Journal of System Simulation

Volume 30 | Issue 2 Article 8

1-2-2019

An Autonomous Cognitive Model Simulating Basal Ganglia Mechanism

Zongshuai Li
1.School of electronic information and automation, Civil Aviation University of China, Tianjin 300300, China; ;

Chen Jing
2.School of information technology engineering, Tianjin University of Technology and Education, Tianjin 300222, China;

Follow this and additional works at: https://dc-china-simulation.researchcommons.org/journal

 Part of the Artificial Intelligence and Robotics Commons, Computer Engineering Commons, Numerical Analysis and Scientific Computing Commons, Operations Research, Systems Engineering and Industrial Engineering Commons, and the Systems Science Commons

This Paper is brought to you for free and open access by Journal of System Simulation. It has been accepted for inclusion in Journal of System Simulation by an authorized editor of Journal of System Simulation.

An Autonomous Cognitive Model Simulating Basal Ganglia Mechanism

Abstract

Abstract: Aiming at autonomous cognition problem, *applying the winner-takes-all(WTA) learning mechanism, using cortical-striatal synaptic modification principle in basal ganglia, as well as based on the operant conditioning reflex theory, the basal ganglia autonomous cognitive model is established,* which is suitable for the cognitive learning with limited actions. Skinner's pigeons experiment was simulated by applying this bionic learning model, which showes a gradual process of adaptive learning. The simulation results reflect that the proposed cognitive model is effective and can be used in the balance learning for a self-balancing robot. *In the premise of unknown robot mathematical model, the purpose of self-balancing learning is achieved through the operant learning in the proposed model.* This research provides a reference for agent's bionic cognitive model.

Keywords

operant learning model, basal ganglia, cognitive model, robotic pigeon, self-balancing robot

Recommended Citation

Li Zongshuai, Chen Jing. An Autonomous Cognitive Model Simulating Basal Ganglia Mechanism[J]. Journal of System Simulation, 2018, 30(2): 427-434.

This paper is available in Journal of System Simulation: https://dc-china-simulation.researchcommons.org/journal/vol30/iss2/8

第 30 卷第 2 期 系统仿真学报© Vol. 30 No. 2
2018 年 2 月 Journal of System Simulation Feb., 2018

一种模拟基底神经节机理的自主认知模型

李宗帅[1]，陈静[2]
(1.中国民航大学电子信息与自动化学院，天津 300300；2.天津职业技术师范大学信息技术工程学院，天津 300222)

摘要：针对智能体自主认知问题，*采用适者生存 WTA 学习机制，利用基底神经节中的皮质-纹状体突触修饰机理，基于操作学习原理提出了基底神经节的自主认知模型*，该模型适用于有限行为的认知学习，应用该仿生学习模型对 Skinner 的鸽子实验进行了模拟，模拟了 Skinner 鸽的渐进自适应学习过程，仿真结果表明本文所提出的行为认知模型是有效的。将该认知模型用于自平衡机器人的自主平衡学习中，*在未知机器人数学模型的前提下通过模型中的操作学习达到自平衡学习的目的*，为智能体仿生认知模型的研究提供了参考。

关键词：操作学习模型；基底神经节；认知模型；机器鸽；自平衡机器人
中图分类号：TP181 文献标识码：A 文章编号：1004-731X (2018) 02-0427-08
DOI: 10.16182/j.issn1004731x.joss.201802008

An Autonomous Cognitive Model Simulating Basal Ganglia Mechanism

Li Zongshuai[1], *Chen Jing*[2]
(1. School of electronic information and automation, Civil Aviation University of China, Tianjin 300300, China;
2. School of information technology engineering, Tianjin University of Technology and Education, Tianjin 300222, China)

Abstract: Aiming at autonomous cognition problem, *applying the winner-takes-all(WTA) learning mechanism, using cortical-striatal synaptic modification principle in basal ganglia, as well as based on the operant conditioning reflex theory, the basal ganglia autonomous cognitive model is established,* which is suitable for the cognitive learning with limited actions. Skinner's pigeons experiment was simulated by applying this bionic learning model, which showes a gradual process of adaptive learning. The simulation results reflect that the proposed cognitive model is effective and can be used in the balance learning for a self-balancing robot. *In the premise of unknown robot mathematical model, the purpose of self-balancing learning is achieved through the operant learning in the proposed model.* This research provides a reference for agent's bionic cognitive model.

