文章编号：2095-2163(2019)01-0014-06 中图分类号：TP18 文献标志码：A

水质监测无人船路径规划方法研究

吕扬民，陆康丽，王 梓
（浙江农林大学 信息工程学院，浙江 临安 311300）

摘 要：水质监测无人船对问题水域进行监测时，由于地形或天气原因使工作人员无法在视野中对水质监测无人船实时操作，需要 USV 的自主路径规划到指定水位进行检测。针对以上问题，本文提出利用无人船在未知水域中获得障碍物的分布信息，通过用 Q 学习算法对数据训练以规划路径，再利用 BP 神经网络的反馈进行权值的调整得到奖励值 R，反馈给 Q 学习算法进行 Q 值迭代，其过程中选择不同的动作方向使 Q 值达到最优，从而使路径达到最优。最后通过实验仿真验证了该算法收敛速度更快，有效地提高路径规划效率，证明了该无人船路径规划算法的可行性。

关键词：水质监测无人船；路径规划；BP 神经网络；强化学习

Research on path planning method of water quality monitoring USV

LV Yangmin, LU Kangli, WANG Zi
(College of Information Engineering, Zhejiang A&F University, Lin'an Zhejiang 311300, China)

【Abstract】 When the water quality monitoring USV monitors the problem waters, due to terrain or weather reasons, the staff cannot monitor the water quality monitoring of the USV in real time. The USV's autonomous path planning is required to detect the water level. In view of the above problems, this article proposes to use the USV to obtain the distribution information of obstacles in the unknown waters. By using the Q learning algorithm to train the data for planning the path, the BP neural network feedback is used to adjust the weight to get the reward value R. The feedback is given to the Q learning algorithm for Q-value iteration, in which different action directions are selected to optimize the Q value, so that the path is optimal. Finally, the experimental results show that the algorithm converges faster and effectively improves the path planning efficiency. The demonstrations in this paper prove the feasibility of the USV path planning algorithm.

【Key words】 water quality monitoring USV; path planning; BP neural network; reinforcement learning

0 引 言

水质监测是水质评价和预防水污染的主要方法。随着工业废水的增多，水体污染的问题则引发高度关注，水污染动态监测的研究已然刻不容缓。但是因为传统的水质监测方法步骤繁多、并耗时不菲，而且获取到的数据多样性、准确性也远远未能满足决策的需求[1]。基于上述问题，多种水质监测方法已陆续进入学界视野。如曹立杰等人[2]提出通过建立传感器网络，得到较为精准的水质反演模型。田野等人[3]提出通过水质模型对卫星数据进行反演，得到监测水域的水质参数分布图。但是以上方法却无法灵活地更换监测水域，工程量大、且时效性欠佳，相较而言水质监测无人船体积小便于携带、监测领域不受地形影响，能连续性原位进行多项水质参数监测，使监测结果更具有多样性和准确性。

无人驾驶船（Unmanned Surface Vehicle，USV）是一种能够在未知水域环境下自主航行，并完成各种任务的水面运动平台[4]，其研究内容主要涉及了自动驾驶、自主避障、航行规划和模式识别等热门方向[5]。故而，目前已广泛应用于军事领域的扫雷、侦察和反潜作战等方面，同时还可以用于民用领域的水文气象探测、环境监测和水上搜救等专项服务中[6-8]。但由于水质的流动性，可以流经多种复杂地形，如流经洞穴时等，工作人员将无法探测；或又由于天气的多变，如水域长期处于多雾天气，致使工作人员视线受阻，因而无法实时准确地掌控操作 USV。综合上述分析后可知，就可以利用 USV 的自主航行到达目标水位进行检测，而自主航行功能的实现即需用到本文下面拟将系统展开研究的路径规划技术。

USV 路径规划技术是指 USV 在作业水域内，按照一定性能指标（如路程最短、时间最短等）搜索得到一条从起点到目标点的无碰路径[9]，是 USV 导航技术中核心组成部分，同时也代表着 USV 智能化水准。目前常用的规划方法主要有粒子群算法[10]、

