DOI:10.19651/j.cnki.emt.2415579

基于改进 Q-Learning 的移动机器人路径规划算法*

王立勇 王弘轩 苏清华 王绅同 张鹏博
(北京信息科技大学现代测控技术教育部重点实验室 北京 100192)

摘 要: 随着移动机器人在生产生活中的深入应用,其路径规划能力也需要向快速性和环境适应性兼备发展。为解决现有移动机器人使用强化学习方法进行路径规划时存在的探索前期容易陷入局部最优、反复搜索同一区域,探索后期收敛率低、收敛速度慢的问题,本研究提出一种改进的 Q-Learning 算法。该算法改进 Q 矩阵赋值方法,使迭代前期探索过程具有指向性,并降低碰撞的情况;改进 Q 矩阵迭代方法,使 Q 矩阵更新具有前瞻性,避免在一个小区域中反复探索;改进随机探索策略,在迭代前期全面利用环境信息,后期向目标点靠近。在不同栅格地图仿真验证结果表明,本文算法在 Q-Learning 算法的基础上,通过上述改进降低探索过程中的路径长度、减少抖动并提高收敛的速度,具有更高的计算效率。

关键词: 路径规划;强化学习;移动机器人;Q-Learning 算法;ε-decreasing 策略
中图分类号: TN711.1;TP242.6 **文献标识码:** A **国家标准学科分类代码:** 520.60

Path planning algorithm of mobile robot based on improved Q-Learning

Wang Liyong Wang Hongxuan Su Qinghua Wang Shentong Zhang Pengbo
(Key Laboratory of Modern Measurement and Control Technology, Ministry of Education, Beijing Information Science and Technology University, Beijing 100192, China)

Abstract: With the in-depth application of mobile robot in production and life, its path planning ability also needs to develop to both rapidity and environmental adaptability. In order to solve the problems existing in the existing mobile robot path planning using reinforcement learning methods, which are easy to fall into local optimization in the early stage of exploration, repeatedly search the same area, and explore the late convergence rate and slow convergence rate, an improved Q-Learning algorithm is proposed in this study. The algorithm improves the Q matrix assignment method to make the exploration process directional in the early iteration and reduces the collision situation; the Q matrix iterative method is improved to make the Q matrix update forward-looking and avoid repeated exploration in a small area; the random exploration strategy is improved to make full use of environmental information in the early iteration and close to the target point in the later stage. The simulation results of different raster maps show that the algorithm in this paper has higher computational efficiency by reducing the path length, reducing jitter and improving the speed of convergence based on the Q-Learning algorithm.

Keywords: path planning;reinforcement learning;mobile robot;Q-Learning algorithm;ε-decreasing strategy

0 引 言

随着计算机技术的快速发展,移动机器人技术成为计算机技术的重要载体,并且在人工智能领域的多项研究中起到重要作用,具有广泛的应用领域和社会价值。在封闭环境中,移动机器人可实现货仓物流车辆自动化物料配送,降低人工劳动在整体劳动强度中所占比例的同时,大大提高工作效率;在危险环境中,移动机器人可以代替人工执行特殊任务,降低危险作业的伤亡率。移动机器人的自主行驶过程主要分为 3 个部分:环境感知、路径规划、横纵向控制。路径规划是处理环境信息的结果,是控制的基础,其目的是找到一条链接起始点和目标点的完整路径,并满足与环境中障碍不发生碰撞、长度尽可能短、符合移动机器人的动力学要求[1]。

路径规划方法有基于图搜索的算法[2]、基于仿生学的算法[3-4]、基于势场的算法[5-6]、基于速度空间的算法[7]、基于采样的算法[8-9]等,各种路径规划算法在处理不同环境时有其优劣势。标准 Q-Learning 算法在进行路径规划时,由

