文章编号:1001-506X(2011)06-1370-07

# 基于遗传算法的 Skinner 操作条件反射学习模型

蔡建羡[1,2], 阮晓钢[1]

(1. 北京工业大学电子信息与控制工程学院, 北京 100124; 2. 防灾科技学院, 河北 三河 065201)

**摘 要**: 以概率自动机(probabilistic automata, PA)为平台,结合遗传算法(genetic algorithm, GA)的进化思想,设计了反映 Skinner 操作条件反射(operant conditioning, OC)思想的仿生学习模型,称为基于遗传算法的操作条件反射概率自动机(genetic algorithm-operant conditioning probabilistic automata,GA-OCPA)学习系统。每一次学习尝试之后,首先,学习系统把通过 OC 学习算法学习得到的信息熵值作为个体适应度;然后,执行遗传算法,搜索最优的个体;最后,再执行 OC 学习算法学习最优个体内的最优操作行为,以得到新的信息熵值。理论上分析了 GA-OCPA 学习系统学习算法的收敛性,通过对两轮机器人运动平衡控制的仿真分析,表明设计的 GA-OCPA 学习系统的学习是一个自动获取知识和提炼的过程,具有高度的自适应能力。

**关键词**: 操作条件反射; 遗传算法; 概率自动机; 运动平衡控制

**中图分类号**: TP 18 **文献标志码**: A **DOI**:10.3969/j.issn.1001-506X.2011.06.34

# Skinner operant conditioning learning model based on genetic algorithm

CAI Jian-xian[1,2], RUAN Xiao-gang[1]

(*1. School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China; 2. Institute of Disaster Prevention, Sanhe 065201, China*)

**Abstract**: Platform on probabilistic automata and combined with evolution thought of genetic algorithm, this paper constructs a bionic learning model which can reflect the essence of Skinner operant conditioning. The designed learning model is named as genetic algorithm-operant conditioning probabilistic automaton (GA-OCPA) bionic autonomous learning system. After each learning trial, the learning system firstly obtains the information entropy value based on operant conditioning (OC) learning result and uses it as the fitness of individual. And then genetic algorithm is performed based on information entropy value to find the optimal individual. At last, the OC learning algorithm is performed to learn the optimal operant action in optimal individual, and correspondingly a new information entropy value will be obtained. The convergence theorems for the learning algorithm of GA-OCPA bionic learning system is presented, and the simulation analyses in motion balancing control of two-wheeled robot demonstrate that the learning of GA-OCPA bionic learning system is a process of autonomously acquiring and epurating knowledge and has high adaptive ability.

**Keywords**: operant conditioning; genetic algorithm; probabilistic automata; motion balancing control

## 0 引 言

对于动力学模型复杂性和环境不确定的系统,由于环境知识获取困难,采用传统的基于模型和专家知识的控制方法往往难以取得良好的控制效果。使机器具有学习能力,模拟或实现人类学习活动为目的的机器学习,是人工智能的一个重要的研究领域。最有代表性的神经网络作为一种重要的机器学习方法,尤其在机器人学领域发挥着重要作用[1-2],但是神经网络学习是基于经验和数据离线学习的,实际当中这些数据或经验是很难获得或需要花费很多时间的。所以,仿生学习作为一种可以不需要环境模型,无导师的在线学习方法,对实现自主机器人自学习、自适应能力具有重要的应用价值。

学习自动机[3-4]和强化学习[5-6]是两种常见的仿生学习机制,非常适合于解决简单的控制任务;但是,当控制任务复杂时,他们将无能为力。学习自动机和强化学习的学习思想主要来源于美国 Harvard 大学心理学教授 Skinner 提出的"操作条件反射"[7-8]。Skinner 操作条件反射(operant conditioning, OC)被视为生物系统最基本的学习形式,其核心内容为:如果在一定的主客观条件下,生物的某种操作性行为(或操作)所导致的后果符合生物的取向性,那么在

**收稿日期**:2010-01-31; **修回日期**:2011-01-06。

**基金项目**:国家自然科学基金(60774077);国家高技术研究发展计划(863 计划)(2007AA04Z226);北京市教委重点项目(KZ200810005002)资助课题

**作者简介**:蔡建羡(1978-),女,讲师,博士,主要研究方向为机器学习、机器人智能性能。E-mail:cjxlaq@163.com

类似的主客观条件下,生物实施类似的行为(或操作)的概率将会上升。因此,基于 Skinner OC 理论设计一种能解决复杂控制任务的仿生学习模型就是本文的出发点。自 1996 年以来,众多学者对 OC 学习行为模型进行了广泛的研究[9-12],并将之成功应用到机器人上。但是,这些模型没有给出具体的评价函数以及动作选择的数学计算模型,使得学习模型不具备泛化能力。

概率自动机(probabilistic automata, PA)[13]的状态转移具有随机性,所以很适于构建仿生学习系统。此外,对于仿生学习系统而言,如果控制效果不佳或者出现新的状态模式时,要求其具有高度的自适应能力,而单纯基于概率自动机构建的学习系统不具备这种能力。遗传算法(genetic algorithm, GA)是模拟生物的进化现象(自然淘汰、交叉、变异等),并采用自然进化机制来表现复杂现象的一种概率搜索方法[14]。GA 应用于机器学习领域,其宗旨在于使复杂环境中行走的机器人像动物一样具有高度学习能力。基于 GA 获得的这种学习能力,不仅能够处理已确定的状态模式,而且能够处理新出现的状态模式。

