Journal of Computer Applications
计算机应用,2013,33(8):2283 -2288
ISSN 1001-9081
CODEN JYIIDU
2013-08-01
http://www.joca.cn

文章编号:1001-9081(2013)08-2283-06
doi:10.11772/j.issn.1001-9081.2013.08.2283

进化操作行为学习模型及在移动机器人避障上的应用

郜园园*,朱　凡,宋洪军
(浙江农林大学 信息工程学院,杭州 311300)
(*通信作者电子邮箱 yuangao84@163.com)

摘　要:针对移动机器人避障上存在的自适应能力较差的问题,结合遗传算法(GA)的进化思想,以自适应启发评价(AHC)学习和操作条件反射(OC)理论为基础,提出了一种基于进化操作行为学习模型(EOBLM)的移动机器人学习避障行为的方法。该方法是一种改进的AHC学习模式,评价单元采用多层前向神经网络来实现,利用TD算法和梯度下降法进行权值更新,这一阶段学习用来生成取向性信息,作为内在动机决定进化的方向;动作选择单元主要用来优化操作行为以实现状态到动作的最佳映射。优化过程分两个阶段来完成,第一阶段通过操作条件反射学习算法得到的信息熵作为个体适应度,执行GA学习算法搜索最优个体;第二阶段由OC学习算法选择最优个体内的最优操作行为,并得到新的信息熵值。通过移动机器人避障仿真实验,结果表明所设计的EOBLM能使机器人通过不断与外界未知环境进行交互主动学会避障的能力,与传统的AHC方法相比其自学习自适应的能力得到加强。

关键词:移动机器人;自适应启发评价;操作条件反射;遗传算法;避障
中图分类号:TP242　　**文献标志码**:A

Evolutionary operant behavior learning model and its application to mobile robot obstacle avoidance

GAO Yuanyuan*, ZHU Fan, SONG Hongjun
(*School of Information Engineering, Zhejiang Agriculture and Forestry University, Hangzhou Zhejiang* 311300, *China*)

Abstract: To solve the problem of poor self-adaptive ability in the robot obstacle avoidance, combined with evolution thought of Genetic Algorithm (GA), an Evolutionary Operant Behavior Learning Model (EOBLM) was proposed for the mobile robot learning obstacle avoidance in unknown environment, which was based on Operant Conditioning (OC) and Adaptive Heuristic Critic (AHC) learning. The proposed model was a modified version of the AHC learning architecture. Adaptive Critic Element (ACE) network was composed of a multi-layer feedforward network and the learning was enhanced by TD(λ) algorithm and gradient descent algorithm. A tropism mechanism was designed in this stage as intrinsic motivation and it could direct the orientation of the Agent learning. Adaptive Selection Element (ASE) network was used to optimize operant behavior to achieve the best mapping from state to actor. The optimizing process has two stages. At the first stage, the information entropy got by OC learning algorithm was used as individual fitness to search the optimal individual with executing the GA learning. At the second stage, the OC learning selected the optimal operation behavior within the optimal individual and got new information entropy. The results of experiments on obstacle avoidance show that the method endows the mobile robot with the capabilities of learning obstacle avoidance actively for path planning through interaction with the environment constantly. The results were compared with the traditional AHC learning algorithm, and the proposed model had better performance on self-learning and self-adaptive abilities.

Key words: mobile robot; Adaptive Heuristic Critic (AHC); operant conditioning; Genetic Algorithm (GA); obstacle avoidance

0　引言

移动机器人研究的最终目标是机器人能够在未知环境导航中通过不断增加经验改善行为而具备高度自治的能力。常用的移动机器人避障方法主要有人工势场法、环境地图法、神经网络法和模糊逻辑算法[1-4]。但由于已有算法不同程度地存在一定局限性,诸如搜索空间大、算法复杂、效率不高等,尤其对于未知环境,不少算法的复杂度会大大增加,甚至无法求解。而机器学习的方法为复杂环境的知识获取提供了有效的解决途径[5-7]。

与已有的监督学习和无监督学习方法不同的是,增强学习可以利用与环境的交互而获得的评价性反馈信号来实现系统优化的性能,是一种试错学习的方式。Sutton等[8]在Barto等的研究基础上提出了自适应启发评价(Adaptive Heuristic Critic, AHC)方法。AHC学习系统通常由自适应评价单元(Adaptive Critic Element, ACE)和动作选择单元(Adaptive Selection Element,ASE)组成。此后,一些学者对AHC学习算法作了进一步的研究,扩大了AHC的应用领域[9-13]。但是,因为传统的AHC方法没有生物学上的约束,只是为解决不同的问题而设计的,还不完全像动物学习。Touretzky等[14]指出

收稿日期:2013-02-26;**修回日期**:2013-05-07。
基金项目:浙江省青年科学基金资助项目(LQ13F030012);浙江农林大学人才启动项目(2013FR023)。
作者简介:郜园园(1984-),女,河南安阳人,讲师,博士,主要研究方向:移动机器人控制、机器学习、智能控制;　朱凡(1979-),女,河南南阳人,讲师,博士,主要研究方向:生物医学信息处理;　宋洪军(1981-),男,山东泰安人,博士研究生,主要研究方向:智能交通、机器视觉。

