基于马氏决策过程模型的动态系统学习控制: 研究前沿与展望

徐 昕[1] 沈 栋[2,3] 高岩青[3,4] 王 凯[3,5]

摘 要 基于马氏决策过程 (Markov decision process, MDP) 的动态系统学习控制是近年来一个涉及机器学习、控制理论和运筹学等多个学科的交叉研究方向, 其主要目标是实现系统在模型复杂或者不确定等条件下基于数据驱动的多阶段优化控制. 本文对基于 MDP 的动态系统学习控制理论、算法与应用的发展前沿进行综述, 重点讨论增强学习 (Reinforcement learning, RL) 与近似动态规划 (Approximate dynamic programming, ADP) 理论与方法的研究进展, 其中包括时域差值学习理论、求解连续状态与行为空间 MDP 的值函数逼近方法、直接策略搜索与近似策略迭代、自适应评价设计算法等, 最后对相关研究领域的应用及发展趋势进行分析和探讨.

关键词 学习控制, Markov 决策过程, 增强学习, 近似动态规划, 机器学习, 自适应控制

引用格式 徐昕, 沈栋, 高岩青, 王凯. 基于马氏决策过程模型的动态系统学习控制: 研究前沿与展望. 自动化学报, 2012, **38**(5): 673−687

DOI 10.3724/SP.J.1004.2012.00673

Learning Control of Dynamical Systems Based on Markov Decision Processes: Research Frontiers and Outlooks

XU Xin[1] SHEN Dong[2,3] GAO Yan-Qing[3,4] WANG Kai[3,5]

Abstract Learning control of dynamical systems based on Markov decision processes (MDPs) is an interdisciplinary research area of machine learning, control theory, and operations research. The main objective in this research area is to realize data-driven multi-stage optimal control for complex or uncertain dynamical systems. This paper presents a comprehensive survey on the theory, algorithms, and applications of MDP-based learning control of dynamical systems. Emphases are put on recent advances in the theory and methods of reinforcement learning (RL) and adaptive/approximate dynamic programming (ADP), including temporal-difference learning theory, value function approximation for continuous state and action spaces, direct policy search, approximate policy iteration, and adaptive critic designs. Applications and the trends for future research and developments in related fields are also discussed.

Key words Learning control, Markov decision processes (MDP), reinforcement learning (RL), approximate dynamic programming (ADP), machine learning, adaptive control

Citation Xu Xin, Shen Dong, Gao Yan-Qing, Wang Kai. Learning control of dynamical systems based on Markov decision processes: research frontiers and outlooks. *Acta Automatica Sinica*, 2012, **38**(5): 673−687

随着控制理论的深入发展, 人们越来越希望控制方法能够处理更为复杂的动态系统, 特别是含有很强非线性的未知系统. 这正是控制理论现在所面临的巨大挑战和机遇. 事实上, 现实系统充满了各种更为复杂的因素需要考虑, 如系统的未知动态、系统动力学的高度非线性、变化的环境影响、系统模型的不确定性等. 数学模型是对现实系统在一定程度上的近似, 但现实系统中可能存在许多未知动态是建模时没有考虑到的, 甚至可能是很难或者无法建模

收稿日期 2011-08-01 录用日期 2011-12-13
Manuscript received August 1, 2011; accepted December 13, 2011
国家自然科学基金 (61075072, 90820302, 60921061), 霍英东青年教师基金优选资助课题 (114005), 教育部新世纪优秀人才支持计划 (NCET-10-0901) 资助
Supported by National Natural Science Foundation of China (61075072, 90820302, 60921061), Fork Ying Tong Education Foundation (114005), and Program for New Century Excellent Talents in University (NCET-10-0901)
本文责任编委 刘德荣
Recommended by Associate Editor LIU De-Rong

1. 国防科学技术大学机电工程与自动化学院自动化研究所 长沙 410073 2. 中国科学院自动化研究所 北京 100190 3. 复杂系统智能管理与控制国家重点实验室 北京 100190 4. 美国亚利桑那大学系统与工业工程学院 图森市 85721-0020 美国 5. 国防科学技术大学军事计算实验与平行系统技术研究中心 长沙 410073
1. Institute of Automation, College of Mechatronic Engineering and Automation, National University of Defense Technology, Changsha 410073, P. R. China 2. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, P. R. China 3. State Key Laboratory of Management and Control for Complex Systems, Beijing 100190, P. R. China 4. Department of Systems and Industrial Engineering, University of Arizona, Tucson AZ85721-0020, USA 5. Center for Military Computational Experiments and Parallel Systems Technology, National University of Defense Technology, Changsha 410073, P. R. China

的, 如交通系统、社会经济系统等. 而对这些复杂系统的控制将会突破传统控制理论的范围, 需要我们发展能够对被控系统具有 "学习" 能力的控制新方法和新理论.

