文章编号: 1000−8152(2012)11−1452−06

一种自治操作条件反射自动机

阮晓钢, 戴丽珍, 于乃功, 于建均

(北京工业大学 电子信息与控制工程学院 人工智能与机器人研究所, 北京 100124)

摘要: 针对仿生自主学习控制问题, 根据自动机的原理, 以操作条件反射学习机制为基础, 运用仿生的自组织学习方法, 提出一种自治操作条件反射自动机(autonomous operant conditioning automata, AOCA)模型, 主要包括: 操作集合、状态集合、“条件–操作”规则集合、可观测的状态转移以及操作条件反射学习律; 定义了基于AOCA状态取向值的操作熵; 给出了AOCA操作熵收敛性证明; 分析了AOCA自组织特性; 规定了AOCA的递归运行程序. 同时, 将其应用于斯金纳动物实验的模拟, 动物分阶段学习, 并且成功习得技能, 实验结果表明AOCA实现了模拟操作条件反射学习机制.

关键词: 自动机理论; 自治; 操作条件反射; 仿生学; 自主学习

中图分类号: TP13 **文献标识码**: A

An autonomous operant conditioning automaton

RUAN Xiao-gang, DAI Li-zhen, YU Nai-gong, YU Jian-jun

(Institute of Artificial Intelligence and Robots, College of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China)

Abstract: To deal with the bionic self-organization learning-control problem, on the basis of the automata principle and the operant conditioning learning mechanism, we make use of the bionic self-organization learning method to build an autonomous operant conditioning automaton (AOCA) model including the action set, the state set, the condition-action rule set, the observed state transition, the operant conditioning learning law as well as a recursive program. We also define the operant entropy based on the orientation values of states in the AOCA model, prove the convergence of the AOCA operant entropy, analyze the self-organization characteristics and develop the recursive operation program for the AOCA. The AOCA model has been applied to simulate the Skinner animal experiment, in which the animal learns the task in stages and succeeds in acquiring the skills eventually. These experiment results demonstrate that the AOCA can realize the learning mechanism of operant conditioning.

Key words: automata theory; autonomy; operant conditioning; bionics; autonomous learning

1 引言(Introduction)

在学习模型中, 自动机是一个十分典型的例子, 已成功地应用于优化问题[1–3]、博弈[4]、模式识别[5]、图像处理与计算机视觉[6]、智能控制[7]等众多不同的领域之中, 但对于自组织模型的研究却相对较少. 从变量集特性的角度来看, 自动机可划分为自治式和非自治式自动机. 与非自治式自动机相比, 自治式自动机的输出不需要外部指令的驱动, 是自动机根据自身的需要而做出的作用于环境的某种行动. 也就是说, 即使外部环境发生改变, 自治式自动机仍然可以照常工作, 而非自治式自动机则需要通过结构模型或参数的改变来适应外部环境的变化.

众所周知, 非自治系统总是可以通过相应的操作从而转化为一个自治系统, 也就是说总可以找到一个自治操作条件反射自动机与相应的非自治操作条件反射自动机相对应. 因此, 自治操作条件反射自动机具有更为广泛的应用前景.

人和动物的诸多技能或行为是在其神经系统自学习和自适应的过程中渐进地形成和发展起来的. 理解与模拟人和动物神经系统内在的学习和组织机制, 并将这种机制赋予机器, 是控制科学、人工智能和机器人学研究的重要课题. 操作条件反射(operant conditioning, OC)是心理学中条件反射的一种重要形式, 也是人和动物神经系统内在的一种重要的学习机制[8–9]. 操作条件反射这一概念由美国哈佛大学心理学家斯金纳于1938年提出, 又称为斯金纳的操作条件反射理论, 是针对人的行为控制而提出的.

在过去的几年中, 伴随着人工智能的发展, 斯金纳操作条件反射理论的应用特别是在机器人理论中的应用取得了长足的进展. 20世纪90年代美国卡

收稿日期: 2011−08−15; 收修改稿日期: 2012−07−20.

基金项目: 国家自然科学基金资助项目(61075110); 北京市自然科学基金资助项目/北京市教育委员会科技计划重点资助项目(KZ201210005001); 北京市自然科学基金资助项目(4102011); 高等学校博士学科点专项科研基金资助项目(20101103110007).