基金项目：浙江省重点研发计划项目（2015C0008）。

作者简介：吕扬民（1993-），男，硕士研究生，主要研究方向：机器学习。

收稿日期：2018-11-18

A * 算法[11]、可视图法[12]、人工势场法[13]、蚁群算法[14]等，但其方法多用于已知环境条件下。

当前对于已知环境下的航迹规划问题已经得到了较好的解决，但 USV 在未知水域作业执行任务之前却无法得到将要监测水域的环境信息，无法通过基于已知环境信息的路径规划方法去求出 USV 航行路线[15]。此外，由于监测水域环境复杂，传感器信息众多，系统的计算工作量大，致使 USV 存在实时性差、障碍物前振荡等缺点。因此 USV 路径规划亟需研究算法简单、实时性强、且能控制系统中的不确定现象的路径规划算法，所以有必要引入具有自主学习能力的方法，其中基于 Q 学习算法的路径规划适合于在未知环境中的路径规划。现阶段研究中，郭娜[16]即在传统 Q 学习算法基础上，采用模拟退火方法进行动作选择，解决探索与利用的平衡问题。陈自立等人[17]提出采用遗传算法建立新的 Q 值表以进行静态全局路径规划。董培方等人[18]把人工势场法加入 Q 学习算法中，以引力势场作为初始环境先验信息，再对环境逐层搜索，加快 Q 值迭代。

在此基础上，本文提出了一种基于 BP 神经网络的 Q 学习强化学习路径规划算法，以神经网络拟合 Q 学习方法中的 Q 函数，使其能够以连续的系统状态作为输入，并通过经验回放和设置目标网络方法显著提高网络在训练过程中的收敛速度。经过实验仿真，验证了本文所提出改进路径规划方法的可行性。

1　问题描述

USV 路径规划的实质是在一定标准下找出从初始位置到最终位置的最佳无碰撞安全路线。

USV 路径规划研究中，首先需要建立一个可航行的环境模型。假设 USV 的航行环境为存在着一定数量的静态障碍物的二维空间，采用栅格法对此区域进行分割。将栅格区域的左下角设为空间坐标系原点，水平向右为 X 轴，垂直向上为 Y 轴，划分为 $n * n$ 栅格坐标系。如此一来，该问题就简化为在静态环境中寻找从开始点到终点的无碰撞的最优路径。

设计中运用神经网络拟合函数，输入为 USV 的当前状态，USV 的状态是以当前位置来表示，即空间坐标 s；USV 路径规划的输出是下一时刻的转角，即动作 a。环境状态信息则是以障碍物的位置和大小及目标点的位置来表示，不同环境下障碍物出现的位置不同，目标点的设定位置也不同。网络的输出个数为动作空间的数量，每个输出表示在当前状态下，采取对应动作后的期望奖励大小 $R(s)$。对此内容可研究详述如下。

2　基于改进 Q 学习的无人船路径规划方法

2.1　基于传统 Q 学习的路径规划方法

Q 学习是基于马尔科夫决策过程（Markov Decision Process）来描述问题，通过 USV 与环境的互动积累经验，同时不断更新 USV 的策略，使其做出的决策能够获得更高的奖励。常用的强化学习方法包括模仿学习、Q 学习及策略梯度法等。而且进一步研究可知，Q 学习方法不需要收集训练数据，且能够生成决定性策略，因而适用于 USV 在未知水域进行路径规划问题。

马尔科夫决策过程包含 4 个元素，分别是：S，A，$P_{s,a}$，R。其中，S 表示 USV 所处的系统状态集合，即 USV 在当前的状态及当前环境的状态，如障碍物的大小和位置；A 表示 USV 所能采取的动作集合，即 USV 转动的方向；$P_{s,a}$ 表示系统模型，即系统状态转移概率，$P(s' \mid s,a)$ 描述了在当前状态 s 下，执行动作 a 后，系统到达状态 s' 的概率；R 表示奖励函数，由当前的状态和所采取的动作决定。把 Q 学习看成找到策略使综合评价最大的增量式规划，Q 学习的设计思想是不考虑环境因素，而是直接优化一个可迭代计算的 Q 函数，定义函数为在状态 s_t 时执行动作 a_t，且此后最优动作序列执行时的折扣累计强化值，即：