学习控制系统是靠自身的学习功能来认识控制对象和外界环境的特性, 并相应地改变自身特性以改善控制性能的系统. 这种系统具有一定的识别、判断、记忆和自行调整的能力[1−2]. 随着被控对象复杂性和不确定性的日益增加, 学习控制的理论与应用逐渐得到学术界的关注. 1971 年著名学者 Fu 从发展学习控制的角度正式提出智能控制这个新兴的学科领域[3]. Saridis 从控制理论发展的观点, 论述了从通常的反馈控制到最优控制、随机控制, 再到自适应控制、自学习控制、自组织控制的发展路线[4]. 20 世纪 80 年代发展起来的迭代学习控制 (Iterative learning control, ILC) 适用于可以重复执行同一跟踪目标的不确定被控系统, 通过对被控系统进行控制尝试, 以输出信号与给定目标的偏差修正不理想的控制信号, 使得系统的跟踪性能得以提高[5]. 迭代学习控制的研究对于具有高精度轨迹跟踪控制要求的动力学系统有着重要的意义, 但与本文所述学习控制在本质上有很大区别.

随着神经网络与模糊系统理论与技术的发展, 基于神经网络与模糊逻辑的自适应控制系统[6] 得到了广泛的研究, 这一类系统通常利用神经网络与模糊系统的学习与逼近能力来实现未知非线性的补偿, 满足一定条件下的系统性能要求.

马氏决策过程 (Markov decision process, MDP) 是动态随机系统建模与优化控制的重要理论框架, 在动态规划理论方面已经取得了大量的研究成果[7−8]. 实际应用中的许多动态随机系统能够建模为 MDP, 并且 MDP 的最优控制策略不仅能够满足系统稳定性的要求, 也满足最优性能指标要求, 因此基于 MDP 的动态系统学习控制对于模型复杂或者不确定的动态系统优化控制具有重要的理论意义和应用价值. 表 1 对三种主要的学习控制方法进行了简单比较. 与迭代学习控制和基于神经网络 (Neural networks, NN) 的自适应控制不同, 基于 MDP 的学习控制能够实现不确定动态随机系统在大规模状态空间的近似最优控制.

目前, 基于 MDP 的动态系统学习控制已经成为一个涉及机器学习、运筹学和控制理论等多个学科的交叉研究领域, 其中增强学习 (Reinforcement learning, RL) 理论与方法成为该领域的主要研究内容之一[9−12]. 增强学习又称为强化学习或者再励学习, 在人工智能的早期研究中一度成为研究热点之一. 到 20 世纪 80 年代末, 增强学习的研究进一步得到重视, 并出现了与运筹学、控制理论、机器人学等交叉综合的特点. 特别是近十年来, 通过与运筹学的马氏决策过程与动态规划理论广泛交叉, 增强学习的理论、算法和应用取得了若干重要的研究成果, 成为近年来机器学习和智能控制领域的前沿和热点[13−15].

早期的增强学习理论与算法研究主要针对离散状态与行为空间 MDP 的学习控制问题, 近年来, 有关近似动态规划 (Approximate dynamic programming, ADP) 的研究成为增强学习、运筹学和优化控制理论等相关领域共同关注的热点之一[10−14, 16]. 所谓近似动态规划, 是指为克服传统基于模型的离散表格式动态规划方法在求解大规模空间 MDP 时面临的 "维数灾难" (Curse of dimensionality) 和 MDP 模型信息不确定时面临的 "模型灾难" (Curse of modeling), 而采用各种函数逼近算法对 MDP 的最优值函数与策略进行近似估计的理论与方法. 实际早在上世纪 90 年代, Wang 与 Saridis 就已经开始研究随机非线性的逼近次优解, 给出了求解广义 HJB 方程的程序[17−18], 完整详细的证明过程可参见文献 [19]. 在 RL 与 ADP 的算法和理论研究中, 通常不要求已知有关 MDP 转移概率和回报函数的先验模型信息, 而更强调在与环境交互中实现数据驱动的学习, 因此可以看作是自适应动态规划方法[20]. 近年来随着相关研究的不断深入, 增强学习中对 MDP 模型信息的估计和利用也得到更多的关注, 如优先遍历 Q-学习算法[9]、自适应评价设计 (Adaptive critic design, ACD) 增强学习算法[21] 等.

作为 MDP 学习控制的主要方法, RL 与 ADP 在理论与算法方面仍然面临许多困难和挑战, 如高维

表 1 常用学习控制方法的比较

Table 1 Comparison of common learning control methods

	运动轨迹	动态系统模型	优化目标
迭代学习控制	重复	多数考虑未知确定性系统	跟踪误差 e
基于神经网络的自适应控制	可不重复	部分已知	保证系统稳定性或一定的鲁棒性
基于 MDP 的学习控制	可不重复	模型未知或部分已知的 MDP, 可为随机动态系统	$J_d = \mathrm{E}\left[\sum_{t=0}^{\infty}\gamma^t r_t\right], J_a = \lim_{N\to\infty}\sup\frac{1}{N}\mathrm{E}\left[\sum_{t=0}^{N-1} r_t\right]$

连续空间的近似最优策略逼近、学习控制系统的稳定性、复杂系统的结构化分层控制、回报函数的设计等. 为进一步推动相关研究, 本文对基于 MDP 的动态系统学习控制理论、算法与应用进行综述, 重点针对连续状态与行为空间 MDP 的学习控制算法, 包括平稳控制策略值函数估计的时域差值 (Temporal-difference, TD) 学习理论与算法、基于值函数逼近的学习控制算法、策略梯度与近似策略迭代算法、以及自适应评价设计算法等, 最后对 RL 与 ADP 的应用及发展趋势进行分析和探讨.