Keywords: operant learning model; basal ganglia; cognitive model; robotic pigeon; self-balancing robot

引言

Skinner 操作条件反射的原理被大量应用于动物训练、教学以及医学中。近年来，研究人员把操作条件反射理论应用于机器人学习和控制方面，进行了大量的研究与实验[1-8]，实现了自主平衡、自主避障以及语音控制等功能。研究发现，对未来的预测在智能体认知中非常重要，指导动物进行下一步的决策。

在国内的相关研究中，北京工业大学阮晓钢等人率先利用概率自动机建立了一种基于遗传算法的 Skinner 操作条件反射学习模型[9]，并模拟了 Skinner 的鸽子实验，表现了该模型具有良好的仿

收稿日期：2016-05-11 修回日期：2016-06-13；
基金项目：国家自然科学基金(61403282)，天津市高等学校科技发展基金(20130807)；
作者简介：李宗帅(1982-)，男，山东德州，硕士，讲师，研究方向为人工智能、自动化、自动控制。

Published by Journal of System Simulation, 2018

Journal of System Simulation, Vol. 30 [2018], Iss. 2, Art. 8

生学习能力，但是该方法在连续状态问题求解时需对状态空间进行分类，容易造成维数灾问题。后来，任红格等人应用操作条件反射原理建立了感觉运动系统的认知模型[10]。徐冰等人采用动机模型提出了一种虚拟人自主行为选择机制[11]。

文献[12]设计了一种操作条件反射模型，通过功能性核磁共振成像技术发现，基底神经节中的腹部纹状体影响着奖赏和动机，脊部纹状体影响着动作和认知控制。

鉴于基底神经节在操作条件反射过程中的神经生理学基础，本文首先建立了基底神经节的操作条件反射学习行为认知模型，并将模拟退火(SA)机制引入到基底神经节的 WTA(Winner-takes-all)机制中，采用斯金纳的鸽子实验，成功模拟了动物的渐进学习和适应性能，最后将该模型用于自平衡学习问题中，针对连续状态问题表现出了良好的认知学习能力。

1 操作条件反射

1.1 操作条件反射基本原理

Skinner 的操作条件反射理论是一种由刺激引起的行为改变的过程与方法，又被称为工具学习(instrumental learning)或操作学习(operant learning)。操作学习理论表明，当某种行为使系统向好的方向发展，或者说某行为正确，则下次在相同状态下实施该行为的概率将增加，否则，当某种行为使系统向不好的方向发展，或者说选择该行为是错误的，则下次在相同状态下实施该行为的概率会减小，通过一定阶段的操作学习训练，生物体(系统)会学会适应环境的操作行为。按照 Björn Brembs 对操作条件反射学习的观点，对未来的预测在操作学习过程中发挥着重要作用。

对于一个还未完全了解的系统，可以通过学习过去的经验预测其未来行为。预测学习的一个重要优点就是它的训练样本直接源于实时输入的时间序列，而不需要特殊的教师信号。根据这一思想，我们首先建立如下操作条件反射学习模型。

1.2 操作学习模型

根据操作条件反射理论，本文建立了基于该理论的学习模型，即：操作学习模型。

定义 1 一个操作学习模型 OLM 可以表示为一个八元组计算模型 $\mathrm{OLM}=\langle S,A,f,\varphi,r,V(S,A),P,L\rangle$，各元素含义如下：

(1) S：OLM 的内部离散状态集合，$S=\{s_i\,|\,i=0,1,2,\cdots,n\}$，$S$ 为系统所有可能离散状态组成的非空集，s_i 表示第 i 个离散状态，n 为离散状态的个数；

(2) A：OLM 的可选操作行为集合，$A=\{a_i\,|\,i=0,1,2,\cdots,m\}$，$a_i$ 表示第 i 个可选操作行为，m 为可选操作行为的个数；

(3) f：OLM 状态转移函数 $f:S(t)\times a(t)\to S(t+1)$，即 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 $a(t)\in A$ 确定，一般由环境或者系统模型决定；

(4) φ：OLM 的取向机制，$\varphi(t)=\varphi(S(t))$ 表示时间 t 时刻系统的取向性，根据系统的状态定义，这里的取向性是从生物学意义上来定义的，环境决定生物进化的方向，即生物的取向性。取向值越大越好，根据不同情况来定义不同的取向函数；

(5) r：$r(t)=r[S(t),A(t)]$ 为系统在 t 时刻 $S(t)$状态下实施操作行为 $A(t)$后状态转移到 $S(t+1)$后的奖赏；根据取向函数来定义，如果 $\varphi(t+1)>\varphi(t)$，表明系统向好的方向发展，$r=1$；如果 $\varphi(t+1)<\varphi(t)$，表明系统向不好的方向发展，$r=0$；

(6) $V(S,A)$：OLM 预测函数，在固定状态 S 下，$V(S,A)=\{v_i(S,a_i)\,|\,i=0,1,2,\cdots,m\}$，可以看作是每一个可选行为对为未来奖赏折扣和的估计值；