$$Q_{t+1}(s_t,a_t) = R(s_t) + \gamma \max_{a_t}\{Q(s_{t+1},a_t)\} \quad (1)$$

其中，γ 为折扣因子，其值 $0 \leqslant \gamma \leqslant 1$；$R(s_t)$ 为奖励函数，其值为正数或者负数。

在初始阶段学习中，Q 值可能是不正确地反映了其所定义的策略，初始 $Q_0(s,a)$ 对于所有的状态和动作假定是给出的。这里，若设给定环境的状态集合为 s，USV 可能的动作集合 A 选择性较多，数据量大，需要用到可观的系统内存空间，且无法被泛化。为了克服上述不足，对传统 Q 学习进行改进，采用 BP 神经网络实现 Q 值迭代，网络的输入对应描述环境的状态，网络的输出对应每个动作的 Q 值。

2.2　改进的 Q 学习路径规划算法

$Q(\lambda)$ 算法借鉴了 $TD(\lambda)$ 算法，通过回溯技术让数据不断地进行传递，使得某一状态的动作决策也会受到其后续状态的影响。如果未来某一决策 π

是失败的,那么当前的决策也要承担相应的惩罚,会把这种影响追加到当前决策;如果未来某一决策 π 是正确的,那么当前的决策也会得到相应的奖励,同样也会影响当前决策。结合改进后能够提高算法的收敛速度,满足学习的实用性。改进的 $Q(\lambda)$ 算法更新规则为:

$$Q(s_t,a_t)=Q(s_t,a_t)+\alpha\delta_t' \tag{2}$$

其中,α 为学习率,δ_t' 为 $TD(\lambda)$ 误差,其值为:

$$\delta_t'=R(s_t)+\gamma V(s_{t+1})-Q(s_t,a_t) \tag{3}$$

其中,值函数 $V(s)=\max_a Q(s,a)$。

另外,也可以把 $TD(0)$ 误差定义为:

$$\delta_{t+1}=R(s_{t+1})+\gamma V(s_{t+2})-V(s_{t+1}) \tag{4}$$

在这一过程中,也应用了折扣因子 $\lambda\in[0,1]$,并以此对将来步骤中的 TD 误差进行折扣,其数学公式可表示为:

$$Q(s_t,a_t)=Q(s_t,a_t)+\alpha\delta_t^{\lambda} \tag{5}$$

文中,$TD(\lambda)$ 的误差 δ_t^{λ} 将定义为:

$$\delta_t^{\lambda}=\delta_t'+\sum_{i=1}^{¥}(\gamma\lambda)^i\delta_{t+i} \tag{6}$$

只要将来的 TD 误差未知,前述的更新就无法进行。但是,通过使用跟踪迹就可以逐步求得其值。下面将 $\eta^t(s,a)$ 定义为特征函数:在 t 时刻 (s,a) 发生,则返回 1,否则返回 0。为了简化,忽略学习效率,对每个 (s,a) 定义一个跟踪迹 $e_t(s,a)$,如式(7)所示:

$$e_t(s,a)=\sum_{t=i}^{t-1}(\gamma\lambda)^{t-i}\eta^t(s,a)$$

$$e_t(s,a)=\begin{cases}1 & s=s_t,a=a_t\\ \gamma\lambda e_{t-1}(s,a) & \text{other}\end{cases} \tag{7}$$

那么在时刻 t 在线更新为:

$$Q(s,a)=Q(s,a)+\alpha[\delta_t'\eta^t(s,a)+\delta_t e_t(s,a)] \tag{8}$$

强化学习希望使系统运行时收获的总体收益期望最大,即 $E(R(s_0)+\gamma R(s_1)+\gamma^2R(s_2)+\cdots)$ 最大。为此需要找到一个最优策略 π,使得当 USV 依照 π 进行决策和运动时,获得的总收益最大。通常,强化学习的目标函数为以下其中之一:

$$\begin{aligned}V^{\pi}(s)&=E(R(s_0)+\gamma R(s_1)+\gamma^2R(s_2)+\cdots\mid s_0=s,\pi)\\ Q^{\pi}(s,a)&=E(R(s_0)+\gamma R(s_1)+\gamma^2R(s_2)+\cdots\mid s_0=s,a_0=a,\pi)\end{aligned} \tag{9}$$