内基梅隆大学的Touretzky等[10–12]开始研究并设计了一种基于操作条件反射理论的计算模型用来进行动物学习的实验, 并成功将其应用于RWIB21移动机器人的学习过程. Daw等[13–14]在巴普洛夫的经典条件反射和斯金纳的操作条件反射理论基础之上提出了基于时间差分的经典条件反射学习模型, 并以Sony AIBO机器人为研究平台进行算法应用研究, 实现了在操作条件反射模型上的人机互动. 2005年Itoh等[15]基于操作条件反射理论设计了一个具有人特点的机器人的行为模型, 并将该模型应用于人形机器人WE-4RII, 使人形机器人WE-4RII能够自主地在事先定义好的行为中选取合适的行为. 以上这些模型是基于操作条件反射理论的计算模型, 具有仿生的特性, 但没有形式化和具体化. 2010年, Sokolowski等[16]采用操作条件反射的原理设计了一套类似于斯金纳箱的设备用于模拟蜜蜂学习采蜜, 但是没有给出操作条件反射模型. 本文作者前期研究中提出了一种基于概率自动机的操作条件反射计算模型[17], 用于模拟斯金纳鸽子试验, 给出实验结果, 但没有分析自组织特性, 并且学习速度较慢; 同时, 在概率自动机的基础上引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[18], 并将其应用于两轮机器人的自平衡控制中, 取得了较好的控制效果.

鉴于此, 本文在自动机的理论中引入斯金纳操作条件反射理论, 提出了一个抽象的自组织模型: 一种自治操作条件反射自动机(autonomous operant conditioning automata, AOCA), 并对其自组织特性进行分析. 采用本文提出的AOCA模型不但可以模拟生物的操作条件反射机制, 还可以描述、模拟、设计各种自组织系统. 通过对斯金纳动物实验的模拟, 证明了AOCA模型的有效性, 体现了其自学习和自适应特性.

2 斯金纳操作条件反射(Skinner operant conditioning)

2.1 斯金纳操作条件反射理论(Theory of skinner operant conditioning)

操作条件作用的学习理论是由美国心理学家斯金纳根据动物学习实验提出的[8–9]. 斯金纳通过研究发现, 有机体做出的反应与其随后出现的结果对行为起着控制作用, 它能影响以后反应发生的概率. 操作条件反射是指在一定的刺激情境中, 有机体的某种反应结果能满足其某种需要, 以后在相同的情境中其反应概率就会提高的现象. 这种有机体自发操作的行为是主动的, 代表着有机体对环境的主动适应. 同时, 操作性行为导致了操作性条件反射. 强化是一个用来描述反应概率增加的术语, 即强化增加的是反应发生的概率, 如何安排强化是学习的核心之所在. 其核心在于动物通过学习或训练, 其小脑皮质的神经组织发生改变, 使得某种感知序列和动作序列联系起来, 即从“感知–动作”不断循环, 如图1所示.

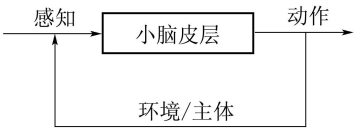

图 1 操作条件反射机制

Fig. 1 Operant conditioning mechanism

2.2 斯金纳操作条件反射实验(Experiment of skinner operant conditioning)

斯金纳用动物做实验的装置称为斯金纳箱(Skinner box), 如图2所示. 该装置是一个特殊条件的控制箱, 箱内隔光、隔音, 并装有自动控制和记录的光、声系统及一套杠杆和喂食器. 只要在箱内按下杠杆, 喂食器就自动供给食丸. 在进行实验时, 首先调试好实验装置, 保证白鼠偶尔按压杠杆后即可获得一粒食丸. 白鼠在刚进入控制箱的时候会到处乱跑并向四周攀附, 当它偶然触压杠杆时便可得到一粒食丸: 在几次偶尔的触压杠杆后, 白鼠会频繁地按压杠杆, 每次都会获得一粒食丸. 食物奖赏强化了白鼠按压杠杆的行为, 并减弱了其他行为(如在箱子四周乱转). 每次实验都对白鼠的行为进行记录.

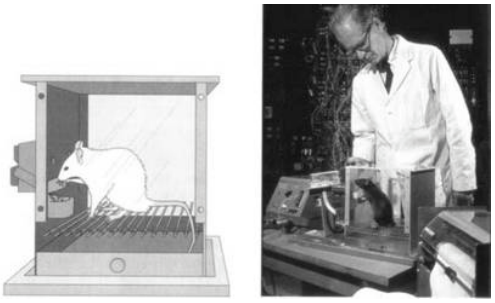

图 2 斯金纳箱

Fig. 2 Skinner box

3 自治操作条件反射自动机模型(Autonomous operant conditioning automata model)

AOCA是一种描述自治式自动机器的离散计算模型, 主要由操作集合、状态集合、“条件–操作”规则集合、可观测的状态转移、操作条件反射学习律, 以及基于AOCA状态取向值的操作熵组成.