(7) P：OLM 从条件状态到可选操作行为的概率矢量，$P=[p(a_1),p(a_2),\cdots,p(a_m)\,|\,S]=[p_{a_1,S},p_{a_2,S},\cdots,p_{a_m,S}]$，行为选择服从 Γ 概率分布，其中，$p_{a_j,S}=p(a=a_j\,|\,S)=e^{\beta V(S,a_j)}/\sum_{a\in A}e^{\beta V(S,a)}$，其含义为在状态处于 S 的条件下，OLM 依据概率 $p(a_j)\in\Gamma$ 实施操作行为 $a_j\in A$；概率矢量中，对每一个状态 $s\in S$ 的概率矢量为条件概率，满足：

https://dc-china-simulation.researchcommons.org/journal/vol30/iss2/8
DOI: 10.16182/j.issn1004731x.joss.201802008

$0 < p(a_m|s) < 1$， $\sum_{i=1}^{m} p(a_i|s) = 1$；

(8) L：OLM 学习机制，$L: P(t) \to P(t+1)$，该学习机制的更新主要是通过评价机制的更新来实现，这里主要根据时间差分学习算法(Temporal-Difference learning，TD learning)更新 OLM 预测函数 $V(S,A)$ 中的权值。

信息论中的熵表征了系统的不确定程度，信息熵越大，系统不确定性也越大。某状态下熵 E_k 定义如下：

$$E_k = -\sum_{i \in U(k)} \pi_k(i) \log \pi_k(i) \tag{1}$$

式中：$\pi_k(i)$ 表示在状态 k 下选择行为 i 的概率。根据熵的定义，我们如下定义条件熵的概念。

定义 2 条件熵 $H(s_i)$ 表示 OLM 在状态 $s_i \in S$ 条件下的操作行为熵：

$$H(s_i) = -\sum_{k=1}^{m} p(a_k|s_i) \log_2(p(a_k|s_i)) \tag{2}$$

整个操作学习过程的基本原理可简述如下：t 时刻，系统状态为 $S(t) = s_i \in S$，根据初始的预测函数和 P 概率矢量确定每个操作行为的选择概率，根据竞争机制依概率选择操作行为 $a_k \in A$，实施该操作，状态发生转移 $S(t+1) \in S$，根据取向信息获得该次此操作的即时评价信息 r，根据 TD 学习更新预测网络 $V(S,A)$，从而形成新的预测估计值，获得新的概率矢量 P，继续下一时刻的行为选择，如此循环，直到更新后的网络能够学习到最优的操作，操作条件反射形成，学习结束。

2 基于操作学习的基底神经节模型

2.1 模型构建

基底神经节(BG)是大脑深部一系列神经核团组成的功能整体。它与大脑皮质、丘脑和脑干相连。据解剖学和生理学可知，BG 的主要功能是控制自主运动，能够进行行为选择和操作条件反射学习。BG 包括尾核、壳核(纹状体)、苍白球、黑质和丘脑底核。纹状体根据胆碱酯酶染色不同分为纹状小体和基质两部分。纹状体是 BG 接收大脑皮层(CC)兴奋性传入的结构，其中基质接收整个大脑皮层的兴奋性传入纤维投射，纹状小体只接收额叶前部的兴奋性传入纤维投射。基质发出抑制性传出纤维至苍白球内侧段和黑质网状部，组成了基底核的输出结构。纹状小体传出至黑质致密部，控制调节黑质-纹状体通路。最终形成了大脑皮层-基底核-丘脑-大脑皮层(CC-BG-TH-CC)的回路结构。其中，动作评估在纹状小体中去学习，行为方式选择在基质中进行，来自黑质致密部的多巴胺能输入被用作动作评估的指导信号，用来改善由动作导致的最大未来奖赏的行为表达，以便获得更加精确的行为结果。

基于操作学习原理和 BG 的神经生理学基础，本文建立了基底神经节的行为认知模型，BG 模型主要包括：感觉皮质(SC)、运动皮质(MC)、纹状体(纹状小体、基质)(STR)和黑质(SN)，可以用式(3)来表示。

$$\begin{aligned} \mathrm{BG} \sim (&SC, MC, STR_{striosome}(CC, CC\text{-}STR_{synapsis}), \\ &STR_{matrix}, SN_{DA}) \end{aligned} \tag{3}$$

各元素的含义为：SC 为感觉运动皮质感知的系统状态信息；MC 为运动皮质感知的行为信息；CC 为大脑皮质的感觉运动皮质信息；$CC\text{-}STR_{synapsis}$ 为大脑皮质与纹状体之间的突触连接；$STR_{striosome}$ 为纹状小体输出，对未来奖赏折扣和的预测；SN_{DA} 黑质纹状体输出的多巴胺能。

BG 与大脑皮质和丘脑共同协调作用，形成了操作条件反射学习机制，整个流程如图 1 所示。实线表示信号流，虚线表示突触修饰，外部虚框部分为整个基底神经节，内部虚框部分为纹状体部分(Striatum)，纹状小体 striosome 和基质 matrix 共同构成纹状体。黑质(Substantia Nigra，SN)部分产生多巴胺能信号，用来修正皮质-纹状体之间的突触连接，丘脑实现两部分功能：奖赏信息的生成以及信息传递。整个认知过程如下：大脑皮质、基底神经节中的纹状体以及丘脑共同协调作用，基底神经节中的纹状小体接收系统状态和可选行为信息，产生对可选行为的预测值，经过基质 matrix 的依概率行为选择策略产生推荐行为，经由丘脑传递至大脑皮质的运动皮层，作用于机器人，与环境交互，

