文章编号: 1001-0920(2011)09-1398-04

基于滑模思想和Elman网络的操作条件反射学习控制方法

阮晓钢, 陈 静

(北京工业大学 电子信息与控制工程学院，北京 100124)

摘 要: 针对一类单输入单输出高阶非线性控制系统, 提出一种基于滑模思想和Elman网络的操作条件反射(OCR)学习控制方法. 该方法采用Elman网络构造滑模面-行为对的评价函数, 通过滑模面的变化设计奖赏函数, 根据奖赏信号更新评价函数, 实现行为选择概率的更新. 通过每轮次熵的定义, 定量分析了所学知识的变化量. 针对行走倒立摆系统的仿真实验结果表明, 采用该仿生的OCR学习控制方法, 可实现行走倒立摆的平衡控制.

关键词: 操作条件反射；滑模控制；Elman网络；熵；倒立摆；平衡控制

中图分类号: TP273 **文献标识码:** A

Operant conditioning reflex learning control scheme based on SMC and Elman network

RUAN Xiao-gang, *CHEN Jing*

(School of Electronic Information and Control Engineering，Beijing University of Technology，Beijing 100124，China. Correspondent：CHEN Jing，E-mail：chenjing0828@139.com)

Abstract: A bionic operant conditioning reflex(OCR) learning control scheme is proposed based on the thought of sliding model control(SMC) and Elman network for a class of SISO higher-order nonlinear control system. In this method, an Elman network is used as an evaluation function of sliding surface and action in the scheme. Reward signal is designed according to the change of sliding surface, and then the evaluation function is updated through the reward stimulation, while the behavior choice probability is changed. Through the definition of entropy for each round, a quantitative analysis about the knowledge change in the learning process is given. The results of the simulation experiment in the walking inverted pendulum system show that, bionic OCR learning control is used, which realizes the balancing control for the walking inverted pendulum system.

Key words: operant conditioning reflex；sliding model control；Elman network；entropy；inverted pendulum；balancing control

1 引 言

仿生自主学习控制是近年来机器人领域的研究热点, 基于操作条件反射(OCR)原理的智能体学习源于心理学, 是一种集计算机技术、自动控制技术、仿生学、心理学、生物学于一体的一项理论, 目前对该理论的研究成果多见于生物实验方面, 在机器人控制方面应用并不多见. 能够体现操作条件反射原理的理论主要有概率自动机理论和离散动作的Q学习理论, 其中重要的一个特点是行为的概率选择机制.

1938年, 美国哈佛大学心理学教授Skinner发表了一部具有影响力的著作[2], 由此创立了Skinner操作条件反射理论(OCR). 操作条件反射[1-2]和经典条件反射[3]是联想学习的两个主要的学习方式, 所有的动物(包括人类)都能体现这两种方式. 不同的是, 操作条件反射(OCR)体现的是行为和行为产生的结果之间的联结关系.

操作条件反射理论得到了国内外学者的广泛关注, 近年来, 人们把操作条件反射理论应用于机器人学习和控制方面, 做了大量的实验和研究. 德国的Brembs等人[4-6]利用果蝇和蜗牛实验, 研究了操作条件反射, 将“纯”操作条件反射和并行操作条件反射在一个生物的飞行仿生器中进行了模拟, 指出智能体

收稿日期: 2010-06-01；**修回日期:** 2010-08-25.

基金项目: 国家863计划项目(2007AA04Z226)；国家自然科学基金项目(60774077)；北京市自然科学基金项目(4102011)；北京市教委重点项目(KZ200810005002).

作者简介: 阮晓钢(1958−), 男, 教授, 博士生导师, 从事机器人、自动控制等研究；陈静(1984−), 女, 博士生, 从事机器人、智能控制等研究.

对未来奖赏的预测非常重要, 它指导自身进行下一步的决策. 2002年, Zalama等人[7]针对机器人自主导航问题, 设计了一种基于OCR理论的刺激反应型神经网络. 2005年, 日本早稻田大学的Itoh等人用基于OCR的Hull行为理论[8], 使机器人WE-4RII学会了用右手和人握手的行为. 然而, 目前OCR学习控制机制在复杂控制对象上的应用还不够成熟. 传统的强化学习方法是针对系统状态进行行为的设计, 通过不断的试错反馈给系统奖惩信号, 许多不必要的信息对于系统而言是非常危险的. 换言之, 为了找到合适的输入序列, 智能体需要进行大量的"试错", 在极端的情况下, 可能会对系统造成很大的破坏. 为此, 寻求一种尽量减少试探次数的学习控制方法无疑对无模型控制是非常必要的.

