DOI:10.16511/j.cnki.qhdxxb.2008.s2.019

ISSN 1000-0054 清华大学学报(自然科学版)2008年 第48卷 第S2期 11/36
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2008, Vol. 48, No. S2 1747-1750

基于神经网络的强化学习在避障中的应用

乔俊飞， 侯占军， 阮晓钢

(北京工业大学 人工智能与机器人研究所，北京 100124)

摘 要：为了提高移动机器人的自学习能力，在基于行为控制结构的基础上设计了智能控制结构，该结构引入了强化学习模块。神经网络具有很好的泛化能力，该文提出了基于神经网络的强化Q学习算法，克服了表格式Q学习算法只能应用到离散的状态中并需要大量存储空间的不足，最后结合智能控制结构应用到移动机器人的避障中。实验结果表明，该方法能够使移动机器人通过自学习实现自主避障。

关键词：移动机器人；强化学习；神经网络；避障

中图分类号：TP 273 **文献标识码**：A

文章编号：1000-0054(2008)S2-1747-04

Neural network-based reinforcement learning applied to obstacle avoidance

QIAO Junfei, HOU Zhanjun, RUAN Xiaogang

(Institute of Artificial Intelligence and Robots, Beijing University of Technology, Beijing 100124, China)

Abstract: An intelligent control architecture with reinforcement learning was designed based on a behavior-based architecture to improve the learning ability of mobile robots. Normal tabular Q-learning can only be applied to discrete states and requires a large memory. Since neural networks have good generalization, a Q-learning system was developed based on a neural network for obstacle avoidance of mobile robots. Experiments show that the mobile robot can then learn to avoid obstacles.

Key words: mobile robot; reinforcement learning; neural network; obstacle avoidance

移动机器人是当今研究热点，而机器人避障技术是移动机器人研究的一个重要方向。针对移动机器人的避障问题，已经提出了许多方法，如栅格法、模拟势场法、基于滚动窗口的规划方法和基于行为的控制方法等等。基于行为的控制方法是由著名的人工智能专家R. Brooks[1]首先提出的，相对于传统的方法，它具有快速性、鲁棒性好等特点，因此得到了广泛的关注。由于移动机器人运行环境是未知的，靠经验对机器人进行编程实现很难应对可能遇到的各种情况，因此很有必要引入自学习功能使得机器人经过一段时间的训练实现自主避障。基于此，强化学习[2]被广泛应用到基于行为的移动机器人控制当中[3-5]，并被证明是一种通过经验提高智能系统的有效计算工具[6]，其中Q学习[7]是最常用的一种强化学习算法。

为了提高移动机器人的自主性能，使得机器人具有自学习能力，有必要在控制结构中引入强化学习模块，设计一种智能控制结构。另外，由于机器人感知到的环境信息是连续的，以往在应用强化学习的时候需要对环境信息进行离散化，并且以表格的形式存储状态值，因此需要很大的存储空间。神经网络有良好的泛化能力可以实现对任意非线性函数的逼近，因此本文利用神经网络来实现Q函数以便解决强化学习在连续状态中的应用问题。

收稿日期：2008-04-18

基金项目：国家自然科学基金资助项目(60375017)；
北京市优秀人才培养资助项目(2006D0501500203)；
北京市教委科技发展计划项目(KM2006100050190)

作者简介：乔俊飞(1968—)，男(汉)，内蒙古，教授。
E-mail：adqiao@sina.com

1 移动机器人本体结构及其环境检测

1.1 移动机器人本体结构描述

本文所用到的仿真移动机器人是类似于Poineer-3。设定移动机器人为直径为0.5 m的柱体形状,机器人体的前半部分装有6个声纳传感器,它的测量范围为0～1.5 m,主要用来测量周围的环境,以便检测到障碍物的信息。除了传感器可以感知外部环境外,摄像头可以实时捕捉环境信息,通过辨识目标位置实现机器人在避开障碍物的前提下实现自主导航,因此所应用的移动机器人配置了摄像头。移动机器人有2个驱动轮用来控制移动机器人的速度与方向,另外还包括一个平衡轮用来保持机器人的平衡。其结构如图1所示。

