一种基于优势更新的机器人平衡控制算法

史 涛,任红格
(河北联合大学 电气工程学院,河北 唐山 063009)

摘 要:针对自平衡机器人运动平衡控制问题,提出了一种基于优势更新的强化学习机制作为机器人的自平衡仿生学习算法。该算法利用优势更新中的基线,结合强化学习中的概率好奇心机制,以一定的概率选择优等行为,剔除劣等行为,从而使机器人在未知环境下可获得像生物一样的自主学习技能,实现机器人的仿生自平衡运动控制。最后,应用该算法对机器人进行自平衡的仿真实验。结果表明,这种基于优势更新的强化学习算法能使机器人获得较强的平衡控制技能,取得了较好的动态性能,体现了机器人的仿生特性。

关键词:优势更新;强化学习;好奇心;仿生;机器人

中图分类号:TP242.6 **文献标志码**:A **文章编号**:1672-3767(2013)03-0017-05

A Balance Control Algorithm of the Robot Based on Advanced Updating

Shi Tao, Ren Hongge
(College of Electrical Engineering, Hebei United University, Tangshan, Hebei 063009, China)

Abstract: Aiming at the movement balance control problem of the self-balance robot, the reinforcement learning mechanism based on the advanced updating was proposed as a self-balance bionic learning algorithm of the robot. This algorithm can choose optimal behavior with the certain probability by using baseline in the advanced updating and combining the probability curiosity mechanism in the reinforcement learning, and get rid of the inferior behavior, so that the robot can obtain the bionic self-learning skills like creature under the unknown environment, and realize the bionic self-balance control of the robot. Finally, the simulation experiment on robot self-balance control was made by use of the bionic learning algorithm, and the result indicates that the robot can obtain the stronger self-learning control skills and gain the better dynamic performance by applying the reinforcement learning algorithm based on the advanced updating; in addition, the bionic characteristic of the robot is embodied.

Key words: advanced updating; reinforcement learning; curiosity mechanism; bionic; robot

强化学习是建立生物体与机器的一个桥梁,而优势更新恰好体现了强化学习的本质特征。1995 年,Baird 和 Harmon 提出了优势学习算法[1],优势学习是优势更新的一种改进[2],优势学习保留了优势更新的特性,而且只需要学习一个函数,而不是两个函数;1996 年,Baird 等[3]对优势更新做了详细的分析。Dickinson 等[4]对人脑进行了核磁共振成像实验,为优势更新模型对操作条件反射中评价和动作这两个行为过程提供了说明。本研究提出了一种基于优势更新的强化学习机制的仿生学习算法,可以把生物体所具有的一些生命特征用强化模型表达出来,并将这种特征模型运用到机器人上,使机器人表现出生物的某些特性。

强化学习主要研究生物的选择行为,这种行为适于表达情感和探索环境。通过与环境的交互学会某种能力,是仿生学习最重要的特征之一。优势更新是建立强化模型的理论基础,强化学习中的评价机制利用优势更新体现了生物特征,但在行为选择方面,仍然面临着探索和利用这对关系的折中问题。2002 年,伦敦大

收稿日期:2013-01-05
基金项目:国家自然科学基金项目(61203343)
作者简介:史 涛(1980—),男,河北定兴人,讲师,博士研究生,主要从事人工智能及认知发育机器人方面的研究.
E-mail: renhg@heuu.edu.cn

学计算神经科学院的 Dayan 等[5]利用优势更新在强化学习中的作用，突出了生物学习的渐进特征，最终使机器人成功完成了走迷宫实验。2004 年，神经学研究所的 Doherty 和伦敦大学的 Dayan 等联合研究，指出优势更新是强化学习中所用的预测误差信号的基础，通过功能性核磁共振成像技术，揭示了强化学习现象与生物体中部分功能的对应关系[6]，体现了用优势更新建立强化学习模型对机器人仿生学习的重要性。

目前，尚未见资料将优势更新与强化学习机制相结合对两轮式机器人自平衡控制展开研究。本研究提出了一种基于优势更新的强化学习机制作为机器人的仿生自平衡学习算法，使机器人通过与环境的交互，学会运动平衡控制，表现出像人或动物一样的仿生学习行为。

