文章编号: 1001-0920(2006)09-1068-05

一种自适应模糊Actor-Critic学习

王雪松[1], 程玉虎[1], 易建强[2]
(1. 中国矿业大学 信息与电气工程学院, 江苏 徐州 221008; 2. 中国科学院 自动化研究所, 北京 100080)

摘 要: 提出一种基于模糊RBF网络的自适应模糊Actor-Critic学习. 采用一个模糊RBF神经网络同时逼近Actor的动作函数和Critic的值函数, 解决状态空间泛化中易出现的"维数灾"问题. 模糊RBF网络能够根据环境状态和被控对象特性的变化进行网络结构和参数的自适应学习, 使得网络结构更加紧凑, 整个模糊Actor-Critic学习具有泛化性能好、控制结构简单和学习效率高的特点. Mountain Car的仿真结果验证了所提方法的有效性.

关键词: Actor-Critic学习; 模糊推理系统; RBF网络; 泛化
中图分类号: TP18 **文献标识码**: A

A Kind of Adaptive Fuzzy Actor-Critic Learning

WANG Xue-song[1], *CHENG Yu-hu*[1], *YI Jian-qiang*[2]
(1. School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou 221008, China; 2. Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China. Correspondent: WANG Xue-song, E-mail: wangxuesongcumt@163.com)

Abstract: An adaptive fuzzy Actor-Critic reinforcement learning based on a fuzzy radial basis function network is proposed, which can solve the 'curse of dimensionality' problem caused by state space generalization. A fuzzy RBF network is used to approximate both the action function of Actor and the value function of Critic simultaneously. The fuzzy RBF network is able to adjust its structure and parameters in an adaptive way with a self-organizing approach according to the change of environment state and the characteristics of the plant during the learning process, which ensures network size is economical. The proposed fuzzy Actor-Critic learning has advantages of perfect generalization ability, simple control structure, and high learning efficiency. Simulation experiment on Mountain Car control shows the validity of the proposed algorithm.

Key words: Actor-Critic learning; Fuzzy inference system; Radial basis function network; Generalization

1 引 言

目前, 强化学习在理论研究和实际应用两方面均取得了较多成果, 但大量研究结果仍然只针对小规模、离散状态和动作空间问题[1]. 为了实现连续空间下的强化学习, 通常采用对高维空间进行离散化的方法, 但离散化操作必然导致"维数灾"问题, 进而在学习时间和存储空间两方面降低强化学习控制系统的性能[2]. 因此, 要实现对连续状态空间或(和)动作空间的最佳逼近, 强化学习智能体必须具备泛化能力, 其本质是用参数化的函数来逼近"状态-动作"的映射关系[3]. 因此, 如何设计具有高泛化能力和高计算效率的函数逼近器成为强化学习理论和应用研究的一个关键问题.

模糊推理系统(FIS)实际上是一种通用的函数逼近器, 能够利用有限的模糊集合来描述连续的状态空间, 适合表达模糊和不确定性知识, 比较符合人类的思维方式, 但缺乏自学习和自适应能力. 径向基(RBF)网络也是一种函数逼近器, 它具有并行计算、容错性和自学习等优点, 但不适合表达知识, 不能充分利用已有的经验知识. 将RBF网络与FIS相结合形成模糊径向基网络(FRBF), 它比单纯的RBF网络与FIS具有更佳的逼近性能[4]. 为此, 本文在

收稿日期: 2005-06-07; **修回日期**: 2005-08-31.
作者简介: 王雪松(1974—), 女, 安徽泗县人, 副教授, 博士, 从事智能控制、机器学习等研究; 程玉虎(1973—), 男, 安徽淮南人, 讲师, 博士, 从事机器学习、智能机器人等研究.

FRBF 网络的基础上, 提出一种自适应模糊Actor-Critic 强化学习并将其应用于Mountain Car 控制系统仿真以验证算法的有效性

2　基于FRBF 网络的模糊Actor-Critic 学习结构

把FIS 与RBF 网络相结合形成FRBF 网络, 可以使得FRBF 网络内部的连接结构意义上更加清晰, 而且具有自适应学习能力. 本文提出的基于FRBF 网络的Actor-Critic 学习结构如图1 所示