3.1 AOCA模型定义(Definition of AOCA model)

本文设计的具有自组织(包括自学习和自适应)功能的自治操作条件反射自动机AOCA结构如图3所示. 该自动机由9元组表示, 即: AOCA$=<t, \Omega, S, \Gamma, \delta, \varepsilon, \eta, \psi, s_0>$. 定义如下:

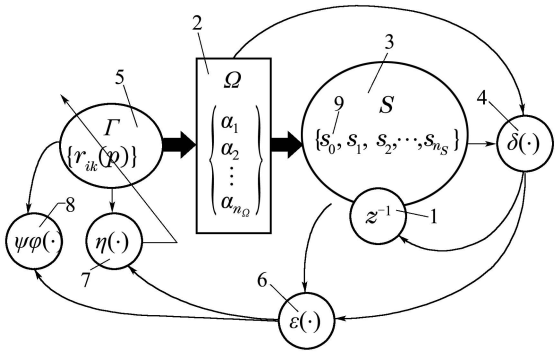

图 3 AOCA的结构示意图

Fig. 3 Structure of autonomous operant conditioning automata

1) AOCA的离散时间: $t \in \{0,1,\cdots,n_t\}$, $t=0$为AOCA的起始时刻.

2) AOCA的操作符号集合: $\Omega = \{\alpha_k | k = 1,2,\cdots,n_\Omega\}$, α_k为AOCA的第k个操作符号.

3) AOCA的状态集合: $S = \{s_i | i = 0,1,2,\cdots,n_S\}$, s_i为AOCA的第i个状态.

4) AOCA的状态转移函数: $\delta : s(t) \times \alpha(t) \to s(t+1)$.

AOCA在$t+1$时刻的状态$s(t+1) \in S$由t时刻的状态$s(t) \in S$和t时刻的操作$\alpha(t) \in \Omega$共同确定, 与其t时刻之前的状态和操作无关. δ所确定的状态转移过程既可以是已知的也可以是未知的, 但其状态转移的结果是能够观测的.

5) AOCA的操作规则集合: $\Gamma = \{r_{ik}(p) | p \in P;\ i \in \{0,1,2,\cdots,n_S\};\ k \in \{1,2,\cdots,n_\Omega\}\}$.

随机"条件–操作"规则$r_{ik}(p) : s_i \to \alpha_k(p)$意味着AOCA在其状态处于$s_i \in S$的条件下依概率$p \in P$实施操作$\alpha_k \in \Omega$, $p = p_{ik} = p(\alpha_k|s_i)$即AOCA在状态处于$s_i$的条件下实施操作$\alpha_k$的概率值, P表示p_{ik}的集合, 称为状态转移概率矩阵.

6) AOCA的状态取向函数: $\varepsilon : S \to E = \{\varepsilon_i | i = 0,1,2,\cdots,n_S\}$. 其中$\varepsilon_i = \varepsilon(s_i) \in E$为状态$s_i \in S$的状态取向值.

7) AOCA的操作条件反射学习律: $\eta : \Gamma(t) \longmapsto \Gamma(t+1)$.

学习律用于调节操作规则$r_{ik}(p) \in \Gamma$的实施概率$p \in P$.

假设t时刻AOCA处于$s_i \in S$条件下的状态为$s(t) = s_i$, 当前实施的选择操作$\alpha(t) = \alpha_k \in \Omega$; $t+1$时刻观测到的状态$s(t+1) = s_j$. 则按照斯金纳的操作条件反射理论可得, 如果$\varepsilon(s(t+1)) - \varepsilon(s(t)) < 0$, 则$p(\alpha(t)|s(t))$倾向于减小; 反之, 如果$\varepsilon(s(t+1)) - \varepsilon(s(t)) > 0$则$p(\alpha(t)|s(t))$倾向于增大.

假设t时刻实施操作的概率值为$p_{ik}(t)$; 则根据状态转移函数进行更新之后$t+1$时刻AOCA实施操作$\alpha_k \in \Omega$的概率值则更新为$p_{ik}(t+1)$.

模拟生物的操作条件反射机制可得$t+1$时刻当前操作的概率将会改变, 即

$$p_{ik}(t+1) = p_{ik}(t) + \Delta,$$

其中: $\Delta = \varphi(\vec{\varepsilon}_{ij}) = a \times \vec{\varepsilon}_{ij} \times (1 - p_{ik}(t))$且满足$0 \leqslant p_{ik}(t) + \Delta \leqslant 1$. Δ与状态取向函数ε有关, 其中: 状态转移取向函数为

$$\vec{\varepsilon}_{ij} = \varepsilon(s_j) - \varepsilon(s_i),$$

是状态取向值的增量; $\varphi(x)$是单调增函数, 同时满足当且仅当$x = 0$时$\varphi(x) = 0$; $a \in (0,1)$是学习率. 由状态取向函数的定义可知, ε越大表明操作的结果越好, 同时Δ也越大. 由于在各个时刻AOCA处于$s_i \in S$的条件下的实施操作$\alpha_k \in \Omega$的概率之和为1, 即

$$\sum_{k=1}^{n_\Omega} p_{ik}(a_k(t)\,|s_i(t)\,) = 1, \tag{1}$$

则$t+1$时刻其他操作的概率将作相应的变化, 即

$$p_{iu}(t+1) = p_{iu}(t) - \Delta \times \zeta,$$

其中: $u \in [1, n_\Omega]$且$u \neq k$, $\zeta = p_{iu}(t) / \sum\limits_{u=1,u\neq k}^{n_\Omega} p_{iu}(t)$. $\sum\limits_{u=1,u\neq k}^{n_\Omega} p_{iu}(t)$表示$t$时刻AOCA状态处于$s_i \in S$的条件下实施操作$\alpha_u \in \Omega$的概率之和.

