用于移动机器人路径规划的改进强化学习算法

张 威[1a,2,3],初泽源[1b],杨玉涛[1a],王 伟[1a]
(1. 中国民航大学 a. 航空工程学院;b. 安全科学与工程学院,天津 300300;2. 中国民航航空地面特种设备研究基地,天津 300300;3. 民航智慧机场理论与系统重点实验室,广州 510470)

摘 要:针对传统 Q-learning 算法规划出的路径存在平滑度差、收敛速度慢以及学习效率低的问题,本文提出一种用于移动机器人路径规划的改进 Q-learning 算法。首先,考虑障碍物密度及起始点相对位置来选择动作集,以加快 Q-learning 算法的收敛速度;其次,为奖励函数加入一个连续的启发因子,启发因子由当前点与终点的距离和当前点距地图中所有障碍物以及地图边界的距离组成;最后,在 Q 值表的初始化进程中引入尺度因子,给移动机器人提供先验环境信息,并在栅格地图中对所提出的改进 Q-learning 算法进行仿真验证。仿真结果表明,改进 Q-learning 算法相比传统 Q-learning 算法收敛速度有明显提高,在复杂环境中的适应性更好,验证了改进算法的优越性。

关键词:强化学习;路径规划;启发式奖励函数;Q 值初始化
中图分类号:TP249 **文献标志码**:A **文章编号**:1674-5590(2024)05-0059-07

Improved reinforcement learning algorithm for mobile robot path planning

ZHANG Wei[1a,2,3], *CHU Zeyuan*[1b], *YANG Yutao*[1a], *WANG Wei*[1a]
(1a. *College of Aeronautical Engineering*, 1b. *College of Safety Science and Engineering, CAUC, Tianjin* 300300, *China*;
2. *Aviation Special Ground Equipment Research Base, CAAC, Tianjin* 300300, *China*;
3. *Key Laboratory of Smart Airport Theory and System, CAAC, Guangzhou* 510470, *China*)

Abstract: Aiming at the problems of poor smoothness, slow convergence speed and low learning efficiency of the paths planned by the traditional Q-learning algorithm, this paper proposes an improved Q-learning algorithm for mobile robot path planning. Firstly, the density of obstacles and the relative position of the start point are considered to select the action set to accelerate the convergence speed of the Q-learning algorithm. Secondly, a continuous heuristic factor is added to the reward function, which consists of the distance between the current point and the end point, and the distance of the current point from all the obstacles in the map as well as the boundary of the map. Finally, a scale factor is introduced into the initialization process of Q-value table to give the mobile robot with a priori environment information, and the proposed improved Q-learning algorithm is simulated and verified in a raster map. The simulation results show that the convergence speed of the improved Q-learning algorithm is significantly improved compared with the traditional Q-learning algorithm, and its adaptability in complex environments is better, which verifies the superiority of the improved algorithm.

Key words: reinforcement learning; path planning; heuristic reward function; Q-value initialization

移动机器人路径规划问题是在有障碍物的环境中,在满足优化条件的前提下,寻找移动机器人从初始位置到期望位置的最优路径[1]。路径规划主要分为全局路径规划和局部路径规划。

局部路径规划是针对移动机器人只了解环境的部分信息,或者只根据传感器获取的信息不断更新环境信息情况下而进行的路径规划[2]。常见的局部路径规划算法有遗传算法[3-4]、人工势场法[5-6]、模糊逻辑算

收稿日期:2023-01-07; **修回日期**:2023-05-04
基金项目:国家自然科学基金民航联合研究基金重点项目(U2033208);天津市研究生科研创新项目(2021YJSS122)
作者简介:张威(1979—),男,湖南衡阳人,教授,博士,研究方向为民航智能装备、机器人学等.

