第 21 卷 第 10 期　　控 制 与 决 策　　2006 年 10 月
Vol. 21 No. 10　　Control and Decision　　Oct. 2006

文章编号: 1001-0920(2006)10-1087-05

基于复杂过程简化模型的DHP学习控制

陈宗海, 文 锋
(中国科学技术大学 自动化系, 合肥 230027)

摘 要: 提出一种基于简化模型的DHP(Dual Heuristic Programming)方法的学习控制, 避免了标准DHP方法需要被控对象的精确模型来求得对于状态和控制动作的Jacobian矩阵, 而是利用简化过程对象模型获得近似Jacobian矩阵, 实现学习控制的需要. 生化反应器定值控制的仿真结果表明, 该方法加快了学习过程, 并对更大范围的参数变化具有鲁棒性.

关键词: 强化学习; DHP方法; 生化反应器; 简化模型
中图分类号: TP13　　**文献标识码**: A

Learning Control of DHP Method Based on Complex Process Simplified Model

CHEN Zong-hai, WEN Feng
(Department of Automation, University of Science and Technology of China, Hefei 230027, China. Correspondent: CHEN Zong-hai, E-mail: chenzh@ustc.edu.cn)

Abstract: The standard DHP method needs accurate plant model to calculate the Jacobian matrix of state and control action, which is difficult to meet. A learning control strategy based on DHP (Dual Heuristic Programming) method of simplified model is proposed, which applies approximate Jacobian matrix to DHP training and thus relaxes this limitation. Simulation results of contrapose Bioreactor show that the proposed method can accelerate learning process and is robust to larger parameter changes.

Key words: Reinforcement learning; DHP method; Bioreactor; Simplified model

1 引　　言

日益上涨的能源费用、产品价格和质量上越来越激烈的全球竞争以及使产品更加环保的要求, 使得生化过程对于现代制造业来说越来越重要. 通过生化反应过程得到的许多产品通常无法使用常规制造手段生产, 而这些产品属于量少而价高的范畴, 如药品等[1]. 一般生化过程都在生化反应器中完成. 为了增加产量, 减少产品质量的波动, 提高资源利用率, 通常需要使生化反应器保持在某个状态下. 生化反应过程多是非线性、大时延等典型的复杂过程对象, 并且在过程中既有细胞水平的菌体代谢, 又有工程水平的物质和能量传送, 难以用常规方法实现控制. 生化反应器的这些特点, 使其非常适合作为学习控制方法的一个测试对象, 用以检验学习算法的性能. Anderson 和Miller[2]给出了作为学习控制基准问题之一的生化反应控制问题, 其中分设定值在状态空间中的稳定区域和不稳定区域两种情况, 后者的控制难度更大. 目前, 还没有其他作者使用学习控制方法解决此问题.

本文提出一种基于简化模型的DHP学习方法, 该方法属于ADP[3~8](Approximate Dynamic Programming)方法中的一种. 标准的DHP方法需要被控对象的精确模型, 用以计算模型对于状态和控制量的Jacobian矩阵. 实际问题中, 一般难以得到被控对象的精确模型, 并且模型的参数还可能出

收稿日期: 2005-07-11; **修回日期**: 2005-08-31.
基金项目: 国家自然科学基金项目(60575033).
作者简介: 陈宗海(1963—), 男, 安徽桐城人, 教授, 博士生导师, 从事复杂系统的建模与控制、智能机器人等研究; 文锋(1978—), 男, 湖南祁东人, 博士生, 从事智能控制等研究.

现漂移, 以及存在噪声影响等. 本文使用近似Jacobian 矩阵, 不需要精确模型, 从而避免了对模型求导的复杂计算, 便于实际应用.