用增强学习方法训练的移动机器人还不完全与动物的先进性、功能性和适应性相像。而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,对实现机器人自学习、自适应能力具有重要的研究价值。Gutnisky 等[15]受神经生理学、心理学和动物行为学启发设计了一种学习避障的行为选择模型,为研究新的仿生学习方法提供了一种思路。本文正是在传统 AHC 方法基础上,引入神经心理学上操作条件反射理论和生物学上进化机制作为约束,建立一种仿生学习方法,使机器人像人或动物一样具有自主学习复杂环境的能力。

Touretzky 等[14]提出的操作条件反射(Operant Conditioning, OC)被认为是生物系统最基本的学习形式,增强学习的思想也来源于此。其核心内容为:某一操作行为一旦受到其结果的强化,则该行为发生的概率就会增加。操作条件反射这一概念的特点在于,它强调行为结果对行为的影响。自 20 世纪 90 年代中期开始,美国卡内基梅隆大学(CMU)机器人学研究所主要研究关于 Skinner OC 的计算理论和计算模型,期望这种模型能复制动物学习操作或控制的实验;然后在机器人上实现这种模型,使其成为可训练的机械[16]。1997 年,美国波士顿大学 Neurobotics 实验室的 Gaudiano 等[17]针对一个实际的轮式机器人 Khepera 的导航问题,建立了一个 Pavlov 理论与 Skinner 理论相结合的神经计算模型,Khepera 不需要任何先验知识和教师信号,即可在巡航过程中学习规避障碍。2005 年,日本早稻田大学机械工程系机器人研究小组 Itoh 等[18]为人性化机器人(Humanoid Robot)设计了一种基于 OC 操作条件反射的新行为模型,发展了 Hull 行为理论来作为 Sinner 操作条件反射理论的数学模型,并使 WE-4RII 能在其预先制定的行为列表范围内,自主地选择合适特定情景的行为模式。但是,这些计算理论和计算模型没有给出具体的数学计算模型,不具备泛化能力,应用受到了限制。

以概率自动机为平台,蔡建羡等[19]用其来模拟操作条件反射机制,设计了相应的仿生系统,给出了具体的数学计算模型,并成功实现了两轮机器人的平衡控制;同时蔡建羡等还把 OC 与 GA 相结合,提出了一种操作条件反射模型,并对其进行了初步的研究,用于解决两轮机器人自平衡问题。但在解决避障问题上引入生物学上的 GA 和 OC,还未见到相关的研究,本文以此为基础,在 AHC 学习框架下,引入了遗传算法的进化思想,提出了一种进化操作学习模型来模拟生物 OC 学习机制,使机器人像动物一样具有高度的自学习自适应能力。评价单元(ACE)采用多层前向神经网络来实现,用 TD(λ)算法和梯度下降法进行权值更新,提高了神经网络的学习速率。动作选择单元(ASE)由遗传算法优化的操作行为规则集合构成,分为两个学习阶段来完成:第一阶段通过操作条件反射学习算法得到的信息熵作为个体适应度,执行 GA 搜索最优个体,从而通过进化得到最优的操作行为集合;第二阶段由 OC 学习算法选择最优个体内的最优操作行为,并得到新的信息熵值,指导最优个体的生成。最后将本文方法应用于移动机器人学习避障行为中,使机器人在无教师信号指导的情况下,通过不断与环境交互来学习行为能力,从而实现在未知障碍物环境中进行无碰自由巡航,习得避障的能力。

1 进化操作行为学习模型结构设计

进化操作学习模型是基于操作条件反射的思想建立的,它是一种仿生的学习模式。本文构建的进化操作学习模型的结构如图 1 所示。评价单元(ACE)采用多层前向神经网络来实现,利用 TD(λ)算法和梯度下降法进行权值更新,其作用是根据外部的原始强化信号 r_t 及当前的状态信号 s_t 来对候选的动作进行评价,其输出为 $V(s_t)$,从而构成内部的二次强化信号 $\hat{r}_t$,在执行某一选择动作时,系统转移到新状态,ACE 单元的输出可用来评价策略的优劣。动作选择单元(ASE)主要用来生成输出动作,通过执行动作使环境状态发生改变,并同时获得来自于环境的外部强化信号 r_t。ASE 分为两个学习阶段来完成:第一阶段通过 OC 学习算法得到的信息熵作为个体适应度,执行 GA 搜索最优个体;第二阶段由 OC 学习算法选择最优个体内的最优操作行为,并得到新的信息熵值。

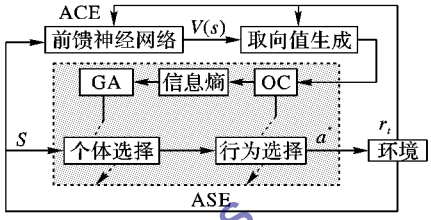

图 1 进化操作学习模型结构

进化操作学习模型主要有三个学习任务,具体实现过程如下:

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。GA 用来优化操作行为集合,即在给定的规则前提条件下,通过进化来得到与前提部分最匹配的结论部分。种群中的每个个体表示一个操作行为集合,在这个学习过程中,个体是通过进化学习得来的,这样可节省确定操作行为集合的实验时间,减小人为的干预,使其主动获取,大大增强了系统的自适应和自组织能力。在 GA 中,种群中的每个个体对操作行为集合进行编码。每个个体都有相应的信息熵值,采用信息熵值作为个体的适应度对个体进行评价,种群中具有最小信息熵值的个体作为最优操作行为集合,以作为下一步选择最优行为的动作集合。

第二个学习任务:基于 OC 学习最优行为,即状态到动作的最佳映射。在 OC 中,从上一步学习得到的最优操作行为集合中,通过随机概率学习获得最优的行为,作为系统的控制信号输出。每一个操作行为都有一个概率值与它对应,表示其被选择的几率,由取向性信息对概率值进行更新,操作行为集合中概率值最大的行为其被选择的次数越多,即认为是当前状态下最优的操作行为。

第三个学习任务:基于前馈神经网络生成取向性信息,即决定生物进化的方向。当执行最优操作行为后,系统转移到新状态,并由环境输出原始强化信号值,通过 TD(λ)和梯度下降法对前馈神经网络权值进行更新。前向神经网络输出行为动作评价值,来评价该操作行为的优劣,从而构成内部二次强化信号,由状态取向函数获得取向性信息,以作为 OC 学习中概率值更新的依据,决定进化的方向。

2 学习算法设计

2.1 ACE 网络学习算法

ACE 评价网络采用三层前向神经网络,根据多步瞬时差分的误差反向传播算法来推导各层的连接权值[20]。输入层和隐含层的神经元节点个数分别为 n 和 m,输出层节点个数为 1;网络的输出层分别与输入层和隐含层连接;$s_1, s_2, \cdots, s_n$ 为网络的输入,$y_1, y_2, \cdots, y_m$ 为隐含层的输出,输出层的输出为 $V(s)$;w, v, u 分别为输入层与隐含层、输入层与输出层、隐

含层与输出层的连接权值。

令 $\sum_{i=1}^{n} w_{ij}s_i(t) = \varphi_1$,可得到 $y_j(t) = f\left[\sum_{i=1}^{n} w_{ij}s_i(t)\right] = f(\varphi_1)$,其中 $f(\cdot)$ 为 sigmoid 函数,定义为

$$f(x) = \frac{1}{1+\exp(-x)} \tag{1}$$

令 $\sum_{i=1}^{n} v_i(t)s_i(t) + \sum_{j=1}^{m} u_j(t)y_j(t) = \varphi_2$,可以得到

$$V(s_t) = f\left[\sum_{i=1}^{n} v_i(t)s_i(t) + \sum_{j=1}^{m} u_j(t)y_j(t)\right] = f(\varphi_2) \tag{2}$$

基于时序差分(Temporal Difference,TD)误差构成二次强化信号为

$$\hat{r}_t = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{3}$$

其中 $\gamma(0 \leqslant \gamma \leqslant 1)$ 为学习算法的折扣率。下面采用误差反向传播算法进行网络权值更新调整。产生使用 TD 算法定义网络的性能函数,即

$$E(t) = \frac{1}{2}\hat{r}_t^2 \tag{4}$$

按照梯度下降算法,各层权值更新如下:

$$\Delta wg(t) = -\alpha\frac{\partial E}{\partial wg(t)} = \alpha\delta_t\frac{\partial V(s_t)}{\partial wg(t)} = \alpha\delta_t\sum_{k=1}^{t}\lambda^{t-k}\nabla wgV(s_k) \tag{5}$$

其中:wg 代表权值符号 w、v 或 u;$\nabla wgV(s_t)$ 是关于 wg 的偏导数;λ 是 TD(λ)算法多步折扣因子;α 是 ACE 学习率。

引入资格迹 e_t 来调整网络权值,不仅能调节当前时刻所对应的学习参数,而且允许调节以前时刻所涉及的学习参数,通过累积当前和过去的梯度值来加快学习的进程。资格迹 e_t 公式如下:

$$e_t = \sum_{k=1}^{t}\lambda^{t-k}\nabla wgV(s_k) = \nabla wgV(s_t) + \sum_{k=1}^{t-1}\lambda^{t-k}\nabla wgV(s_k) = \nabla wgV(s_t) + \lambda e_{t-1} \tag{6}$$

输入层与输出层连接权值修正公式如下:

$$\begin{cases}\nabla v_j(t)V(s_t) = \dfrac{\partial V(s_t)}{\partial v_j(t)} = f'(\varphi_2)y_j(t)\\ v_j(t) = v_j(t-1) + \alpha\hat{r}_t[\nabla v_j(t)V(s_t) + \lambda e_{t-1}]\end{cases} \tag{7}$$

输入层与隐含层连接权值修正公式如下:

$$\nabla w_{ij}(t)V(s_t) = \frac{\partial V(s_t)}{\partial w_{ij}(t)} = \frac{\partial V(s_t)}{\partial y_j(t)}\frac{\partial y_j(s_t)}{\partial w_{ij}(t)} = f'(\varphi_2)v_j(t)f'(\varphi_1)y_j(t);$$

$$w_{ij}(t) = w_{ij}(t-1) + \alpha\hat{r}_t[\nabla w_{ij}(t)V(s_t) + \lambda e_{t-1}] \tag{8}$$

