文章编号: 1001-0920 (2014) 06-1016-05 DOI: 10.13195/j.kzyjc.2013.0522

一种基于操作条件反射原理的学习模型

阮晓钢[a], 黄 静[a,b], 范青武[b], 魏若岩[a]

(北京工业大学 a. 电子信息与控制工程学院, b. 实验学院, 北京 100124)

摘 要: 针对认知机器人的自主学习问题, 提出一种基于操作条件反射原理的学习模型 (OCLM). 该模型采用状态空间、操作行为空间、概率分布函数、仿生学习机制、系统熵等进行描述, 给出状态的“负理想度”的概念, 定义了取向函数的计算方法. 运用模型对机器人避障导航问题进行仿真实验, 并对参数设置进行了讨论. 实验结果表明, 基于 OCLM 模型的机器人能通过与环境的交互获得认知, 成功避障到达目的地, 具有一定的自学习能力, 从而表明了模型的有效性.

关键词: 学习模型; 操作条件反射; 自学习; 仿生; 避障

中图分类号: TP273 **文献标志码:** A

A learning model based on operant conditioning principles

RUAN Xiao-gang[a], *HUANG Jing*[a,b], *FAN Qing-wu*[b], *WEI Ruo-yan*[a]

(a. College of Electronic Information and Control Engineering, b. Pilot College, Beijing University of Technology, Beijing 100124, China. Correspondent: HUANG Jing, E-mail: mymailhj@sohu.com)

Abstract: Inspired by Skinner's operant conditioning theory, an operant conditioning learning model is presented to deal with the autonomous learning problem in cognitive robotics. The model is described by nine elements, including the space set, the action set, the bionic learning function and the system entropy etc. A notion "negative ideal rate" is defined to compute the orientation function. The OCLM is applied to solve obstacle avoidance and navigation problems for mobile robots. The experiment results show that the robot based on the model can autonomously learn how to arrive at the goal in a collision-free way through interaction with the environment, and show the effectiveness of the proposed model.

Key words: learning model; operant conditioning; autonomous learning; bionics; obstacle avoidance

0 引 言

心理学发展至今, 其理论成果深刻地影响着人工智能、认知机器人学等相关领域的研究. 学习借鉴心理学的经典理论, 对其进行数学建模, 用于解决机器人的实际控制问题, 已成为人工智能、认知机器人研究的一种新思路.

1938 年, Skinner[1]首次提出了操作条件反射的概念, 并由此创立了操作条件反射理论. 他借鉴巴甫洛夫的“强化”概念, 并将这一概念的内涵进行了革新. 将“强化”分为正强化和负强化两种, 正强化促使有机体对刺激的反应概率增加, 负强化促使有机体消除该刺激的反应增加. 刺激产生反应, 反应影响刺激出现的概率, 这正是斯金纳操作条件反射理论的核心. Skinner 的这一理论对智能体的学习行为给出了清晰的描述, 吸引了很多学者对其进行研究. Zalama 等[2]基于 Grossberg 的条件反射模型研究了机器人的避障问题, 该模型借用经典条件反射理论中“条件刺激”和“非条件刺激”的概念, 以距离数据为条件刺激, 碰撞为非条件刺激, 使训练后的机器人能在无导师信号的情况下学会在任意位置的避障. 此后, Gaudiano 等[3-4]进一步发展了该模型, 将其与人工神经网络相结合, 应用在实物机器人 Pioneer 1 和 Khepera 上, 进行了避障方面的实验, 效果良好, 但是, 模型更侧重于对经典条件反射的建模, 对操作条件反射涉及较少. Ishii 等[5]为了研究动物与机器人之间的交互问题, 基于操作条件反射理论, 在机器人 WM-6 和老鼠之间对斯金纳老鼠实验进行了复现, 对比实验表明, 机器人与老鼠之间的交互加快了操作条件反射建立的速度, 提高

收稿日期: 2013-04-27; **修回日期:** 2013-07-30.

基金项目: 国家自然科学基金项目(61075110); 北京市自然科学基金项目(KZ201210005001); 国家 973 计划项目(2012CB720000); 高等学校博士学科点专项科研基金项目(20101103110007).

