第24卷 第5期
1998年9月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 24, No. 5
Sept., 1998

短 文

基于 Q 学习算法和 BP 神经网络的倒立摆控制[1)]

蒋国飞　　吴沧浦

（北京理工大学自动控制系　北京　100081）

摘　要　Q 学习是 Watkins[1]提出的求解信息不完全马尔可夫决策问题的一种强化学习方法. 将 Q 学习算法和 BP 神经网络有效结合，实现了状态未离散化的倒立摆的无模型学习控制. 仿真表明：该方法不仅能成功解决确定和随机倒立摆模型的平衡控制，而且和 Anderson[2]的 AHC (Adaptive Heuristic Critic)等方法相比，具有更好的学习效果.

关键词　Q 学习，BP 网络，学习控制，倒立摆系统，高斯噪声.

1　引言

在各种非线性系统中，倒立摆是一个十分典型的例子. 用强化学习方法来实现倒立摆的平衡控制，迄今已经取得了不少成果. 1983年 Barto 等人[3]设计了两个单层神经网络，采用 AHC(Adaptive Heuristic Critic)学习算法实现了状态离散化的倒立摆控制. 1989年，Anderson[2]进一步用两个双层神经网络和 AHC 方法实现了状态未离散化的倒立摆的平衡控制. 最近，Peng[4]通过将状态离散化成为162个区域，用 Lookup 表表示 Q 值的方法实现了基于 Q 学习算法的倒立摆的平衡控制. 然而在那些有连续状态的问题中，如果采用离散化这些连续量再用 Lookup 表来表示的方法，则 Q 学习算法和常规动态规划方法一样，存在状态变量的空间复杂性问题，即所谓的维数灾问题. 解决方法之一是用参数化的结构来表示 Q 值，如低阶多项式、决策树等. 本文通过训练 BP 网络来逼近 Q 值函数并利用 BP 网络的泛化能力，实现了基于 Q 学习算法的状态未离散化的确定和随机倒立摆的无模型学习控制. 本文的目的在于用倒立摆控制问题来证实：用 Q 学习和神经网络结合的方法去实现某些状态连续控制系统的无模型控制的可行性.

2　Q 学习

在介绍 Q 学习方法前，先简述有限马氏决策问题的模型：在每个时间步 $k=1,2,\cdots,$

1)国家自然科学基金重点资助项目.

收稿日期　1997-01-22

控制器观察马氏过程的状态为 x_k,选择决策 a_k,收到即时报酬 r_k,并使系统转移到下一个状态 y_k,转移概率为 $P_{x_k y_k}(a_k)$. 控制的目的是寻求一个最优控制策略,使未来每个时间步所获报酬的折扣和的期望最大,即极大化 $E\{\sum_{i=0}^{\infty}\gamma^i r_{k+i}\}$,其中$0\leq\gamma<1$为折扣因子.

给定一个策略 π,定义 Q 值为

$$Q^{\pi}(x,a)=R(x,a)+\gamma\sum_{y}P_{xy}(a)V^{\pi}(y), \tag{1}$$

其中 $R(x,a)=E\{r\,|\,x,a\}$,$V^{\pi}(y)=\max_{b}Q^{\pi}(y,b)$.

换言之,Q 值即在状态 x 执行控制 a 及后续策略 π 的报酬折扣和的期望. Q 学习的目的就是在转移概率和所获报酬未知的情况下来估计最优策略的 Q 值. 为方便起见,定义 $Q^{*}(x,a)\equiv Q^{\pi^{*}}(x,a)$,$\forall\ x,a$. 其中 π^{*} 表示最优策略.

在线 Q 学习方法实现如下:在每个时间步 k,观察当前状态 x_k,选择和执行控制 a_k,再观察后继状态 y_k 及接受即时报酬 r_k,然后根据下式调整 Q_{k-1} 值:

$$Q_k(x,a)=\begin{cases}(1-\beta_k)Q_{k-1}(x,a)+\beta_k[r_k+\gamma V_{k-1}(y_k)], & \text{如果}(x,a)=(x_k,a_k),\\ Q_{k-1}(x,a), & \forall\ (x,a)\neq(x_k,a_k),\end{cases} \tag{2}$$

其中 β_k 为学习因子,$V_{k-1}(y)\equiv\max_{b}\{Q_{k-1}(y,b)\}$.