Journal of System Simulation, Vol. 30 [2018], Iss. 2, Art. 8

状态发生转移，通过丘脑产生奖赏信号，在黑质 SN 处生成多巴胺能信号，修正 $CC\text{-}STR_{synapsis}$ 的突触连接，修正对状态行为的预测，随着学习的进行，机器人能够通过所建立的基底神经节操作学习模型逐渐习得技能。

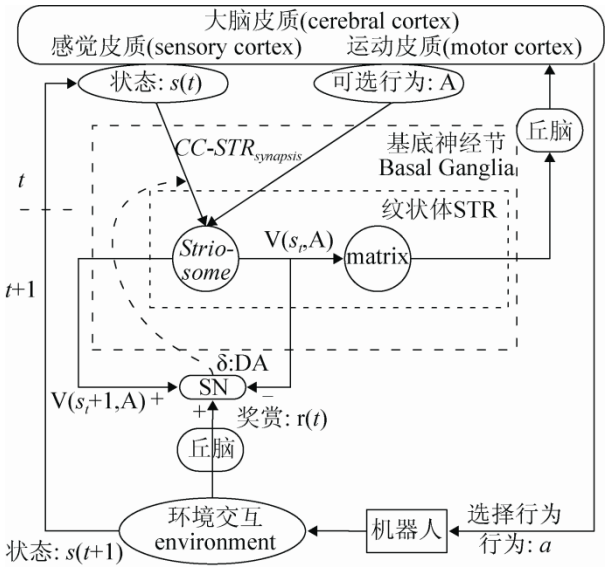


图 1 基于操作学习原理的基底神经节认知结构

Fig.1 Basal ganglia cognitive structure based on the principle of instrumental learning

2.2 基于递归神经网络的纹状体实现

2.2.1 基于递归神经网络的纹状体结构模拟

从皮质层到纹状小体之间的神经网络用递归神经网络来实现，网络中间的连接权值表示了皮质-纹状小体的突触连接，这种递归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，能够体现人类大脑记忆的稳定性。

2.2.2 纹状体 STR 输出

皮质层的输入信息包括感觉皮质信息和运动皮质信息，因此，定义如下符号：

$$CC=[SC;MC] \tag{4}$$

$$STR_{\text{striosome}}=STR(CC,CC\text{-}STR_{\text{synapsis}}) \tag{5}$$

式中：$STR_{\text{striosome}}$ 值为对未来奖赏折扣和的预测，

$$STR_{\text{striosome}}(t)=r(t+1)+\gamma r(t+2)+\gamma^2 r(t+3)+\cdots \tag{6}$$

则 $t+1$ 时刻未来奖赏折扣和为：

$$STR_{\text{striosome}}(t+1)=r(t+2)+\gamma r(t+3)+\gamma^2 r(t+4)+\cdots \tag{7}$$

纹状体 STR 输出采用上述的递归神经网络实现，其中隐含层表示大脑皮质中的颗粒细胞，神经网络权值表示皮质与纹状体的突触连接。由式(6)和式(7)可知，$STR_{striosome}(t)=r(t+1)+\gamma STR_{striosome}(t+1)$，这表明 t 时刻的动作评价值 $STR_{striosome}(t)$ 可以用 $t+1$ 时刻的评价值 $STR_{striosome}(t+1)$ 来表示，但由于在预测初期必然会存在一个误差，使得用评价值 $STR_{striosome}(t+1)$ 表达的 $STR_{striosome}(t)$ 与实际的 $STR_{striosome}(t)$ 并不相等，这样由纹状体输和丘脑输出的奖赏信息在黑质处进行处理，产生了多巴胺能响应，用 SN_{DA} 来表示，则

$$SN_{DA}(t)=r(t+1)+\gamma STR_{striosome}(t+1)-STR_{striosome}(t) \tag{8}$$

2.3 CC 与 STR 的突触修饰

黑质产生的多巴胺能反馈至纹状体，用于修饰皮质-纹状体突触，形成黑质-纹状体环路(Nigra-Strio loop)，依据时间差分学习算法，用 $r(t+1)+\gamma STR_{\text{striosome}}(t+1)$ 估计 $STR_{\text{striosome}}(t)$ 形成 $SN_{DA}(t)$ 误差，基于权值的突触修饰机制为式(9)和(10)，

$$CC\text{-}STR_{\text{synapsis}}(t)\leftarrow CC\text{-}STR_{\text{synapsis}}(t)+\Delta CC\text{-}STR_{\text{synapsis}}(t) \tag{9}$$

$$\Delta CC\text{-}STR_{\text{synapsis}}(t)=\alpha\cdot SN_{DA}\cdot\partial STR/\partial CC\text{-}STR_{\text{synapsis}} \tag{10}$$