收稿日期:2024-03-05
* 基金项目:基础加强计划基金(2021JCJQJJ0022)、国家自然科学基金(52175074)项目资助

于对各种规划陷阱处理不当，存在收敛率较低、收敛速度慢、学习时间长等问题[10]。田晓航等[11]引入蚁群算法(ant colony optimization，ACO)的信息素机制，提出一种寻优范围优化方法，减少智能体的无效探索次数，并结合信息素浓度，设计了一种动态调整探索因子的方法，以提升在不同环境中算法的有效性和可行性。薛颂东等[12]提出一种改进 ACO 算法与动态 Q-Learning 融合的机器人路径规划算法，利用信息素矩阵为 Q 矩阵赋值，减少机器人初期的无效探索，并设计一种动作动态选择策略，提高算法的收敛速度。此类结合 ACO 算法的 Q-Learning 算法需要调节的参数较多，对复杂环境中的障碍物变化较为敏感。Li 等[13]提出了一种改进 Q-Learning 算法，将 ε-greedy 探索策略与玻尔兹曼探索相结合，并且结合启发式搜索策略来减小搜索空间并限制方向角的变化范围。尹旷等[14]针对 Q-Learning 算法中探索与利用的平衡问题提出了 ε-decreasing 策略，使随机探索概率随 Q 矩阵的更新呈下降的趋势，从而主导探索前期全面的感知环境，后期快速收敛至目标状态。但是探索前期具有较大的盲目性，增加了不必要的计算量。Shi 等[15]提出一种基于先验知识的改进 Q-Learning 算法。利用先验知识初始化 Q 值，从而引导智能体从算法初期以更大的概率向目标方向运动，消除大量无效迭代。段建民等[16]引入环境的势能值作为搜索启发信息对 Q 值进行初始化，从而在学习初期便能引导移动机器人快速收敛，改善了传统强化学习过程的盲目性，更大程度上利用了已知地图环境中的先验信息。王慧等[17]借鉴增加学习层和先验知识启发搜索的思想，利用不同栅格在地图中的位置为 Q 矩阵初始化，并且根据不同的状态在构建的标量场中的位置获得不同的奖赏值使得探索动作具有一定的指向性。此类结合已知环境中先验信息的 Q-Learning 算法在探索前期迭代效率较高，同时在复杂度较高的栅格地图环境下容易陷入局部最优。李威等[18]在标准 Q-Learning 算法中引入一种经验记忆力机制，增加经验记忆力表和指令表提高收敛的速度。并通过引入双重奖励机制，防止机器人在未知环境中的盲目搜索和过度搜索，解决算法的过估计问题。Meng 等[19]提出了一种基于势能函数的反思性奖励设计方法，对比当前决策和历史决策序列的关系的优劣性，使用反思性奖励协助完成值函数的更新，使智能体在充分利用奖赏信息后具有人类学习中的反射特性。

本文在标准 Q-Learning 算法的基础上，借鉴启发式搜索的思想，提出一种改进 Q-Learning 算法。利用环境地图中的先验信息对栅格进行评价，并根据可通行区域与障碍物的位置关系设计衰减函数作为初始 Q 矩阵，以此将各部分栅格加以区分，降低距离目标点较远的栅格的搜索优先级，一定程度规避无效搜索，从整体上降低迭代过程中的搜索次数；分析连续的两次动作选择后产生的收益，避免当前动作收益较大而陷入局部最优的情况，从而减少无效搜索；在动作选择环节设计改进 ε-decreasing 策略，使探索过程兼顾前期的全面性和接近收敛时的指向性，以迭代过程初期搜索次数上升为代价全面评估环境信息，并在迭代过程后期快速向目标点靠近，减少收敛过程的抖动。

1　标准 Q-Learning 路径规划方法

强化学习的核心机制在于通过持续与环境交互，即探索环境和利用环境的过程。在试错的过程中通过预设的奖励函数逐步调整和优化智能体的行为策略。这一过程旨在发现最佳策略，确保智能体能够获取最大的累积奖励，进而达成最大化奖励或实现预设目标的目的。在强化学习算法中，马尔可夫决策过程(markov decision process，MDP)为其提供了基础模型框架[20]，智能体的状态、动作、策略以及奖励构成了 4 个关键要素。其基本过程如图 1 所示。

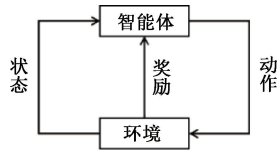


图 1　强化学习算法基本过程

在基于值的强化学习算法中，Q-Learning 算法具有简洁直观的特点，可以由工程师调整算法的迭代函数和学习率等参数。其显著优点包括参数需求少、无需预先构建环境模型，并支持离线操作，因此 Q-Learning 算法在强化学习路径规划算法的场景中有较多应用。如果能够合理利用传统路径规划方法加以结合，就能展现出快速的收敛速度和较高质量的路径规划能力。Q-Learning 算法的核心在于构建 Q 矩阵，其中每个元素代表在特定状态下采取特定行动所能获得的最大期望累积回报。智能体通过随机探索策略和 Q 矩阵的指导，在每个状态下做出最佳行动选择。随着迭代过程的进行，Q 矩阵逐渐稳定，进而通过其值分布确定最终路径。Q-Learning 算法的流程如图 2 所示。

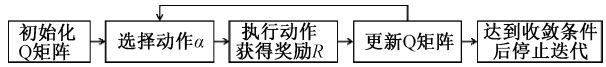


图 2　Q-Learning 算法

智能体依靠奖励函数 $R(s)$ 得到最优策略：其中，s 为智能体的状态集，并且根据迭代过程初始状态确定其 s_0。每一个状态下含有一组动作集 a，包含当前状态下智能体所有可选择的动作。智能体根据策略 π 选择动作，并按照状态转换映射关系 $s'=P(s,a)$ 到达下一状态 s'，此时根据奖励函数 $R(s)$ 获得环境反馈的奖励 R。

