第 17 卷 第 4 期　　控 制 与 决 策　　2002 年 7 月
Vol. 17 No. 4　　Control and Decision　　July 2002

文章编号: 1001-0920(2002)04-0473-03

一种基于增强学习的自适应控制方法

顾冬雷, 陈卫东, 席裕庚
(上海交通大学 自动化研究所, 上海 200030)

摘　要: 针对模型未知时变非线性对象的控制问题, 提出一种直接的自适应控制策略。该策略基于径向基神经网络并结合增强学习的自调节能力, 无需知道控制对象的动态特性, 而是通过在线试错在控制过程中不断积累与问题相关的信息, 用以产生可接受的逐步优化的解。

关键词: 增强学习; 径向基神经网络; 自适应控制
中图分类号: TP 24　　**文献标识码**: A

A novel adaptive control algorithm based on reinforcement learning

GU Dong-lei, CHEN Wei-dong, XI Yu-geng
(Institute of Automation, Shanghai Jiaotong University, Shanghai 200030, China)

Abstract: A direct adaptive control schema for model unknown time dependent dynamical plant is presented. Combining the ability of radial based function network and the self tune property of reinforcement learning, it needs not to know the priori knowledge of the plant's dynamical property. The knowledge related to the problem is accumulated gradually by trial and error method and then the acceptable optimal solution can be obtained.

Key words: reinforcement learning; radial based function network; adaptive control

1 引　　言

智能控制方法将一些启发式方法融入控制策略, 以模拟人的决策过程, 能在一定程度上解决复杂动态系统的控制问题[1]。本文利用增强学习方法[2]能在动态环境中不断进行试错学习而得到合适行为的能力, 设计出一种智能控制策略, 以控制未知动态特性的对象。该控制策略采用的增强学习方法基于 Actor/Critic 框架结构[3,4]。其中, Actor 部分由一个径向基神经网络(RBFN)构成, 它的学习过程包括结构适应(不断增长网络规模以容纳新的经验)和参数适应(产生逐步优化的策略); Critic 部分则比较被控对象的输出与参考输入之间的差距, 检验当前控制策略的效果, 产生"奖励/惩罚"值作为反馈, 以进行自适应学习。仿真结果显示了该控制策略良好而迅速的适应能力。

2 Actor/Critic 增强学习算法

2.1 基本算法流程

如图1所示, Actor/Critic算法是一种自适应

收稿日期: 2000-11-07; **修回日期**: 2001-04-19
基金项目: 国家自然科学基金项目(60105005)
作者简介: 顾冬雷(1972—), 男, 江苏启东人, 博士生, 从事增强学习、机器人协调控制的研究; 席裕庚(1946—), 男, 上海人, 教授, 博士生导师, 从事预测控制、智能机器人等研究。

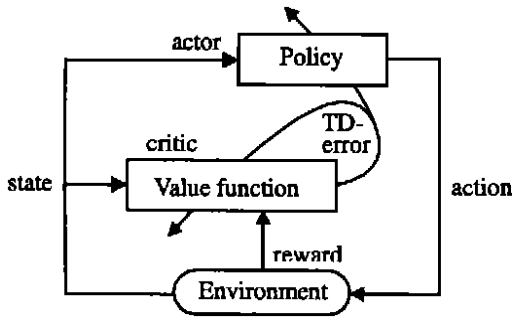


图 1 Actor/Critic 体系结构

策略迭代方法[2]。它由两部分组成: 策略评价部分和策略改进部分。其中, Actor 实现将感知状态映射到动作状态执行概率的随机策略, Critic为当前策略得到的评价函数。文献[4]利用 Actor's eligibility trace 的能力, 在 Critic 中近似函数非常不精确的情况下学到较好的控制策略。本文针对过程控制的状态空间和动作空间均为连续的特点, 对此算法作了一些处理, 使它可以用于连续过程的自适应控制。

2.2 基于 RBFN 的 Actor

Actor/Critic 方法容易在 Actor 中采用传统的有监督的学习技术, 将专家知识融入系统[5]。如图 2 所示, 本文采用增长的 RBFN 作为算法中的 Actor 部分, 对系统的经验逐步积累, 从而学到合适的控制策略。

