文章编号:1006-2467(2009)11-1751-05

基于强化学习的自主移动机器人反应式自救控制

王忠巍, 曹其新, 栾 楠, 张 蕾
(上海交通大学 机器人研究所, 上海 200240)

摘 要: 为了解救陷入环境障碍的自主移动机器人,提出了一种基于强化学习的自救脱困控制方法. 该方法通过移动机器人与环境的交互作用,能够在线学习实现脱困自救的运动控制策略,并利用机器人自身条件克服环境障碍,避免了实施救援机器人的行动和终止其作业任务所造成的损失. 利用工作环境的先验知识指导,设计含有启发信息的强化学习系统回报函数,保证搜索和学习控制策略向正确方向进行,同时提高学习控制器的适应性和鲁棒性. 数字仿真证明了通过自学习控制策略实现自救脱困的可行性.

关键词: 自主移动机器人; 反应式控制; *Q* 学习; 启发式回报函数
中图分类号: TP 242 **文献标志码**: A

Reactive Self-rescue Control for Autonomous Mobile Robot Based on Reinforcement Learning

WANG Zhong-wei, CAO Qi-xin, LUAN Nan, ZHANG Lei
(Research Institute of Robotics, Shanghai Jiaotong University, Shanghai 200240, China)

Abstract: A control technique to achieve self-rescue of autonomous mobile robot from obstacle environment based on reinforcement learning was proposed. Motion control strategy of self-rescue was got on line through the interaction between the mobile robot and the obstacle environment, so the self-rescue control technique helps to overcome obstacle environment by the autonomous mobile robot independently and avoid the loss from rescue activity and task failure. Prior knowledge of working environment was applied to direct the design of heuristic reward function for the reinforcement learning system, which guarantees the correct direction of searching and learning control strategy. The simulation experiments indicate that it is feasible to achieve self-rescue by self learning control strategy.

Key words: autonomous mobile robot; reactive control; *Q*-learning; heuristic reward function

在很多不适于人类工作或者难以到达的环境中,需要特种机器人执行特殊的任务,如海底管道检测机器人,水下机器人,星球探索机器人等. 这些机器人在自主移动、作业过程中不可避免地会遇到各种未知的环境障碍,一旦陷入障碍又无法脱困,就会造成严重的设备损失和任务失败损失. 因此,有效应对环境障碍成为自主移动机器人领域的一个研究热点. 目前,对于这一问题的研究主要集中在避障导航的方法[1-2],提高越障能力的新颖运动机构设计[3-4],以及专用救援装置[5-6]等方面,而自主移动机器人在

收稿日期:2008-10-14
基金项目:国家自然科学基金资助项目(50705054),国家高技术研究发展计划(863)项目(2006AA040203)
作者简介:王忠巍(1978-),男,辽宁辽阳人,博士生,主要从事自主移动机器人智能控制技术和嵌入式控制技术的研究.
曹其新(1960-),男,教授,博士生导师,电话(Tel.):021-62932750;E-mail:qxcao@sjtu.edu.cn.

陷入环境障碍时能够基于自身运动控制策略脱离困境,继续完成作业任务,无疑更具吸引力.

在未知环境中作业的移动机器人控制研究中,Brooks[7]提出了基于行为的反应式控制思想,使得机器人对未知环境具有一定的适应性,但其缺乏有效的高层知识表示,仅能完成一些基本的类似昆虫的智能行为,只适用于在未知环境下执行较简单的任务.随着智能控制和计算智能方法研究的发展,各种用于自主移动机器人反应式智能控制的方法也被提出,如基于模糊逻辑的控制器[8]、基于神经网络的控制器[9]、利用模糊神经网络构建控制器[10]等.这些智能控制方法需要各种状态下的教师信号进行监督学习,然后应用于未知环境中,本质上是凭借其鲁棒性与泛化能力处理未经过学习的状态.

基于反复试错的强化学习算法(Reinforcement Learning,RL)能够实现 Agent 与环境交互中在线学习理想控制策略,是一种将面向任务的有意识行为与基于传感器的反应式行为有机结合的自主机器人控制系统设计方法[11],它为实现故障环境中移动机器人的自救控制提供了理论基础.然而,传统的 RL 算法都假设无先验知识、系统参数未知,由此造成单纯强化学习算法,没有任何基础地搜索最优策略,搜索范围大、实时性差,无法满足设计移动机器人智能控制器的需要.实际上,在设计移动机器人智能控制器时,总会具有关于控制策略和应用环境的先验知识,因此,合理的 RL 学习系统应该在先验知识的基础上搜索最优控制策略,先验知识的融入能够有效提高 RL 系统的学习效率[12-13].本文提出基于先验知识指导,设计含有启发信息的 RL 系统回报函数,加快搜索有效的控制策略,并用数字仿真陷入环境障碍的过山车自救控制为例,证明了通过自学习控制策略实现自救脱困的可行性.