使用 $Q(s,a)$，即以状态－动作对的值作为函数，学习过程不断迭代以使 Q 矩阵收敛。$Q(s,a)$ 如式(1)所示。

$$Q^*(s,a)=R(s,a)+\sum T(s,a,s')\max Q^*(s',a') \tag{1}$$

其中，$Q^*(s,a)$ 是在 s 状态下采取行动 a 得到的最优

奖励值之和,定义 $V^*(s)$ 为该状态下的最优值函数。如式(2)所示。

$$V^*(s) = \max Q^*(s,a) \tag{2}$$

根据贝尔曼方程,在算法决策过程中求解最优决策序列。$V^\pi(s)$ 代表在 s 状态下运用策略选择一个动作所获得的奖励。如式(3)所示。

$$V^\pi(s) = E_{i\sim\pi(s)}[r(s,\pi(s)) + \gamma V^\pi(s')] \tag{3}$$

Q-Learning 算法的迭代公式和 ε-greedy 策略如式(4)和(5)所示。

$$Q'(s,a) = Q(s,a) + \alpha(r + \gamma \max Q(s',a') - Q(s,a)) \tag{4}$$

$$prob(a_i) = \begin{cases} 1-\varepsilon, & a = \operatorname{argmax} Q(s,a_i) \\ \varepsilon, & a \neq \operatorname{argmax} Q(s,a_i) \end{cases} \tag{5}$$

其中,考虑由 ε-greedy 策略在状态 s 下选择动作 a,转移到状态 s',得到奖励 R,α 为学习率,γ 为折扣率,ε 为探索因子,表示移动机器人随机探索概率。

在 Q-Learning 算法中,首先需要初始化 Q 矩阵。然后,智能体基于 ε-greedy 策略来选择在不同状态下的动作,通过前期设计 ε 的大小决定智能体选择预期回报最大动作的概率。一旦智能体执行某个动作并达到新的状态,它便会利用 $Q(s,a)$ 来获取实际的迭代值,并据此更新 Q 矩阵。当智能体顺利达到目标状态时,本轮迭代结束。此时,智能体会返回至初始状态,并开启下一轮的迭代过程。这一循环往复的过程将持续进行,直至智能体完成整个学习过程,从而不断优化其决策策略。

2 改进 Q-Learning 算法

标准 Q-Learning 算法及部分改进算法仍存在一些问题:

1)Q 矩阵的初值选取与环境地图不匹配,导致探索初期的动作选择较为盲目;

2)Q 矩阵更新时只考虑当前动作收益,未评估后续状态的缺陷,容易在某一状态附近发生过多抖动;

3)探索因子和探索阶段不匹配,导致探索前期容易陷入局部最优,探索后期不易收敛。

本章提出 Q 矩阵初值选取、Q 矩阵更新策略以及动态探索因子设计 3 部分优化方案。

2.1 Q 矩阵初值选取

通过在进入 Q-Learning 学习的初始阶段设置带有一定指向性的 Q 矩阵,实现对标准 Q-Learning 算法的改进:为可以通行的栅格根据其与起始点、目标点和障碍物的位置关系赋予一定的 Q 值。这可以避免 Q-Learning 算法在探索过程前期的盲目性,并在后续探索过程中在一定程度上对靠近障碍物的栅格起过滤作用。具体工作如下:

假设栅格地图横纵坐标方向上长度比例保持一致、每个栅格均为边长为 1 的正方形,如图 3 所示。其中,黄色栅格是起始点 s,紫色栅格是目标点 g,黑色栅格是障碍物 o,白色栅格是可通行区域 k。

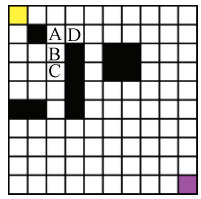


图 3 栅格地图

分别计算每个栅格与起始点、目标点之间的欧氏距离 $D(k,s)$、$D(k,g)$,如式(6)和(7)所示。

$$D(k,s) = \sqrt{(x_k - x_s)^2 + (y_k - y_s)^2} \tag{6}$$

$$D(k,g) = \sqrt{(x_k - x_g)^2 + (y_k - y_g)^2} \tag{7}$$

根据每个栅格与起始点、目标点的位置关系,对地图中的栅格进行初步区分,如式(8)所示。

$$D_0(k) = \frac{1}{D(k,s)} + \frac{\eta}{D(k,g)} \tag{8}$$

其中,$\eta \in (1,10)$,用于调整目标点对移动机器人在探索初期的吸引作用。此初值保证越靠近起始点和目标点连线的栅格 Q 值越大,并且可以人为调整探索初期向目标点靠近的趋势,以此引导机器人寻找长度较短的成功路径。

