未知环境中基于强化学习的移动机器人路径规划*

梁 泉
（南京农业大学 工学院，江苏 南京 210031）

摘要： 为解决未知环境中移动机器人的自适应路径规划问题，提出了一种基于Q学习算法的自主学习方法。首先设计了未知环境中基于传感器信息的移动机器人自主路径规划的学习框架，并建立了学习算法中各要素的数学模型；然后利用模糊逻辑方法解决了连续状态空间的泛化问题，有效地降低了Q值表的维数，加快了算法的学习速度；最后在不同障碍环境中对基于Q学习算法的自主学习方法进行了仿真实验，仿真实验中移动机器人通过自主学习较好地完成了自适应路径规划。研究结果证明了该自主学习方法的有效性。

关键词： 未知环境；Q学习算法；移动机器人；路径规划
中图分类号： TP242 **文献标志码：** A **文章编号：** 1001-4551(2012)04-0477-05

Reinforcement learning based mobile robot path planning in unknown environment

LIANG Quan
(College of Engineering, Nanjing Agricultural University, Nanjing 210031, China)

Abstract: In order to solve problem of the adaptive path planning on mobile robot in unknown environments, a self-learning method based on Q-learning algorithm was proposed. Firstly, the learning framework was designed for the adaptive path planning of mobile robot in unknown environments based on sensor information, and the mathematics model for each element of learning algorithm was proposed. Then the generalization problem of continuous state space of reinforcement learning system was solved by fuzzy logic, the size of the Q-table was reduced and the speed of the learning algorithm was increased. Finally, the simulation of self-learning method based on Q-learnning algorithm was carried in the environment with different obstacles, the adaptive path planning was achieved by the mobile robot through self-learning. The research results certify the validity of this method.

Key words: unknown environment; Q-learning algorithm; mobile robot; path planning

0 引 言

路径规划是移动机器人导航研究的一个重要环节和课题，可分为基于地图的全局路径规划和基于传感器的局部路径规划。在移动机器人导航控制理论和方法的研究中，环境已知条件下的离线全局路径规划方法已经取得大量成果，如人工势场法、拓扑法和栅格法等。然而，未知环境中移动机器人只有较少的先验知识，很难获取环境的精确数学模型，因此如何使移动机器人自主地学习，感知环境并做出决策成为当前研究的重点[1-3]。

根据学习方法的不同，机器人学习可以分成监督学习、进化方法和强化学习3类。强化学习的优势在于通过和环境交互地试错进行在线学习。例如，Lavi Michael Zamstein[4]采用Q学习算法实现了移动机器人路径规划。然而，强化学习在处理连续状态空间时易出现维数灾问题，因此泛化成为强化学习方法的一种重要能力[5]。例如，Hee Rak Beom等人[6]采用模糊逻

收稿日期： 2011-10-17
基金项目： 国家高技术研究发展计划（“863”计划）资助项目（2006AA10A304）
作者简介： 梁 泉（1986-），男，山东青岛人，主要从事机器人运动规划方面的研究. E-mail:njau_lq@yahoo.com.cn

辑，秦政等人[7]采用BP神经网络对强化学习进行泛化处理完成了移动机器人的自主导航。

本研究将强化学习与模糊逻辑相结合，利用模糊逻辑对环境状态进行离散化处理，以保证状态划分的合理性。利用强化学习选择策略完成移动机器人避障，实现未知环境中的自适应路径规划，具有学习速度快，收敛性好，稳定性高等优点。

1 强化学习及结构设计

1.1 强化学习原理

Watkins[8]指出在强化学习系统中，智能体主动对环境做出试探，并从环境对试探动作产生的评价中获得知识，改进行动策略以达到预期目的。强化学习的基本原理是：如果智能体的某个行为策略导致环境对智能体正的奖赏，则智能体以后采取这个行为策略的趋势会加强。

常用的强化学习算法有TD算法、Sarsa算法和Q算法等[9]。本研究采用Q学习算法，它作为一种最有效的与环境模型无关的算法，具有在线学习的特点，使得移动机器人通过自主学习适应未知环境[10]。