1 机器人系统结构及数学模型

1.1 机器人系统结构

两轮自平衡机器人的动系统采用两个电动机同轴差分驱动的驱动方式，通过增量式光电编码器计算得到车轮的转速信息，并用姿态传感器实时检测系统的姿态信息。机器人控制系统的核心部件是一块放置在机器人车体底部的 TMS320F2812 DSP 处理器，作为系统的控制器，它通过采集姿态传感器的信息，计算得到实时的关于车体姿态的数字信号，按照一定的算法计算出控制量控制电动机的转速和转向，从而驱动机器人前进或后退，以控制其运动平衡。如图 1 所示。

自平衡机器人选择了一个倾角传感器和一个陀螺仪作为检测车体姿态变化的传感器。假设机器人偏离竖直方向的角度为 θ，机身角速度为 $\dot{\theta}$。图 1 中，倾角传感器用来测量角度量 θ，陀螺仪用来测量角速度量 $\dot{\theta}$。

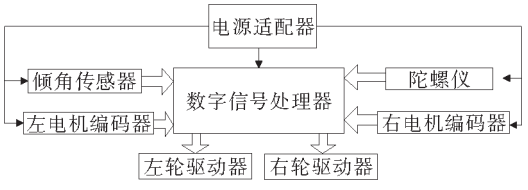

图 1 机器人系统的结构图

Fig. 1 The structure of robot system

1.2 机器人动力学建模

根据 Lagrange 方程建立了自平衡机器人的动力学模型[7]，由此可知，要实现机器人的运动平衡控制需要四个姿态信息，即机器人偏离竖直方向的倾斜角度、角速度以及左右两个轮子的角速度。根据系统的动力学方程选择状态变量为 $\boldsymbol{X}=[\dot{\theta}_l,\dot{\theta}_r,\dot{\theta},\theta]^T$，控制量为 $\boldsymbol{U}=[u_l,u_r]^T$，经计算得到系统的状态方程为：

$$\dot{\boldsymbol{X}}=\boldsymbol{A}\boldsymbol{X}+\boldsymbol{B}\boldsymbol{U}。\tag{1}$$

对上述系统状态方程进行线性化处理，即对于平衡点附近 $|\theta|\leqslant 10°$ 时，令 $\sin\theta=\theta$，$\cos\theta=1$，整理可得到该模型的参数矩阵分别为：

$$\boldsymbol{A}=\begin{bmatrix} 0.2381 & -0.6734 & 0 & -87.3042 \\ -0.6734 & -0.2381 & 0 & -87.3042 \\ 0.9320 & 0.9320 & 0 & 136.6289 \\ 0 & 0 & 1 & 0 \end{bmatrix};\tag{2}$$

$$\boldsymbol{B}=\begin{bmatrix} -0.1327 & 0.9781 \\ 0.9781 & -0.1327 \\ -2.1384 & -2.1384 \\ 0 & 0 \end{bmatrix}。\tag{3}$$

2 基于优势更新的仿生自平衡控制器设计

优势更新能够体现强化学习的本质特征，而强化学习是生物学习的一种形式，它允许智能体根据好奇心机制调整自己的行为，以获得最优行为[8-9]。

为了使机器人在未知环境能像生物一样自主地完成运动平衡控制任务，实现机器人自平衡控制的目的，采用了一种基于优势更新的强化学习机制作为机器人的自平衡仿生学习算法。这种算法可以使机器人经过仿生学习过程，最终掌握运动平衡控制的技能。

2.1 仿生控制器的结构设计

基于优势更新的强化学习机制的机器人自平衡算法，根据强化学习理论，利用优势更新体现生物学习的渐进特征，利用 Boltzmann 机作为好奇心机制进行动作的概率式选取，其控制器的系统结构如图 2 所示。

该仿生自主学习控制器主要由检测装置、控制器和执行器三部分构成，其中：检测装置用来使机器人感知外界环境的变化；控制器根据检测的状态量改进各种控制性能，生成满足要求的控制信号；执行器根据控制信号所发出的动作量执行相应的动作，直至机器人学会保持控制运动平衡的状态。