Watkins[1]证明了学习因子序列$\{\beta_k\}$在满足一定的条件下,如果任一个(x,a)二元组能用等式(2)进行无穷多次迭代,则当 $k\to\infty$时,$Q_k(x,a)$以概率1收敛于 $Q^{*}(x,a)$.

3　倒立摆系统

倒立摆系统是指图1所示的非线性系统. 小车可以自由地在限定的轨道上左右移动. 小车上的倒立摆一端被铰链在小车顶部. 另一端可以在小车轨道所在的垂直平面上自由转动. 控制的目的在于通过推动小车向左或向右移动,使倒立摆平衡并保持小车不和轨道两端相撞.

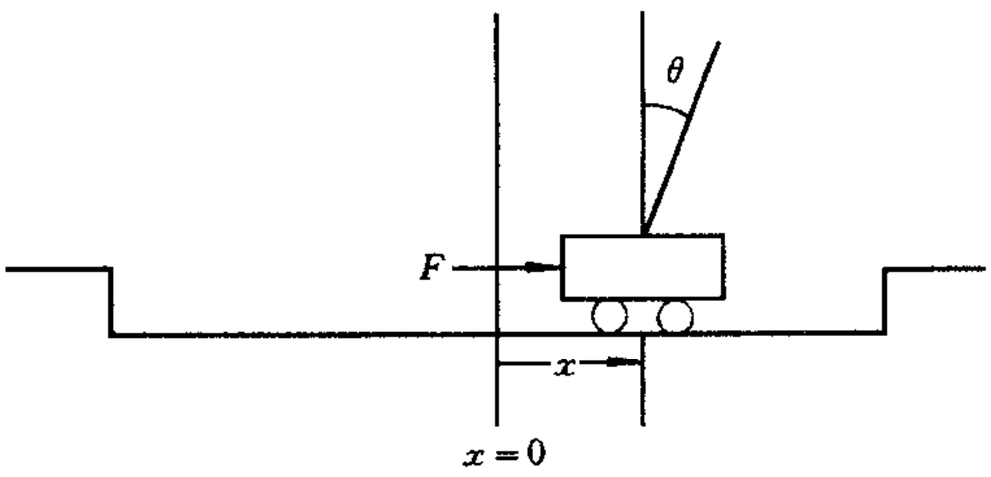

图1　倒立摆系统

一般情况下,倒立摆系统有四个状态变量

x:小车在轨道上的位置;　　θ:倒立摆偏离垂直方向的角度;

$\dot{x}$:小车的运动速度;　　$\dot{\theta}$:倒立摆的角速度.

倒立摆系统可以用以下运动方程来描述

$$\ddot{\theta}_t=\frac{g\sin\theta_t+\cos\theta_t\left[\dfrac{-F_t-ml\dot{\theta}_t^2\sin\theta_t+\mu_c\mathrm{sgn}(\dot{x}_t)}{m_c+m}\right]-\dfrac{\mu_p\dot{\theta}_t}{ml}}{l\left[\dfrac{4}{3}-\dfrac{m\cos^2\theta_t}{m_c+m}\right]}, \tag{3}$$

$$\ddot{x}=\frac{F_t+ml[\dot{\theta}_t^2\sin\theta_t-\ddot{\theta}_t\cos\theta_t]-\mu_c\mathrm{sgn}(\dot{x}_t)}{m_c+m}, \tag{4}$$

其中 $g=9.8\mathrm{m/s^2}$,重力加速度;$m_c=1.0\mathrm{kg}$,小车质量;$m=0.1\mathrm{kg}$,倒立摆质量;$l=0.5\mathrm{m}$,倒立摆的一半长度;$\mu_c=0.000\ 5$,小车和轨道的摩擦系数;$\mu_p=0.000\ 002$,倒立摆和小车