作者简介: 阮晓钢(1958−), 男, 教授, 博士生导师, 从事控制科学与工程、人工智能与认知科学、机器人学与机器人技术等研究; 黄静(1979−), 女, 博士生, 从事人工智能与认知科学、智能控制的研究.

了学习速率. Itoh 等[6]基于操作条件反射的 Hull 理论进行建模, 将其应用于仿人机器人 WE-4RII 上, 使机器人能从预先定义的行为列表中自动选择适合特定情境的行为模式. Taniguchi 等[7]提出了一种利用尖峰神经元网络实现的强化学习模式, 该模式对操作条件反射进行了建模, 使机器人建立起行为与听觉输入之间的联系, 实现了对机器人的语音控制. Salotti 等[8]提出了一种基于事件预测的模型, 该模型由 Rescorla 等的经典条件反射模型发展而来, 结合贝叶斯网络对其进行改进, 最终实现了对操作条件反射和经典条件反射的整合. 虽然以上学者都针对不同应用问题对操作条件反射进行了建模, 但没有给出统一的、形式化的数学模型. 阮晓钢等[9]和蔡建羡等[10]针对基于操作条件反射的自适应学习作了深入研究, 提出了一个形式化的、基于自动机的学习模型, 复现了 Skinner 的鸽子实验, 并将其应用于两轮自平衡机器人的运动控制问题. 但是, 针对该模型的应用主要局限于机器人平衡控制问题, 同时该模型在理论构建上存在缺陷, 所提出的取向函数与生物取向性在概念上不一致, 未能完全反映操作条件反射理论.

为了解决以上不足, 本文提出一个新的基于操作条件反射的学习模型, 并将其应用于移动机器人的避障问题, 再现了人或动物的自主学习行为. 仿真实验结果表明, 机器人在无导师信号的前提下能通过"试错式"的方式与环境交互, 建立操作条件反射, 最终成功避障到达目的地.

1　模型构建

基于操作条件反射的学习模型 (OCLM) 可以定义为一个九元组 $\mathrm{OCLM}=\langle t,S,A,P,f,\varepsilon,\delta,L,H\rangle$, 各元素含义解释如下:

1) t 为 OCLM 模型的时间参数. t 也用来表示模型迭代轮数, $t=\{t_i|i=0,1,\cdots,n_t\}$, t_0 表示模型初始建立的时间.

2) S 为 OCLM 模型的状态空间. $S=\{s_i|i=1,2,\cdots,n_s\}$. 其中: s_i 为状态空间中第 i 个状态; n_s 为状态空间大小, 即状态数目.

3) A 为 OCLM 模型的操作集合. $A=\{a_k|\,k=1,2,\cdots,n_a\}$. 其中: a_k 为 OCLM 模型的第 k 个操作, n_a 为操作行为的总数.

4) P 为 OCLM 模型的概率分布. $P:S\times A\to P=\{p_{ik}|\,i=1,2,\cdots,n_s,\ k=1,2,\cdots,n_a\}$, 其中 $p_{ik}=p(a_k|\,s_i)$ 表示在状态 $s_i\in S$ 下, OCLM 以概率 p_{ik} 选择动作 $a_k\in A$ 执行. P 也可用向量形式表示为 $P=\{P_1,P_2,\cdots,P_{n_s}\}$, 其中 $P_i=\{p_{i1},p_{i2},\cdots,p_{ir}\}\in P$ 为第 i 个状态对应的概率向量, r 为该状态所有可选择操作的个数, 显然有 $0<p_{ik}<1$, 且 $\sum_{k=1}^{r}p_{ik}=1$.

5) f 为 OCLM 模型的状态转移函数. $f:\ S\times A|_P\to S$ 表示状态 $s_i\in S$ 依概率 $p_{ik}\in P$ 选定某动作 $a_k\in A$ 后, 状态变化为 $s_j\in S$.

6) ε 为OCLM 模型中每个状态的负理想度. "负理想度"是本文为了计算取向函数、进而反映所感知到刺激是否为正强化而设立的概念, 记作 $\varepsilon=\varepsilon(S)=\{\varepsilon(s_i)|i=1,2,\cdots,n_s\}\in R$, 用来表征状态 s_i 远离理想状态的程度, 数值越大, 状态 s_i 相对设定目标越不理想.