2.4 纹状体基质中的行为选择机制

在基底神经节的操作学习过程中，纹状体的基质部分司职于行为选择机制，操作学习过程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：

$$P(a=a_i\mid SC(t))=\frac{e^{STR_{striosome}(SC(t),a_k)/T}}{\sum_{a\in A}e^{STR_{striosome}(SC(t),a)/T}} \tag{11}$$

式中：T>0 为温度常数，表示行为选择的随机程度，当 T 趋近于零时，选择最大 $STR_{striosome}(SC(t),a_k)$ 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。系统初始行为选择的随机性通过随机初始化递归神经网络的权值实现，T 的降温过程如式(12)所示，

https://dc-china-simulation.researchcommons.org/journal/vol30/iss2/8
DOI: 10.16182/j.issn1004731x.joss.201802008

$$\begin{cases} T_0 = T_{\max}, \\ T_{t+1} = T_{\min} + \beta(T_t - T_{\min}) \end{cases} \tag{12}$$

式中：$0 \leqslant \beta \leqslant 1$ 为退火因子。

2.5 基底神经节中的操作学习流程

由上所述，基底神经节中的操作学习过程如下：

Step1：初始化，迭代学习步数初值 $t=0$；迭代学习次数为 Maxstep；初始化皮质-纹状体突触权值为零，则一开始选取初始操作行为的概率相等，即：$\forall s \in S, p(a_k \mid s) = 1/m(k=1,2,\cdots,m)$，其中，$m$ 为可选操作行为的个数。选取各操作行为的初始概率相等，意味着在初始状态下，BG 不含有任何预定的决策。此外，初始化时还要给定皮质层感知的初始状态 $SC(0)$、操作行为集合 A 及 OCR 学习机制的学习系数。

Step2：获得纹状体的输出值及基质中所选择的行为，每个行为的纹状体输出值 $STR_{striosome}(SC(t), a_k)$，根据式(11)计算每一行为的选择概率，通过纹状体中 Matrix 的行为选择机制，输出行为 $a(t)$，并存储 $STR_{striosome}(SC(t), a(t))$；

Step3：实施选取的操作行为，执行行为 $a(t)$，状态发生转移 $SC(t+1)$，获得即时奖赏 r_{t+1}，按照 Step2 方式进行行为选择 $a(t+1)$，求得 $t+1$ 时刻纹状体预测值 $STR_{striosome}(SC(t+1), a(t+1))$；

Step4：操作条件反射，根据式(9)，(10)进行皮质-纹状体突触权值更新，依据式(12)进行降温，从而行为选择概率发生变化，操作条件反射形成；

Step5：时间更新，如果 t<Maxstep，t=t+1，转移至 Step2；如果 t=Maxstep，转移至 Step5；

Step6：结束。

3 行为认知模型在 Skinner 鸽实验中的应用

为了验证设计的认知模型体现出操作条件反射的行为特性，对 Skinner 鸽子实验进行了复现。

3.1 Skinner 鸽实验模型建立

关于 Skinner 鸽子实验是 Skinner 针对操作条件反射提出的一个经典的行为学习实验，即：开始时鸽子啄红，黄和蓝三个按钮是随机的，但是，它啄红色按钮时得到食物(正强化刺激)，啄黄色按钮时无任何刺激，啄蓝色按钮时给予电击(负强化刺激)，一段时间之后，鸽子啄取红色按钮的次数明显高于啄取其它两个按钮的次数。鸽子通过操作行为获取了知识，避免电击，饥饿时能够自主获取食物。

为了建立 Skinner 鸽子实验的数学模型，首先对状态和动作进行编码，建立简化的离散数学模型。

3.1.1 状态和行为编码

设鸽子有三个状态，分别为：痛苦、渴望食物和满足状态，痛苦状态 $s_0 = 1$；渴望食物状态 $s_1 = 2$；满足状态 $s_2 = 3$。状态值越大，表明鸽子的状态越好。鸽子有三个行为分别为：啄红色按钮、啄黄色按钮、啄蓝色按钮，分别编码为：$a_0 = 1, a_1 = 2, a_2 = 3$。鸽子的取向性：最大化满足自身需求，状态越好取向性也越大。

3.1.2 状态转移

通过对鸽子实验进行建模分析，将其行为学习过程简化为状态转移方程 f：$S(t) \times A(t) \to S(t+1)$ 具体表示为：

$$\begin{cases} f(s_0, a_0) = s_1 \quad f(s_1, a_0) = s_2 \quad f(s_2, a_0) = s_2 \\ f(s_0, a_1) = s_0 \quad f(s_1, a_1) = s_0 \quad f(s_2, a_1) = s_1 \\ f(s_0, a_2) = s_0 \quad f(s_1, a_2) = s_0 \quad f(s_2, a_2) = s_0 \end{cases} \tag{13}$$