由于栅格地图中存在障碍物区域,而特殊的障碍物位置分布会形成震荡类型的“规划陷阱”。此类“规划陷阱”形成的原因是机器人进入障碍物相距较近而形成的小入口中,小入口内靠近目标点的一侧又有障碍物阻挡,将退出小入口当作当前状态下的收益最大的动作,以此多次往复之后才会远离小口。在图 3 中,初始时刻机器人位于起始点,先向右行进 2 次,再向下行进 1 次到达点 A 处时,认为向下行进至点 B 处为回报最大的动作;在点 B 处时认为向上返回至点 A 处为回报最大的动作。即使是在 ε-greedy 策略下,有一定可能逃离循环,也有可能到达点 C 处,在下一次行动选择时进入循环,或者到达点 D,可能进入到右侧的“规划陷阱”中。

为减少这种情况对规划过程带来的负面影响,分析处于可通行区域的栅格,根据与其曼哈顿距离小于 5 的区域内栅格的数量和距离对此栅格的 $D_0(k)$ 衰减至 $D'(k)$,如式(9)所示。

$$D'(k) = D_0(k) \times \prod_{i=1}^{N_o}[1 - \mu(\delta - D(k,o_i))] \tag{9}$$

式中:N_o 为与此可通行栅格曼哈顿距离小于 5 的障碍物个数,$\mu \in (0,0.05)$,用于调整衰减的程度;$\delta \in \left(1,\frac{1}{\mu}\right)$,用于调节衰减度上限,$D(k,o_i)$ 为此可通行栅格到障碍物 i 的曼哈顿距离,如式(10)所示。

$$D(k,o_i) = |x_k - x_{oi}| + |y_k - y_{oi}| \tag{10}$$

引入式(9)对 Q 矩阵进行调整,可以使距离障碍物越近、附近障碍物越多的状态对应的 Q 值受到越大的衰减,引导机器人远离障碍物形成的小入口,从而减少陷入“规划

陷阱”的可能性。

使用最大最小标准化函数对 Q 矩阵进行归一化处理，将衰减后的每一个 Q 值 $D'(k)$ 线性地转化到 $D(k)\in(0,1)$，作为每一个栅格初始的 Q 值，即：$Q(s,a)=D(k)$。最大最小标准化函数如式(11)所示。

$$x'=\frac{x-\min(x)}{\max(x)-\min(x)} \tag{11}$$

2.2　Q 矩阵更新策略

由式(4)可以分析得到，标准 Q-Learning 算法根据 Q 矩阵作为指导选择回报值最高的动作时，只考虑当前动作集合中的最佳动作，并未考虑下一个时刻的情况。在一个已经对可通行栅格进行初始化，赋予其一定的回报值的栅格地图中，假定当前状态所选择的一个动作后，了解执行此动作后所在状态动作集合的所有回报值是十分容易的。

本文算法将假设的下一状态的动作集合考虑到当前所选动作的回报中，需要根据当前状态所选动作后的状态以及再下一步动作后的状态对各种情况进行分析。

1)执行当前状态所选动作后机器人处于障碍物或目标点，此时的回报值应为设定的 R 值，如式(12)所示。

$$Q(s,a)=Q(s,a)+\alpha[R(s,a)-Q(s,a)] \tag{12}$$

2)执行当前状态所选动作后机器人处于可通行区域，执行下一状态所选动作后机器人处于障碍物或目标点，此时的回报值应为 Q 矩阵中的 Q 值，下一状态所选动作的回报值应为设定的 R 值，如式(13)所示。

$$Q(s,a)=Q(s,a)+\alpha[R(s,a)+\gamma(\omega\max_{a'\in A}Q(s',a')+(1-\omega)R(s',a'))-Q(s,a)] \tag{13}$$

3)执行当前状态所选动作后机器人处于可通行区域，执行下一状态所选动作后机器人仍处于可通行区域，此时以及下一状态所选动作的回报值均应为 Q 矩阵中的 Q 值如式(14)所示。

$$Q(s,a)=Q(s,a)+\alpha[R(s,a)+\gamma(\omega\max_{a'\in A}Q(s',a')+(1-\omega)\max_{a''\in A'}Q(s'',a''))-Q(s,a)] \tag{14}$$

式中：$\omega\in(0.5,0.8)$，为平衡当前状态所选动作的回报值与下一状态所选动作回报值的比例，a' 为下一状态所选动作，由式(15)确定。

$$a'=\{a_0\mid a_0=\operatorname{argmax}Q(s,a_0),a_0\in A\} \tag{15}$$

2.3　动态探索因子设计

在进行 Q-Learning 算法迭代过程中，为了得到一条优质的路径，需要同时兼顾两方面的内容：探索环境信息的全面性以及收敛的快速性。所以，要尽可能地多尝试动作集中的各种动作，并且还需要重视价值函数最大的动作所带来的高回报。

具体实现方法如下：在迭代的初始阶段，智能体对环境信息的了解为零，此时应该以全面探索环境信息为主，随着对环境信息的不断感知，智能体需要向最优解的迭代方向靠近，此时应该选择回报最大的动作以尽快得到最优解。因此，在每一次迭代过程中，ε 应在一定的次数内保持较高的水平，随着对环境的进一步感知而有所下降。如式(16)所示。

$$\varepsilon=\mathrm{e}^{-\frac{|(1-\xi)(Q_{\max}(s,a)-Q(s,a))+\xi(Q_{\max}(s',a')-Q(s',a'))|}{n}} \tag{16}$$