7) δ 为 OCLM 模型中的取向函数. 取向函数 δ 模拟了自然界中生物的取向性, 表示为 $\delta=\delta(S,A)=\{\delta_{ik}|i=1,2,\cdots,n_s,\ k=1,2,\cdots,n_a\}$, 其中 δ_{ik} 为状态 $s_i\in S$ 执行动作 $a_k\in A$ 后系统性能的变化. 与生物取向性概念一致, $\delta>0$ 为正取向, 表明系统性能趋向变好; $\delta<0$ 为负取向, 表明系统性能趋向变差; $\delta=0$ 为零取向, 表明系统性能没有变化. 基于上述关于负理想度 ε 的定义, 取向函数 δ 的计算方法表示为

$$\delta_{ik}=\delta(\Delta\varepsilon_{ij})\begin{cases}>0,\ \Delta\varepsilon_{ij}<0;\\ =0,\ \Delta\varepsilon_{ij}=0;\\ <0,\ \Delta\varepsilon_{ij}>0.\end{cases}\tag{1}$$

其中 $\Delta\varepsilon_{ij}=\varepsilon(s_j)-\varepsilon(s_i)$. 取向函数 δ 在定义区间上连续, 为对 $\Delta\varepsilon_{ij}$ 单调递减函数, 其绝对值随 $\Delta\varepsilon_{ij}$ 绝对值单调递增. 当 $\Delta\varepsilon_{ij}>0$ 时, 负理想度增大, 系统性能趋向变差, 因此取向函数 $\delta<0$, 且 $\Delta\varepsilon_{ij}$ 越大, 取向函数 δ 越小; 当 $\Delta\varepsilon_{ij}<0$ 时, 负理想度变小, 系统性能趋向变好, 因此取向函数 $\delta>0$, 且 $\Delta\varepsilon_{ij}$ 越大, 取向函数 δ 越小; 当 $\Delta\varepsilon_{ij}=0$ 时, 负理想度不变, 系统性能趋向也不变化, 因此取向函数 $\delta=0$.

8) L 为 OCLM 模型基于操作条件反射原理的学习机制. $L:\ P(t)\to P(t+1)$ 表示根据操作条件反射原理调整模型的概率分布, 即"正强化时, 动作概率增加; 负强化时, 概率减少". 设 t 时刻状态 s_m 选择动作 a_k 执行, 感知到来自环境的刺激记为 θ, 同时状态转移到 s_n, 有:

① 若 θ 为正强化 ($\delta_{mk}>0$), 则当 $a(t)=a_k$ 时, 有

$$p_{mk}(t+1)=p_{mk}(t)+\frac{1-p_{mk}(t)}{1+\exp(-\eta_1\delta_{mk}t)};\tag{2}$$

当 $a(t)\neq a_k$ 时, 有

$$p_{mk'}(t+1)=p_{mk'}(t)-\frac{1-p_{mk}(t)}{1+\exp(-\eta_1\delta_{mk}t)}\frac{1}{n_a-1}.\tag{3}$$

② 若 θ 为负强化 ($\delta_{mk}<0$), 则当 $a(t)=a_k$ 时, 有

$$p_{mk}(t+1)=p_{mk}(t)-\frac{p_{mk}(t)}{1+\exp(\eta_2\delta_{mk}t)};\tag{4}$$

当 $a(t)\neq a_k$ 时, 有

$$p_{mk'}(t+1)=p_{mk'}(t)+\frac{p_{mk}(t)}{1+\exp(\eta_2\delta_{mk}t)}\frac{1}{n_a-1}.\tag{5}$$

③ 若θ为非强化刺激($\delta_{mk}=0$), 则概率保持不变, 即

$$L:\begin{cases} p_{mk}(t+1)=p_{mk}(t),\ a(t)=a_k; \\ p_{mk'}(t+1)=p_{mk'}(t),\ a(t)\neq a_k. \end{cases} \tag{6}$$

其中: $p_{mk}(t)$为t时刻状态s_m选择动作a_k执行的概率; η_1和η_2为学习速率, 且$\eta_1,\eta_2>0$.