3.1.3 取向函数和奖赏机制

针对 Skinner 鸽定义如式(14)所示的取向性泛函：

$$\varphi(t) = s(t) \tag{14}$$

该取向性泛函说明，当鸽子处于满足状态时，$s(t) = 3, \varphi(t) = 3$，取向性最好；当鸽子处于渴望食物状态时，$s(t) = 2, \varphi(t) = 2$，取向性较差；当鸽子处于痛苦状态时，$s(t) = 1, \varphi(t) = 1$，取向性最差；

根据取向函数定义奖赏机制，如果 $\varphi(t+1) \geqslant \varphi(t)$，表明系统向好的方向发展，$r = 1$；如果 $\varphi(t+1) < \varphi(t)$，表明系统向不好的方向发展，$r = 0$，即：

$$r_{t+1} = \begin{cases} 1(\text{代表奖赏}) & if\ (\varphi(t+1) > \varphi(t)), \\ 0(\text{代表惩罚}) & else. \end{cases} \tag{15}$$

3.2 实验参数及过程

将基于操作条件反射学习机制的 BG 模型用于 Skinner 鸽子实验，递归神经网络输入层、隐含层、输出层神经元个数 $r=2,h=5,m=1$，初始化皮质-纹状体突触权值为零(保证初始时刻的行为选择概率为 1/3)；设定最大运行步数 Maxstep=2 000；随机初始化鸽子状态 S(0)，经过 2.5 节提出的操作学习流程，突触权值不断变化，最终恒定，行为学习结束。

3.3 实验结果与分析

按照上述实验步骤，设定采样周期 $T_c=1s$，在 MATLAB 环境下，进行实验模拟仿真。在仿真过程中，记录下每一时刻下机器鸽的状态和实际采取的行为，分别在时刻 200 s，400 s，600 s，800 s，1 200 s，2 000 s 对选取三个行为的次数进行了统计，该 6 次的统计结果如图 2 所示，图 3 给出了鸽子啄不同颜色按钮概率的变化情况以及采用经典概率自动机方法的对比结果。

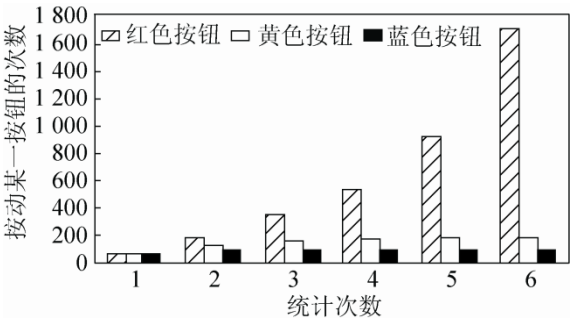
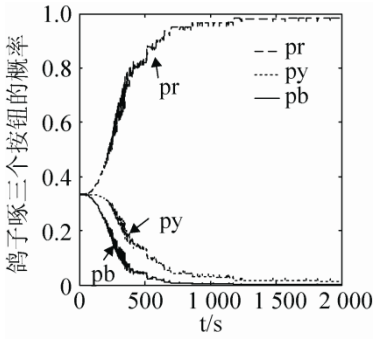
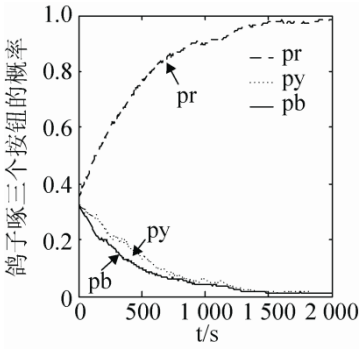


图 2 选取不同行为的次数
Fig.2 Number of selecting different behavior

从图 2 和图 3 的仿真结果可以看出，在训练初期阶段，因选取不同操作行为的初始概率相同，所以鸽子啄三个按钮的次数基本上是相同的，均为 0.333 3。但随着鸽子与随机环境的交互，皮质-纹状体的突触不断发生变化，引发了行为选择概率的变化，鸽子选择啄红色按钮的概率和次数逐渐增加，而选择啄黄色和啄蓝色按钮的概率和次数逐渐降低。当进行到 2 000 s 左右时，鸽子选择啄红色按钮的概率已经远远大于选择啄另外两个按钮的概率，鸽子啄红色按钮的次数持续增加，而啄黄色、蓝色两个按钮的次数基本上不再变化，由此，基底神经节中的操作条件反射形成。通过图 3 的仿真对比结果也可以看出，与概率自动机方法相比，本文方法在学习初期学习速度较慢，但是由于网络的泛化能力，在学习的后期概率更新速度较快。



