文章编号: 1001-0920(2001)01-0007-05

马尔可夫决策过程自适应决策的进展

李江洪, 韩正之
(上海交通大学 智能工程研究所, 上海 200030)

摘 要: 在介绍一般马尔可夫决策过程的基础上, 分析了当前主要马尔可夫过程自适应决策方法的基本思想、具体算法实现以及相应结论, 总结了现有马尔可夫过程自适应决策算法的特点, 并指出了需要进一步解决的问题。

关键词: 马尔可夫过程; 部分可观马尔可夫过程; 自适应决策
中图分类号: TP 217.2 文献标识码: A

New Achievements in Adaptive Markov Decision Process

LI Jiang-hong, HAN Zheng-zhi
(Research Institute of Intelligent Engineering, Shanghai Jiaotong University, Shanghai 200030, China)

Abstract: Based on an introduction of MDP, main results of algorithms, principles, implementation and conclusions for adaptive MDP are analyzed in detail. The characters of these algorithms are summarized. Problems needing further discussing for adaptive MDP are also pointed out.

Key words: Markov decision process (MDP); partial observable MDP (POMDP); adaptive decision

1 引 言

马尔可夫决策过程(Markov Decision Process, MDP)是一种应用广泛的随机决策过程。自 Wald 开创 MDP 研究后, MDP 受到人们的普遍重视, 国内也有不少学者投入到 MDP 的研究, 并取得了一定成果[1~5]。目前, MDP 不仅在理论研究上成果显著[6], 而且在社会生产实践中也得到广泛应用[7~10]。

本文首先介绍一般马尔可夫决策过程, 然后较详细地阐述了当前主要马尔可夫过程自适应决策方法的基本思想、具体算法实现以及相应结论, 最后总结了马尔可夫过程自适应算法的特点, 并指出了进一步的研究方向。

2 MDP 问题的提法

MDP 模型一般指 5 元组

$$\langle X, U, \{U(x)\,|\,x; x \in X\}, \{p(i,j,u)\,|\,i,j \in X, u \in U\}, \{c(x,u)\,|\,x \in X, u \in U\}\rangle \quad (1)$$

其中, X 为状态空间, $x \in X$ 称为状态; U 为控制空间, $u \in U$ 称为控制; $U(x)$ 是状态为 x 时的可行控制集; $\{p(i,j,u)\,|\,i,j \in X, u \in U\}$ 中 $p(i,j,u)$, 表示在控制 u 下从状态 i 转移到 j 的概率; $\{c(x,u)\,|\,x \in X, u \in U\}$ 中 $c(x,u)$, 表示在状态为 x 时由于采用控制 u 而产生的立即损失。

用 $t \in R^+$ 表示时间, t 时的决策规则 π 是一概率分配函数, 它决定可行控制集 $U(x_t)$ 中各个控制取

收稿日期: 1999-08-02; 修回日期: 1999-11-15
基金项目: 国家自然科学基金项目(69874025)
作者简介: 李江洪(1970—), 男, 湖南长沙人, 博士生, 从事随机决策和智能控制的研究; 韩正之(1947—), 男, 浙江慈溪人, 教授, 博士生导师, 从事控制理论与应用等研究。

为实际控制 u_t 的概率。决策规则列 $\pi = \{\pi_t\}$ 称为策略, MDP 中的策略常为 Markov 策略。MDP 常见的决策目标函数有总损失、无界平均损失以及无界折扣损失等[11]。

部分可观 MDP(POMDP) 是 MDP 的一种特殊情形, 其特殊性在于核过程的Markov过程状态不可观测, 而只能观测到与核过程存在一定概率关系的观测过程状态。在POMDP 中, 常用信息状态表示状态空间中各状态为核过程实际状态的概率, 并且信息状态本身是一种 Markov 过程。当 MDP 目标函数以信息状态为状态时, 便可得到概率 POMDP 的目标函数。详细分析参见文献[12]。

3 MDP 自适应决策的研究内容与一般方法

在 MDP 中, 状态转移概率直接影响目标函数, 但却很难准确地获得状态转移概率。MPD 自适应决策就是研究在状态转移概率不确定的情形下, 根据系统的当前信息进行决策, 优化目标函数。

在控制 $u \in U$ 作用下, 记 Markov 过程从状态 i 转移到 j 的概率为 $p(i,j;u,\theta)$, 其中参数 $\theta \in \Theta$, Θ 称为参数空间。设真实参数为 θ^0, 由于 θ^0 未知, 所以实际系统的状态转移概率也未知。MDP 的目标函数通常为样本路径平均损失, 即

$$J(\pi, x) = \lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} c(x_t, u_t) \qquad (2)$$