本文以 PA 为平台,结合 GA 的进化思想,设计了反映 Skinner OC 机理的仿生学习模型,称为基于遗传算法的操作条件概率自动机(genetic algorithm-operant conditioning probabilistic automaton, GA-OCPA)学习系统。该学习系统不仅能仿生地学习到最优的操作行为,而且还可以对操作行为集合进行优化。学习系统的学习是一个自动知识获取和提炼的过程,具有高度的自适应能力。整个学习过程可分为两个阶段,第一阶段主要基于遗传算法对代表操作行为集合的个体进行寻优,在遗传的种群中,每一个个体对操作行为集合进行编码,并且每一个个体和一个信息熵值密切相关,把信息熵作为遗传算法进化的适应度,反映的是个体的自组织程度。第二个阶段则是基于操作 OC 学习机制从最优个体中学习最优的操作行为,作用于环境后收到环境的反馈信息,基于该反馈信息更新 OC 学习机制,进而更新信息熵值。本文从理论上分析了 GA-OCPA 学习系统学习算法的收敛性,通过对两轮直立式机器人运动平衡控制的仿真分析,验证了 GA-OCPA 学习系统的可行性和有效性。

## 1　仿生学习系统设计

仿生学习是一种生物体自发的主动行为,它是通过生物体自身的努力而获得、形成的行为,可用产生式系统的规则表示,这种表示虽然简单,但计算完备,便于处理,其形式为

IF　< condition >　THEN　< action >

指当条件满足时,即规则被触发,就可能采取行动。所以,仿生学习的目标就是学习一个操作规则集合。和模糊控制不同的是,操作规则是通过随机学习得到的。仿生学习的本质是:如果在一定的主客观条件下,生物的某种操作性行为(或操作)所导致的后果符合生物的取向性,那么在类似的主客观条件下,生物实施类似的行为(或操作)的概率将会上升。

以 PA 为平台,结合 GA 的进化思想,构建的 GA-OCPA 仿生学习系统的结构如图 1 所示,图中的未知系统相当于环境,GA-OCPA 学习系统类似于智能体,通过与环境的交互来适应环境。

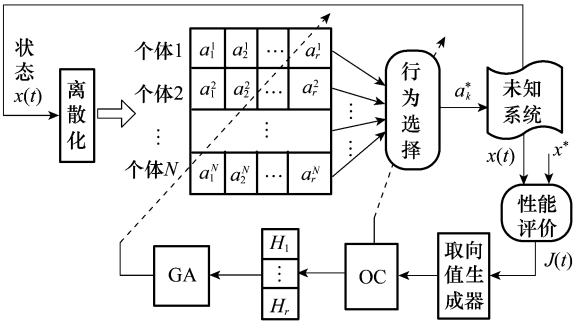



图 1　GA-OCPA 学习系统结构

整个 GA-OCPA 学习系统的学习过程主要分为两个任务,这两个学习任务的具体实现过程如下:

(1) 第一个学习任务:基于 GA 学习最优的个体,即:最优的操作行为集合。

种群中的一个个体代表一个操作行为集合,在这个学习过程中,个体是通过学习进化而得到的。这样不仅可以节省大量确定操作行为集合的实验时间;而且操作行为集合内部的行为个数不必太多,加速学习收敛的时间。由于个体具有进化能力,所以大大加强了学习系统的自适应性。在 GA 中,种群中的每个个体对操作行为集合进行编码。每个个体都有一个相应的信息熵值,采用信息熵值对种群推荐的个体进行评价。在每一个学习步,种群中具有最大信息熵值的个体被选中,作为最优的操作行为集合。

(2) 第二个学习任务:基于 OC 学习最佳的"状态-行为"映射规则。在 OC 中,从学习得到的最优的操作行为集合中,学习某条件状态对应的最佳的操作行为,作为控制系统的控制信号。每一个操作行为有一个相应的概率值,概率值的目的是对个体中的操作行为进行评价。在每一个学习步,操作行为集合中具有最大概率值的操作行为被选中的频次最高。

OC 学习算法的操作行为作用于环境后,首先,利用环境反馈的性能评价信息,计算状态取向值;然后,依据取向值信息,更新 OC 学习算法,调整操作行为的概率矢量。由调整的概率矢量获得操作行为集合的新信息熵值,作为遗传算法个体的适应度。执行以新信息熵值作为适应度的 GA,产生一个新的种群,新的个体取代具有低信息熵值的个体,如此循环,直至学习到最优的个体和最优的操作行为。

由此可见,学习系统的两个学习任务的实现虽然采用的是不同算法,但是两个学习过程互相影响,互相制约。

GA 学习的结果决定了 OC 学习的操作行为区域，而 OC 学习的结果将导致信息熵值的变化，也即个体适应度的变化，进而影响 GA 学习的结果。

按照概率自动机定义的形式，图 1 所示的 GA-OCPA 学习系统的数学定义为

**定义 1** GA-OCPA 学习系统是一个九元组计算模型，表示为：GA-OCPA=$<x,S,Q,\Gamma,f,\varphi,L,H,G>$，各部分说明如下：

(1) GA-OCPA 学习系统的内部状态 $x(t)$，为检测到的控制系统的实际状态值。

(2) GA-OCPA 学习系统的内部离散状态集合为 $S=\{s_i \mid i=1,2,\cdots,n\}$，$S$ 为系统所有可能的离散状态组成的非空集，$s_i \in S$ 表示第 $i$ 个离散状态，$n$ 为离散化的个数。

离散状态集合 $S$ 是系统实际检测到的连续状态 $x$ 离散化的结果。本文采用一个很简单的离散化方法：依据 $x$ 实际的取值范围，在闭区间$[s_{\min},s_{\max}]$内等分为 $n$ 个区段，$n=\frac{s_{\max}-s_{\min}}{w}$，$s_{\min}$ 为实际状态 $x$ 的下限值，$s_{\max}$ 为实际状态 $x$ 的上限值；$w$ 为每个区段的长度；一个区段代表一个离散化状态 $s_i$。