本文其余部分内容安排如下: 第 1 节介绍了基于机器学习的控制系统设计与分析的基本思想; 第 2 节给出动态学习控制的 MDP 模型; 进而第 3 节至第 6 节分别重点阐述四类基于 MDP 的学习预测或控制算法, 即平稳控制策略值函数估计的 TD 学习理论与算法、基于值函数逼近的学习控制算法、基于直接策略搜索与近似策略迭代的学习控制算法以及具有自适应评价设计结构的学习控制算法; 第 7 节与第 8 节分别讨论 RL 及 ADP 的应用和研究方向展望; 最后第 9 节为本文结论.

1 基于机器学习的控制系统设计与分析

从 20 世纪 50 年代起, 随着计算机的飞速发展, 现代控制理论发展起来. 一系列控制思想和控制方法, 如鲁棒控制、自适应控制、随机控制、最优控制等, 都得到了广泛的关注并取得重要进展. 一般而言, 传统的控制系统都是根据物理系统给出的数学模型, 而很多控制方法都是基于某个具体的数学模型来构造控制器, 并分析其稳定性、收敛性、最优性等各种评价指标. 因此从某种程度上而言, 为了实现良好的控制性能, 数学模型需要 “足够精确”, 以保证其能反映实际系统的主要特性, 从而将控制器应用到实际系统时仍能达到预先的性能要求. 但另一方面, 复杂的动态系统数学模型往往带来控制器设计的困难, 并且实际系统运行过程中的各种不确定性也难以准确描述, 这些都对动态系统的建模提出了挑战.

在实际应用中, 因为数学模型往往无法精确地描述实际系统, 所以有必要深入研究对系统结构信息有较少依赖特别是无依赖的控制器设计理论和方法. 1977 年著名控制专家 Saridis 出版了关于随机系统自组织控制的专著[22−23], 对自组织控制的思想进行了系统的阐述. 自组织控制主要针对动态特性完全或部分未知的复杂不确定对象, 强调控制器结构与参数的自组织调整, 其结构与参数的自组织特性也可以看作是一种学习控制的能力. 近年来, 随着控制对象的规模和复杂性不断增加, 基于数据驱动 (Data-driven) 的控制器设计与分析理论开始得到学术界日益广泛的关注. 数据驱动型的控制器设计是指仅依赖于可以获得的系统输入/输出或状态数据以及其他可以获得的观测数据来实现控制器的设计与优化任务. 迭代学习控制就是一种典型的数据驱动型控制方法. 迭代学习控制由 Arimoto 等在 20 世纪 80 年代提出[24], 适用于运行过程能够在固定时间长度内运行完毕并不断重复的系统. 迭代学习控制的显著优势是其控制律设计仅需要跟踪目标及输入/输出信号, 对系统信息要求很少, 但算法简单而有效, 是典型的数据驱动型控制器. 目前, 迭代学习控制在理论和应用方面都得到了深入研究, 发展了各种分析方法[5, 22−29]. 对于线性确定系统, 相关的结果已经发展的比较完善并且有大量应用. 而对于一般的非线性系统, 目前的结果多需要非线性特性满足全局 Lipschitz 条件, 部分文章放宽了这一条件[27−29], 在这方面还需要更深入的研究. 另外, 针对含有随机噪声系统的迭代学习控制, 目前的研究结果还很少, 其中 Saab 和 Chen 等针对离散时间随机系统给出了一些开拓性的结果[28−31], 针对连续时间系统就作者所知还无相应的结果. 需要指出, 国内学者 Chen 等的工作[29, 31] 针对随机系统首次得到了以概率 1 收敛到最优控制的迭代学习更新律. 重复学习控制 (Repetitive learning control, RLC) 是另一类带有学习特性的控制方法, 与迭代学习控制的不同之处在于, 它所适用的系统并不要求在有限时间内运行完毕, 即系统运行时间可以到无穷, 但要求跟踪信号为周期信号[26, 32−35].

作为人工智能的重要分支之一, 机器学习的主要目标是使计算机程序或者系统能够根据经验或者观测数据不断改善自身的性能. 学术界把已提出的机器学习方法按照与环境交互的特点分为监督学习、无监督学习和增强学习三类. 其中监督学习方法是目前研究得较为广泛的一种, 该方法要求给出学习系统在各种环境输入信号下的期望输出 (即教师信号), 典型的监督学习方法包括神经网络的反向传播算法、决策树学习算法、支持向量机算法等. 无监督学习方法主要包括各种自组织学习方法如聚类学习、自组织神经网络学习 (SOM、ART-1、ART-3) 等. 无监督学习系统的输入仅有环境的状态信息, 也不存在与环境的交互.