式中：$Q_{\max}(s,a)-Q(s,a)$、$Q_{\max}(s',a')-Q(s',a')$ 分别为当前状态和下一状态动作集中最大回报值与随机动作回报值之差，用于表征当前状态和下一状态探索指向性的强弱；$\xi\in(0,1)$，用于加权当前状态所选动作与下一状态所选动作指向性的评估；n 为随着迭代次数增加而减小的参数，如式(17)所示。

$$n=\frac{T-t}{T}\times n_0 \tag{17}$$

其中，T 表示最大迭代次数，t 表示当前迭代次数。从式(16)可以看出，当迭代环节开始时，智能体还不了解环境信息，由于栅格地图对应的 Q 矩阵刚刚初始化结束，$Q_{\max}(s,a)-Q(s,a)$ 与 $Q_{\max}(s',a')-Q(s',a')$ 的值都处于较低的水平，所以探索因子 ε 接近于 1，此时智能体在随机选择探索方向。理论上随着探索次数的增加，动作集中的动作会被几乎全部选中，这说明智能体对环境的探索具有全面性。随着迭代次数的增加，此时智能体已经尽可能收集环境信息，不再需要探索无用区域。n 逐渐减小，而 $Q_{\max}(s,a)-Q(s,a)$ 与 $Q_{\max}(s',a')-Q(s',a')$ 的值逐渐增大，ε 就会不断减小，智能体开始高频率的选择回报值较高的动作来获得更高的 R 值，直至得到最优路径。以此解决错过最优的全局路径和陷入局部最优的问题。

2.4　算法流程

1)载入栅格地图；

2)根据栅格地图中障碍物位置生成初始的 Q 矩阵，移动机器人置于起始点，按照四邻域搜索方式进行搜索；

3)对于机器人在探索过程中的任一状态 s，采用上文 2.3 节中所设计的改进迭代过程进行动作选择，执行动作 a 后，机器人进入状态 s'，并根据 2.2 节中设计的迭代公式更新 Q 矩阵；

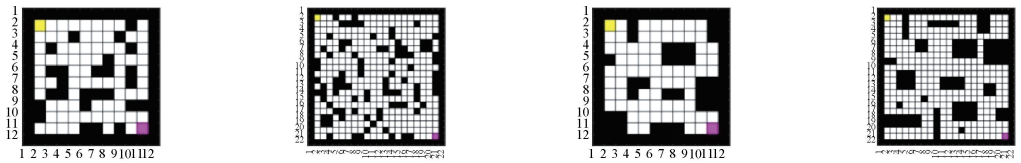
4)如果机器人已经到达目标点位置或目前走过的路径长度达到单次探索最大路径长度，则执行下一步骤，否则继续循环上一步直至满足上述状态；

5)如果 Q 矩阵已达到收敛状态或者已达到最大迭代次数，则应输出当前的最优路径，否则应该将机器人置于起始点重新进行学习过程。

3　实验结果及分析

为证明本文算法在多种环境下的可行性，本实验设计 4 种栅格地图环境，分别是小型随机障碍物地图(10×10)、中型随机障碍物地图(20×20)、小型规则障碍物地图(10×10)、中型规则障碍物地图(20×20)，如图 4 所示。

其中，黄色栅格为起始点，紫色栅格为目标点，黑色栅格为障碍物，白色栅格为可通行区域。



(a) 小型随机障碍物地图　(b) 中型随机障碍物地图　(c) 小型规则障碍物地图　(d) 中型规则障碍物地图

图 4　实验栅格地图

本文所设定的 Q-Learning 算法参数值如表 1 所示。

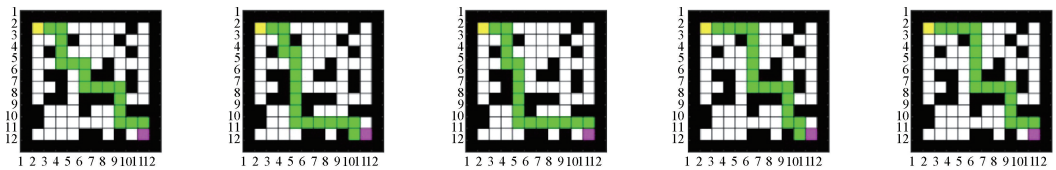
表 1　Q-Learning 算法参数值

参数	值
小型栅格地图尺寸	10×10
中型栅格地图尺寸	20×20
奖励值	100
惩罚值	−50
学习率 α	0.1
探索因子 ε	0.05
折扣率 γ	0.9
目标点吸引力因子 η	2
衰减度因子 μ	0.01
衰减上限因子 δ	90
前瞻性因子 ω	0.6
探索率指向性因子 ξ	0.2
单次探索最大路径长度	3 000
最大迭代次数	1 000