(3) GA-OCPA 学习系统的种群为 $Q=\{A_j \mid j=1,2,\cdots,N\}$，$A_j$ 表示种群的第 $j$ 个个体，一个种群产生 $N$ 个个体，信息熵值用来对个体进行评价。个体 $A_j$ 即为学习系统的操作行为集合，$A_j=\{a_{jk} \mid k=1,2,\cdots,r\}$，其中 $a_{jk}$ 表示第 $j$ 个个体中的第 $k$ 个可选操作行为。每个个体对 $r$ 个操作行为进行编码。

个体中的可选操作行为是通过随机选取概率 $P_j^i$ 决定的，$p_{jk}^i \in P_j^i=\{p_{j1}^i,p_{j2}^i,\cdots,p_{jr}^i\}$ 表示个体中行为 $a_{jk}$ 被选取的趋势。对于具有高概率值的操作行为，认为该行为能获得较低的取向值。在每一个学习步中，基于概率 $P_j^i$，$r$ 个行为中的一个被随机选中。因此，个体和操作行为选取互相竞争的映射规则表示为

$$\begin{aligned}R_i(P_j^i)&:\text{ If } s_i(t),\ \text{Then } A \text{ is } A_j \text{ with}\\ &H_{\min} \text{ and } a \text{ is } a_{j1}(t) \text{ with } p_{j1}^i\\ &\text{or } a \text{ is } a_{j2}(t) \text{ with } p_{j2}^i\\ &\cdots\cdots\\ &\text{or } a \text{ is } a_{jr}(t) \text{ with } p_{jr}^i\end{aligned}$$

其中，$H_{\min}$ 表示个体最小的信息熵值。

由于当学习得到的操作行为 $a$ 足够逼近 $a^*$ 时，学习系统随机选取行为的探索能力，将使得系统的输出不再稳定。所以对 $a$ 采取滤波的操作，可以得到更精确、更平滑的输出响应。本文采用对各操作行为按概率加权求和的方法进行滤波，所以，GA-OCPA 学习系统的最终输出为

$$a^*(t)=\sum_{k=1}^{r}a_{jk}(t)p_{jk}^i \tag{1}$$

(4) GA-OCPA 学习系统的“状态-行为”映射规则为 $\Gamma=\{R_i(P_j^i)\}(P_j^i \in P^i=\{P_1^i,\cdots,P_N^i\},R_i \in R=\{R_1,\cdots,R_u\})$。$R_i(P_j^i)$ 为状态处于 $s_i(t)$ 的条件下，在最优的个体 $A_j$ 内，学习系统依据 $P_j^i$ 实施操作行为 $a_{jk} \in A_j$。$p_{jk}^i(a_{jk} \mid x) \in P_j^i=\{p_{j1}^i,p_{j2}^i,\cdots,p_{jr}^i\}$ 表示 GA-OCPA 学习系统在状态 $s_i(t)$ 的条件下实施操作行为 $a_{jk}$ 的概率值($0<p_{jk}^i<1$，$\sum_{k=1}^{r}p_{jk}^i=1$)。

(5) GA-OCPA 学习系统的状态转移函数为 $f:s_i(t)\times a_{jk}(t)\rightarrow s_i(t+1)$，$t+1$ 时刻的状态 $s_i(t+1)$ 由 $t$ 时刻的状态 $s_i(t)$ 和 $t$ 时刻的操作行为 $a_{jk}(t)$ 确定，与 $t$ 时刻之前的状态和操作行为无关。

(6) GA-OCPA 学习系统的状态取向函数为 $\varphi=\{\varphi_1,\varphi_2\cdots,\varphi_n\}$，$\varphi_i$ 表示对状态 $s_i(t)$ 的取向值，表征的是对状态的取向程度($0\leqslant\varphi_i\leqslant1,i=1,2,\cdots,n$)。基于上述条件，设计的取向值表达式为

$$\varphi_i(t)=\left|\frac{e^{\gamma\varepsilon_i(t)}-e^{-\gamma\varepsilon_i(t)}}{e^{\gamma\varepsilon_i(t)}+e^{-\gamma\varepsilon_i(t)}}\right| \tag{2}$$

式中，$\varepsilon_i(t)=\dot{e}_i(t)+\zeta e_i(t)$，$e_i(t)=x_i(t)-x_{id}$，$x_{id}$ 为期望状态值，$\zeta$ 为误差的权重系数；$\gamma$ 为取向值系数。当取向值趋于 0 时，表示学习性能好，对该状态的取向程度高；反之，当取向值趋于 1 时，表示学习性能差，对该状态的取向程度低。所以，利用取向值的变化趋势来更新概率矢量 $P_j$。

(7) GA-OCPA 学习系统的 OC 学习算法为 $L:\Gamma_j(t)\rightarrow\Gamma_j(t+1)$，由 OC 学习算法实现选取最优的操作行为。

(8) GA-OCPA 学习系统的信息熵为 $H^i=\{H_1^i,H_2^i,\cdots,H_N^i\}$ 是处于状态 $s_i(t)$ 下的操作行为熵，$H_j^i(t) \in H^i$ 表示第 $j$ 个操作行为集合 $A_j$ 的行为熵，为

$$\begin{aligned}H_j^i(t)&=H_j^i(A_j(t) \mid s_i(t))=\\ &-\sum_{k=1}^{r}p_{jk}^i(a_k \mid s_i(t))\log_2 p_{jk}^i(a_k \mid s_i(t))\end{aligned} \tag{3}$$

这里的“操作行为熵”就是度量系统无组织程度的“信息熵”，“信息熵”的减小是系统自组织的特征。当所有操作行为 $a_{jk}$ 可能出现的概率相等时，操作行为熵最大。操作行为熵越大，说明操作行为 $a_{jk}$ 的不确定性越大，系统获得的信息越少。所以，把信息熵值作为个体的适应度，信息熵值越小，说明个体的自适应度越高，通过 OC 学习的结果来反馈 GA 学习的效果。

