第 27 卷第 7 期
2010 年 7 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 27 No. 7
Jul. 2010

文章编号: 1000−8152(2010)07−0960−05

模糊操作条件概率自动机仿生自主学习系统和机器人自平衡控制

阮晓钢, 蔡建羡

(北京工业大学 电子信息与控制学院, 北京 100124)

摘要: 为了实现两轮机器人的自平衡控制, 利用Skinner操作条件反射机理, 以概率自动机为平台, 融入模糊推理, 构造了模糊操作条件概率自动机(OCPA)仿生自主学习系统. 该学习系统是一个从状态集合到操作行为集合的随机映射, 采用操作条件反射学习机制, 从操作行为集合中随机学习作为控制系统控制信号的最优行为, 并利用学习到的操作行为取向值信息, 调整操作条件反射学习算法. 此外, 学习系统还引入行为熵, 以验证其自学习和自组织能力. 应用于两轮机器人自平衡控制的仿真结果, 验证了模糊OCPA学习系统的可行性.

关键词: 操作条件反射; 概率自动机; 模糊推理; 仿生自主学习系统; 熵; 自平衡控制

中图分类号: TP273　　**文献标识码**: A

Fuzzy operant conditioning probabilistic automaton bionic autonomous learning system and robot self-balancing control

RUAN Xiao-gang, CAI Jian-xian

(College of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China)

Abstract: A fuzzy operant conditioning probabilistic automaton(OCPA) bionic autonomous learning system is constructed based on Skinner operant conditioning theory and combined with the probabilistic automaton and fuzzy inference for realizing a two-wheeled robot self-balancing control. The learning system is a stochastic mapping from state sets to operant action sets. The optimal action for controlling the system is stochastically learned from the operant action set by adopting operant conditioning learning algorithm; in the same time the orientation value information of the learned operant action is used to adjust the operant conditioning learning algorithm. In addition, the action entropy is added to verify the self-learning and self-organizing ability of the learning system. In the simulation, a two-wheeled robot self-balancing control is realized, demonstrating the feasibility of the fuzzy OCPA learning system.

Key words: operant conditioning; probabilistic automaton; fuzzy inference; bionic autonomous learning system; entropy; self-balancing control

1 引言(Introduction)

两轮自平衡机器人是本质上多变量、强耦合和非线性的复杂动态系统, 其核心问题是运动平衡控制. 围绕这一核心问题, 国内外都开展了两轮机器人自平衡控制的研究[1~4]. 但是, 这些研究仍是以传统的控制技术和控制方法为主, 这导致动态性能和静态性能都较差. McFarland D[5]认为, 以仿生的思路设计机器人要优于传统的设计方式, 让机器人具有类似人和动物的自主学习机制, 能自主地学习平衡控制技能, 是人工智能的一个重要研究领域.

在心理学历史上, 美国Harvard大学心理学教授Skinner提出的操作条件反射(operant conditioning, OC)[6]是一种重要的条件反射理论, 被视为生物系统最基本的学习形式. 这是因为人或动物的平衡控制技能在很大程度上是基于这种学习机制自组织地渐进形成、发展和完善的. 自20世纪90年代中期开始, 美国卡内基梅隆大学(CMU)机器人学研究所, 主要研究关于Skinner OC的计算理论和计算模型, 期望这种模型能复制动物学习操作或控制的实验; 然后在机器人上实现这种模型, 使其成为可训练的机械[7~10]. 但是, 这些计算理论和计算模型没有给出具体的数学计算模型, 不具备泛化能力, 应用受到了限制.

基于上述现状, 本文利用Skinner OC机理, 以概

收稿日期: 2009−04−11; 收修改稿日期: 2009−08−10.
基金项目: 国家自然科学基金资助项目(60774077); 国家“863”计划资助项目(2007AA04Z226); 北京市教委重点资助项目(KZ200810005002).

