DOI: [illegible] cnki jdxbgxb[illegible]
第 36卷 增刊 2
2006年 9月
吉 林 大 学 学 报 (工 学 版)
Journal of Jilin University (Engineering and Technology Edition)
Supplement 2
Sept. 2006

文章编号: 1671-5497(2006) Suppl 2-0108-05

基于分层强化学习的多移动机器人避障算法

祖丽楠, 田彦涛, 梅 昊

(吉林大学 通信工程学院, 长春 130022)

摘 要: 介绍了一种基于分层思想的强化学习方法, 即将机器人的复杂行为分解为一系列简单的行为进行离线独立学习, 并分别设计了每个层次的结构、参数及函数。这种学习方法能够减小状态空间并简化强化函数的设计, 从而提高了学习的速率以及学习结果的准确性, 并使学习过程实现了决策的逐步求精。最后以多机器人避障为任务模型, 将避障问题分解为躲避静态和动态障碍物以及向目标点靠近 3 个子行为分别进行学习, 实现了机器人的自适应行为融合, 并利用仿真实验对其有效性进行了验证。

关键词: 自动控制技术; 避障; 强化学习; Q-学习; 分层学习

中图分类号: TP24 文献标识码: A

Obstacle avoidance of multi mobile robots based on hierarchical reinforcement learning

Zu Li-nan, Tian Yan-tao, Mei Hao

(*College of Communication Engineering, Jilin University, Changchun 130022, China*)

Abstract: A reinforcement learning algorithm based on the idea of partition layer was proposed that decomposing the complicated problem into a series of simple portions to be learned independently. The structures, parameters and functions of every level were designed. This learning algorithm could reduce the status space and predigest the design of reinforcement functions so as to improve the learning speed and the veracity of learning results. Also, it could realize the accuracy of the learning process step by step. Finally, the method was used for adaptive action fusion of mobile robot in an "obstacle avoidance" task by decomposing it into avoiding static and dynamic obstacle and closing to object actions. And its efficiency was shown by simulation results.

Key words: automatic control technology; obstacle avoidance; reinforcement learning; Q-learning; multi-level learning

对多移动机器人避障问题的研究人们常采用 Brooks[1] 提出的基于行为的反应式控制、基于规则的控制策略[2] 等方法。这些方法直接、确定, 机器人表现出了较快的反应性和实时性。但当任务和环境变得复杂时, 它们将会无法控制, 从而产生错误的决策以至死锁。另外, 这些方法完全依靠设计者给出完整的控制策略, 需要大量的经验和领域知识, 因此, 使机器人具备自学习能力是提

收稿日期: 2005-08-14.

基金项目: 吉林省科技发展计划重大项目 (20050326).

作者简介: 祖丽楠 (1979-), 女, 博士研究生. 研究方向: 分布式智能系统与网络控制. E-mail: zulinan-zln@163.com

通讯联系人: 田彦涛 (1958-), 男, 教授, 博士生导师. 研究方向: 分布式智能系统. E-mail: tianyt@jlu.edu.cn

高其对环境适应性和鲁棒性等智能特性的有效方法。在诸多学习方法中, 强化学习得到了最广泛的关注[3]。但在具体应用时, 若输入状态选择不合适, 学习系统的状态空间可能会很大; 同时强化函数的设计不同, 其对行为的评价也不同。这都导致了该方法收敛速度慢、学习效果不确定等问题[4]。

为了解决上述问题, 本文以多机器人避障问题为应用背景研究机器人独立学习算法。

1　分层强化学习算法

强化学习是一种以环境反馈作为输入的[5]、实时的、无模型的增量式学习方法。由于强化学习系统很少依赖外部的指导信息、不需要建立环境和任务的精确数学描述以及具有较高的适应性和较快的反应性等特性, 使其在大空间、复杂非线性系统中得到广泛的应用, 并已扩展到智能探索、监控学习和结构控制等领域, 尤其在多机器人协作行为中的应用最广泛[6-7]。

本文以多机器人系统避障任务为模型, 设计了学习系统的结构、学习器的参数以及相关函数, 并采用强化学习方法中较为常用的 Q 学习[8]方法对上述行为进行离线独立学习。