1.2 强化学习系统结构

除了智能体和环境，一个强化学习系统还包括4个主要组成要素：环境模型、奖惩函数、值函数以及动作选择策略。

1.2.1 环境模型

移动机器人环境模型如图1所示。本研究设定环境中存在移动机器人、障碍物（动态或静态）和目标点，且每一时刻移动机器人均可通过传感器得到障碍物和目标点的相关信息。

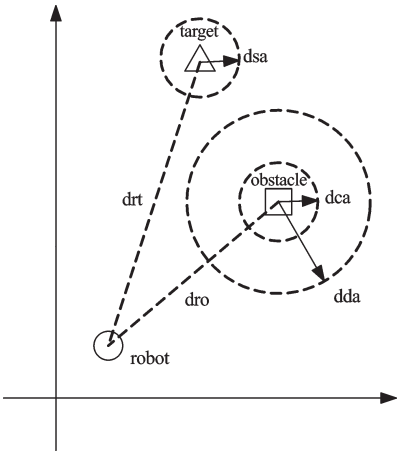


图1 移动机器人环境模型

1.2.2 奖惩函数

由于Q学习算法是基于有限状态和有限动作的有限MDP模型的，输入/输出必须是有限离散状态，而大多数实际系统本身的输入/输出是连续的，因此需要进行泛化处理。利用模糊逻辑实现系统连续状态空间离散化，是最为广泛的一种方法[11]。在图1中，本研究设定移动机器人到障碍物距离 d_{ro}，移动机器人到目标点距离 d_{rt}，设定成功距离 $d_{sa}=1$，冲突距离 $d_{ca}=2$，危险距离 $d_{da}=4$，可将环境状态 S 离散为成功状态（WS）、安全状态（SS）、危险状态（DS）、失败状态（FS），具体定义如下：

$$S=\begin{cases}\text{WS} & d_{rt}\leqslant d_{sa}\\ \text{SS} & d_{ro}>d_{da}\\ \text{DS} & d_{ca}<d_{ro}\leqslant d_{da}\\ \text{FS} & d_{ro}\leqslant d_{ca}\end{cases}\tag{1}$$

根据状态转移，移动机器人从SS移动到WS的奖惩值为100；从DS移动到SS的奖惩值为10；从SS移动到DS的奖惩值为-10；从DS移动到FS的奖惩值为-100。当移动机器人处于DS状态时，若 $n+1$ 时刻移动机器人与障碍物距离 $d_{ro}(n+1)$ 小于 n 时刻距离 $d_{ro}(n)$，则奖惩值为-10，否则奖惩值为0。综上所述，奖惩函数可定义为：

$$r=\begin{cases}100 & S\subset \text{SS}\rightarrow\text{WS}\\ 10 & S\subset \text{DS}\rightarrow\text{SS}\\ -10 & S\subset \text{SS}\rightarrow\text{DS}\\ -10 & S\subset \text{DS}\rightarrow\text{DS},d_{ro}(n+1)<d_{ro}(n)\\ 0 & S\subset \text{DS}\rightarrow\text{DS},d_{ro}(n+1)\geqslant d_{ro}(n)\\ -100 & S\subset \text{DS}\rightarrow\text{FS}\end{cases}\tag{2}$$

1.2.3 Q值函数

Waktins定义Q值函数为在状态 s_t 时执行动作 a_t 的评价函数，且此后按最优动作序列执行时的强化信号折扣和，即：

$$Q(s_t,a_t)=r_t+\gamma\max_{a\in A}Q(s_{t+1},a)\tag{3}$$

在实际应用中，Q值函数的更新规则为：

$$Q_t(s_t,a_t)=(1-\alpha)Q_{t-1}(s_t,a_t)+\alpha[r_t+\gamma\max_{a\in A}Q_{t-1}(s_{t+1},a)]\tag{4}$$

