文章编号：1000－8152(2005)02－0291－04

基于强化学习的模型参考自适应控制

郭红霞[1]，吴　捷[1]，王春茹[2]
(1. 华南理工大学 电力学院，广东 广州 510640；2. 广东工业大学 自动化学院，广东 广州 510090)

摘要：提出了一种基于强化学习的模型参考自适应控制方法，控制器采用自适应启发评价算法，它由两部分组成：自适应评价单元及联想搜索单元．由参考模型给出系统的性能指标，利用系统反馈的强化信号在线更新控制器的参数．仿真结果表明：基于强化学习的模型参考自适应控制方法可以实现对一类复杂的非线性系统的稳定控制和鲁棒控制，该控制方法不仅响应速度快，而且具有较高的学习速率，实时性较强．

关键词：强化学习；模型参考自适应控制；联想搜索单元；自适应评价单元
中图分类号：TP273　　**文献标识码**：A

Model reference adaptive control based on reinforcement learning

GUO Hong-xia[1], WU Jie[1], WANG Chun-ru[2]
(1. College of Electrical Engineering, South China University of Technology, Guangzhou Guangdong 510640, China;
2. College of Automation, Guangdong University of Technology, Guangzhou Guangdong 510090, China)

Abstract: Aiming at adaptive control problems of a sort of nonlinear system, model reference adaptive control based on reinforcement learning is proposed. The controller uses adaptive heuristic critic algorithm, which consists of two elements: adaptive critic element, associative search element. The desired performance index is presented by the reference model, and the controller parameters are updated by reinforcement signal given by system. The simulation shows that the proposed method is efficient for a class of complex nonlinear system, and it has a high learning rate, which is important to online learning.

Key words: reinforcement learning; model reference adaptive control; associative search element; adaptive critic elements

1　引言(Introduction)

近来，神经网络由于具有很强的非线性映射能力、并行处理能力和自适应、自学习能力，被广泛用于非线性系统的自适应控制中[1,2]．在这些非线性自适应控制系统中，神经网络大多以监督学习和非监督学习方式，通过精确的训练样本学习隐含在样本中的有关非线性系统本身的内在规律性，以调整网络连接权系数．但在一些复杂的实际应用中，精确的训练样本通常难以获得，或其代价昂贵．强化学习作为一种重要的学习方法，不需要外部环境的数学模型，只是把控制系统的性能指标要求直接转换为一种评价指标，当系统性能指标满足要求时，所施加的控制动作得到奖励，否则得到惩罚．控制器通过奖罚学习，最终获到对系统的最优控制动作[3,4]．

本文将强化学习用于一类非线性系统的自适应控制中，提出了一种基于强化学习的模型参考自适应控制方法．控制器采用自适应启发评价算法，同时由参考模型给出系统的性能指标，利用奖罚信号训练网络参数．最后通过仿真实验，并与一般的模糊神经网络模型参考自适应控制器的仿真结果相比较，验证了该算法的正确性和优越性．

2　基于强化学习的模型参考自适应控制系统(Model reference adaptive control based on reinforcement learning)

基于强化学习的MRAC(Model Reference Adaptive Control)系统的结构如图1所示．

图中，model是参考模型，它的输出 Y_m 作为系统期望的闭环响应，与被控对象的输出 y 相比较，由奖罚机构给出控制效果好坏的评价信号(奖或罚)r，当 y 不满足性能指标时，$r=-1$，否则 $r=0$．控制器采用自适应启发评价(adaptive heuristic critic)算法，它由两部分组成：自适应评价单元ACE(Adaptive Critic Element)及联想搜索单元ASE(Associative Search Element)．

收稿日期：2003－05－08；收修改稿日期：2004－06－29．
基金项目：国家自然科学基金资助项目(60174025)；国家重点基础研究专项经费资助项目(G1998020308)．

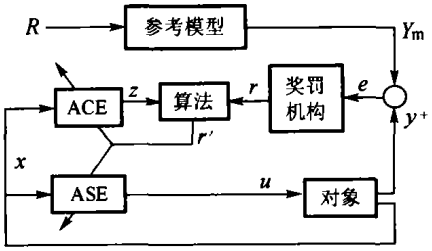

图1 基于强化学习的模型参考自适应控制

Fig. 1 Model reference adaptive control based on reinforcement learning

3 控制器的设计及其算法(Design and arithmetic of controller)

3.1 ACE的学习算法(Learning algorithm of ACE)

ACE单元由3层前馈神经网络构成.其中,输入层共有 $n+1$ 个节点,分别是 $x_0, x_1, \cdots, x_n$. x_0 是偏置量,$x = [x_1, \cdots, x_n]$ 是被控对象的状态矢量.