9) H为OCLM模型的系统熵. 本文以"系统熵"H描述模型的自组织程度, 进而表明模型的自适应性. 系统熵$H(t)$用来描述t时刻OCLM模型的乱度, 为状态熵$HS(t)$的数学期望, 即

$$H_t=-\sum_{i=1}^{n_s}p(s_i)\sum_{k=1}^{n_a}p(a_k|s_i)\log_2 p(a_k|s_i). \tag{7}$$

显然, 系统熵H_t与t时刻系统的自组织程度成反比, H_t越小, 系统自组织程度越高.

OCLM模型的工作原理可以描述如下: 设t时刻, 系统状态s_i的负理想度为ε_i, 按概率p_{ik}从集合A中选择动作a_k执行, 根据函数f的定义, 状态随之转移到s_j, 负理想度为ε_j. 由此可计算出取向函数$\delta(\Delta\varepsilon)=\delta(\varepsilon_j-\varepsilon_i)$, 从而按照学习机制$L$对概率分布函数$P$进行调整. 经过多轮学习后, 系统熵$H$收敛达到最小, 系统自组织程度达到最大, 操作条件反射形成, 模型习得评价最优的动作. 算法流程如图1所示.

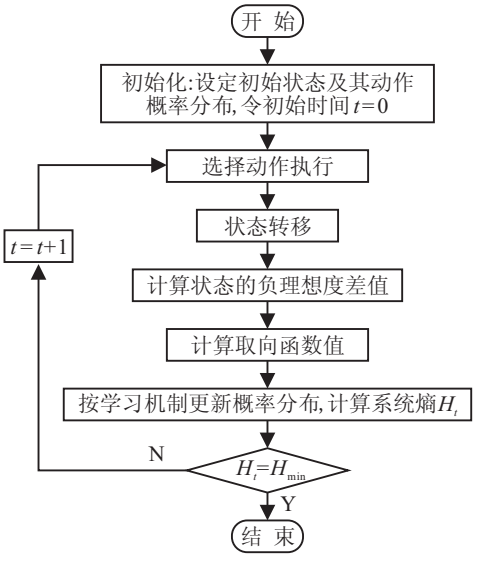

图1 OCLM模型算法流程

2 仿真实验

OCLM作为一种自组织系统, 具有一定自学习能力, 本节将OCLM用于机器人避障仿真实验, 令机器人在没有导师信号的情况下, 通过"试错式"的方式与环境交互, 建立操作条件反射, 完成避障及导航, 从而验证模型的有效性. 下文中所用符号含义与前相同, 不再赘述.

仿真实验基于一个圆形移动机器人进行, 该机器人半径为0.2 m, 周围均匀分布6个声纳传感器, 可通过发射和接受超声波测定前方障碍物距离, 其有效测量距离为15 cm～8 m(覆盖仿真实验环境); 行走机构采用双轮差动式运动底盘, 在机器人左右两侧安装轮w_L和w_R, 由直流伺服电机驱动, 尾部有一个起支撑作用的万向轮w_F, 前进速度设定为0.2 m/s. 该机器人机械结构简化示意图如图2所示(图中编号圆圈表示声纳传感器).

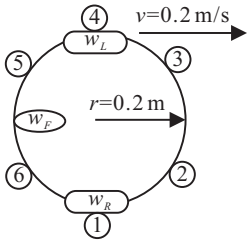

图2 机器人结构俯视图

实验中机器人面临的环境地图如图3所示. 机器人被放置在一个8 m×8 m大小的空间, 从出发点到目的地沿途设置了10个障碍. 为了增加实验的难度, 以更好地验证模型的有效性, 实验中对障碍的放置进行设计. 其中: 1号和3号障碍与目的地横坐标相同, 纵坐标不同; 2号和4号障碍与目的地纵坐标相同, 横坐标不同, 使得1号～4号障碍从4个方向包围目的地; 出于同样的目的, 5号和6号障碍放置在出发点的正右方和正下方(以坐标方向看); 7号～10号障碍散放在出发点和目的地的直线距离两侧.