隐含层与输出层权值修正公式如下:

$$\nabla u_i(t)V(s_t) = \frac{\partial V(s_t)}{\partial u_j(t)} = f'(\varphi_2)s_i(t);$$

$$u_i(t) = u_i(t-1) + \alpha\hat{r}_t[\nabla u_i(t)V(s_t) + \lambda e_{t-1}] \tag{9}$$

其中:$i = 1,2,\cdots,n$,$j = 1,2,\cdots,m$。ACE 通过式(7)~(9)完成评价网络参数学习。

2.2　ASE 学习算法

遗传算法中种群为所有可选状态和操作行为的集合,为了加快进化过程,将整个种群分成若干个子种群,每一个子种群用来进化一种状态下的操作行为,子种群的染色体具有相同的状态变量部分和不同的行为变量部分。用 $Q = \{A_j^i \mid i = 1,2,\cdots,N;\ j = 1,2,\cdots,M\}$ 来表示种群,A_j 表示种群 i 中第 j 个个体,N 为子种群的总个数即状态数,M 为每一个子种群中的个体总数,每一个个体产生一个信息熵值用来作为个体的自适应度,对个体进行评价。个体 A_j^i 即为系统状态与操作行为的一个集合,$A_j^i = \{s_i, a_{jk}^i\}(k = 1,2,\cdots,c)$,其中:$s_i$ 表示系统当前状态,a_{jk}^i 表示第 i 个种群内第 j 个个体中的第 k 个可选操作行为。每个个体对 c 个操作行为进行编码。

定义信息熵为 $H^i = \{H_1^i, H_2^i, \cdots, H_M^i\}$ 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,那么种群 i 中个体 j 的适应度函数 f_j^i 设计如下:

$$f_j^i = 1/H_j^i \tag{10}$$

其中当所有操作行为 a_{jk}^i 可能出现的概率相等时,信息熵最大。信息熵越大说明操作行为 a_{jk}^i 的不确定性越大,系统获得的信息越少,也就是个体的适应度越小;信息熵越小,说明个体的适应度越高,系统的自组织程度越强。

为提高遗传算法的搜索速度,采用实数编码的方法。开始时,种群中染色体个数为 0,当机器人进行行为学习时,获取相应的状态变量,并建立相应的子种群,该种群中每一个染色体具有相同的状态变量部分和各不相同的行为变量部分。随着机器人对环境的探索,新的染色体将不断添加。第 i 个子种群内第 j 个染色体编码形式为:$|s_i|a_{j1}^i|a_{j2}^i|a_{j3}^i|\cdots|a_{jc}^i|$,其中 s_i 为系统所处状态,在该状态下相应的子种群被激发,a_{jk}^i 对应输出为机器人的操作行为,$k \in [1,c]$,c 为每一个染色体内的操作行为总个数。在进行 GA 之前,首先对交叉概率 p_c 和变异概率 p_m 进行初始化。选择算子采用和适应度值成比例的概率方法进行选择,即轮盘赌法。交叉变异采用简单的单点交叉算子和位点(基本位)变异法。

本文设计了一种取向性函数,用来作为 OC 学习的内发动机。用 $\Psi = \{\psi_1, \psi_2, \cdots, \psi_n\}$ 来表示取向性函数,ψ_i 为系统处于状态 s_i 下的取向性值,定义 $E(t)$ 为系统 TD 性能指标的函数,TD 性能指标越小,表示系统所处状态的取向性越大,取向值越接近于 0;反之,取向性越小,取向值越接近于 1。

$$\psi_i(t) = \left|\frac{1-e^{-\xi E(t)}}{1+e^{-\xi E(t)}}\right| \tag{11}$$

其中:ξ 为取向值系数,设 $\xi = 1.0 \leqslant \psi_i(t) \leqslant 1(i = 1,2,\cdots,N)$,取值为 0 时表示取向性最大,取值为 1 时最向性最小,N 为状态总数。

P_j^i 为系统处理状态 s_i 下第 j 个个体的操作行为概率集合。$P_j^i = \{p_{j1}^i, p_{j2}^i, \cdots, p_{jc}^i\}$,其中 c 为每一个个体内操作行为的总数。p_{jk}^i 表示状态 s_i 下第 j 个个体内操作行为 a_k 选择的概率值。操作行为选择依下述规则[14]进行:

$R_i(P_j^i)$:IF $s_i(t)$ 及个体 A_j 被选中,

THEN a is $a_1(t)$ with p_{j1}^i

OR a is $a_2(t)$ with p_{j2}^i

…

OR a is $a_c(t)$ with p_{jc}^i

经过一段时间学习和训练后,行为选择逐渐趋于最优,如果再采用随机动作选择策略则可能会由于小概率事件的发生而导致系统输出不稳定。这里采用概率加权平均的方式来获取最优操作行为,若个体 j 是状态 s_i 下由 GA 进化所得的最优个体,即最优操作行为集合,那么通过 OC 学习从该最优个体中所获得的状态 s_i 下的最优操作行为是 $a_i^*(t) \in A^*$,A^* 是所有状态下的最优操作行为。设学习完成后最优操作行为集合中存在任意一个操作行为的概率满足 $p_{jk}^i(t) > p_\tau$,则得到