在生化反应器的控制问题上, 本文分别使用基于简化模型的DHP 方法和标准的DHP 方法进行了仿真实验. 结果表明, 两种方法都能通过学习实现生化反应器的设定值控制, 但基于简化模型的DHP 方法大大加快了学习过程, 在不稳定区域设定值控制情况下表现尤为明显; 同时由于其对模型信息的要求较少, 对于更大范围的参数变化具有鲁棒性.

2 生化反应器控制问题[2]

生化反应过程大多反应缓慢, 既有细胞水平的菌体代谢过程, 又有工程水平的物质和能量传送过程, 使得这类对象具有强非线性、大时延等特性. 生化反应器由一个水箱构成, 箱中装有水、营养物质和生物细胞的混合物, 其状态是细胞的浓度和营养物质的浓度. 反应器的输入和输出流量相等, 以保持液位不发生变化, 此流量即为对象的控制量. 虽然生化反应器控制问题的状态个数少, 但是存在很强的非线性, 且在控制量的改变和细胞浓度的变化之间存在较大的时延.

生化反应器的离散时间模型[1] 为

$$\begin{aligned} c_1(k+1) &= c_1(k) + \Delta[-c_1(k)u(k) + \\ &\quad c_1(k)(1-c_2(k))\mathrm{e}^{c_2(k)/\gamma}], \\ c_2(k+1) &= c_2(k) + \Delta[-c_2(k)u(k) + \\ &\quad c_1(k)(1-c_2(k))\mathrm{e}^{c_2(k)/\gamma}(1+ \\ &\quad \beta)/(1+\beta-c_2(k))], \end{aligned}$$

其中: c_1 为细胞浓度, c_2 为营养物质浓度, 满足 $0 \le c_1, c_2 \le 1$; u 为流量, $0 \le u \le 2$; $\beta = 0.02$ 为细胞生长参数; $\gamma = 0.48$ 为营养对于细胞生长抑制作用的参数; 采样周期 $\Delta = 0.01$ s.

生化反应器控制的设置[2] 为: 每隔 50 个时刻计算一次控制量, 然后保持控制量不变, 直到下一次重新计算控制量. 对象的初始状态都选取在设定值的 ±10% 之间的随机值. 控制的目标是通过改变输入输出流量, 使反应器中的细胞浓度维持在设定的水平上. 定义费用函数为

$$E = \sum_{k=1}^{\infty}[c(50k) - c_d(50k)]^2,$$

其中 c_d 为设定值, 每隔 50 个时刻计算一次费用.

Anderson 和Miller[2] 给出了生化反应器控制的两种设定值控制情况: 1) 设定值 $c_d = (0.120\,7, 0.880\,1)^{\mathrm{T}}$, 处于状态空间中的稳定区域; 2) 设定值 $c_d = (0.210\,7, 0.722\,6)^{\mathrm{T}}$, 处于状态空间中的不稳定区域.

3 基于简化模型DHP 方法的学习控制

DHP 方法通常归属于强化学习[3], 利用两个神经网络来进行学习控制: 动作网络产生控制动作, 评价者网络对动作网络的性能进行评估. 评价者网络用于估计 cost-to-go 函数对于状态的梯度, 其训练使微分Bellman 方程的误差最小; 动作网络的训练则使当前状态的 cost-to-go 函数最小. 相比于其他ADP 方法, Prokhorov[4] 指出DHP 方法更容易实现, 并且由于DHP 方法的评价者网络直接估计 cost-to-go 函数对于状态的梯度, 更容易得到 cost-to-go 函数对于控制量的梯度, 因而DHP 方法的学习过程也相对较快. Prokhorov[4,5] 给出了DHP 方法基于回归神经网络的实现, Lendaris[9] 则给出了DHP 方法基于多层感知器网络的实现.

