基于最先策略增强学习的 ART2 神经网络*

樊 建[1,2] 吴耿锋[1]

[1](上海大学 计算机工程与科学学院 上海 200072)
[2](南京陆军指挥学院 南京 210045)

摘 要 提出一种基于最先策略增强学习的 ART2 神经网络 FPRL-ART2(Foremost-Policy Reinforcement Learning based ART2 neural network),并介绍其学习算法.为了达到在线学习的目的,在 FPRL-ART2 中,从状态到行为值之间的映射中,选择第一个得到奖励的行为,而不是选择诸如 1-step Q-Learning 中具有最优行为值的行为.ART2 神经网络用于存储分类模式,其权重通过增强学习增强或减弱,达到学习的目的.并将 FPRL-ART2 运用到移动机器人避碰撞问题的研究中.仿真实验表明,引入 FPRL-ART2 后减少移动机器人与障碍物发生碰撞的次数,具有良好的避碰效果.

关键词 增强学习,ART2 神经网络,最先策略,避碰撞
中图法分类号 TP18

Foremost-Policy Reinforcement Learning Based ART2 Neural Network

FAN Jian[1,2], WU Geng-Feng[1]

[1](*School of Computer Engineering and Science, Shanghai University, Shanghai* 200072)
[2](*Nanjing Army Command College, Nanjing* 210045)

ABSTRACT

A foremost-policy reinforcement learning based ART2 neural network (FPRL-ART2) and its learning algorithm are proposed in this paper. To fit the requirement of real time learning, the first awarded behavior based on present states is selected in our Foremost-Policy Reinforcement Learning (FPRL) in stead of the optimal behavior in 1-step Q-Learning. The algorithm of FPRL is given and it is integrated with ART2 neural network. The stored weights of classified pattern in ART2 is increased or decreased by reinforcement learning. The FPRL-ART2 is successfully used in collision avoidance of mobile robot and the simulation experiment indicates that the times of collision between robot and obstacle is effectively decreased. The FPRL-ART2 makes favorable result of collision avoidance.

Key Words Reinforcement Learning, ART 2 Neural Network, Foremost - Policy, Collision Avoidance

1 引 言

基于自适应谐振理论的 ART2 神经网络[1]是采用竞争学习和自稳机制原理实现稳定的分类,被广泛应用于模式识别与分类.黎明、严超华等人提出一种具有更严格警戒测试准则的 ART2 神经网

* 上海市科学技术发展基金项目(No. 015115042)、上海市教委第 4 期重点学科建设项目(No. B682)资助
收稿日期:2004-12-19;修回日期:2005-05-08
作者简介 **樊建**,男,1978 年生,博士研究生,主要研究方向为智能信息处理、机器人控制. E-mail: jfan@mail.shu.edu.cn.
吴耿锋,男,1945 年生,教授,博上生导师,主要研究方向为智能控制、神经元网络、模糊逻辑和专家系统.

络[1],使其具有更高的准确识别率.

增强学习作为一种无监督的学习方法,因其普遍适用性而得到广泛关注.在机器人控制领域内,它是一个有效的学习方法.如肖南锋等人提出一种增强学习算法[11],并成功应用于未知环境下的机器人控制.

本文提出一种基于最先策略增强学习的 ART2 神经网络 FPRL-ART2 (Foremost-Policy Reinforcement Learning based ART2 neural network),给出该神经网络的学习算法.通过最先策略增强学习使 ART2 神经网络增强或减弱已存储的分类模式,使其适合于在线学习.

2　基于最先策略的增强学习方法 FPRL

2.1　增强学习

增强学习[10]又称为强化学习或再励学习,是不同于监督学习和无监督学习的另一大类机器学习方法.由于增强学习方法能够通过与环境的交互来实现行为决策的优化,因此在求解复杂的优化控制问题中具有广泛的应用价值.