式中：a —可执行动作；A —可执行动作集合；$\alpha\in[0,1]$ —学习率，它控制着学习的速度，α 越大则收敛越快，但过大的 α 可能导致算法不收敛；r_t —t 时刻环境返回学习系统的强化值；γ —折扣因子。

1.2.4 动作选择策略

在强化学习中，移动机器人一方面需要尽可能地选择不同的动作，以找到最优的策略，即探索；另一方面又要考虑选择 Q 值函数最大的动作，以获得大的奖赏，即利用。设计合理的动作选择机制，控制探索和利用的平衡，对保证算法能快速收敛到最优 Q 值函数具有重要的意义。

本研究采用Boltzmann分布机制选取 Q 值最大的动作进行移动机器人避障。在Boltzmann分布中，在状

态 s 下,动作 a_i 被选择的概率取为:

$$Pr(a_i)=\frac{e^{Q(s,a_i)/T}}{\sum_{a_k\in A}e^{Q(s,a_k)/T}} \tag{5}$$

式中:a_k —属于动作集合 A 的某一可执行动作;T—可调节的温度参数,且 $T>1$。

可定义为:

$$T=T_0n^{-1/\beta} \tag{6}$$

随着学习次数 $n\to\infty$,T 从 T_0 趋近于0,常数 β 越小则曲线的曲率越大。

2　仿真实验及分析

为了验证该方法的有效性,本研究利用Matlab软件对未知环境中移动机器人自主路径规划进行了不同场景的仿真。Q学习算法的仿真程序流程如图2所示。

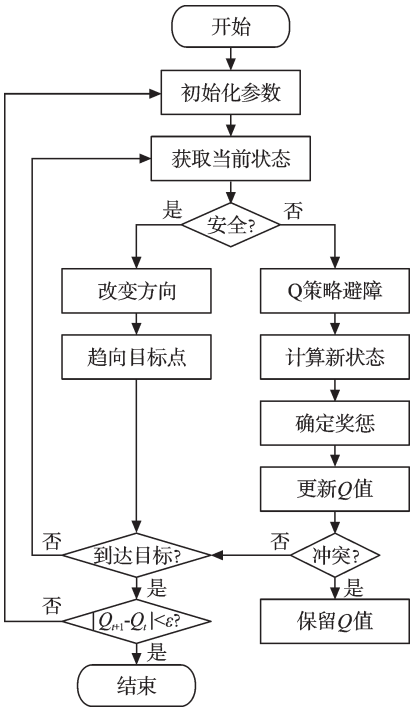


图2　Q学习算法仿真

移动机器人自适应路径规划的强化学习框架由两部分组成:趋向目标和避障。当环境中不存在障碍物或者障碍物较远时,移动机器人调整方向驶向目标点;当环境中存在障碍物且障碍物较近时,移动机器人利用Q学习策略进行避障,可选择动作包括左转和右转。

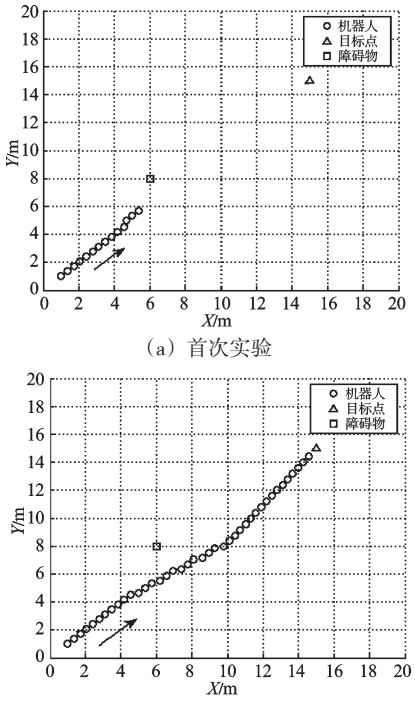
本研究对移动机器人路径规划进行仿真,首先对参数进行设置:学习率 $\alpha=0.5$,折扣因子 $\gamma=0.8$,参数 $\varepsilon=0.02$,移动机器人速度 $v_r=0.5$ m/s,动态障碍物速度 $v_o=0.5$ m/s。某一时刻 t 的移动机器人位置及动态障碍物位置可由下式求得:

$$\begin{bmatrix}p_{xt}\\p_{yt}\end{bmatrix}=\begin{bmatrix}p_{x(t-1)}\\p_{y(t-1)}\end{bmatrix}+v\begin{bmatrix}\cos\theta\\\sin\theta\end{bmatrix}\Delta T \tag{7}$$