8) AOCA的操作熵: $\psi : P \times E \to \mathbb{R}^+$.

其中: $\mathbb{R}^+$是正的实数集, AOCA在t时刻的操作熵$\psi(t)$表示t时刻状态处于s_i条件下的操作熵之和, 即

$$\psi(t) = \psi(\Omega(t)|S) = \sum_{i=0}^{n_S} p_i \psi_i(t) = \sum_{i=0}^{n_S} p(s_i)\psi_i(\Omega(t)|s_i). \tag{2}$$

$\psi(t)$由t时刻处于状态$s(t) = s_i$条件下的状态转移概率矩阵和取向函数集合决定. $\psi_i(t)$是AOCA处于状态s_i条件下的操作熵

$$\psi_i(t) = \psi_i(\Omega(t)|s_i) = -\sum_{k=1}^{n_\Omega} p_{ik} \log_2 p_{ik} = -\sum_{k=1}^{n_\Omega} p(\alpha_k|s_i) \log_2 p(\alpha_k|s_i). \tag{3}$$

将式(3)代入式(2)即可得AOCA在t时刻的操作熵

$$\psi(t) = -\sum_{i=0}^{n_S} p(s_i) \sum_{k=1}^{n_\Omega} p(\alpha_k|s_i) \log_2 p(\alpha_k|s_i). \tag{4}$$

9) AOCA的起始状态: $s_0 = s(0) \in S$.

3.2 程序步骤(Program steps)

自治操作条件反射自动机AOCA的工作步骤如下所示:

Step 1 初始化.

令$t = 0$, 并随机给定AOCA的初始状态$s(0)$, 学

习率a, 给定初始操作概率$p_{ik}(0)=1/n_\Omega(i=0,1,2,\cdots,n_S,\ k=1,2,\cdots,n_\Omega)$; 给定停机时间$T_{\rm f}$.

Step 2　选择操作.

依操作集合Γ中的“条件–操作”规则$r_{ik}(p):s_i\to\alpha_k(p)$进行选择操作.

Step 3　实施操作.

t时刻, AOCA处于状态$s(t)\in S$实施上一步已选中的操作$\alpha(t)\in\Omega$, 当前状态发生转移$\delta(s(t),\alpha(t))=\delta(s_i,\alpha_k)$.

Step 4　观测状态.

依AOCA的状态转移函数$\delta:s(t)\times\alpha(t)\to s(t+1)$, 状态转移的结果是完全能够观测的, 即存在$j\in\{0,1,2,\cdots,n_S\}$使得$s(t+1)=s_j$.

Step 5　操作条件反射.

在t时刻实施操作, 不仅AOCA的状态发生转移, 它的各个操作在下一时刻的实施概率也发生改变, 则依操作条件反射学习律$\eta:\Gamma(t)\longmapsto\Gamma(t+1)$调节操作规则$r_{ik}(p)\in\Gamma$的实施概率$p\in P$. t时刻$s(t)=s_i$且$\alpha(t)=\alpha_k$; 那么$t+1$时刻的操作概率依操作条件反射学习律η进行更新.

$$\eta:\begin{cases}p_{ik}(t+1)=p_{ik}(t)+\Delta,\\ p_{iu}(t+1)=p_{iu}(t)-\Delta\times\zeta,\ u\neq k.\end{cases}\tag{5}$$

Step 6　计算操作熵.

根据定义的操作熵的式(4)计算t时刻的操作熵, 其中: $p(s_i)$是AOCA状态$s_i\in S$的出现概率在t时刻的值, $p(\alpha_k|s_i)$是AOCA状态处于$s_i\in S$的条件下实施操作$\alpha_k\in\Omega$的概率在t时刻的值.

Step 7　递归转移.

如果$t+1\leqslant T_{\rm f}$, 那么$t=t+1$时刻重复Step 2–Step 7; 否则, 停机.

3.3 AOCA自组织特性分析(Self-organization characteristics analysis of AOCA)

根据操作熵的定义可知, 在每个状态下的操作熵已知的情况下再对其进行加权求和即可得出AOCA在t时刻的操作熵. 如果AOCA的操作熵$\psi(t)$随着时间的增长越来越小, 并且在$t\to\infty$时趋向于最小, 那么就可以说明AOCA操作熵$\psi(t)$是收敛的.