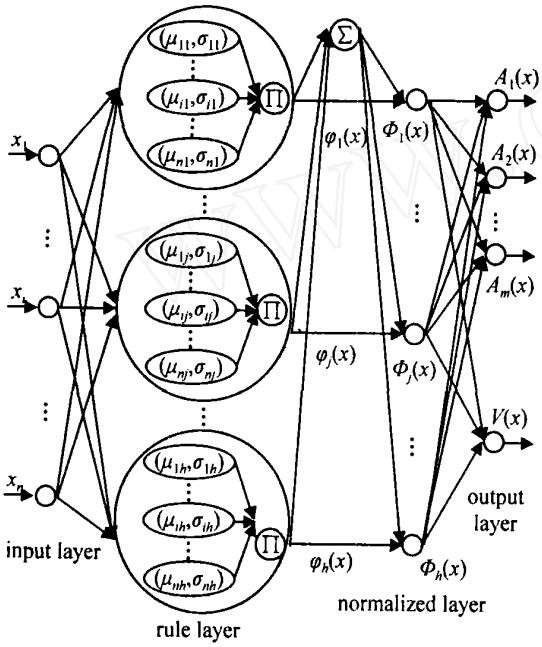


图1　基于FRBF 网络的Actor-Critic 学习结构

第1 层: 输入层. 该层的每个神经元代表一个输入变量 x_i, 这里 x_i 是清晰值, $x = (x_1, x_2, \cdots, x_n)^{\mathrm{T}} \in R^n$ 为直接传递给下一层的输入向量

第 2 层: 规则层. 该层的每个节点表示一条模糊规则的前件部分, 每个节点均具有 n 个高斯型隶属度函数, 有

$$MF_{ij} = \exp\left(-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}\right), \quad j = 1, 2, \cdots, h. \tag{1}$$

式中: MF_{ij} 表示第 j 个规则节点内的第 i 个隶属度函数, μ_{ij} 和 σ_{ij} 分别表示隶属度函数的中心和宽度, n 和 h 分别为输入层和规则层的节点个数

当时间为 t, 输入为 x_t 时, 第 j 个节点的输出 $\varphi_j(x_t)$ 等于其内部 n 个隶属度函数的乘积, 表示第 j 条模糊规则的适应度, 即

$$\varphi_j(x_t) = \prod_{i=1}^{n} MF_{ij}(x_i) = \exp\left(-\sum_{i=1}^{n}\frac{(x_{it} - \mu_{ij})^2}{2\sigma_{ij}^2}\right). \tag{2}$$

式中 x_{it} 为 t 时刻输入向量的第 i 个分量

第 3 层: 归一化适应度计算层. 该层节点数等于规则层的节点数, 该层的作用是统一度量每个规则的适应度, 对规则的适应度进行归一化操作. 虽然归一化操作增加了计算复杂性, 但是对位于两个或多个接收区域的重叠区域中的点具有更好的插值效果. 该层第 j 个神经元的归一化输出 Φ_j 表示第 j 条模糊规则的归一化适应度, 有

$$\Phi_j(x_t) = \frac{\varphi_j}{\sum_{l=1}^{h}\varphi_l} = \frac{\exp\left(-\sum_{i=1}^{n}\frac{(x_{it} - \mu_{ij})^2}{2\sigma_{ij}^2}\right)}{\sum_{l=1}^{h}\exp\left(-\sum_{i=1}^{n}\frac{(x_{it} - \mu_{il})^2}{2\sigma_{il}^2}\right)},$$

$$\forall x, \sum_{j=1}^{h}\Phi_j(x) = 1. \tag{3}$$

第 4 层: 输出层, 由Actor 和Critic 两部分组成. 其中, Critic 部分对状态值函数进行逼近, 将状态映射为期望的评价值 $V(x_t) \in R^1$; 而Actor 部分则产生一个合理的动作, 将状态映射为动作, 实现从感知空间 $x_t = [x_1, x_2, \cdots, x_n] \in R^n$ 到动作空间 $A(x_t) \in R^m$ 的映射, n, m 分别对应输入、输出空间的维数. Actor 网络输出中第 k 个动作的选择函数 $A_k(x_t)$ 和Critic 网络的输出值函数 $V(x_t)$ 分别计算如下:

$$A_k(x_t) = \sum_{j=1}^{h} w_{kj}\Phi_j(x_t), \tag{4}$$

$$V(x_t) = \sum_{j=1}^{h} v_j\Phi_j(x_t). \tag{5}$$

式中 w_{kj} 和 v_j 分别对应于归一化层第 j 个节点到Actor 网络第 k 个输出节点的权值和到Critic 网络输出层的权值

Actor 网络的输出并不直接作用于被控对象, 而是在Actor 网络所推荐的控制动作 A_k 上叠加一个高斯干扰 n_k, 以解决强化学习中的"探索 - 利用"两难问题. 高斯干扰的大小取决于 $V(x_t)$, $V(x_t)$ 大则干扰小, 反之则干扰大. 具体方法如下:

$$A'_k(x_t) = A_k(x_t) + n_k(0, \sigma_V(t)), \tag{6}$$

式中

$$\sigma_V(t) = \frac{1}{1 + \exp(2V(x_t))}.$$

在Actor-Critic 结构中, Critic 与Actor 均采用TD 法来学习值函数和动作概率函数. 在TD 学习算法中, TD 误差 δ_{TD} 的计算由状态转移中两相邻状态值函数的时间差分实现, 即

$$\delta_{\mathrm{TD}} = r_t + \gamma V(x_{t+1}) - V(x_t). \tag{7}$$

式中 $0 < \gamma < 1$ 为折扣系数, 用来确定延迟回报与立即回报的比例. TD 误差实际上反映了所选动作的

优劣程度, 基于 TD 误差分别给出 Actor 和 Critic 网络的权值更新公式如下:

$$v_j(t+1) = v_j(t) + \alpha_c \delta_{TD} \Phi_j(x_t), \tag{8}$$

$$w_{kj}(t+1) = w_{kj}(t) + \alpha_A \delta_{TD} \Phi_j(x_t). \tag{9}$$

式中 α_c 和 α_A 分别为 Critic 和 Actor 的学习率.

3 模糊 Actor-Critic 学习算法

在用神经网络构造模糊推理系统时所要解决的问题是采取何种策略训练神经网络使其结构和参数得到优化, 所以, 本文提出的模糊 Actor-Critic 强化学习的学习过程实际上是 FRBF 网络的学习过程, 包括网络结构学习和参数学习两部分.

3.1 结构学习

通过使 FRBF 网络规则层和归一化层保持最优节点数可以获得更好的泛化能力和更快的收敛速度. 网络结构学习包括增加和合并节点两种操作.

3.1.1 增加节点

(1) TD 误差标准

由 TD 误差的定义可知, 当 Actor-Critic 算法收敛后, TD 误差应当具有随时间趋向于零的特性. 但是, 当状态空间的精度不够时, 即使部分观测空间的值函数学习收敛以后, TD 误差的方差可能仍然很大, 表明这些区域的 TD 误差还存在较大的波动. 这时必须通过增加新的节点以更好地覆盖当前输入, 获得更好的动作输出. TD 误差标准的具体计算是通过定义每个基函数节点的局部误差来进行的, 有如下的定义[5]:

定义 1 FRBF 网络节点的局部误差: 对每个基函数 Φ_j, 移动平均局部加权 TD 误差 f_j 和平方 TD 误差 g_j 分别为

$$f_j(t+1) = (1-\gamma_c\Phi_j) f_j(t) + \gamma_c \Phi_j \varphi_j \delta_{TD}, \tag{10}$$

$$g_j(t+1) = (1-\gamma_c\Phi_j) g_j(t) + \gamma_c \Phi_j \varphi_j \delta_{TD}^2. \tag{11}$$

增加节点的标准为 g_j 与 f_j 的比率 $L_j(t) = g_j(t)/f_j(t)$ 大于一阈值 θ. 但是, 必须考虑当环境为静态时状态转移概率为常数的情况, 此时平均 TD 误差收敛到零会导致 L_j 的发散. 为了避免这种情况, 添加一条停止增加节点的标准: 当平方 TD 误差 g_j 小于一个常量 θ_g 时即停止增加节点的操作. 因此, 根据 TD 误差增加节点的标准为

$$L_j > \theta, \ g_j > \theta_g. \tag{12}$$

(2) if-part 标准

直观上, 一个 FIS 应该对过程的每一状态都能推理出一个合适的控制作用, 这个性质称为"完整性". 模糊控制器的完整性与数据库和规则库有关: 在数据库方面主要指的是基本模糊集赖以定义的支撑集应该以一定程度 ϵ 覆盖有关论域, 这种性质称为 ϵ 完整性[6]. 这样, 使得总有一条规则的适应度值大于 ϵ, 即总有一条主导规则; 在规则库方面, 完整性体现在当模糊条件尚未包含在规则库或者当输入与预定模糊条件之间的匹配度低于 ϵ 时, 可以增加额外的规则, 否则, 在后者的情况下将无主导规则. 因此, 当

$$\varphi = \arg\max_j \varphi(x_t) < \epsilon \tag{13}$$

成立时, 意味着对于当前的输入 x_t, 在 FRBF 网络的规则层没有节点能够很好地覆盖当前输入, 这时就该考虑是否增加新的规则节点以更好地覆盖当前输入, 确保每一个输入变量的隶属度值不小于 ϵ.