监督学习方法能够根据经验数据对系统和环境进行建模和预测, 这种特性引起了控制界的兴趣并将其应用于解决各种辨识和控制问题. 近二十年来, 机器学习在系统辨识和控制中的应用得到深入研究, 其中有代表性的是神经网络在系统辨识和控制方面的研究和应用. Werbos 结合神经网络和控制理论, 阐述了将人工神经网络 (Artificial neural networks) 应用到控制和辨识中的重要性及研究方

法[36]. Antsaklis 等讨论了控制技术面临的挑战与神经网络方法的适用性[37]. 文献 [38] 从模型的角度入手, 阐明了神经网络用于系统辨识和控制的基本概念及理论问题, 分析了静态与动态反向传播方法, 深刻揭示了神经网络在系统辨识和控制方面的有效性. 此外, Liu[39] 讨论了如何应用神经网络方法来处理非线性辨识和非线性控制问题. 文献 [40] 对一般非线性系统, 基于递归神经网络进行辨识, 并保证估计的稳定性. 近年来, 随着统计学习理论的发展, 支持向量机在系统辨识与自适应控制系统中的应用也得到了广泛关注. 基于统计学习的结构风险极小化原理, 支持向量机在有限数据样本条件下能够获得更好的推广或泛化性能, 并且避免了神经网络梯度学习算法存在的局部极值问题, 因此能够更好的用于控制系统的模型辨识和非线性特性的逼近. 文献 [41] 基于最小二乘支持向量机 (Least squares support vector machines) 对 Hammerstein 模型进行辨识, 同时考虑了 SISO 与 MIMO 情形, 对非线性函数要求放宽且对输入的要求也较弱, 但没有给出明确的收敛性分析. 此外, 近年来应用 SVM 进行辨识的研究工作还包括文献 [42−45]. 除监督学习外, 无监督学习在控制系统的模型辨识与状态特征抽取等也得到了若干研究和应用[46].

与监督学习和无监督学习不同, 增强学习基于学习心理学的有关原理, 强调在与环境的交互中学习, 学习过程中仅要求获得评价性的反馈信号 (称为回报或增强信号, Reward/Reinforcement signal), 以极大化或极小化未来的回报为学习目标. 增强学习由于不需要给定各种状态下的教师信号, 因此对于求解复杂的优化决策问题具有广泛的应用前景. 增强学习的最初思想来自于心理学的 "试误法" 学习理论, 即利用外界刺激实现对行为选择的强化. 但目前增强学习方法的内涵已远远超出了早期心理学和神经科学研究的简单 "试误法" 学习机制, 而主要强调以不确定条件下序贯决策的优化为目标, 以 MDP 为系统优化控制的描述模型, 成为基于 MDP 的复杂系统自适应优化控制的一类重要方法[47−48].

2 动态系统学习控制的 MDP 模型

作为描述动态随机系统优化决策问题的一类基本数学模型, MDP 模型[8] 通常用四元组 $\{S, A, P, R\}$ 表示, 其中 S 为状态空间, A 为行为空间, P 为状态转移概率, R 为回报函数, 状态转移概率满足马氏性即无后效性. MDP 的优化控制目标函数为 J. 定义行为策略为从状态集合到行为选择概率的映射, 即

$$\pi:\ S \to P(a) \tag{1}$$

在运筹学与动态规划的研究中, 通常假设 MDP 状态转移方程为离散时间方程的形式, 并且状态空间与行为空间为有限或可数的离散空间. 对于上述 MDP, 当状态转移为确定性函数时, 可以描述为如下的离散时间动态系统的最优控制问题:

$$x_{k+1} = f(x_k, u_k) \tag{2}$$

设时刻 t 的回报为 r_t, 则 MDP 优化控制的性能指标可以具有折扣总回报或者平均期望回报两种形式:

1) MDP 折扣总回报目标

$$J_d = \mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right],\ 0 < \gamma < 1 \tag{3}$$

2) MDP 平均期望回报目标

$$J_a = \lim_{N\to\infty} \sup \frac{1}{N} \mathrm{E}\left[\sum_{t=0}^{N-1} r_t\right] \tag{4}$$

设 π 为平稳策略, 则 Markov 决策过程的状态值函数定义为

$$V^{\pi}(x) = \mathrm{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t | x_0 = x\right] \tag{5}$$

其中, 数学期望 $\mathrm{E}_{\pi}[\cdot]$ 定义在状态转移概率 P 和平稳策略 π 的分布上.

MDP 的最优策略 π^* 定义为极大化或者极小化目标函数的平稳行为策略. 根据动态规划理论, 上述 MDP 最优策略 π^* 的值函数满足如下的 Bellman 方程:

$$V^{\pi^*}(s_t) = \max_{a_t} \mathrm{E}\left[r(s_t, a_t) + \gamma V^{\pi^*}(s_{t+1})\right] \tag{6}$$

在 RL 与 ADP 的研究中, MDP 的状态空间与行为空间也可以是连续空间, 并且状态转移时间也可以是连续的. 对于连续时间 MDP, 通常可以为描述如下的连续时间系统最优控制问题:

$$\dot{x} = f(x, u) \tag{7}$$

对于连续时间 MDP, 平稳策略 π 的值函数定义如下:

$$V^{\pi}(x) = \int_{\tau=t}^{\infty} r(x(\tau), u(\tau))\mathrm{d}\tau \tag{8}$$

对于连续时间 MDP 的最优控制问题, 为有效设计 RL 与 ADP 算法, 可以采用如下的区间 (Interval) Bellman 方程[14]:

$$V^{\pi}(x) = \int_{\tau=t}^{t+T} r(x(\tau), u(\tau))\mathrm{d}\tau + V^{\pi}(x(t+T)) \tag{9}$$

为实现模型不确定条件下的优化控制, 基于 MDP 的学习控制器设计目标是利用 MDP 的观测数据 (包括状态转移、控制输入、回报函数等) 实现对最优值函数与最优策略的估计或者逼近. 因此, MDP 的值函数估计问题, 特别是连续状态与行为空间条件下的值函数逼近问题, 成为 RL 与 ADP 的核心理论问题之一.