MDP 自适应决策就是在 θ^0 未知时, 逐步求解策略 Ψ, 使 $J(\Psi, x)$ 等于或接近 θ^0 已知时目标函数的最优值, 即实现决策自优[13]。

MDP 自适应决策算法可分为间接决策法和直接决策法两大类。间接决策法是在参数估计的基础上, 根据肯定等价原理进行决策; 直接决策法则不经过参数估计而直接进行决策。下面分别加以介绍。

3.1 间接自适应决策算法

间接自适应决策算法一般根据肯定等价原理[13], 在估计 θ^0 的基础上对 Markov 过程进行决策。该决策算法的基本形式如下:

1) 确定参数估计周期 m 及参数估计器, 任意选取参数初始值 $\theta_0 \in \Theta$;

2) 在时刻 $t = km (k = 0, 1, \cdots)$, 由参数估计器得到未知参数估计值 θ, 视 θ 为未知参数的真实值, 计算以 θ_t 为参数的 MDP 最优策略 g_θ;

3) 当 $km \le t < (k+1)m$ 时, 根据状态 x_t, 由策略 g_{θ_t} 从可行控制集 $U(x_t)$ 中选择控制 u_t, 对 Markov 过程进行决策。

各种具体决策算法之间的差别在于参数估计器的不同。通常称采用极大似然参数估计器

$$\theta_{km} = \arg\max_{\theta\in\Theta} \prod_{t=0}^{km} \frac{p(x_t, x_{t+1}; u_t, \theta)}{p(x_t, x_{t+1}; u_t, \theta^0)} \qquad (3)$$

估计参数的算法为基本自适应决策算法。

假定 MDP 满足辨识条件, 文献[14] 证明采用基本自适应决策算法, 不仅估计参数 $\{\theta_k\}$ 能收敛到 θ^0, 而且能够实现决策自优。然而当辨识条件不满足时, 文献[15] 证明此时不能实现决策自优, 所能得到的最好结果是对任意 $i, j \in X$, $p(i,j;g_{\theta^*}(i), \theta^*) = p(i,j;g_{\theta^*}(i), \theta^0)$, 其中 $\theta^* \in \Theta$ 为 $\{\theta_k\}$ 的收敛值。由于辨识条件不成立时基本自适应决策算法无法实现决策自优, 因此, 自适应决策算法都是研究在辨识条件不成立时如何实现决策自优。下面介绍其中的一些主要算法。

3.1.1 随机化方法

在 MDP 自适应决策过程中, 对所有 $x \in X$ 和 $u \in U$, 事件 $(x_t = x, u_t = u)$ 经常发生是获取系统未知信息的必要条件。但在基本自适应决策算法中, 由于 $u_t = g_{\theta_t}(x_t)$ 而排除了其它控制, 如果 $\lim_{t\to\infty}\theta_t = \theta^*$, 那么只有 $(x_t, g_{\theta_t^*}(x_t))$ 经常发生, 从而 $\{\theta_k\}$ 不一定收敛到 θ^0。随机化方法的基本思想就是改变控制的这种排它性, 即实际控制不一定取 $u_t = g_{\theta_t}(x_t)$, 而是从 $u_t = g_{\theta_t}(x_t)$ 的邻域中随机选取某一容许控制, 这样其它控制也有机会成为实际控制。

估计参数随机化法是一种随机化 MDP 自适应决策算法, 特点是不取极大似然参数估计值作为实际估计值, 而是从使似然函数几乎极大化的所有参数中随机选取某参数作为估计值。具体算法如下:

1) 给定 $\epsilon > 0$, 定义似然函数 $L_k(\theta) = \prod_{t=0}^{k-1} p(x_t, x_{t+1}; u_t, \theta)$, $k = 1, 2, \cdots$;

2) 定义 $\Gamma_0(\epsilon) = \Theta$, $\Gamma_k(\epsilon) = \{\theta \in \Theta \mid L_k(\theta) > \max_{\theta\in\Theta} L_k(\theta) - k\epsilon\}$, $k = 1, 2, \cdots$, 估计参数随机化策略 $\Psi = (\mu_0, \mu_1, \cdots)$, 其中 μ_k 是在 $\Gamma_k(\epsilon)$ 上的均匀概率测度;