1 环境障碍中的过山车自救控制任务

陷入环境障碍的过山车如图 1 所示.图中,曲线代表一个山谷的地形,S 为山谷最低点,A 为左端最高点,G 为右端最高点.过山车陷入山谷中,其任务是从谷底 S 点出发并尽快运动到右端最高点 G,达到脱离环境障碍的目的.在这里,过山车被设计成为一个动力不足的系统,用来表明所遇到的环境障碍已超过其设计指标,所以一直控制过山车加速前进不能使它到达 G 点,而必须首先让它后退,反方向爬坡一定高度以积累一定势能,然后才能控制其到达 G 点.

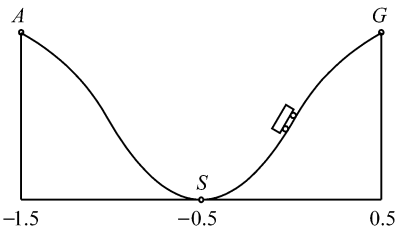


图 1 陷入环境障碍的过山车示意图

Fig. 1 Schematic of mountain-car being in environmental obstacles

陷入环境障碍的过山车数字仿真系统动力学特性方程为[11]

$$\left.\begin{aligned} x_{t+1} &= \mathrm{bound}[x_t + v_{t+1}] \\ v_{t+1} &= \mathrm{bound}[v_t + 0.001u - g\cos 3x_t] \end{aligned}\right\} \quad (1)$$

系统具有二维的连续状态空间,分别用小车的水平位移 x 和水平运动速度 v 表示.小车控制量 u 具有 3 个离散的取值,$u \in \{+1, -1, 0\}$,分别代表加速、减速和匀速 3 种控制行为.设重力 $g = 0.0025$;G 点、S 点和 A 点的 x 取值分别为 0.5、-0.5 和 -1.5.式(1)中 bound[ ]运算将 x 和 v 限制在它们各自的范围内:如果 $x \leq -1.5$,则设置 $x = -1.5$,同时 $v_{t+1} = 0$;如果 $x \geq 0.5$,则过山车自救控制任务完成.$v \in [-0.07, 0.07]$.

2 基于先验知识的强化学习算法

强化学习是指从环境状态到动作映射的学习,它不同于监督学习技术那样通过正、反例告知采取何种行为,而是基于不断地和环境交互得到外部环境评价信号,并通过谋求最大化环境评价来发现最优行为策略,因此,强化学习具有自学习和在线学习的特点.

2.1 Q 学习算法

Q 学习算法[11]是强化学习的主要算法之一,是一种无模型的强化学习方法,它提供 Agent 在马尔可夫环境中,利用经历的动作序列选择最优动作的一种学习能力.Q 学习算法所依赖的离散马尔可夫过程可解释为:在时刻 t,Agent 在有限动作集合中选取动作 a_t,环境接受该动作后由状态 s_t 转移到 s_{t+1},同时给出评价 r_t,r_t 和 s_{t+1} 的概率分布取决于 a_t 和 s_t.Q 学习算法实际上是马尔可夫过程的一种变化形式,但它不去估计环境模型,而是直接优化一个可迭代计算的 Q 函数,此 Q 函数为在状态 s_t 时执行动作 a_t,且此后按最优动作序列执行时的折扣累计强化值,

$$Q_{t+1}(s_t, a_t) = r_t + \gamma \max_{a_t}\{Q(s_{t+1}, a_t) \mid a_t \in A\} \quad (2)$$

由于 Q 学习的无导师自适应能力,故基于 Q 学习 Agent 可以实现其行为自主性.

2.2　带有启发式回报函数的 Q 学习算法

强化学习系统与外界环境之间的交互是其获得智能的主要来源,回报函数是学习系统获取外部环境信息的方式和评估动作效果的依据. 在学习过程中,学习系统收到有效回报的多少直接决定了学习算法的优劣,因此,回报函数的设计是关系到学习效果的关键因素,也是构建强化学习系统中最重要和最困难的方面. 一般情况下回报函数的设计只与最终目标相关联,如 Sutton 等[11]的过山车仿真程序,其采用的回报函数为

$$r(t) = \begin{cases} -1, & x < 0.5 \\ 0, & x \geq 0.5 \end{cases} \tag{3}$$

即当过山车越过环境障碍时获得零回报,其余每步均为负回报. 该回报函数形式简单、意义明确,但提供给学习系统的与环境交互过程信息非常少,学习系统只有多次遍历所有状态后,才能学习到较好的控制规律.

