径

1) 创建环境，生成并初始化智能体；

2) 初始化 critic 网络 $Q(s, a|\theta^Q)$，actor 网络 $\mu(s|\theta^\mu)$；

3) **for** episode=1, M **do**

4) 初始化环境状态 s_1；

5) **for** t=1, T **do**

6) 根据策略和噪音选取 $a_t = \mu(s_t|\theta^\mu) + N_t$；

7) 在环境中执行 a_t 得到奖赏 r_t 和 s_{t+1}；

8) 在记忆池中存储样本 (s_t, a_t, r_t, s_{t+1})；

9) 从记忆池抽取 N 个样本 (s_j, a_j, r_j, s_{j+1})；

10) 设置 $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})\theta^{Q'})$；

11) 更新 critic 网络：

$$L = \frac{1}{N}\sum\nolimits_i (y_i - Q(s_i, a_i|\theta^Q))^2$$

12) 更新 actor 网络：

$$\nabla_{\theta^\mu} J \approx \frac{1}{N}\sum_i \nabla_a Q(s, a|\theta^Q)|_{s_i, \mu_i} \nabla_{\theta^\mu} \mu(s|\theta^\mu)|s_i$$

13) 更新目标网络：

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

14) **end for**

15) **end for**

为验证本文所提出方案的有效性，分别在有

障碍和无障碍、单平台和多平台的情况下对实验方案进行测试, 在单平台无障碍情况下对 DQN 和 DDPG 算法进行对比。实验场景设置为经度 124.0~129.0°E, 纬度 16.0~21.0°N, 分辨率为 0.1°的海区, 模拟移动观测平台从西向东进行海洋环境要素观测。

实验中的参数设置如表 1 所示。

表 1　智能体参数设置
Table 1　Agent parameter setting

参数名	参数值
学习效率α	0.01
衰减度γ	0.95
样本池容量	10e+6
最小取样样本数	2e+10
每回合最大时间步	20
随机种子数	2
总训练回合数	500

单平台的环境参数设置如表 2 所示。

表 2　环境参数设置 (单平台)
Table 2　Environmental parameter setting (single platform)

参数名	参数值
起始位置	[18.5°N,124.0°E]
状态维数	(3,51,51)(3,5,5)(1,2)
动作维数	2
动作范围(°,km/h)	(−60,60),(2,5)

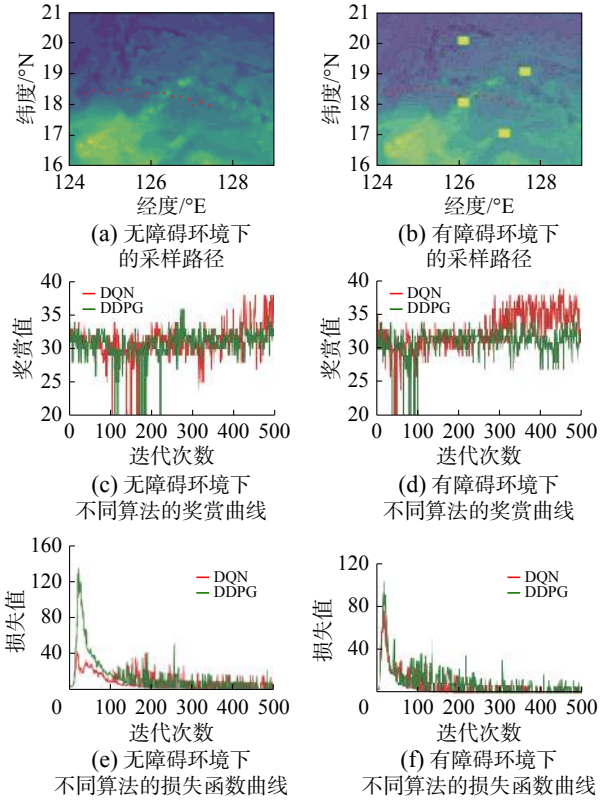
多平台的环境参数设置如表 3 所示。

表 3　环境参数设置 (多平台)
Table 3　Environmental parameter setting (multi-platform)

参数名	参数值
起始位置1	[16.5°N,124.0°E]
起始位置2	[17.5°N,124.0°E]
起始位置3	[18.5°N,124.0°E]
起始位置4	[19.5°N,124.0°E]
起始位置5	[20.5°N,124.0°E]
状态维数	(3,51,51)(3,5,5)(1,2)
动作维数	2
动作范围(°,km/h)	(–60,60),(2,5)