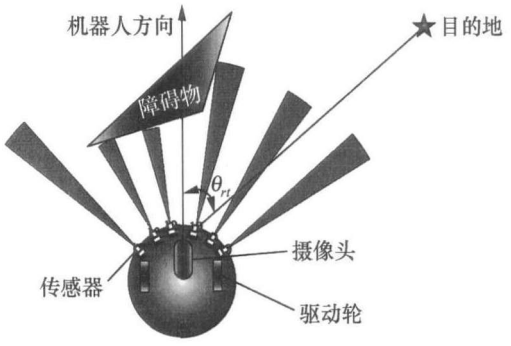

图1 移动机器人结构

1.2 机器人的环境检测

机器人是通过传感器和摄像头来感知外部的环境信息从而做出决策来执行相应的动作。移动机器人从外界得到的信息包括障碍物相对于机器人本身的距离以及所在的方向,这一部分信息通过声纳传感器来获得,另外为了实现机器人的导航,移动机器人需要得到目标点的方位信息,这可以通过安置的摄像头来获得。

2 移动机器人的控制结构设计

基于行为的控制结构起源于R. Brooks的包容式结构。包容式结构的基本思想是把智能体分解成若干完成某个具体任务的基本行为模块,它是一种"由底层到上层"控制结构,不同于传统的"自上向下"的控制结构。由于纯粹的基于行为的控制结构[8]需要设计者根据移动机器人所遇到的各种情况进行考虑并通过人工对各个具体行为进行编程规划,这是一个相当复杂的任务,因此为了提高移动机器人的智能性,引入学习功能显得尤为重要。基于此,设计了一种带有强化学习的控制结构,如图2所示。控制结构包含5个模块:感知模块、行为模块、强化模块、选择模块以及执行模块。

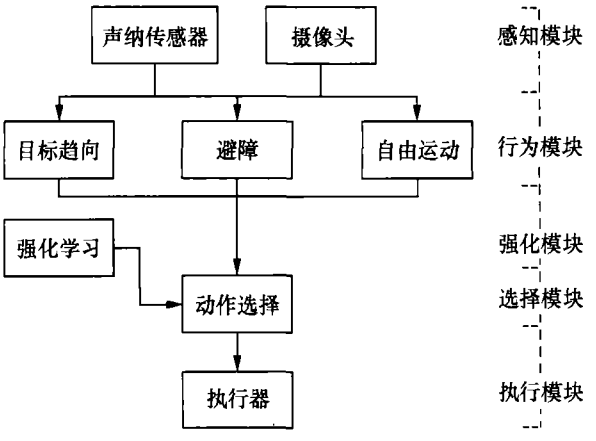

图2 移动机器人的控制结构

3 基于神经网络的强化学习

3.1 强化 Q 学习

强化学习技术是从控制理论、统计学、心理学等相关学科发展而来的。所谓强化学习是指从环境状态到行为映射的学习,以使系统行为从环境中获得累积奖赏值。它不同于监督学习技术那样通过正例、反例来告知采取何种行为,而是通过试错(trial-and-error)的方法来发现最优策略。强化学习系统接受环境状态的输入为 s,根据内部的推理机制,系统输出相应的行为动作为 a。环境在系统动作作用 a 下,变迁到新的状态 s'。系统接受环境新状态的输入,同时得到环境对于系统的顺时奖惩反馈 r。对于强化学习系统来讲,其目标是学习一个行为策略 π: $s\rightarrow a$,使系统选择的动作能够获得环境奖赏的累计值最大。换言之,系统要最大化式(1),其中 γ 为折扣因子。在学习过程中,强化学习技术的基本原理是:如果系统某个动作导致环境正的奖赏,那么系统以后产生这个动作的趋势便会加强,反之系统产生这个动作的趋势便减弱。这和生理学的条件反射原理是接近的。

$$V^{\pi}(s_t)\equiv r_t+\gamma r_{t+1}+\gamma^2 r_{t+2}+\cdots. \tag{1}$$

Q 学习是一类被广泛应用的强化学习,它不直接应用上面的值函数,而是利用一个类似的 Q 函数,其表达式如下:

$$Q(s_t,a_t)\leftarrow r_t+\gamma V(s_{t+1}), \tag{2}$$

式中 a_t 是时刻 t 从动作集 A 被选中的动作。由于系统的目的是使总的奖励值为最大,因此用 $\max\limits_{a\in A}Q(s_{t+1},a)$ 取代式中的 $V(s_{t+1})$。得到表达式