率自动机(probabilistic automaton, PA)为平台[11], 融入模糊推理[12], 构造了模糊OCPA(operant conditioning probabilistic automaton)仿生自主学习系统. 该学习系统是一个从状态集合到操作行为集合的随机映射, 设计OC学习机制, 从操作行为集合中随机学习作为控制系统控制信号的最优行为; 利用操作行为取向信息调整OC学习算法. 同时, 引入行为熵来评价学习系统的自学习和自组织能力. 应用于两轮机器人自平衡控制的仿真结果, 验证了模糊OCPA学习系统的可行性.

2 模糊OCPA学习系统(Fuzzy OCPA learning system)

PA由于主要研究所处环境具有有限随机因素的自动机, 所以比较适合于构建仿生学习模型, 且能使学习算法具有全局性能; 而模糊推理的过程和OC的形成过程类似, 所以融入模糊推理, 能把学习系统学习的经验更形象、直观的表示出来. 基于Skinner OC理论, 以PA为平台, 并融入模糊推理机制, 构造出的模糊OCPA学习系统结构如图1所示.

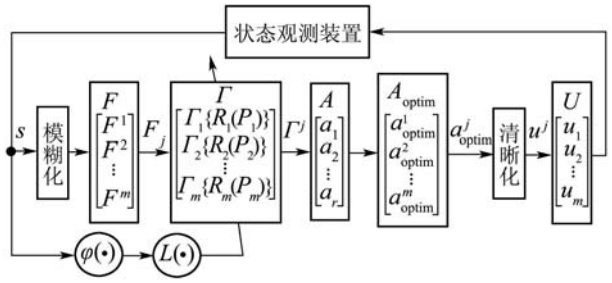

图 1　模糊OCPA学习系统结构

Fig. 1　Structure of fuzzy OCPA learning system

图1中, 采用有限的模糊集合来描述状态和动作空间, 模糊推理视为状态空间到行为空间的映射关系. 按照PA定义的形式, 模糊OCPA学习系统定义如下:

定义 1　模糊OCPA学习系统是一个九元组离散计算模型: $\mathrm{OCPA} = \langle U, A, s, F, \Gamma, f, \varphi, L, H\rangle$, 各部分含义说明如下:

1) 模糊OCPA学习系统的清晰化操作行为集合: $U = \{u_j \,|\, j = 1, 2, \cdots, m\}$, u_j等价于模糊推理系统的第j个输出控制信号, m表示输入状态到输出行为的映射数目;

2) 模糊OCPA学习系统的模糊操作行为集合: $A = \{a_k \,|\, k = 1, 2, \cdots, r\}$, a_k表示模糊OCPA学习系统的第k个可选模糊操作行为, 等价于模糊推理的可能后件;

3) 模糊OCPA学习系统的内部状态: $s(t)$, 它表示状态观测装置检测到的实际状态值, 等价于模糊推理系统的状态输入;

4) 模糊OCPA学习系统的内部模糊状态集合: $F = \{F^j \,|\, j = 1, 2, \cdots, m\}$, F为s模糊化后的模糊子集, 等价于模糊推理系统的前件;

5) 模糊OCPA学习系统的模糊状态－模糊操作行为随机映射集合: $\{R_1(P_1), R_2(P_2), \cdots R_{\mathrm{m}}(P_m)\}$, 其中, R_j表示第j个映射; $P_j \in P = \{P_1, P_2, \cdots, P_m\}$表示第$j$个映射的概率矢量. 这里的映射集合等价于模糊规则集合, 所以m亦表示规则总数. 综上, 随机映射关系可描述如下:

$$\begin{aligned} R_j(P_j):\ & \text{If } s(t) \text{ is } F^j, \text{ Then } a \text{ is } a_1(t)|p_{j1},\\ & \text{or } a \text{ is } a_2(t)|p_{j2},\\ & \quad\vdots\\ & \text{or } a \text{ is } a_r(t)|p_{jr}. \end{aligned}$$