$$a_i^*(t) = \sum_{k=1}^{c} a_{jk}^i(t)p_{jk}^i(t) \tag{12}$$

其中:p_τ 为最大概率阈值,$0.5 \leqslant p_\tau \leqslant 1$,该值的选择与具体

实验有关。

执行最优操作行为 $a_i^*(t)$ 后，系统由 t 时刻状态 s_i 转移到新的状态 $s(t+1)$。按照 OC 理论，如果行为的结果使得取向性增强，则该行为得到强化，即 $\psi_i(t+1)-\psi_i(t)>0$ 时，该操作行为选择概率增加；反之如果行为的结果使得取向性减弱，则该行为得到惩罚或消退，即 $\psi_i(t+1)-\psi_i(t)<0$ 时，该操作行为选择概率减小；若 $\psi_i(t+1)-\psi_i(t)=0$，则概率不变。

基于 OC 理论的概率选择更新机制如下：

前提：若由 GA 进化得到第 j 个操作行为集合是最优的，在该操作行为集合内选择最优的操作行为。

如果 $\psi_i(t+1)-\psi_i(t)<0$，则

$$\begin{cases} p_{jk}^i(t+1)=p_{jk}^i(t)+\Delta_1, & a(t)=a_k \\ p_{jk'}^i(t+1)=p_{jk'}^i(t)-\Delta_1', & a(t)\neq a_k \end{cases} \tag{13}$$

增量部分设计如下：

$$\Delta_1=\beta(1-p_{jk}^i(t)),\Delta_1'=\beta p_{jk'}^i(t)$$

如果 $\psi_i(t+1)-\psi_i(t)>0$，则

$$\begin{cases} p_{jk'}^i(t+1)=p_{jk'}^i(t)-\Delta_2', & a(t)=a_{k'} \\ p_{jk}^i(t+1)=p_{jk}^i(t)+\Delta_2, & a(t)\neq a_k \end{cases} \tag{14}$$

增量部分设计如下：

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

其中 β 为 OC 算法的学习率，$0<\beta<1$。OC 学习通过式(13)、(14)来完成操作行为选择的概率更新，以此来获取最优的操作行为。概率更新后，每个个体的信息熵也得到更新。定义信息熵为 $H^i=\{H_1^i,H_2^i,\cdots,H_M^i\}$ 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，则

$$H_j^i(t)=H_j^i(A_j(t)\mid s_i(t))=-\sum_{k=1}^{c}p_{jk}^i(a_{jk}^i\mid s_i(t))\mathrm{lb}\,p_{jk}^i(a_{jk}^i\mid s_i(t)) \tag{15}$$

3 基于 EOBLM 的移动机器人学习避障

3.1 移动机器人系统

考虑到移动机器人的对称结构与传感器均匀分布特征，为针对避障和导航设计一个简单易行的模糊控制器，本文简化了输入维数。文中采用机器人半径为 $R=20$ cm，前方的 18 个传感器每 6 个一组共分为三组，分别为 SL、SF 和 SR。每一个传感器可探测的距离范围是 10～250 cm。对于每一个传感器 $x_i(i=1,2,\cdots,24)$，其覆盖角度为 15°，在其范围内得到与障碍物的距离信息 l_i。三个方向的距离对应的表示为：左侧距离 dL，前方距离 dF 以及右侧距离 dR。三个方向的距离值为机器人的半径与每一个方向六个传感器所检测到距离的最小值之和，如式(16)所示。

$$\begin{cases} dR=R+\min\limits_{i=1,2,\cdots,6}(l_i) \\ dF=R+\min\limits_{i=7,8,\cdots,12}(l_i) \\ dL=R+\min\limits_{i=13,14,\cdots,18}(l_i) \end{cases} \tag{16}$$

如图 2 中，本文采用两个坐标系，一个是用 XOY 表示的世界坐标系，一个是用 xoy 表示的机器人坐标系，其中 o 是机器人的中心，当机器人的后面两个轮子与 y 轴在同直线时则会径直向前走。机器人的控制变量是机器人运动的转角 $\Delta\theta$。在机器人行为控制结构中，决定机器人合适的动作 $\Delta\theta$ 即可实现避障的目的。