法[7]、强化学习算法[8-9]等。其中，遗传算法收敛速度慢且收敛效果不稳定；人工势场法依赖于环境提供的信息，适应新环境的能力差；模糊逻辑算法的精度有限，适合与其他算法相结合作为初步的路径规划算法；强化学习算法的适应性好，能在陌生环境中快速学习，可通过与环境不断交互来获取动态数据[10]。强化学习算法主要有两大类：①基于直接搜索策略的强化学习算法[11]；②基于值函数的强化学习方法[12]，其核心是计算值函数的期望，无模型的强化学习算法是整个强化学习算法的核心[13]。

Q–learning 算法是目前移动机器人路径规划算法中最有效的强化学习算法[14]。Low 等[15]提出了一种将花授粉算法（FPA, flower pollination algorithm）与Q–learning算法相结合的融合算法，该方法用 FPA 适当初始化 Q 值，加快了 Q–learning 算法的收敛速度，但没有调整动作状态空间，机器人的可选动作有限，规划出的路径难以应用于复杂环境。宋勇等[16]提出利用人工势场初始化 Q 值的方法，使算法收敛速度有了一定提升，但未考虑路径的平滑度，规划出的路径中存在直角转折，无法直接用于指导机器人运动。张福海等[17]将移动机器人周围障碍物信息与目标点的位置离散成有限个状态，设计了连续的报酬函数，提高了算法训练效率，但没有赋予机器人先验知识。徐晓苏等[18]在 Q 值初始化的过程中引入人工势场法中的引力势场，调整移动机器人的动作状态空间，解决了路径平滑度较差的问题，提高了算法的训练速度，但并没有给机器人连续的奖励函数，使机器人无法获得即时的奖励反馈。

针对上述研究存在的不足，本文提出了一种改进 Q–learning 算法。首先，增加了机器人的动作状态空间，提高了路径平滑度；其次，引入了启发式奖励函数，提高了算法的收敛速度；最后，优化了 Q 值表的初始化进程，提高了机器人在训练初始阶段的学习效率。

1　Q–learning 算法原理

作为一种典型的无模型（model–free）[19] 算法，Q–learning 算法的基本思想是将移动机器人的经验存储在 Q 值表中，表中的值表示在某一状态下执行动作集中某一动作的长期奖励值。根据 Q 值表，Q–learning 算法可以告诉移动机器人在不同状况下选择哪个动作可以获得最大的预期回报，状态–动作值 $Q(s,a)$ 更新规则如下

$$Q^{\text{new}}(s_t,a_t)=\alpha[R(s_t,a_t)+\gamma\max_a Q^{\text{old}}(s_{t+1},a_t)-Q^{\text{old}}(s_t,a_t)]+Q^{\text{old}}(s_t,a_t) \tag{1}$$

式中：s_t 和 a_t 分别为当前状态和动作；s_{t+1} 和 a_{t+1} 为下一状态和动作；Q^{new} 和 Q^{old} 分别表示当前时刻的状态–动作值和上一时刻的状态–动作值；R 为状态 s_t 下执行动作 a_t 所获得的奖励；α 为学习率，代表利用已知信息的程度；γ 为折扣因子，代表考虑未来奖励的程度；α 和 γ 的取值范围应在 0 ~ 1 之间。

2　改进 Q–learning 算法

Q–Learning 算法主要基于 Q 值和 R 值来进行迭代。R 值是根据当前状态对不同的动作赋予相应的奖励，对地图中每个状态所可能发生的所有动作赋予的奖励就构成了奖励矩阵；Q 值是对状态动作综合评价的期望值，其由当前状态动作、下一状态动作和当前状态的最大奖励值来决定。改进 Q–learning 算法将主要围绕着优化 Q 值初始化过程以及加入启发式奖励函数来进行。

2.1　增加动作状态空间

相对于移动机器人初始位置（图 1 中黑色方块），传统 Q–learning 算法的动作集通常只包含 4 个动作，分别是上、下、左、右，如图 1 中深灰色方块所示，在规划路径时会出现路径转折过多、路径平滑度较差的情况，无法直接用于指导机器人的运动。针对这一问题，本文在传统算法的基础上增加了 4 个斜向动作，分别为左上、右上、左下、右下，如图 1 中浅灰色方块所示。

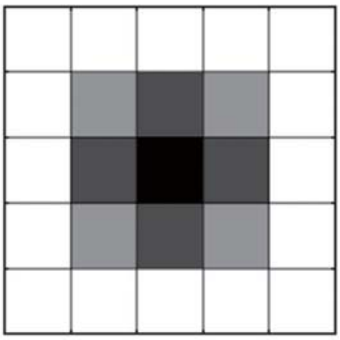


图 1　改进 Q–learning 算法的 8 个动作

Fig.1　Eight actions of the improved Q–learning algorithm

移动机器人在选择动作时，会有一定概率选择使其离终点更远的动作。为了提高算法的收敛效率，让其始终向着终点的方向前进，将 8 个动作分为 4 种动作集，如图 2 所示。