理想的回报函数不仅与最终目标相关联,还应能指导学习系统选择更为有利的动作,快速获得好的控制策略. 因此,本文在设计陷入环境障碍的过山车自救控制学习算法时,利用先验知识设计含有启发信息的回报函数,刺激某个状态下执行特定动作,引导学习算法更快地学会自救控制策略,相应的强化学习模型如图 2 所示.

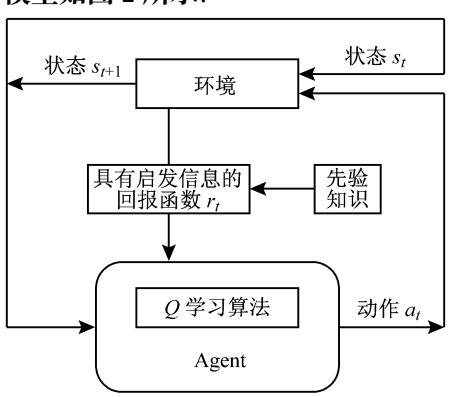


图 2　基于启发式回报的 Q-学习算法框架结构

Fig. 2　Framework of Q-learning based on heuristic reward function

图 1 中的过山车在每个状态可选择加速、减速和匀速 3 个控制行为,由生活中的经验知识很容易联想:当采取的控制行为与速度方向一致时,通常会使过山车的动能更强,由此克服环境障碍的能力就更强. 基于该启发信息设计的回报函数可表述为:①如果采取的动作使过山车沿初始运动方向的位移增加,则取得正回报,其余情况为负回报;②在符合正回报条件下,速度为正方向时,则取得回报值 +2;速度为负方向时,则取得回报值 +1. 具体数学表达如下:

$$r(t) = \begin{cases} +2, & (x_{t+1} - x_t)\,u \geq 0, v(t) \geq 0 \\ +1, & (x_{t+1} - x_t)\,u \geq 0, v(t) < 0 \\ -1, & \text{其　他} \end{cases} \tag{4}$$

由前面分析可知,过山车的控制量始终沿一个方向,不能使其脱离环境障碍. 因此,依据启发式回报函数指导选择特定控制行为的同时,智能控制器应以一定的概率 ε 随机选择其他可能的动作,通过探索不同的控制行为序列,学习有效的脱困自救控制策略. 利用先验知识设计启发式回报函数的 Q 学习算法如下:

(1) 初始化状态空间、算法参数.

(2) 观察当前的状态 s_t. ①依据当前 Q 值表,以概率 $1-\varepsilon$ 按贪心策略选择并执行一个动作 a_t;②以概率 ε 随机选择并执行一个动作 a_t.

(3) 观察下一个状态 s_{t+1},依据式(4)计算立即强化信号 r_t.

(4) 根据强化信息 r_t,调整当前 Q 值表,

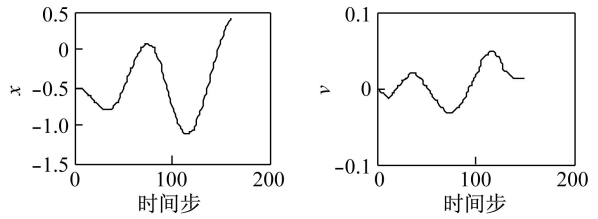
$$Q_t(s_t, a_t) = \begin{cases} (1-a_t)\,Q_{t-1}(s_t, a_t) + a_t[r_t + gV(s_{t+1})] \\ \qquad s = s_t, a = a_t \\ Q_{t-1}(s_t, a_t) \qquad \text{其　他} \end{cases}$$

其中,$V(s_{t+1}) = \max\limits_{a \in A}\{Q_{t-1}(s_{t+1}, a)\}$.

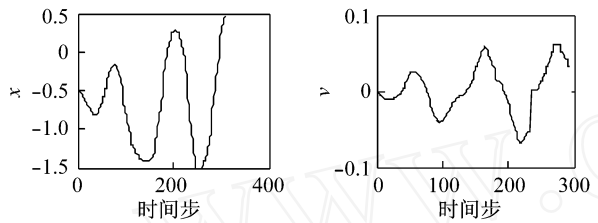
3　仿真实验

针对过山车自救脱困控制任务,分别设计了基于普通回报函数的 Q 学习控制器和基于启发式回报函数的 Q 学习控制器. 目的是通过对比两者的控制效果,表明含有经验知识的启发式回报函数具有提高 RL 算法学习效率的作用.