王雪松等人[9]基于Elman网络研究了非线性系统增强式学习控制, 用于小车爬山问题, 达到了预期的学习效果, 控制对象相对比较简单; Obayashi等人[11]采用基于滑模思想的RBF网络在倒立摆的平衡姿态学习中得到了成功的应用, 学习速度快, 但控制效果不好, 控制初期系统震荡较为明显.

基于上述方法各自的优缺点, 本文将滑模的思想应用于OCR机制中, 利用Elman网络良好的动态特性, 提出一种基于Elman网络和滑模思想的OCR学习控制方法. 用Elman网络构造滑模面-行为对的评价函数, 通过滑模面的变化来设计奖赏信号, 根据奖赏信号不断更新评价函数, 从而实现行为选择概率的更新. 将该方法用于行走倒立摆的平衡控制中, 采用仿真实验验证了所提出方法的有效性和快速性.

2　OCR学习机制介绍

本文以一个经典的鸽子实验来说明操作条件反射理论的学习机制. 该实验是这样一个过程: 鸽子在一个空间里, 其中安装有红、黄、蓝3个按钮, 当鸽子啄红色按钮时, 会有食物出来; 啄黄色按钮时, 无现象; 啄蓝色按钮时, 会有突然的电击作为惩罚. 在鸽子不断的按钮过程中, 逐渐学会了获取食物. 按红色按钮的概率趋近于1, 按其他两个按钮的概率趋近于0. 机器鸽的这种行为是一个试探性的, 通过操作后的结果来引导以后的操作.

基于操作条件反射原理的学习控制主要有3个要素: 行为选择机制(依概率选择行为), 评价机制和取向机制. 取向机制是学习的核心环节, 用来更新行为选择策略. 图1为基于OCR原理的学习控制机制示意图.

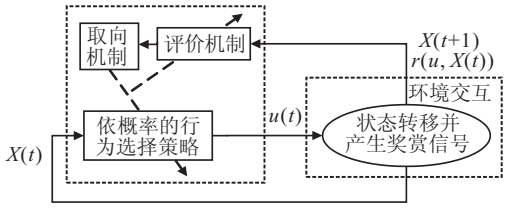


图 1　基于OCR原理的学习机制

3　滑模思想和Elman网络

3.1　非线性控制系统中的滑模思想

考虑如下所示的n阶非线性微分方程:

$$.x^{(n)} = f(\boldsymbol{x}) + b(\boldsymbol{x})u. \tag{1}$$

其中: $\boldsymbol{x} = [x, \dot{x}, \ddot{x}, \cdots, x^{(n-1)}]^{\mathrm{T}}$为系统的状态向量, 假设所有的状态是可观测的; u为控制输入; $f(\boldsymbol{x})$和$b(\boldsymbol{x})$为未知的连续函数.

控制目标为: 期望系统跟踪目标状态x_d, 定义误差向量$\boldsymbol{e} = [x - x_d, \dot{x} - \dot{x}_d, \ddot{x} - \ddot{x}_d, \cdots, x^{(n-1)} - x_d^{(n-1)}]^{\mathrm{T}}$.

滑模面定义为

$$H : \{\boldsymbol{e} | s(\boldsymbol{e}) = 0\}, \tag{2}$$

$$s(\boldsymbol{e}) = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{e}, \tag{3}$$

其中$\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \alpha_2, \cdots, \alpha_{(n-1)}]^{\mathrm{T}}$. 选取合适的$\boldsymbol{\alpha}$使多项式方程$\alpha_{(n-1)}p^{(n-1)} + \alpha_{(n-2)}p^{(n-2)} + \cdots + \alpha_1 p + \alpha_0 = 0$的根在相平面的左半平面, 即满足Hurwitz稳定性条件.

当满足以上条件时, 系统稳定性的条件可以转化为$\lim\limits_{t\to\infty} s(\boldsymbol{e}) = 0$.

3.2　Elman网络

Elman网络是回归神经网络的一种, 如图2所示. 从系统观点看, 它不仅含有输入层、隐层和输出层的节点, 而且含有与隐层节点数相同的反馈节点, 用来记忆隐层单元以前时刻的输出值, 可认为是一个时延算子. 信号能够在神经元之间往返传递, 整个网络处在一种不断改变的动态之中, 能够更直接更生动地反映系统在计算过程中的动态特性, 比前向神经网络具有更强的动态行为和计算能力.