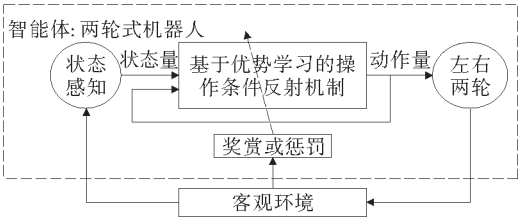

图 2 控制器系统结构框图

Fig. 2 Structure diagram of controller system

2.2 仿生算法的设计

2.2.1 概率式学习的好奇心机制设计

优势更新[6]为强化学习中的两个过程提供了说明：第一部分是评价，用一个时间差分预测误差信号来更新与外部状态和内部环境（由刺激的排列所决定）相关的未来奖赏的相继预测；另一部分是行为，以策略的形式修正刺激—响应或刺激—响应—回报联想，与最大奖赏相关的行为在随后的实验中被更加频繁地选择。优势量用来评估行为，行为完全根据其优势进行选择，而选择依赖于好奇心机制。

好奇心是生物进化过程中形成的适应环境和生存的习性，对于自平衡机器人，其主要的好奇心就是直立、平衡和稳定。好奇心机制中的迁移概率是依据 Boltzmann 机的整合激发概率推导出来的，它能够使系统的状态服从 Boltzmann-Gibbs 分布，并最终趋于热平衡状态[10]，即

$$P_{\mathrm{TR}}(s\rightarrow s')=\begin{cases}1, & \Delta\xi<0\\ \exp[-\Delta\xi/T], & \Delta\xi\geqslant 0\end{cases}。\tag{4}$$

其中：P_{TR}是状态迁移概率；s 是当前状态，s'是随机生成的状态；$\Delta\xi=\xi(s')-\xi(s)$是状态迁移造成的能量变化。$\Delta\xi\geqslant 0$ 时，依 Boltzmann-Gibbs 分布单调递减地接受随机生成的状态 s'；$\Delta\xi<0$ 时，以概率 1 无条件接受随机生成的状态 s'。这表明好奇心机制中的迁移概率随着温度 T 的减小，选择次最优动作的概率会越来越小，相对来说，得到强化的动作的选择概率就会增加。这与强化学习的基本原理一致，即，如果产生的后果是正强化信号，那么下次出现该行为的概率会增加；反之，如果产生的后果是负强化信号，那么下次出现该行为的概率会减少。而 Boltzmann 机作为好奇心机制，具备了强化学习理论中动作选择的特征。

2.2.2 基于优势更新的仿生学习算法的设计

基于优势更新的强化学习机制的仿生学习算法是根据被称为优势的量来评估行为，行为优势值大的会优先考虑。如果行为优势值为负值，则会被遗弃。当生物做得更好时，在该状态下选择此行为的频率会更高。通过动作选择变得更好的过程是一种策略改进的形式，因此，采用 Boltmann 机作为机器人动作概率选择的好奇心机制。

当某个状态的行为选择发生变化时，状态的值和该状态所有行为的优势也变化了。行为执行的相同预测误差信号的值实际上就是该行为优势的目标。因此，优势的误差是基于在状态 b 处选择行为后所出现的 TD(temporal difference)预测误差 δ。优势误差信号可以被用来改变权值，它依赖于评价函数 J：

$$J(t)=r(t+1)+\gamma\cdot r(t+2)+\gamma^2\cdot r(t+3)+\cdots\tag{5}$$

其中：r 指回报信号，是对所选动作的作用效果的一种评价。

机器人处于平衡态指的是它的各状态量满足：摆杆倾角 $\theta<0.0523$ rad，且机器人摆杆角速度 $\dot{\theta}$、左右轮角速度 $\dot{\theta}_l$ 和 $\dot{\theta}_r$ 均小于 3.489 rad/s，折扣因子 $\gamma=0.9$ 表示近期回报预测和远期回报预测的重要程度，动作评价函数的目的是使长期的评价函数值 $J(t)$最大，即表明该动作所产生的效果最优，采样时间为 0.02 s。

3 仿真实验与分析

为了验证基于优势更新的机器人仿生自平衡算法的有效性，对机器人做了运动平衡控制的自学习仿真实验，并给出了相应的仿真结果。

在仿真实验中[11]，随机给机器人一个初始值，使之开始试探性学习，结果表明，机器人在理想情况下经过 172 次试探，在第 173 次试探过程中运行了 2000 步，最终通过不断学习，学会了运动平衡控制的技能，其最终状态曲线如图 3 所示。

