第 27 卷第 10 期
2010 年 10 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 27 No. 10
Oct. 2010

文章编号: 1000−8152(2010)10−1423−06

基于Skinner操作条件反射的两轮机器人自平衡控制

任红格, 阮晓钢

(北京工业大学 电子信息与控制工程学院, 北京 100124)

摘要: 针对两轮自平衡机器人的运动平衡控制问题, 采用了基于Skinner操作条件反射理论的自回归神经网络学习算法作为机器人的学习机制, 利用自回归神经网络对评价函数进行逼近, 以实现对行为决策的优化, 从而使机器人能够在无需外部环境模型的情况下, 通过学习和训练, 获得像人或动物一样的自主学习技能, 解决了两轮机器人的运动平衡控制问题. 最后分别在无扰动和有扰动的两种状态下设计了仿真实验并进行了比较. 结果表明, 该操作条件反射学习机制具有较快的自主平衡控制技能和较好的鲁棒性能, 体现了较高的理论研究意义和工程应用价值.

关键词: Skinner操作条件反射; 自回归神经网络; 两轮机器人; 自平衡控制; 鲁棒性

中图分类号: TP391　　**文献标识码:** A

Self-balance control of two-wheeled robot based on Skinner's operant conditioned reflex

REN Hong-ge, RUAN Xiao-gang

(School of Electronic & Control Engineering, Beijing University of Technology, Beijing, 100124, China)

Abstract: For the movement-balance control of a two-wheeled self-balancing robot, we adopt the autoregression neural-network-learning-algorithm based on Skinner's operant conditioned reflex theory as the learning mechanism, and use the autoregression neural-network to approximate the critic function in the optimization of behavioral decision, so that a two-wheeled robot can obtain self-learning skills like a human being or an animal through studying and training in a model-free external environment to realize the movement balance control. Two simulation experiments are separately performed in the states with and without disturbance, respectively. The comparison of the respective results shows that learning mechanism with Skinner's operant conditioned reflex has a faster control skill in self-balance and a high robustness. This exhibits great research significance in theory and practice.

Key words: Skinner's operant conditioned reflex; autoregression neural-network; two-wheeled robot; self-balance control; robustness

1 引言(Introduction)

感觉运动系统是一个综合了感受器功能和运动神经机能的神经生理组织, 是一个由感觉到运动的系统. 感觉运动系统的基本功能是“感知–行动”, 它的基本神经活动和学习机制就是条件反射. 关于神经系统, 存在两种重要的条件反射理论, 即Pavlov的经典条件反射理论和Skinner的操作条件反射理论.

人或动物的运动平衡控制技能来自于小脑感觉运动系统, 而Skinner操作条件反射是感觉运动系统最为基本的和重要的学习机制. Skinner操作条件反射是由美国哈佛大学教授斯金纳(B.F. Skinner)利用斯金纳箱而建立的[1], 它的基本原理是: 如果一个操作发生后, 接着呈现一个强化刺激, 则这个操作的强度(反应发生的概率)就会增大, 反之则减小[2]. 关于操作条件反射, 人们已经做了大量的研究. 1988年, 美国California大学的Rose等人利用Skinner操作条件反射理论, 通过奖赏和惩罚训练, 使机器人做了一些预先指定的行为[3]; 1997年, 美国Boston大学神经机器人学实验室的P. Gaudiano 等人把Skinner操作条件反射理论应用到真实机器人Pioneer 1和Khepera上, 实现了机器人自适应避障的功能[4]; 2005年, 日本早稻田大学的Itoh等人, 利用了基于Skinner操作条件反射原理的Hull行为理论, 使机器人WE-4RII学会了用右手和人握手的行为[5].

但目前还没有人采用Skinner操作条件反射理论对两轮机器人自平衡运动控制问题展开研究, 因此, 将Skinner操作条件反射理论应用于机器人系统, 是文章的动机和出发点. 针对两轮自平衡机器人的运

收稿日期: 2008−12−03; 收修改稿日期: 2010−03−02.