对于当前输入 x_t, 如果满足式(12) 所示的 TD 误差标准和式(13) 所示的 if-part 标准, 则按照下式设置新增节点的中心与宽度:

$$\begin{cases} \mu_{new} = x_t, \\ \sigma_{new} = \tau \| x_t - \mu_{nearest} \|. \end{cases} \tag{14}$$

式中: $\| x_t - \mu_{nearest} \|$ 表示当前输入与其最近节点之间的欧氏距离; τ 表示重叠系数, 使相邻节点的基函数之间保持适当的重叠.

在学习初始阶段, 网络规则层无节点, 接收到的第一个输入作为第一个节点, 该节点的中心 μ 取为系统的初始输入变量, 而其宽度 σ 则可根据实际情况(如在输入状态变量的范围内) 随机取值. 对于以后的输入, 根据以上两个标准来衡量是否需要增加规则层节点数.

3.1.2 合并节点

为简化网络结构, 减少多余的规则, 在学习过程中, 对于作用相似的规则节点进行合并. 根据判断规则节点的输入隶属度函数的相似度来确定是否进行节点的合并. 判断的依据是: 若两个高斯型隶属度函数的中心点位置与宽度基本相同, 可认为这两个规则节点具有相似性. 此处采用 k- 近邻算法来计算规则节点的相似性. 假定所有的实例对应于 n 维空间中的点, 规则节点间的近邻关系是根据标准欧氏距离定义的, 定义任意两个规则节点间的距离 $d(\varphi_i, \varphi_p)$ 为

$$d(\varphi_i, \varphi_p) = \left[\sum_{j=1}^{n} \left((\mu_{ij} - \mu_{ip})^2 + (\sigma_{ij} - \sigma_{ip})^2 \right) \right]^{\frac{1}{2}}. \tag{15}$$

如果两个规则节点之间的距离 $d(\varphi_i, \varphi_p)$ 小于事先规定的阈值 d_ϵ, 则进行节点的合并, 相应的规则层和归一化层的节点数减 1. 合并以后的归一化层节点与输出层的连接权值为

$$\begin{cases} v = (v_j + v_p)/2, \\ w_k = w_{kj} + w_{kp}. \end{cases} \tag{16}$$

3.2 参数学习

可调参数包括归一化层与输出层的连接权值、规则层节点的中心与宽度. 本文采用梯度下降法进行网络参数学习. 定义误差函数为

$$E(t) = \delta_{TD}^2/2 \tag{17}$$

规则层节点的中心和宽度更新为

$$\mu_{ij}(t+1) = \mu_{ij}(t) + \eta_\mu \delta_{TD} w_{kj} \Phi_j(x_i)[1 - \Phi_j(x_i)] \frac{x_i - \mu_{ij}}{\sigma_{ij}^2}, \tag{18}$$

$$\sigma_{ij}(t+1) = \sigma_{ij}(t) - \eta_\sigma \frac{\partial E(t)}{\partial \sigma_{ij}} = \sigma_{ij}(t) + \eta_\sigma \delta_{TD} w_{kj} \Phi_j(x_i)[1 - \Phi_j(x_i)] \frac{(x_i - \mu_{ij})^2}{\sigma_{ij}^3}. \tag{19}$$

式中 η_μ 和 η_σ 分别为中心和宽度的学习率.

4 仿真研究

为验证本文所提算法的有效性, 针对一个称为 Mountain Car 或小车爬山的学习控制问题进行仿真研究. Mountain Car 的学习控制在有关强化学习的文献中, 通常作为一个典型的连续状态空间强化学习问题来验证算法的学习效率和泛化性能[7].

小车爬山问题具有两维的连续状态空间, 且除了系统的状态观测值以外, 假定没有任何有关系统动力学模型的先验知识, 因此采用传统的基于模型的最优控制方法仍然难以求解.

图 2 为Mountain Car 学习控制问题的示意图, 图中曲线代表一个山谷的地形, 其中 S 为山谷最低点, G 为右端最高点.