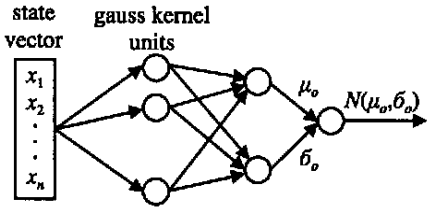


图 2 基于 RBFN 的 Actor

RBFN 对连续函数进行近似, 其输出为

$$\hat{y}(x) = \sum_{k=1}^{K} w_k \phi_k(x) \tag{1}$$

其中, x 为输入向量, $\hat{y}$ 为网络输出向量, ϕ 为基函数, w 为权重, K 为基函数的数目。基函数

$$\phi_k(x) = \exp\left(-\frac{\| x - q_k \|^2}{2\sigma^2}\right) \tag{2}$$

其中, q_k 代表基函数的中心, σ 代表它的宽度, 范数 $\| \cdot \|$ 为欧几里德距离。由于采用增长的 RBFN 结构, 基函数的中心数 K 不是预先设定的, 而是随着学习过程不断增长的值。唯一预先设置的值是 σ, 它可以看作模型精度。

关于 RBFN 的规模增长, 首先初始化第一个输入向量 x_1, 将它作为 RBFN 第 1 个基函数核。如果有新的样本 $(x_t, t = 2, 3, \cdots)$ 到来, 当它在当前所有的基函数 $(\phi_k(x_t), k = 1, 2, \cdots, K)$ 中的激活度小于某个阈值 θ 时, 则将它作为一个新的核, 即

$$\max_k \phi_k(x_t) < \theta \Rightarrow q_{K+1} = x_t \tag{3}$$

这样, 随着学习过程的不断进行, 学习系统以增量的方式逐渐学到过程对象的动态特性并加以记录。θ 应取合适的值, 以确保各基函数之间合适的相互覆盖程度。

如果将学习系统的输出作为一个概率密度分布, 则 Actor/Critic 方法可以处理连续的动作空间。令网络输出 $\mu_o(x)$ 和 $\sigma_o(x)$ 分别为正态分布随机变量的均值和方差, 控制量 a_t 的概率密度函数满足正态分布 $N(\mu_o, \sigma_o)$, 即

$$p(a_t) = \frac{1}{\sigma_o\sqrt{2\pi}}\exp\left(\frac{-(a_t - \mu_o)^2}{2\sigma_o^2}\right) \tag{4}$$

其中, 均值 μ_o 代表当前认为比较合适的系统输出, 方差 σ_o 代表系统对不同控制量的探索程度。随着学习的不断进行, 方差越来越小, 即系统逐渐收敛到某个策略。这种方法将增强学习对未知知识的探索和对已知知识的利用自然地结合在一起。

2.3 Critic 部分

Actor/Critic 算法的 Critic 部分从环境中得到立即奖励, 产生 TD-error, 对 Actor 中网络的权重进行更新, 使系统获得适应性。

立即奖励 r_t 通过比较实际输出 y_t 与参考输出 ref_{t-1} 之间的差值而获得。为了一致起见, 将奖励限制在 $[0, 1]$ 之间, 令 $r_t = \exp(-|y_t - \mathrm{ref}_{t-1}|)$。这样便使实际输出与参考输出比较接近的区域得到更多的关注, 因为优化更多地体现在这个区域的接近上。

增强学习算法需要迭代计算代价函数 $\hat{V}(x_t)$, 但是当 x 是连续的状态空间时, 无法得到一个普适的函数。为此采用针对 Actor 的 eligibility trace 方法[4], 当 $\hat{V}(x_t)$ 相当不精确时, 也能得到较好的策略。假定 $\hat{V}(x_t)$ 已收敛到近似最优, 则 $\hat{V}(x_{t+1}) - \hat{V}(x_t)$ 近似为常数。于是用常数项代替 TD-error 中的该项, 避免了求解 $\hat{V}(x_t)$, 从而避免了连续的状态空间带来的麻烦, 使它可以用于连续过程的控制。此处 TD-error $= r_t - b$, 其中 b 为常数项。网络最终的权值更新公式为

$$\Delta w_{\mu i} = (r_i - b)\frac{a_t - \mu_o}{\sigma_o}\phi_i(x_t) \tag{5}$$

$$\Delta w_{\sigma i} = (r_t - b)\frac{(a_t - \mu_o)^2 - \sigma_o^2}{\sigma_o^2}\phi_i(x_t) \tag{6}$$