2 种 Q 学习算法的参数均设定为:学习率 $\alpha = 0.5$,折扣因子 $\gamma = 0.95$,贪心概率 $\varepsilon = 0.1$. 过山车的初始状态为 $x = -0.5$ 和 $v = 0$. 图 3 给出基于 2 种 Q 学习算法的学习控制器数字仿真结果. 其中,图 3(a)是在经过 200 次监督学习后得到的过山车自救脱困过程曲线(监督学习:当小车到达 G 点或时间步数超过 1 000 步,结束一次学习实验,然后重新初始化系统状态,开始下一次学习). 图 3(b)是在未经过监督学习的一次随机实验中所得到的过山车自救脱困过程曲线.



(a) 采用普通回报函数



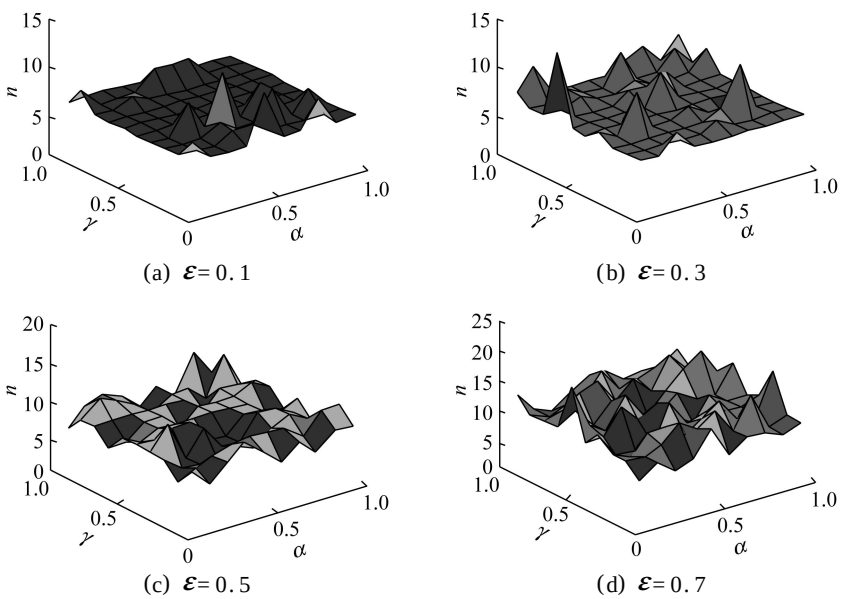
(b) 采用启发式回报函数

图 3 采用不同回报函数的过山车位置、速度曲线

Fig. 3 Variation curve of position and velocity of mountain car with different reward function

由图可见,基于普通回报函数的 Q 学习算法经过监督学习后,能够以最优的控制行为实现自救脱困任务,且用时最少;基于启发式回报函数的 Q 学习算法虽然不能以最优控制行为完成自救脱困,但能够通过数个试探动作,学会反方向爬坡积累势能,实现自救脱困的控制技巧,其无需进行监督学习便能快速、有效地实现脱离障碍环境的目的. 产生如此好效果的原因在于:基于先验知识设计的启发式回报函数为 Q 学习控制器搜索控制策略提供了一个其本正确的方向,而 Q 学习算法的适应能力弥补了先验知识的不完善性,因此,能快速搜索到好的自救控制策略.

理想的移动机器人自救控制策略是在尽量少的动作试探后实现脱离障碍环境. 一般来说,控制算法中的参数设置对控制器性能的影响较大,而移动机器人可能遇到的环境障碍是不确定的,控制器参数也就无法事先设计成最优,所以理想的控制器需要对参数设置不敏感,并表现出很强的适应性和鲁棒性. 文中采用过山车在实现脱困过程中经过位置最低点 $x = -0.5$ 的次数(动作试探次数 n)来衡量学习控制器的优劣,并通过数字仿真得到过山车脱困试探次数 n 与控制器参数的关系,如图 4 所示.



(a) $\varepsilon = 0.1$ (b) $\varepsilon = 0.3$

(c) $\varepsilon = 0.5$ (d) $\varepsilon = 0.7$

图 4 过山车脱困试探次数 n 与控制算法参数的关系

Fig. 4 Relationship between trials number of mountain car's overcoming obstacle and parameters of control algorithm

由图可见,即使在设置最糟的控制器参数条件下,过山车做 20 次左右的试探动作后仍能实现自救脱困. 这说明基于先验知识设计的启发式回报函数在确保学习向正确方向进行的同时,降低了控制器对控制参数的敏感性,提高了学习算法的适应性和鲁棒性.