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

在时间 t，智能体根据当前所处的状态选择一个动作 a，然后根据以下的表达式来更新 Q 值

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha[r_t+\gamma\max_{b\in A}Q(s_{t+1},b)-Q(s_t,a_t)], \tag{4}$$

b 为在 $t+1$ 状态所选择的动作。

Q 学习的运行步骤如下。

步骤 1：随机初始化 $Q(s,a)$。

步骤 2：重复一下步骤。

1）获得智能体所处的当前状态；

2）选择一个动作并执行；

3）获得新的状态并同时得到奖励值；

4）根据式(4)更新 Q 值。

3.2 基于神经网络的强化学习

强化学习一般应用到状态信息是离散的情况下，并把相应的状态-动作对值以表格的形式存储到内存当中，这样不但占用大量的内存空间而且学习收敛速度非常慢，状态信息连续将无法实现。基于此，本文提出了基于神经网络的 Q 学习，即用神经网络来逼近 Q 函数，从而可以克服图表存储 Q 值所存在的缺陷。应用一个 3 层的 BP 神经网络，其输入端输入移动机器人所感知的状态信息，输出端输出每个动作所对应的 Q 值。其框图如图 3 所示。

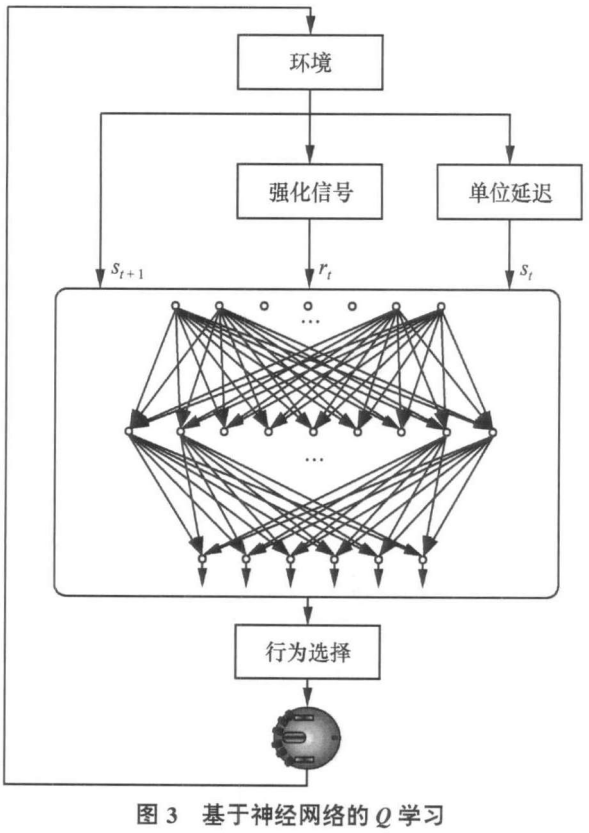

图 3 基于神经网络的 Q 学习

其具体执行步骤如下。

步骤 1：初始化神经网络和运行中所用到的参数。

步骤 2：初始化移动机器人的状态。

步骤 3：获得移动机器人的当前状态信息 s_t。

步骤 4：把状态信息输入到神经网络中，根据获得 Q 值选择动作。

步骤 5：执行动作使得移动机器人达到一个新的状态 s_{t+1}，同时获得反馈强化信号值。若发生碰撞，则返回到移动机器人的初始位置再重新开始。

步骤 6：根据 BP 算法训练神经网络。

步骤 7：重复步骤 3—6 直到学习完毕。

3.3 行为选择策略

在学习的初始阶段，由于其 Q 值是随机初始化的，因此不具有任何意义。为了能够探索到所有可能的动作，本文引入 Boltzmann 分布来实现初始阶段对动作的随机选择，某一个动作被选中的概率为

$$P(a_t|s)=\frac{e^{Q(s,a_t)/T}}{\sum_{a\in A}e^{Q(s,a)/T}}, \tag{5}$$

式中，T 为虚拟问题，随着温度的增加，选择的随机性越强。

随着学习的进行，Q 值慢慢趋向于所期望的状态-动作值，这时候根据贪婪策略来选择动作，即选

择最大 Q 值所对应的动作。

$$a = \arg\max_{b \in A} Q(s,b). \tag{6}$$

4 仿真试验

为了验证方法的可行性,在仿真环境下对其进行验证。首先把移动机器人分解为 7 个基本动作:左转 15°、30°、45°,右转 15°、30°、45°和前行。输入神经网络的信息为每个声纳的距离信息 $\boldsymbol{d}=\{d_i, i=1,2,3,4,5,6\}$以及移动机器人的运动方向和目标方向的夹角 θ_{rt}。