1.1　分层强化学习系统结构

强化学习系统结构如图 1所示。在避障任务中, 机器人要学习躲避环境中的各种障碍物。而学习时若将所有的障碍物都视为一种情况, 则机器人必须将复杂的感知信息直接映射到庞大的行为集上。这样, 不仅输入的状态空间非常大, 并且由于对不同障碍物的躲避方式不同, 导致对其行为的评价方式也不同。若通过一个强化函数将所有的信息描述出来, 其设计也是相当复杂的, 并且会降低机器人避障的准确度。因此, 本文采用分层的思想, 将机器人的避障行为分解成 3个简单的子行为: 躲避静态障碍物行为、躲避动态障碍物行为和向目标点靠近行为。为了减小状态空间以及简化强化函数的设计, 将学习划分为两个层次: 静态避障 Q-学习层和动态避障 Q-学习层, 将向目标点靠近行为融合到每个层次中。静态避障 Q-学习层需要实现躲避静态障碍物以及向目标点靠近, 动态避障 Q-学习层需要实现躲避动态障碍物以及避障行为的融合。

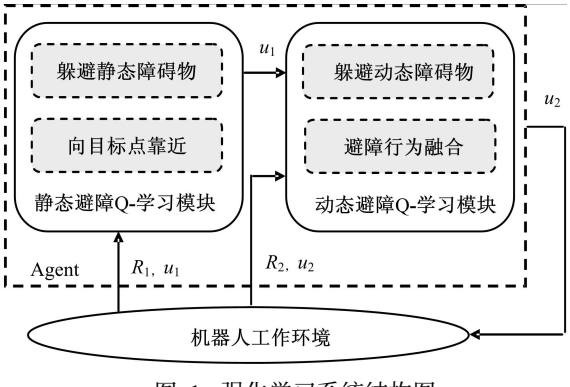


图 1　强化学习系统结构图

Fig 1　Structure of reinforcement learning system

1.2　静态避障 Q-学习模块

1.2.1　环境信息的获取

首先假设传感器在理想状态下工作, 机器人能实时检测自身以及目标和障碍物的状态信息。为简化问题, 只考虑二维环境中机器人运动方向正前方 135°范围内的避障问题, 如图 2所示。

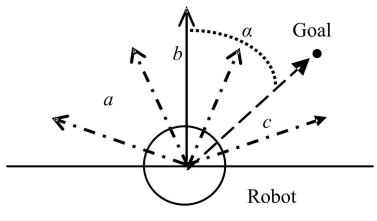


图 2　机器人传感器配置

Fig 2　Sensor configure of robot

将该范围划分为 3个分别为 45°的区域: *a*、*b*、*c*, 在每个区域中依具体任务设置不同数量的传感器构成一个传感器组, 同时给出如下一组定义: m 为机器人到障碍物间的距离; n 为机器人到目标点间的距离; $\boldsymbol{v}$ 为机器人运动方向; α 为 $\boldsymbol{v}$ 与机器人到目标点间连线的夹角; a_i 为权重系数; r_{min} 为最小危险距离 (本文设为 10); r_{max} 为安全距离 (本文设为 35); $r_m \in (r_{min}, r_{max})$ 为避碰区域; π /6 为 $\boldsymbol{v}$ 与障碍物间的安全角度。

1.2.2　输入和输出状态空间设计

为了简化状态空间, 同时还能给机器人提供避障所需的必要信息, 将 3个传感器组分别探测到的机器人与静态障碍物间的距离 $m_i(t)$ (i=1, 2, 3)和 α 作为静态避障模块的输入, 分别表示为 $m(i)$ 和 An。对连续的输入信号采用 Box方法进行量化, 结果为

$$m(i)=\begin{cases}0 & m_i \leqslant 10\\ 1 & 8<m_i\leqslant 35,\ i=a,\ b,\ c\\ 2 & 35<m_i\end{cases}$$

$$An=\begin{cases}0 & -\pi/36\leqslant\alpha\leqslant\pi/36\\ 1 & -\pi/6\leqslant\alpha<-\pi/36\\ 2 & \pi/36<\alpha\leqslant\pi/6\\ 3 & -\pi\leqslant\alpha\leqslant-\pi/6\\ 4 & \pi/6<\alpha\leqslant\pi\end{cases}$$

