DOI:10.19651/j.cnki.emt.2106005

基于强化学习的移动机器人路径规划优化*

尹 旷 王红斌 方 健 莫文雄 叶建斌 张 宇
(广东电网有限责任公司广州供电局电力试验研究院 广州 510410)

摘 要：随着信息化程度不断加深,移动机器人的应用越来越广泛,但在很多情况下,移动机器人需要工作在不断变化且复杂的环境中,由于无法提前获取环境信息,往往难以对移动机器人进行路径规划并寻找到一条合适的路径。针对这一问题,提出了一种移动机器人路径规划方法。该方法运用栅格法建立环境模型,利用探索步数定义回报值,并通过强化学习不断优化路径。针对强化学习中对环境的探索与利用的平衡问题,提出一种变化的 ε-decreasing 动作选择策略和学习率选择方法,使探索因子随着智能体对环境探索程度的增加而动态变化,从而加快了学习算法的收敛速度。仿真结果表明,该方法能够实现移动机器人在复杂的环境下的自主导航和快速路径规划,在获得相同路径长度的前提下,迭代次数相比于传统强化学习算法减少了约 32%,有效地加快了收敛速度。

关键词：路径规划;环境信息;强化学习;探索因子
中图分类号：TP29;TP242.6 **文献标识码**：A **国家标准学科分类代码**：520.60

Optimization of robot path planning based on reinforcement learning

Yin Kuang Wang Hongbin Fang Jian Mo Wenxiong Ye Jianbin Zhang Yu
(Electric Power Test Research Institute of Guangzhou Power Supply Bureau of Guangdong Power Grid Co., Ltd., Guangzhou 510410, China)

Abstract: As the degree of informatization continues to deepen, the application of robots is becoming more and more extensive. However, in many cases, robots need to work in a constantly changing and complex environment. Because of the inability to obtain environmental information in advance, it is often difficult to plan a suitable path for a robot. To solve this problem, this paper proposes an method for robot path planning. This method uses the grid method to establish an environmental model, uses the number of exploration steps to define the return value, and continuously optimizes the path through reinforcement learning. At the same time, aiming at the problem of the balance between the exploration and utilization of the environment in reinforcement learning, a variable ε-decreasing action selection strategy and learning rate selection method are proposed to make the exploration factor dynamically change as the agent explores the environment, thereby accelerating the convergence speed of the learning algorithm. Simulation results show that this method can realize autonomous navigation and fast path planning of mobile robots in complex environments, compared with traditional algorithms, under the premise of obtaining the same path length, the number of iterations is reduced by approximately 32%, effectively speeding up the convergence speed.

Keywords: path planning;environment information;reinforcement learning;exploration factor

0 引 言

移动机器人路径规划[1]是指在处于有障碍物的环境中,移动机器人按照给定的任务或条件,寻找一条从起点到终点、平顺无碰撞的最优路径[2]。路径规划的主要目标是机器人在障碍环境中根据绩效指标如距离、时间和能量,找到从起点到终点的最佳路线。根据对环境的了解情况,路径规划分为两个研究方向:全局路径规划是智能体基于环境先验完全信息的路径规划;局部路径规划是环境模型对于智能体而言是部分未知,或者全部未知情况下的路径规划。在实际应用中,机器人需要具有适应未知环境的能力,因此解决机器人在未知环境中的路径规划,对机器人技术的应用和普及具有重大意义,是移动机器人各项应用研究的前提和基础。