强化信号的获得是 Q 学习的一个关键问题,设定合适的强化信号可以提高学习的收敛速度。根据移动机器人所要完成的任务,把它的强化信号分为 2 部分,一部分是根据移动机器人相对与障碍物所产生的强化值 r_{ro},另一部分是根据移动机器人相对于目标点距离所得的强化值 r_{rt},最后总的强化信号是两者之和。

$$r_{ro} = \begin{cases} -0.5, & \text{靠近障碍物;} \\ +0.5, & \text{远离障碍物;} \\ -1.0, & \text{碰撞障碍物;} \\ 0, & \text{其他情况.} \end{cases}$$

$$r_{rt} = \begin{cases} +0.3, & \text{接近目标点;} \\ -0.3, & \text{远离目标点;} \\ 0, & \text{其他情况.} \end{cases}$$

移动机器人在一个存在随机设置障碍物的环境下运行,初始阶段,由于移动机器人是处于随机选择动作的阶段,因此所运行的路线不平滑且经常碰到障碍物。经过上百次的学习,机器人可以实现在避开障碍物的情况下顺利到达目标点,并且运行轨迹比较平滑,如图 4 所示。随着学习的进行,其运行效果越来越好。以同样的方法,更换了不同的环境,移动机器人都顺利地完成了任务。

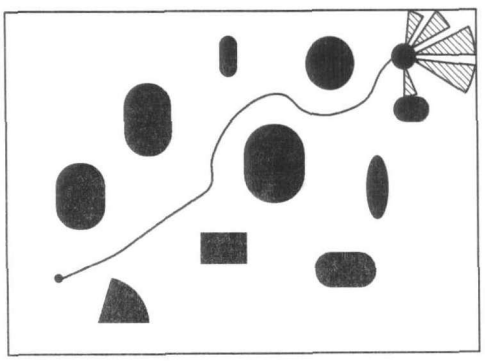

图 4 移动机器人运行轨迹

5 结 论

本文根据基于行为的控制结构提出了一种具有强化学习功能的控制结构,同时把基于神经网络的 Q 学习与其结合起来应用到未知环境下移动机器人的避障中。试验结果证明了其有效性。本文主要把基于神经网络的 Q 学习应用到相对比较简单的存在静态障碍物的情形下,以后的工作是把这种方法应用到较为复杂的移动机器人系统当中,实现更多的功能,例如在存在动态障碍物的环境中实现自主避障。

参考文献 (References)

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

[2] Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey [J]. *Journal of Artificial Intelligence Research*, 1996, **4**(3): 237 - 285.

[3] Mahadevan S, Connell J. Automatic programming of behavior-based robots using reinforcement learning [J]. *Artificial Intelligence*, 1992, **55**(2 - 3): 311 - 365.

[4] Sutton R. Reinforcement learning architectures for animats [C] // Proc of the International Conf Simulation of Adaptive Behavior. Paris, France: MIT Press, 1991: 288 - 296.

[5] Maes P, Brooks R. Learning to coordinate behaviours [C] // Proc Eighth National Conf Artificial Intelligence. Boston. MA: MIT Press, 1990: 796 - 802.

[6] Kumar S. Neural Network: A Classroom Approach, International Edition [M]. New Delhi: Tata McGraw-Hill Publishing Company Limited, 2004.

[7] Watkins C, Dayan P. Q-Learning [J]. *Machine Learning*, 1992, **8**: 279 - 292.

[8] Schoppers M J. Universal plans for reactive robot in unpredictable environments [C]// Proc of the 10th International Joint Conference on Artificial Intelligence. Milan, Italy: Morgan Kaufmann, 1987: 1039 - 1046.