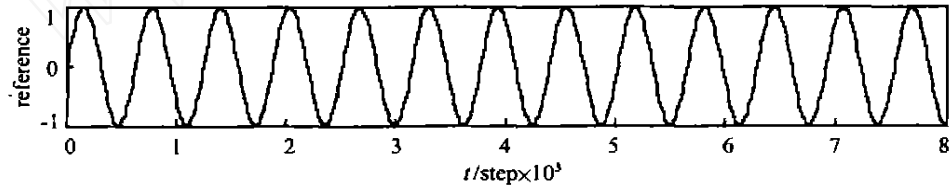
3　仿真结果

采用本文控制策略对如下时变的对象进行仿真研究

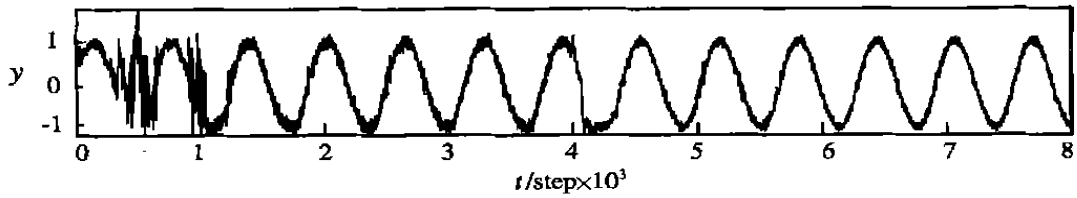
$$y(t+1)=\begin{cases}1.5y(t)/(1+y^2(t))+0.3\cos(y(t))+1.2u(t), & t\le 4\,000\\ 1.5y(t)/(1+y^2(t))+0.5\sin(y(t))+0.8u(t), & t>4\,000\end{cases}$$

假定参考输入 $\mathrm{ref}(t)=\sin(t/100)$, 输入向量 $x=(y_{t-1},y_{t-2},y_{t-3},\mathrm{ref}_{t-1},\mathrm{ref}_{t-2},\mathrm{ref}_{t-3},a_{t-1},a_{t-2},a_{t-3})^{\mathrm{T}}$。令式(3) 中的 $\theta=0.6$, 式(5) 中的 $b=0.95$, 对控制量 u 的限幅为[-1, 1]。仿真结果如图 3 所示。

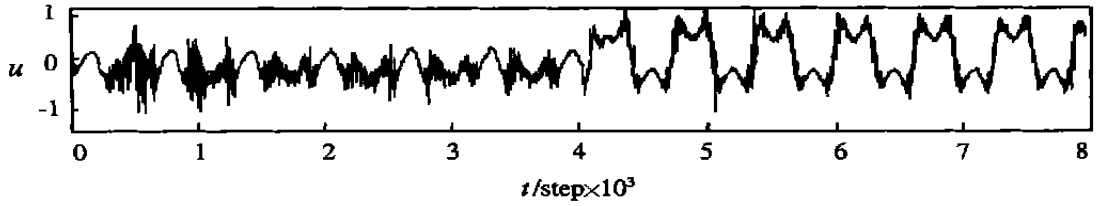
由图 3 可见, 初始状态学习系统对过程对象特性一无所知, 但随着网络规模不断增大, 学习系统能不断积累对过程对象动态特性的经验。到 1 000 步时, 网络规模基本不再增长, 学习系统控制下的对象输出能较好地跟随参考曲线。在 4 000 步时, 被控对象的动态特性发生改变, 系统能很快地学到新的动态特性, 表现出 K 值有所增大, 但网络的规模则不再变大。被控对象动态特性的变化导致控制模式的显著改变, 控制量 u 的模式发生了明显的改变。当被控对象的输出受到干扰时, 在经历了短暂的波动之后, 系统能很好地跟踪参考输入。