根据上述改进方法,以表 1 中的学习率和探索因子为基础进行变换,得到改进的 Q-Learning 算法。使用 Q-Learning 算法和本文提出的方法规划的路径如图 5～8 所示,路径长度随迭代过程的变化如图 9～12 所示。

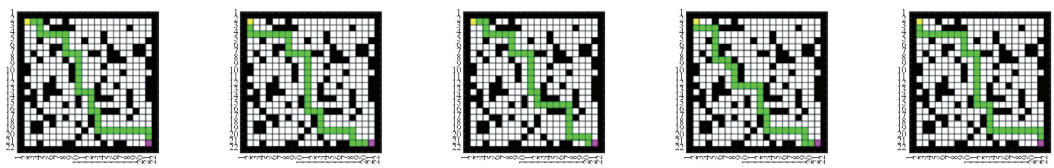
其中,Q-Learning 算法的规划结果如图 5～12 中(a)所示,单独启用改进 Q 矩阵赋值模块算法的规划结果如图 5～12 中(b)所示,单独启用改进 Q 矩阵迭代模块算法的规划结果如图 5～12 中(c)所示,单独启用改进随机探索策略模块算法的规划结果如图 5～12 中(d)所示,单独启用综合上述模块算法的规划结果如图 5～12 中(e)所示。Q-Learning 算法与本文算法规划结果的详细数据如表 2 所示。

从图 5～8 可以看出使用 Q-Learning 算法和本文的改进算法可以在这 4 种环境中成功地规划机器人的路径,并且小型地图中路径长度为 18,中型地图中路径长度为 38,这说明上述所有算法均可以在本文设计的 4 种环境中规划出最优路径。



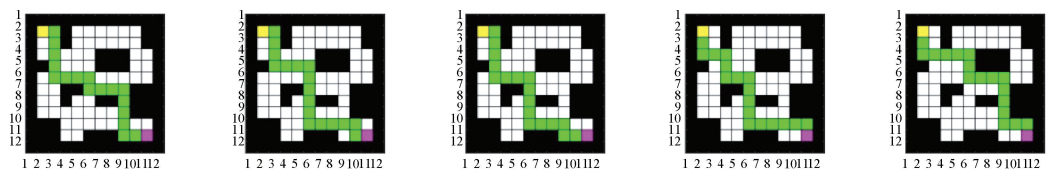
(a) Q-Learning　(b) 改进Q矩阵赋值模块　(c) 改进Q矩阵迭代模块　(d) 改进随机探索策略模块　(e) 本文完整算法