其中: $p_{jk} \in P_j = \{p_{j1}, p_{j2}, \cdots, p_{jr}\}$表示学习系统在模糊状态处于$F^j$的条件状态下, 实施模糊操作行为$a_k$的概率值, 满足: $0 < p_{jk} < 1$, $\sum_{k=1}^{r} p_{jk} = 1$;

6) 模糊OCPA学习系统的模糊状态转移函数: f: $F^j(t) \times a_k(t) \to F^j(t+1)$, $t+1$时刻的状态$F^j(t+1)$由t时刻的状态$F^j(t)$和t时刻的操作行为$a_k(t) \in A$确定, 与t时刻之前的状态和行为无关;

7) 模糊OCPA学习系统的取向函数: $\varphi = \{\varphi_1, \varphi_2 \cdots, \varphi_m\}$, $\varphi_j \in \varphi$表示对状态F^j的取向程度, 满足: $0 < \varphi_j < 1$, 其变化趋势用来更新概率矢量P_j. 若定义系统实际状态$s(t)$与期望状态$s^*(t)$的误差: $e(t) = s(t) - s^*(t)$, 则取向值定义如下:

$$\begin{cases} \varphi(e(t)) = \gamma e(t) + 1, & 0 < e(t) < 1/\gamma,\\ \varphi(e(t)) = -\gamma e(t) + 1, & -1/\gamma < e(t) < 0, \end{cases} \tag{1}$$

其中$\gamma e(t) \in [-1, 1]$, γ值依据实际的情况选取.

8) 模糊OCPA学习系统的OC学习算法: L: $\Gamma_j(t) \to \Gamma_j(t+1)$, 按照Skinner OC 理论, 若实施行为$a_k(t)$后, 相邻时刻取向值的差$\varphi(e(t+1)\,|F^{j'}(t+1)) - \varphi(e(t)\,|F^j(t)) < 0$, 则实施该行为的概率$p(a_k(t)|F^j(t))$倾向于减小, 反之亦然, 所以OC学习机制可形式化的用公式描述为

$$\begin{cases} \text{If } \varphi(e(t+1)\,|F^{j'}(t+1)) - \varphi(e(t)\,|F^j(t)) > 0,\\ \text{Then} \begin{cases} p_{jk}(t+1) = p_{jk}(t) + \Delta_1, & a(k) = a_k,\\ p_{jk'}(t+1) = p_{jk'}(t) - \Delta_1', & a(k) \neq a_k. \end{cases} \end{cases} \tag{2}$$

上式增量部分设计为: $\begin{cases}\Delta_1=\alpha(t)[1-p_{jk}(t)]\\ \Delta_1'=\alpha(t)p_{jk'}(t)\end{cases}$, 其中$\alpha(t)=\lambda\varphi\left(e(t+1)\left|F^{j'}(t+1)\right.\right)$;

$$\begin{cases}\text{If } \varphi(e(t+1)\left|F^{j'}(t+1)\right.)-\varphi(e(t)\left|F^{j}(t)\right.)<0,\\ \text{Then}\begin{cases}p_{jk'}(t+1)=p_{jk'}(t)-\Delta_2', & a(k)=a_{k'},\\ p_{jk}(t+1)=p_{jk}(t)+\Delta_2, & a(k)\neq a_k.\end{cases}\end{cases}\tag{3}$$

上式增量部分设计为: $\begin{cases}\Delta_2'=\beta(t)p_{jk'}(t),\\ \Delta_2=\beta(t)[\dfrac{1}{r-1}-p_{jk}(t)],\end{cases}$ 其中$\beta(t)=\eta\varphi(e(t+1)\left|F^{j'}(t+1)\right.)$.

公式(2)(3)中, $\alpha(t)$,$\beta(t)$为学习速率函数, 满足: $0<\alpha(t)<1, 0<\beta(t)<1$; λ,η为学习系数, 满足: $0<\lambda<1, 0<\eta<1$.