基金项目: 国家“863”计划资助项目(2007AA04Z226); 国家自然科学基金资助项目(60774077); 北京市教委重点资助项目(KZ200810005002); 北京市人才强教计划资助项目; 高等学校博士学科点专项科研基金资助课题.

动平衡控制问题, 复制人或动物小脑感觉运动系统的组织和结构, 模拟这种组织和结构中的Skinner操作条件反射机能, 为机器人设计人工小脑感觉运动系统, 从而使机器人能够像人或动物一样, 通过学习和训练, 自组织地渐进形成、发展和完善其运动平衡控制技能.

2 两轮机器人系统结构及动力学模型(Structure feature and dynamic model of the two-wheeled robot)

2.1 两轮自平衡机器人的结构特点(Structure feature of the two-wheeled robot)

两轮自平衡机器人系统是一个高阶次、不稳定、多变量、非线性、强耦合的系统, 它实际上是一个可以行走的一级倒立摆, 它以双轮差速方式布置, 每个轮子由直流电机通过减速器直接驱动, 以电机轴心线为中心前后转动. 如图1所示. 对于两轮机器人来说, 在静止状态下不能稳定平衡, 若要其稳定必须采用动态平衡[6], 机器人的平衡是一个动态过程, 机器人在平衡点附近不停的变化进行调节以保持平衡.

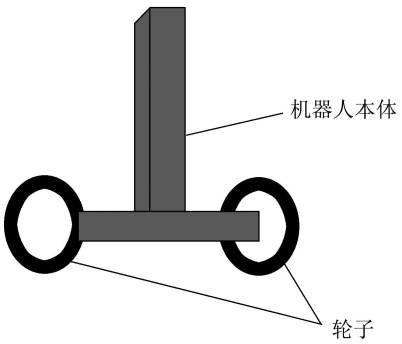


图 1 机器人系统的结构图

Fig. 1 The structure drawing of the robot

2.2 两轮自平衡机器人的动力学模型(dynamic model of the two-wheeled robot)

采用Lagrange方法对两轮自平衡机器人进行系统动力学建模, 其数学表达式如式(1)所示:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial T}{\partial \dot{q}_k}\right)-\left(\frac{\partial T}{\partial q_k}\right)=Q_k,\ k=1,2\cdots. \tag{1}$$

式中: T为系统的总动能, q_k为系统的广义坐标, Q_k为广义力, 其中: 系统的3个广义坐标q_k为左轮角度θ_{l}、右轮角度θ_{r}和摆杆角度θ, 广义坐标下的系统广义力Q_k为左轮转矩$Q_{\theta_{\mathrm{l}}}$、右轮转矩$Q_{\theta_{\mathrm{r}}}$和车体作用在x轴的转矩Q_θ, 根据以上系统的动力学方程可得到多输入多输出非线性动力学模型. 选择系统动力学方程选择的状态变量为$X=(\dot{\theta}_{\mathrm{l}},\dot{\theta}_{\mathrm{r}},\dot{\theta},\theta)^{\mathrm{T}}$, 分别表示机器人左右两轮角速度, 机器人摆杆角速度和角度. 控制量$u=(u_{\mathrm{l}},u_{\mathrm{r}})^{\mathrm{T}}$, 分别表示加在左右车轮电机上的电压. 经过线性化处理, 在平衡点附近, 即$|\theta\leqslant 5\,^\circ|$处, 令$\sin\theta=\theta$, $\cos\theta=1$, 可得到系统的状态方程为(2)和(3)所示. 其中, 两轮自平衡机器人动力学模型的状态方程中的参数是根据欧鹏公司生产的两轮自平衡机器人的实际模型测量和计算得到的:

$$\begin{pmatrix}\ddot{\theta}_{\mathrm{l}}\\ \ddot{\theta}_{\mathrm{r}}\\ \ddot{\theta}\\ \dot{\theta}\end{pmatrix}=\begin{pmatrix}0.0195 & -0.2293 & 0 & -65.2850\\ -0.2293 & -0.0195 & 0 & -65.2850\\ 0.3450 & 0.3450 & 0 & 244.8912\\ 0 & 0 & 1 & 0\end{pmatrix}\begin{pmatrix}\dot{\theta}_{\mathrm{l}}\\ \dot{\theta}_{\mathrm{r}}\\ \dot{\theta}\\ \theta\end{pmatrix}+\begin{pmatrix}-0.0588 & 0.6900\\ 0.6900 & -0.0588\\ -1.0381 & -1.0381\\ 0 & 0\end{pmatrix}\begin{pmatrix}u_{\mathrm{l}}\\ u_{\mathrm{r}}\end{pmatrix}, \tag{2}$$

$$y=\begin{pmatrix}-0.075 & 0 & 0 & 0\\ 0 & -0.075 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1\end{pmatrix}\begin{pmatrix}\dot{\theta}_{\mathrm{l}}\\ \dot{\theta}_{\mathrm{r}}\\ \dot{\theta}\\ \theta\end{pmatrix}+\begin{pmatrix}0 & 0\\ 0 & 0\\ 0 & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}u_{\mathrm{l}}\\ u_{\mathrm{r}}\end{pmatrix}. \tag{3}$$

3 Skinner操作条件反射控制器的设计(Controller designing of Skinner's operant conditioned reflex)

Skinner操作条件反射理论是关于人或动物学习操作技能的理论, 它允许智能体去调整自己的行为, 以便从环境中获得最大回报, 这对于机器人运动控制, 特别是运动平衡控制, 有重要的指导意义.

针对两轮机器人, 采用了一种新型的基于Skinner操作条件反射理论的自回归神经网络学习机制, 利用这种学习机制, 可以使机器人在未知环境中通过自主学习, 最终获得像人或动物一样的自主运动平衡控制技能.

3.1 自回归神经网络(autoregression neural network)

神经网络由于具有很强的非线性映射能力、并行处理能力和自适应、自学习能力, 被广泛应用于非线性系统的自适应控制中[7,8]. 为了克服神经网络收敛速度慢、不能保证收敛到全局最小点, 以及学习、记忆不稳定性等缺陷, 本文提出了一种基于Skinner操作条件反射理论的自回归神经网络学习

机制.

在自回归神经网络中, 隐层节点不仅接收来自输入层的输出信号, 还接收隐层节点自身的一步延时输出信号, 使网络在原来的学习模式中连同新加入的新学习模式一起进行训练. 采用内部回归网络, 可以增强网络本身处理动态信息的能力, 使其更适合复杂系统的稳定控制. 这种学习模式与人类大脑的学习模式相似, 新信息的记忆不会影响已记忆的信息, 这就是人类大脑记忆的稳定性[9]. 其网络结构图如图2所示.

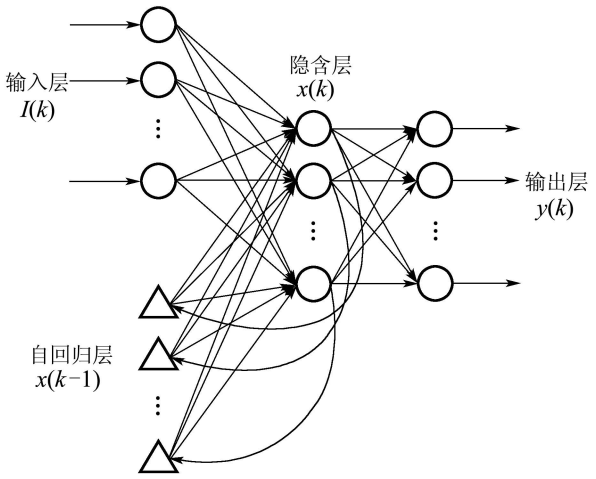


图 2　自回归神经网络结构图

Fig. 2　The structure chart of the autoregression neural network

3.2　基于Skinner操作条件反射理论的自回归神经网络控制器的设计(Designing of the autoregression neural network controller based on Skinner's operant conditioned reflex)