3.1 评价者网络和动作网络的实现

动态规划中, 定义 cost-to-go 函数为

$$J(k) = \sum_{i=0}^{\infty}\gamma^i U(k+i),$$

其中 $U(k)$ 为费用函数, $J(k)$ 满足Bellman 方程: $J(x(k)) = U(k) + \gamma J(x(k+1))$. $J(k)$ 对状态 $x(k)$ 的导数, 记为 $\lambda(x(k)) = \partial J(x(k))/\partial x(k)$, 满足Bellman 方程的微分形式:

$$\lambda(x(k)) = \frac{\partial U(k)}{\partial x(k)} + \gamma\frac{\partial J(k+1)}{\partial x(k)}. \tag{1}$$

本文用多层感知器网络分别实现评价者网络和动作网络. 其中: 评价者网络用于估计函数 $\lambda(k)$, 并使得此估计能满足式(1); 动作网络则用于产生控制动作, 使 $J(k)$ 最小化. 评价者网络和动作网络中均含有一个隐含层, 其激活函数为 $f(x) = (1+\mathrm{e}^{-x})^{-1}$. 评价者网络的输出层为线性输出; 动作网络的输出层的激活函数与隐含层相同.

3.2 DHP 方法训练评价者网络和动作网络

评价者网络估计函数 $\lambda(k)$ 应满足式(1). 其训练最小化误差为

$$\begin{aligned} E_c &= \frac{1}{2}\sum_t e_c^{\mathrm{T}}(t)e_c(t), \\ e_c(k) &= \lambda(k) - \frac{\partial U(k)}{\partial x(k)} - \gamma\frac{\partial J(k+1)}{\partial x(k)} = \\ &\quad \lambda(k) - \lambda^d(k), \end{aligned}$$

其中 $\lambda^d(k)$ 为评价者的理想输出.

使用链式规则求导, 得到

$$\begin{aligned} \lambda^d(k) &= \frac{\partial U(k)}{\partial x_i(k)} + \gamma\frac{\partial J(k+1)}{\partial x_i(k)} = \\ &\quad \sum_{j=1}^{n}\gamma\lambda_j(k+1)\Big[\frac{\partial x_j(k+1)}{\partial x_i(k)} + \\ &\quad \sum_{l=1}^{m}\frac{\partial x_j(k+1)}{\partial u_l(k)}\frac{\partial u_l(k)}{\partial x_i(k)}\Big] + \end{aligned}$$

$$\left[\frac{\partial J(k)}{\partial x_i(k)} + \sum_{l=1}^{m} \frac{\partial J(k)}{\partial u_l(k)} \frac{\partial u_l(k)}{\partial x_i(k)}\right]$$

动作网络用于计算控制量, 其输出 $u(k)$ 应使 $J(k)$ 最小, 即令 $\partial J(k)/\partial u(k) = 0$. 因此动作者网络训练应最小化误差

$$E_a = \frac{1}{2}\sum_i e_a^{\mathrm{T}}(t) e_a(t),$$

$$e_a(k) = \frac{\partial J(k)}{\partial u(k)} - 0 =$$

$$\frac{\partial J(k)}{\partial u(k)} + \gamma\lambda(k+1)\frac{\partial x(k+1)}{\partial u(k)}.$$

评价者网络和动作网络学习过程均可以用图来表示. 动作网络的学习过程如图 1 所示, 其中虚线表示训练信号的产生过程

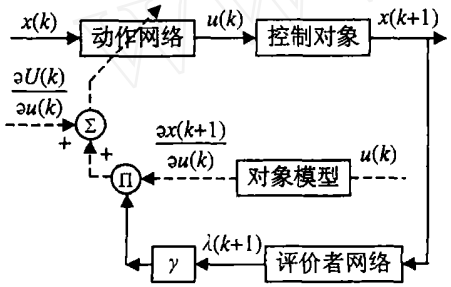

图 1　动作网络训练示意图

3.3　基于简化模型的DHP 方法

从DHP 算法可以看出, 其中需要用到被控对象的精确模型, 以计算模型对于状态和控制动作 Jacobian 矩阵, 即 $\partial x(k+1)/\partial u(k)$ 和 $\partial x(k+1)/\partial x(k)$. 对于生物反应器, 记 $x(k) = [c_1(k), c_2(k)]^{\mathrm{T}}$ 为对象的状态, 则有

$$\frac{\partial x(k+1)}{\partial x(k)} = \begin{bmatrix} \frac{\partial c_1(k+1)}{\partial c_1(k)} & \frac{\partial c_1(k+1)}{\partial c_2(k)} \\ \frac{\partial c_2(k+1)}{\partial c_1(k)} & \frac{\partial c_2(k+1)}{\partial c_2(k)} \end{bmatrix},$$