(a) 本文方法仿真结果



(b) 概率自动机方法仿真结果

图 3 概率的变化
Fig.3 Changing curve of probability

通过式(2)可以计算出每一时刻系统的熵，绘制出熵随时间的变化情况，如图 4 所示。通过熵的变化趋势可以看出，初始时，机器鸽的行为选择的随机性比较大，通过一段时间基底神经节的操作条件训练，神经结构(主要指皮质-纹状体的突触权值)发生变化，熵逐渐向减小的方向发展，这也说明了系统的学习是熵减小的过程，是有序的自组织过程。

概率自动机方法在针对连续状态问题求解时需要对状态空间进行分类，容易造成维数灾问题，本文方法不需要对状态进行分类，从具体实现上要优于概率自动机方法。

https://dc-china-simulation.researchcommons.org/journal/vol30/iss2/8
DOI: 10.16182/j.issn1004731x.joss.201802008

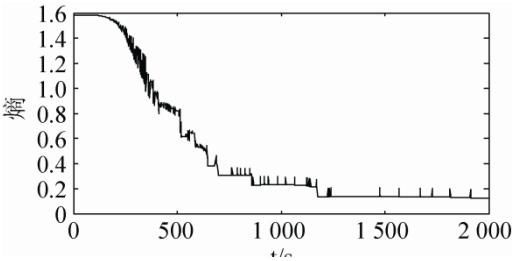


图 4 熵的变化
Fig.4 Changing curve of entropy

4 行为认知模型在自平衡学习中的应用

为了进一步验证本文所提出的认知模型的有效性，将其应用于自平衡机器人的自平衡学习中，系统模型参考文献[13]，学习过程中选用如式(16)所示的特征函数，

$$J(t)=k_1\theta_b^2(t)+k_2\dot{\theta}_b^2(t) \tag{16}$$

式中：$k_1>0,k_2>0$ 为学习控制参数，可调节；$\theta_b,\dot{\theta}_b$ 分别为机器人本体的倾角和倾角速度。

基于特征函数的奖赏定义如式(17)所示，

$$r(t)=\begin{cases}0, & if \quad J(t+1)\leqslant J(t)\\ -1, & else\end{cases} \tag{17}$$

系统的初始角度为 π/18，+10N·m 和–10 N·m 为两个可选择行为，首先进行了在固定状态下的离线学习，两个行为的学习过程以及系统熵的变化如图 5 所示。

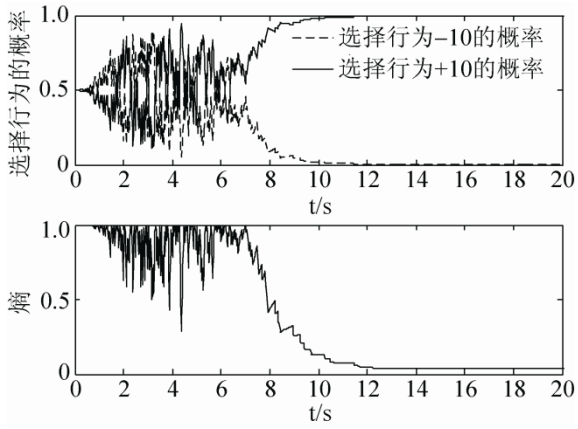


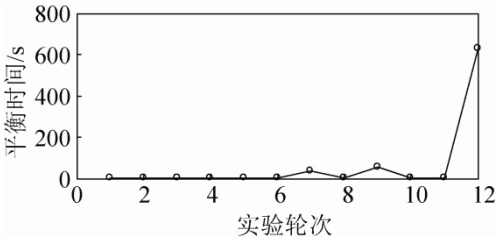
图 5 两个离散行为下的离线学习结果
Fig.5 Off-line learning results under two discrete behaviors

在学习过程中，由于学习初期没有先验知识，学习过程中的行为选择的随机性变化较大，随着学习的推进，在某一状态下选取合适行为的概率将会逐渐趋向于 1，而系统熵逐渐趋向于 0，系统向确定性方向变化。

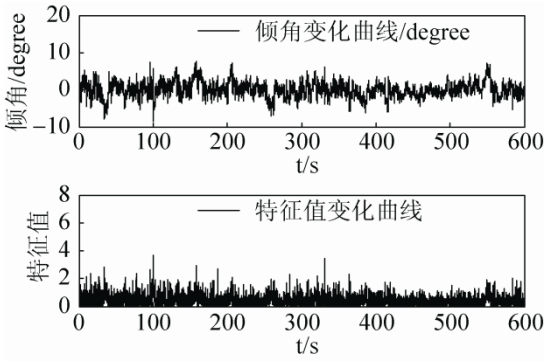
为了实现机器人的自平衡学习，将本文方法用于系统的在线学习，实验过程如下：

参数设置：择 5 个可选行为：{–10, –5, 0, 5, 10} (N·m)，学习过程中评价网络的递归神经网络选用 3-5-1 的结构，即输入神经元 3 个，隐层神经元 5 个，输出神经元 1 个。