根据Skinner操作条件反射理论的基本原理, 设计了一种新型的基于Skinner操作条件反射理论的自回归神经网络学习机制, 即在不需要外部环境数学模型的情况下, 把控制系统的性能指标要求直接转换为一种评价指标, 当系统性能指标满足一定误差要求时, 所施加的控制动作得到奖励, 否则给以惩罚. 控制器通过奖罚学习, 使评价函数(未来奖赏的累积和)最大, 以获得对系统的最优控制动作[10,11], 从而, 使人或动物学会某种行为或控制某种行为的发生. 其系统控制结构图如图3所示.

基于Skinner操作条件反射理论的自回归神经网络学习机制由两部分构成: 评价神经网络(CNN)和行为神经网络(ANN), 并且, 这两种网络都采用了内部回归神经网络. 评价神经网采用时间差分方法(temporal difference, 简称为TD方法)对评价函数进行逼近, 将状态映射为期望的评价值, CNN利用直接从环境中获取的评价性反馈信号, 并积累反馈信号未来值的加权, 为动作神经网络提供一个更具信息量的评价函数来评估当前动作的好坏[12], 并且Watkins等人[13]已经证明了评价函数可收敛到最优解; 行为神经网络ANN利用评价神经网络CNN的输出来实现行为决策的优化, 从而使获得正强化的行为更容易再次发生, 即增大了这种行为发生的概率. 这正是操作条件反射学习机制基本原理的体现.

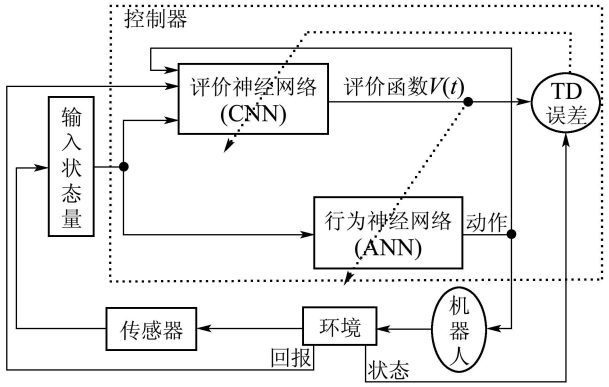


图 3　基于Skinner操作条件反射理论的自回归神经网络学习系统结构图

Fig. 3　The structure chat of the autoregression neural network leaning system based on Skinner's operant conditioned reflex

3.3　收敛性证明(Convergence proof)

以动作评价函数$V(t)$进行的Skinner操作条件反射学习机制的收敛性和稳定性证明如下所示:

证　定义动作评价函数$V(t)$为

$$V(t)=r(t+1)+\gamma\cdot r(t+2)+\gamma^2\cdot r(t+3)+\cdots,$$

其中回报函数$r(t)$定义为

$$r(t)=\begin{cases}0, & X\leqslant \text{scales},\\ -1, & \text{其他}.\end{cases}$$

即, 当状态变量X在直立平衡范围内时给与奖赏, 否则给与惩罚, 并且, 折扣因子γ一般取值为$0<\gamma<1$, 因此, $r(t)\leqslant 0$, 且$0<\gamma<1$, 则, $V(t)\leqslant 0$.

当$t\to\infty$, 机器人学会了自主平衡控制, 能够选择出最优动作, 得到最大奖赏, 即: $r(t)=0$; 从而使累积折扣奖赏最大化, 即得到最优动作评价函数$V^*(t)$, 使得$V^*(t)=0$. 即, 系统趋于平衡稳定状态.

故以动作评价函数$V(t)$进行的Skinner操作条件反射学习机制在$t\to\infty$时是收敛的, 同时, 系统处于平衡稳定状态, 且当且仅当$r(t)=0$时, 存在最优动作评价函数$V^*(t)=0$.　　证毕.

