文章编号：1000－8152(2001)05－0748－03

强化学习在导弹制导中的应用*

周 锐 陈宗基
(北京航空航天大学自动控制系·北京，100083)

摘要：简述了强化学习的基本原理和特点，讨论了强化学习中评价函数的神经网络近似问题，重点分析了采用多神经网络近似评价函数的学习问题，实现了状态空间或任务的自动分解，提高了评价函数的推广能力.网络的学习是离线进行，并作为反馈控制器在线应用.并以 A-学习为例，将强化学习应用于导弹的制导问题，仿真结果表明了强化学习在导弹制导或控制问题中的应用前景和有效性.

关键词：神经网络；强化学习；微分对策；导弹制导
文献标识码：A

Application of Reinforcement Learning in Missile Guidance

ZHOU Rui and CHEN Zongji
(Department of Automatic Control, Beijing University of Aeronautics and Astronautics·Beijing, 100083, P. R. China)

Abstract: Principle and characteristic of reinforcement learning are outlined. The value function approximation of reinforcement learning with neural networks is studied, and the learning algorithm using modular neural networks to approximate the value function is emphatically analyzed, which decomposes the state space automatically and increases the generalizing ability of the neural networks. The neural networks are trained offline, and is used online as a feedback controller. The A-learning algorithm is applied in the missile guidance problem, and the simulation results show the good performance and effectiveness of the application of reinforcement learning in those problems of missile guidance and control.

Key words: neural networks; reinforcement learning; differential games; missile guidance

1 引言(Introduction)

强化学习(reinforcement learning,简称 RL)来源于行为心理学,它把学习看作是试探评价过程.它与以往的各种导师学习有着本质的区别,强化学习系统也许根本不知道正确的答案或策略是什么,只要给计算机一个要达到的目标,则强化学习就可通过反复试验和来自环境的连续反馈来学习怎样达到该目标.强化学习不像传统的人工智能那样,完全基于符号运算,用自上而下的方式,而是像生物体那样自下而上,以感觉和动作为基础,在与环境的交互中学习,这种学习更接近人脑的思维过程[1].

2 强化学习原理(Principle of reinforcement learning)

基于神经网络的强化学习基本原理如图 1 所示[2],其中神经网络用于存储在学习过程中所学习到的经验和信息.在 t 时刻,网络根据接收到的环境状态 x_t 和当前奖励或强化信号 r_t,选择控制作用 a_t 作用于环境.环境接收到控制作用 a_t 以后状态变为 x_{t+1},同时学习系统接收到下一个奖励信号 r_{t+1}.系统在一系列控制作用下,从某一状态开始进行到末端状态这一过程中所接收到的奖励总和可表示为

$$V(x_t) = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} = r_{t+1} + \gamma V(x_{t+1}). \quad (1)$$

其中 $0 < \gamma \leqslant 1$ 为加权系数，$V(x_t)$ 称为状态评价函数.学习的目的就是要使得总的奖励之和最大,这正是动态规划的基本原理.

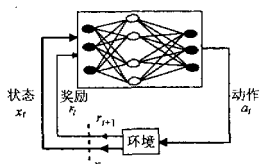

图 1 强化学习基本原理
Fig. 1 Principle of reinforcement learning

* 基金项目:国家自然科学基金(69904002);国防预研基金及航天科技创新基金资助项目.
收稿日期:2000－01－10;收修改稿日期:2000－11－22.

若某一策略 π 所对应的评价函数用 $V^{\pi}(x)$ 表示，则最优策略 π^* 所对应的评价函数称为最优评价函数，用数学描述为：

$$V^*(x) = \max_{\pi} V^{\pi}(x), \quad \forall x \in X.$$

评价函数是状态或状态/作用对的函数，通过定义不同的评价函数，可得到不同的强化学习算法，如常见的 Q-学习[3]．神经网络的作用就是用来学习、近似和推广这些评价函数．

3　A-学习(A-Learning)

在对 Q-学习算法研究时发现，Q-学习对于小的时间步长情况学习速度极其缓慢，甚至不能收敛．而 A-学习(advantage learning，简称 A-学习)则克服了这一不足[4]．在该算法中，状态评价函数 $V^*(x_t)$ 和 A 函数 $A^*(x_t, u_t)$ 分别定义如下

$$V^*(x_t) = \max_{u} A^*(x_t, u_t),$$

$$A^*(x_t, u_t) = V^*(x_t) + \frac{\langle r(x_t, u_t) + \gamma V^*(x_{t+1}) \rangle - V^*(x_t)}{\Delta t K}.$$