强化学习系统的输出为机器人的旋转角度:±20°、±10°和0°(正号表示机器人以自身运动方向为 y 轴向左偏转;负号则相反)和速度大小:0.5.10 m/s 同样采用 Box方法对其量化为15个输出值,分别为速度和旋转角度的组合对。

1.2.3 强化信号的计算

避障任务是一个多目标行为,即远离障碍物和趋近目标点行为,所以强化信号应该取两个目标函数的加权和。考虑到机器人与障碍物间的距离不同导致其行为的侧重点也应该不同,所以奖励函数应依据具体情况而定。本文在设计躲避静态障碍物行为的奖励函数 R_t^1 时,依据机器人与障碍物间的距离 m 给出了其奖励方式

$$R_t^1=\begin{cases}-10 & m<r_{\min}\\ f(m,n) & \text{otherwise}\\ f(n,\alpha) & m>r_{\max}\end{cases}\qquad(1)$$

当 $m<r_{\min}$ 时,机器人与障碍物间发生碰撞的可能性很大,所以应立即给予最大程度的惩罚,本文将其设为 -10;当 $m>r_{\max}$ 时,机器人与障碍物间发生碰撞的可能性很小,所以此时应该重点考虑其趋近目标点行为,即机器人与目标点间的距离和角度的关系,令奖励函数为 $f(n,\alpha)$,则

$$f(n,\alpha)=w_1\,\mathrm{sign}(n_t-n_{t+1})(n_{t+1}-n_t)^2+w_2(\alpha_t-\alpha_{t+1})\qquad(2)$$

当 $r_{\min}\leqslant m\leqslant r_{\max}$ 时,机器人与障碍物间存在碰撞的潜在危险,但情况不是很紧急,所以应令其在避障的同时向目标点靠近,这样既不会使机器人过远地偏离目标,而且即使在要发生碰撞的时候也有足够的时间进行修正。令奖励函数为 $f(m,n)$,则

$$f(m,n)=w_3\,\mathrm{sign}(m_{t+1}-m_t)(m_{t+1}-m_t)^2+w_4\,\mathrm{sign}(n_t-n_{t+1})(n_{t+1}-n_t)^2\qquad(3)$$

式中:w_i($i=1,2,3,4$)为权重系数。

1.2.4 动作选择策略

本文采用 lookup表作为Q值表的存储方式。学习系统根据存储在Q值表中当前状态下与各个动作对应的Q值来选择动作。本文选择了Boltzmann方法作为行为探索的策略。对于某个状态 s,由式(4)确定要执行的动作 a_i

$$p(s,a_i)=e^{Q(s,a_i)/T}\Big/\sum_{a_k\in A}e^{Q(s,a_k)/T}\qquad(4)$$

式中:$Q(s,a_i)$为"状态-动作"的值函数;T为"温度"系数。

1.3 动态避障Q-学习模块

1.3.1 环境信息的获取

为分析机器人避障条件,给出如下定义。

定义1 以机器人运动方向作为 y 轴建立一个坐标系,称之为虚拟坐标系,并将障碍物的实际运动方向映射到虚拟坐标系的各个象限中。

与1.2.1节相同,机器人通过自身配置的传感器来探测环境信息。同时增加了两个变量的定义:v_o 为障碍物的速度;γ 为动态障碍物运动方向在虚拟坐标系的对应角度。

1.3.2 输入和输出状态空间设计

为了回避直接行为融合所产生的误差,将静态避障Q-学习模块的输出结果作为动态避障学习模块的输入,间接对两种避障行为的融合进行学习。为了简化输入状态空间,本节将如下几个变量作为该模块的输入:机器人通过三组传感器测得的与动态障碍物间距离的综合信息 $m(t)$、障碍物位置与机器人运动方向的相对位置(在其左、右两侧)及障碍物运动方向在虚拟坐标系中所处的象限的组合、障碍物的速率 $|v_o(t)|$ 以及静态障碍物的输出结果 St。同样,采用Box方法对其进行量化后分别表示为