式中:$\begin{bmatrix}p_{xt}\\p_{yt}\end{bmatrix}$—$t$ 时刻的位置,$\begin{bmatrix}p_{x(t-1)}\\p_{y(t-1)}\end{bmatrix}$—$t-1$时刻的位置,$\Delta T$ —时间间隔,θ —ΔT 时间内的航向角。

2.1　静态障碍

本研究设定移动机器人初始位置(1,1),静态障碍物位置(6,8),目标点位置(15,15)。首次实验时,移动机器人在第14个时间步与障碍物发生冲突,经过50次训练后实验,移动机器人40个时间步后顺利到达目标点,路径如图3所示。随后进行100次验证性实验,移动机器人到达目标点的成功率为100%,到达目标点平均所需时间步为41个,验证性实验如图4所示。



(a) 首次实验

(b) 训练50次后实验

图3　移动机器人路径(静态障碍)

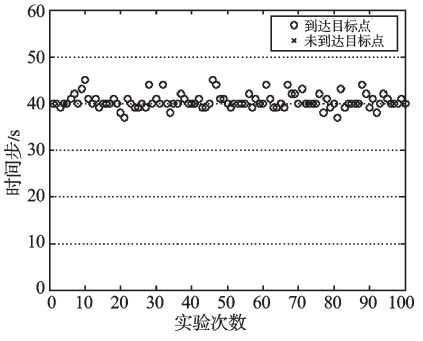
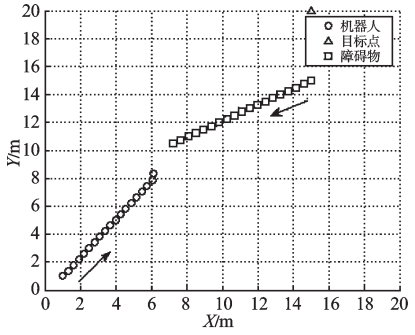


图4　验证性实验(静态障碍)

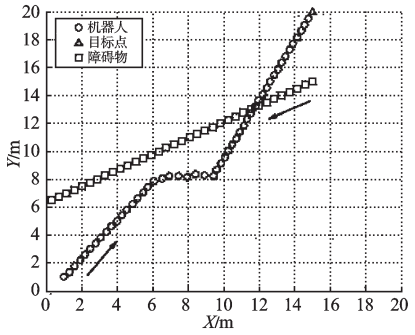
2.2　动态障碍

本研究设定移动机器人初始位置(1,1),动态障碍物初始位置(15,15),目标点位置(15,20)。首次实验时,移动机器人在第19个时间步与障碍物发生冲突,经过50次训练后实验,移动机器人50个时间步后顺利到达目标点,路径如图5所示。本研究随后进行100次

验证性实验，移动机器人到达目标点的成功率为94%，到达目标点平均所需时间步为52个，每实验5次后计算未到达目标点的失败率，随着实验次数的增加，失败率明显降低，验证性实验如图6所示。

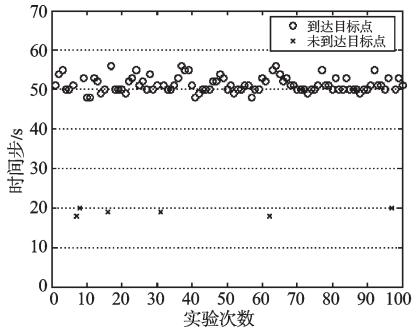


（a）首次实验

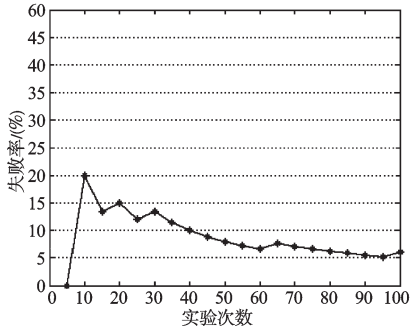


（b）训练50次后实验

图5 移动机器人路径(动态障碍)