增强学习的基本框架就是由一个学习 agent 观察当前环境状态 s_t,并根据行为选择策略 π 得到相应的行为 b_t,随后 agent 给出一个外部评价 r_{t+1}.例如对于障碍物避碰,与物体相碰时 $r=-1$,否则 $r=0$,同时观察当前环境的新状态 $s_{t+1}=T(s_t,b_t)$,T 为在状态 s_t、行为 b_t 下到新状态 s_{t+1} 之间的映射.增强学习的目标就是根据从实际环境中得到的数组(s_t,b_t,r_{t+1},s_{t+1})学习从状态到行为值(可由行为值函数得到)之间的映射,也就是特征函数.如文献[5]中介绍,其中应用比较广泛的 1-step Q-Learning 的行为值函数如式(1)、特征函数如式(2)所示:

$$Q_{t+1}(s_t,b_t)=(1-\alpha)Q_t(s_t,b_t)+\alpha(r_t+\gamma V_t(s_{t+1})), \tag{1}$$

$$V_{t+1}(s)=\max_{b\in B}Q_{t+1}(s,b). \tag{2}$$

式(1)中,α 为学习率,γ 为影响因子,r_t 为时刻 t 时的评价值,其表示在状态 s_t 时,采取行为 b_t 后达到新状态 s_{t+1} 的期望值.一旦得到如式(2)所示的特征函数 $V_{t+1}(s)$,最优行为选择策略 $\pi_{t+1}(s)$ 就很容易得到.式(2)中 B 为所有系统可能采取行为 b 的集合.如式(3)所示,从一个状态的所有可能行为中选出具有最大行为值的一个,即最优行为选择策略,

$$\pi_{t+1}(s)=b \text{ 如果 } Q_{t+1}(s,b)=V_{t+1}(s). \tag{3}$$

2.2　基于最先策略的增强学习方法 FPRL

从上节我们可以看到 1-step Q-Learning 算法是基于最优策略的,需要很长时间去学习得到当前状态下的最佳行为,比较适合于离线学习.为了适应在线学习的需要,我们提出一种基于最先策略的增强学习方法 FPRL (Foremost-Policy Reinforcement Learning),即在当前状态时,并不选择具有最优行为值的行为,而是选择第一个得到奖励的行为.实验表明虽然它不能达到最佳的效果,但却大大减少了学习时间,适合于在线学习.

定义特征函数为得到第一个评价值为正数的行为值

$$\operatorname{firstR}_{b\in A}Q_{t+1}(s,b), \tag{4}$$

学习策略 $\pi_{t+1}(s)$ 如式(5)所示:

$$\pi_{t+1}(s)=b \text{ 如果 } Q_{t+1}(s,b)=\operatorname{firstR}_{b\in A}Q_{t+1}(s,b). \tag{5}$$

最先学习策略为,从一个状态的所有可能行为中选出评价值第一个为正值的行为.

行为值函数则如式(6)所示:

$$Q_{t+1}(s_t,b_t)=(1-\alpha)Q_t(s_t,b_t)+\alpha(r_t+\gamma \operatorname{firstR}_{b\in A}Q_{t+1}(s,b)). \tag{6}$$

3　基于最先策略增强学习的 ART2 神经网络 FPRL-ART2

3.1　自适应谐振理论

自适应谐振理论(Adaptive Resonance Theory,ART)是 1976 年由美国 Boston 大学的 Grossberg 提出[8,9],之后他又提出了 ART 神经网络.发展至今,ART 神经网络可分为 ART1、ART2 和 ART3,其中 ART1 可用于处理二进制值,ART2 用于处理任意模拟量和二进制,ART3 则将人脑模型中的神经元突触的生物化学运行机理应用到人工神经网络中,可进行分级搜索.

ART2 网络模型由两个子系统构成:注意子系统和定向子系统(如图 1 对应的虚线部分).注意子系统对输入进行预处理后,通过竞争选择与输入模式最匹配的模式原型(即聚类中心).定向子系统对选出的模式原型进行相似度警戒测试,如通过警戒测试则系统进入共振状态学习并调整权矢量,否则屏蔽当前激活节点,搜索其他的模式原型.如所有的模式原型均不匹配,则开辟新的输出端点.此中,权值的学习与调整算法直接影响模式原型与实际聚类

中心的接近程度，而相似度的警戒测试是 ART2 网络学习和分类稳定性的保证.