图 5　小型随机障碍物地图中 Q-Learning 算法与本文算法规划结果对比



(a) Q-Learning　(b) 改进Q矩阵赋值模块　(c) 改进Q矩阵迭代模块　(d) 改进随机探索策略模块　(e) 本文完整算法

图 6　中型随机障碍物地图中 Q-Learning 算法与本文算法规划结果对比



(a) Q-Learning　(b) 改进Q矩阵赋值模块　(c) 改进Q矩阵迭代模块　(d) 改进随机探索策略模块　(e) 本文完整算法

图 7　小型规则障碍物地图中 Q-Learning 算法与本文算法规划结果对比

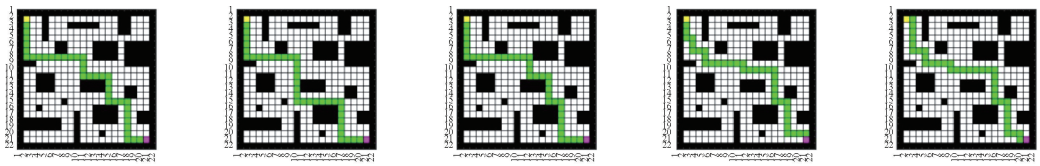


图 8　中型规则障碍物地图中 Q-Learning 算法与本文算法规划结果对比

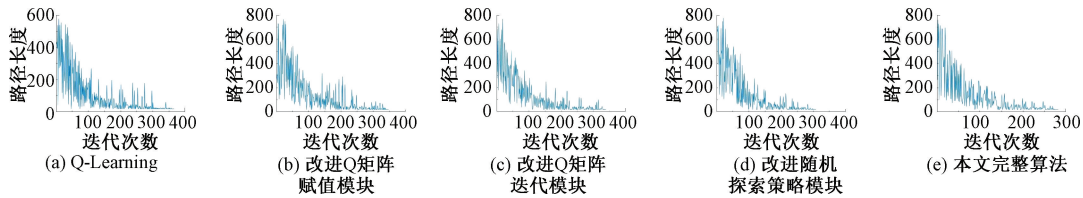


图 9　小型随机障碍物地图中 Q-Learning 算法与本文算法路径长度随迭代过程变化对比

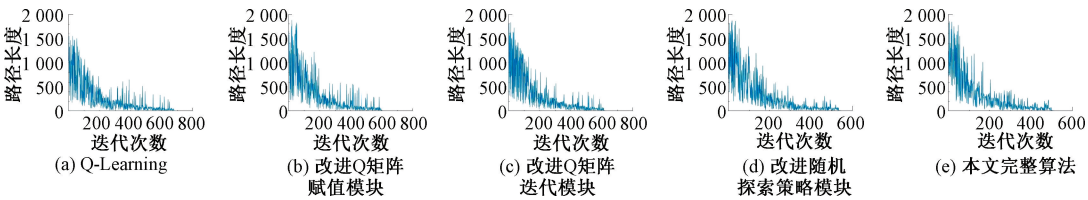


图 10　中型随机障碍物地图中 Q-Learning 算法与本文算法路径长度随迭代过程变化对比

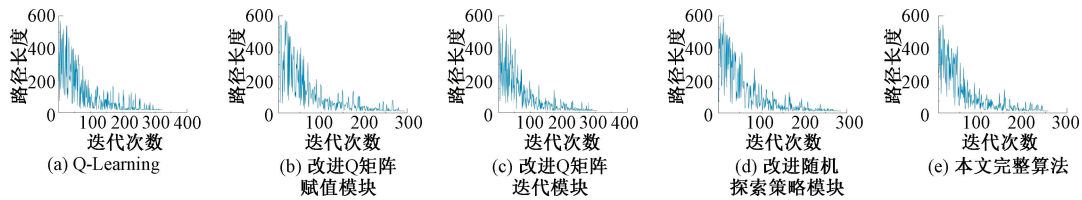


图 11　小型规则障碍物地图中 Q-Learning 算法与本文算法路径长度随迭代过程变化对比

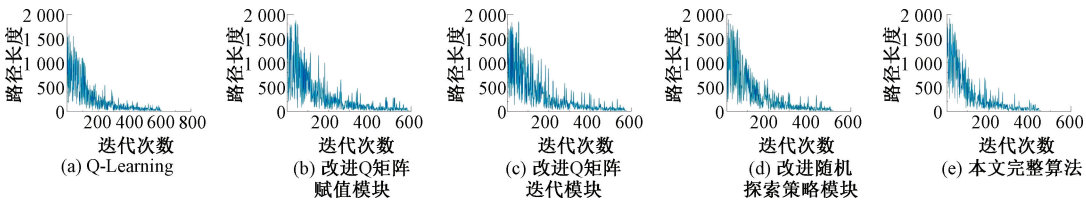


图 12　中型规则障碍物地图中 Q-Learning 算法与本文算法路径长度随迭代过程变化对比

从图 9～12 可以看出单独使用本文算法的不同模块的优化作用，并且整体算法有较好的优化效果。单独使用改进 Q 矩阵赋值策略模块（如图 9(b)～12(b)所示）使迭代过程前期路径长度减少，同时使规划出路径所需的迭代次数有一定程度的降低。出现这种现象的原因是为 Q 矩阵赋值时，越靠近起始点和目标点连线的位置对应状态的值越高，附近障碍物越多、距离障碍物越近的位置对应状态的值受到的衰减作用越大。通过此方法进行赋值后，缩减迭代前中期搜索的区域，从而减少路径长度。单独使用改进 Q 矩阵迭代模块（如图 9(c)～12(c)所示）使迭代次数有所减少。出现这种现象的原因是在移动机器人选择当前状态的动作后，更新 Q 矩阵时考虑到达下一状态所选动作的收益，减少移动机器人在某些障碍物分布下陷入局部最优的情况，从而整体上减少规划路径所需的迭代次数。单独使用改进随机探索策略模块（如图 9(d)～12(d)所示）使探索前期的步长略有升高，但是规划路径所需的迭代次数有所降低，并且在接近收敛时抖动更少。出现这种现象的原因是探索因子的初值较高，在迭代过程初始阶段，路径长度长且变化大，有少数情况路径长度较短。经过这一阶段的探索，整体的环境信息基本探索完毕，此时探索因子

表 2 Q-Learning 算法与本文算法规划结果对比

地图类型	评价参数	Q-Learning 算法	本文算法			
			单个优化模块			完整算法
			Q 矩阵初值	Q 矩阵更新	探索因子	
小型随机障碍物地图	路径长度	18	18	18	18	18
	迭代次数	366	348	339	308	282
中型随机障碍物地图	路径长度	38	38	38	38	38
	迭代次数	687	605	621	542	507
小型规则障碍物地图	路径长度	18	18	18	18	18
	迭代次数	324	294	306	285	256
中型规则障碍物地图	路径长度	38	38	38	38	38
	迭代次数	612	589	577	515	454