(9) GA-OCPA 学习系统的 GA 为 $G:s_i(t)\rightarrow A_j$，由 GA 通过进化得到最优个体 $A_j$。$A_j$ 的适应度值用信息熵 $H_j^i(t)$ 表示，但关系为反比关系，信息熵值越大，说明适应度越小，即：最优个体 $A_j$ 的信息熵值满足 $H_j^i=\min(H_1^i,H_2^i,\cdots,H_N^i)$。

## 2 学习算法设计

### 2.1 遗传算法设计

GA 以操作行为的信息熵值作为个体的自适应度，对操作行为集合进行训练。本文设计的 GA 与传统的 GA 相比，主要区别在于：前者以信息熵值作为个体的适应度，所以每经过一次实验就会产生新的一代；而后者，需要经过种群中的 $N$ 个个体的 $N$ 次实验后，才产生新的一代。GA 包括 3 个主要操作：选择、交叉和变异。GA 的学习过程如下：

(1) 编码

为了提高遗传算法的搜索效率,采用实数编码的方法。第 $j$ 个个体具有形式:$|a_{j1}|a_{j2}|a_{j3}|\cdots|a_{jr}|$。在进行 GA 之前,首先对交叉概率 $P_c$、变异概率 $P_m$ 进行初始化。在每一个学习步,种群的一个个体被选中,并把该个体作为 OC 学习机制的操作行为集合。将通过 OC 学习算法学习到的最优操作行为作用于环境,采用环境反馈的信息对 OC 机制进行调整,以实现对所有个体熵值的更新。尽管个体熵值不会在一次实验中收敛,但它们从一定程度上象征了个体适应度的大小程度。由这些个体的适应度值,在一次实验之后产生一个新的种群。

(2) 选择

选择主要是模仿自然选择中适者生存的现象。适应度相当于自然界中一个生物适应环境的各项能力的大小,它决定了个体的生存或淘汰。本文采用竞争选择的方法来实现选择,在竞争选择中,种群中的两个或多个个体被选中,对选中的个体的适应度进行比较,具有最高适应度的个体作为一个父代,其他父代选取的方法相同。

(3) 交叉

交叉概率决定了被选中的两个父代是否交叉或直接繁殖下一代。本文采用的是单点交叉方法,也就是随机的选中一个交叉点,两个个体中这个点之前的子代部分互相交换。

选择和交叉后,性能差的个体将会被新产生的子代所代替。一般来说,新个体的信息熵值大于旧个体,这意味着获得了具有更佳性能的个体。

(4) 变异

变异发生时,新产生的基因代替旧的基因。本文采用最简单的位点(基本位)变异法。

### 2.2 操作条件反射算法设计

**定义 2**　状态 $s_i(t)$ 被奖赏的概率为

$$d_{ik}(t) = w_{ik}(t) / z_{ik}(t) \tag{4}$$

式中,$z_{ik}(t)$ 表示在状态 $s_i(t)$ 的条件下,行为 $a_{jk}$ 被选择的次数;$w_{ik}(t)$ 表示行为 $a_{jk}$ 被奖赏的总和。引入奖赏概率的目的是考虑了环境的动态性,此外奖赏概率可以视为对行为的一个跟踪评价信号。

假设 $t$ 时刻的状态为 $s_i(t)$,实施操作行为 $a_{jk}^*(t)$ 后,$t+1$ 时刻观测到的状态为 $s_{i'}(t+1)$。按照 Skinner OC 理论,如果实施操作行为 $a_{jk}^*(t)$ 后,相邻时刻状态的取向值之差满足 $\varphi_{i'}(t+1)-\varphi_i(t)<0$,则该行为的实施概率 $p_{jk}(t)$ 倾向于增大;反之,如果 $\varphi_{i'}(t+1)-\varphi_i(t)>0$,则行为 $p_{jk}(t)$ 倾向于减小。因此,OC 学习算法可以形式化地描述为

(1) $d(t)$ 更新算法

IF $a(t)=a_{jk}$

THEN $w_{ik}(t+1)=w_{ik}(t)+1/\varphi(t)$

$$z_{ik}(t+1) = z_{ik}(t)+1 \tag{5}$$

因此,$d_{ik}(t+1)=\dfrac{w_{ik}(t+1)}{z_{ik}(t+1)}$;

IF $a(t)=a_{jk'}$ and $k\neq k'$

THEN $w_{ik}(t+1)=w_{ik}(t)$

$$z_{ik}(t+1) = z_{ik}(t) \tag{6}$$

因此,$d_{ik}(t+1)=d_{ik}(t)$。

(2) 行为概率 $p(t)$ 更新算法

IF $\phi_{i'}(t+1)-\phi_i(t)<0$

THEN $p_{jk}^i(t+1)=p_{jk}^i(t)+\Delta_1$, $a(t)=a_{jk}$

$$p_{jk'}^i(t+1) = p_{jk'}^i(t)-\Delta_1',\ a(t)\neq a_{jk} \tag{7}$$

增量部分设计为

$$\Delta_1 = \alpha(t)[1-p_{jk}^i(t)]$$
$$\Delta_1' = \alpha(t)p_{jk'}^i(t)$$

其中,$\alpha(t)=\dfrac{\eta_1}{1+\exp[\varphi_{i'}(t+1)]}$。

IF $\phi_{i'}(t+1)-\phi_i(t)>0$

THEN $p_{jk'}^i(t+1)=p_{jk'}^i(t)-\Delta_2' \quad a(k)=a_{jk'}$

$$p_{jk}^i(t+1) = p_{jk}^i(t)+\Delta_2 \quad a(k)\neq a_{jk} \tag{8}$$

增量部分设计为

$$\Delta_2' = \beta(t)p_{jk'}^i(t)$$
$$\Delta_2 = \beta(t)\left[\frac{1}{r-1}-p_{jk}^i(t)\right]$$