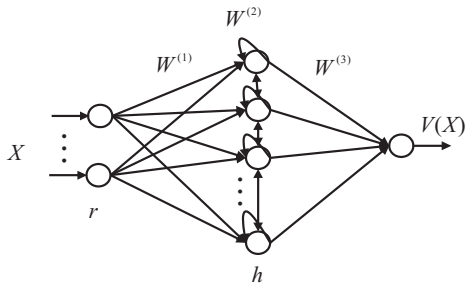


图 2　Elman网络拓扑结构

4　OCR学习机制的实现

Elman网络的输入, $X(t) = [s, a]^{\mathrm{T}} \in \Re^{2\times 1}$(其中$s$按照式(3)计算)表示在$t$时刻的滑模面-动作对, $V(\cdot)$表示对t时刻滑模面-动作对的评价, 同时也是对未来奖赏的一个预测. 隐含层节点个数为h, 整个网

络的连接权值为 $W^{(1)}$, $W^{(2)}$, $W^{(3)}$, 分别为 $2\times h$ 维, $h\times h$ 维, $h\times 1$ 维矩阵. 输出层有一个节点, 激活函数为线性加权函数. 网络数学模型为

$$V(s,a)=(W^{(3)})^{\mathrm{T}}\boldsymbol{H}(t)=$$
$$(W^{(3)})^{\mathrm{T}}f((W^{(2)})^{\mathrm{T}}\boldsymbol{H}(t-1)+(W^{(1)})^{\mathrm{T}}X(t)). \quad (4)$$

$\boldsymbol{H}(t)$ 为隐含层的输出, $\boldsymbol{H}(t-1)$ 为前一时刻隐含层的输出. 其中: 隐含层激活函数为 Sigmoid 函数, 即

$$f(x)=\frac{1}{1+\mathrm{e}^{-x}},\ \frac{\mathrm{d}f(x)}{\mathrm{d}x}=f(x)[1-f(x)].$$

4.1 评价机制

针对系统 (1) 设计评价机制, 根据滑模面的变化实施奖惩, 如果实施某一行为后, s 越靠近 0, 则给予奖赏; 反之, 则给予惩罚. 0 代表奖励, -1 代表惩罚.

$$r(t)=\begin{cases}0,\ s_{t+1}^2\leqslant s_t^2;\\ -1,\ s_{t+1}^2>s_t^2.\end{cases} \quad (5)$$

4.2 取向机制

通过更新网络参数修正行为网络的概率选择策略. 这里用 Q 学习中常用的 TD 误差进行修正, TD 误差的平方和 $e(t)=\frac{1}{2}\Delta_{\mathrm{TD}}^2(t)$, 采用梯度下降和资格迹相结合的方法进行权值的更新, 加速网络学习过程. 网络权值更新法则为

$$\begin{cases}\Delta_{\mathrm{TD}}(t)=r_t+\gamma V(s_{t+1},a_{t+1})-V(s_t,a_t),\\ e_c(t)=\displaystyle\sum_{k=1}^{t}(\gamma\lambda)^{t-k}\frac{\partial V(s_k,a_k)}{\partial W(k)}=\\ \qquad\gamma\lambda e_c(t-1)+\dfrac{\partial V(s_t,a_t)}{\partial W(t)},\\ \Delta W(t)=\eta\Delta_{\mathrm{TD}}(t)e_c(t).\end{cases} \quad (6)$$

其中: $0<\gamma<1$ 为折扣因子, 表示学习系统的远视程度, 如果取值较小, 则表示系统比较关注最近动作的影响; 如果取值较大, 则对比较长时间内的动作都很关注. λ 为资格迹的折扣因子 (或称传导率), 一般情况下 $0\leqslant\lambda\leqslant 1$, $e_c(t)$ 为 $W(t)$ 的资格迹. 资格迹的引入解决了时间信度分配的问题, 利用累加当前和过去的梯度值, 加快了学习速度.

4.3 行为选择机制

在 OCR 学习机制中, 最重要的特点是依概率选择行为, 这里用 Boltzmann-Gibbs 概率分布来定义选择行为的概率, 表达式为

$$P(a=a_t|s_t)=\frac{\mathrm{e}^{V(s_t,a_t)/T}}{\displaystyle\sum_{a\in A}\mathrm{e}^{V(s_t,a_t)/T}}. \quad (7)$$

其中: $T>0$ 为温度常数, 表示行为选择的随机程度, T 越大, 行为选择的随机程度越大; T 越小, 行为选择的随机程度越小; 当 T 趋近于零时, 选择最大 V 值对应的行为的概率为 1. T 是随着时间递减的, 表示系统在学习过程中获得了越来越多的经验知识, 从一个不确定性系统逐渐演化为确定性系统. 为了表现生物系统的活性, 温度常数不能太小, T 的范围为 $[0.1\sim 0.5]$, 系统初始行为选择的随机性通过随机初始化 Elman 网络的权值来实现.