其中

$$\frac{\partial c_1(k+1)}{\partial c_1(k)} = 1 + \Delta[-u(k) + (1 - c_2(k))e^{c_2(k)/\gamma}],$$

$$\frac{\partial c_1(k+1)}{\partial c_2(k)} = \Delta[c_1(k)(1 - c_2(k))/\gamma - c_1(k)]e^{c_2(k)/\gamma},$$

$$\frac{\partial c_2(k+1)}{\partial c_1(k)} = \Delta(1 - c_2(k))e^{c_2(k)/\gamma}(1 + \beta)/(1 + \beta - c_2(k)),$$

$$\frac{\partial c_2(k+1)}{\partial c_2(k)} = 1 + \Delta[-u(k) - c_1(k)e^{c_2(k)/\gamma}(1 + \beta) \times (c_2^2(k) - c_2(k)\beta - 2c_2(k) + 1)/(1 + \beta - c_2(k))^2],$$

$$\frac{\partial x(k+1)}{\partial u(k)} = \begin{bmatrix} \frac{\partial c_1(k+1)}{\partial u(k)} \\ \frac{\partial c_2(k+1)}{\partial u(k)} \end{bmatrix} = \begin{bmatrix} -\Delta c_1(k) \\ -\Delta c_2(k) \end{bmatrix}.$$

在实际控制问题中, 对象的精确模型一般难以得到, 为此, 本文提出对模型进行简化, 使用近似 Jacobian 矩阵进行DHP 训练. 其分析过程如下.

对于生化反应器这类对象的连续时间模型可表示为

$$\dot{x} = f(x) + g(x)u,$$

对应的离散时间模型可表示为

$$x(k+1) = x(k) + \int_T^{(k+1)T}[f(x(\tau)) + g(x(\tau))u(\tau)]\mathrm{d}\tau =$$

$$x(k) + \int_T^{(k+1)T} f(x(\tau))\mathrm{d}\tau + \int_T^{(k+1)T} g(x(\tau))u(\tau)\mathrm{d}\tau =$$

$$x(k) + \int_T^{(k+1)T} f(x(\tau))\mathrm{d}\tau + u(k)\int_T^{(k+1)T} g(x(\tau))\mathrm{d}\tau, \quad (2)$$

其中 T 为采样周期, 并利用控制量在采样间隔之间不发生变化的事实, 得 $u(\tau) = u(k)$, 当 $kT \le \tau \le (k+1)T$ 时.

由式(2) 得

$$\partial x(k+1)/\partial x(k) = I + \partial\int_T^{(k+1)T} f(x(\tau))\mathrm{d}\tau/\partial x(k).$$

一般来说, 生化反应过程变化缓慢, 可以认为 $\partial\int_T^{(k+1)T} f(x(\tau))\mathrm{d}\tau/\partial x(k) \approx 0$, 于是得到近似值 $\partial x(k+1)/\partial x(k) \approx I$.