的摩擦系数；$F_t=\pm 10.0N$，在时刻 t 作用于小车质心的力. 小车轨道长度为4.8米. 通过 Euler 方法数值近似，可用以下差分方程来仿真倒立摆系统

$$x(t+1)=x(t)+\tau\dot{x}(t), \tag{5}$$

$$\dot{x}(t+1)=\dot{x}(t)+\tau\ddot{x}(t), \tag{6}$$

$$\theta(t+1)=\theta(t)+\tau\dot{\theta}(t), \tag{7}$$

$$\dot{\theta}(t+1)=\dot{\theta}(t)+\tau\ddot{\theta}(t). \tag{8}$$

时间步 τ 一般设为0.02秒. 显然以上给出的倒立摆系统是一个确定性系统. 本文为了说明基于 Q 学习和神经网络的方法同样适用于连续随机系统的无模型控制，在以上确定性倒立摆模型中引入一个噪声信号来构成一个随机倒立摆模型，即在仿真中用以下方程来代替方程(6).

$$\dot{x}(t+1)=\dot{x}(t)+\tau\ddot{x}(t)+\xi(\mu,\sigma^2), \tag{9}$$

其中 $\xi(\mu,\sigma^2)$ 为高斯噪声.

4 基于 Q 学习和 BP 网络的倒立摆控制

和其他实现倒立摆控制的方法不同，在强化学习方法中，控制器唯一能从环境得到的反馈是当倒立摆偏离垂直方向的角度超出 $\pm 12°$ 或小车在 ± 2.4 米处和轨道两端相撞时环境给出的一个失败信号. 因此本文定义即时报酬 r_t 为

$$r_t=\begin{cases}-1, & \text{如果}|\theta_t|>12°\text{ 或 }|x_t|>2.4\text{m}.\\ 0, & \text{其他}.\end{cases}$$

由于控制器是在执行了一系列决策后才得到这个延迟的失败信号，则控制器必须解决奖励或惩罚随时间分配的问题，即确定在这过程中哪些决策应该对最后的失败负责. 实际上 Q 学习算法是在各时间步 Q 值的更新迭代中将这失败信号进行反传并根据 Q 值来确定相应决策的优劣. 本文由于设倒立摆平衡失败时的即时报酬为负($r_t=-1$)，因此对应 Q 值较小的决策就更有可能导致倒立摆系统的平衡失败. 同时在实现状态未离散化的倒立摆控制时，控制器还涉及状态空间很大时 Q 值函数的泛化问题(也叫奖励或惩罚随结构分配问题).

本文提出的基于 Q 学习和 BP 网络的状态未离散化倒立摆控制系统的结构如图2所示. 为方便起见，在图中定义状态 $X=(x,\dot{x},\theta,\dot{\theta})^T$. BP 网络的输入为状态 X 和决策 a，输出为 $Q(X,a,W)$，其中 W 为网络权重. 整个控制系统工作如下：在每个时间步 k，观测倒立摆的当前状态为 X_k，根据 BP 网的实际输出 $Q(X_k,a,W_k)$ 值并按照某种探索策略来选择当前决策 a_k. 然后观测倒立摆的后继状态 Y_{k+1} 并检测是否有失败信号(确定即时报酬 r_k). 系统再根据(10)式更新二元组 (X_k,a_k) 的 Q 值，然后利用误差信号 $e=Q(X_k,a_k)-Q(X_k,a_k,W_k)$ 更新 BP 网的权重 W_k 为 W_{k+1}，使 BP 网实际输出逼近更新后的理想输出 $Q(X_k,a_k)$，然后再转到状态 Y_{k+1} 继续以上的过程. 由于未对状态空间离散化，在系统中利用了 BP 网的泛化能力来求解未曾训练过的状态—决策二元组的 Q 值. 另外，在 BP 网络的权重学习中，对任一状态和决策所对应的 Q 值进行逼近都可能会影响该状态和另一控制所对应的 Q 值，所以图2中所示的 BP 网实际上可以用两个 BP 网来代替(每个控制一个)，这样可以期望得到更好的学习效果

$$Q(X_k,a_k)=(1-\beta_k)Q(X_k,a_k,W_k)+\beta_k(r_k+\gamma\max_b Q(Y_{k+1},b,W_k)). \tag{10}$$