AOCA是一个基于斯金纳操作条件反射理论的自组织系统, 具有自学习和自适应功能. 系统自组织的过程是吸取信息的过程, 是吸取负熵的过程, 是消除不确定性的过程. 下文将通过对AOCA操作熵$\psi(t)$的收敛性分析, 说明AOCA的自组织特性.

定义 1　设AOCA$=<t,\Omega,S,\Gamma,\delta,\varepsilon,\eta,\psi,s_0>$是一个自治操作条件反射自动机, 其状态转移过程$\delta:s(t)\times\alpha(t)\to s(t+1)$是确定性的, $s_i\in S$, $\{k=1,2,\cdots,n_\Omega\}$, 则

$$\alpha_k\in\begin{cases}\Omega_i^-,\ \text{iff}\ \vec{\varepsilon}_{ik}<0,\\ \Omega_i^0,\ \text{iff}\ \vec{\varepsilon}_{ik}=0,\\ \Omega_i^+,\ \text{iff}\ \vec{\varepsilon}_{ik}>0,\end{cases}$$

其中$\Omega_i^-,\Omega_i^0,\Omega_i^+\in\Omega$分别称为$s_i$的负取向操作集合、零取向操作集合和正取向操作集合; 同时满足$p(\Omega_i^-\cup\Omega_i^0\cup\Omega_i^+|s_i)=1$.

定义 2

$$\begin{cases}\vec{\varepsilon}_i^{\,-}=\vec{\varepsilon}(\Omega_i^-)=\sum\limits_{\alpha_k\in\Omega_i^-}\vec{\varepsilon}_{ik},\\ \vec{\varepsilon}_i^{\,+}=\vec{\varepsilon}(\Omega_i^+)=\sum\limits_{\alpha_k\in\Omega_i^+}\vec{\varepsilon}_{ik},\end{cases}$$

其中$\vec{\varepsilon}_i^{\,-}$和$\vec{\varepsilon}_i^{\,+}$分别称为$\Omega$关于$s_i$的负取向值和正取向值.

定义 3

$$\begin{cases}p_i^-=p(\Omega_i^-|s_i)=\sum\limits_{\alpha_k\in\Omega_i^-}p_{ik},\\ p_i^0=p(\Omega_i^0|s_i)=\sum\limits_{\alpha_k\in\Omega_i^0}p_{ik},\\ p_i^+=p(\Omega_i^+|s_i)=\sum\limits_{\alpha_k\in\Omega_i^+}p_{ik},\end{cases}$$

其中p_i^-,p_i^0,p_i^+分别为AOCA处于状态s_i条件下$\Omega_i^-,\Omega_i^0,\Omega_i^+$中操作被激发的概率.

定义 4　令$\xi(\vec{\varepsilon}_{ij})=\dfrac{a\times\vec{\varepsilon}_{ij}}{1-a\times\vec{\varepsilon}_{ij}}$, $\sigma_{ij}=1-a\times\vec{\varepsilon}_{ij}$, 则激发函数可写成如下形式:

$$\eta:\begin{cases}p_{ik}(t+1)=p_{ik}(t)+\sigma_{ij}\xi(\vec{\varepsilon}_{ij})(1-p_{ij}(t)),\\ p_{iu}(t+1)=\sigma_{ij}p_{iu}(t),\ u\neq k,\end{cases}$$

其中: $\xi(\vec{\varepsilon}_{ij})\geqslant-p_{ij}(t)$, $\sigma_{ij}=1/(1+\xi(\vec{\varepsilon}_{ij}))>0$.

引理 1　设AOCA$=<t,\Omega,S,\Gamma,\delta,\varepsilon,\eta,\psi,s_0>$是一个自治操作条件反射自动机, 其状态转移过程$\delta:s(t)\times\alpha(t)\to s(t+1)$是确定性的, 给定$i\in\{0,1,2,\cdots,n_S\}$, 设$p(s_i)\neq0$, 并且$\Omega_i^+(t=0)\neq\varnothing$, 则$t\to\infty$时$p_i^+=p(\Omega_i^+|s_i)=1$成立. 即: 当$t\to\infty$时, AOCA依概率1选取正取向状态转移操作.

注 1　$\Omega_i^+(t=0)\neq\varnothing$意味着存在$\alpha_k$使得$\xi(\vec{\varepsilon}_{ik})>0$, 并且保证$p_{ik}(t=0)\neq0$.

证　令l为s_i出现次数的序号, $p_i^+(l)$为状态s_i第l次出现时集合Ω_i^+中的随机操作被选中的概率, 初始概率为$p_i(0)$.