3　平稳控制策略值函数估计的 TD 学习理论

在增强学习与近似动态规划中, 时域差值学习[49−51] 占有重要地位, 是自适应评价器等学习控制方法对 MDP 值函数进行估计和逼近的关键算法, 并且在人类大脑的学习功能与信号研究中已经发现了 TD 模型描述了人类对复杂时序事件进行行为适应的高阶学习过程 (有关成果发表在 2004 年的 *Nature* 杂志上[50]).

TD 学习理论的初步建立以 Sutton 首次给出时域差值学习的形式化描述和 TD(λ) 学习算法为标志[49], 已取得了许多研究成果, 并成为其他增强学习算法如 Q 学习算法[52]、Sarsa 学习算法的基础[53].

针对连续空间的 MDP, 采用各种值函数逼近方法的 TD 学习算法得到了广泛研究. 采用神经网络的非线性 TD 逼近方法往往存在收敛性问题[54−55], 并且提高泛化性能需要进行大量的结构和参数选择与优化. 线性 TD(λ) 算法[56] 采用如下的线性逼近器对 MDP 平稳策略的值函数进行逼近:

$$\widetilde{V}_t(x_t) = \boldsymbol{\phi}^{\mathrm{T}}(x_t)\boldsymbol{\theta}_t \tag{10}$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{\alpha}_t\delta_t\boldsymbol{z}_{t+1} \tag{11}$$

$$\delta_t = r_t + \gamma\widetilde{V}_t(x_{t+1}) - \widetilde{V}_t(x_t) \tag{12}$$

其中, $\boldsymbol{z}_{t+1} = \gamma\lambda\boldsymbol{z}_t + \boldsymbol{\phi}(x_t)$ 为适合度轨迹 (Eligibility traces) 向量, x_t 为 Markov 链在时刻 t 的状态, $\boldsymbol{\phi}(x) = [\phi_1(x), \phi_2(x), \cdots, \phi_n(x)]^{\mathrm{T}}$ 为状态的基函数向量, $\boldsymbol{\alpha}_t$ 为学习步长, $\boldsymbol{\theta}_t$ 为权值向量.

针对 TD(λ) 学习算法在求解平稳策略 MDP 值函数预测时的收敛性, Tsitsiklis 等证明了线性 TD(λ) 算法在概率 1 意义下的收敛性并给出了收敛解的逼近误差上界[56]. 根据文献 [56] 的分析, 线性 TD 算法以概率 1 收敛到如下的不动点方程的解:

$$\mathrm{E}[A(X_t)]\boldsymbol{\theta}^* - \mathrm{E}[b(X_t)] = 0 \tag{13}$$

$$A(X_t) = \boldsymbol{z}_t(\boldsymbol{\phi}^{\mathrm{T}}(x_t) - \gamma\boldsymbol{\phi}^{\mathrm{T}}(x_{t+1})) \tag{14}$$

$$b(X_t) = \boldsymbol{z}_t r_t \tag{15}$$

其中, x_t, x_{t+1} 分别为时刻 t 与 $t+1$ 的 MDP 状态, 并且收敛后的值函数与真实值函数的误差上界为

$$\|\boldsymbol{\Phi}\theta^* - V^*\|_D \le \frac{1-\lambda\gamma}{1-\gamma}\|\Pi V^* - V^*\|_D \tag{16}$$

其中, $\boldsymbol{\Phi}$ 为线性特征基函数对应的向量, $\boldsymbol{\theta}^*$ 为算法收敛后的权值向量, V^* 为真实的值函数, $D = \mathrm{diag}\{\pi_i\}$ (π_i 为状态 x_i 的平稳概率分布), $\|X\|_D = \sqrt{X^{\mathrm{T}}DX}$, $\Pi = \boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}}D\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}D$.