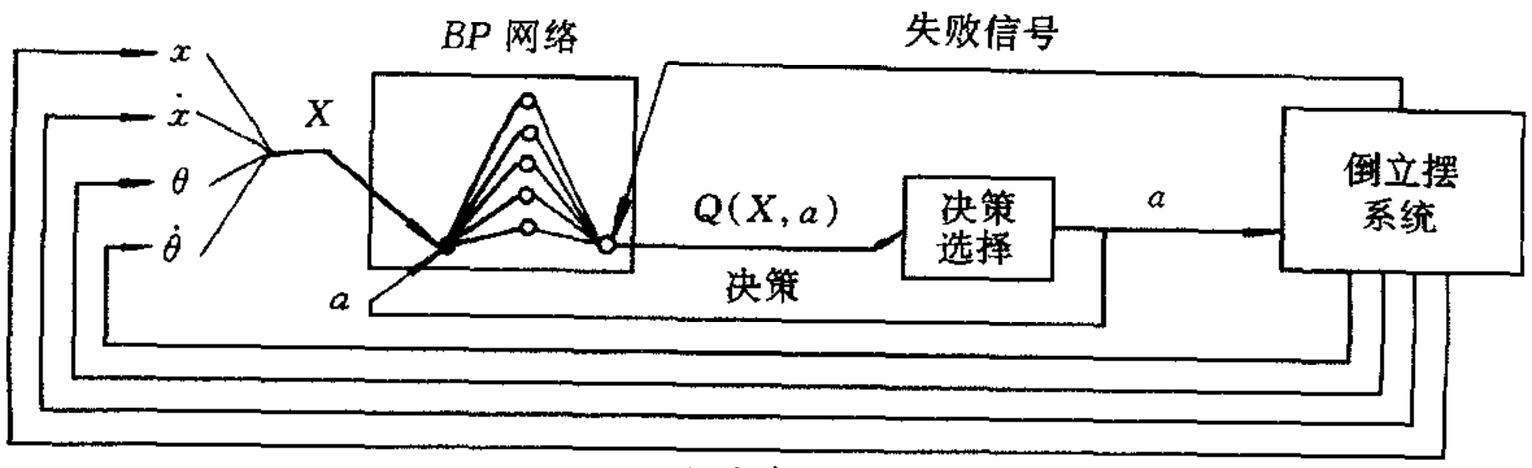

图2　基于 Q 学习和 BP 网络的状态未离散化倒立摆控制系统的结构图

5　仿真及结果

如上节所述,在实际仿真中,采用了两个 BP 网络. 每个网络分三层,输入层和隐层各有五个结点,输出层有一个结点. 对 BP 网络的实际输入进行了标准化,使其分布在[−1.0,1.0]之间. Q 学习算法的学习因子 $\beta=0.2$,折扣因子 $\gamma=0.95$. 在随机倒立摆模型的仿真中,高斯噪声 $\xi(\mu,\sigma^2)$ 的均值 $\mu=0$,标准差 $\sigma=0.1$(一般 $\dot{x}(t)$ 的值在(−1.5,1.5)之间). 在倒立摆控制系统中可行控制只有两个(左推或右推),因此本文直接选择对应 Q 值较大的控制为当前控制. 将随机发生器的种子和 BP 网初始权重设为不同值,对确定性和随机倒立摆模型各做十次试验,每次试验当倒立摆的试探次数(失败次数)超过100 00次或一次试探的平衡步数超过500 000步时,中止倒立摆的学习并重新开始另一次试验. 在仿真中,如果倒立摆在一次试探中能保持500 000步不到,就认为本次试验已经能成功控制倒立摆平衡了. 在状态离散化的倒立摆控制中,每次平衡失败后,倒立摆的初始状态一般设在 $X=0$的位置. 在仿真中,为了保证倒立摆得到在各种场合的控制经验,在每次平衡失败后,将初始状态复位为一定范围内的随机值. 平均十次试验的结果,得到基于 Q 学习和 BP 网络的状态未离散化倒立摆控制的结果如图3所示.

在同样条件下,将基于 Q 学习方法、AHC 方法和随机控制(无学习)方法实行状态未离散化的确定性倒立摆模型控制的结果进行比较,发现基于 Q 学习和 BP 网络的方法学习效果最好,每次试验在平均1 000次失败后就可以成功控制倒立摆平衡. 而用 Anderson[2]的两层网络和 AHC 方法则大约要6 000次. 随机控制方法不能控制倒立摆平衡,每次试探最多只能运行几百步. 基于 Q 学习和 BP 网络的方法同样可以实现随机倒立摆模型的平衡控制,但由于随机噪声的引入增加了学习难度,每次试验平均要在2 600次失败后才能控制倒立摆平衡.