因为$\Omega_i^+(t=0)\neq\varnothing$, 所以$p_i(0)\neq0$; 又$p_i^+(l)\geqslant0$, 故对于$\forall l=0,1,2,\cdots$有$p_i^+(l)\neq0$.

因为$p(s_i)\neq0$, 所以当$t\to\infty$时, AOCA状态s_i出现的频次也趋于无穷; 同时, $p_i^+(0)\neq0$使得Ω_i^+被选中的频次也趋于无穷.

假定AOCA在第l次处于状态s_i的条件下选择操作$\alpha_k\in\Omega$, 则

1)　若$\alpha_k\in\Omega_i^-$, 则$\vec{\varepsilon}_{ik}<0$, $\xi(\vec{\varepsilon}_{ik})<0$, $\sigma_{ik}=$

$1/(1+\xi(\vec{\varepsilon}_{ik}))>1$, 所以

$$\Delta p_i^+(l)=p_i^+(l+1)-p_i^+(l)=$$
$$\sigma_{ik}p_i^+(l)-p_i^+(l)=$$
$$(\sigma_{ik}-1)p_i^+(l)>0.$$

2) 若$\alpha_k\in\Omega_i^0$, 则$\vec{\varepsilon}_{ik}=0$, $\xi(\vec{\varepsilon}_{ik})=0$, $\sigma_{ik}=1/(1+\xi(\vec{\varepsilon}_{ik}))=1$, 所以

$$\Delta p_i^+(l)=p_i^+(l+1)-p_i^+(l)=0.$$

3) 若$\alpha_k\in\Omega_i^+$, 则$\vec{\varepsilon}_{ik}>0$, $\xi(\vec{\varepsilon}_{ik})>0$, $\sigma_{ik}=1/(1+\xi(\vec{\varepsilon}_{ik}))>0$, 所以

$$\Delta p_i^+(l)=p_i^+(l+1)-p_i^+(l)=$$
$$\sigma_{ik}\xi(\vec{\varepsilon}_{ik})(1-p_i^+(l))\geqslant 0,$$

其中, 当且仅当$p_i^+(l)=1$时$\Delta p_i^+(l)=0$成立.

综上所述, $\Delta p_i^+(l)\geqslant 0$.

当$t\to\infty$时, Ω_i^+被选中的频次也趋于无穷, 即: 当$t\to\infty$时, $\Delta p_i^+(l)>0$的情形可发生任意多次. 因此, p_i^+将不断升高. 又p_i^+有上界, 且上界为1, 故p_i^+的增长将直至$p_i^+=1$为止.

由信息熵的性质可知, 当所有操作行为$\alpha_k(t)$可能出现的概率$p_k(t)$相等时, 操作熵$\psi(t)$最大. 所以, 一般情况下在学习的初始时刻所有操作行为$\alpha_k(t)$设定为相同的操作概率值$p_k(t)$.

对式(3)重新进行整理可得

$$\psi_i(t)=\psi_i(\Omega(t)|s_i)=-\sum_{k=1}^{n_\Omega}p_{ik}(t)\log_2 p_{ik}(t)=$$
$$-p_{ik}(t)\log_2 p_{ik}(t)-\sum_{v=1,v\neq k}^{n_\Omega}p_{ik}(t)\log_2 p_{ik}(t). \quad (6)$$

随着学习的进行, 概率函数$p_{ik}(t)$被更新, 根据引理1可知$p_{ik}(t)$最终趋向于极大值1, 而$p_{iv}(t)$最终趋向于极小值0, 即

$$\begin{cases}\lim\limits_{t\to\infty}p_{ik}(t)\to 1,\\ \lim\limits_{t\to\infty}p_{iv}(t)\to 0.\end{cases} \quad (7)$$

将式(7)代入式(6), 整理可得

$$\lim_{t\to\infty}\psi_i(t)=$$
$$\lim_{t\to\infty}[-p_{ik}(t)\log_2 p_{ik}(t)-\sum_{v=1,v\neq k}^{n_\Omega}p_{ik}(t)\log_2 p_{ik}(t)]=$$
$$-\lim_{t\to\infty}p_{ik}(t)\log_2 p_{ik}(t)-$$
$$\lim_{t\to\infty}\sum_{v=1,v\neq k}^{n_\Omega}p_{iv}(t)\log_2 p_{iv}(t)\to 0$$

证毕.

4 仿真实验(Simulation experiment)

AOCA能够模拟生物的操作条件反射机制, 具有仿生的自主学习能力. 本节在一个最小系统即具有学习能力的小白鼠的基础上用AOCA来模拟斯金纳小白鼠实验, 证明该模型的有效性. 除特殊说明, 该实验中的符号与AOCA定义相同.