3.2　单平台实验结果

在单平台实验中, 选定移动平台运动初始位置, 分别进行有障碍和无障碍的实验。最终得到奖赏函数曲线和损失函数曲线, 并画出单平台采样路径, 如图 8 所示。

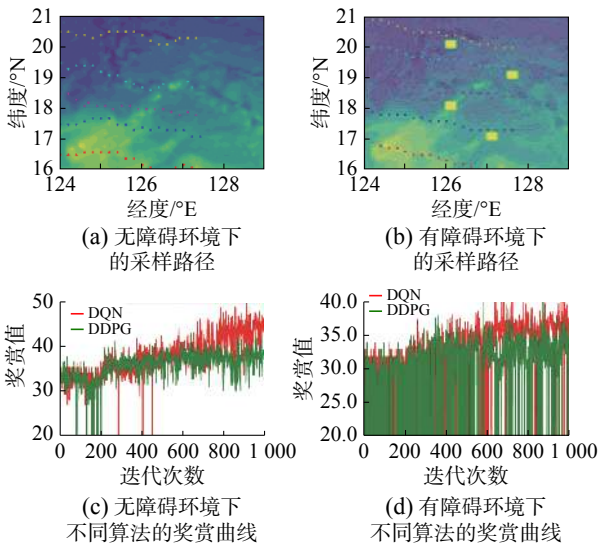


(a) 无障碍环境下的采样路径　(b) 有障碍环境下的采样路径
(c) 无障碍环境下不同算法的奖赏曲线　(d) 有障碍环境下不同算法的奖赏曲线
(e) 无障碍环境下不同算法的损失函数曲线　(f) 有障碍环境下不同算法的损失函数曲线

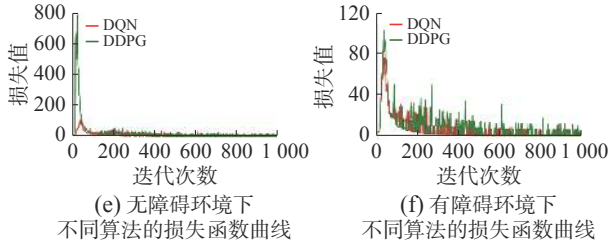
图 8　单平台实验
Fig. 8　Single platform experiment

3.3　多平台实验结果

多平台实验是选取 5 个移动观测平台, 设置 5 个起始点, 分别使用 DQN 算法和 DDPG 算法进行有障碍和无障碍采样实验。多平台进行 1 000 次迭代。实验结果得到奖赏函数曲线、损失函数曲线和多平台采样路径, 如图 9 所示。



(a) 无障碍环境下的采样路径　(b) 有障碍环境下的采样路径
(c) 无障碍环境下不同算法的奖赏曲线　(d) 有障碍环境下不同算法的奖赏曲线



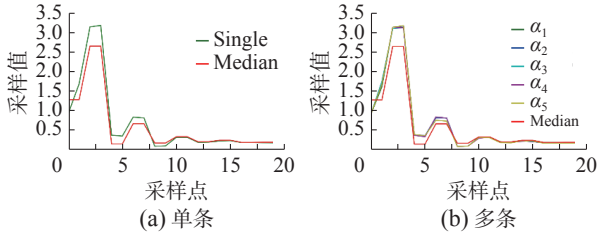
(e) 无障碍环境下不同算法的损失函数曲线　　(f) 有障碍环境下不同算法的损失函数曲线

图 9　多平台实验

Fig. 9　Multi-platform experiment

3.4　实验结果分析

对单平台和多平台通过 DQN 算法得到的采样结果与背景场平均温度进行对比。结果如图 10 所示。



(a) 单条　　(b) 多条

图 10　采样结果对比

Fig. 10　Comparison of sampling results

对于单平台实验，如图 8，分别采用 DQN 和 DDPG 算法在有障碍和无障碍的情况下进行了实验，在进行 500 次的实验迭代后，实验结果表明，采用离散动作空间的 DQN 算法能够得到较好的奖赏曲线，并且其损失函数值相对较小。实验表明，在本观测平台采样任务中，采取离散的动作空间更有利于找到较高的奖赏值，即可以采集到更多的观测信息。

对于多平台实验，如图 9，可明显看出总的奖赏函数的上升趋势。当进行有障碍实验时，由于当路径碰撞障碍物或出界时环境会自动给出负的奖赏值来“警告”智能体，因此可以看到前期奖赏曲线会有比较稠密的负值。随着训练的不断进行，可以观察到负值明显减少，并且奖赏值有比较明显的提高。

在单平台和多平台实验中，观测平台在有障碍的环境下，基于离散动作的 DQN 算法都能有效地避开障碍，对海洋环境信息进行有效采集。

通过对单平台和多平台得到的采样结果与背景场平均温度对比分析，如图 10，单平台和多平台的采样结果都要高于背景场的平均温度，说明基于离散动作的深度强化学习的海洋移动观测平台可以在有限资源条件下采集更多的海洋环境信息，进一步说明 DQN 算法在海洋移动观测网络观测路径规划中的可行性和有效性。