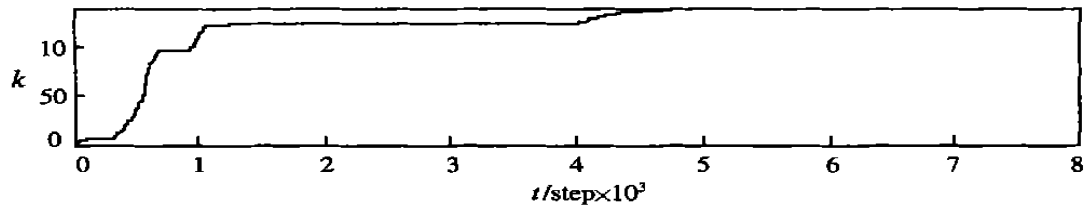
(a)　参考输入



(b)　对象输出 y



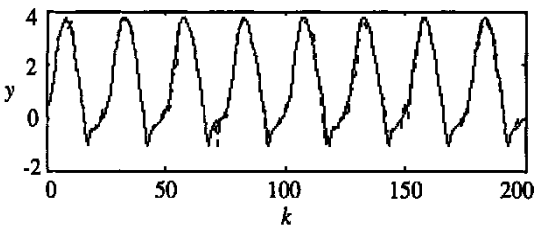
(c)　控制量 u



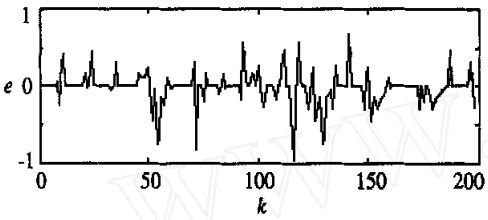
(d)　学习特性 K

图 3　系统仿真结果

(下转第 479 页)



(a)　验证曲线



(b)　验证误差曲线

图 2　仿真验证结果

5　结　　论

本文分析了当模糊逻辑系统采用非单点模糊产生器对输入进行模糊化时, 非单点模糊逻辑系统所具有的前置滤波特性, 并对此进行数学描述和解释。通过仿真验证以及与单点模糊逻辑系统的性能比较, 充分说明了非单点模糊逻辑系统具有较强的前置滤波能力和鲁棒性。该系统能较好地解决工程应用中输入噪声干扰问题, 具有较强的实际工程应用价值。

参考文献(References):

[1] 王立新. 自适应模糊系统与控制[M]. 北京: 国防工业出版社, 1995.

[2] Mouzouris G C, J M Mendel. Non-singleton fuzzy logic systems: Theory and appl[J]. *IEEE Trans on Fuzzy Systems*, 1997, 5(1): 56-71.

[3] 张骏, 王楠, 魏炳波, 等. 随机模糊神经网络在快速凝固研究中的应用[J]. 西北工业大学学报, 2000, 18(2): 336-339.
(J Zhang, N Wang, B Wei, et al. Application of stochastic fuzzy neural networks in rapid dendritic solidification [J]. *J of Northwestern Polytec Univ*, 2000, 18(2): 336-339.)

[4] J S Jang, C T Sun, E Mizutani. *Neuro-fuzzy and Soft Computing* [M]. Tokyo: Prentice Hall, 1997.

[5] Ronald R Yager, Dimitar P Filev. Approximate clustering via the mountain method[J]. *IEEE Trans on Syst, Man & Cybern*, 1994, (8): 1274-1284.

(上接第 475 页)

4　结　　语

本文利用 Actor/Critic 增强学习算法并结合 RBFN 神经网络构成一个直接的自适应控制器, 用于控制动态特性时变的未知过程对象。仿真结果表明, 该学习算法能较快地适应动态特性的变化, 学到合适的控制策略。对于那些难以建模的复杂系统, 当控制精度要求不高时, 本文方法是一种富有吸引力的选择。

参考文献(References):

[1] D A Linkens, H O Nyongesa. Learning systems in intelligent control: An appraisal of fuzzy, neural and genetic algorithm control applications[J]. *IEE Proc on Control Theory and Appl*, 1996, 143(4): 367-386.

[2] Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey[J]. *J of Artificial Intell Res*, 1996, 4: 237-277.

[3] Barto A G, Sutton R S, Anderson C W. Neuronlike adaptive elements that can solve difficult learning control problems[J]. *IEEE Trans on Syst, Man & Cybern*, 1983, 13(5): 834-846.

[4] Kimura H, Kobayashi S. An analysis of actor/critic algorithms using eligibility traces: Reinforcement learning with imperfect value function[A]. *15th Int Conf on Machine Learning* [C]. Madison, 1998. 278-286.

[5] Clouse J A, Utogoff P E. A teaching method for reinforcement learning[A]. *Proc of the 9th Int Conf on Machine Learning* [C]. San Mateo, 1992. 93-101.