在选择动作集时，引入障碍物密度作为重要评价因素之一。障碍物密度的计算公式为

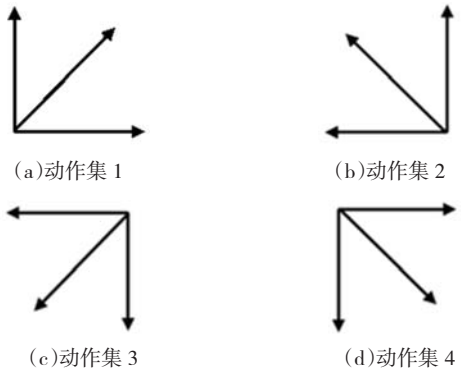


图 2　4 种动作集

Fig.2　4 kinds of action sets

$$f(\text{obs})=\begin{cases}\dfrac{N_{\text{obs}}}{L^2} & L\leqslant 3\\ 0 & L>3\end{cases} \tag{2}$$

式中：obs 表示障碍物；L 表示移动机器人当前位置到最近障碍物的距离；N_{obs} 表示 3 个步长范围内的障碍物个数。移动机器人通过扫描 3 个步长范围内的障碍物个数计算当前所处位置的障碍物密度，若 $f(\text{obs})\leqslant 0.5$，则按照以下规则选择动作集，即

$$\begin{cases}\text{选择动作集 1} & \sin\theta>0 \text{ 且 } \cos\theta>0\\ \text{选择动作集 2} & \sin\theta>0 \text{ 且 } \cos\theta<0\\ \text{选择动作集 3} & \sin\theta<0 \text{ 且 } \cos\theta<0\\ \text{选择动作集 4} & \sin\theta<0 \text{ 且 } \cos\theta>0\\ \text{到达终点} & x_{\text{now}}=x_{\text{end}} \text{ 且 } y_{\text{now}}=y_{\text{end}}\end{cases} \tag{3}$$

式中：$(x_{\text{now}},y_{\text{now}})$为移动机器人当前所在的坐标；$(x_{\text{end}},y_{\text{end}})$为终点所在坐标；$\theta$ 为当前点到终点的连线与水平线所成的夹角。

若 $f(\text{obs})>0.5$，则以行进方向为 y 轴正向，以左轮到右轮的方向为 x 轴正向，如图 3 所示，将扫描空间分为 4 个象限，其中 1、2、3 分别表示移动机器人扫描的 1 个步长范围、2 个步长范围、3 个步长范围。4 个象限与 4 种动作集相对应，分别计算从第一象限到第四象限的障碍物密度，取障碍物密度最小的象限对应的动作集为所选择的动作集。

2.2　改进奖励函数

在求解最优路径时通常都会采用离散的奖励函数来评判动作的优劣，其具体应用方式是：上位机在接收到移动机器人的动作后，根据其状态的更迭，给出相应的奖励信息；同时，移动机器人也会根据奖励机制来选择奖励值最大的动作，从而实现其对环境的学习与反馈。传统 Q-learning 算法在移动机器人训练的初始阶段，不清楚什么动作能获得最大的奖励值。这时，如果奖励反馈是离散的，那么移动机器人将无法充分利用所获得信息。所以，本文设计了一种连续的启发式奖励函数，定义了几种不同状态下的奖励 r，即

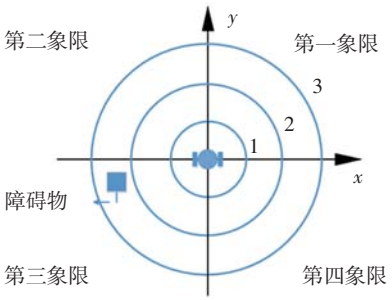


图 3　机器人扫描 3 个步长范围内障碍物

Fig.3　Robot scanning for obstacles within 3 steps

$$r=\begin{cases}-\infty & \text{遇到障碍物}\\ 0 & \text{其他状况}\\ 100 & \text{到达终点}\end{cases} \tag{4}$$

通过计算当前坐标点与终点坐标点的欧氏距离，为奖励函数引入了启发因子，欧式距离计算公式为

$$D=\sqrt{(x_{\text{now}}-x_{\text{end}})^2+(y_{\text{now}}-y_{\text{end}})^2} \tag{5}$$