4　结束语

本文主要研究在有限资源条件下如何对移动海洋观测平台进行合理有效的设计，使得观测平台可以对庞大海洋环境中采集更多的有效信息。本文分别设计了基于离散动作的 DQN 算法和基于连续动作的 DDPG 算法对海洋环境移动观测网络进行规划，并对通过算法得到的采样结果的有效性进行了分析。

首先通过获取海洋环境数值预报信息，基于 RankGaussion 对预报信息进行数据预处理，在此基础上结合海洋环境信息和移动观测平台的碰撞及能量约束设置奖赏函数，采用 DQN 和 DDPG 算法最终从与环境的交互信息中学习路径规划策略完成单智能体路径规划任务。在此基础上，构建基于行为分析的多平台观测网络，通过将具有完全合作关系的移动观测平台奖赏进行结合，指导多个移动观测平台各自的采样路径规划。实验结果表明，采用基于离散动作的深度强化学习算法能够有效提高观测效率。

本文将在以下几个方面展开更深入的研究：

1) 针对多观测平台，设计基于协作的多智能体强化学习算法，对移动海洋观测网络进行规划，以期获得更多有效的观测信息；

2) 海洋环境信息复杂，不同的奖赏函数设计都会影响观测效果，下一步将考虑更多的环境因素，研究设计更合理有效的奖赏函数；

3) 将观测方案结果与海洋环境数值预报系统进行深度结合，使得观测数据更好服务于海洋环境数值预报系统。

参考文献：

[1] 王建友.习近平建设海洋强国战略探析 [J]. 辽宁师范大学学报 (社会科学版), 2019, 42(5): 103–112.
WANG Jianyou. Analysis on xi jinping's strategy of building a maritime power [J]. Journal of Liaoning Normal University (social science edition), 2019, 42(5) : 103–112.

[2] 尹路,李延斌,马金钢.海洋观测技术现状综述 [J].舰船电子工程, 2013, 33(11): 4–7, 13.
YIN Lu, LI Yanbin, MA Jingang. A review of ocean observation technology [J]. Ship electronic engineering, 2013, 33(11): 4–7, 13.

[3] 张燕武.自适应海洋观测 [J].地球科学进展, 2013, 28(5): 537–541.

ZHANG Yanwu. Adaptive ocean observation [J]. Advances in earth science,2013,28(5):537–541.

[4] 李颖虹, 王凡, 任小波. 海洋观测能力建设的现状、趋势与对策思考 [J]. 地球科学进展, 2010, 25(7): 715–722.
LI Yinghong, WANG Fan, REN Xiaobo. Current situation, trend and countermeasures of ocean observation capacity construction[J]. Advances in earth science, 2010, 25(7): 715–722.

[5] 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. Multiobjective tracking method based on multi-agent cooperative reinforcement learning[J]. Computer engineering,2020, 46(11):90–96.

[6] 韩向敏, 鲍泓, 梁军, 等. 一种基于深度强化学习的自适应巡航控制算法 [J]. 计算机工程, 2018, 44(7):32–35.
HAN Xiangmin, BAO Hong, LIANG Jun, et al. An adaptive cruise control algorithm based on deep reinforcement learning [J]. Computer engineering, 2018, 44(7): 32–35.

[7] YAN C, XIANG X, WANG C. Towards real-time path planning through deep reinforcement learning for a UAV in dynamic environments[J]. Journal of intelligent & robotic systems, 2020, 98(2): 297–309.

[8] WEN S, ZHAO Y, YUAN X, et al. Path planning for active SLAM based on deep reinforcement learning under unknown environments[J]. Intelligent service robotics, 2020, 13(2): 263–272.

[9] YAO Q, ZHENG Z, QI L, et al. Path planning method with improved artificial potential field—A reinforcement learning perspective[J]. IEEE access, 2020, 8: 135513–135523.

[10] LI B, WU Y. Path planning for UAV ground target tracking via deep reinforcement learning[J]. IEEE access, 2020, 8: 29064–29074.

[11] JIANG J, ZENG X, GUZZETTI D, et al. Path planning for asteroid hopping rovers with pre-trained deep reinforcement learning architectures[J]. Acta astronautica, 2020, 171: 265–279.

[12] WANG B, LIU Z, LI Q, et al. Mobile robot path planning in dynamic environments through globally guided reinforcement learning[J]. IEEE robotics and automation letters, 2020, 5(4): 6932–6939.

[13] WEI Y, ZHENG R. Informative path planning for mobile sensing with reinforcement learning[C]//IEEE INFOCOM 2020-IEEE Conference on Computer Communications. Beijing, China, 2020: 864–873.

[14] JOSEF S, DEGANI A. Deep reinforcement learning for safe local planning of a ground vehicle in unknown rough terrain[J]. IEEE robotics and automation letters, 2020, 5(4): 6748–6755.