4.4 熵

熵在信息论中是表征系统不确定程度的物理量, 有概率的地方就有熵的存在, 熵越大, 系统的不确定性越大. 而学习系统熵的变化应该是减小的, 表示学习系统知识量的增加. 在某一状态下熵 E_k 定义如下:

$$E_k=-\sum_{i\in U(k)}\pi_k(i)\log\pi_k(i), \quad (8)$$

其中 $\pi_k(i)$ 表示在状态 k 下选择行为 i 的概率.

为了定量分析学习的过程, 这里定义一个学习周期的熵

$$E_{\mathrm{period}}=\sum_k E_k. \quad (9)$$

5 实验仿真及分析

将上述基于 Elman 网络和滑模思想的 OCR 学习控制方法应用于行走倒立摆系统, 如图 3 所示.

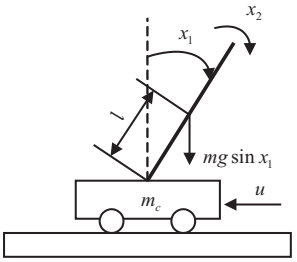


图 3 行走倒立摆系统

x_1 是摆的角度, x_2 是角速度, 其动态方程为

$$\begin{cases}\dot{x}_1=x_2,\\ \dot{x}_2=f(x_1,x_2)+g(x_1,x_2)u.\end{cases} \quad (10)$$

其中

$$f(x_1,x_2)=\frac{(m_c+m)g\sin x_1-mlx_2^2\cos x_1\sin x_1}{l\Big(\dfrac{4}{3}(m_c+m)-m\cos^2 x_1\Big)},$$
$$g(x_1,x_2)=\frac{\cos x_1}{l\Big(\dfrac{4}{3}(m_c+m)-m\cos^2 x_1\Big)};$$

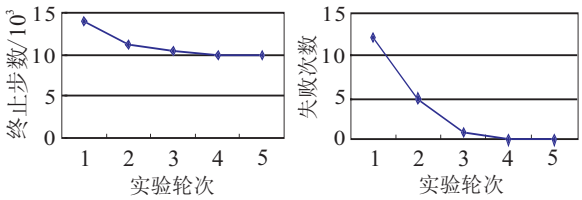
$g=9.8\,\mathrm{m/s^2}$ 是重力加速度, m_c 是小车质量, m 是倒立摆的质量, l 是倒立摆长度的一半, u 是施加的外力 (控制), 取值范围为 $[-10,10]$ N. 这里取 $m_c=1\,\mathrm{kg}$, $m=0.1\,\mathrm{kg}$, $l=0.5\,\mathrm{m}$.

采样时间 $T_s=0.02\,\mathrm{s}$, 如果摆杆角度 $|x_1|>\pi/6$, 则将摆杆拉回到初始角度, 再继续学习; 如果摆杆能够连续保持 1 000 步 (20 s) 不倒, 则实验成功结束. 学习过程中采用的参数: $h=5$, $\eta=0.05$, $\lambda=0.8$, $\gamma=0.9$, $\alpha_0=4(\mathrm{or}\ 1)$, $\alpha_1=1$, $s=\alpha_0x_1+\alpha_1x_2$, $a=u=$

$-10, -5, 0, 5, 10$.

实验发现, 不同的滑模系数对控制结果有一定的影响. 统计实验显示: 当$\alpha_0 = 4$, $\alpha_1 = 1$时, 系统成功学习结束时的平均运行步数为1 404.111步, 平均失败14.666 67次; 当$\alpha_0 = 1$, $\alpha_1 = 1$时, 系统成功学习结束时的平均运行步数为5 223.222步, 平均失败40.222 22次. 究其原因, 不同的滑模系数使得多项式方程$\alpha_{(n-1)}p^{(n-1)} + \alpha_{(n-2)}p^{(n-2)} + \cdots + \alpha_1 p + \alpha_0 = 0$的根在相平面左半平面的位置不同. 通过实验也可以看出, 离虚轴越远, 学习算法的稳定性越好, 速度也越快, 平均失败次数越少.