9) 模糊OCPA学习系统的模糊操作行为熵: $H=\{H_1,H_2,\cdots,H_m\}$, $H_j(t)\in H$表示学习系统处于模糊状态F^j条件下的操作行为熵,

$$\begin{aligned}&H_j(t)=H_j(A(t)|F^j)=\\&-\sum_{k=1}^{r}p_{jk}\log_2 p_{jk}=\\&-\sum_{k=1}^{r}p(a_k|F^j)\log_2 p(a_k|F^j),\end{aligned}\tag{4}$$

这里的操作行为熵就是度量系统无组织程度的“信息熵”, 当所有操作行为$a_k(t)$可能出现的概率相等时, 操作行为熵最大. 行为熵的减小是系统自组织和自学习的特征.

综上, 模糊OCPA学习系统的学习步骤如下:

Step 1 初始化: 由于一开始操作行为的取向信息是未知的, 所以选取初始操作概率为: $p_{jk}(0)=\dfrac{1}{r}$. 选取各行为初始概率相同, 意味着初始状态下, 模糊OCPA学习系统不含有任何预定的决策.

Step 2 选取操作行为: 假设t时刻系统模糊状态为F^j, 依据随机映射Γ_j中的概率矢量$p_{jk'}(t)$, 随机选取操作行为$a_k(t)$.

Step 3 执行操作行为: 实施选取的操作行为$a_k(t)$, 并观测$t+1$时刻的状态$F^{j'}(t+1)$.

Step 4 取向值增量计算: 依据取向函数$\varphi(e)$, 分别计算状态$F^j(t)$和$F^{j'}(t+1)$的取向值, 由此得取向值增量: $\vec{\varphi}(e)=\varphi(e(t+1)\left|F^{j'}(t+1)\right.)-\varphi(e(t)|F^j(t))$

Step 5 操作条件反射: 据式(2)和式(3)的OC学习算法, 调整操作行为$a_k(t)$的实施概率矢量$p_{jk}(t+1)$.

Step 6 递归转移: 如果不满足$\lim\limits_{t\to\infty}p_k(t+1)\approx 1$, $\lim\limits_{\substack{t\to\infty\\k\neq k'}}p_{k'}(t+1)\approx 0$, 则更新时刻变量$t=t+1$, 转Step 2, 按照修改后的概率矢量$p_{jk}(t+1)$随机选择新的控制量$a_{k'}(t+1)$. 反之, 则说明学习系统已经学习到了最优的操作行为a^j_{optim}, 转向Step 7.

Step 7 结束.

3 学习算法收敛性证明(Convergence proof of learning algorithm)

引理 1 模糊OCPA$=\langle U,A,s,F,\Gamma,f,\varphi,L,H\rangle$是一个仿生自主学习系统, 则成立:

$$\begin{cases}\lim\limits_{t\to\infty}p_{jk}\left(a_k(t)\left|F^j(t)\right.\right)=1,\\ \lim\limits_{\substack{t\to\infty\\k\neq k'}}p_{jk'}\left(a_{k'}(t)\left|F^j(t)\right.\right)=0.\end{cases}\tag{5}$$

其中: $a_k(t)$表示使取向值趋于极大的操作行为, $a_{k'}(t)$表示使取向值趋于极小的操作行为.

证 假定t时刻, 模糊OCPA学习系统在模糊状态F^j的条件下选取操作行为$a_k\in A$, 则

1) 若$a(t)=a_k(t)$, 由式(2)得

$$\begin{aligned}&\Delta p_{jk}(t)=p_{jk}(t+1)-p_{jk}(t)=\\&\alpha(t)-\alpha(t)p_{jk}(t)=\alpha(t)(1-p_{jk}(t),\end{aligned}\tag{6}$$

因$0<p_{jk}(t)<1$, 且$0<\alpha(t)<1$, 得: $\Delta p_{jk}(t)\geqslant 0$.