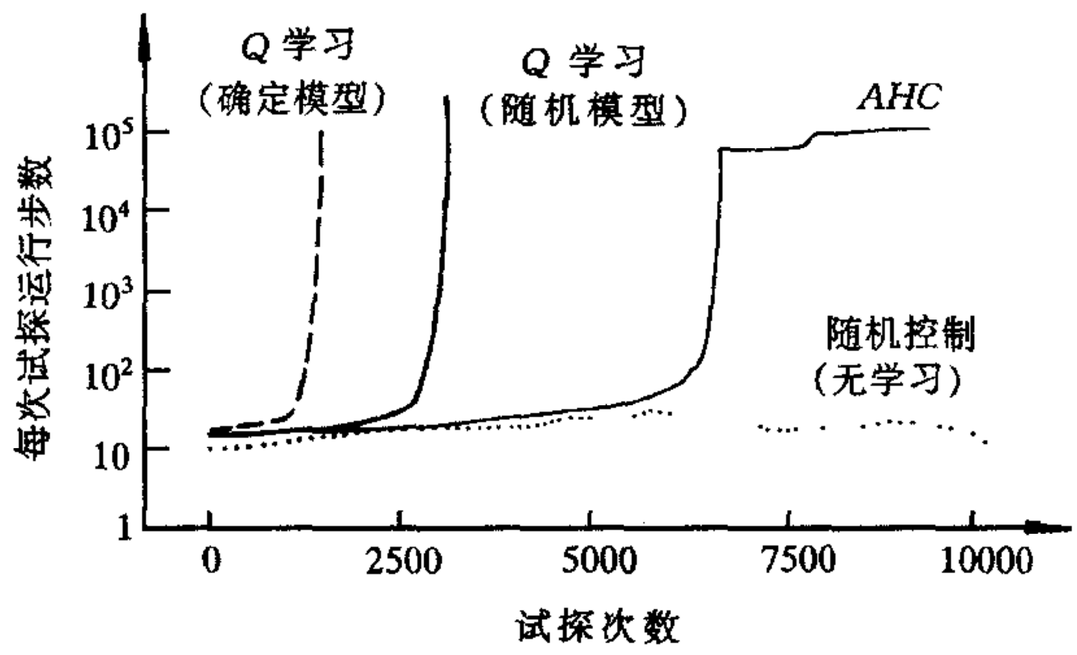

图3　各方法实现状态未离散化倒立摆控制的学习曲线

需要说明的是,尽管基于 Q 学习和 BP 网络的方法取得较好控制效果,但作者更关注的是验证了这种方法在实现某些状态连续控制系统的无模型控制的可行性. 实际上,倒立

摆控制问题只是以上问题的一个例子.

参 考 文 献

1 Watkins C J C H. Learning from delayed rewards[Ph. D. Dissertation]. UK: King's College, 1989

2 Anderson C W. Learning to control an inverted pendulum using nerual networks. *IEEE Control System Magazine*, 1989, **9**(3): 31—37

3 Barto A G, Sutton R S, Anderson C W. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Trans. on SMC*, 1983, **13**(5): 834—846

4 Peng J. Efficient dynamic programming-based learning for control[Ph. D. thesis]. USA: Northeastern University, 1993

LEARNING TO CONTROL AN INVERTED PENDULUM USING *Q*-LEARNING AND NEURAL NETWORKS

JIANG GUOFEI WU CANGPU

(*Department of Automatic Control, Beijing Institute of Technology, Beijing* 100081)

Abstract *Q*-learning is a reinforcement learning method to solve Markovian decision problems with incomplete information. This paper presents a novel method to control an inverted pendulum with unquantized states by using *Q*-learning and neural networks. Simulation results are included to show that the new method can not only balance the determined or stochastic inverted pendulums successfully but also lead to a better effect of learning when compared with Anderson's AHC method.

Key words *Q*-Learning, BP neural network, learning control, inverted pendulum, Gaussian noise.