$$Com=\begin{cases}0 & \text{左,第一象限}\\ \vdots & \\ 3 & \text{左,第四象限}\\ 4 & \text{右,第一象限}\\ \vdots & \\ 7 & \text{右,第四象限}\end{cases}\quad St=\begin{cases}0,(0°,-20°)\\ \vdots\\ 14,(10°,20°)\end{cases}$$

$$|v_o(t)|=\begin{cases}0 & \text{大}\\ 1 & \text{中}\\ 2 & \text{小}\end{cases}\quad m(t)=\begin{cases}0 & m_t\leqslant 10\\ 1 & 8<m_t\leqslant 35\\ 2 & 35<m_t\end{cases}$$

其中,Com 中的"左(右)、第 i 象限"表示障碍物当前位置处于机器人运动方向的左(右)侧,并且障碍物的运动方向处于虚拟坐标系的第 i 象限。

学习系统的输出与静态避障Q-学习模块的输出形式相同。

1.3.3 强化信号计算

动态避障Q-学习模块的学习过程需要对每个行为同时进行评价,其强化信号应该是多个行为的强化值之和。

令 R_t 为动态避障 Q-学习模块的奖励函数，R_t^2 为单独躲避动态障碍物的奖励函数，则有

$$R_t = R_t^1 + R_t^2 \tag{5}$$

其中，R_t^1 的计算如式(1)，R_t^2 的计算如下

$$R_t^2 = \begin{cases} -10, & m < r_{\min} \\ f(m, \gamma), & \text{otherwise} \end{cases} \tag{6}$$

当障碍物位置处于机器人运动方向右侧时有

$$f(m, \gamma) = \begin{cases} a_1 \operatorname{sign}(m_{t+1} - m_t)(m_{t+1} - m_t)^2 + \\ a_2(\frac{\pi}{2} - \gamma)/\frac{\pi}{2}, \quad 0 \leqslant \gamma < \frac{\pi}{2} \\ a_1 \operatorname{sign}(m_{t+1} - m_t)(m_{t+1} - m_t)^2 + \\ a_2(\gamma - \frac{3\pi}{2})/\frac{\pi}{2}, \quad \frac{3\pi}{2} \leqslant \gamma < 2\pi \\ a_1 \operatorname{sign}(m_{t+1} - m_t)(m_{t+1} - m_t)^2 - \\ a_2(\gamma - \frac{\pi}{2})/\frac{\pi}{2}, \quad \frac{\pi}{2} \leqslant \gamma < \pi \\ a_1 \operatorname{sign}(m_{t+1} - m_t)(m_{t+1} - m_t)^2 - \\ a_2(\frac{3\pi}{2} - \gamma)/\frac{\pi}{2}, \quad \pi \leqslant \gamma < \frac{3\pi}{2} \end{cases} \tag{7}$$

当障碍物位置处于机器人运动方向左侧时有

$$f(m, \gamma) = \begin{cases} a_1 \operatorname{sign}(m_{t+1} - m_t)(m_{t+1} - m_t)^2 - \\ a_2(\frac{\pi}{2} - \gamma)/\frac{\pi}{2}, \quad 0 \leqslant \gamma < \frac{\pi}{2} \\ a_1 \operatorname{sign}(m_{t+1} - m_t)(m_{t+1} - m_t)^2 - \\ a_2(\gamma - \frac{3\pi}{2})/\frac{\pi}{2}, \quad \frac{3\pi}{2} \leqslant \gamma < 2\pi \\ a_1 \operatorname{sign}(m_{t+1} - m_t)(m_{t+1} - m_t)^2 + \\ a_2(\gamma - \frac{\pi}{2})/\frac{\pi}{2}, \quad \frac{\pi}{2} \leqslant \gamma < \pi \\ a_1 \operatorname{sign}(m_{t+1} - m_t)(m_{t+1} - m_t)^2 + \\ a_2(\frac{3\pi}{2} - \gamma)/\frac{\pi}{2}, \quad \pi \leqslant \gamma < \frac{3\pi}{2} \end{cases} \tag{8}$$