其中,$\beta(t)=\dfrac{\eta_2}{1+\exp[\varphi_{i'}(t+1)]}$。

式中,$\eta_1>0$,$\eta_2>0$;$\alpha(t)$、$\beta(t)$ 为学习速率函数,$0<\alpha(t)<1$,$0<\beta(t)<1$。

把取向值函数 $\varphi_i(t)$ 融入到概率更新公式中,不仅起到影响学习速度的作用,而且更能使学习系统体现出类似于动物的取向特性。例如,当某状态的取向值增大时,$\alpha(t)$ 值变小,导致行为概率更新的幅度减小,从而减慢了学习的速度;反之,当某状态的取向值减小时,将加快学习速度。

概率矢量更新后,第 $j$ 个个体的熵值更新为

$$H_j^i(t+1) = H_j^i(A_j(t)\mid s_i(t)) =$$
$$-\sum_{k=1}^{r} p_{jk}^i(t+1)\log_2 p_{jk}^i(t+1) \tag{9}$$

### 2.3 学习算法收敛性证明

**定理 1**　当 $t\to\infty$ 时,GA-OCPA 学习系统依概率 1 选取具有最大奖赏概率的操作行为

$$\lim_{t\to\infty} p_{jm}^i(a_{jm}(t)\mid s_i(t))\to 1 \tag{10}$$

式中,$m$ 是具有最大奖赏概率 $d_{jm}(t)$ 的操作行为的索引。

**证明**　假定 $t$ 时刻,GA-OCPA 学习系统选取操作行为 $a_{jk}(t)$,则

(1) 若 $a(t)=a_{jk}(t)$,由式(7)得

$$\Delta p_{jk}^i(t) = p_{jk}^i(t+1)-p_{jk}^i(t) =$$
$$\alpha(t)-\alpha(t)p_{jk}^i(t) = \alpha(t)(1-p_{jk}^i(t)) \tag{11}$$

因为 $0<p_{jk}^i(t)<1$,$0<\alpha(t)<1$,所以可得 $\Delta p_{jk}^i(t)\geqslant 0$;

(2) 若 $a(t)=a_{jk}(t)$,由式(8)得

$$\Delta p_{jk}^i(t) = p_{jk}^i(t+1)-p_{jk}^i(t) =$$
$$\beta(t)-\beta(t)p_{jk}^i(t) = \beta(t)(1-p_{jk}^i(t)) \tag{12}$$

因为 $0<\beta(t)<1$,所以得 $\Delta p_{jk}^i(t)\geqslant 0$。

由式(11)和式(12)知,操作行为 $a_{jk}(t)$ 的概率增量 $\Delta p_{jk}^i(t)\geqslant 0$,所以当 $t\to\infty$ 时,$a_{jk}(t)$ 被选中的频次逐渐增加直至趋于无穷,这意味着 $\Delta p_{jk}^i(t)\geqslant 0$ 的情形可发生任意多

次，而且发生的频次越来越高，因此 $p_{jk}^{i}(t)$将不断升高。此外，由于 $\Delta p_{jk}(t)\geqslant 0$ 的等号仅当 $p_{jk}^{i}(t)=1$ 时成立，所以 $p_{jk}^{i}(t)$的增长将直至 $p_{jk}^{i}(t)=1$ 为止，所以必存在一个最优行为 $a_{jm}$，使$\lim\limits_{t\to\infty} p_{jm}^{i}(a_{jm}(t)\mid x_i(t))\to 1$ 成立。 证毕

**定理 2** 当 $t\to\infty$时，GA-OCPA 学习系统处于状态 $s_i$ 的操作行为熵 $H_j^i(\{A_j\}\mid s_i)$收敛至极小，即

$$\lim_{t\to\infty} H_j^i(t)=H_{j\min}^i \tag{13}$$

**证明** 由于 GA-OCPA 学习系统是一个基于 Skinner OC 理论的自学习自组织系统，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。所以，若处于状态 $s_i$ 下的操作行为熵$\lim\limits_{t\to\infty} H_j^i=H_{j\min}^i$，则说明 GA-OCPA 学习系统具有自组织特性，同时也表明个体的适应度达到最大值。

当所有操作行为 $a_{jk}(t)$可能出现的概率 $p_{jk}^{i}(t)$相等时，操作行为熵最大。所以，一般在学习的初始时刻，所有操作行为 $a_{jk}(t)$选取相同的选择概率 $p_{jk}^{i}(t)$，对式(3)重新进行整理得

$$\begin{aligned}H_j^i(t)&=H_j^i(A(t)\mid s_i)=\\&-\sum_{k=1}^{r} p_{jk}^{i}(a_{jk}\mid s_i)\log_2 p_{jk}^{i}(a_{jk}\mid s_i)=\\&-\left[p_{jm}^{i}(a_{jm}\mid s_i)\log_2 p_{jm}^{i}(a_{jm}\mid s_i)+\right.\\&\left.\sum_{k'=1,k'\neq m}^{r} p_{jk'}^{i}(a_{jk'}\mid s_i)\log_2 p_{jk'}^{i}(a_{jk'}\mid s_i)\right]\end{aligned} \tag{14}$$

由定理 1 可知$\lim\limits_{t\to\infty} p_{jm}^{i}(a_{jm}(t)\mid s_i(t))\to 1$，对应的很容易得 $\sum\limits_{k'=1,k'\neq m}^{r} p_{jk'}^{i}(a_{jk'}(t)\mid s_i(t))\to 0$，代入式(14)得

$$\begin{aligned}H_j^i(\infty)=\lim_{t\to\infty}&\left[-p_{jm}^{i}(a_{jm}\mid s_i)\log_2 p_{jm}^{i}(a_{jm}\mid s_i)-\right.\\&\left.\sum_{k'=1,k'\neq m}^{r} p_{jk'}^{i}(a_{jk'}\mid s_i)\log_2 p_{jk'}^{i}(a_{jk'}\mid s_i)\right]\to 0\end{aligned} \tag{15}$$