隐含层共有 m 个结点,每个结点的输入和输出之间的关系是

$$v_j^{(2)}(t+1) = g\Big(\sum_{i=0}^{n} w_{ji}^{(12)}(t)x_i(t+1)\Big), \quad (1)$$

$$g(t) = \frac{1}{1+\exp(-t)}. \quad (2)$$

其中:$v_j^{(2)}(t+1)$ 是 $t+1$ 时刻隐层的输出,$\omega_{ji}^{(12)}(t)$ 是 t 时刻第一层第 i 个结点和第二层第 j 个结点之间的连接权系数,$j = 1,2,\cdots,m$.

输出层只有一个结点,它接受来自第二层的输出信号,输出对评价信号 r 的预报值 z,即

$$z(t+1) = \sum_{i=1}^{m} w_i^{23}(t)v_i(t+1). \quad (3)$$

其中 $\omega_i^{23}(t)$ 是 t 时刻第二层第 i 个结点和输出层之间的连接权系数.

算法单元根据ACE输出的预报评价信号 z 和评价信号 r 来产生内部强化信号 r'. 其算法如下:

$$r'(t+1) = r(t+1) - z(t+1). \quad (4)$$

其中:$r(t+1)$ 是 $t+1$ 时刻的评价信号,它在正常状态(输出满足系统要求的性能指标)时为0,在错误状态(输出不满足系统要求的性能指标)时为 -1. 对于ACE单元来说,$z(t+1)$ 是用来评价系统的运行状况,并产生一个对评价信号 r 的预报.

ACE单元学习的目的是通过调整网络权值使得内部强化信号 r' 最大.因此,$\omega_i^{23}(i = 1,2,\cdots,m)$ 采用以下学习算法:

$$w_i^{23}(t+1) = w_i^{23}(t) + \beta r'(t+1)v(t,t+1). \quad (5)$$

其中 $\beta > 0$ 是学习率.

$\omega_{ji}^{(12)}$ 采用改进的BP算法进行学习:

$$w_{ji}^{(12)}(t+1) = w_{ji}^{(12)}(t) + \alpha r'(t+1)v(t+1) * (1 - v(t+1)\mathrm{sgn}(w_i^{23}(t)x_i(t+1))). \quad (6)$$

其中 $\alpha > 0$ 是学习率.采用 $\mathrm{sgn}(\omega_i^{23}(t))$ 是为了算法更具鲁棒性.

3.2 ASE的学习算法(Learning algorithm of ASE)

ASE采用6层模糊神经网络实现,ASE中包含控制规则集,它反映人的操作经验.采用已有的控制经验可加速学习速度.

在ASE单元中,第1层的输入是被控对象的状态矢量 x.

第2层是模糊化层,每个节点的输出是各模糊集合的隶属度 $A_i^j(x_i)(i = 1,2,\cdots,n, j = 1,2,\cdots,m)$,$A_i^j$ 代表 x_i 的第 j 个模糊集合.它采用三角形隶属函数,有3个参数 (a_j, b_j, c_j),模糊集合的隶属度为

$$A_i^j(x_i) = \begin{cases} 1 - |x_i - c_j|/b_j, & x_i \in [c_j, c_j + b_j], \\ 1 - |x_i - c_j|/a_j, & x_i \in [c_j - a_j, c_j], \\ 0, & \text{其他}. \end{cases} \quad (7)$$

其中:c_j 是三角形的中心坐标,a_j, b_j 分别是 c_j 点左边和右边的宽度.

第3层是规则层.此层输出为第 k 条规则的激活度 α_k.

模糊规则的一般形式为

R_i^k:(if x_1 is A_1^i and x_2 is A_2^i and $\cdots$ and x_n is A_n^i then y is B_k) is t_i^k.

其中:$i = 1,2,\cdots,L$,L 是规则数(即该层的结点数),该层每一个结点代表一条规则;B_k 是 y 的模糊集合;t_i^k 是第 i 条规则当结论为 B_k 时的可信度(truth value),它是[0,1]区间的值;$t_i^k = 1$ 表示该规则绝对正确,$t_i^k = 0$ 表示该规则完全不对.