该实验中的小白鼠有两个操作行为: 压杠杆α_1和不压杠杆α_2, 即操作集合$\Omega=\{\alpha_1,\alpha_2\}$, 其概率分别用$p_1$, p_2表示; 状态集合$S=\{s_0,s_1\}$, 其中: s_0表示饥饿状态, s_1表示非饥饿状态; 操作规则集合$\Gamma=\{r_{ik}(p)|p\in P;\ i\in\{0,1\};\ k\in\{1,2\}\}$. 其状态转移函数$\delta: s(t)\times\alpha(t)\to s(t+1)$, 具体情况如下:

$$s_0\times\alpha_1\to s_1,\ s_0\times\alpha_2\to s_0,$$
$$s_1\times\alpha_1\to s_1,\ s_1\times\alpha_2\to s_0.$$

其状态取向函数为$\varepsilon: S\to E=\{\varepsilon_i|i=0,1\}=\{0,1\}$. 具体如表1所示.

表 1 状态转移规则和取向机制

Table 1 State transition rules and orientation mechanism

δ	α_1	α_2
s_0	$s_1(p_{01}),\vec{\varepsilon}_{01}=1$	$s_0(p_{02}),\vec{\varepsilon}_{02}=-1$
s_1	$s_1(p_{11}),\vec{\varepsilon}_{11}=0$	$s_0(p_{12}),\vec{\varepsilon}_{12}=0$

本实验给定学习率$a=0.01$, 初始时刻两个行为的概率均为0.5. 实验按照AOCA的程序步骤递归地运行和实施.

实验中只要小白鼠触压杠杆就能获得奖赏, 通过学习其触压杠杆的概率就会增加. 即下一时刻小白鼠选择压杠杆的可能性增加, 其概率依操作条件反射学习律$\eta: \Gamma(t)\to\Gamma(t+1)$更新, 经过反复不断地学习, 小白鼠选择压杠杆的概率p_1越来越大. 在初始阶段, 小白鼠选择压杠杆的概率和不压杠杆的概率一样, 经过160步的学习, 小白鼠压杠杆的概率达到90%; 经过300步的学习, 小白鼠基本上完全学会触压杠杆获取食物, 如图4所示, 小白鼠压杠杆的概率p_1最终趋向于1. 在实验过程中, 根据定义的操作熵公式(4)计算出了每个时刻的操作熵, 随着时间的推移AOCA的操作熵$\psi(t)$越来越小并且在$t\to\infty$时趋向于最小, 如图5所示.

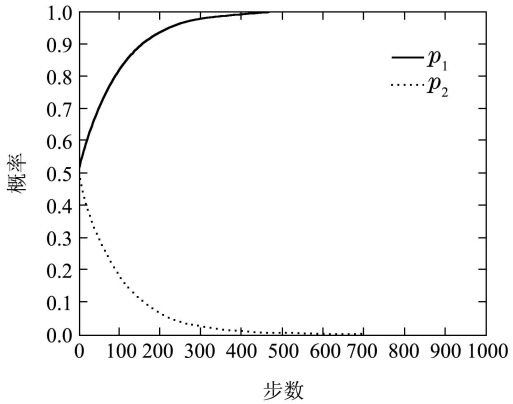

图 4 小白鼠的操作概率曲线

Fig. 4 Curve of operation probability

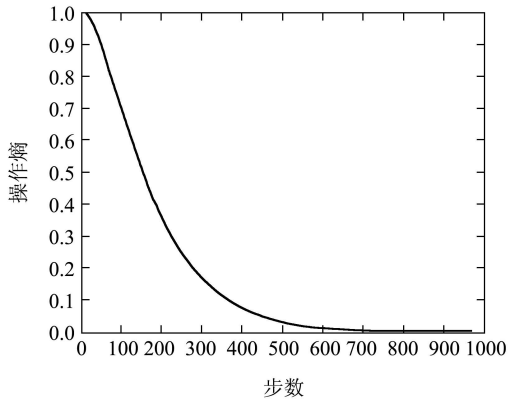

图 5 小白鼠的操作熵曲线

Fig. 5 Curve of operation entropy

AOCA操作熵$\psi(t)$明显是收敛的, 与理论证明结论相符.

斯金纳小白鼠实验的仿真结果表明, 小白鼠在设计的自动机AOCA的模拟下, 能够通过自主的学习, 逐渐增加具有正取向的操作行为的次数, 降低具有负取向的操作行为次数. 这说明小白鼠已经了解和适应所处的环境, 具有自主根据经历改变操作行为的倾向, 这正体现了操作条件反射的本质, 与实际实验中的情况是吻合的, 证明该模型是有效的.