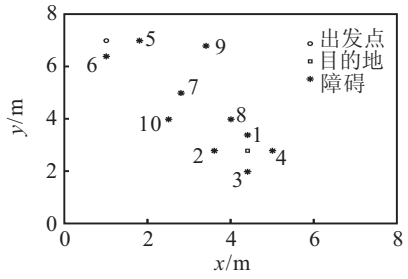

图3 机器人避障巡航环境地图

在本实验中, 机器人在t时刻所处状态s_t由其平面坐标决定, 即$s_i|t=(x_i,y_i)$, 其可能执行的操作为选择何种方向前进. 不失一般性, 在本实验中被简化为从1～6个传感器分布位置中选择相应角度前进, 即$A=\{a_k|k=1,2,\cdots,6\}$, a_k表示选择第k个传感器方向. 在每个状态初始时刻, 动作选择概率$p(a_k|s_i)$均匀分布. 当动作a_k被选择执行, 即发生状态转移时, 按下式完成:

$$x_{\rm new}=x_{\rm old}+vt_s\cos\theta_k,\ y_{\rm new}=y_{\rm old}+vt_s\sin\theta_k. \tag{8}$$

其中: v为机器人前进的线速度, t_s为采样间隔(本实验中设定为1 s), θ_k为第k个感知器在以机器人圆心为极点、前进方向为极轴建立的坐标系中所处位置的弧度值.

每个状态的负理想度$\varepsilon(s_i)$由其与目的地和障碍之间的距离共同决定, 距离目的地越远, 离障碍越近, 负理想度值越大. 因此, 设计负理想度计算公式为

$$\begin{aligned}&\text{if } \min(d_1,d_2,\cdots,d_{10})>r,\\&\quad \varepsilon(s_i)=w_1d_{\text{goal}}+w_2\exp(-\min(d_1,d_2,\cdots,d_{10}));\\&\text{else } \varepsilon(s_i)=100\,000+w_1d_{\text{goal}}+\\&\qquad w_2\exp(-\min(d_1,d_2,\cdots,d_{10})).\end{aligned}\tag{9}$$

其中: d_{goal}为机器人在该状态与目的地的距离, $d_k(\cdot)$为返回机器人与第k个障碍的距离值, r为机器人半径, w_1和w_2为修饰权重. 负理想度计算公式的含义是: 当机器人与周边障碍没有发生碰撞时, 负理想度正比于与目的地的距离, 反比于与障碍物的距离, 一旦发生碰撞, 负理想度在此基础上增加100 000.

取向函数可以定义为

$$\delta_{ik}=\delta(\Delta\varepsilon_{ij})=\begin{cases}\exp(1/\Delta\varepsilon_{ij}),\ \Delta\varepsilon_{ij}<0;\\0,\ \Delta\varepsilon_{ij}=0;\\-\exp(-1/\Delta\varepsilon_{ij}),\ \Delta\varepsilon_{ij}>0.\end{cases}\tag{10}$$

其中$\Delta\varepsilon_{ij}=\varepsilon_j-\varepsilon_i$, 即状态转移后的负理想度变化. 当选择动作使状态发生转移后, 引起负理想度变化, 此时可由取向函数对该动作进行评价, 进而根据式(2)～(6)调整动作概率.

学习以不断试错的方式进行, 机器人从6个方向中随机进行选择, 计算可能的位移, 进而计算出负理想度变化, 按操作条件反射理论完成动作概率修正, 更新状态熵. 当状态熵收敛至规定精度时, 学习结束, 机器人选择概率最大的动作执行, 直至抵达终点.

算法中有两组参数变量会对实验结果产生影响, 一组是负理想度计算公式中的修饰权重, 即w_1和w_2; 一组是正强化学习速率η_1和负强化学习速率η_2. 为了确定两组参数变量的取值, 设计一系列实验进行分析验证. 每组实验只对一组参数变量进行讨论, 保持其他变量值不变, 以程序运行时间(单位为s)和巡航途中与障碍碰撞次数两个数据说明参数变化对实验结果的影响. 同时, 为了排除偶然因素的影响, 两个数据都取程序运行100次后的平均值.

第1组实验$\eta_1=\eta_2=1$, 运行结果如表1所示. 由表1可见, 当修饰权重w_1和w_2相等并同步增长时, 平均运行时间并没有随之减少, 但是平均碰撞次数有减少的趋势, 表明修饰权重增加对减少碰撞次数有一定作用.