（a）实验结果统计

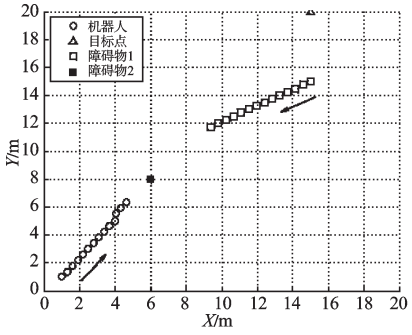


（b）失败率变化

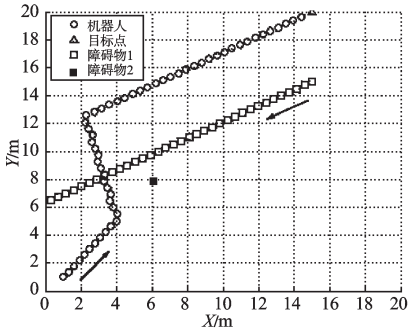
图6 验证性实验(动态障碍)

2.3 动静态障碍

本研究设定移动机器人初始位置(1,1)，静态障碍物位置(6,8)，动态障碍物初始位置(15,15)，目标点位置(15,20)。根据规划，移动机器人优先躲避距离最近的障碍物。首次实验时，移动机器人在第14个时间步与静态障碍物发生冲突，经过50次训练后实验，移动机器人55个时间步后顺利到达目标点，路径如图7所示。本研究随后进行100次验证性实验，移动机器人到达目标点的成功率为83%，到达目标点平均所需时间步为56个，每实验5次后计算未到达目标点的失败率，随着实验次数的增加，失败率明显降低，验证性实验如图8所示。

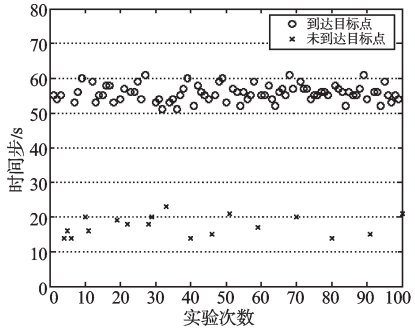


（a）首次实验图

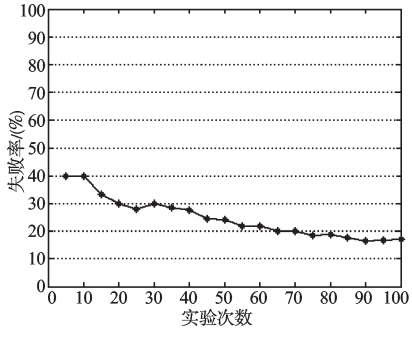


（b）训练50次后实验

图7 移动机器人路径(动静态障碍)



（a）实验结果统计



（b）失败率变化

图8 验证性实验(动静态障碍)

通过分析上述3组实验可以看出,实验初期,由于移动机器人对未知环境的先验知识不足,最终导致路径规划陷入冲突。多次训练后,移动机器人通过在线学习累积经验,完成了对未知环境的自适应,最终顺利地通过障碍环境到达目标点。另外,随着未知环境中障碍物复杂程度的增加,移动机器人到达目标点的成功率有所降低,但仍可以体现出移动机器人对未知环境有较高的自适应能力。

3　结束语

本研究提出了一种基于Q学习算法的路径规划方法,根据Q学习算法的必要因素设计学习系统结构,结合模糊逻辑方法对学习系统的输入状态进行泛化处理,在不同障碍环境中进行了仿真实验,结果表明了移动机器人具有较强的自学习能力,通过在线学习能顺利完成未知环境中的自适应路径规划,证明了该方法的有效性。

本研究为移动机器人平台的实地实验提供了理论基础和数据参考。

参考文献(References):

[1] ELSHAML A. Mobile Robots Path Planning Optimization in Sstatic and Dynamic Environments[D]. The University of Guelph,2004.