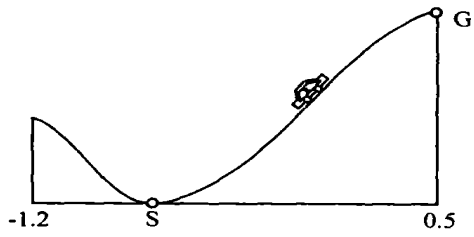


图 2　Mountain Car 示意图

小车的任务是在动力不足的条件下, 从谷底的 S 点以尽量短的时间运动到最高点 G. 小车的控制量 u 具有 3 个离散的取值, 即 + 1, 0, - 1. 系统的状态由两个连续变量 x 和 v 表示, 状态空间满足

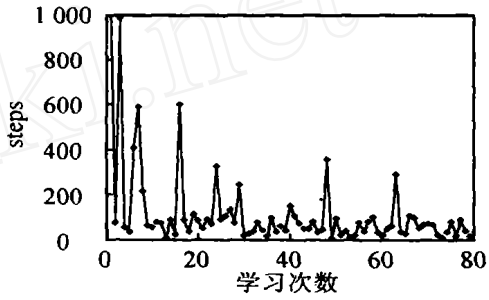
$$\{(x, v) \in R^2 \mid -1.2 \le x \le 0.5, \quad -0.07 \le v \le 0.07\}, \tag{20}$$

式中: x 为小车的水平位移, v 为小车的水平运动速度. 当小车位于 S 点, G 点和左端最高点时, x 的取值分别为 - 0.5, 0.5 和 - 1.2. 由文献[7]有关的动力学方程为

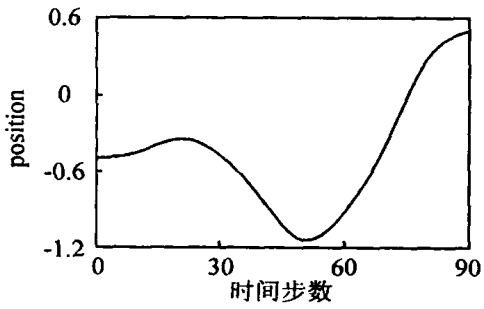
$$\begin{cases} \dot{x} = v, \\ \dot{v} = 0.001u - g\cos 3x, \end{cases} \tag{21}$$

式中: $g = 0.0025$ 为重力, u 为控制量.

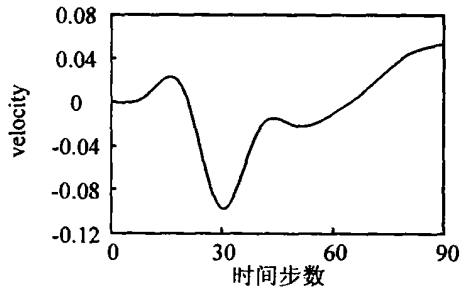
学习控制器的目标是在没有任何模型先验知识的前提下, 实现小车从 S 点运动到 G 点的最短时间控制. 上述学习控制问题可以用一个确定性MDP 来建模, MDP的状态空间由连续变量x 和v构成, 行为



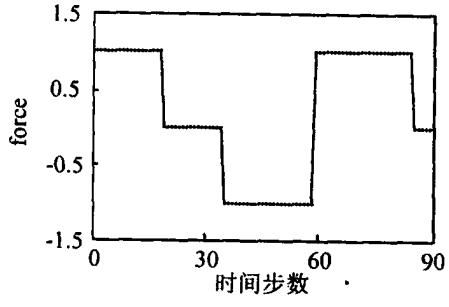
(a)　Mountain Car 的运行学习曲线



(b)　小车位置变化曲线



(c)　小车速度变化曲线



(d)　控制量变化曲线

图 3　Mountain Car 仿真曲线

空间由 3 个离散值 + 1, - 1 和 0 构成. 回报函数设计为

$$r_t = \begin{cases} -1, x < 0.5; \\ 0, x \geq 0.5. \end{cases} \tag{22}$$

每次学习实验小车的初始状态为 $x = -0.5, v = 0$, 当小车到达 G 点或时间步数超过设定值时, 当前学习结束. 学习系统的性能由每次实验中小车从起始点运动到 G 点的时间步数来评价. 仿真中, 网络权值初始化为 $[-1, 1]$ 内的随机值, $\gamma = 0.98$, $\alpha_C = 0.2$, $\alpha_A = 0.3$, $\gamma_c = 0.35$, $\theta_c = 3.0$, $\theta_a = 0.01$, $\epsilon = 0.135\,4$, $\tau = 0.45$, $d_\epsilon = 0.01$, $\eta_c = 0.05$, $\eta_a = 0.03$, 采样周期为 0.02 s.