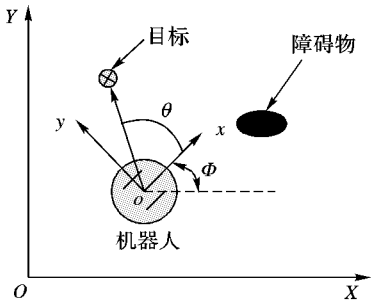

图 2 机器人的结构与坐标图

3.2 仿真实验与分析

机器人状态变量即三个方向上与障碍物距离 dL，dF 和 dR 均离散化为 2 个，动作变量 $\Delta\theta$ 离散化为 5 个。这里定义如果传感器所探测障碍物距离小于 50 cm 则 N(Near)，大于 50 cm 为 F(Far)，$S=\{dL,dF,dR\}$，所以输入状态离散化的数目 $sn=8$，输出动作离散化的数目为 $sm=5$。由状态变量和动作变量组成表示为染色体。开始时，种群中染色体的个数为 0，当机器人进行行为学习时，获取相应的状态变量，并建立相应的子种群，该种群中每一个染色体具有相同的状态变量部分和各不相同的行为变量部分。随着机器人对环境的探索，新的染色体将不断添加。共有 $N=8$ 个子种群，每一个子种群的规模为 $M=50$，每一个个体内的操作行为个数为 $c=5$，具体值为在 $[-65,65]$ 区间随机生成 5 的倍数的整数，其中负数表示机器人向右偏转角度，正数表示向左偏转角度。操作行为的初始概率 $p_{jk}^i(0)=\frac{1}{5}$，对应的初始熵 $H_j^i(0)=-\sum_{j=1}^{5}\frac{1}{5}\times\mathrm{lb}\frac{1}{5}=2.32$，初始适应度值为 $f_j^i(0)=1/H_j^i(0)=0.43$，其中 $i=1,2,\cdots,N,j=1,2,\cdots,M,k=1,2,\cdots,c$。

在每一代进化中，随着遗传操作产生优秀个体的加入，遗弃适应度值较大的个体(本文适应度值越大，适应性越差)，以保持种群规模的不变。进化操作学习结束后，选择每个子种群中具有最小适应度值的个体作为机器人路径规划的最优操作行为集合。在下一阶段学习中，机器人从最优操作行为集合中按操作条件反射概率选择最优的行为作为路径规划的控制器作用于环境。

具体实现算法过程中，采样时间取 1 s，速度 v 取固定的值 0.2 m/s。ACE 学习算法中参数初始化为：$\lambda=0.95,\alpha=0.3,n=3,m=16$；OC 学习中参数 $\beta=0.05,\xi=1,p_\tau=0.98$；GA 学习算法中：交叉概率 $p_c=0.8$，变异概率 $p_m=0.1$。

设定机器人每进化 300 次为一次实验学习，一次学习完成后重新回到起始点进行环境探索开始新一轮的学习，共进行 10 次学习。在每一次学习过程中，当机器人与障碍物距离小于最小安全距离时，回到前一时刻碰撞前的位置重新选择动作进行学习直到避开障碍物。

本文遗传算法中共产生 8 个子种群，分别对应机器人的 8 种输入状态。机器人在环境探索过程中，根据输入状态变量激活相应的子种群，然后根据遗传操作进行相应的进化。图 3 显示了种群中某个子种群在进化过程中历代最大适应度值(对应最小熵值)和最小适应度值(对应最大熵值)的变化曲线。机器人经过 3 000 次的进化代数，仿真结果显示，机器人处于状态 $S2=(N,N,F)$ 下的子种群被进化的次数为 298 次。当进化到一定代数后，算法逐渐收敛，个体的最小适应度得到稳定，即在该种群中得到了优化的操作行为集合。

图 4 为学习过程中,图 3 对应的优化操作行为集合的最大操作行为概率值和最小操作行为概率值的变化曲线。可以看出,在该子种群被激活的条件下,某个体被选择的次数越多,该个体内操作行为被选择的个数也随之增多,好的操作行为被选择的概率逐渐增大,相应地不好的操作行为被选择的概率逐渐减小。直到好的操作行为被选择的概率达到 p_τ,相应的个体的适应度达到最大,机器人也通过不断进化学习到了最优的操作行为。

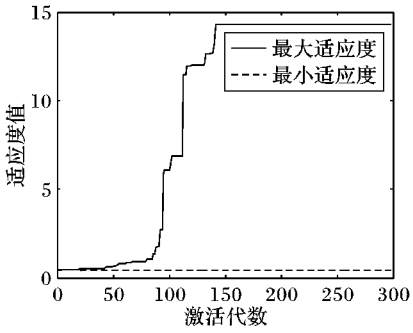

图 3　适应度值变化曲线

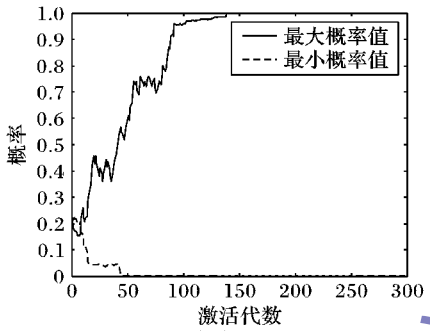

图 4　概率值学习曲线

当学习完成后,每一个系统状态下都得到了相应的最优操作行为,实现了从感知空间到动作空间的最佳映射。从表 1 可以看出,由 GA 进化后所得的最优操作行为集合内获得的最优操作行为其选择概率均为该集合内的最大值,所对应所属的个体适应度值也相应最小,达到了预期的实验效果。

表 1　各状态下学习完成后的最优操作行为

系统状态	最优行为/(°)
*S*1(N, N, N)	45
*S*2(N, N, F)	-55
*S*3(N, F, N)	0
*S*4(N, F, F)	-10
*S*5(F, N, N)	50
*S*6(F, N, F)	55
*S*7(F, F, N)	-15
*S*8(F, F, F)	5