3) 在时刻 $k = 1, 2, \cdots$, 通过以 μ_k 为概率测度的随机试验, 从 $\Gamma_k(\epsilon)$ 中选取参数估计值 θ_k, 计算以 θ_k 为参数的 MDP 最优策略 g_{θ_k}, 根据 g_{θ_k} 对 Markov 过程进行决策。

文献[16] 证明参数随机化方法下的估计参数列 $\{\theta_k\}$ 将收敛到 θ^0; [17] 研究了 MDP 分别在估计参

数随机化和控制随机化下的自适应决策, 得到$\{\theta_k\}$将收敛到θ^0并能以任意精度实现决策自优的结论。

3.1.2 强制选择法

强制选择法要求MDP的状态空间、控制空间和参数空间均有限, 并且所有状态转移概率大于0。基本思想是保证事件$(x_t = i, u_t = u)$无穷尽地经常发生, 从而满足$\{p(i,j;u,\theta^0)\}$可辨识的必要条件。算法的特点是当时间为某些稀疏的给定时刻时, 强制地从控制空间中循环选取实际控制, 而在其它时刻仍采用基本自适应决策算法。由于强制选择时刻的稀疏性, 强制选择的控制不但不会影响目标函数值, 而且可使估计参数列收敛到真实参数[13]。

3.1.3 有偏极大似然法

前已指出, 当辨识条件不成立时, 采用基本自适应决策算法能得到的最好结果是$p(i,j;g_{\theta^*}(i),\theta^*) = p(i,j;g_{\theta^*}(i),\theta^0)$。记以$\theta$为参数的MDP在策略$\pi$下的目标函数值为$J(\pi,\theta)$, 定义$J^*(\theta) = \min_{\pi} J(\pi,\theta)$, 则由"最好结果"可进一步得到

$$J^*(\theta^*) = J(g_{\theta^*};\theta^*) = J(g_{\theta^*};\theta^0) \geq J^*(\theta^0) \tag{4}$$

显然, 如能设计参数估计器, 使得参数估计值偏向于满足$J^*(\theta) \leq J^*(\theta^0)$的$\theta$, 则必有$J^*(\theta^*) = J(g_{\theta^*};\theta^*) = J(g_{\theta^*};\theta^0) \leq J^*(\theta^0)$, 并且保证参数估计器渐近于极大似然估计器, 使式(4)成立。此时便可得到$J^*(\theta^*) = J(g_{\theta^*};\theta^*) = J(g_{\theta^*};\theta^0) = J^*(\theta^0)$, 从而实现决策自优[18]。实际上这就是有偏极大似然自适应决策的基本思想。

如何使参数估计器在偏向"最好"参数的同时, 又渐近地保持一般极大似然估计器的特性, 是设计有偏极大似然估计器的关键。有偏极大似然估计器的构造如下

$$\theta_k = \arg\max_{\theta\in\Theta}\left[\frac{f(J^*(\theta^0))}{f(J^*(\theta))}\right]^{o(k)} \Lambda_k(\theta) \tag{5}$$

其中

$$\Lambda_k(\theta) = \prod_{t=0}^{k-1}\frac{p(x_t,x_{t+1};u_t,\theta)}{p(x_t,x_{t+1};w,\theta^0)}$$

为一般似然函数, f 和 o 均为实值函数。要求f 严格单调增加且对任意$\theta\in\Theta, f(J^*(\theta)) > 0$; 要求$o(k) > 0, \lim_{k\to\infty} o(k) = \infty, \lim_{k\to\infty} o(k)/k = 0$。可以证明, 即使$\{\theta_k\}$不收敛到$\theta^0$, 有偏极大似然自适应决策仍能实现决策自优[19]。

文献[19～22]给出了适用于各种不同MDP的有偏极大似然自适应决策算法。

3.2 直接自适应决策算法

直接MDP自适应决策算法是近年出现的一种新的决策方法, 其特点是将学习法引入Markov决策。与间接决策算法相比, 直接决策法既不需要MDP的解析模型, 也不需要计算最优策略, 它适合于状态转移概率以及报酬函数结构未知的MDP。

3.2.1 ***Q*-学习法**

当MDP状态空间X和控制空间U均有限且目标函数为无界折扣报酬时, 对每个状态$x \in X$和控制$u \in U$, 定义状态-控制价值

$$Q^*(x,u) = \sum_{y\in X} p(x,y,u)c(x,u) + \alpha\sum_{y\in X} p(x,y,u)V^*(y) \tag{6}$$