又由式(2) 得

$$\partial x(k+1)/\partial u(k) = \int_T^{(k+1)T} g(x(\tau))\mathrm{d}\tau$$

对模型进行简化, 可用 $\mathrm{sgn}\left(\int_T^{(k+1)T} g(x(\tau))\mathrm{d}\tau\right)$ 近似 $\partial x(k+1)/\partial u(k)$, 其中 $\mathrm{sgn}(\cdot)$ 为符号函数, 即利用其定性信息

使用简化模型训练评价者网络, 其训练误差为

$$e_c(k) = \lambda(k) - \lambda^{d\prime}(k),$$

$$\lambda_i^{d\prime}(k) = \sum_{j=1}^{n}\gamma\lambda_j(k+1)\left[I_{ij} + \sum_{l=1}^{m}\mathrm{sgn}\left(\frac{\partial x_j(k+1)}{\partial u_l(k)}\right)\frac{\partial u_l(k)}{\partial x_i(k)}\right] + \left[\frac{\partial J(k)}{\partial x_i(k)} + \sum_{l=1}^{m}\frac{\partial J(k)}{\partial u_l(k)}\frac{\partial u_l(k)}{\partial x_i(k)}\right],$$

其中 $I_{ij} = \begin{cases} 1, i = j; \\ 0, i \neq j. \end{cases}$

使用简化模型训练动作网络, 其训练误差为

$$e_a(k) = \frac{\partial \alpha(k)}{\partial u(k)} + \gamma\lambda(k+1)\,\mathrm{sgn}\left(\frac{\partial \alpha(k+1)}{\partial u(k)}\right).$$

对于生化反应器对象, 其简化模型为

$$\frac{\partial \alpha(k+1)}{\partial \alpha(k)} \approx I,$$

$$\mathrm{sgn}\left(\frac{\partial \alpha(k+1)}{\partial u(k)}\right) = \mathrm{sgn}\begin{pmatrix} -\Delta c_1(k) \\ -\Delta c_2(k) \end{pmatrix}.$$

由以上分析可见, 本文提出的基于简化模型的DHP 方法, 只需知道控制量对于状态转移影响的定性信息, 解除了标准DHP 方法需要对象精确模型的限制, 使该方法更符合实际控制的要求

4 仿真实验

分别使用基于精确模型的标准DHP 方法和基于简化模型的DHP 方法对生化反应器进行学习控制 初始的评价者网络和动作网络权值是随机初始化的, 表明初始时刻没有任何关于对象的知识 为了比较两种方法, 使用了相同的初始网络权值以及相同的初始状态 其他参数为 $\gamma = 0.9$, 网络的学习率为 0.01.

4.1 设定值控制实验

定义一次学习过程为从初态开始, 对生化反应器进行 100 步控制, 相当于 50 s 仿真控制过程 在稳定区域设定值控制下, 每组实验包含 500 次学习; 在不稳定区域设定值控制下, 每组实验包含 10 000 次学习. 实验结果表明, 两种设定值情况下, 标准DHP 方法和基于简化模型的DHP 方法都能够学习控制对象状态并达到设定值 图 2 和图 3 给出了一组控制结果