表 1　第 1 组实验运行结果

参数取值	平均运行时间/s	平均碰撞次数
$w_1=w_2=1$	3.909 6	1.20
$w_1=w_2=5$	12.647 5	0.76
$w_1=w_2=50$	4.046 5	0.76

第2组实验$\eta_1=\eta_2=1$, 运行结果如表2所示. 由表2可见, 当障碍距离权重w_2不变, 只增加目标距离权重w_1时, 平均运行时间增加, 而碰撞次数有缓慢减少的趋势.

表 2　第 2 组实验运行结果

参数取值	平均运行时间/s	平均碰撞次数
$w_1=5,w_2=1$	2.091 2	0.94
$w_1=10,w_2=1$	3.260 5	0.87
$w_1=100,w_2=1$	6.934 7	0.85

第3组实验$\eta_1=\eta_2=1$, 运行结果如表3所示. 由表3可见, 当目标距离权重w_1不变, 障碍距离权重w_2增加时, 较之前2组实验, 程序平均运行时间明显增加, 且w_2越大, 收敛越慢; 但是, 平均碰撞次数得到减少, 且w_2越大, 碰撞次数越少, 避障效果越好.

表 3　第 3 组实验运行结果

参数取值	平均运行时间/s	平均碰撞次数
$w_1=1,w_2=5$	103.777 7	0.51
$w_1=1,w_2=10$	1 170.650 0	0.10

以上3组实验数据表明, 在学习速率不变的情况下, 修饰权重增加对于程序运行时间有不利的影响, 类似正比的关系; 目标距离权重w_1和障碍距离权重w_2对碰撞次数均有影响, 但障碍距离权重w_2影响更为明显, 增加w_2可以明显减少碰撞次数.

综合收敛时间和避障效果的考虑, 对修饰权重设定值选择$w_1=5,w_2=1$的组合. 在此设定下, 再执行3组实验, 分别考察学习速率同步增长、正强化学习速率增长(负强化学习速率保持不变)和负强化学习速率增长(正强化学习速率保持不变)对实验结果的影响. 第4组实验$w_1=5,w_2=1$, 运行结果如表4所示. 第5组实验$w_1=5,w_2=1$, 运行结果如表5所示. 第6组实验$w_1=5,w_2=1$, 运行结果如表6所示.

表 4　第 4 组实验运行结果

参数取值	平均运行时间/s	平均碰撞次数
$\eta_1=\eta_2=5$	1.785 0	0.80
$\eta_1=\eta_2=10$	1.626 0	0.63
$\eta_1=\eta_2=100$	1.647 5	0.72

表 5　第 5 组实验运行结果

参数取值	平均运行时间/s	平均碰撞次数
$\eta_1=5,\eta_2=1$	2.834 2	0.89
$\eta_1=10,\eta_2=1$	2.563 4	0.71

表 6　第 6 组实验运行结果

参数取值	平均运行时间/s	平均碰撞次数
$\eta_1=1,\eta_2=5$	2.611 2	1.01
$\eta_1=1,\eta_2=10$	2.567 4	0.89

由表4～表6可见, 增大学习速率可以加快程序收敛速度, 减少运行时间, 同时还可以减少碰撞次数,

但当学习速率过大时(如100), 运行时间和碰撞次数均有所回升. 在收敛速度上, 两种学习速率的影响无明显区别; 在避障效果上, 增大正强化学习速率比增大负强化学习速率有效.

综上, 两组参数值最终确定为 $w_1 = 5$, $w_2 = 1$, $\eta_1 = \eta_2 = 10$, 在此情况下执行程序, 效果如图4所示. 由图4可见, 机器人从出发点出发, 基于操作条件反射理论对周围环境进行认知, 成功避开障碍到达目的地.

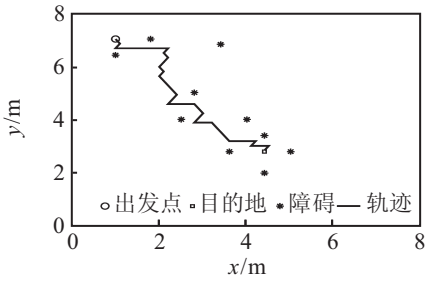

图 4 机器人避障巡航轨迹

图5为一次学习过程中系统熵的变化情况. 由图5可见, 模型经30次学习后达到收敛, 系统熵收敛至0, 表明机器人已习得最优动作, 建立起操作条件反射, 同时系统自组织程度也达到最大. 避障导航的过程是机器人自学习、自组织、自适应的过程, 在这个过程中, 自学习是手段, 自组织是表现, 自适应才是目的.