2) 若$a(t)=a_k(t)$, 由式(3)得

$$\begin{aligned}&\Delta p_{jk}(t)=p_{jk}(t+1)-p_{jk}(t)=\\&\beta(t)-\beta(t)p_{jk}(t)=\beta(t)(1-p_{jk}(t),\end{aligned}\tag{7}$$

又因$0<\beta(t)<1$, 得$\Delta p_{jk}(t)\geqslant 0$.

由式(6)(7), 可知行为$a_k(t)$的概率增量$\Delta p_{jk}(t)\geqslant 0$, 所以当$t\to\infty$时$a_k(t)$被选中的频次逐渐增加, 对应$p_{jk}(t)$也将不断升高. 又由式(6)(7)可知: $\Delta p_{jk}(t)\geqslant 0$的等号当且仅当$p_{jk}(t)=1$时成立, 所以$p_{jk}(t)$的上界为1, $p_{jk}(t)$的增长将直至$p_{jk}(t)=1$为止. 同理可得, $p_{jk'}(t)$的降低将直至$p_{jk'}(t)=0$为止.

证毕.

定理 1 模糊OCPA $=\langle U,A,s,F,\Gamma,f,\varphi,L,H\rangle$ 是一个仿生自主学习系统, 则模糊OCPA学习系统处于状态F^j的操作行为熵$H_j(\{A(t)\}|F^j)$随时间t收敛至极小: $\lim\limits_{t\to\infty}H_j(t)=H_{j\min}$.

证 学习初始阶段, 所有操作行为$a_k(t)$可能出现的概率$p_{jk}(t)$相等, 操作行为熵最大. 随着学习的进行, 若熵值能趋于极小, 则可验证学习系统的自学习和自组织能力. 对式(4)重新进行整理得

$$\begin{aligned}H_j(t)=&-\sum_{k=1}^{r}p(a_k|F^j)\log_2 p(a_k|F^j)=\\&-[p(\alpha_k|F^j)\log_2 p(\alpha_k|F^j)+\end{aligned}$$

$$\sum_{k'=1,k'\neq k}^{r} p(a_{k'}|F^j)\log_2 p(a_{k'}|F^j)]. \tag{8}$$

把引理1中式(5)代入上式, 整理得

$$\lim_{t\to\infty} H_j(t) = H_{j\min} \approx 0, \tag{9}$$

证毕.

4　仿真分析(Simulation analysis)

本文以北工大人工智能与机器人研究所研制的两轮机器人为研究对象, 如图2所示.

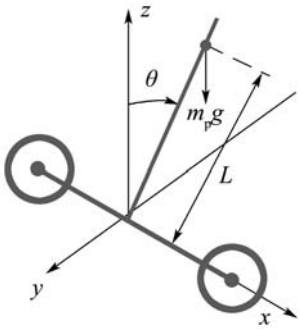

图 2　两轮机器人示意图

Fig. 2　Schematic diagram of two-wheeled robot system

图2中: θ为杆偏离垂直方向的角度, $m_{\rm p}$为机器人机身的质量; L为质心到轮轴的距离. 两轮机器人是在控制其不倒的前提下控制其轨迹的, 所以平衡控制是其根本控制任务. 机器人的倾角θ和控制力矩分别和模糊OCPA学习系统的状态F和操作行为a相对应, 并均模糊化为5个模糊子集: $F=\{\rm NB, NS, ZE, PS, PB\}$, 论域为$X=[-2,-1,0,1,2]$, 各模糊变量均采用三角形隶属函数. 结合机器人系统的实际情况, 取向函数中的系数: $\gamma=\dfrac{1}{12}$; OC学习算法中的常系数: $\lambda=0.1, \eta=0.2$; 初始状态: $[\theta\ \ \dot{\theta}]=[10^\circ\ 0]$; 操作行为集合$A=[-10,-2,0,2,10]$; 采样时间$t_{\rm s}=0.01$ s; 操作行为的初始概率均为$p_{jk}(a_k(0)|F^j(0))=\dfrac{1}{5}$; 由初始行为概率值, 计算得初始行为熵值:

$$H(0)=\sum_{i=1}^{25}(-\sum_{j=1}^{5}\frac{1}{5}\times\log_2\frac{1}{5})=58.05,$$

此时熵值最大.