注:正数表示向左偏转角度,负数表示向右偏转角度。

为了验证本文所提出的进化操作学习模型方法的学习性能,同经典的 AHC 方法相比。对比实验中,AHC 方法没有引入 GA 学习和 OC 学习,其他参数与本文所提出的 EOBLM 方法相同。图 5 显示机器人在 20 次学习过程中分别采用本文方法和传统的 AHC 方法碰撞次数的变化曲线,可以看出,随着机器人不断地进化,其与障碍物的碰撞次数逐渐减小并最终实现无碰巡航。共进行 20 次的实验,每一次实验的代数为 300。从图中可以看出,EOBLM 方法由于在 ASE 单元中加入 GA 学习和 OC 学习,通过进化有效地缩小了搜索空间,提高了学习速率,使得从实验初期开始其碰撞次数就明显少于 AHC 方法;随着训练时间增加,其碰撞次数也逐渐减小到 0。

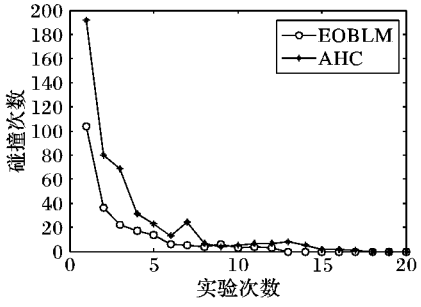

图 5　碰撞次数变化曲线

图 6 为机器人经过 20 轮学习完成后在不同环境的无障碍巡航路线图。图(a)为复杂障碍物环境下学习完成后的机器人自由巡航运行轨迹,可以看出已习得躲避障碍物的能力并能从图右下角的 U 型障碍物环境中成功走出,并且运动轨迹较为光滑。图(b)为学习完成后机器人在窄道环境中的运行轨迹,窄道宽度为 40 cm,可以看出,已习得技能的机器人能适应新的环境,成功避开两侧的障碍物,无碰撞地穿过窄道。

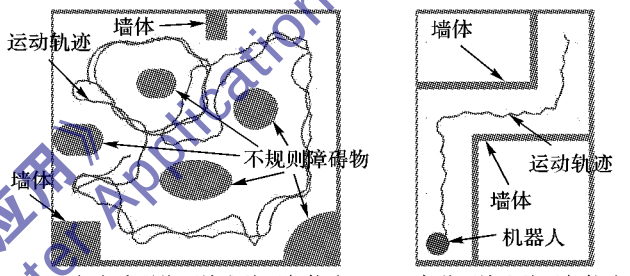

图 6　学习完成后运动轨迹

从上述实验结果可以看出,本文所设计的进化操作学习的方法可以较好地实现机器人避碰实验。在与环境不断交互过程中通过经验的不断积累,ACE 单元对行为动作结果进行评价,作用于 ASE 动作单元使其动作选择得到更新学习。ASE 包括 GA 学习和 OC 学习两个阶段,GA 学习阶段获得最优操作行为集合,OC 学习阶段习得最优操作行为。结果显示,本文所提出的进化操作学习是一个动态的学习过程,机器人通过不断与环境交互激发子种群从而得到优化的操作行为集合和最优的操作行为,并以进化到的最优操作行为作为避碰行为的控制器,从而完成无障碍巡航的任务。与传统的 AHC 学习相比,具有高度的自主学习性和自适应性,鲁棒性也较强。

4　结语

本文结合 GA 的进化思想,模拟操作条件反射学习机制,以自适应启发评价(AHC)学习为框架,设计了一种进化的操作学习模型,并将其应用于移动机器人学习避障行为。与传统的强化学习方法相比,该方法有效地提高了 ACE 单元的学习速率;同时 ASE 单元通过自主学习行为动作,无需教师信号或专家知识,具有高度的自主性和自适应能力。使用进化的操作学习方法使机器人学习避碰行为,学习得到的最优操作行为作为机器人避碰行为的控制器,仿真结果表明该方法能够有效地实现无碰巡航,提高了机器人反映的灵活性和对环境的适应性。下一步将重点研究本文方法的可扩展性,使其能够应用于实际复杂的两轮机器人系统,使其在实现运动平衡的同时又能够具有实时避障能力。

参考文献：

[1] 王志文，郭戈. 移动机器人导航技术现状与展望[J]. 机器人，2003，25(5)：470－474.

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

[3] YEN J，PFLUGER N. A fuzzy logic based extension to Payton and Rosenblatt' s command fusion method for mobile robot navigation [J]. IEEE Transactions on Systems，Man and Cybernetics，1995，25(6)：971－978.

[4] KERMICHE S，SAIDI M L，ABBASSI H A. Gradient descent adjusting Takagi-Sugeno controller for a navigation of robot manipulator [J]. Journal of Engineering and Applied Science，2006，1(1)：24－29.

[5] JOO ER M，CHANG D. Obstacle avoidance of a mobile robot using hybrid learning approach [J]. IEEE Transactions on Industrial Electronics，2005，52(3)：898－905.

[6] JOO ER M，ZHOU Y. Automatic generation of fuzzy inference systems via unsupervised learning [J]. Neural Networks，2008，21(10)：1556－1566.

[7] BOUBERTAKH H，TADJINE M，GLORENNEC P-Y. A new mobile robot navigation method using fuzzy logic and a modified Q-learning algorithm [J]. Journal of Intelligent & Fuzzy Systems，2010，21(1/2)：113－119.