收稿日期:2021-03-16
* 基金项目:中国南方电网有限责任公司科技项目(GZHKJXM20180069)资助

目前，针对一些环境信息全部未知或部分未知的场景下机器人的路径规划，国内外研究人员进行了很多研究并取得了相应的成果。文献[3]提出一种改进人工势场法的路径规划方法，采用设置中间目标点的方式，给机器人一个外力以避免其在局部最小点处停止或者徘徊，克服了传统人工势场法在寻找最优路径时存在的收敛速度过慢、容易陷入局部最优等缺点。文献[4]提出一种基于改进蚁群算法的机器人路径规划方法，其运用增加阈值和路径权重的方法改进蚁群算法[5]中信息素更新公式，增加优秀蚂蚁对路径规划带来的影响，减少迭代次数。文献[6]提出一种基于模糊逻辑的智能车局部路径规划方法，其针对模糊逻辑在智能车局部路径规划中存在的问题，引入多次转动避障的策略，有效避免死锁发生，进而实现智能车的路径规划。文献[7]提出一种基于混合遗传算法的移动机器人路径规划方法，其将遗传算法[8]和模拟退火算法[9]相结合，采用栅格法对环境建立模型，同时在遗传算子中添加插入算子和删除算子以优化路径，相对于基本遗传算法，其收敛速度和搜索质量等有了明显的提高。这些方法在一定程度上克服了环境先验信息不足的问题，加快求解路径规划问题时的收敛速度，规划出一条较优的路径。但由于算法本身与环境交互性差，无法充分从环境获取信息，在面对更复杂环境时，需要耗费大量时间在无谓的搜索上，从而导致出现收敛速度变慢，容易出现局部最优的情况。

随着机器学习[10-13]的快速发展，利用基于强化学习的方法来解决路径规划问题越来越受到人们的重视，强化学习强调在与环境交互的过程中“试错”和“改进”，特点是可以在没有系统模型的条件下实现无导师的在线学习，十分符合移动机器人路径规划的需求。本文先用栅格法[14]建立模型，利用模型自动生成状态转移矩阵，并通过智能体与环境的交互定义奖赏函数更新回报值，使得智能体在对环境进行探索时不会盲目进行移动，有利于寻找最优路径，提高路径解的精确度。通过改进动作选择策略，根据环境的探索情况指导智能体的动作选择，从而进行移动机器人路径规划，进一步减少智能体在路径规划后期对环境进行的无效探索，改善算法收敛速度。为验证该方法的可行性和灵活性，选用 MATLAB 软件对其进行仿真计算。

1　强化学习算法原理

Q-learning 是一种基于值迭代来学习动作策略[15]的强化学习算法。该算法利用 Q 函数[16]寻找最优的动作-选择策略，其核心是不断更新一个由状态，动作和奖赏三者组成的表格(又称为 Q 矩阵)。Q 矩阵中的 Q 值作为在某一个状态下采取动作的质量的度量，通过使用 Q 值度量方式来寻找每一个状态下所应该采取的最佳动作。随着程序的不断迭代，Q 矩阵最终会趋于收敛。

算法中采用 $Q(s,a)$，即状态-动作对的值作为估计函数，它的学习过程是不断地通过公式迭代来使相邻 Q 值趋于一致。其基本形式如下：

$$Q^*(s,a)=R(s,a)+\gamma\sum T(s,a,s_t)\max Q^*(s_t,a_t) \tag{1}$$

式中：$Q^*(s,a)$ 为状态 s 下采用动作 a 所得到的最优奖赏值的和；$R(s,a)$ 表示状态 s 下采用动作 a 所得到的即时奖赏值；$\sum T(s,a,s_t)\max Q^*(s_t,a_t)$ 对应表示非即时奖赏值，即当前时刻之后采用不同动作 a_t 所获得的累计奖赏值；再定义 $V^*(s)$ 为在状态 s 下的最优值函数，则有：

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

引入贝尔曼方程，通过贝尔曼方程对算法决策过程中最佳决策序列进行求解，贝尔曼方程如下：

$$V^{\pi}(s)=E_{i\sim\pi(s)}[R_t\mid s_t=s] \tag{3}$$

式中：$V^{\pi}(s)$ 为状态 s 以及后继状态下继续以策略 π 去选择执行动作所获得的奖赏；$E_{i\sim\pi(s)}$ 为策略 π，其中 $R_t=\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}$，$r_{t+1}$ 为状态 s_t 转移到 s_{t+1} 时获得的回报，γ 为折扣，取值在 0～1 之间。