其中,$V^{\pi}(s)$ 表示在当前初始状态 s 下,依照策略 π 的决策运动所能获得的期望收益;而 $Q^{\pi}(s,a)$ 表示在当前状态 s 下采取动作 a,之后所有的状态下都按照策略 π 的决策运动所能获得的期望收益。在 Q 学习中目的就是找到最优策略 π^*,使得 $\pi^*=\arg\max_{\pi}Q^{\pi}(s,a)$。

定义 $Q^*(s,a)=\max_{\pi}Q^{\pi}(s,a)$,$Q^{\pi}(s,a)$ 指在状态 s 时执行动作 a,而此后的所有状态下都遵循最优策略进行运动所能收获的期望收益大小。假设 $Q^*(s,a)$ 已知,那么可以很容易地由 $Q^*(s,a)$ 生成 π^*,只要对每个 s 均使得 $\pi^*(s)=\arg\max_{\pi}Q^*(s,a)$ 成立。这样,求取最优策略的问题就转化为求 $Q^*(s,a)$。由于有:

$$Q^*(s,a)=R(s_0)+\gamma E(R(s_1)+\gamma R(s_2)+\cdots\mid s_1,a_1) \tag{10}$$

且 a_1 由 π^* 决定,则:

$$a_1=\pi^*(s_1)=\arg\max_a Q^*(s_1,a) \tag{11}$$

那么,可得出:

$$Q^*(s_0,a_0)=R(s_0)+\gamma\max_a Q^*(s_1,a)+\alpha[\delta_1'\eta^1(s,a)+\delta_1 e_1(s,a)] \tag{12}$$

式(12)称为贝文曼方程。该方程是以递归的形式定义了 $Q^*(s,a)$,从而使得 Q 函数可以被迭代求出。

传统 Q 学习算法中,Q 函数是以表格的形式保存并更新,但在 USV 避障路径规划中,遭遇的障碍物可能出现在空间中任意位置,若以表格的形式 Q 函数将难以描述在连续空间中出现的障碍物。针对这一状况,本文在 Q 学习基础上,将展开深度 Q 学习以 BP 神经网络来拟合 Q 函数,其输入状态 s 是连续变量。通常,以非线性函数拟合 Q 函数时学习过程难以收敛,对此研究就采用了经验回放和目标网络的方法改善学习稳定性。

2.3 BP 神经网络结构与奖励函数设定

在强化学习中,奖励函数的设计将直接影响学习效果的好坏。通常,奖励函数对应着人对某项任务的描述,通过奖励函数的设计即可将任务的先验知识成功融入学习中。在 USV 路径规划中,本次研究在致力于使 USV 尽快到达目标位置的同时,还期望在航行过程中能够保证安全,并避免与障碍物相撞。为此本文将奖励函数分为 3 种,具体就是:USV 与目标位置的距离进行奖励、USV 到达目标位置进行奖励、USV 与障碍物相撞进行惩罚。文中,可将奖励函数写为如下数学形式:

$$R(s)=\begin{cases}-0.1 & \text{其它}\\ 10 & \text{无人船到达目标位置}\\ -10 & \text{无人船与障碍物相撞}\end{cases} \tag{13}$$

从量级上看,第一、二种的奖励值比第三的奖励值大。因为对于 USV 避障任务来说,其主要目标就

是避开障碍物且达到目标位置，而不是仅仅缩短 USV 与目标位置的距离。加入此项的原因在于，如果仅仅对 USV 达到目标位置和 USV 撞上障碍物进行奖励和惩罚，那么在运动过程中将会有大量的步骤所得奖励皆会为 0，这会使得 USV 在大部分情况下不会启用改进策略，学习效率偏低。而加入该项奖励相当于加入了人对此项任务的先验知识，使得 USV 在学习和探索时更有效率。综上可得，本文研发算法的整体设计流程如图 1 所示。