其中, $c(x,u)$是立即报酬, $p(x,y,u)$是控制为u时Markov过程状态从x转移到y的概率, $V^*(y)$是以y为初始状态的MDP目标函数最优值。显然, $V^*(x) = \max_{u\in U} Q^*(x,u)$, 如果$Q^*(x,u)$已知, 则很容易计算出最优策略。但在转移概率和报酬函数未知时, 将无法计算$Q^*(x,u)$。

Q-学习法的基本思想是估计$Q^*(x,u)$, 并利用Sutton预测误差$c(x,u) + \alpha\hat{V}(y) - \hat{Q}^*(x,u)$更新$\hat{Q}^*(x,u)$。其中, $\hat{Q}^*(x,u)$是$Q^*(x,u)$的估计值, y是控制u下自状态x转移后的状态, $\hat{V}(y) = \max_{u\in U}\hat{Q}(x,u)$。$\hat{Q}^*(x,u)$的更新算法如下

$$\hat{Q}^*_{t+1}(x,u) = \hat{Q}^*_t(x,u) + \beta[c(x,u) + \alpha\hat{V}(y) - \hat{Q}^*_t(x,u)] \tag{7}$$

其中$\beta\in(0,1]$为学习速率。当β以适当速率收敛到0, 并且$\hat{Q}^*(x,u)$无穷地经常更新时, $\hat{Q}^*(x,u)$将收敛到$\hat{Q}^*(x,u)$, $\hat{V}(x)$将收敛到$V^*(x)$, 因此最终将获得最优策略。

在Q-学习法的基础上, 文献[23～25]分别提出了各自的MDP自适应决策算法。

3.2.2 分散学习决策法

分散学习自适应决策算法的基本思想是运用多自动机对策原理实现MDP决策。在该算法中, 决策由多个自动机和一个协调机来实现, 每个自动机根据其控制集上各控制的分配概率负责对Markov过程的一种状态决策, 并且自动机在决策之后不能立刻知道系统的响应。协调机的作用是根据Markov过程的状态通知负责该状态的自动机, 并同时向自动机传送当前总报酬、运行总时间等信息。

自动机在决策的同时, 根据这些信息并按L_{R-I}学习规则调整其控制集上的概率分布。

设自动机第k次决策时采用的控制为$\alpha(k) =$

a_i, 第 $k+1$ 次收到的 MDP 报酬累计量和时间累计量分别为 $\eta(k+1)$ 和 $\rho(k+1)$; 设自动机第 l 次决策时, 控制 $a_h(l)$ 的被选概率为 $p_h(l)$。自动机根据 L_{R-I} 学习规则, 调整其控制集上概率分布的算法为

$$p_i(k+1) = p_i(k) + \alpha\frac{\rho(k+1)}{\eta(k+1)}[1 - p_i(k)] \quad (8)$$

$$p_j(k+1) = p_j(k) + \alpha\frac{\rho(k+1)}{\eta(k+1)}p_j(k), \quad j \neq i \quad (9)$$

当 Markov 过程满足在任意确定性平稳策略下状态是遍历的条件时, 则分散学习决策的目标函数将收敛到最优值[26]。

4 POMDP 自适应决策

POMDP 自适应决策研究在状态转移矩阵受未知参数的影响下如何实现决策自优。现有的 POMDP 自适应决策算法都是基于肯定等价原理的间接决策法。文献[18] 介绍了POMDP自适应决策的一般方法。各种算法之间的差别是参数估计器。

4.1 有偏极大似然法

在 POMDP 中, 有一类是核过程状态在一给定的常返区域内完全可观, 而在状态空间其它区域内不可观。有偏极大似然法可用于该类 POMDP 自适应决策, 其基本思想与 MDP 有偏极大似然法相同。

记 POMDP 的状态空间为 X, 控制空间为 U, 参数空间为紧集 Θ, 核过程状态转移概率为 $p(x, y, u; \theta)$, 其中 $x, y \in X$, $u \in U$, $\theta \in \Theta$; 完全可观常返区域为 Γ。设 λ_k^θ 为 k 时刻以 θ 为参数 POMDP 的信息状态, 决策目标是使无界平均损失最小。设 θ^0 为真实参数, ϵ 为决策精度, 有偏极大似然决策算法如下[27]:

1) 离散化参数空间: 根据决策精度 ϵ, 将参数空间 Θ 划分为 q 个互不相交的子集 $\{\Theta_1, \Theta_2, \cdots, \Theta_q\}$, 并在子集 $\Theta_j (j = 1, 2, \cdots, q)$ 中确定满足一定条件的参数 θ^j, 计算以 θ^j 为参数的 POMDP 最优策略 π_{θ^j};

2) 估计参数: 在参数估计时刻 k, 采用类似于式 (1) 的有偏极大似然估计器得到估计参数 $\hat{\theta}_k$;

3) 决策: 设 $\hat{\theta}_k \in \Theta_d, d \in \{1, 2, \cdots, q\}$, 则由 Θ_d 可得参数 θ^d 及策略 π_{θ^d}, 视 θ^d 为 POMDP 的真实参数, 计算信息状态 $\lambda_k^{\theta^d}$, 根据 $\lambda_k^{\theta^d}$ 由 π_{θ^d} 确定控制 u_k 对 MDP 决策。

可以证明, 有偏极大似然决策算法下的目标函数值与目标函数最优值之间的差值最多为 ϵ。文献[27, 28] 分析了具体的有偏极大似然决策算法。

4.2 递推预测误差法

递推预测误差法是估计参数的一种方法。文献[29] 将它用于一类设备维护 POMDP 自适应决策, 对影响核过程状态转移概率的未知参数 $\theta \in [0, 1]$ 进行估计。具体算法如下

$$R_{n+1} = R_n + \frac{1}{n+1}(\varphi_n^2 - R_n), \quad R_1 = 1 \quad (10)$$

$$\varphi_n = -\frac{\partial}{\partial\theta_n}\epsilon_n(\theta_n) = 2q - 1 \quad (11)$$

$$\theta_{n+1} = \Pi_\Theta\left(\theta_n + \frac{1}{n+1}R_{n+1}^{-1}\varphi_n\epsilon_n(\theta_n)\right), \quad \theta_0 \in \Theta \quad (12)$$

其中 $\epsilon_n(\theta)$ 是 n 时刻的观测过程状态预测误差平方和。在估计参数 θ_n 的基础上, 根据肯定等价原理进行决策。在该算法下估计参数序列 $\{\theta_n\}$ 将依概率收敛到真实参数, 并能实现决策自优。

5 评论与展望

现有的 MDP 自适应决策算法都能在辨识条件不成立的条件下实现最优决策。直接决策算法的特点是将学习法引入 MDP 决策, 不需要 MDP 的解析模型和计算最优策略, 适合于状态转移概率以及报酬函数结构未知的 MDP; 其缺点是要求 MDP 的状态和控制均有限。

间接决策算法的基础是肯定等价原理。尽管各间接决策算法互不相同, 但它们都是对基本自适应决策算法的改进, 并具有以下共同特点; 1) 要求转移概率大于0, 且当 Θ 为连续空间时, 转移概率是 θ 的连续函数; 2) 要求真实参数 $\theta^0 \in \Theta$, 且 Θ 为紧集; 3) 目标函数为无界的平均或折扣损失。要求间接决策算法具有这些特点是因为: 1) 保证马尔可夫过程不可约, 从而平稳最优策略存在; 2) 保证估计参数列的极限点存在。由于估计参数列只能渐近地收敛到极限点, 因此只能研究无界目标函数。对于 POMDP 自适应决策算法亦可得到类似的结论。

通过对现有 MDP 自适应决策算法的分析可以发现, 间接决策算法对 Θ 为连续空间时的自适应决策研究尚不完善, 这主要是因为在间接自适应决策过程中需要计算相对于估计参数的最优策略, 而在连续空间中具有无穷多个点, 从而使计算复杂化, 影响了决策效率。如何解决这一问题是今后需要进一步研究的课题。此外, 无论间接决策法还是直接决策法, 目前只研究了无界规划水平下的 MDP 自适应决

策。由于实际系统的规划水平常常是有界的,将这些算法应用于社会生产实践存在一定的困难,因此,对有界规划水平 MDP 自适应的研究具有重要的现实意义,也是今后 MDP 自适应决策研究的方向。

参考文献:

[1] 董泽清,宋京生.无界报酬半马氏折扣模型的初等方法[J].科学通报,1987,32(11):809-812.

[2] 宋京生,董泽清.连续时间总报酬马氏决策规划[J].科学通报,1987,32(16):1201-1205.