5 结束语(Conclusions)

本文基于斯金纳的操作条件反射理论在自动机的理论框架中, 提出一种自治操作条件自动机AOCA, 同时基于该模型设计了一种仿生自主学习控制的方法, 并通过对斯金纳小白鼠实验的模拟来验证该模型和学习方法的有效性. 仿真结果表明, AOCA充分体现了动物的操作条件反射机制, 并且具有和实际类似的学习效果, 从而说明了AOCA能够实现操作条件反射学习的机制.

为了进一步说明AOCA能够对自治系统的设计发挥指导作用, 在后续研究中将把AOCA模型和仿生自主学习方法应用到智能机器人的自主学习控制中, 以期将AOCA发展成为实现真正的智能机器人的理论基础.

参考文献(References):

[1] MISRA S, OOMMEN B J. Dynamic algorithms for the shortest path routing problem: learning automata-based solutions [J]. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2005, 35(6): 1179 – 1192.

[2] NAJIM K, PIBOULEAU L, LE LANN M V. Optimization technique based on learning automata [J]. *Journal of Optimization Theory and Applications*, 1990, 64(2): 331 – 347.

[3] 于乃功, 王胜, 阮晓钢. 基于细胞自动机的移动机器人路径规划算法 [J]. 控制与决策, 2010, 25(07): 1055 – 1058.
(YU Naigong, WANG Sheng, RUAN Xiaogang. Robot path planning algorithm based on cellular automata [J]. *Control and Decision*, 2010, 25(7): 1055 – 1058.)

[4] BANKS J S, SUNDARAM R K. Repeated games, finite automata, and complexity [J]. *Games and Economic Behavior*, 1990, 2(2): 97 – 117.

[5] KARHUMAKI J, PLANDOWSKI W, RYTTER W. Pattern-Matching problems for two-dimensional images described by finite automata [J]. *Nordic Journal of Computing*, 2000, 7(1): 1 – 13.

[6] ROSIN P L. Image processing using 3-state cellular automata [J]. *Computer Vision and Image Understanding*, 2010, 114(7): 790 – 802.

[7] LAMEGO M M. Automata control systems [J]. *IET Control Theory & Applications*, 2007, 1(1): 358 – 371.

[8] SKINNER B F. Two types of conditioned reflex and a pseudo type [J]. *Journal of General Psychology*, 1935, 12(1): 66 – 77.

[9] SKINNER B F. *The Behavior of Organisms an Experimental Analysis* [M]. New York & London: D. Appleton-Century Compan, 1938.

[10] DAW N D, TOURETZKY D S. Operant behavior suggests attentional gating of dopamine system inputs [J]. *Neurocomputing*, 2001, 38–40: 1161 – 1167.

[11] SAKSIDA L M, RAYMOND S M, TOURETZKY D S. Shaping robot behavior using principles from instrumental conditioning [J]. *Robotics and Autonomous Systems*, 1997, 22(3/4): 231 – 249.

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

[13] DAW N D, TOURETZKY D S. Behavioral considerations suggest an average reward TD model of the dopamine system [J]. Neurocomputing, 2000, 32: 679 – 684.

[14] COURVILLE A C, DAW N D, TOURETZKY D S. Bayesian theories of conditioning in a changing world [J]. *Trends in Cognitive Sciences*, 2006, 10(7): 294 – 300.

[15] ITOH K, MIWA H. Behavior model of humanoid robots based on operant conditioning [C] *//The 5th IEEE-RAS International Conference on Humanoid Robots*. New York: IEEE, 2005: 220 – 225.

[16] SOKOLOWSKI M B C, ABRAMSON C I. From foraging to operant conditioning: a new computer-controlled skinner box to study free-flying nectar gathering behavior in bees [J]. *Journal of Neuroscience Methods*, 2010, 188(2): 235 – 242.

[17] 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(08): 1025 – 1030.
(RUAN Xiaogang, CAI Jianxian, DAI Lizhen. Compute model of operant conditioning based on probabilistic automata [J]. *Journal of Beijing University of Technology*, 2010, 36(8): 1025 – 1030.)

[18] 阮晓钢, 蔡建羡. 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 [J]. 控制理论与应用, 2010, 27(7): 960 – 964.
(RUAN Xiaogang, CAI Jianxian. Fuzzy operant conditioning probabilistic automaton bionic autonomous learning system and robot self-balancing control [J]. *Control Theory & Applications*, 2010, 27(7): 960 – 964.)

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, E-mail: adrxg@bjut.edu.cn;

戴丽珍 (1983–), 女, 博士研究生, 从事模式识别、人工智能、认知科学等研究, E-mail: alice.dai2011@gmail.com, 通讯作者;

于乃功 (1966–), 男, 教授, 博士, 从事机器人、复杂系统建模与控制等研究, E-mail: yunaigong@bjut.edu.cn;

于建均 (1965–), 女, 副教授, 硕士生导师, 从事控制理论与控制工程、模式识别等研究, E-mail: yujianjun@bjut.edu.cn.