第 i 条规则的激活度 α_k 定义[6]为

$$\alpha_i = \Big(\sum_{k=1}^{n} A_k^i(x_k)\mathrm{e}^{-A_k^i(x_k)}\Big)/\sum_{k=1}^{n}\mathrm{e}^{-A_k^i(x_k)}. \quad (8)$$

第4层求当 y 的结论为 B_k 时的总激活度 α_k'.该层的节点数等于输出变量 y 的模糊集合数 $m1$.其输出是当结论为 B_k 时的总激活度 α_k':

$$\alpha_k' = \sum_{i=1}^{L}\Big(\frac{\alpha_i}{\sum_{j=1}^{L}\alpha_i} * t_i^k\Big). \quad (9)$$

其中:$k = 1,2,\cdots,ml$,ml 为输出变量 y 的模糊集合数.

第 5 层是解模糊层.其输出是 y 论域中的一个值,即 $B_k^{-1}(\alpha_k')$.

该层的节点数等于输出变量 y 的模糊集合数 ml,其采用三角形隶属函数,有 3 个可调参数(a_k, b_k, c_k),其中 c_k 是三角形的中心位置,a_k 和 b_k 为左右两边的宽度. 由式(7) 可以求出 $B_k^{-1}(\alpha_k')$. 当 $B_k(\alpha_k)$ 的隶属函数是三角形,$B_(^{-1}\alpha_k')$ 一般有两个解,应该取这两个解的平均值,即

$$B_k^{-1}(\alpha_k') = c_k + \frac{1}{2} * (1 - \alpha_k') * (b_k - a_k). \tag{10}$$

当 $B_k(\alpha_k')$ 的隶属函数为梯形时,$B_k^{-1}(\alpha_k')$ 只有唯一解,即

$$B_k^{-1}(\alpha_k') = c_k + a_k * (\alpha_k' - 1). \tag{11}$$

第 6 层是输出层.它输出对系统的控制作用 u:

$$u = \left(\sum_{k=1}^{ml} \alpha_k' B_k^{-1}(\alpha_k')\right) / \sum_{k=1}^{ml} \alpha_k'. \tag{12}$$

ASE 网络的权系统均为 1,其可调参数 ω_i 有:第 4 层的参数 t_i^k、第 5 层的输出隶属度函数参数(a_k, b_k, c_k). 它学习的目的是使 ACE 网络的输出 z 最大,采用梯度法进行学习,则

$$\omega_i(t+1) = \omega_i(t) + \Delta\omega_i, \tag{13}$$

$$\Delta\omega_i = \gamma \partial z / \partial \omega_i = \gamma \partial z / \partial u * \partial u / \partial w_i. \tag{14}$$

其中 $\gamma > 0$ 为学习率.由于 z 和 u 之间没有明确的关系,所以采用近似的方法求出 $\partial z / \partial u$, 即

$$\frac{\partial z}{\partial u} = \operatorname{sgn}\left(\frac{z(t+1) - z(t)}{u(t+1) - u(t)}\right). \tag{15}$$

其中:当 $B_k(\alpha_k)$ 的隶属函数是三角形时,$\partial u / \partial w_i$ 可由式(8) ~ (12)求出,即

$$\frac{\partial u}{\partial a_k} = \frac{\alpha_k'}{2 * \sum_{k=1}^{ml} \alpha_k'} * (\alpha_k' - 1), \tag{16a}$$

$$\frac{\partial u}{\partial b_k} = \frac{\alpha_k'}{2 * \sum_{k=1}^{ml} \alpha_k'} * (1 - \alpha_k'), \tag{16b}$$

$$\frac{\partial u}{\partial c_k} = \frac{\alpha_k'}{\sum_{k=1}^{ml} \alpha_k'}, \tag{16c}$$

$$\frac{\partial u}{\partial t_k^i} = \frac{cc * [B_k^{-1}(\alpha_k') + 0.5 * (a-b) * \alpha_k'] - p}{cc^2} * \bar{\alpha}_i. \tag{16d}$$