证毕

### 2.4 学习流程

基于 GA-OCPA 学习系统，实现两轮机器人运动平衡控制的学习流程如下：

**步骤 1** 初始化

迭代学习步数 $t=0$。

GA 部分：采用实数编码方式，随机产生由 $r$ 个初始数据构成的个体 $A_j=\{a_{jk}\mid k=1,2,\cdots,r\}$，若干个个体构成一个初始种群 $N$。

OC 部分：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，GA-OCPA 学习系统不含有任何预定的决策，其采用任何决策的概率是相等的。

**步骤 2** 基于 GA 训练最优个体 $A_j$

(1) 由初始的概率值，带入式(3)，可得到初始熵值，作为个体的适应度值；

(2) 遗传操作：包括选择、交叉、变异等操作。经遗传操作，选择出最优个体(假设为 $A_j$)，作为 OC 学习算法的操作行为集合。

**步骤 3** 基于 OC 选取最优的操作

(1) 从步骤 2 得到的操作行为集合 $A_j$ 中，依据操作行为概率，选择操作行为(假设为 $a_{jk}(t)\in A_j$)；

(2) 实施操作行为 $a_{jk}(t)$，观测 $t+1$ 时刻的状态，计算取向值增量：$\tilde{\varphi}(t)=\varphi(t+1)-\varphi(t)$，并判断正负；

(3) 按照式(7)和式(8)设计的 OC 学习算法，调节操作行为 $a_{jk}(t)$的实施概率 $P_{jk}^{i}(t+1)$。判断是否结束一次实验，若结束，则转步骤 4，反之，转步骤 2。

**步骤 4** 判断是否满足停止准则

停止准则一般设计为 $|e(t)|<\kappa$($\kappa$ 为一个很小的正实数)，若满足，转向步骤 5；否则，依据更新的概率，计算个体的适应度值，并产生新一代群体，转步骤 2。

**步骤 5** 结束

## 3 仿真实验

### 3.1 两轮机器人系统

本文以某人工智能与机器人研究所研制的两轮机器人为研究对象，简化图如图 2 所示。图中，$m_P$ 为车体的重量；$\theta$ 为杆偏离垂直方向的角度；$L$ 为质心距车轮轴中心的距离。

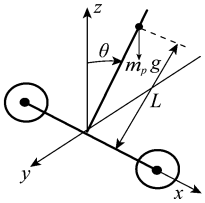



图 2 两轮机器人系统简化图

基于 GA-OCPA 学习系统实现两轮自平衡机器人运动平衡控制的结构如图 3 所示，机器人通过倾角传感器和陀螺仪获得倾角 $\theta$ 和角速度 $\dot{\theta}$ 的模拟值，通过编码器获得左轮速度 $v_l$ 和右轮速度 $v_r$ 的模拟值，经 A/D 转换，采用 GA-OCPA学习系统对检测的信息进行处理，以得到控制机器人的左驱动电机和右驱动电机的控制信号。

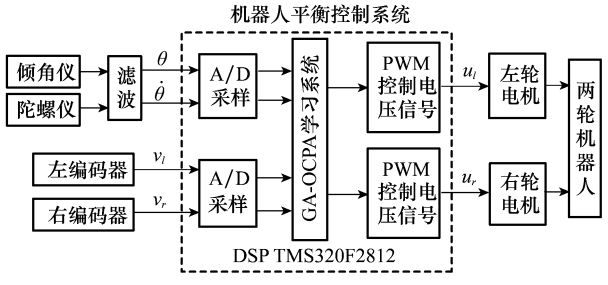


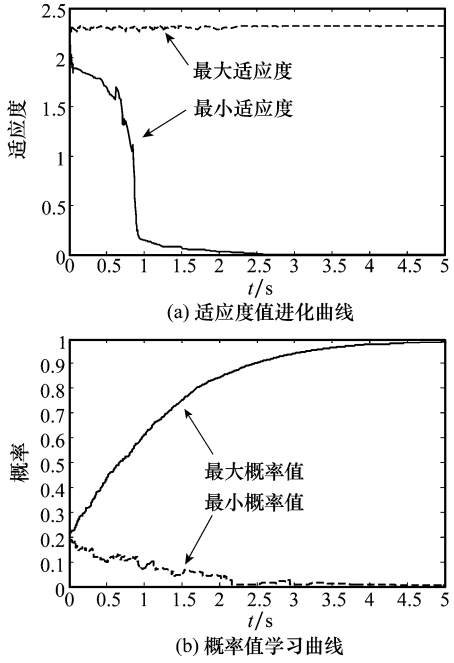

图 3 两轮自平衡机器人的控制结构

### 3.2 仿真实验和分析

取采样时间 $t_s=0.02$ s，实际状态的上限值 $s_{\max}=0.3$ rad，下限值 $s_{\min}=-0.2$ rad，每个区段的长度 $w=0.1$ rad，所以输入状态离散化的数目 $n=5$，种群中个体的个数 $N=20$，每个个体内操作行为的个数 $r=5$，在$[-10,10]$之间随机生成，操作行为的初始概率 $p_{jk}(0)=\dfrac{1}{5}$，对应的初始熵 $H_j(0)=-\sum\limits_{j=1}^{5}\dfrac{1}{5}\times\log_2\dfrac{1}{5}=2.32$；取向值计算公式中，$\zeta=0.8$，$\gamma=0.03$；OC 学习算法公式中，$\eta_1=0.05$，$\eta_2=0.01$；