奖励函数的计算方式为

$$R=\begin{cases}r+\eta\left(\dfrac{1}{D^2}m-\dfrac{1}{E^2}n\right) & \text{当前状态非终点状态}\\ 100 & \text{当前状态为终点状态}\end{cases} \tag{6}$$

式中：E 为当前点距地图边界以及地图中所有障碍物的欧氏距离的总和；η 为考虑启发因子影响效果的尺度系数；m 为考虑当前坐标与终点坐标的欧氏距离权重；n 为考虑当前坐标到地图边界以及所有障碍物的欧氏距离权重，且 $m+n=1$。为了不影响已有的奖励 r，同时保证算法随机探索的特性，经过反复测试后，m 取值为 0.7，n 值为 0.3。

启发式奖励函数能够在学习过程中给机器人持续性即时奖励，让机器人快速判断出在当前状态下动作集中最有价值的动作，使机器人每一步都始终向终点前进，提高了算法收敛效率。

2.3　优化 $\boldsymbol{Q}$ 矩阵初始过程

传统 Q-learning 算法中对 $\boldsymbol{Q}$ 矩阵的设计是设置一个统一的初始值，通常设置为 0 或 1，经过每次迭代后再去更新 Q 值。而机器人在最初开始学习的过程中，对环境的认知一片空白，不清楚什么动作能达到最佳状态，所以在很长一段时间中都处于盲目探索，这势必会浪费大量的时间。因此，借助合理的手段将 $\boldsymbol{Q}$ 矩阵的初始值进行优化，就能加快算法的收敛速度，提高算法运行效率。

针对这一问题，本文设计了一个函数。首先设置 $\boldsymbol{Q}$

矩阵的初始值为1,已知起点和终点的位置,将二维坐标转换为一维序列,由于起点随机,所以需要比较起点序列号的大小。根据起点与终点的相对位置,每个方格所对应的 Q 值都相应地增加或减少,优化 $\boldsymbol{Q}$ 矩阵初始过程的函数表示为

$$\begin{cases} Q^{\text{new}}(s_t,s_{t+1})=Q^{\text{old}}(s_t,s_{t+1})+\dfrac{\sqrt{(x_{\text{now}}-x_{\text{end}})^2+(y_{\text{now}}-y_{\text{end}})^2}}{|D_{\text{EP}}-D_{\text{SP}}|}c & D_{\text{SP}}<D_{\text{EP}} \\ Q^{\text{new}}(s_t,s_{t+1})=Q^{\text{old}}(s_t,s_{t+1})-\dfrac{\sqrt{(x_{\text{now}}-x_{\text{end}})^2+(y_{\text{now}}-y_{\text{end}})^2}}{|D_{\text{EP}}-D_{\text{SP}}|}c & D_{\text{SP}}>D_{\text{EP}} \end{cases} \tag{7}$$

$$\begin{cases} D_{\text{SP}}=(x_{\text{start}}-1)\,n+y_{\text{start}} \\ D_{\text{EP}}=(x_{\text{end}}-1)\,n+y_{\text{end}} \end{cases} \tag{8}$$

式中:SP表示起点;EP表示终点;D_{SP} 表示起点位置的序列号;D_{EP} 表示终点位置的序列号,$|D_{\text{EP}}-D_{\text{SP}}|$ 表示起点到终点距离的绝对值;$(x_{\text{start}},y_{\text{start}})$ 为起点坐标;c 为非零常数,是调整函数效果的偏移度。

优化初始 Q 值表能够让机器人在原本杂乱无序的试探中脱离开来,减少不必要的探索行为;同时,加快学习速度,强化高收益的行为。

综上所述,改进Q–Learning算法流程如下:

(1)载入环境信息,确定起点,同时赋给 m、n、α、γ 初始值,并将 Q 初始值设为1;

(2)根据式(7)优化 Q 值表;

(3)观察当前状态 s_t;

(4)继续在环境中探索,在当前状态下选择下一个动作 a_{t+1} 并执行,通过式(6)反馈,得到动作的即时奖励 R;