在贝尔曼方程的基础上，状态 s 由 ε-greedy 策略选择动作 a，转移到状态 s_t，Q-learning 的迭代公式和 ε-greedy 策略基本表达式如下：

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

$$prob(a_i)=\begin{cases}1-\varepsilon, & a=\arg\max Q(s,a_i)\\ \varepsilon, & \text{其他}\end{cases} \tag{5}$$

其中，α 为学习率，决定每一次动作对 Q 值的影响程度大小，r 为执行动作后得到的奖赏值。

Q-learning 算法首先对 Q 矩阵中的 Q 值进行初始化，然后智能体在不同状态下根据 ε-greedy 策略[17]选择动作，ε 的大小决定了智能体选择的动作，程序执行动作到达新的状态后通过 Q-learning 迭代公式得到实际迭代值并对 Q 矩阵进行更新，当到达目标状态时此次迭代过程结束，智能体自动选择下一个样本再继续从初始状态进行迭代，直至整个学习过程结束。

2　基于 ε-decreasing 的强化学习路径规划

2.1　改进动作-选择策略

在强化学习中，智能体通过其选择的动作序列影响训练样本的分布。一方面，智能体需要尽可能多地尝试不同的动作以探索环境信息。另一方面，也需要考虑选择值函数最大的动作以带来较大的回报。探索和利用环境之间的平衡问题实质上是智能体每个决策中如何选择动作的问题。广泛的探索延长了智能体的训练时间，而过度地利用导致程序过早收敛，从而错过最优路径。因此需要对动作-选择策略进行改进，使得 Q-learning 算法在探索和利用环境之间达到平衡，既防止过多无意义的探索降低学习算法的性能，又避免一味的选择最有利的动作而陷入局部最优。

具体实现方法如下。

在算法运行初始,由于智能体面对的是一个陌生的环境,此时需要尽可能的通过随机动作来探索环境,随着对环境信息的掌握程度不断增加,智能体逐渐减少对环境的探索,转而选择最有利的行动以带来更大的回报。因此希望 ε 在整个算法的运行过程中随着对环境了解保持一种下降的趋势,称为 ε-decreasing。

$$\varepsilon = \mathrm{e}^{\left|\frac{Q(S_t,a_b)\ Q(S_t,a_t)}{n}\right|} \tag{6}$$

其中,$Q(s_t,a_b)$ 代表当前状态 S 下的选择对应最大动作的值,$Q(s_t,a_t)$ 代表当前状态 S 下一个随机动作的值,n 则代表一个随迭代次数增加而不断减少的值,其计算方式如下:

$$n = \frac{M-m}{M} \tag{7}$$

其中,M 代表算法最大迭代次数,m 代表当前的迭代次数。由式(6)可以看到,当智能体对环境一无所知时,$Q(s_t,a_b)$ 和 $Q(s_t,a_t)$ 的数值相等,此时 ε 因子等于 1,代表智能体此时所选择的所有动作都在探索环境,而随着迭代次数的增加和智能体对环境信息的熟悉,n 的数值变小而 $Q(s_t,a_b)$ 和 $Q(s_t,a_t)$ 的差值增大,ε 因子不断变小,代表智能体逐渐开始选择最有利的行动获取最优路径。

2.2　动态学习率选择

式(4)中的 α 为学习率,其决定了每一次行动之后智能体从环境学习得到的信号的多少,α 越大每一次动作对 Q 值的影响越大。但在算法运行过程当中,并不是每一次的动作的学习率都要保持一致,当对环境进行探索时,由于需要对环境增加了解,故设置较大的学习率,而当利用环境时则设置较小的学习率,防止陷入局部最优,因此学习率 α 的选择设置如下:

$$\alpha = \begin{cases} 0.7, & \varepsilon_0 < \varepsilon \\ 0.1, & \text{其他} \end{cases} \tag{8}$$