其中 cc, p, $\bar{\alpha}_i$ 为

$$cc = \sum_{k=1}^{ml} \alpha_k', \quad \bar{\alpha}_i = \frac{\alpha_i}{\sum_{j=1}^{L} \alpha_j}, \quad p = \sum_{k=1}^{ml} \alpha_k' B_k^{-1}(\alpha_k^{-1}). \tag{17}$$

当 $B_k(\alpha_k)$ 的隶属函数是梯形时可相应的得到 $\partial u / \partial w_i$. 为使参数的学习具有自适应性,式(14)采用内部强化学习信号,即

$$\Delta\omega = \gamma r'(t) \frac{\partial z}{\partial \omega} = \gamma r'(t) \cdot \frac{\partial z}{\partial u} \cdot \frac{\partial u}{\partial \omega}. \tag{18}$$

4 仿真实验与分析(Simulation and Analysis results)

在本文提出的控制方案中,由参考模型 model 的阶跃响应给出系统要求的闭环响应 Y_m. 控制目标是使系统的实际输出 y 与 Y_m 之间的误差 e 在允许误差 ε 的范围之内. 当 e 不在 ε 的范围之内时,对当前的控制行为给予惩罚 $r = -1$;当 e 在允许的误差范围之内时对控制行为给予奖励 $r = 0$. 即

$$e = Y_m - y, \quad r = \begin{cases} 0, & |e| \leqslant \varepsilon, \\ -1, & |e| > \varepsilon. \end{cases} \tag{19}$$

4.1 仿真实验(Simulation experiment)

考虑如下的非线性系统

$$\begin{cases} \dot{x}_1 = a x_2, \\ \dot{x}_2 = 2 x_1 x_2 x_3 + u, \\ \dot{x}_3 = -x_3 + x_3 u, \\ y = x_1. \end{cases} \tag{20}$$

其中系统参数 $a = 1$.

参考模型为

$$Y_m = \frac{1}{s^2 + 1.2s + 1}. \tag{21}$$

允许误差设为 $\varepsilon = \pm 0.02$. ACE 的结构采用 $N^3[4,5,1]$,学习率 $\alpha = \beta = 0.2$. ASE 的结构采用 $N^6[3,11,18,5,5,1]$,学习率 $\gamma = 0.008$;

在 MATLAB 环境下进行仿真.采样步长设为 0.001s.仿真结果如图 2 所示.图中:y 为系统实际输出,Y_m 为参考模型输出,R 为参考模型的阶跃输入,e 为系统实际输出与参考模型输出之间的误差. 在图中曲线 1 是采用本文提出的强化学习控制算法得到的仿真结果,曲线 2 为无强化学习算法,系统采用一般的模糊神经网络模型参考自适应控制器的仿真结果.

考察系统的鲁棒性,系统在 $t = 14$ s 时,参数 a 发生摄动,$a = 2 + \frac{8}{1 + e^{-(t-2)}}$. 仿真结果如图 3 所示.图中曲线 1 是采用本文提出的强化学习算法时得到的结果,曲线 2 是无强化学习算法而采用一般的模糊神经网络模型参考自适应控制器得到的结果.