4 仿真实验设计及结果分析(Simulation experiment design and result analysis)

4.1 实验设计(Experiment design)

以两轮机器人在未知环境中通过自主学习达到运动平衡作为控制目标, 在Skinner操作条件反射学习机制中, 评价神经网络CNN采用3层神经网络$N^3[6,8,1]$的结构, 输入为两轮机器人的4个状态量和行为神经网络ANN的2个输出, 即左右车轮电机上的电压$u_{\rm l}$和$u_{\rm r}$, CNN的输出是评价函数$V(t)$, 可表示为

$$V(t)=r(t+1)+\gamma\cdot r(t+2)+\gamma^2\cdot r(t+3)+\cdots .$$

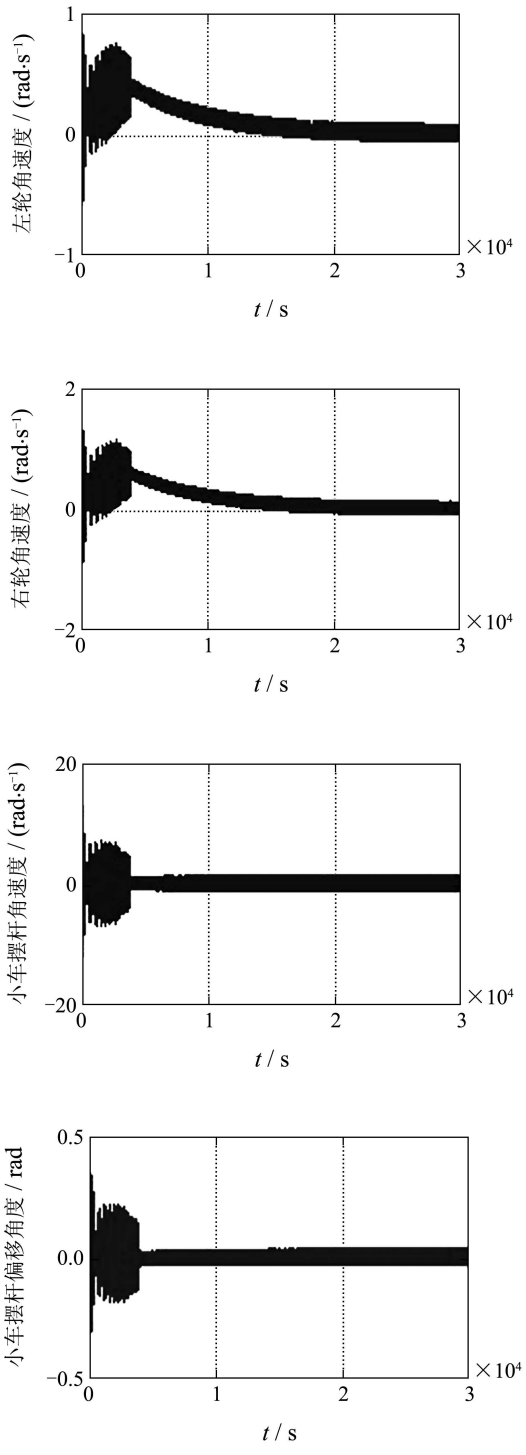
其中: r为回报, 当机器人状态量满足摆杆倾角$\theta<0.0523$ rad, 且机器人摆杆角速度$\dot{\theta}$、左右轮角速度$\dot{\theta}_{\rm l}$和$\dot{\theta}_{\rm r}$均小于3.489 rad/s时, 给机器人一个奖赏信号, 即$r=0$, 否则, $r=-1$. 选取折扣因子$\gamma=0.9$(一般取$0<\gamma<1$), 采样时间$T=0.01s$. 行为神经网络ANN采用3层神经网络$N^3[4,8,2]$的结构, 输入为机器人的4个状态量, 输出为机器人左右车轮电机上的电压$u_{\rm l}$和$u_{\rm r}$, 即选择使评价函数$V(t)$的值最大的动作. 该学习机制是在线学习过程, 且环境未知, 但机器人当前的状态量是可获取的.