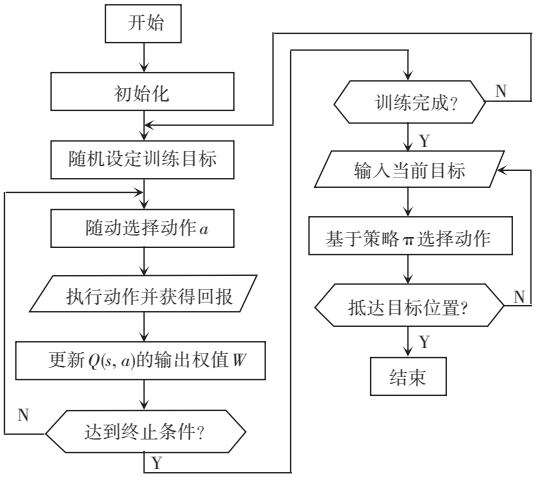


图 1　整体流程图

Fig. 1　Overall flow chart

由图 1 可见，对流程中各步骤可做阐释解析如下。

Step1　初始化经验回放存储区 D。

Step2　初始化 Q 网络，状态、动作赋初始值。

Step3　随机选择动作 a_t，得到当前奖励 r_t，下一时刻状态 s_{t+1}，将 (s_t, a_t, r_t, s_{t+1}) 存入 D。

Step4　从存储区 D 中随机采样一批数据 (s_t, a_t, r_t, s_{t+1}) 进行训练。当 USV 达到目标位置，或超过每轮最大时间时的状态都认为是最终状态。

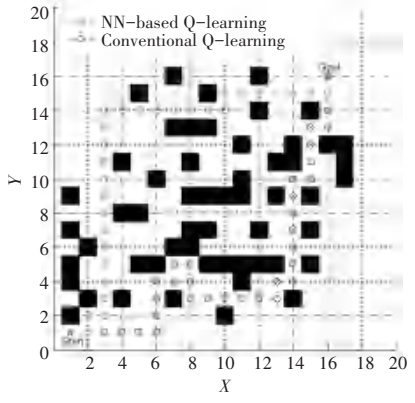
Step5　如果 s_{t+1} 不是最终状态，则返回 Step3；若 s_{t+1} 是最终状态，则更新 Q 网络参数，并返回 Step3。重复一定轮数后，算法结束。

D 为经验回放存储区，用来存储 USV 航行过程，并采集训练样本。经验回放的存在使得每次训练时的多个样本在时间上不是连续的，从而最小化样本之间的相关性，而且也增强了样本的稳定性和准确性。

3　实验仿真

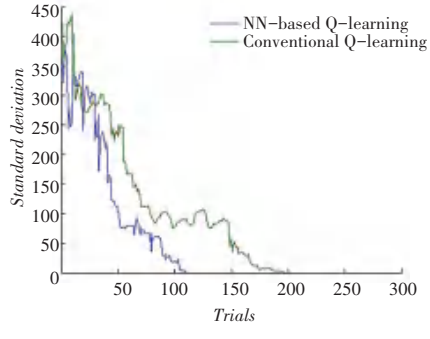
为了检验本文研发设计的路径规划算法性能，本文在 Matlab2014a 软件上进行仿真实验。在实验中，仿真环境为 20 * 20 的区域，折扣因子 γ 取值为 0.9，存储区 D 大小设为 40 000，循环次数 1 000，神经网络第一层有 64 个神经元，第二层有 32 个神经元。在训练的每一轮中，每当 USV 撞到障碍物或 USV 到达目标位置时，该轮都立即结束，并返回一个奖励。

为验证本文方法的准确性，将采用文献[16]中的迷宫地形来构建实验，但是由于文献[16]中的迷宫地形偏于简单，本文将设计 3 种不同地形来进行算法的比较，图 2 为复杂水域地形，图 3 为简易同心圆迷宫地形，图 4 为复杂迷宫地形。对本文改进算法与传统 Q 学习算法在以上地形进行仿真，由路径图可以看出，蓝色代表的改进算法路线相比传统 Q 学习算法仿真的路线，路径长度更短，更加简捷。由标准误差图可以看出，改进算法比传统 Q 学习算法提前三分之一进入收敛稳定状态。