(5)上位机接收到动作后更新为新状态 s_{t+1},得到更新 Q 值为

$$Q(s_{t+1},a_{t+1})=(1-\alpha)Q(s_t,a_t)+\alpha(r+\gamma\max_{a_{t+1}}Q(s_{t+1},a_{t+1}))$$

(6)观察新状态 s_{t+1};

(7)判断机器人是否已经到达终点,或系统是否已经达到了设定的最大迭代次数,若满足其中一项,则学习结束,否则将返回步骤(3)继续学习。

3　仿真实验分析

3.1　实验环境

为验证本文提出的改进Q–learning算法的可行性与优越性,针对路径规划问题在二维栅格地图中进行仿真实验。基于实验平台MatlabR2020分别设计一个简单栅格地图和一个复杂栅格地图,将传统Q–learning算法与改进的Q–learning算法分别在两种环境中进行对比实验。简单栅格地图如图4所示,复杂栅格地图如图5所示。其中,深色方格代表障碍物,浅色区域代表可移动空间。

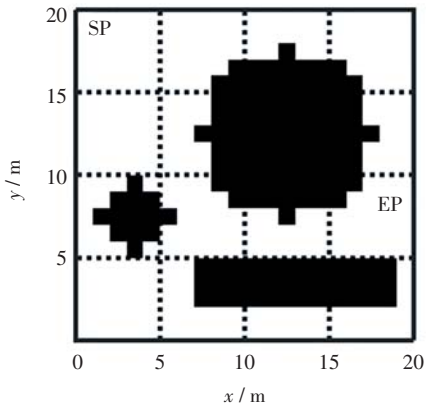


图4　简单栅格地图

Fig.4　Simple raster map

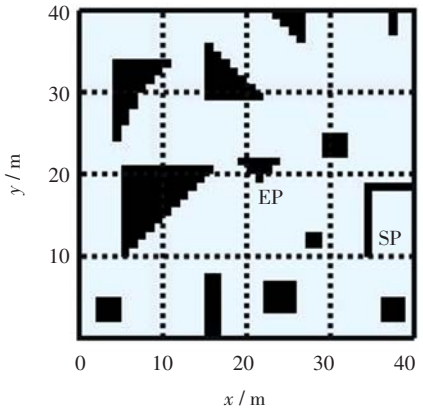


图5　复杂栅格地图

Fig.5　Complex raster map

在栅格地图上随机确定一组起点测试两种算法在未知环境中运行的效果,比较参数为平均迭代次数、平均路径长度和平均规划时间。算法中相关参数设置如下:学习率 $\alpha=0.6$,折扣因子 $\gamma=0.9$,到达终点的奖励值 $r_{\text{des}}=100$,遇到障碍物的奖励值 $r_{\text{obs}}=-\infty$,设定每次实验迭代次数上限为500次。每种算法在相同的地图上随机选择一组起点进行50次实验,取其中一次的迭代结果和所寻得的路径绘制对比图。

3.2　简单栅格地图实验结果与分析

首先,将传统Q–learning算法和改进Q–learning算法在简单栅格地图中进行训练,然后用训练得到的模型进行全局路径规划,最终,得到传统Q–learning算法和改进Q–learning算法规划出的全局路径。

传统Q–learning算法与改进Q–learning算法分别在图4环境中进行了50组实验,表1是两种算法在进行50组实验后所得到的参数平均值。由表1可知,

表 1　两种算法在简单栅格地图中路径规划的性能比较

Tab.1　Comparison of the path planning performance between 2 algorithms in simple raster map

算法	迭代次数/次	路径规划时间/s	路径长度/m
改进 Q-learning 算法	42.70	12.36	10.18
传统 Q-learning 算法	293.16	56.23	46.07

改进 Q-learning 算法相较于传统 Q-learning 算法在迭代次数、路径规划时间、路径长度方面分别提升了 85.43%、78.02%、77.90%。

取其中一次的迭代结果和所寻得的路径进行分析。图 6 为在图 4 仿真环境中得到的部分传统初始 Q 值以及优化后的部分 Q 值。

1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1

（a）传统 Q-learning 算法

1	13.580 4	1	1	1	1	1	1	1
11.767 0	1	14.740 7	1	1	1	1	1	1
1	13.666 1	1	14.413 1	1	1	1	1	1
1	1	13.319 4	1	13.735 3	1	1	1	1
1	1	1	13.096 9	1	13.524 0	1	1	1
1	1	1	1	12.786 5	1	12.998 2	1	1
1	1	1	1	1	11.895 1	1	11.013 0	1
1	1	1	1	1	1	10.874 9	1	10.311 6
1	1	1	1	1	1	1	8.468 4	1
1	1	1	1	1	1	1	1	9.097 6