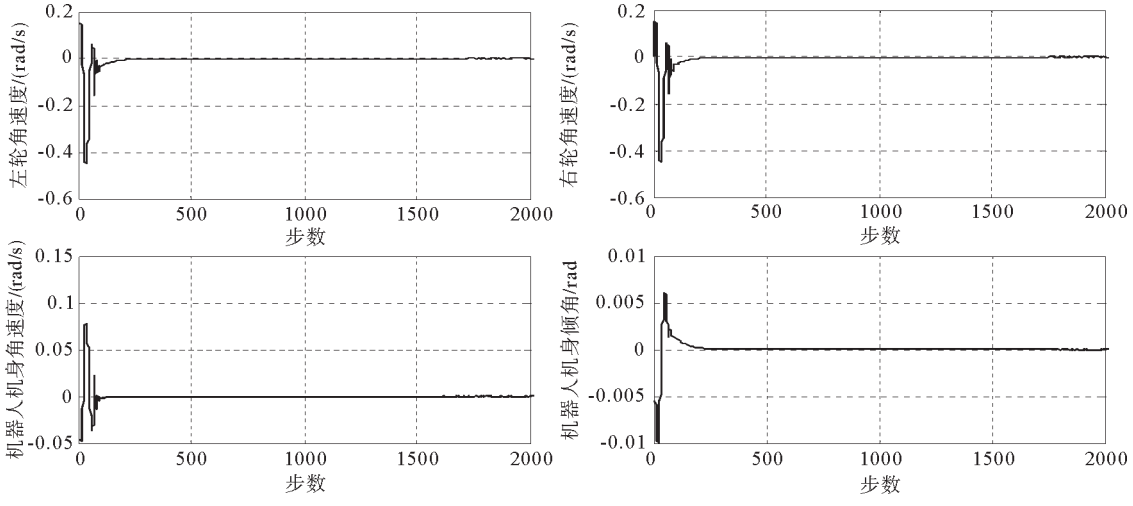

图 3　机器人各状态量的变化曲线图

Fig. 3　Variation curves of robot's state variable

外界干扰在机器人运动过程中有着非常严重的影响，这些干扰不仅会改变机器人的运动特性，还会威胁到机器人的平衡和稳定，因此，抗干扰能力是两轮自平衡机器人的一个重要性能指标。为了验证该算法的抗干扰能力，在机器人平衡运动 10 000 步，即 100 s 时，加入一个幅值为 20 的脉冲扰动，机器人响应干扰时各状态量变化曲线如图 4 所示。由仿真曲线可知，在施加脉冲干扰后，机器人大约经历 300 步，即 3 s 的时间就可以恢复到平衡状态。表明在有先验经验模型的基础上进行再学习，它的学习能力和速度会更快，同时也体现了该学习算法具有良好的抗干扰性。

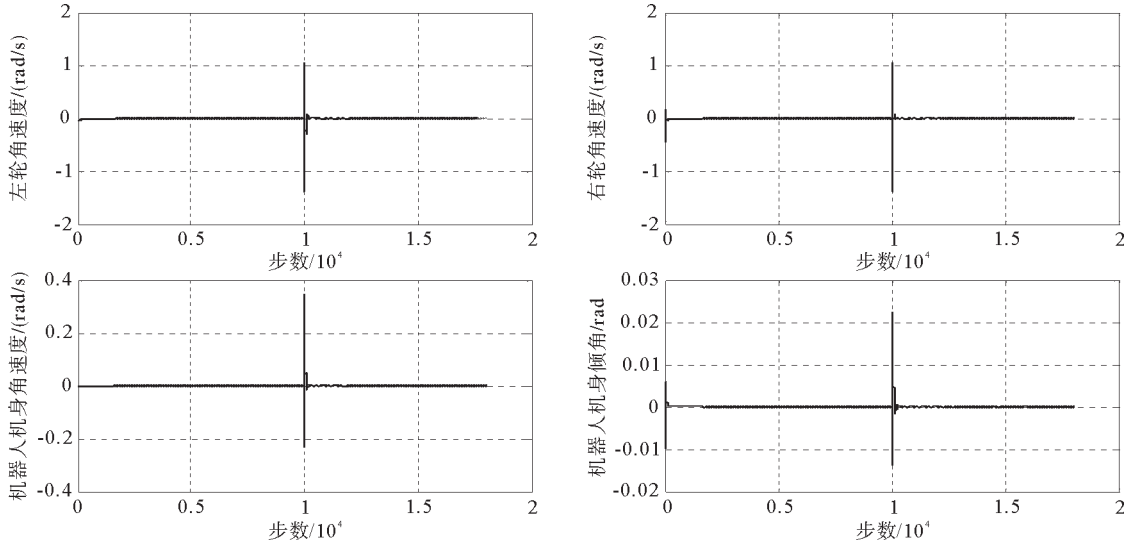

图 4　干扰动实验中机器人状态量仿真曲线图

Fig. 4　Simulation curves of the robot state quantities under the disturbance experiment