停止准则中 $\kappa=0.001$;初始状态$[\theta \quad \dot{\theta}]=[0.2\ \text{rad}\ 0\ \text{rad/s}]$。

在仿真中进行了 30 次实验,因为本文设计的 OC 学习算法是模拟动物的一种仿生学习算法,所以有失败的可能,但主要发生在实验初期,随着学习经验的积累,成功的概率越来越高,最后趋于 1。所以,从中抽取一次成功实验,来分析 GA-OCPA 学习系统的学习效果。图 4 为适应度和概率值的仿真结果。




(a) 适应度值进化曲线

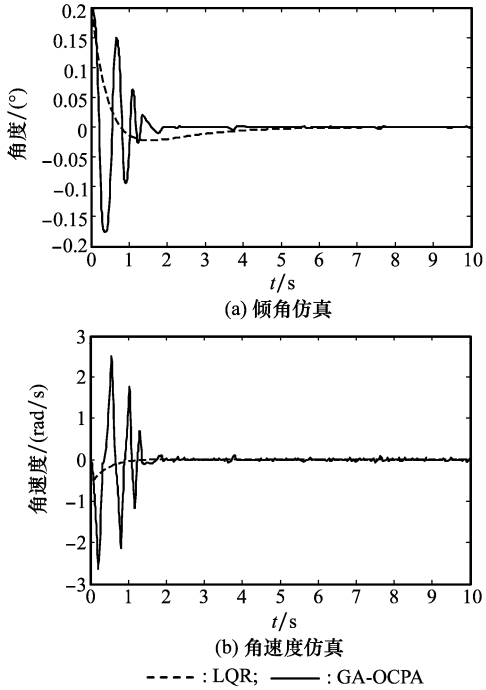
(b) 概率值学习曲线

图 4　适应度值和概率值的变化曲线

图 4(a)显示了种群在进化过程中历代最大个体适应度(对应最小熵值)和最小个体适应度(对应最大熵值)的变化曲线。由结果可以看出,若个体被选中的次数增多,则导致对该个体内操作行为选取的机率增大,当学习到最优操作行为时,由于其选择概率接近 1,因此该个体的熵值接近于 0,对应个体的自适应度达最大值。反之,若某个体被选中的次数少,则其内操作行为概率更新的次数少,概率变化不大,对应的个体的熵值也基本保持不变,始终在最大值上下浮动,最后达最大值。图 4(a)的结果表明,当进化到一定代数后,GA 算法逐渐收敛,个体的最大适应度值到达稳定,此时能够得到优化的个体(操作行为集合)。

图 4(b)显示了操作行为集合在学习过程中最大操作行为概率值和最小操作行为概率值的变化曲线。由结果可以看出,随着学习的进行,会出现选取概率逐渐增加的操作行为,当其概率增长到 0.8 左右时,已基本能实现成功控制机器人,最终该操作行为概率趋于 1。这表明随着学习经验的积累,OC 学习算法逐渐收敛,操作行为的最大概率值基本到达稳定,由一开始的随机学习转移到确定性的学习,此时能够得到最优的操作行为(控制信号)。同时也表明 OC 的学习过程是一个变化的动态学习过程,随着学习经验的积累,其学习的策略也在更新。也就是说初始阶段主要还是探索的阶段,随着学习的进行,探索性减少,更多的体现出的是自组织和自适应性。

机器人运动平衡控制实验中,首先只利用倾角仪和陀螺仪检测到的倾角和角速度信息来实现机器人的自由平衡控制模式。机器人的倾角和角速度仿真结果如图 5 所示,为了验证 GA-OCPA 学习系统的性能,同线性二次型调节器(linear quadratic regulator, LQR)控制进行了比较,图 5(a)为倾角仿真结果,图 5(b)为角速度仿真结果。




(a) 倾角仿真

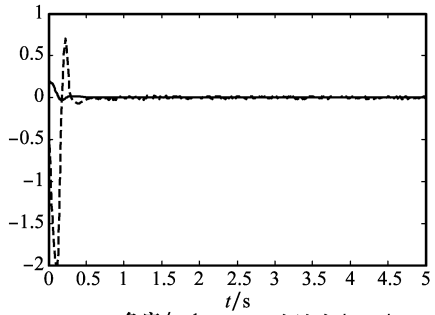
(b) 角速度仿真

图 5　倾角和角速度仿真实验

由图 5 可以看出,初始阶段,由于 GA-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,大约 2 s 后机器人的角度恢复到 0;LQR 控制与其相比,虽然初始阶段振荡较小,但大约经过 3 s 后机器人的角度才恢复到 0。这说明在本文设计的方案下,机器可以快速的自主学习到最优的操作行为,成功实现自平衡控制。

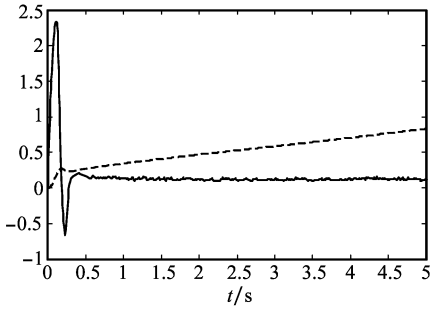
为了进一步验证 GA-OCPA 学习系统处理机器人不同运动模式的能力,加入编码器检测到的轮子的速度信息,实现机器人的直线运动和转向运动模式。图 6 给出了机器人直线运动的仿真结果,直线运动中,两个轮子给定相同的跟踪速度 0.15 m/s。由图 6 可以看出,姿态角大约 0.6 s 就迅速恢复到 0,同时两个轮子也跟踪上了给定的速度,误差均在允许范围内。同时也可以看到,图 6(a)中姿态平衡控制的效果和图 5(a)相比,在响应时间和控制精度上明显变好,这是加入编码器检测到的速度信息所致。