将每次学习所得费用和作为两种方法的性能指标, 进行 5 组实验, 并对数据求平均值, 得到在两种设定值情况下两种方法的学习过程, 如图4和图5

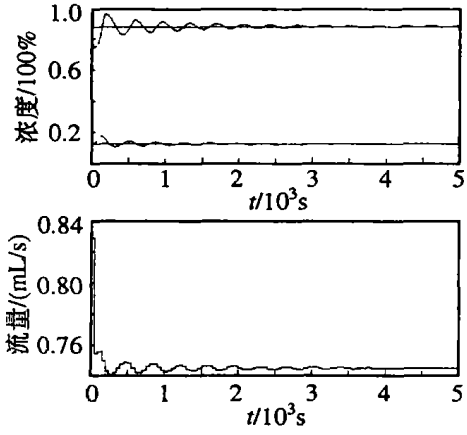

图 2 稳定区域设定值控制结果

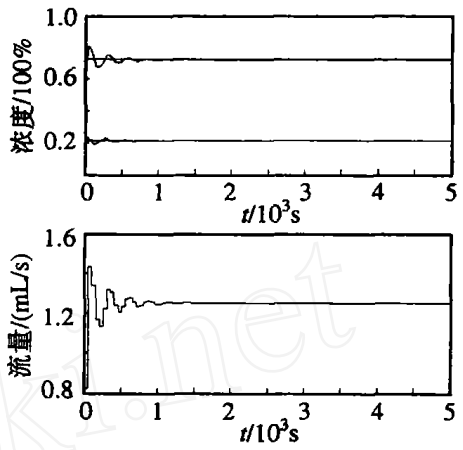

图 3 不稳定区域设定值控制结果

所示 可以发现, 不稳定区域设定值控制情况下平均需要 10 000 次学习, 远远大于稳定区域设定值控制情况下平均需要的 50 次学习. 这也验证了不稳定区域控制的难度比稳定区域控制的难度更大 在两种情况下, 基于简化模型的DHP 方法的学习过程都比标准DHP 方法的学习过程快, 在不稳定区域设定值控制的情况下表现尤为明显 实验达到了较好的控制效果

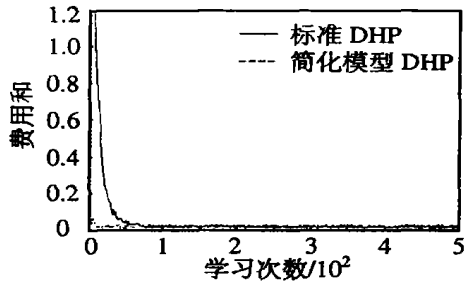

图 4 稳定区域设定值控制学习过程比较

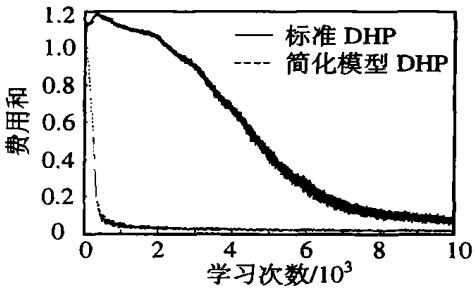

图 5 不稳定区域设定值控制学习过程比较

4.2 参数变化影响的实验

由于生化反应器的强非线性, 如果参数发生变化, 将严重影响控制器的性能 例如在采用内模控制时, 若参数 γ 变化 2%, 则设定值的误差将超过 50%. 为了研究参数变化时DHP 方法的性能, 设计两组实验, 对象模型中使用原来的参数 $\gamma = 0.48$, $\beta = 0.02$, 而仿真对象模型为:

1) 设置参数 $\gamma = 0.65$, $\beta = 0.021$, 其中: 参数 γ

增大了 35%, 参数 β 增大了 5%, 设定值为原稳定区域设定值 $c_d = (0.120\ 7, 0.880\ 1)^T$. 分别使用标准 DHP 方法和基于简化模型的DHP 方法进行 10 000 次学习, 其控制结果如图 6 所示. 可见, 两种方法经过学习均能实现良好的控制

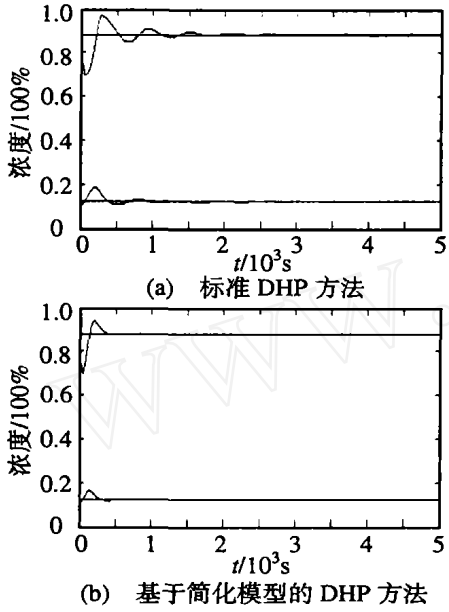

图 6　参数变化 1) 下的控制结果

2) 设置参数 $\gamma = 0.42$, $\beta = 0.019$, 其中: 参数 γ 减小了 12.5%, 参数 β 减小了 5%, 设定值为原稳定区域设定值 $c_d = (0.120\ 7, 0.880\ 1)^T$. 分别使用标准DHP 方法和基于简化模型的DHP 方法进行 10 000 次学习, 其控制结果如图 7 所示. 可见, 本文提出的方法仍能实现良好控制, 而标准DHP 方法在设定值附近出现了振荡