学习过程设置：设置机器人的初始角度为 $\pi/18$，极限倾角为 $\pi/6$，当机器人角度大于极限倾角，认为学习失败，重新返回初值继续学习，如果机器人能够保持平衡 600 s，则认为机器人已经习得平衡技能，记录每次试探学习保持平衡的时间，通过实验发现，经过多次试探，机器人经过 12 次试探(如图 6(a))，便习得了平衡技能，可保持平衡不倒 10 分钟，达到了学习目的(如图 6(b))。经过本文所提出的基底神经节行为认知模型，机器人能维持在(–5，+5)度之间保持平衡。



(a) 每轮学习机器人的平衡时间



(b) 600 s 内的平衡效果

图 6 自平衡机器人平衡学习结果
Fig.6 Balance learning result of self balancing robot

Journal of System Simulation, Vol. 30 [2018], Iss. 2, Art. 8

5 结论

本文基于操作条件反射原理建立了基底神经节的行为选择学习模型，首先，应用预测机制建立了操作条件反射学习模型，该学习模型与基底神经节的行为选择机理相对应，形成了基于操作学习的基底神经节认知模型。其中，皮质-纹状体的突触由递归神经网络权值来实现，皮质感觉输入信息为状态-行为对，纹状体中的纹状小体输出对动作的评价，纹状体中的基质执行行为选择，经由丘脑传递至运动皮质，输出所选行为，作用于环境，形成了皮质-纹状体-丘脑-皮质回路。通过黑质产生的多巴胺能对皮质-纹状体的突触进行修饰，形成了黑质-纹状体回路。

所构建基于操作条件反射的基底神经节行为认知模型适用于有限行为的认知学习，对系统状态个数没有限制，应用该仿生学习模型对 Skinner 的鸽子实验进行了模拟，通过仿真实验可以看出，所构建的基底神经节模型有效地模拟了生物的操作条件反射现象，表现出了良好的自组织、自适应和自学习能力，并且成功应用于机器人的自平衡学习中，进一步验证了认知模型的有效性。

参考文献:

[1] Bruce E R, James M G, Jacques J V. Machine operant conditioning[C]// Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 1988: 1500-1501.

[2] Zalama E, Gaudiano P, Coronado J L. Obstacle avoidance by means of an operant conditioning model[C]// International Workshop on Artificial Neural Networks Malaga-Torremolinos, Spain, 1995, 930: 471-477.

[3] Gaudiano P, Chang C. Adaptive obstacle avoidance with a neural network for operant conditioning: Experiments with real robots[C]// IEEE International Symposium on Computational Intelligence in Robotics and Automation, Monterey, 1997: 13-187.

[4] Björn Brembs, brembs.net:Research on Learning, Memory and Evolution [EB/OL]. http://brembs.net/.2014.

[5] Björn Brembs. Spontaneous decisions and operant conditioning in fruit flies[J]. Behavioural Processes (S0376-6357), 2011, 87(1): 157-164.

[6] Zalama E, Gomez J, Paul M, et al. Adaptive behavior navigation of a mobile robot[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part A – Systems and Humans, 2002, 32(1): 160-169.

[7] Itoh K, Miwa H, Matsumoto M, et al. Behavior model of humanoid robots based on operant conditioning[C]// IEEE/RAS International Conference on Humanoid Robots. Piscataway, NJ, USA: IEEE, 2005: 220-225.

[8] Tadahiro T, Tetsuo S. 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: 1177-1199.

[9] 蔡建羡, 阮晓钢. 基于遗传算法的 Skinner 操作条件反射学习模型[J]. 系统工程与电子技术, 2011, 33(6): 1370-1376.
Cai J X, Ruan X G. Skinner operant conditioning learning model based on genetic algorithm[J]. System Engineering and Electronics, 2011, 33(6): 1370-1376.

[10] 任红格, 史涛, 张瑞成. 基于操作条件反射机制的感觉运动系统认知模型的建立[J]. 机器人, 2012, 34(3): 292-298.
Ren H G, Shi T, Zhang R C. Foundation of the sensorimotor system cognitive model with operant conditioning mechanism[J]. Robot, 2012, 34(3): 292-298.

[11] 徐冰, 刘肖健. 基于动机模型的自主性虚拟人行为选择研究[J]. 计算机应用与软件, 2012, 29(4): 71-74.
Xu B, Liu X J. Research on motivation model based behavior selection of autonomous virtual human[J]. Computer application and software, 2012, 29(4): 71-74.

[12] J John O'Doherty, Peter Dayan, et al. Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning [J]. Science (S0036-8075). 2004, 304(5669): 452-454.

[13] Ruan X, Chen J, Yu N. Thalamic cooperation between the cerebellum and basal ganglia with a new tropism-based action-dependent heuristic dynamic programming method[J]. Neurocomputing (S0925-2312), 2012, 93: 27-40.

https://dc-china-simulation.researchcommons.org/journal/vol30/iss2/8
DOI: 10.16182/j.issn1004731x.joss.201802008