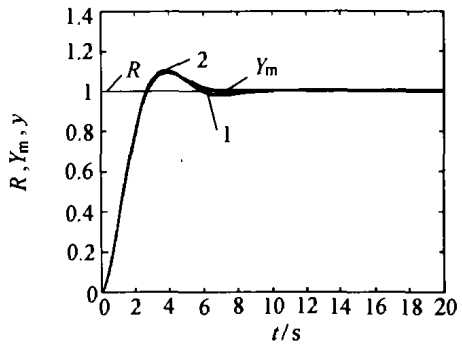

图 2(a) $\varepsilon = \pm 0.02$ 时,系统的输出响应 y 曲线

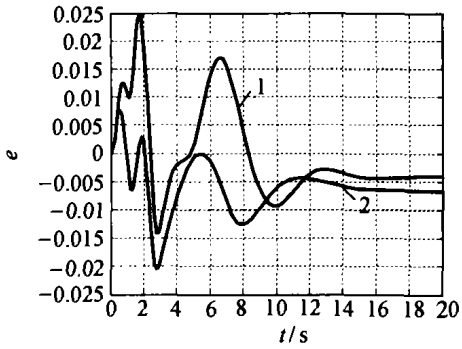

图 2(b) $\varepsilon = \pm 0.02$ 时,系统的误差 e 曲线

图 2 $\varepsilon \pm 0.02$ 时,系统的输出响应 y 及误差 e 曲线

Fig. 2 Resulting output and error response of plant

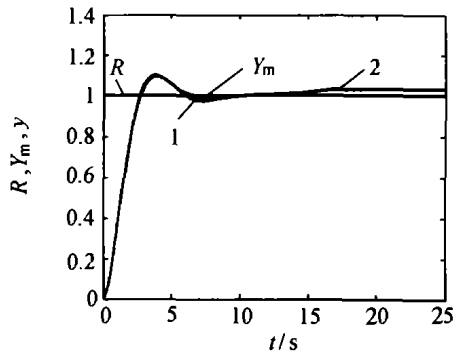

图 3(a) 参数摄动时系统的输出响应 y 曲线

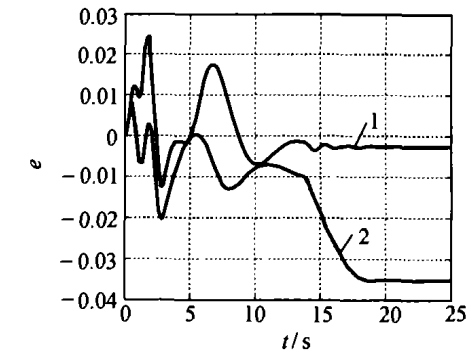

图 3(b) 参数摄动时系统的误差 e 曲线

图 3 参数摄动时系统的输出响应 y 及误差 e 曲线

Fig. 3 Resulting output and error response of plant when exists parameter perturbation

4.2 仿真结果分析(Simulation results analysis)

从上面的仿真结果可以看到:

1) 如图 2 的曲线 1 所示,采用本文提出的基于强化学习的模型参考自适应控制方法,可以使被控对象满足所给定的性能指标,并可实现系统的稳定控制,控制系统具有快速的学习能力和适应能力;

2) 不需要知道被控对象的数学模型,同时也不需要精确的训练样本,在没有足够的知识的条件下通过系统给出的强化信号来逐步改善控制器参数以达到预期目的.这种自学习自适应控制方法为复杂非线性系统的控制提供了新的思路;

3) ASE 结构中的学习率 γ 不能取的过大,否则使得系统的响应速度过慢,选择合适的 γ 可加速响应速度;

4) 由图 3 可知:当系统存在干扰时,本文提出的算法使系统具有较强的抗干扰能力.虽然稳态误差不为零,但是误差在系统的性能指标之内.因此基于强化学习的模型参考自适应控制可以实现对一类非线性系统的鲁棒控制.

5 结论(Conclusion)

本文提出了一种基于强化学习的模型参考自适应控制方法,用模糊神经网络构造控制器,通过系统给出的强化学习信号在线修改控制器的参数,以使控制系统达到预期的控制效果.仿真实验表明:该方法不仅是有效的,而且具有较高的学习速率,实时性较强.它为非线性系统的学习控制提供了新的思路.强化学习作为一种有前景的学习控制方法,目前引起了广泛的研究兴趣.然而强化学习控制系统中还有待于进一步的深入研究,比如怎样提高学习速度、如何确定最优的学习策略等.

参考文献(References):

[1] 叶其革,王晨皓,吴捷.基于自组织模糊神经网络电力系统稳定器的设计[J].控制理论与应用,1999,16(5):688 - 695.
(YE Qige, WANG Chenhao, WU Jie. Design of self-organizing power system stabilizer based on fuzzy neural network [J]. *Control Theory & Applications*, 1999, 16(5): 688 - 695.)

[2] NARENDRA K S, PARTHASARATHY K. Identification and control of dynamic systems using neural networks [J]. *IEEE Trans on Neural Networks*, 1990, 1(1): 4 - 27.

[3] JAGANNATHAN S, LEWIS F L. Multilayer discrete-time neural-net controller with guaranteed performance [J]. *IEEE Trans on Neural Network*, 1996, 7(1): 107 - 130.

[4] ZOMAYA Y. Reinforcement learning for the adaptive control of nonlinear systems [J]. *IEEE Trans on Systems, Man, and Cybernetics*, 1994, 24(2): 357 - 363.

[5] 阎平凡.再励学习——原理、算法及其在智能控制中的应用[J].信息与控制,1996,25(1):28 - 34.