2.3　算法主要步骤

1)对当前环境进行栅格初始化,确定起始点和目标点坐标;

2)根据栅格化环境建立 Q 矩阵,并初始化 Q 矩阵,机器人从起始点出发对环境进行探索;

3)根据机器人所处状态 s,利用上文所述改进动作选择策略选择机器人在状态 s 下所需要执行的动作 a;

4)执行动作 a 后移动到状态 s' 位置,并根据 Q-learning 的迭代公式更新 Q 矩阵的对应值;

5)判断当前位置是否为目标点坐标或是达到移动机器人最大尝试步长数,如果不是,则返回步骤 3),否则进入到步骤 6);

6)判断 Q 矩阵是否收敛或者达到最大迭代次数,如果是,则结束整个学习过程,输出最优路径,否则返回步骤 2)进行下一次尝试学习。

3　仿真试验与分析

为了验证本文提出的基于强化学习的路径规划优化方法在探索未知环境上的可行性,使用 MATLAB 对其进行仿真试验。表 1 所示为基础 Q-learning 算法所需要用到的参数。

表 1　基础 Q-learning 算法参数

参数名称	数值
地图尺寸	20×20
奖励值	100
惩罚值	−100
学习率	0.1
探索因子	0.2
折扣因子	0.9
最大步长	2 000
最大迭代次数	1 000
动作选择	4

将表 1 的探索因子和学习率根据本文提出的方法进行变换即为改进 Q-learning 算法,其中探索因子初始值为 1,根据智能体对环境的熟悉程度而动态下降,具体计算公式如式(6)和(7)所示,学习率的初始值为 0.7,随着探索因子的动态减小,当小于 0.2 时学习率动态设置为 0.1。仿真环境设置如图 1 所示,利用基础 Q-learning 算法和本文提出方法规划的路径分别如图 2 和 3 所示。

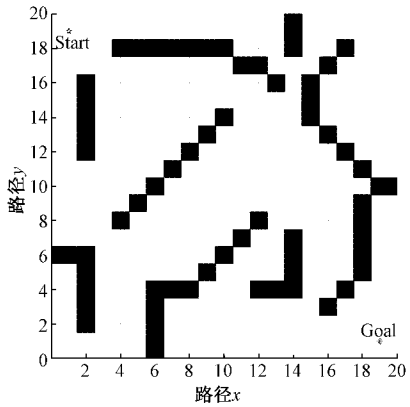

图 1　仿真环境

从图 2 和 3 中可以看到,两种方法都能够成功地在未知环境下进行路径规划,虽然路径规划的路线有一定的出入,但路径长度均为 36,说明两个方法均能够在该环境下找到最优路径。

为了进一步对比两种方法的性能,图 4 和 5 给出两种方法的 Q-learning 步长收敛图,可以看到在开始迭代时,两种强化学习方法的路径长度抖动都较大,随着迭代次数的增加,路径长度均呈现下降的趋势,但基础 Q-learning 算法