大约学习120步后, 操作行为的概率变化曲线如图3所示.

由图3所示结果, 可得操作行为的概率值近似为

$$\lim_{t\to\infty} p_{j3}(a_3(t))\approx 0.96,\ \ \lim_{\substack{t\to\infty\\ k\neq 3}} p_{jk}(a_k(t))\approx 0.01,$$

代入操作行为熵计算公式得: $H(120)\approx\sum_{i=1}^{5}(-0.96\times\log_2 0.96)\approx 0.2$, 达到最小值.

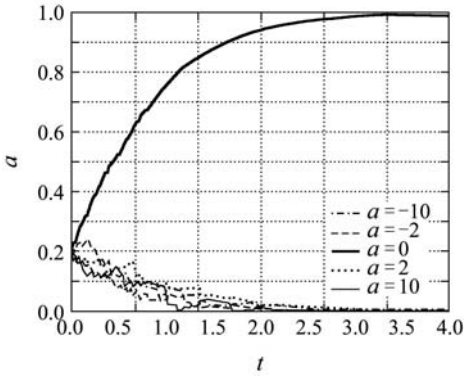

图 3　概率变化曲线

Fig. 3　Cure of probability change

上述仿真结果中, 操作行为概率值的变化证明了设计的OC学习算法是收敛的; 操作行为熵值的变化则验证了模糊OCPA学习系统能成功的模拟生物的OC机制, 具有仿生的自组织、自学习功能.

两轮机器人的倾角和角速度仿真结果如图4和图5所示, 为了验证模糊OCPA学习系统的性能, 和LQR控制进行了比较.

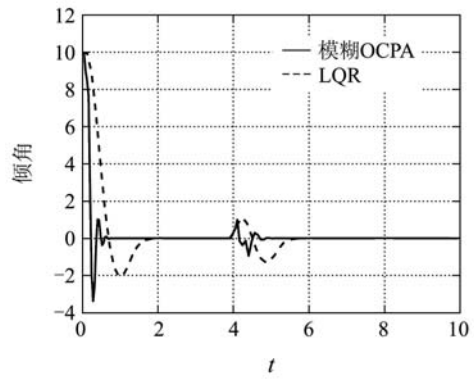

图 4　倾角仿真曲线

Fig. 4　Result of dip angle error

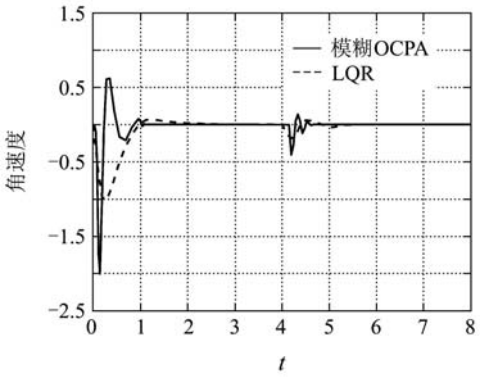

图 5　角速度仿真曲线

Fig. 5　Result of angle speed

由图4、图5可以看出, 学习初始阶段, 由于模糊OCPA学习系统没有任何学习经验, 所以振荡较

大, 但学习速度很快, 机器人大约1.2 s后就恢复到平衡状态; LQR控制和其相比, 虽然初始阶段振荡较小, 但大约经过2 s后才恢复到平衡状态. 这说明在本文设计的方案下, 机器人可以快速的自主学习到最优的操作行为, 成功实现其自平衡控制.