(a)路径仿真图

(a) Map of path simulation

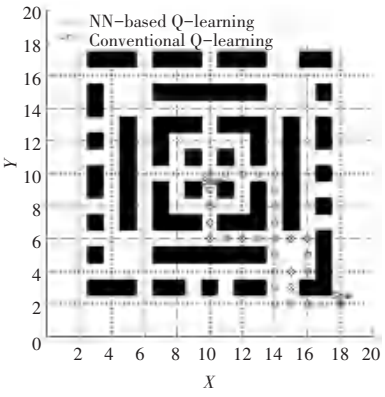


(b)误差分析图

(b) Map of standard error

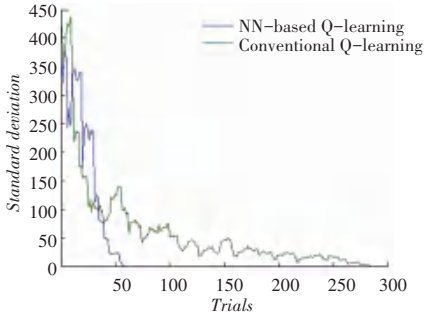
图 2　复杂水域地形仿真

Fig. 2　Map of complex water terrain simulation



(a)路径仿真图

(a) Map of path simulation

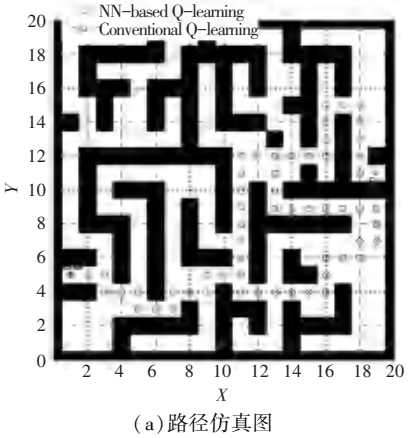


(b)误差分析图

(b) Map of standard error

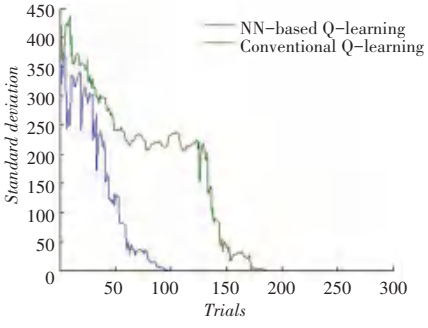
图 3 简易同心圆迷宫仿真

Fig. 3 Map of simple concentric circle maze simulation



(a)路径仿真图

(a) Map of path simulation



(b)误差分析图

(b) Map of standard error

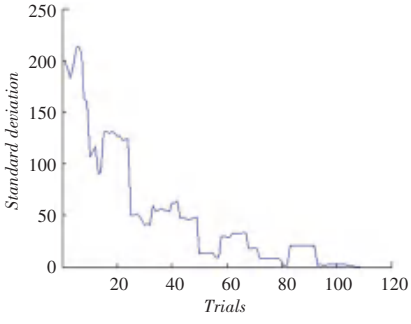
图 4 复杂迷宫仿真图

Fig. 4 Map of complex maze simulation

在前述基础上,再以临安东湖水域实际环境背景为例进行实验仿真。从图 5(a)中看出,USV 在仿真过程中并未出现与障碍物相撞且路径规划过程简单、且快捷。图 5(b)为标准误差曲线。由图 5 可以看出,在训练次数达到 56 次时,曲线趋于平稳,说明已经大致规划出一条安全快捷的整体路线,此时在多数情况下 USV 都能避开障碍物到达目标位置。由此可以推出如下结论,基于 BP 神经网络的改进 Q 学习算法比传统 Q 学习算法,学习收敛速度更快,路径更优化。

(a)路径仿真图

(a) Map of path simulation



(b)误差分析图

(b) Map of standard error

图 5 东湖背景仿真

Fig. 5 Map of east lake background simulation

4 结束语

本文用强化学习的方法解决水质监测无人船在未知水域进行水质监测时自主导航路径规划问题,通过 BP 神经网络拟合 Q 函数,在训练后即能根据当前环境中障碍物的实时信息做出正确决策。仿真结果表明,该方法能够使水质监测无人船在未知环境根据不同的状态规划出可行路径,决策时间短、路线更优化,而且能够满足在线规划的实时性要求,从而克服传统 Q 学习路径规划方法计算量大、收敛速度慢的缺点,能在第一时间实现问题水域的有效监测。

参考文献

[1] 李俊生. 高光谱遥感反演内陆水质参数分析方法研究-以太湖为例[D]. 北京:中国科学院,2007.

(下转第 23 页)