（b）改进 Q-learning 算法

图 6　Q 值表的对比

Fig.6　Comparison of Q-value tables

由图 6(a)可知，传统 Q-learning 算法赋予 Q 值表的初值均为 1，即机器人对环境空间的先验知识为空白。所以在学习的初始阶段，机器人会花费大量的时间盲目探索。由图 6(b)中可知，通过重新计算起点到终点路线中所有途经方格的状态-动作值，能够给予机器人一定的先验信息，从而优化 Q 值的初始过程。

传统 Q-learning 算法和改进 Q-learning 算法在简单环境中的收敛过程如图 7 所示。

从图 7 可知，首先，传统 Q-learning 算法在寻找最短路径的迭代过程中，学习初期的曲线波动幅度较大，说明该算法利用新知识的水平较低，需要多次学习才能收敛；其次，由于没有合适的奖励函数，在选择动作时比较盲目，延长了探索时间。改进 Q-learning 算法由于获得了先验知识，在学习初期就能很快地收敛并找到最短路径；再次，由于连续的奖励函数，机器人能够即时判断出动作状态的优劣，从而选择最优的

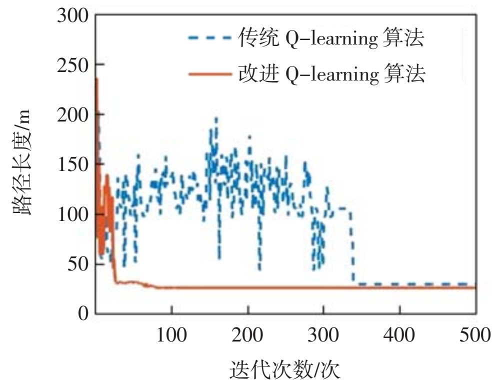
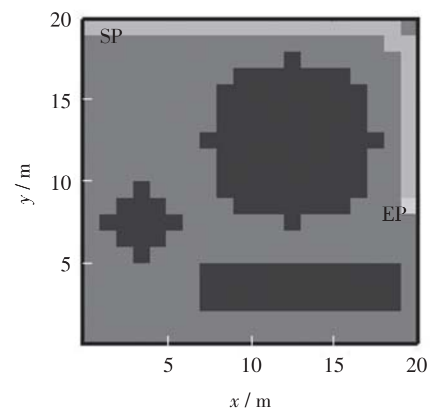


图 7　两种算法在简单环境中的收敛过程

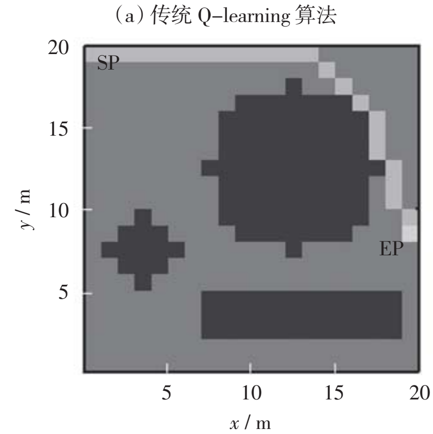
Fig.7　Convergence process of 2 algorithms in simple environment

动作，所以在收敛前的几次学习中路径长度波动较小，收敛稳定性较高。

传统 Q-learning 算法与改进 Q-learning 算法50 次实验其中一次所规划出的路径如图 8 所示。由图8 可知，改进 Q-learning 算法所规划出的路径与传统 Q-learning 算法规划出的路径相比长度缩短了 3.60 m。实验结果验证了改进后 Q-learning 算法的优越性。



（a）传统 Q-learning 算法



（b）改进 Q-learning 算法

图 8　两种算法在简单环境中规划出的路径

Fig.8　Paths planned by 2 algorithms in simple environment

3.3　复杂栅格地图实验结果与分析

为了进一步验证改进 Q-learning 算法在复杂环境

中的适应性，将两种算法在地图面积更大、障碍物更为复杂的 40 × 40 的栅格地图中进行 50 次仿真试验，实验环境如图 5 所示。

表 2 中的值为两种算法在复杂环境下进行 50 次实验得出的平均值。相较于传统 Q-learning 算法，改进 Q-learning 算法迭代次数平均缩减了 26.10%、路径规划时间平均减少了 30.90%、路径长度平均缩短了63.78%，进一步验证了改进 Q-learning 算法的优越性。