如图 1 所示，ART2 系统分为 $F1$ 层和 $F2$ 层，$F1$ 层有 W、X、U、V、P、Q、R 这 7 个子层，$F2$ 层由 Y 子层组成. 信号 S 从 $F1$ 层输入，经其处理之后通过由下往上的连接权重 b_{ij} 的加权组合传递到 $F2$ 层，$F2$ 层中的各节点相互竞争产生优胜单元，之后通过自上而下的连接权重 t_{ji} 再将信号传回 $F1$ 层. 在 $F1$ 层计算 t_{ji} 与输入模式的匹配程度并与某一阈值做比较. 如果匹配值大于该值，则修改 b_{ij} 与 t_{ji} 之权重值；若小于该值，则向 $F2$ 层发出重置信号以抑制优胜单元，并继续寻找其它优胜单元.

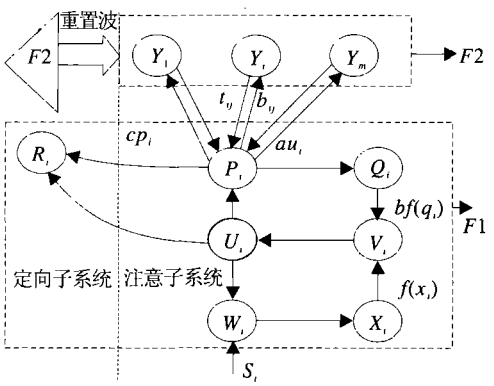

图 1 ART2 神经网络结构

Fig. 1 ART2 neural network structure

3.2 FPRL-ART2 神经网络

为了减小避碰撞系统对内存空间的需要，我们采用 ART2 神经网络存储大量的避碰分类模式[3]. 面对大量的分类模式，用手工评估分类模式是非常困难的，需要一个选择和评估机制. 为此我们在 ART2 神经网络中引入最先策略增强学习(FPRL)机制，解决如何评估和选择已存储在 ART2 中的分类模式的问题.

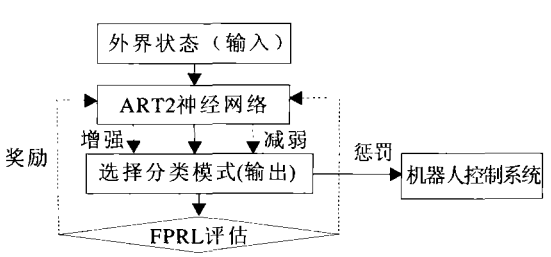

图 2 FPRL-ART2 学习流程图

Fig. 2 Learning flow chart of FPRL-ART2

FPRL-ART2 流程如图 2 所示，当外界状态输入 ART2 神经网络时，神经网络选择相应分类模式，然后由 FPRL 增强学习评估模块对该分类模式运行效果进行增强学习评估. 当效果良好时，增强该分类模式，即在 ART2 神经网络中增大相应网络权重. 否则惩罚该模式，即减小相应网络权重. 当学习达到一定次数时，转向选择其它分类模式.

3.3 FPRL-ART2 学习算法

如图 1 所示，$F1$ 层各特征值公式如下：

$$u_i = \frac{v_i}{e + \| V \|}, \tag{7}$$

$$w_i = s_i + au_i, \tag{8}$$

$$p_i = u_i + dt_{ji}, \tag{9}$$

$$x_i = \frac{w_i}{e + \| W \|}, \tag{10}$$

$$q_i = \frac{p_i}{e + \| P \|}, \tag{11}$$

$$v_i = f(x_i) + bf(q_i), \tag{12}$$

$$f(x) = \begin{cases} x, & \text{if } x \geqslant 0 \\ 0, & \text{if } x < 0 \end{cases} \tag{13}$$

学习算法步骤如下：

1. 参数与权值初始化. 初始化 a、b、d、e、α、ρ、ω，其中 $a = b$，一般取值为 10；d 为权重修改参数；e 为误差参数；α 为学习率；ρ 为阈值，一般介于 0.7 到 1 之间；权值初始化为 $t_{0i} = 0$，$b_{i0} = \frac{1}{(1-d)\sqrt{n}}$，其中 n 为输入参数个数，ω 为权重 t_{ji} 与 b_{ij} 之间的比例因子.