CNN和ANN的网络权系数的初始值在[–0.3,0.3]中随机选取, 机器人状态变量的初始值取一定范围内的随机值, CNN和ANN同时在线更新. 每次实验当机器人的试探次数(失败次数)超过100次或一次试探的平衡步数超过30000步, 则中止机器人的学习并重新开始另一次实验. 如果机器人在其中一次试探中能保持30000步不倒, 则认为机器人在未知环境中已经学会控制自身平衡了. 每次平衡失败后, 将初始状态及权值复位为一定范围内的随机值, 重新学习[14]. 总结多次实验结果, 机器人平均经过71次失败学习后就可以自主控制其运动平衡了, 表现了其较快的自主学习能力. 仿真结果如图4所示.

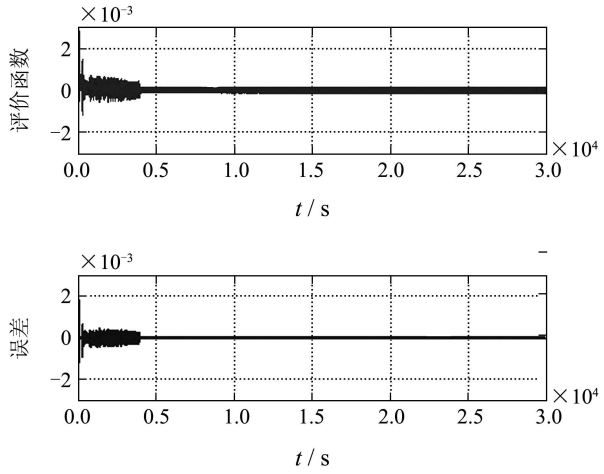
但在实际环境中, 传感器检测到的两轮机器人状态会受到外界的干扰或传感器本身的不精确, 都会使传输的状态量产生一定的测量误差, 为了模拟真实环境, 在机器人已经学会自主控制系统平衡后的任意一段时间(保持系统20000次不倒时)内, 向输入的状态量中加入幅值为4的冲激函数, 此时, 仿真结果表明系统仍能保持平衡, 如图5所示.

4.2 结果分析(Result analysis)

图4(a)表示机器人各状态量随步数增加的变化曲线, 表明了机器人经过约4000步的自主学习, 学会了控制其运动平衡, 使机器人左右两轮在平衡位置附近不停的变化进行调节, 从而保持机器人不倒. 图4(b)表示机器人系统的评价函数曲线以及估计评价函数与前一时刻实际评价函数的误差曲线, 评价函数曲线表明动作神经网络ANN所选择的行为可以使机器人获得的累计回报(正强化)r最大, 其值接近于零. 误差曲线表明评价函数的时间差分误差最小, 接近于零, 即评价神经网络CNN可以近似逼近评价函数$V(t)$.