表 2　两种算法在复杂栅格地图中路径规划的性能比较

Tab.2　Comparison of the path planning performance between 2 algorithms in complex raster map

算法	迭代次数/次	路径规划时间/s	路径长度/m
改进 Q-learning 算法	105.78	183.29	23.44
传统 Q-learning 算法	143.14	265.24	64.71

两种算法在复杂环境中的收敛过程如图 9 所示。由图 9 可知，改进 Q-learning 算法在复杂环境中的适应度更好，通过优化初始 Q 值表，加快了其在训练前期的收敛速度，启发式奖励函数使算法更精确地选择下一步动作，缩短了探索时间，提高了收敛速度。

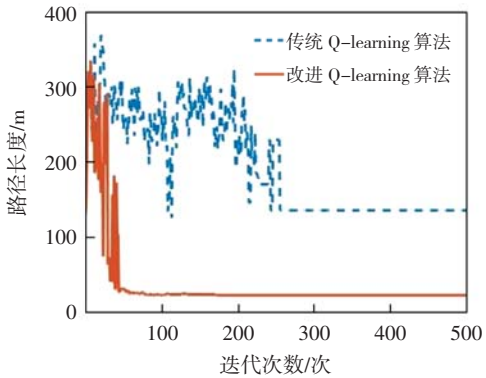
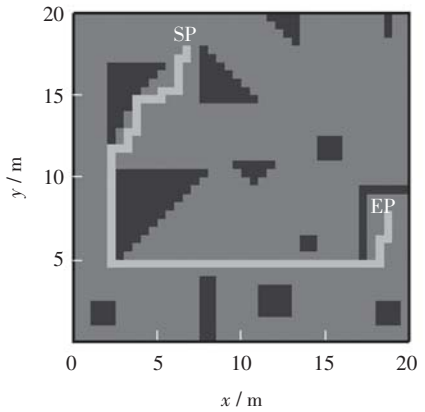


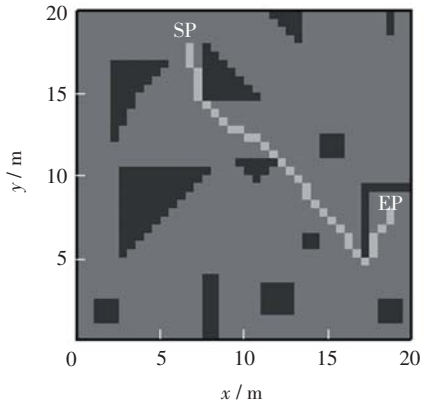
图 9　两种算法在复杂环境中的收敛过程

Fig.9　Convergence process of 2 algorithms in complex environment

图 10 为取 50 次实验其中一次的实验结果绘制的对比图，两种算法都各自规划出了一条从起点到终点最短的路径，但改进 Q-learning 算法所规划出的路径更为平滑。由于传统 Q-learning 算法只有 4 个动作，所以在规划路径时就会出现较多的冗余路径，这会延长路径规划的时间，这些冗余路径造成的额外转角还会影响机器人的实际行走。与传统 Q-learning 算法相比，改进 Q-learning 算法增加了 4 个斜向的动作，这有效地避免了冗余路径的产生，给机器人提供了更短的路线。其次，通过启发式函数持续给机器人的每个动作不同的奖励，可以加快机器人的学习速度，避免在环境中盲目的探索，让机器人的每一步都向着终点的方向前进，算法收敛更快。



（a）传统 Q-learning 算法



（b）改进 Q-learning 算法

图 10　两种算法在复杂环境中规划的路径

Fig.10　Paths planned by 2 algorithms in complex environment

综合两次实验结果，改进 Q-learning 算法比传统 Q-learning 算法在迭代次数上平均减少了 55.77%、路径长度平均缩短了 70.84%、规划时间平均降低了 54.46%。