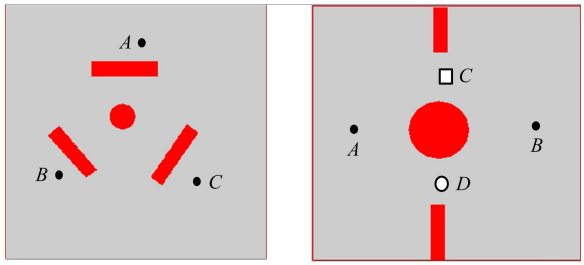
1.3.4　动作选择策略

该模块在学习过程中采用了 lookup 表作为 Q 值表的存储方式，与 1.2.4 节相同，采用 Boltzmann 分布方法作为其动作选择策略。

2　仿真实验及其结果

2.1　任务描述及训练策略

图 3(a)为静态避障行为的训练环境。图中有三个点 A，B，C。机器人从某一点出发，同时任意选择另外两点之一作为其目标点，到达之后，将其作为出发点并重复上述过程直至仿真过程结束。图 3(b)为动态避障行为的训练环境。环境中的 A、B 两点为起始点和目标点，机器人在两点之间作往返运动，在环境中的 C、D 两处各有一动态障碍物作上下往返运动以阻止机器人通过。



(a)　　(b)

图 3　躲避静态、动态障碍物训练环境

Fig. 3　Training environment for avoiding static and dynamic obstacles

2.2　仿真结果

机器人在图 3(a)所示的任务环境中训练躲避静态障碍物行为，经过 1000 次后停止学习，对其学习结果的测试如图 4 所示。图中 A、B 分别为起始点和目标点。从图中的运行轨迹可以看到，机器人能够平滑地绕开静态障碍物，并且在避障的同时始终保持着向目标点靠近的状态。

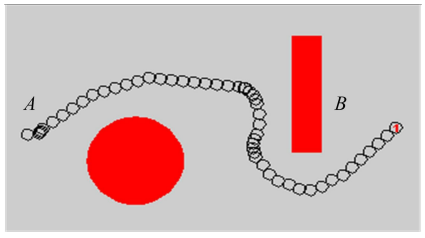


图 4　机器人躲避静态障碍物

Fig. 4　Robot avoid static obstacles

机器人在图 3(b)所示的任务环境中训练躲避动态障碍物行为，经过 1000 次后停止学习，对其学习结果的测试如图 5、6所示。在图 5中，机器人分别在 A、D 两处和 B、C 两处向相对方向行进，途中 4个机器人相遇并进行躲避。图 6中机器人由 A 点向 B 点行进，途中不仅存在静态障碍物，而且在 C 点处遇到动态障碍物并进行躲避。从上述图中可以看到，机器人与障碍物或其他机

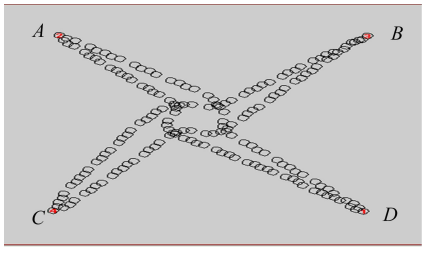


图 5　机器人之间避障

Fig. 5　Robots avoid each other

器人间不仅能够成功地进行躲避,而且在避障的同时都能朝着目标点行进并最终到达,成功地实现了对多个行为的融合。

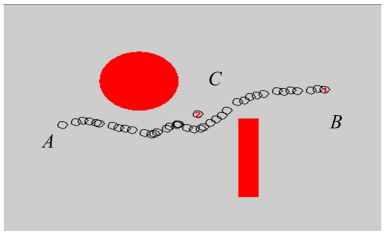


图 6 机器人躲避各种障碍物

Fig. 6 Robot avoid variety of obstacle

2.3 算法收敛性分析

图 7为学习躲避静态和动态障碍物行为时系统每步获得的平均奖励随试验次数的变化曲线。图 8为机器人学习躲避静态障碍物行为和动态障碍物行为时策略改变次数随试验次数变化的曲线。由图可以看出,学习系统收到的平均奖励次数随试验次数的增加而增加;策略平均改变次数随试验次数的增加而减小。初始阶段曲线变化较大,这是由于此时学习策略还不成熟,具有较强的随机性,但随着学习的进行算法开始收敛,策略逐渐趋于平稳,并最终保持在“无策略改变”的状态。此时,平均奖励也收敛在一个稳定的范围内。