图 7 所示结果为机器人转向运动时的仿真结果,给定右轮跟踪速度为 0.6 m/s,左轮跟踪速度为 0.15 m/s。由仿真结果可以看出,大约 1 s 后,姿态角恢复到 0;左轮和右轮也都跟踪上了期望的速度,并且误差都在允许范围内。由于给定两个轮子的速度不相等,所以会发生转向运动,在 $X$-$Y$ 相平面的轨迹是一个圆。




——: 角度/rad; - - - -: 角速度/(rad/s)

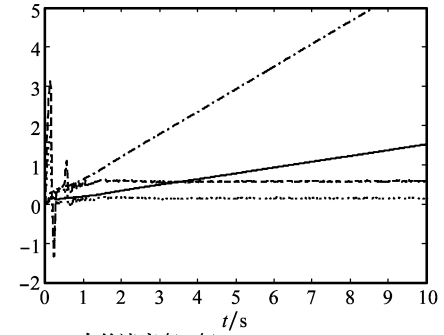
(a) 倾角和角速度仿真




——: 速度/(m/s); - - - -: 位移/m

(b)位移和速度仿真

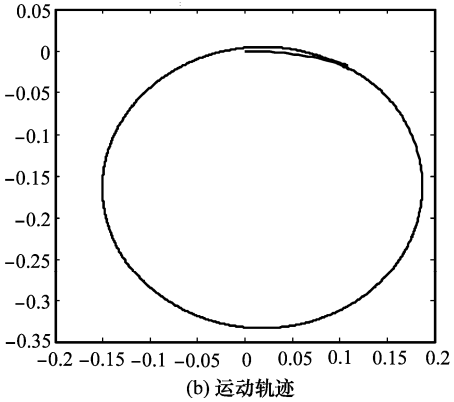
图 6 直线运动仿真实验




·····: 左轮速度/(m/s); - - - -: 右轮速度/(m/s);

——: 左轮位移/m; - · - · -: 右轮位移/m

(a) 位移和速度仿真



(b) 运动轨迹

图 7 转向运动仿真实验

由图 5～图 7 还可以看出,仅采用本文设计的GA-OCPA学习系统,就可以同时实现机器人的自由平衡、直线运动和转向运动 3 种运动模式,而这是学习自动机或强化学习不能实现的。因为,每种运动模式对应的行为集合和强化信号都不相同,所以采用学习自动机和强化学习的方法,必须针对每种运动模型分别设计对应的学习方案。

## 4 结束语

本文结合 GA 的进化思想,设计了反映操作条件反射机理的 GA-OCPA 学习系统,该学习系统采用具有进化特性的遗传算法,以信息熵值作为适应度,训练最优个体;采用具有仿生特性的 OC 学习算法,以取向值作为对学习好坏程度的评价,学习最优操作行为。学习系统的学习是一个自动获取知识和提炼的过程,并且仅仅需要取向值这一微弱的信息,大大降低了对教师信号质量和数量的要求。理论分析和仿真实验结果均验证了 OC 学习算法的收敛性。对两轮机器人进行的自由平衡控制、直线运动和转向运动 3 种运动模式的仿真实验,表明仅仅采用本文设计的 GA-OCPA 学习系统,就可以实现机器人的各种运动模式,体现出较强的自适应能力。

## 参考文献:


[1] Yilddirim S. Design of adaptive robot control system using recurrent neural network[J]. *Journal of Intelligent & Robotic Systems*, 2005, 44(3): 247 - 261.

[2] Floreano D, Mondada F. Evolutionary neuro-controller for autonomous mobile robots[J]. *Neural Networks*, 1998, 11(7 - 8): 1461 - 1478.

[3] 蒋宗礼,姜守旭. 形式语言与自动机理论[M]. 北京: 清华大学出版社, 2007: 35 - 63. (Jiang Z L, Jiang S X. *Formal language and automata theory*[M]. Beijing: Tsinghua University Press, 2007: 35 - 63.)

[4] Holocombe W M L. *Algebraic automata theory*[M]. London: Cambridge University Press, 1982: 25 - 42.

[5] Sutton R S, Barto A G. *Reinforcement Learning*[M]. London: MIT Press, 1998: 1 - 12.

[6] Kondo T, Ito K. A reinforcement learning with evolutionary state recruitment strategy for autonomous mobile robots control[J]. *Robotics and Autonomous Systems*, 2004, 46(2): 111 - 124.

[7] Skinner B F. *The behavior of organisms*[M]. New York: Appleton-Century-Crofts, 1938: 283 - 286.

[8] Skinner B F. Two types of conditioned reflex and a pseudo type[J]. *Journal of General Psychology*, 1935, 12(3): 66 - 77.

[9] Touretzky D S, Saksida L M. Skinnerbots[C]// *Proc. of the Fourth International Conference on Simulation of Adaptive Behavior*, 1996: 285 - 294.

[10] 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, 3209(15): 55 - 66.

[11] Touretzky D S, Saksida L M. Operant conditioning in Skinnerbots[J]. *Adaptive Behavior*, 1997, 5(3/4): 219 - 247.

[12] Touretzky D S, Daw N D, Thompson E J T. Combining configured and TD learning on a robot[C]// *Proc. of the 2nd International Conference on Development and Learning*, 2002: 47 - 52.

[13] 陶仁骥. 自动机引论[M]. 北京: 科学出版社, 1986: 43 - 62. (Tao R J. *Automata introduction*[M]. Beijing: Science Press, 1986: 43 - 62.)

[14] Juang C F. Combining of online clustering and Q-valued based GA for reinforcement fuzzy system design[J]. *IEEE Trans. on Fuzzy System*, 2005, 13(3): 289 - 302.