4　结语

本文围绕移动机器人的路径规划问题，提出了一种用于移动机器人路径规划的改进强化学习算法。在传统 Q-learning 算法的基础上调整了动作状态空间、增加了启发式函数、优化了 $\boldsymbol{Q}$ 矩阵初始过程。实验结果表明，改进 Q-learning 算法在迭代次数、路径长度、规划时间和路径平滑度上都有显著提升，能够提高路径规划的效率、加快收敛速度，且更适合复杂环境。后续工作将围绕真实移动机器人来展开，考虑移动机器人的运动学特性和底盘的角加速度限制，使得移动机器人更适应随机变化的复杂环境。

参考文献：

[1] 朱大奇, 颜明重. 移动机器人路径规划技术综述[J]. 控制与决策, 2010, 25(7): 961–967.

[2] SARIFF N, BUNIYAMIN N. An overview of autonomous mobile robot path planning algorithms[C]//2006 4th Student Conference on Research and Development, June 27–28, 2006, Shah Alam, Malaysia. IEEE, 2006: 183–188.

[3] ELSHAMLI A, ABDULLAH H A, AREIBI S. Genetic algorithm for dynamic path planning[C]//Canadian Conference on Electrical and Computer Engineering 2004, May 2–5, 2004, Niagara Falls, ON, Canada. IEEE, 2004: 677–680.

[4] 段俊花, 李孝安. 基于改进遗传算法的机器人路径规划[J]. 微电子学与计算机, 2005(1):70–72, 76.

[5] 朱　毅, 张　涛, 宋靖雁. 非完整移动机器人的人工势场法路径规划[J]. 控制理论与应用, 2010, 27(2): 152–158.

[6] BOUNINI F, GINGRAS D, POLLART H, et al. Modified artificial potential field method for online path planning applications[C]//2017 IEEE Intelligent Vehicles Symposium(IV), June 11–14, 2017, Los Angeles, CA, USA. IEEE, 2017: 180–185.

[7] 李　擎, 张　超, 韩彩卫, 等. 动态环境下基于模糊逻辑算法的移动机器人路径规划[J]. 中南大学学报(自然科学版), 2013, 44(S2): 104–108.

[8] BAI Y F, DING X F, HU D S, et al. Research on dynamic path planning of multi–AGVs based on reinforcement learning[J]. Applied Sciences, 2022, 12(16): 8166.

[9] CAMPBELL J S, GIVIGI S N, SCHWARTZ H M. Multiple model Q–learning for stochastic asynchronous rewards[J]. Journal of Intelligent & Robotic Systems, 2016, 81(3): 407–422.

[10] 王　珂, 卜祥津, 李瑞峰, 等. 景深约束下的深度强化学习机器人路径规划[J]. 华中科技大学学报(自然科学版), 2018, 46(12): 77–82.

[11] 刘朝阳, 穆朝絜, 孙长银. 深度强化学习算法与应用研究现状综述[J]. 智能科学与技术学报, 2020, 2(4): 314–326.

[12] 刘建伟, 高　峰, 罗雄麟. 基于值函数和策略梯度的深度强化学习综述[J]. 计算机学报, 2019, 42(6): 1406–1438.

[13] 王子强, 武继刚. 基于 RDC–Q 学习算法的移动机器人路径规划[J]. 计算机工程, 2014, 40(6): 211–214.

[14] RAJA P. Optimal path planning of mobile robots: a review[J]. International Journal of Physical Sciences, 2012, 7(9): 1314–1320.

[15] LOW E S, ONG P, CHEAH K C. Solving the optimal path planning of a mobile robot using improved Q–learning[J]. Robotics and Autonomous Systems, 2019, 115: 143–161.

[16] 宋　勇, 李贻斌, 李彩虹. 移动机器人路径规划强化学习的初始化[J]. 控制理论与应用, 2012, 29(12): 1623–1628.

[17] 张福海, 李　宁, 袁儒鹏, 等. 基于强化学习的机器人路径规划算法[J]. 华中科技大学学报(自然科学版), 2018, 46(12): 65–70.

[18] 徐晓苏, 袁　杰. 基于改进强化学习的移动机器人路径规划方法[J]. 中国惯性技术学报, 2019, 27(3): 314–320.

[19] 尹　旷, 王红斌, 方　健, 等. 基于强化学习的移动机器人路径规划优化[J]. 电子测量技术, 2021, 44(10): 91–95.

（责任编辑：刘智勇）