(a) 机器人状态量变化曲线



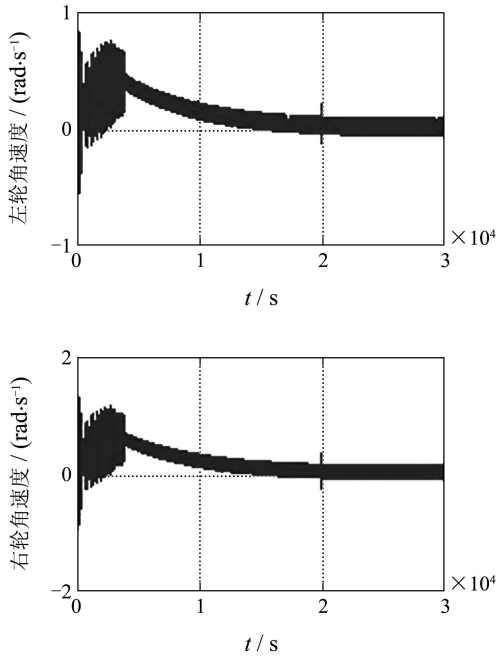
(b) 评价函数及其误差变化曲线

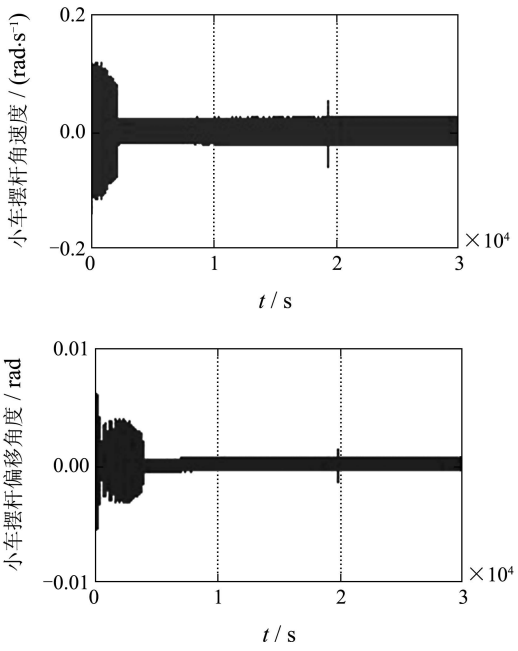
图 4　有扰动时机器人仿真曲线

Fig. 4　The simulation curve of the robot with the perturbation

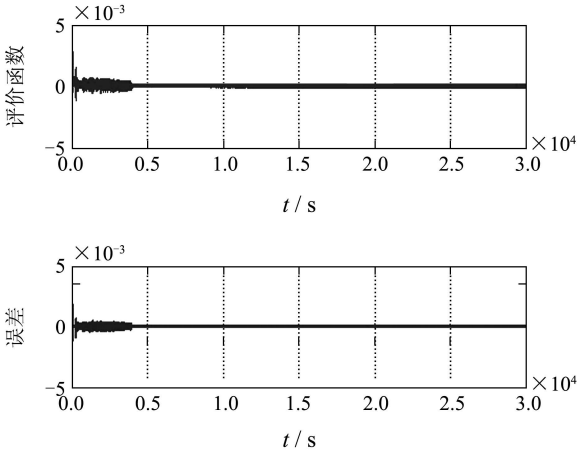
图5显示了向输入信号中加入扰动信号之后的机器人状态量的变化曲线和评价函数及其误差变化曲线, 仿真结果表明机器人能很快的回到平衡位置, 体现了由自回归神经网络建立的Skinner操作条件反射学习系统具有较强的抗扰能力.

由图4和图5可知, 该学习算法在没有扰动的情况下, 可以使机器人在71次失败之后, 最终以大约40 s的学习时间学会了控制其自主平衡, 表现了该算法较快的自主学习能力; 当受到外界扰动时, 机器人能很快的恢复到平衡位置, 表现了该算法具有较好的鲁棒性能, 总结这两个实验可推出, 该操作条件反射学习机制能够使机器人具有较强的自主学习能力.





(a) 加扰动后的机器人状态量变化曲线



(b) 加扰动后的评价函数及其误差变化曲线

图 5　有扰动时机器人仿真曲线

Fig. 5　The simulation curve of the robot with the perturbation

5　结论(Conclusion)

根据动物能够通过基于认知心理学原理的操作条件反射的训练快速学习新行为的这一现象, 提出了基于Skinner操作条件反射理论的自回归神经网络学习机制, 并把这一思想应用到两轮机器人上, 通过网络的记忆和调整, 使机器人经过学习和训练, 最终能在未知环境下, 获得像人或动物一样自组织的渐进形成、发展和完善其运动平衡控制技能. 仿真实验表明, 基于Skinner操作条件反射理论的自回归神经网络学习机制, 能成功的实现两轮机器人的自主平衡控制, 保持其姿态在原点附近平衡不倒, 满足预期控制目标, 具有较强的自主学习能力和鲁棒性能, 并有较高的理论和应用研究价值.