图 3 给出了仿真曲线. 图 3(a) 为模糊 Actor-Critic 强化学习算法在 Mountain Car 问题中的一次典型运行的学习曲线, 横坐标为学习次数, 纵坐标为每次学习实验中小车从 S 点运动到 G 点所需的时间步数. 从图中可以看出, 学习系统在经过大约 60 次学习后, 小车可以在 100 个时间步内到达目标点, 表明小车已获得了有效的快速爬山控制策略. 图 3(b)～(d) 给出了学习控制器在经过 80 次学习后系统的有关状态变化数据, 其中图 3(b) 为小车的位置曲线, 从图中可以看出, 小车在动力不足的情况下, 必须利用反向运动来积蓄爬山所要求的势能. 由上述曲线可以看出, 模糊 Actor-Critic 学习算法在有收敛性保证的同时, 具有良好的学习效率和泛化性能.

5 结 语

本文充分利用模糊推理的可理解性与 RBF 神经网络的学习能力, 采用 FRBF 神经网络建立的模糊推理系统来实现模糊 Actor-Critic 学习, 解决状态空间泛化中易出现的"维数灾"问题. FRBF 网络具有根据环境状态和被控对象特性的变化进行网络规则层节点动态、自适应地增加和合并的能力, 文中分别给出了节点增加的 TD 误差标准和 if-part 标准, 以及节点合并的模糊相似性测度准则. 另外, 采用梯度下降法对网络规则层节点的中心和宽度进行在线调整. Mountain Car 控制的仿真结果验证了本文算法的有效性.

参考文献(References)

[1] Creighton D C, Nahavandi S. Optimizing Discrete Event Simulation Models Using a Reinforcement Learning Agent[A]. *Proc of Winter Simulation Conf*[C]. San Diego, 2002: 1945-1950.

[2] 李晓萌, 杨煜普, 许晓鸣. 基于递阶强化学习的多智能体AGV 调度系统[J]. ***控制与决策***, 2002, 17(3): 292-296.
(Li X M, Yang Y P, Xu X M. Multiagent AGV Dispatching System Based on Hierarchical Reinforcement Learning [J]. *Control and Decision*, 2002, 17(3): 292-296.)

[3] Ster B. An Integrated Learning Approach to Environment Modeling in Mobile Robot Navigation[J]. *Neurocomputing*, 2004, 57(1-4): 215-238.

[4] 秦斌, 吴敏, 王欣. 模糊神经网络模型混沌混合优化学习算法及应用 [J]. ***控制与决策***, 2005, 20(3): 261-265.
(Qin B, Wu M, Wang X. Hybrid Chaos Optimization Algorithm for Fuzzy Neural Network Model and Its Applications[J]. *Control and Decision*, 2005, 20(3): 261-265.)

[5] Samejima K, Omori T. Adaptive Internal State Space Construction Method for Reinforcement Learning of a Real-world Agent[J]. *Neural Networks*, 1999, 12(7): 1143-1155.

[6] Meesad P, Yen G G. Accuracy, Comprehensibility and Completeness Evaluation of a Fuzzy Expert System [J]. *Int J of Uncertainty, Fuzziness and Knowledge-based Systems*, 2003, 11(4): 445-466.

[7] Lee Y A, Chung T C. A Function Approximation Method for *Q*-learning of Reinforcement Learning[J]. *J of KISS: Software and Applications*, 2004, 31(11): 1431-1438.

(上接第 1067 页)

[8] Byrnes C I, Hu X. The Zero Dynamics Algorithm for General Nonlinear Systems and Its Application in Exact Output Tracking [J]. *J of Mathmatic Systems, Estimation and Control*, 1993, 3(1): 51-72.

[9] 佘焱, 张嗣瀛. 一般非线性系统的相关阶与标准型[J]. ***自动化学报***, 1998, 24(4): 570-572.
(She Y, Zhang S Y. Relative Degree and Normal Form for General Nonlinear Systems [J]. *Acta Automatica Sinica*, 1998, 24(4): 570-572.)

[10] 佘焱, 张嗣瀛. 一般非线性控制系统的局部反馈渐近镇定[J]. ***控制与决策***, 1998, 13(6): 624-628.
(She Y, Zhang S Y. Local Feedback Asymptotic Stabilization for General Nonlinear Systems [J]. *Control and Decision*, 1998, 13(6): 624-628.)

[11] Hu X. Some Results in Nonlinear Output Regulation and Feedback Stabilization [J]. *Automatica*, 1994, 59(4): 1085-1093.