[3] 伍从斌,张继红.报酬无界的连续时间折扣马氏决策规划[J].应用概率统计,1997,13(1):1-10.

[4] 胡奇英.状态部分可观察的无界报酬马氏决策规划[J].数理统计与应用概率,1998,13(3):251-258.

[5] 郭先平.一般 MDP 策略的唯一性[J].应用概率统计,1998,14(3):258-265.

[6] Arapostathis A, Bokar V S, Fernandez G E *et al*. Discrete-time controlled Markov processes with average cost criterion: A survey[J]. SIAM J Contr Opt, 1993, 31(2):282-344.

[7] Shin K G, Krishna C M, Lee Y H. Optimal dynamic control of resources in a distributed system[J]. IEEE Trans on Software Eng, 1989, 15(10):1188-1198.

[8] Bournas R M, Beather F J, Teneketzis D. Optimal flow control allocation policies in communication networks with multiple message classes[J]. IEEE Trans on Autom Contr, 1993, 38(3):390-403.

[9] Lam Y. An optimal maintenance model using a number of different actions[J]. Microelectronics and Reliability, 1997, 37(4):549-712.

[10] Wallace J H, Yar lin K. An optimal structured policy for maintenance of partially observable aircraft engine components[J]. Naval Research Logistics, 1998, 45(4):335-352.

[11] Hernandez L O, Lasserre J B. Discrete-time Markov control processes[M]. New York: Springer-Verlag, 1996. 1-11.

[12] Monahan G. A survey of partially observable Markov decision processes: Theory, models and algorithms [J]. Management Science, 1982, 28(1):1-16.

[13] Kumar P R, Varaiya P. Stochastic systems: Estimation, identification and adaptive control[M]. New Jersey: Prentice-Hall, 1986.

[14] Mandl P. Estimation and control in Markov chains [J]. Adv Appl Prob, 1974, 16(1):40-60.

[15] Borkar V, Varaiya P. Adaptive control of Markov chains-I: Finite parameter set[J]. IEEE Trans on Autom Contr, 1979, 24(6):953-958.

[16] Doshi B, Shreve S E. Strong consistency of a modified maximum likelihood estimator for controlled Markov chains[J]. J Appl Prob, 1980, 17(3):726-734.

[17] Borkar V, Varaiya P. Identification and adaptive control of Markov chains[J]. SIAM J Contr and Opt, 1982, 20(4):470-489.

[18] Hernandez L O, Marcus S J. Adaptive control of Markov processes with incomplete state information and unknown parameters[J]. J of Optim Theory and Appl, 1987, 52(2):227-241.

[19] Kumar P R, Becker A. A new family of optimal adaptive controllers for Markov chains[J]. IEEE Trans on Autom Contr, 1982, 27(1):137-146.

[20] Kumar P R, Lin W. Optimal adaptive controllers for Markov chains[J]. IEEE Trans on Autom Contr, 1982, 27(4):765-774.

[21] Kumar P R. Simultaneous identification and adaptive control of unknown systems over finite parameter sets [J]. IEEE Trans on Autom Contr, 1983, 28(1):68-76

[22] Stettner L. On nearly optimizing strategies for a discrete-time uniformly ergordic adaptive model[J]. Appl Math Optim, 1993, 27(2):161-177.

[23] Watkins C J, Dayan P Q. Learning[J]. Machine Learning, 1992, 8(2):279-292.

[24] Santharam G. Sastry P S. A reinforcement learning neural network for adaptive control of Markov chains [J]. IEEE Trans on Syst, Man and Cybern, 1997, 27(5):588-600.

[25] Sridhar Mahadevan. Average reward reinforcement learning: Foundation, algorithms and empirical results[J]. Machine Learning, 1996, 22(1-3):159-195.

[26] Wheeler R M, Narendra K S. Decentralized learning in finite Markov chains[J]. IEEE Trans on Autom Contr, 1986, 31(6):519-526.

[27] Duncan T E, Pasik D B, Stettner L. Adaptive control of a partially observed discrete time Markov process [J]. Appl Math Optim, 1998, 37(3):269-293.

[28] Duncan T E, Pasik D B, Stettner L. On the ergodic and adaptive control of stochastic differential delay systems[J]. J of Optim Theory and Appl, 1994, 81(3):509-531.

[29] Fernandez G E, Arapostathis A, Marcus S I. Analysis of an adaptive control scheme for a partially observed controlled Markov chain[J]. IEEE Trans on Autom Contr, 1993, 38(6):987-993.