参考文献(References):

[1] 林永惠. 经典性条件反射同操作性条件反射的异同[J]. 渤海学刊, 1997, 1(17): 73 – 76.
(LIN Yonghui. Classical conditioned reflex of interoperability with the conditioned reflex of the similarities and differences[J]. *Bohai Journal*, 1997, 1(17): 73 – 76.)

[2] SKINNER B F. *The Behavior of Organisms*[M]. New York, USA: Copley Publishing Group, 1938.

[3] ROSEN B E, GOODWIN J M, VIDAL J J. Machine operant conditioning[C] //*Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. New Orleans, LA, USA: IEEE, 1988, 3: 1500 – 1501.

[4] GAUDIANO P, CHANG C. Adaptive obstacle avoidance with a neural network for operant conditioning: experiments with real robots[C] //*Proceedings of 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97*. Monterey, CA, USA: IEEE Computer Society Press, 1997, 1: 13 – 18.

[5] ITOH K, MIWA H, MATSUMOTO M, et al, Behavior model of humanoid robots based on operant conditioning[C] //*Proceedings of 2005 5th IEEE-RAS International Conference on Humanoid Robots*. Tsukuba, Japan: Institute of Electrical and Electronics Engineers Computer Society Press, 2005, 1: 220 – 225.

[6] 屠运武, 徐俊艳, 张培仁, 等. 自平衡控制系统的建模与仿真[J]. 系统仿真学报, 2004, 16(4): 830 – 841.
(TU Yunwu, XU Junyan, ZHANG Peiren, et al. Model and simulation of self-balance control system[J]. *Journal of System Simulation*, 2004, 16(4): 830 – 841.)

[7] 叶其革, 王晨皓, 吴捷. 基于自组织模糊神经网络电力系统稳定器的设计[J]. 控制理论与应用, 1999, 16(5): 688 – 695.
(YE Qige, WANG Chenhao, WU Jie. Design of self-organizing power system stabilizer based on fuzzy neural network[J]. *Control Theory & Applications*, 1999, 16(5): 688 – 695.)

[8] NARENDRA K S, PARTHASARATHY K. Identification and control of dynamic systems using neural networks[J]. *IEEE Transactions on Neural Networks*, 1990, 1(1): 4 – 27.

[9] 闻新, 周露, 王丹力, 等. MATLAB神经网络应用设计[M]. 北京: 科学出版社, 2001: 107 – 117.
(WEN Xin, ZHOU Lu, WANG Danli, et al. *MATLAB Neural Network Applications*[M]. Beijing: Science Press, 2001: 107 – 117.)

[10] JAGANNA T S, LEWIS F L. Multilayer discrete-time neural-net controller with guaranteed performance[J]. *IEEE Transactions on Neural Network*, 1996, 7(1): 107 – 130.

[11] ZOMAYA Y. Reinforcement learning for the adaptive control of non2 linear systems[J]. *IEEE Transactions on Systems, Man, and Cybernetics*, 1994, 24(2): 357 – 363.

[12] WANG R X. *Reinforcement learning based on the inverted pendulum control*[D]. Beijing: UBeijing University of Technology, 2005.

[13] WATKINS C J C H, DAYAN P. Q-learning[J]. *Machine Learning*, 1992, 8(3): 279 – 292.

[14] 王瑞霞, 孙亮, 阮晓钢. 基于内部回归神经网络的强化学习[J]. 控制工程, 2005, 12(2): 138 – 140.
(WANG Ruixia, SUN Liang, RUAN Xiaogang. Recurrent neural networks based on the internal reinforcement learning[J]. *Control Project*, 2005, 12(2): 138 – 140.)

作者简介:

任红格 (1979—), 女, 博士研究生, 研究方向为机器人及其智能控制等, E-mail: renhongge@emails.bjut.edu.cn;

阮晓钢 (1958—), 男, 教授, 博士生导师, 研究方向为机器人、自动控制与人工智能等, E-mail: adrxg@bjut.edu.cn.