[8] SUTTON R S，BARTO A G. Reinforcement learning [M]. London：MIT Press，1998：1－12.

[9] SU S F，Hsieh S H. Embedding fuzzy mechanisms and knowledge in box-type reinforcement learning controllers [J]. IEEE Transactions on Systems，Man and Cybernetics：Part B，2002，32(5)：645－653.

[10] ZEYBEK Z. Role of adaptive heuristic criticism in cascade temperature control of an industrial tubular furnace [J]. Applied Thermal Engineering，2006，26(2/3)：152－160.

[11] MUCIENTES M，ALCALA-FDEZ J，ALCALA R，*et al*. A case study for learning behaviors in mobile robotics by evolutionary fuzzy system [J]. Expert Systems with Application，2010，37(2)：1471－1493.

[12] DESOUKY S F，SCHWARTZ H M. Self-learning fuzzy logic controllers for pursuit-evasion differential games [J]. Robotics and Autonomous Systems，2011，59(1)：22－33.

[13] KNUDSON M，TUMER K. Adaptive navigation for autonomous robots [J]. Robotics and Autonomous Systems，2011，59(6)：410－420.

[14] TOURETZKY D S，SAKSIDA L M. Operant conditioning in Skinnerbots [J]. Adaptive Behavior，1997，5(3/4)：219－247.

[15] GUTNISKY D A，ZANUTTO B S. Learning obstacle avoidance with an operant behavior model [J]. Artificial Life，2004，10(1)：65－81.

[16] SAKSIDA L M，RAYMOND S M，TOURETZKY D S. Shaping robot behavior using principles from instrumental conditioning [J]. Robotics and Autonomous Systems，1998，22(3/4)：231－249.

[17] GAUDIANO P，CHANG C. Adaptive obstacle avoidance with a neural network for operant conditioning：Experiments with real robots [C]// CIRA 97：Proceedings of 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation. Piscataway：IEEE，1997：13－18.

[18] ITOH K，MIWA H，MATSUMOTO M，*et al*. Behavior model of humanoid robots based on operant conditioning [C]// Proceedings of the 5th IEEE-RAS International Conference on Humanoid Robots. Piscataway：IEEE，2005：220－225.

[19] 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.

[20] 段勇，崔宝侠，徐心如. 进化强化学习及其在机器人路径跟踪中的应用[J]. 控制与决策，2009，24(4)：532－536.

（上接第 2279 页）

4 结语

本文在充分考虑知识库描述语言线性特征前提下，提出了一种有效的义原描述式权重分配方案，并结合二部图的最大权匹配算法以及现有方法进行词汇的语义相似度计算。实验结果表明，采用本文方法计算得到的词汇语义相似度能够更合理地体现词语间语义上的差异性，更加符合人们的主观理解。接下来，将深入研究《知网》对词汇的描述特点，从而更进一步改善词汇语义相似度计算的合理性。

参考文献：

[1] ZHU Z Y，DONG S J，YU C L，*et al*. A text hybrid clustering algorithm based on HowNet semantics [C]// ICAMCS 2011：2011 International Conference on Advanced Materials and Computer Science. Zurich：Trans Tech Publications Ltd，2011：474－476.

[2] 荀恩东，颜伟. 基于语义网计算英语词语相似度[J]. 情报学报，2006，25(1)：43－48.

[3] 刘群，李素建. 基于《知网》的词汇语义相似度计算[C] // 第三届汉语词汇语义学研讨会论文集. 台北：[出版者不详]，2002：59－76.

[4] 董强，董振东. 知网简介[EB/OL]. [2013－01－29]. http://www.keenage.com/.

[5] 李峰，李芳. 中文词语语义相似度计算——基于《知网》2000[J]. 中文信息学报，2007，21(3)：99－105.

[6] DAI L L，LIU B，XIA Y N，*et al*. Measuring semantic similarity between words using HowNet [C]// ICCSIT'08：2008 International Conference on Computer Science and Information Technology. Washington，DC：IEEE Computer Society，2008：601－605.

[7] 刘青磊，顾晓峰. 基于《知网》的词语相似度算法研究[J]. 中文信息学报，2010，24(6)：31－36.

[8] 王小林，王义. 改进的基于知网的词语相似度算法[J]. 计算机应用，2011，31(11)：3075－3077.

[9] 郝长伶，董强. 知网知识库描述语言[C]// 全国第七届计算语言学联合学术会议论文集. 北京：清华大学出版社，2003：371－377.

[10] 龚劬. 图论与网络最优化算法[M]. 重庆：重庆大学出版社，2009：86－95.

[11] 李荣陆. 中文文本分类语料 2003 [DB/OL]. [2013－01－29]. http://www.nlpir.org/download/tc-corpus-answer.rar.

[12] 余刚，裴仰军，朱征宇，等. 基于词汇语义计算的文本相似度研究[J]. 计算机工程与设计，2006，27(2)：241－244.

[13] HAN J W，KAMBER M. 数据挖掘：概念与技术[M]. 范明，译. 2版. 北京：机械工业出版社，2007：263－266.

[14] LARSEN B，AONE C. Fast and effective text mining using linear-time document clustering [C]// Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York：ACM，1999：16－22.