[15] 杜威. 多智能体强化学习研究 [D]. 徐州:中国矿业大学, 2020.
DU Wei. Research on multi-agent reinforcement learning [D]. Xuzhou: China University of Mining and Technology, 2020.

[16] 卜祥津. 基于深度强化学习的未知环境下机器人路径规划的研究 [D]. 哈尔滨: 哈尔滨工业大学, 2018.
BU Xiangjin. Research on robot path planning in unknown environment based on deep reinforcement learning [D]. Harbin: Harbin Institute of Technology, 2018.

[17] 向卉. 基于深度强化学习的室内目标路径规划研究 [D]. 桂林: 桂林电子科技大学, 2019.
XIANG Hui. Research on robot path planing under unknown environment based on deep reforcement learning [D]. Guilin: Guilin University of Electronic Technology, 2019.

[18] 姚君延. 基于深度增强学习的路径规划算法研究 [D]. 成都: 电子科技大学, 2018.
YAO Junyan. Research of path planning algorithms based on deep reinforcement learning [D]. Chengdu: University of Electronic Science and Technology of China, 2018.

[19] 李艳庆. 基于遗传算法和深度强化学习的多无人机协同区域监视的航路规划 [D]. 西安: 西安电子科技大学, 2018.
LI Yanqing. Route planning for multi-UAV cooperative area surveillance based on genetic algorithm and deep reinforcement learning [D]. Xi'an: Xidian University, 2018.

[20] 邓悟. 基于深度强化学习的智能体避障与路径规划研究与应用 [D]. 成都: 电子科技大学, 2019.
DENG Wu. Research and application of obstacle avoidance and path planning for agents based on deep reinforcement learning [D]. Chengdu: University of Electronic Science and Technology of China, 2019.

[21] 王毅然, 经小川, 田涛,等. 基于强化学习的多 Agent 路径规划方法研究 [J]. 计算机应用与软件, 2019, 36(8):7.
WANG Yiran, JING Xiaochuan, TIAN Tao, et al. Research on multi-agent path planning method based on re-

inforcement learning [J]. Computer applications and software, 2019, 36(8):7.

[22] 林桢祥. 基于深度增强学习的图像语句描述生成研究 [D].长沙：国防科技大学, 2017.
LIN Zhenxiang. Research on image description generation based on deep reinforcement learning [D]. Changsha: National University of Defense Technology,2017.

[23] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[J]. Computer science, 2013, 2: 45–66.

[24] 郭宪. 基于 DQN 的机械臂控制策略的研究 [D]. 北京：北京交通大学, 2018.
GUO Xian. Research on control strategy of manipulator based on DQN [D]. Beijing: Beijing Jiaotong University, 2018.

[25] 李季. 基于深度强化学习的移动边缘计算中的计算卸载与资源分配算法研究与实现 [D].北京：北京邮电大学, 2019.
LI Ji. Research and implementation of computing unloading and resource allocation algorithm in moving edge computing based on deep reinforcement learning [D]. Beijing: Beijing University of Posts and Telecommunications, 2019.

作者简介：

赵玉新，教授，博士生导师，工业和信息化部高技术船舶通信导航与智能系统专业组秘书长、中国航海学会理事、中国运筹学会决策科学分会常务理事 、IET(英国工程技术学会)Fellow、IEEE 高级会员，主要研究方向为水下导航技术及应用、业务化海洋学、智能航海技术。主持国防 973 课题、国家重大专项课题、国家自然科学基金等项目。发表学术论文 100 余篇，出版学术著作 4 部。

杜登辉，硕士研究生，主要研究方向为强化学习算法、海洋观测网。

刘延龙，博士研究生，主要研究方向为智能算法、业务化海洋学、海洋观测网。

《认知基础》正式出版

认知科学 (Cognitive Science) 是研究心智和智能的科学，包括从感觉的输入到复杂问题求解，从人类个体到人类社会的智能活动，以及人类智能和机器智能的性质。它是现代心理学、人工智能、神经科学、语言学、人类学乃至自然哲学等学科交叉发展的结果。认知科学研究的目的是解释人在完成认知活动时是如何进行信息加工的。认知科学的兴起标志着对以人类为中心的心智和智能活动的研究已进入到新的阶段，认知科学的发展将进一步为信息科学技术的智能化作出巨大贡献。认知基础系统地介绍认知科学的概念和方法，反映认知科学、脑科学、人工智能等领域的最新研究成果，综合地探索人类智能和机器智能的性质和规律。

由史忠植编著、机械工业出版社出版的《认知基础》，可作为大学本科和研究生的认知科学、认知心理学、认知信息学、智能科学、智能机器人等课程的教材，也可作为从事认知科学、认知心理学、认知信息学、认知神经科学、人工智能、智能科学、智能系统、智能控制、智能机器人等领域的研究人员参考书。