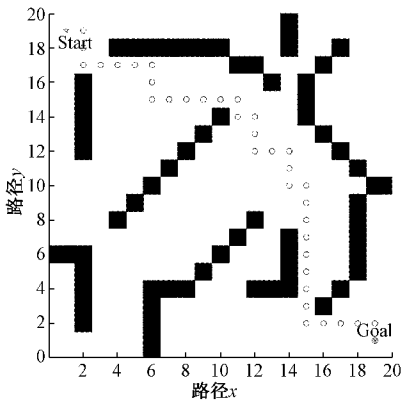

图 2　基础 Q-learning 算法路径规划图

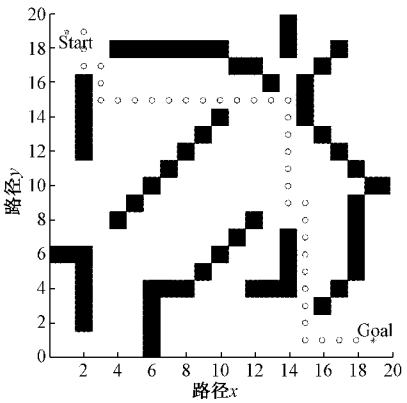

图 3　改进 Q-learning 算法路径规划图

在收敛后期还存在较频繁且剧烈的抖动，而改进 Q-learning 算法则平缓进行了收敛。造成这种现象的原因是在初期，智能体需要对环境进行尽可能多的探索从而充分的获取环境信息，因此两种 Q-learning 算法在迭代伊始同时出现剧烈的路径长度变化。随着迭代次数增加，两种算法对应的智能体对环境的了解逐渐加深，“学习经验”开始逐渐指导路径规划，因此规划得到的路径长度呈现下降的趋势。而到了后期，虽然基础 Q-learning 算法已经进行了足够多次的迭代，但由于采取的 ε-greedy 策略中 ε 的值不变，仍然保持对环境进行随机探索，所以后期还是会出现较剧烈的抖动，最终经过 595 次迭代后完成收敛。改进 Q-learning 算法的 ε-decreasing 策略随着迭代次数的增大，动态缩小 ε 的数值，使智能体探索环境的动作不断转换成利用环境的动作，因而路径长度在迭代后期的波动逐渐减小并最终趋于平缓，且仅经过 407 次迭代后就达到收敛状态。

两种算法在该环境下进行 20 次试验，记录下路径长度和迭代次数，并计算平均路径长度和平均迭代次数，如表 2 所示。

从表 2 中可以看出，两种方法计算所得的平均路径长度均为 36，但本文提出的改进 Q-learning 方法所需的平均迭代次数仅为 402 次，相比于基础 Q-learning 方法迭代次数减少率为 32.0%，有效地减少了迭代次数，从而提高收敛速度。

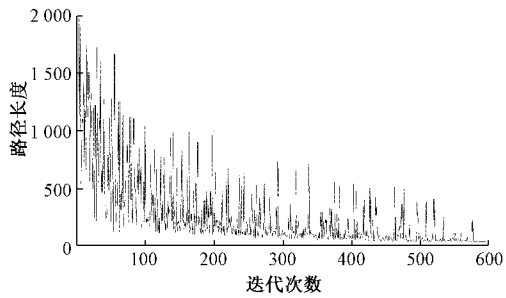

图 4　基础 Q-learning 算法路径规划收敛图

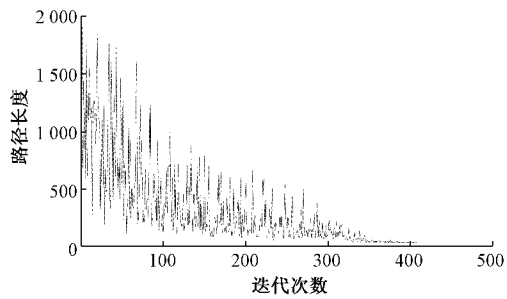

图 5　改进 Q-learning 算法路径规划收敛图

表 2　算法平均路径长度和迭代次数对比

栅格环境	平均路径长度	平均迭代次数	与基础 Q-learning 算法对比迭代次数减少率/%
基础 Q-learning 方法	36	591	0
本文方法	36	402	32.0