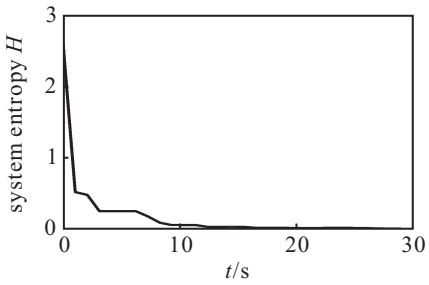

图 5 学习过程中系统熵变化情况

3 结 论

受Skinner操作条件反射原理启发, 本文提出了一种具有仿生自主学习能力的学习模型, 该模型采用九元组进行描述, 其核心是学习机制的设计和定义, 该学习机制以“正强化操作概率增加, 负强化操作概率减少”为原则调整各状态操作行为概率分布. 提出了状态的“负理想度”的概念, 并由此计算取向函数. 取向函数与生物取向性概念一致, 通过取向函数值可区分出正强化和负强化操作, 最终完成学习机制的定义. 将模型用于机器人避障导航仿真实验, 对参数设置进行系列讨论. 实验结果表明, 所提出模型能很好地模拟人和动物的操作条件反射行为, 使机器人经过学习训练后能够自主避开障碍到达目的地, 表现出较强的自学习能力, 具有一定理论研究和工程应用价值.

参考文献(References)

[1] Skinner B F. The behavior of organisms: An experimental analysis[M]. New York: Appleton-Century Company, 1938: 110-150.

[2] Zalama E, Gaudiano P, Coronado J L. Obstacle avoidance by means of an operant conditioning model[M]. Berlin: Springer, 1995: 471-477.

[3] Gaudiano P, Chang C. Adaptive obstacle avoidance with a neural network for operant conditioning: experiments with real robots[C]. IEEE Int Symposium on Computational Intelligence in Robotics and Automation. Monterey: IEEE Press, 1997: 13-18.

[4] Gaudiano P, Zalama E, Chang C, et al. A model of operant conditioning for adaptive obstacle avoidance[C]. From Animals to Animats. Cambridge: MIT Press, 1996: 373-381.

[5] Ishii H, Nakasuji M, Ogura M, et al. Accelerating rat's learning speed using a robot: The robot autonomously shows rats its functions[C]. Proc of the 2004 IEEE Int Workshop on Robot and Human Interactive Communication. Roman: IEEE Press, 2004: 229-234.

[6] Itoh K, Miwa H, Matsumoto M, et al. Behavior model of humanoid robots based on operant conditioning[C]. The 5th IEEE-RAS Int Conf on Humanoid Robots. Tsukuba: IEEE Press, 2005: 220-225.

[7] Taniguchi T, Sawaragi T. Incremental acquisition of behaviors and signs based on a reinforcement learning schemata model and a spike timing-dependent plasticity network[J]. Advanced Robotics, 2007, 21(10): 1177-1199.

[8] Salotti J M, Lepretre F. Classical and operant conditioning as roots of interaction for robots[C]. Proc of the Workshop From Motor to Interaction Learning in Robots Conf on Intelligent Robotics Systems. Nice: Springer, 2008: 124-133.

[9] 阮晓钢, 蔡建羡, 戴丽珍. 基于概率自动机的操作条件反射计算模型[J]. 北京工业大学学报, 2010, 36(8): 1025-1030.
(Ruan X G, Cai J X, Dai L Z. Compute model of operant conditioning based on probabilistic automata[J]. J of Beijing University of Technology, 2010, 36(8): 1025-1030.)

[10] 蔡建羡, 阮晓钢. OCPA仿生自主学习系统及在机器人姿态平衡控制上的应用[J]. 模式识别与人工智能, 2011, 24(1): 138-146.
(Cai J X, Ruan X G. OCPA bionic autonomous learning system and its application to robot poster balance control[J]. Pattern Recognition and Artificial Intelligence, 2011, 24(1): 138-146.)

(责任编辑: 郑晓蕾)