为了进一步验证模糊OCPA学习系统的性能, 在4 s时施加一个正脉冲, 由仿真结果可以看出, 模糊OCPA学习系统大约在0.7 s后使机器人迅速恢复到了平衡状态, 而LQR控制大约经过1.5 s后才恢复到平衡状态. 这说明外界环境发生变化时, 采用模糊OCPA学习系统能更快地适应环境的变化, 并具有较强的鲁棒性和自适应能力.

5 结论(Conclusion)

本文基于Skinner OC原理, 以概率自动机为平台, 同时融入模糊推理机制, 构造了模糊OCPA学习系统, 以实现两轮机器人的自平衡控制. 模糊OCPA学习系统实际上就是一个从状态集合到操作行为集合的随机映射, 采用OC学习算法可以学习到最优的操作行为. 在两轮机器人的自平衡控制上的仿真结果表明, 设计的模糊OCPA学习系统不需要系统的模型, 和传统的控制算法相比, 具有较快的学习收敛速度; 并表现出类似动物操作条件反射的自主学习特性; 当施加外部扰动时, 能迅速适应环境的变化, 表现出较强的鲁棒性和自适应能力.

参考文献(References):

[1] KIM D H, OH J H. Tracking control of a two-wheeled mobile robot using input-output linearization[J]. *Control Engineering Practice*, 1999, 7(3): 369 – 373.

[2] URAKUBO T, TSUCHIYA K, TSUJITA K. Motion control of a two-wheeled mobile robot[J]. *Advanced Robotics*, 2001, 15(7): 711 – 728.

[3] ZHOU J, WU W. The application of disturbance observer in two-wheeled mobile robot[C] *//2004 IEEE Conference on Robotics, Automation and Mechatronics*. Singapore: IEEE, 2004, 1: 171 – 174.

[4] KOZLOWSKI K, PAZDERSKI D. Stabilization of two-wheeled mobile robot using smooth control law[C] *//The 2006 IEEE International Conference on Robotics and Automation*. Orlando: Institute of Electrical and Electronics Engineers Inc, 2006: 3387 – 3392.

[5] MCFARLAND D, BOSSER T. *Intelligent Behavior in Animals and Robots*[M]. Cambridge: MIT Press, 1993.

[6] SKINNER B F. *The Behavior of Organisms*[M]. New York: Appleton-Century-Crofts, 1938.

[7] SAKSIDA L M, TOURETZKY D S. Application of a model of instrumental conditioning to mobile robot control[J]. *Sensor Fusion and Decentralized Control in Autonomous Robotic Systems*, 1997, 32(9): 55 – 66.

[8] TOURETZKY D S, SAKSIDA L M. Operant conditioning in Skinnerbots[J]. *Adaptive Behavior*, 1997, 5(3/4): 219 – 247.

[9] SAKSIDA L M, RAYMOND S M, TOURETZKY D S. Shaping robot behavior using principles from instrumental conditioning[J]. *Robotics and Autonomous Systems*, 1998, 22(3/4): 231 – 249.

[10] TOURETZKY D S, DAW N D, TIRA-THOMPSON E J. Combining configured and TD learning on a robot[C] *// The 2nd International Conference on Development and Learning*. Pittsburgh: IEEE, 2002: 47 – 52.

[11] 陶仁骥. 自动机引论[M]. 北京: 科学出版社, 1986.
(TAO Renji. *Automata Introduction*[M]. Beijing: Science Press, 1986.)

[12] 杨苹, 许礼尉. 聚焦式模糊变结构控制及其在主汽温控制中的应用[J]. 控制理论与应用, 2007, 24(1): 137 – 142.
(YANG Ping, XU Liwei. Focusing variable structure fuzzy control and its application to steam temperature control of a boiler[J]. *Control Theory & Applications*, 2007, 24(1): 137 – 142.)

作者简介:

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

蔡建羡 (1978—), 女, 讲师, 博士研究生, 目前研究方向为自治机器人智能性能的研究, E-mail: cjxlaq@163.com.