(下转第 300 页)

trol, 1999, 44(9): 1711 – 1713.

[4] TRINH H. Linear functional state observer for time-delay systems [J]. *Int J Control*, 1999, 72(8): 1642 – 1658.

[5] DAROUACH M. Linear functional observers for systems with delays in state variables [J]. *IEEE Trans on Automatic Control*, 2001, 46(3): 491 – 496.

[6] DAROUACH M. Corrections to "linear functional observers for systems with delays in state variables" [J]. *IEEE Trans on Automatic Control*, 2001, 46(10): 1677.

作者简介:

朱淑倩 (1979—),女,山东大学数学与系统科学学院控制论专业博士研究生,研究方向为奇异系统、时滞系统,E-mail: sduzsq@mail.sdu.edu.cn;

冯俊娥 (1971—),女,山东大学数学与系统科学学院副教授,2003年获山东大学博士学位,研究领域包括奇异系统、时滞系统、随机系统,E-mail: thefengs@163.com;

程兆林 (1939—),男,山东大学数学与系统科学学院教授,博士生导师,长期从事多变量控制系统的理论与应用、奇异系统、时滞系统、非线性系统等方面的研究,E-mail: chengzha@jn-public.sd.cninfo.net.

(上接第290页)

4 结论(Conclusion)

1) 本文提出的零相差自适应跟踪控制,没有附加状态补偿器,而是直接利用广义输出误差及其各阶导数的值来设计自适应律,在结构上比较简单,在工程应用中易于实现;

2) 此类系统针对具有未知恒定或缓慢时变参数的系统,具有良好的实时性和自适应性,能够改善它的动态性能,具有广泛的应用价值;

3) 从仿真结果上看到,通过合理地选择自适应律中的各个系数的值,便能达到良好的跟踪效果,具有良好的跟踪性能.

参考文献(References):

[1] 吴士昌,臧瀛芝,方敏.一种使用低阶参考模型的MRACS的设计方法及其在液压伺服系统上的试验[J].控制理论与应用,1988,5(2):111 – 116.
(WU Shichang, ZANG Yingzhi, FANG Min. A method of MRACS using the lower reference model and test on the hydraulic servo-mechanism [J]. *Control Theory & Applications*, 1988, 5(2): 111 – 116.)

[2] TOMIZUKA M. Zero phase error tracking algorithm for digital control [J]. *ASME J of Dynamic Systems, Measurement, and Control*, 1987, 109(1): 65 – 68.

[3] LANDAU I D. *Adaptive Control-the Model Reference Approach* [M]. New York: Marcel Dekker, 1979: 51 – 123.

[4] HANG C C, PARKS P C. Comparative studies of model reference adaptive control systems [J]. *IEEE Trans on Automatic Control*, 1973, 18(5): 419 – 428.

作者简介:

王 茂 (1965—),男,哈尔滨工业大学控制科学与工程系教授,博士生导师,从事自适应控制、变结构控制和惯性技术研究,E-mail: wangmao0451@sina.com;

游文虎 (1972—),男,哈尔滨工业大学控制科学与工程系博士生,讲师,从事最优控制、自适应控制、组合导航和惯性技术研究,E-mail: houhainan@0451.com;

黄丽莲 (1972—),女,哈尔滨工业大学控制科学与工程系博士生,从事混沌控制、非线性控制、自适应控制及惯性技术研究,E-mail: lilian_huang@163.com.

(上接第294页)

(YAN Pingfan. Reinforcement learning – theory, arithmetic and its application in intelligent control [J]. *Information and Control*, 1996, 25(1): 28 – 34.)

[6] 张乃尧,阎平凡.神经网络与模糊控制[M].北京:清华大学出版社,1998:236 – 257.
(ZHANG Naiyao, YAN Pingfan. *Neural network and fuzzy control* [M]. Beijing: Tsinghua University Press, 1998: 236 – 257.)

作者简介:

郭红霞 (1971—),女,博士研究生,研究方向为多Agent系统、强化学习及其在电力系统中的应用,E-mail: ghx9@163.com;

吴 捷 (1937—),男,华南理工大学教授,博士生导师,1961年毕业于哈尔滨工业大学,主要研究方向为电力系统自动化、非线性控制、自适应控制和电力系统自动化;

王春茹 (1968—),女,讲师,博士研究生,研究方向为网络控制.