4　结　　论

本文针对移动机器人在位置环境中的路径规划问题，在传统 Q-learning 算法的基础上提出了 ε-decreasing 动作选择策略，该策略以对环境信息的掌握程度为导向，自适应调整智能体对环境的探索和利用两种状态，有助于减少过多无谓的探索从而加快收敛速度；同时根据智能体的状态动态设置学习率，最大程度地避免智能体陷入局部最优情况的发生。与传统 Q-learning 算法相比，本文所提方法在保证得到最优路径的前提下，收敛速度更快，表现出了更好的路径规划性能。但本文所提方法在面对特大环境时，往往存在着生成的 Q 矩阵过大的问题，因此在后续的改善中可以进一步考虑如何有效地压缩 Q 矩阵从而提高算法收敛速度。

参考文献

[1]　FETHI M, BOUNMEDYEN B, BRAHIM M, et al. Contribution to the path planning of a multi-robot system: Centralized architecture[J]. Intelligent Service Robotics, 2020, 13(1): 147-158.

[2]　WAN Y, WANG M, YE Z, et al. A feature selection

method based on modified binary coded ant colony optimization algorithm[J]. Applied Soft Computing, 2016, 49:248-258.

[3] 任彦, 赵海波. 改进人工势场法的机器人避障及路径规划[J]. 计算机仿真, 2020, 37(2):365-369.

[4] 王猛,邢关生. 基于改进蚁群算法的机器人路径规划[J]. 电子测量技术,2020,43(24):52-56.

[5] 李志锟,黄宜庆,徐玉琼. 改进变步长蚁群算法的移动机器人路径规划[J]. 电子测量与仪器学报,2020,34(8):15-21.

[6] 朱曼曼, 杜煜, 张永华,等. 基于模糊逻辑的智能车局部路径规划[J]. 北京联合大学学报, 2016, 4(4):35-38.

[7] 裴以建, 杨亮亮, 杨超杰. 基于一种混合遗传算法的移动机器人路径规划[J]. 现代电子技术, 2019, 42(2):183-186.

[8] HAO K, ZHAO J, YU K, et al. Path planning of mobile robots based on a multi-population migration genetic algorithm[J]. Sensors, 2020, 20(20):5873.

[9] 王存华,王伟. 基于模拟退火优化 BP 算法的指纹地图构建方法[J]. 国外电子测量技术,2020,39(3):17-24.

[10] WAI R J, PRASETIA A S. Adaptive neural network control and optimal path planning of UAV surveillance system with energy consumption prediction[J]. IEEE Access, 2019(99):1-1.

[11] ZHANG Y, WANG Y, LANG H, et al. Visual avoidance of collision with randomly moving obstacles through approximate reinforcement learning [J]. Instrumentation, 2019, 6(3):59-66.

[12] CHEN H, JI Y, NIU L. Reinforcement learning path planning algorithm based on obstacle area expansion strategy [J]. Intelligent Service Robotics, 2020, 13(6):1-9.

[13] 李卫硕,孙剑,陈伟. 基于 BP 神经网络机器人实时避障算法[J]. 仪器仪表学报,2019,40(11):204-211.

[14] JUNG J H, KIM D H. Local path planning of a mobile robot using a novel grid-based potential method [J]. International Journal of Fuzzy Logic and Intelligent Systems, 2020.

[15] CHEN C, CHEN X, MA F, et al. A knowledge-free path planning approach for smart ships based on reinforcement learning[J]. Ocean Engineering, 2019, 189:106299.

[16] 董培方, 张志安, 梅新虎,等. 引入势场及陷阱搜索的强化学习路径规划算法[J]. 计算机工程与应用, 2018,54(16):129-134.

[17] 刘辉,肖克,王京擘. 基于多智能体强化学习的多 AGV 路径规划方法[J]. 自动化与仪表,2020,35(2):84-89.

作者简介

尹旷(通信作者),工程师,主要研究方向为智能电网、高电压与绝缘技术。

E-mail:yinkuang1232021@163.com

王红斌,教授级高级工程师,主要研究方向为高电压与绝缘技术。

E-mail:564053662@qq.com

方健,高级工程师,主要研究方向为高电压与绝缘技术。

E-mail:723750991@qq.com