2. 按照标准 ART2 算法得到输出单元，即分类模式.

3. 根据 2.2 节的 FPRL 增强学习方法得到该分类模式评价值 r_t.

4. 按式(14)、(15) 更改单元 J 之权重值.

$$t_{ji} = \alpha d u_i + \{1 + \alpha d(d-1)\} t_{ji} + r_t, \tag{14}$$

$$b_{ij} = \alpha d u_i + \{1 + \alpha d(d-1)\} b_{ij} + \omega r_t. \tag{15}$$

5. 按式(7) ～ (12)，更改 $F1$ 层之特征值.

6. 测试是否达到更改权值之次数.

7. 测试是否达到学习循环之次数.

4 仿真实验分析

为了验证 FPRL-ART2 的有效性，我们在移动机器人避碰撞仿真实验中引入 FPRL-ART2，通过 ART2 神经网络存储避碰撞行为 CAB(Collision Avoidance Behavior)，当机器人 R 探测到将与障碍物 O 碰撞时，就根据当前状态从 ART2 神经网络中选择 CAB，并使用 FPRL 评价避碰结果，当避碰效

果不理想时,减小神经网络中相关权重,转向选择其它 CAB,直至找到效果理想的 CAB,否则增大神经网络中相关权重,当下次有相同状态输入神经网络时,效果理想的 CAB 将被再次选择. 仿真流程可参阅图 2.

仿真实验采用 TeamBots 开源机器人仿真软件,仿真环境设置如图 3 所示,机器人 R(黑色圆点)和障碍物 O(灰色圆点)在 20×20 单位的平面上运动. R 运动速度为 0.5 个单位,运动方向用极坐标表示为 $\pi/2$,且保持不变;两个 O 在平面上的位置运动方向用极坐标表示分别为 0 和 π,速度为 0.15.

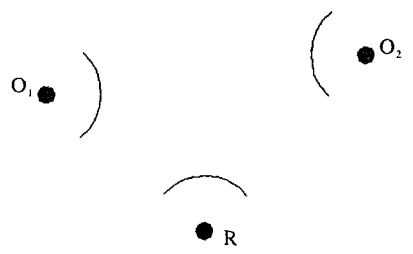

图 3　仿真环境设置

Fig. 3　Setting of simulation environment

仿真实验 1 的主要目的是演示 FPRL 评价结果对 CAB 的影响. 设有 2 个 O 和 1 个 R,R 和 O 起始位置如图 3 所示. 我们提出基于规则的避碰撞方法[3],一条避碰撞规则由 O 状态、R 状态和一个 CAB 组成,其中 CAB 由两部分组成:偏转角度和运动加速度. 在仿真实验中,R 偏转角度最大为 $\pm 30°$,最大运动加速度为 $\pm 0.02°$,行为空间中偏转角度间隔取 5°,加速度间隔取 0.005,并将 R 行为空间划分为 117 个 CAB. R 和 O 状态都用 4 个参数(速度,方向,x 坐标,y 坐标)来表示,FPRL-ART2 神经网络 $F1$ 层节点 12 个(对应于 1 个 R 和 2 个 O 的 12 个状态变量),$F2$ 层节点 117 个(对应于 R 行为空间的 117 个 CAB). 当 R 与 O 将要发生碰撞时,系统根据 R 与 O 当前的状态通过 FPRL-ART2 选取 CAB,并根据避碰结果评估 CAB,增大或减小相关权重. FPRL-ART2 中各参数取值如下:$\alpha = 0.9$,$a = b = 10$,$d = 0.9$,$\rho = 0.9$,$e = 0.05$,$\omega = 0.1$.