[2] YEN G G,HICKY T W. Reinforcement learning algorithms for robotic navigation in dynamic environments[J]. **ISA Transactions**,2004(43):217-230.

[3] 单建华. 未知环境下移动机器人实时模糊路径规划[J]. 机电工程,2009,26(1):1-4.

[4] ZAMSTEIN L M. Path Planning Using Reinforcement Learning on a Real Robot in a Real Environment[D]. The University of Florida,2009.

[5] 陈春林. 基于强化学习的移动机器人自主学习及导航控制[D]. 合肥:中国科学技术大学信息科学技术学院,2006.

[6] BEOM H R,CHO H S. A sensor-based navigation for a mobile robot using fuzzy-logic and reinforcement learning[J]. **IEEE Transactions on Systems, Man, and Cybernetics**, 1995,25(3):464-477.

[7] 秦　政,丁福光,边信黔. 强化学习在移动机器人自主导航中的应用[J]. 计算机工程与应用,2009,43(18):215-217.

[8] WATKINS J C H. Q-learning[J]. **Machine Learning**,1992(8): 279-292.

[9] 王炎欢,陈阿三,刘鑫茂. 直角坐标机器人控制系统的研制[J]. 轻工机械,2010,28(4):67-69.

[10] MARTINEZ-MARIN T,RODRIGUEZ R. Navigation of Autonomous Vehicles in Unknown Environments using Reinforcement Learning[C]//Proceedings of the IEEE Intelligent Vehicles Symposium Istanbul,2007.

[11] 陈卫东,李宝霞,朱奇光. 模糊控制在移动机器人路径规划中的应用[J]. 计算机工程与应用,2009,45(31):221-223.

[编辑:张　翔]

(上接第476页)

参考文献(References):

[1] SNYDER J M,WOODBURY A R,FLEISCHER K,et al. Interval Methods for Multi-Point Collisions between Time-Dependent Curved Surfaces[C]//Proceedings Siggraph 93,New York,1993:321-334.

[2] LENNERZ C,SCHÖMER E. Efficient Distance Computation for Quadratic Curves and Surfaces[C]//2nd Conference on Geometric Modeling and Processing, GMP'02, 2002: 60-69.

[3] CHEN X D,YONG J H,ZHENG G Q,et a1. Computing minimum distance between two implicit algebraic surfaces[J]. **Computer-Aided Design**,2006,38(10):1053-1061.

[4] JIMNEZ P,THOMAS F,TORRAS C. 3D collision detection: a survey[J]. **Computers & Graphics**, 2001, 25(2): 269-285.

[5] JOHNSON D E,COHEN E. A Framework for Efficient Minimum Distance Computations[C]//Proceedings of the 1998 IEEE International Conference on Robotics & Automation, Leuven,998:3678-3684.

[6] 冯果忱. 非线性方程组的迭代解法[M]. 上海:上海科学技术出版社,1989.

[7] 易大义,沈云宝,李有法. 计算方法[M]. 杭州:浙江大学出版社,2002.

[8] 周立春,陈　健,林海波,等. 内螺旋曲面数控拉刀设计及应用研究[J]. 轻工机械,2011,29(3):35-38.

[9] 钱　春. 基于区间牛顿法的点到参数曲线最小距离的计算方法[J]. 机电工程,2010,27(1):82-84.

[10] 廖朵朵,张华军. OpenGL三维图形程序设计[M]. 北京:星球地图出版社,1996.

[11] 李　颖,薛海斌,朱伯立,等. OpenGL函数与范例解析手册[M]. 北京:国防工业出版社,2002.

[编辑:张　翔]