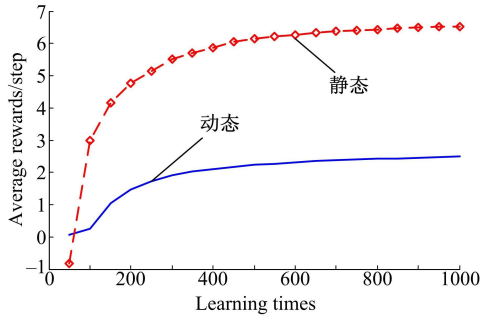


图 7 学习避障行为时平均奖励变化曲线

Fig. 7 Curves of average rewards for obstacle avoid learning

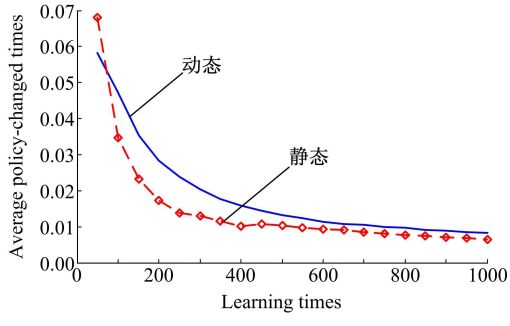


图 8 学习避障行为时策略平均改变次数变化曲线

Fig. 8 Curves of average policy changed versus obstacle avoid learning

3 结束语

本文基于分层的思想设计了一种避障学习结构,对各个模块的输入、输出状态以及强化信号的计算方法分别进行了设计。通过计算表明,若将所有障碍物视为一种情况进行学习,输入状态空间的大小为:3×3×3×5×3×3×8=9720个状态;而采用本文设计的分层结构进行学习时,其输入空间最大为 1080个状态。由此可见分层学习结构明显减小了输入状态空间的大小,从而提高了学习速度。

参考文献:

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

[2] 宋梅萍, 顾国昌, 张汝波. 移动机器人的自适应式行为融合方法[J]. 哈尔滨工程大学学报, 2005, 26(5): 586-613.
Song Meiping, Gu Guochang, Zhang Rubo. Adaptive action fusion method for mobile robot[J]. Journal of Harbin Engineering University, 2005, 26(5): 586-613.

[3] 陈卫东, 席玉庚, 顾东雷. 自主机器人的强化学习研究进展[J]. 机器人, 2001, 23(4): 379-384.
Chen Weidong, Xi Yugeng, Gu Donglei. A survey of reinforcement learning in autonomous mobile robots[J]. Robot, 2001, 23(4): 379-384.

[4] 张汝波. 强化学习理论及应用[M]. 哈尔滨: 哈尔滨工程大学出版社, 2001.

[5] 高阳, 陈世福, 陆鑫. 强化学习研究综述[J]. 自动化学报, 2004, 30(1): 86-100.
Gao Yang, Chen Shifu, Lu Xin. Research on reinforcement learning technology: a review[J]. Acta Automatica Sinica, 2004, 30(1): 86-100.

[6] Mataric M J. Reinforcement learning in the multi-robot domain[J]. Autonomous Robots, 1997, 4(1): 73-83.

[7] Piao Songhao, Hong Bingrong. Fast reinforcement learning approach to cooperative behavior acquisition in multi-agent system[C] // Proceedings of the 2002 IEEE/RSJ. Lausanne, Switzerland, 2002.

[8] Watkins P, Dayan O. Q-learning[J]. Machine Learning, 1992, 8(3): 279-292.