表 1 给出的是当 FPRL-ART2 选择 CAB 后,转换到其它 CAB 时神经网络需执行的次数. 图 4 给出的是 FPRL 评价值对 CAB 的影响.

从图 4 可以看出,当评价值小于 −0.15 时,FPRL-ART2 才会转换至下一 CAB,大于 −0.15 时,则增强当前 CAB. 从表 1 可以看出,增强当前 CAB 评价值和转换 CAB 评价值分别为 0.1 和 −0.2 时,神经网络只需学习一次即可转换到下一 CAB,而采用其它评价值组合时,神经网络则需要学习多次(3～7 次). 从图 4 和表 1 可知,在使用 FPRL-ART2 时,FPRL 的评价值应在一定范围内,这样才有利于 CAB 的转换,当超出一定范围时,则学习效果不好.

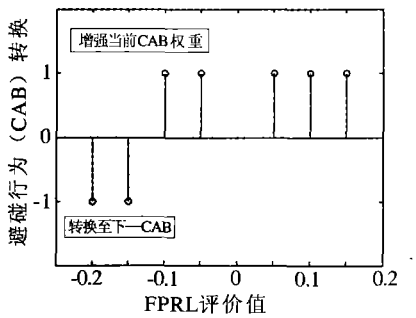

图 4　FPRL 评价值对 CAB 的影响

Fig. 4　Influence of FPRL evaluation result upon CAB

表 1　转换 CAB 时 FPRL-ART2 执行次数统计

Table 1　Execution times statistic of FPRL-ART2 when switching CAB

增强当前 CAB 评价值	转换 CAB 的评价值	FPRL-ART2 执行次数
0.1	−0.15	3
0.1	−0.20	1
0.20	−0.15	7
0.2	−0.20	3
0.3	−0.20	4
0.3	−0.25	3
0.4	−0.25	3
0.4	−0.30	3

仿真实验 2 的主要目的是演示 FPRL-ART2 神经网络的避碰效果. 设有 2 个 O,1 个 R,神经网络各层节点数、仿真环境和仿真实验 1 相同,神经网络各参数也与实验 1 相同,但 O 的运动状态(起始位置和方向)是随机取值的. 每次仿真时间为 4min,FPRL 评价函数如下:

$$r = \begin{cases} 0.1, & \text{if 没有碰撞} \\ -0.25, & \text{if 发生碰撞} \end{cases} \tag{16}$$

图 5 为采用 FPRL-ART2 和未采用 FPRL-ART2 时在仿真时间内 R 与 O 发生碰撞的次数. 从图中可以看出,当采用 FPRL-ART2 后,有效减小 R 与 O 发生碰撞的次数.

图 6 为未使用 FPRL-ART2 时 R 与 O 发生碰

撞的情景，图 7 为采用 FPRL-ART2 时 R 成功躲避 O 时的情景. 图中所示黑色圆点为机器人，灰色圆点为障碍物.