其中 K 是时间比例因子，$\langle \cdot \rangle$ 表示对所有状态和作用的期望值．于是，一个状态/作用对的 A 函数可以看作是 x_t 的状态评价函数 $V(x_t)$ 与执行动作 u_t 而使得加权强化信号总和增加的期望速率之和．将上面两式合并得到[4]

$$A^*(x_t, u_t) = \max_{u_t} A(x_t, u_t) + \frac{\langle r(x_t, u_t) + \gamma \max_{u_{t+1}} A(x_{t+1}, u_{t+1}) \rangle - \max_{u_t} A(x_t, u_t)}{\Delta t K}. \tag{2}$$

从式(2)可以看出，A-学习与 Q-学习算法很相似，只是增加了时间步长 Δt，且 Δt 可以取得很小，以适应小步长情况．

4　评价函数的神经网络近似(Approximation of value function using neural networks)

由于神经网络具有任意逼近、容错和推广等特点，因此，用神经网络来近似评价函数，即可存储所学习过的经验和信息，也可对没有学习到的状态进行推广．但是，对于状态空间比较大的强化学习系统而言，一个神经网络可能只对某一状态子空间具有较好的学习和推广性能，而对其它状态子空间的推广能力可能较差．因此，对于这样的系统可以采用多神经网络结构来近似评价函数，每个神经网络对应某一局部状态子空间，实现的途径就是采用一个控制网络自适应进行选通或加权，一般原理如图 2 所示[5]．图中 $A_i(i = 1,2,\cdots,n)$ 为第 i 个评价子网络的输出，$s_i(i = 1,2,\cdots,n)$ 为选通网络的输出．

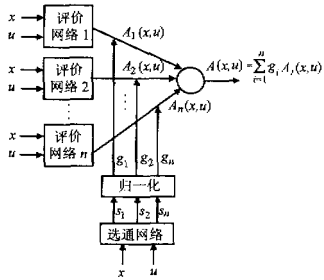

图 2　评价函数的多神经网络近似原理
Fig. 2　Approximating value function using modular neural networks

$$s_i = \sum_{j=1}^{p} x_j w_{ij} + \sum_{j=1}^{q} u_j w_{i(p+j)},$$

其中 w_{ij} 为选通网络的第 j 个输入节点与第 i 个输出节点之间的连接权值，p，q 分别为状态向量 x 和控制向量 u 的维数．假设评价函数子网络的选通概率符合正态分布，对选通网络的输出 s_i 进行归一化处理得到

$$g_i = \frac{e^{s_i}}{\sum_{j=1}^{n} e^{s_j}}.$$

总的近似评价函数为各个子评价网络输出的加权之和

$$A(x_t, u_t) = \sum_{i=1}^{n} A_i(x_t, u_t), \tag{3}$$

此时评价函数的理想值或整个评价网络的理想输出由式(2)确定，进而可以算出整个网络组学习的均方残差

$$E(x_t, u_t, w_i) = \frac{1}{N} \sum_{x_t} [A^*(x_t, u_t) - A(x_t, u_t)]^2.$$

其中 N 为学习过的状态总数，w_i 表示整个网络组的可调权值向量．可以采用 BP 学习算法将均方残差在各评价子网络和选通网络中同时进行反向传播，得到权值向量 w_i 修正算法

$$\Delta w_i = -\alpha \{ [r(x_t, u_t) + \gamma \max_{u_{t+1}} A(x_{t+1}, u_{t+1})] \frac{1}{\Delta t K} + \left(1 - \frac{1}{\Delta t K}\right) \max_{u_t} A(x_t, u_t) - A(x_t, u_t) \} \cdot \left\{ \frac{\gamma \partial \max_{u_{t+1}} A(x_{t+1}, u_{t+1})}{\partial w_i} \frac{1}{\Delta t K} + \right.$$

$$\left(1-\frac{1}{\Delta tK}\right)\frac{\partial \max\limits_{u_t} A(x_t,u_t)}{\partial w_t}-\frac{\partial A(x_t,u_t)}{\partial w_t}\Bigg\}.$$

其中 α 为学习速率,可采用自适应学习速率[2].