逐渐下降,移动机器人在动作集中更可能选择收益最大的动作,所以路径长度明显下降,抖动也相对较少。相比于 Q-Learning 算法的 ε-greedy 策略中 ε 保持不变的策略,本文算法随着迭代次数的增加,动态减小 ε 的值,使机器人倾向于选择收益更大的动作,减少了迭代过程后期的抖动现象,加速了收敛的过程。本文整体算法(如图 9(e)~12(e)所示)的规划效果综合上述模块的优势,相较于 Q-Learning 算法,规划效果有显著提升。

4 结 论

本文提出一种 Q 矩阵初始化与迭代过程具有指向性的改进 Q-Learning 算法,在原有栅格地图中根据可通行区域栅格与起始点、目标点和障碍物的位置关系进行初始化,使机器人进行初期探索过程中获得一定的指向作用,进而更快寻找到目标位置;并根据机器人的状态设置动态的探索因子,在规划出全局路径的基础上,有效改善 Q-Learning 算法容易陷入局部最优的情况。本文改进的算法在保证路径质量的前提下,具有更快的收敛速度,且收敛过程中产生的抖动更小。

参考文献

[1] MATOUI F, BOUSSAID B, METOUI B, et al. Contribution to the path planning of a multi-robot system: centralized architecture[J]. Intelligent Service Robotics, 2020, 13: 147-158.

[2] LI C, HUANG X, DING J, et al. Global path planning based on a bidirectional alternating search A* algorithm for mobile robots [J]. Computers & Industrial Engineering, 2022, 168: 108123.

[3] 李培英. 基于改进遗传算法的移动机器人路径规划[J]. 国外电子测量技术,2022,41(6):38-44.

[4] FERNANDES P B, OLIVEIRA R C L, NETO J V F. Trajectory planning of autonomous mobile robots applying a particle swarm optimization algorithm with peaks of diversity[J]. Applied Soft Computing, 2022, 116: 108108.

[5] SOUZA R M J A, LIMA G V, MORAIS A S, et al. Modified artificial potential field for the path planning of aircraft swarms in three-dimensional environments [J]. Sensors, 2022, 22(4): 1558.

[6] DUHE J F, VICTOR S, MELCHIOR P. Contributions on artificial potential field method for effective obstacle avoidance[J]. Fractional Calculus and Applied Analysis, 2021, 24(2): 421-446.

[7] FOX D, BURGARD W, THRUN S. The dynamic window approach to collision avoidance [J]. IEEE Robotics & Automation Magazine, 1997, 4 (1): 23-33.

[8] KARAMAN S, FRAZZOLI E. Sampling-based algorithms for optimal motion planning [J]. The International Journal of Robotics Research, 2011, 30(7): 846-894.

[9] ARSLAN O, TSIOTRAS P. Dynamic programming guided exploration for sampling-based motion planning

algorithms[C]. IEEE International Conference on Robotics and Automation(ICRA), 2015.

[10] DUNG L T, KOMEDA T, TAKAGI M. Reinforcement learning for POMDP using state classification[J]. Applied Artificial Intelligence, 2008, 22(7/8): 761-77.

[11] 田晓航,霍鑫,周典乐,等.基于蚁群信息素辅助的 Q 学习路径规划算法[J].控制与决策,2023,38(12):3345-3353.

[12] 薛颂东,余欢.改进蚁群与动态 Q 学习融合的机器人路径规划[J].计算机系统应用,2023,32(8):189-197.

[13] LI S, XU X, ZUO L. Dynamic path planning of a mobile robot with improved Q-learning algorithm[C]. IEEE International Conference on Information and Automation, 2015.

[14] 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.

[15] SHI Z, WANG K, ZHANG J. Improved reinforcement learning path planning algorithm integrating prior knowledge[J]. PloS One, 2023, 18(5): e0284942-e0284942.

[16] 段建民,陈强龙.利用先验知识的 Q-Learning 路径规划算法研究[J].电光与控制,2019,26(9):29-33.

[17] 王慧,秦广义,夏鹏,等.基于改进强化学习算法的移动机器人路径规划研究[J].计算机应用与软件,2022,39(7):269-274.

[18] 李威,张晓东,姜学峰,等.基于改进强化学习的机器人路径规划研究[J].制造业自动化,2023,45(3):148-151,172.

[19] MENG H, ZHANG H. Mobile robot path planning method based on deep reinforcement learning algorithm[J]. Journal of Circuits, Systems and Computers, 2022, 31(15): 85-103.

[20] 许宏鑫,吴志周,梁韵逸.基于强化学习的自动驾驶汽车路径规划方法研究综述[J].计算机应用研究,2023, 40(11): 3211-3217.

作者简介

王立勇,博士,教授,主要研究方向为无人车辆设计与自动驾驶技术、装备服役性能退化评估与调控方法。

E-mail: wangliyong@bistu.edu.cn

王弘轩,硕士研究生,主要研究方向为移动机器人路径规划算法。

E-mail: 1254765156@qq.com

苏清华(通信作者),博士,副研究员,主要研究方向为无人驾驶环境感知。

E-mail: suqinghua@bistu.edu.cn