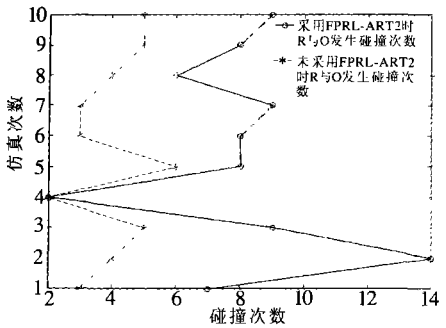

图 5 采用 FPRL-ART2 和未采用 FPRL-ART2 时 R 与 O 发生碰撞次数

Fig. 5 Collision times between R and O with and without FPRL-ART2

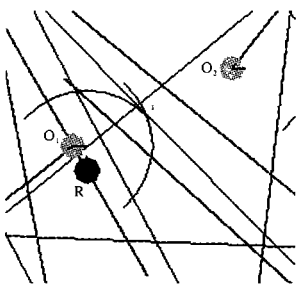

图 6 未采用 FPRL-ART2 发生碰撞

Fig. 6 Collision without FPRL-ART2

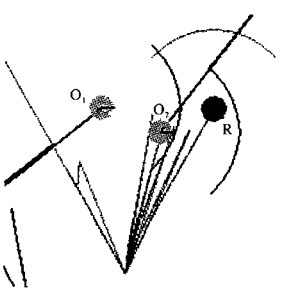

图 7 经过学习成功避碰

Fig. 7 Successful collision avoidance through learning

5 结束语

本文提出一种基于最先策略增强学习方法的 ART2 神经网络 FPRL-ART2，并将其成功用于移动机器人多障碍物避碰撞问题的研究中. 我们利用 ART2 神经网络存储移动机器人避碰行为，使用基于最先策略增强学习方法 FPRL 对避碰结果进行评价，成功避碰则增大当前避碰行为相关权重，发生碰撞则减小当前避碰行为相关权重，直至找到成功的避碰行为. 当下次有相同状态输入神经网络时，成功避碰的行为将被再次选择. 仿真实验表明使用 FPRL-ART2 神经网络后，有效减少了移动机器人与障碍物发生碰撞的次数，取得良好的效果.

参 考 文 献

[1] Carpenter G A, Grossberg S. ART2: Stable Self-Organization of Category Recognition Codes for Analog Input Patterns. Applied Optics, 1987, 26(23): 4919－1930

[2] Liu X H, Yu Z Z, Duan J, *et al*. Face Recognition Using Adaptive Resonance Theory. In: Proc of the International Conference on Machine Learning and Cybernetics. Xi'an, China, 2003, V: 3167－3171

[3] Fan J, Wu G F, *et al*. Reinforcement Learning and ART2 Neural Network Based Collision Avoidance System of Mobile Robot. In: Yin F L, Wang J, Guo C G, eds. Lecture Notes in Computer Science. 2004, 3174: 35－40

[1] Li M, Yan C H, Liu G H. ART2 Neural Networks with More Vigorous Vigilance Test Criterion. Journal of Image and Graphics, 2001, 6(1): 81－85 (in Chinese)
(黎 明，严超华，刘高航. 具有更严格警戒测试准则的 ART2 神经网络. 中国图象图形学报, 2001, 6(1): 81－85)

[5] Whitehead S D, Sutton R S, Ballard D H. Advances in Reinforcement Learning and Their Implications for Intelligent Control. In: Proc of the 5th IEEE International Symposium on Intelligent Control. Philadelphia, USA, 1990, Ⅱ: 1289－1297

[6] Suwimonteerabuth D, Chongstitvatana P. Online Robot Learning by Reward and Punishment for a Mobile Robot. In: Proc of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Lausanne, Switzerland, 2002, Ⅰ: 921－926

[7] Fujimori A, Tani S. A Navigation of Mobile Robot with Collision Avoidance for Moving Obstacles. In: Proc of the IEEE International Conference on Industrial Technology. Bangkok, Thailand, 2002, Ⅰ: 1－6

[8] Grossberg S. Adaptive Pattern Classification and Universal Recoding, I: Parallel Development and Coding of Neural Feature Detectors. Biological Cybernetics, 1976, 23(3): 121－134

[9] Grossberg S. Adaptive Pattern Classification and Universal Recoding, II: Feedback, Expectation, Olfaction, Illusions. Biological Cybernetics, 1976, 23(4): 187－202

[10] Sutton R S, Barto A G. Reinforcement Learning: An Introduction. Cambridge, USA: MIT Press, 1998

[11] Xiao N F, Nahavandi S. A Reinforcement Learning Approach for Robot Control in an Unknown Environment. In: Proc of the IEEE International Conference on Industrial Technology. Bangkok, Thailand, 2002, Ⅱ: 1096－1099