4 结 语

本文提出了一种基于 Q 学习算法的学习控制器,并利用工作环境的先验知识作为启发信息设计 Q 学习控制器的回报函数,保证了搜索和学习控制策略向正确方向进行. 数字仿真陷入环境障碍的过

山车自救控制表明:融入先验知识的 Q 学习控制器能够快速、在线学习到好的自救控制技巧,同时表现出良好的适应性和鲁棒性.

参考文献:

[1] Tsalatsanis A, Valavanis K, Tsourveloudis N. Mobile robot navigation using sonar and range measurements from uncalibrated cameras [J]. **Journal of Intelligent and Robotic Systems**, 2007, 48(2): 253-284.

[2] Mata M, Armingol J M, Fernndez J, *et al*. Object learning and detection using evolutionary deformable models for mobile robot navigation [J]. **Robotica**, 2008, 26(1): 99-107.

[3] Lan G P, Ma S G. Development of a novel crawler mechanism with polymorphic locomotion [J]. **Advanced Robotics**, 2007, 21(3/4): 421-440.

[4] 邓宗全,高海波,王少纯,等. 行星轮式月球车的越障能力分析[J]. **北京航空航天大学学报**,2004,30(3): 197-201.
DENG Zong-quan, GAO Hai-bo, WANG Shao-chun, *et al*. Analysis of climbing obstacle capability of lunar rover with planetary wheel [J]. **Journal of Beijing University of Aeronautics and Astronautics**, 2004, 30(3): 197-201.

[5] 向先波,徐国华,蔡　涛,等. 水下机器人智能自救系统[J]. **华中科技大学学报(自然科学版)**, 2006, 34(7):111-114.
XIANG Xian-bo, XU Guo-hua, CAI Tao, *et al*. Intelligent self-rescue system for underwater vehicles [J]. **Journal of Huazhong University of Science and Technology (Natural Science Edition)**, 2006, 34(7): 111-114.

[6] Wang Z W, Cao Q X, Luan N, *et al*. Development of new pipeline maintenance system for repairing early-built offshore oil pipelines [C]// **IEEE ICIT 2008**. Piscataway, United States: IEEE Publishers, 2008: 1-6.

[7] Brooks R A. Robust layered control system for a mobile robot [J]. **IEEE Journal of Robotics and Automation**, 1986, 2(1): 14-23.

[8] Faress K N, EI Hagry M T, EI Kosy A A. Trajectory tracking control for a wheeled mobile robot using fuzzy logic controller [J]. **WSEAS Transactions on Systems**, 2005, 4(7): 1017-1021.

[9] Wang X Q, Hou Z G, Zou A M, *et al*. A behavior controller based on spiking neural networks for mobile robots [J]. **Neurocomputing**, 2008, 71(4-6): 655-666.

[10] Er M J, Tan T P, Loh S Y. Control of a mobile robot using generalized dynamic fuzzy neural networks [J]. **Microprocessors and Microsystems**, 2004, 28(9): 491-498.

[11] Sutton R S, Barto A G. Reinforcement Learning: An Introduction [M]. Cambridge, MA: MIT Press, 1998.

[12] Mataric M J. Reward functions for accelerated learning [C]// **Proceedings of the Eleventh International Conference on Machine Learning**. San Francisco, CA: Morgan Kaufmann Publishers, 1994: 181-189.

[13] 魏英姿,赵明扬. 强化学习算法中启发式回报函数的设计及其收敛性分析[J]. **计算机科学**,2005,32(3): 190-193.
WEI Ying-zi, ZHAO Ming-yang. Design and convergence analysis of a heuristic reward function for reinforcement learning algorithms [J]. **Computer Science**, 2005, 32(3): 190-193.

下期发表论文摘要预报

基于 Zernike 矩的有限角度 CT 重建方法

戴修斌,　舒华忠,　罗立民

(东南大学 计算机科学与工程学院 影像科学与技术实验室, 江苏 南京 210096)

摘　要:提出了一种基于 Zernike 旋转正交矩的有限角度扇形束投影数据 CT 图像重建方法. 该方法建立了扇形束投影数据和 Zernike 图像矩之间的关系,并利用这种关系从已知角度投影数据中估计未知角度投影数据,从而达到提高重建图像质量的目的. 实验证明,所提出的方法可以有效地估计未知投影数据,并能获得更好的重建结果.