4 结束语

根据仿生学特征，提出了基于优势更新的机器人仿生自平衡算法，通过与环境的交互，使机器人学会运动平衡控制的技能。强化学习研究生物的选择行为，这种行为是适于表达情感和探索环境的。强化学习允许生物体通过学习，认识到自身响应和奖赏或者惩罚结果之间的代价费用。而优势更新的过程体现了生物特征，再结合 Boltzmann 策略的好奇心机制进行动作选择，这就实现了完整的强化学习，即把生物的行为过程用更加明确的模型表示出来，并应用在机器人上，使机器人具有生物所特有的某些智能化特征，机器人能在未知环境下，经过自主学习和训练获得类似生物一样的运动平衡控制技能，较好地满足了预期控制目标，体现了机器人的仿生自主学习能力。

参考文献：

[1]Harmon M E, Baird L C, Klopf A H. Advantage updating applied to a differential game[M]//Advances in Neural Information Processing Systems. Cambridge MA: MIT Press, 1995.

[2]Baird L C. Advantage learning[R]. Technical Report WL-TR-93-1146, Wright-Patterson Air Force Base, Dayton, OH, 1993: 13-27.

[3]Harmon M E, Baird III L C. Multi-player residual advantage learning with general function approximation[R/OL]. Technical Report WL-TR-1065, Wright Laboratory, WL/AACF, Wright-Patterson Air Force Base, OH, 1996: 1-13. [2012-12-05] http://www.leemon.com/papers/sim_tech/sim_tech.pdf.

[4]Dickinson A, Smith J, Mirenowicz J. Dissociation of pavlovian and instrumental incentive learning under dopamine antagonists[J]. Behavioral Neuroscience, 2000, 114: 468-483.

[5]Dayan P, Balleine B W. Reward, motivation, and reinforcement learning: Review[J]. Neuron, 2002, 36(2): 285-298.

[6]John O D, Peter D, Johannes S, et al. Dissociable roles of ventral and dorsal striatum in instrumental conditioning[J]. Science, 2004, 304(5669): 452-454.

[7]张晓华. 系统建模与仿真[M]. 北京：清华大学出版社，2006：224-232.

[8]Touretzky D S, Saksida L M. Operant conditioning in skinnerbots[J]. Adaptive Behavior, 1997, 5(3-4): 219-247.

[9]徐昕. 增强学习及其在移动机器人导航与控制中的应用研究[D]. 长沙：国防科学技术大学，2002：10.

[10]阮晓钢. 神经计算科学：在细胞的水平上模拟脑功能[M]. 北京：国防工业出版社，2006：553-596.

[11]任红格，史涛，张瑞成. 基于操作条件反射机制的感觉运动系统认知模型的建立[J]. 机器人，2012，34(3)：292-298.
Ren Hongge, Shi Tao, Zhang Ruicheng. Foundation of the sensorimotor system cognitive model with operant conditioning mechanism[J]. Robot, 2012, 34(3): 292-298.

（责任编辑：吕文红）

（上接第 16 页）

[39]Guo Y, Woo P. An adaptive fuzzy sliding mode controller for robotic manipulators[J]. IEEE Transaction on Systems, Man, and Cybernetics, 2003, 33(2): 149-159.

[40]Mu X J, Chen Y Z. Neural sliding mode control for multi-link robots[C]//Chinese Control and Decision Conference. Yantai, July 2-4, 2008: 3513-3517.

[41]Lin C M, Chen L Y, Chen C H. RCMAC hybrid control for MIMO uncertain nonlinear systems using sliding-mode technology[J]. IEEE Transactions on Neural Networks, 2007, 18(3): 708-720.

[42]Tao G, Kokotovic P V. Adaptive control of plants with unknown dead-zones[J]. IEEE Transactions on Automatic Control, 1994, 39(1): 59-68.

[43]Selmic R R, Lewis F L. Deadzone compensation in motion control systems using neural networks[J]. IEEE Transactions on Automatic Control, 2000, 45(4): 602-614.

（责任编辑：吕文红）