在线性 TD(λ) 算法研究的基础上, 最小二乘不动点 TD 学习 (以下简称最小二乘 TD 学习) 也得到了关注[57−59]. Xu 等[59] 提出了多步递推最小二乘 TD 学习算法 RLS-TD(λ), 该算法同时结合了递推最小二乘 TD 学习方法和适合度轨迹机制, 从而能够获得比已有算法更好的收敛性能, 并且结合递推最小二乘参数辨识理论的有关成果和学习预测的实验研究, 进一步分析了递推方差阵初值对算法暂态性能的影响. RLS-TD(λ) 的主要迭代公式如下:

$$K_{t+1} = \frac{P_t\boldsymbol{z}_t}{\mu + (\boldsymbol{\phi}^{\mathrm{T}}(x_t) - \gamma\boldsymbol{\phi}^{\mathrm{T}}(x_{t+1}))P_t\boldsymbol{z}_t} \tag{17}$$

$$\boldsymbol{\theta}_{t+1} = \theta_t + K_{t+1}(r_t - (\boldsymbol{\phi}^{\mathrm{T}}(x_t) - \gamma\boldsymbol{\phi}^{\mathrm{T}}(x_{t+1}))\theta_t) \tag{18}$$

$$P_{t+1} = \frac{1}{\mu}\big[P_t - P_t\boldsymbol{z}_t[\mu + (\boldsymbol{\phi}^{\mathrm{T}}(x_t) - \gamma\boldsymbol{\phi}^{\mathrm{T}}(x_{t+1})) \times P_t\boldsymbol{z}_t]^{-1}(\boldsymbol{\phi}^{\mathrm{T}}(x_t) - \gamma\boldsymbol{\phi}^{\mathrm{T}}(x_{t+1}))P_t\big] \tag{19}$$

其中, K_t 为递推增益矩阵, P_t 为方差矩阵, $1 \ge \mu > 0$ 为遗忘因子.

由于实际学习控制问题中往往面临具有非线性值函数的 MDP, 线性 TD 学习算法存在逼近能力和特征基函数的局限性. Xu 等研究了 TD 学习中的核方法, 提出了核最小二乘 TD 学习算法 K-LSTD[60], 并且对核 TD 学习中核矩阵的稀疏化方法进行了研究[51]. 在 K-LSTD 中, MDP 的近似值函数采用如下的再生核 Hilbert 空间基函数的组合表达式:

$$\widetilde{V}(x) = \sum_{i=1}^{t}\alpha_i k(x, x_i) \tag{20}$$

其中, $k(\cdot,\cdot)$ 为 Mercer 核函数, α_i 为组合系数, x_i 为观测样本子集的样本点. 为实现核矩阵的稀疏化, 提高泛化能力, KLS-TD 采用如下的近似线性相关 (Approximately linear dependence, ALD) 分析[61] 技术来进行观测样本子集的选择:

$$\delta_t = \min_c \left\| \sum_j c_j\varphi(x_j) - \varphi(x_t) \right\|^2 \le \mu \tag{21}$$

在式 (21) 中, μ 为选择样本阈值, φ 为核函数对应的非线性特征映射. 满足式 (21) 的样本与已有的样本子集具有线性相关性, 将不作为新样本加入样本子集.

有关最小二乘 TD 的改进算法还包括具有 O(n) (n 为线性特征的个数, RLS-TD 的计算复杂性为

$O(n^2)$) 计算复杂性的增量 LS-TD 算法 iLS-TD[62], 以及结合残差梯度与直接梯度的混合最小二乘 TD 学习算法[63] 等.

当采用高维值函数逼近器时, 在线增强学习算法需要进一步降低计算复杂性. 文献 [64] 研究了具有 $O(n)$ 计算复杂性的梯度增强学习算法 GTD2 和 TDC, 采用如下的投影 Bellman 残差性能指标:

$$J = \left\| \widetilde{V} - \Pi T \widetilde{V} \right\|_D^2 \tag{22}$$

其中, $TV = R + \gamma PV$ 为 Bellman 算子, P 为转移概率矩阵, $\Pi = \boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}} D \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^{\mathrm{T}} D$ 为线性基函数对应的投影算子, $\|v\|_D = D^{\mathrm{T}} v D$, d_i 为状态 x_i 的采样概率分布.

基于性能指标 (22), GTD2 算法的迭代公式为

$$\mu_{t+1} = \mu_t + \beta_t(\delta_t - \phi_t^{\mathrm{T}} \mu_t)\phi_t \tag{23}$$

$$\theta_{t+1} = \theta_t + \alpha_t(\phi_t - \gamma\phi_{t+1})\phi_t^{\mathrm{T}} \mu_t \tag{24}$$

其中, α_t, β_t 为学习步长.

在采用线性值函数逼近器的条件下, GTD2 和 TDC 已被证明能够保证 Off-policy (即观测数据与行为策略不一致) 学习预测的收敛性, 而线性 TD 算法只能保证 On-policy (即观测数据与行为策略完全一致) 学习预测的收敛性.

4 基于值函数逼近的 MDP 学习控制算法

TD 学习算法主要针对 MDP 控制策略为平稳策略时的值函数估计问题, 是一种多步学习预测算法[9, 65], 并且可以看作是求解 MDP 最优策略即学习控制问题的子问题. 因此, 在 TD 学习理论研究的基础上, 用于求解 MDP 学习控制问题的增强学习理论与算法也取得了一系列重要进展. 早期的 RL 研究主要针对表格型学习控制算法及其收敛性理论, 如 Q-学习算法、Sarsa(0) 学习算法等[52−53]. 目前, 基于值函数逼近的增强学习与近似动态规划方法取得了许多研究进展, 包括基于多层前馈神经网络 (Multi-layer feed-forward neural networks, MLFNN) 的近似梯度学习算法等[66]. 在基于 MLFNN 的增强学习算法和应用的研究中, 结合 Q-学习与 Sarsa-学习的基本原理, 通常采用一种称为直接梯度 (Direct gradient)[9] 的学习算法. 式 (25) 给出了基于 Sarsa-学习的直接梯度下降算法.

$$\Delta w_k = \alpha_t(r(x_t, a_t) + \gamma Q(x_{t+1}, a_{t+1}) - Q(x_t, a_t))\frac{\partial Q(x_t, a_t)}{\partial w_k} \tag{25}$$

其中, w_k 为神经网络的输出层权值, 对于隐含层的权值可以利用误差反向传播算法学习.