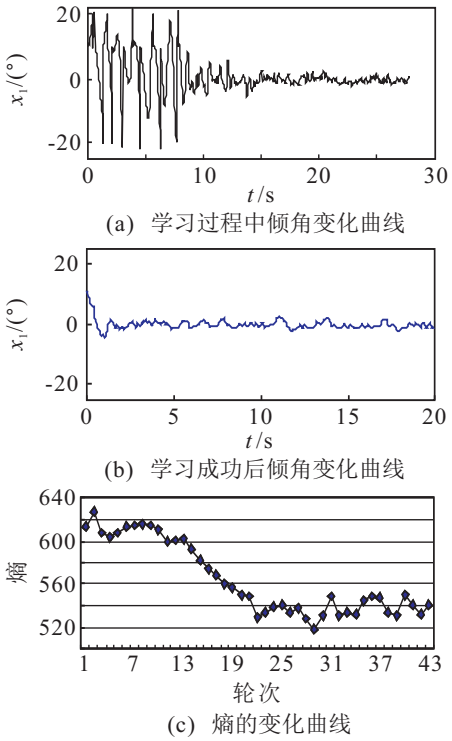
由以上结果, 选择滑模系数$(\alpha_0 = 4, \alpha_1 = 1)$进行仿真实验, 学习算法经过一轮次学习, 在个别状态下的学习次数有限, 因此需要进行多轮次的学习, 从而保证系统尽可能地遍历更多的状态. 需要在前轮实验学好权值的基础上进行下一轮次的实验, 经过4轮次的实验, 倒立摆能够以越来越少的学习步数学习成功, 失败次数为0, 如图4所示.



(a) 成功试验终止步数　　(b) 成功试验失败步数

图 4　多轮学习对比分析图

图5为一次典型实验中的倾角控制结果图, 其中: 图5(a)为第1轮次实验结果, 可以看出, 经过若干次的失败后, 倒立摆能够逐渐学会正确的行为; 图5(b)为5轮次学习结束后的控制结果, 在该轮次学习中失败次数为0, 因为行为的个数有限, 所以角度有些抖动, 但是能够保持在容许的角度范围内. 为了定量分析学习的过程, 进行了43轮次的学习, 按照式(9)计算出每轮次熵的大小. 从图5(c)中可以看出, 通过学习, 系统的熵有一个减小的趋势, 说明系统学习到了一定的知识.



(a) 学习过程中倾角变化曲线

(b) 学习成功后倾角变化曲线

(c) 熵的变化曲线

图 5　一次典型实验中的学习结果图

以上结果是以滑模面-行为对作为Elman网络输入的仿真结果, 与以状态-行为对作为网络输入(Elman网络输入神经元变为3)进行了对比试验, 实验对比数据如表1所示.

表 1　对比实验数据

实验轮次	网络输入为滑模面-行为对		网络输入为状态-行为对	
	终止步数	失败次数	终止步数	失败次数
1	1 393	12	1 332	13
2	1 113	5	1 168	8
3	1 044	1	1 211	7
4	1 000	0	1 115	4
5	1 000	0	1 013	1
6	/	/	1 043	2
7	/	/	1 031	1
8	/	/	1 000	0
9	/	/	1 000	0

由表1可以看出, 采用滑模面-行为对设计的网络仅需要2个神经元作为网络输入, 较之状态-行为对的$n+1$个神经元的网络输入复杂度低; 前者经过4轮次的学习能够成功控制倒立摆的平衡, 后者则需要8轮次学习, 学习速度快于后者; 学习成功后能100%成功控制行走倒立摆的平衡, 算法稳定.

6　结　论

本文针对一类单输入单输出高阶非线性控制系统, 提出了基于滑模思想和Elman网络的OCR学习控制方法, 其优点在于: 1) 基于滑模思想和Elman网络的OCR学习控制方法学习速度快、算法稳定; 2) 采用滑模的思想对控制输入进行了融合, 有效地降低了控制的复杂度, 实验结果优于Obayashi等[11]所采用的方法; 3) 该学习控制方法, 经过多轮次的学习过程, 学习系统能够尽可能遍历所有状态, 达到最佳的学习效果. 仿真实验结果验证了所提出方法的有效性.

参考文献(References)

[1] Thorndike E L. Animal intelligence: Experimental studies[M]. New York: Macmillan, 1911.

[2] Skinner B F. The behavior of organisms: An experimental analysis[M]. New York: Appleton-Century-Crofts, 1938.

(下转第1406页)