5 导弹制导问题(Problem of missile guidance)

考虑 X-Y 平面上导弹和飞机的二维制导问题,假设导弹和飞机的速度 $v_m, v_p(v_m > v_p)$ 为常值,速度方向与 X 轴之间的夹角分别为 θ_m, θ_p,则导弹和飞机状态由下列非线性动态系统支配

$$\begin{cases}\dot{x}_i = v_i\cos\theta_i,\\ \dot{y}_i = v_i\sin\theta_i,\end{cases}\quad i = m,p.$$

假设导弹和飞机只对其方向进行控制,令 $u = [u_m, u_p]^{\mathrm{T}}$ 分别表示导弹和飞机的法向加速度,且为开关量控制,定义

$$\theta_i = \begin{cases}\theta_i + 90^\circ, & u_i = 0.5,\\ \theta_i, & u_i = 0,\\ \theta_i - 90^\circ, & u_i = -0.5,\end{cases}\quad i = m,p.$$

强化函数 $r(x,u_m,u_p)$ 定义成导弹和飞机之间相对距离 d_{mp} 的函数,即

$$r(x,u_m,u_p) = \begin{cases}1, & d_{mp} > d_1(\text{飞机安全逃逸}),\\ -1, & d_{mp} < d_2(\text{导弹有效杀伤}),\\ 0, & d_1 \leqslant d_{mp} \leqslant d_2.\end{cases}$$

导弹试图极小化强化函数,而飞机则试图极大化强化函数,这是微分对策问题.此时 A 函数为

$$A^*(x_t,u_t^m,u_t^p) =$$

$$\min_{u_m}\max_{u_p}A(x_t,u_m,u_p) + \frac{1}{\Delta tK}[r(x_t,u_m,u_p) +$$

$$\gamma \min_{u_m}\max_{u_p}A(x_{t+1},u_m,u_p) - \min_{u_m}\max_{u_p}A(x_t,u_m,u_p)].$$

$A(x,u_m,u_p)$ 采用神经网络组来学习和推广,评价子网络的个数 $n = 4$,结构采用单隐层 BP 网络.所有网络的输入皆为状态量和控制量组合 $[x_m - x_p, \dot{x}_m - \dot{x}_p, y_m - y_p, \dot{y}_m - \dot{y}_p, u_m, u_p]^{\mathrm{T}}$ 网络的学习离线进行,但作为反馈控制器在线应用.

图3给出了均方残差的变化过程,图4为网络收敛后导弹和飞机在给定初始条件下的飞行轨迹,从仿真结果可以看出:当系统经过足够训练以后,导弹学会了怎样追击飞机,而飞机也学会了怎样逃避导弹.

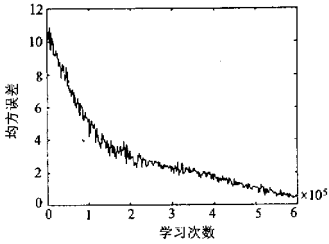

图3 均方残差变化趋势

Fig. 3 Figure of mean square error

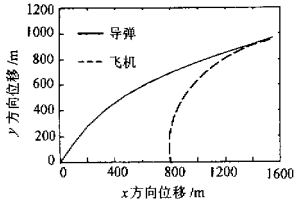

图4 导弹和飞机运动轨迹

Fig. 4 Trajectory of missile and plane

参考文献(References)

[1] Yan P F. Reinforcement learning - principles, algorithms and its applications in intelligent control [J]. Information and Control, 1996, 25(1):28 - 34(in Chinese)

[2] Xu B Z, Zhang B L and Wei G. Neural Networks and Its Applications[M]. Guangzhou: South China University of Technology Press, 1994

[3] Watkins J C H and Dayan P. Technical note: Q-learning [J]. Machine Learning, 1992, 8(4):279 - 292

[4] Baird L. Residual algorithms: Reinforcement learning with function approximation [A]. Proceedings of the Twelfth International Conference on Machining Learning [C], Morgan Kaufman Publishers, San Francisco, CA, 1995

[5] Jacobs R A and Jordan M I. Learning piecewise control strategies in a modular neural network architecture [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1993, 23(2):337 - 345

本文作者简介

周 锐 1968年生.北京航空航天大学自动控制系副教授.主要从事飞行控制,制导与智能控制等问题的研究.发表学术论文20多篇.

陈宗基 1943年生.北京航空航天大学自动控制系教授,博士生导师.主要研究方向:鲁棒与自适应控制,智能控制,混合系统,现代飞行控制系统设计及现代仿真技术等.发表学术论文100多篇.