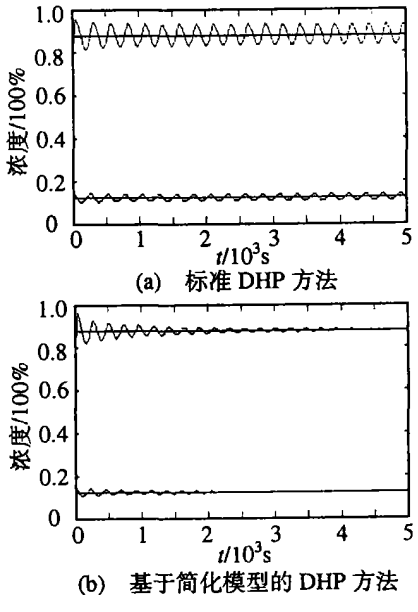

图 7　参数变化 2) 下的控制结果

5　结　　语

生化反应器对象是生化反应过程中的重要装置, 为了保持产品的稳定性和提高生产率, 需要生化反应器保持在某个设定的状态下. 利用本文提出的基于简化模型的DHP学习控制方法对生化反应器分别在稳定区域和不稳定区域进行设定值控制, 仿真结果表明, 该方法能够通过在线训练, 学习到使费用函数最小的控制策略, 达到控制目标

本文提出的基于简化模型的DHP 方法, 相比于标准DHP 方法, 只需要控制量对于状态变化影响的定性信息, 而不需要对象的精确模型. 仿真实验表明, 本文方法的学习过程要快于标准DHP 方法, 在更加困难的不稳定区域设定值控制下表现尤为明显. 并且由于该方法本身对于对象模型的要求更少, 因此能够容忍更大范围的参数变化, 提高了控制策略的鲁棒性

参考文献(References)

[1] Soni A S, Parker R S. Fed-batch Bioreactor Control Using a Multi-scale Model[A]. *Proc of the American Control Conf* [C]. Denver: IEEE press, 2003: 2371-2376.

[2] Anderson C W, Miller W T. Challenging Control Problems[A]. *Neural Networks for Control* [C]. Cambridge: MIT Press, 1991: 475-510.

[3] Si Jennie, Wang Y T. Online Learning Control by Association and Reinforcement[J]. *IEEE Trans on Neural Networks*, 2001, 12(3): 264-276.

[4] Prokhorov D V. *Adaptive Critic Designs and Their Applications* [M]. Lubbuck Texas Tech University, 1997.

[5] Prokhorov D V, Wunsch D. Adaptive Critic Designs [J]. *IEEE Trans on Neural Networks*, 1997, 8(5): 997-1007.

[6] Perkins T J, Barto A G. Lyapunov Design for Safe Reinforcement Learning Control [J]. *J of Machine Learning Research*, 2002, 3(6): 803-832.

[7] Liu D R. Approximate Dynamic Programming for Self-learning Control[J]. *Acta Automatica Sinica*, 2005, 31(1): 13-18.

[8] Wen F, Chen Z H, Wang A Q. An Improvement to Fast-AHC Algorithm [J]. *Information and Control*, 2003, 32(7): 652-656.

[9] Lendaris G G, Shannon T T. Application Considerations for the DHP Methodology[A]. *Proc of the Int Joint Conf on Neural Networks* [C]. Anchorage, IEEE Press, 1998: 1013-1018.