对于采用值函数逼近器的增强学习控制算法, 目前在收敛性分析理论方面还比较缺乏. Heger 研究了值函数逼近误差上界与策略性能误差上界的关系, 指出当值函数逼近误差上界较小时, 获得的近似最优策略具有性能保证, 从而为基于值函数逼近的增强学习泛化方法提供了理论分析基础[67]. Baird 提出的残差梯度算法采用式 (26) 进行神经网络权值的学习, 但仅能保证在平稳学习策略条件下的局部收敛性, 无法实现对马氏决策过程最优值函数的求解[54].

$$\Delta w = -\alpha_t \frac{\partial J_t}{\partial w} = -\alpha_t(r(x_t, a_t) + \gamma Q(x_{t+1}, a_{t+1}) - Q(x_t, a_t)) \times \left(\gamma \frac{\partial Q(x_{t+1}, a_{t+1})}{\partial w} - \frac{\partial Q(x_t, a_t)}{\partial w}\right) \tag{26}$$

文献 [65] 提出的非平稳策略残差学习算法进一步改善了算法对局部最优策略的逼近性能, 但主要针对离散行为空间 MDP, 采用的神经网络等函数逼近器仍然存在结构优化选择等结构风险控制问题. 核方法 (Kernel methods, 或称为核函数方法)[68] 是近年来机器学习领域的研究热点, 其核心思想是通过引入 Mercer 核函数及其对应的非线性映射, 来代替线性学习算法如支持向量机 (Support vector machines, SVMs)、主成分分析 (Principal component analysis, PCA) 中的内积运算, 从而实现在新的非线性特征空间即再生核 Hilbert 空间的高效学习算法, 以提高学习机器的非线性逼近能力和泛化性能. 目前, 核方法已经在以支持向量机为代表的监督学习和以核主成分分析 (Kernel PCA, KPCA) 为代表的无监督学习领域得到了成功的应用, 并且相关理论和算法日益完善, 显示了核方法在提高机器学习算法的非线性逼近与泛化性能方面的优势[69]. 最近, 核方法在增强学习与近似动态规划领域的应用得到学术界日益广泛的关注, 如 2006 国际机器学习会议 (ICML′06) 的核增强学习研讨会 (Kernel RL Workshop) 等[70], 相关成果包括前面提到的 KL-STD 算法、基于核的增强学习[71] 以及后面介绍的核最小二乘策略迭代算法等[72].

5 基于直接策略搜索与近似策略迭代的学习控制算法

与值函数逼近方法不同, 基于直接策略搜索的学习控制方法通过神经网络等函数逼近器直接在 MDP 的策略空间搜索, 但存在如何估计策略梯度 (Policy gradient) 的困难. REINFORCE 算法[73] 只能处理有限的周期性 MDP 问题, Baxter 等在文献 [74] 中提出了 GPOMDP 算法, 把 REINFORCE

算法推广到能处理无限时间域的 MDP 问题, 同时在该方法中结合了适合度轨迹机制, 以提高算法的性能. 文献 [75] 研究了策略梯度学习算法的最优回报基线 (Baseline), 但仍然需要解决策略梯度算法收敛速度缓慢的问题. Schraudolph 等利用增益向量自适应的方法来提高策略梯度算法的收敛速度, 并且在仿真实验中获得了比 GPOMDP 和自然梯度策略梯度方法更快的学习收敛速度[76].

针对策略梯度与值函数的关系, Sutton 等证明了对于任意给定的离散状态与行为空间 MDP, 无论是折扣型回报还是平均型回报, 都有下式成立[77]:

$$\frac{\partial \rho}{\partial \theta} = \sum_x d^{\pi}(x) \sum_a \frac{\partial \pi(x,a)}{\partial \theta} Q(x,a) \tag{27}$$

其中, $\frac{\partial \rho}{\partial \theta}$ 为控制性能对策略参数的导数, $\pi(x,a)$ 为行为选择策略, $d^{\pi}(x)$ 为状态的平稳分布.

策略迭代 (Policy iteration) 算法是动态规划的重要算法, 在 MDP 模型已知时能够收敛到 MDP 的最优策略. 在 RL 与 ADP 的研究中, 近似策略迭代算法得到了广泛的研究. 近似策略迭代算法包括策略评价与策略改进两个主要模块. 在策略评价模块中, 通常采用时域差值学习 TD 算法进行平稳策略值函数的估计. 值函数的估计可以针对状态值函数, 也可以针对行为值函数, 即 $Q^{\pi(t)}$. 直接对行为值函数进行估计的优点是可以很方便地进行策略迭代和优化, 因为在获得了策略 $\pi(t)$ 的行为值函数后, 可以根据下面的策略优化公式选择行为, 从而获得一个新的策略 $\pi(t+1)$:

$$\pi(t+1) = \arg\max_a Q^{\pi(t)}(s,a) \tag{28}$$

优化后的贪心策略 $\pi(t+1)$ 为一个确定性策略, 并且如果行为值函数估计 $Q^{\pi(t)}$ 能够以高精度地逼近策略 $\pi(t)$ 的真实行为值函数, 则新的策略 $\pi(t+1)$ 的性能至少不会比前一次迭代的策略 $\pi(t)$ 差. 这样的迭代过程一直重复下去, 直到连续两次获得的策略 $\pi(t)$ 与 $\pi(t+1)$ 完全相同或者基本相同为止, 此时策略迭代算法收敛到一个最终策略. 如果基于时域差值学习的策略评价能够以高精度逼近每次迭代的行为值函数, 则策略迭代算法能够在很少的迭代次数内收敛到马氏决策过程的最优或近似最优策略.

针对连续状态空间 MDP 的学习控制问题, 文献 [78] 提出了基于线性值函数逼近的最小二乘策略迭代算法 (Least-squares policy iteration, LSPI), 并且证明了如果对于任意第 m 次迭代获得的近似值函数 $\widetilde{Q}_m$ 满足式 (29), 则近似策略迭代算法收敛后获得的近似最优策略与真实最优策略 Q^* 的误差上界可以由式 (30) 给出:

$$\forall\, m, \quad \left\| \widetilde{Q}_m - Q^* \right\|_{\infty} \leq \delta \tag{29}$$

$$\lim_{m\to\infty} \sup \left\| \widetilde{Q}_m - Q^* \right\|_{\infty} \leq \frac{2\gamma\delta}{(1-\gamma)^2} \tag{30}$$

其中, $1 > \gamma > 0$ 为折扣因子, $\delta > 0$ 为值函数逼近误差上界.

LSPI 算法虽然具有较好的近似最优策略收敛性, 但由于采用人工选择的线性基函数, 仍然存在特征选择与非线性空间的泛化问题. 文献 [79] 研究了基于高斯过程 (Gaussian processes) 的策略迭代增强学习算法, 但其计算效率和收敛性还需要进一步深入研究. 文献 [72] 提出了基于核的最小二乘策略迭代增强学习算法 KLSPI, 并且分析了算法的收敛性. KLSPI 采用了基于核的行为值函数逼近算法 KLSTD-Q 作为贪心策略迭代的策略评价算法, 并且结合了基于近似线性相关分析的核稀疏化方法来提高核方法的泛化性能, 因此能够在高精度的非线性值函数逼近的基础上, 实现对最优策略的高效逼近, 实验结果表明 KLSPI 往往能够在基函数自动构造的基础上, 获得优于 LSPI 的学习控制性能. 最近的相关研究成果还包括连续行为空间近似策略迭代算法 CLSPI (Continuous-action LSPI)[80] 和基于 Lapclace 几何框架进行基函数构造的表示策略迭代 RPI (Representation policy iteration, RPI) 算法等[81].

6　具有自适应评价设计结构的学习控制

目前策略梯度增强学习算法存在的主要问题是学习效率不高、收敛速度慢, 因此将策略估计与值函数估计相结合的混合结构学习控制即执行器–评价器增强学习算法[82−84] 成为学术界关注的热点方向. 在执行器–评价器学习控制算法中, 评价器用于对马氏决策过程的策略进行评价, 即利用 TD 学习理论估计与策略对应的值函数; 而执行器则用于根据值函数估计的结果选择和优化策略, 实现对最优策略的搜索. 前面讨论的近似策略迭代方法也可以看作是一类采用贪心策略搜索的执行器–评价器学习控制算法.

近年来得到广泛研究的自适应评价设计 (通常可简称为 ACD) 学习控制理论与方法[85−88] 就是基于评价器–执行器结构的思想, 其基本结构如图 1 所示. 并且 ACD 可以结合基于神经网络的动态系统建模技术, 也称为近似动态规划方法. 因此, 近似动态规划与已有的增强学习理论研究有着不可分割的联系, 并且直接针对实时动态系统优化控制的应用需求. 文献 [59] 在递推最小二乘 TD 学习算法的基础上, 提出了快速自适应评价学习控制算法 FastAHC, 极大地改进了 AHC (Adaptive heuristic

critic) 算法的性能, 并且被成功应用于建筑物能源优化控制等现实世界的学习控制问题中[89].

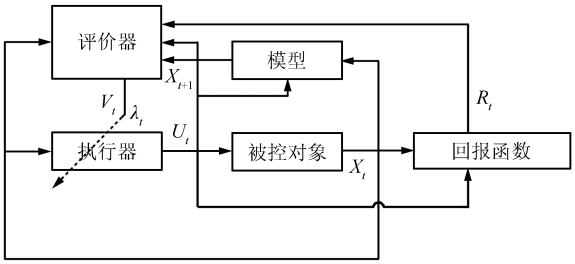

图 1 自适应评价设计的基本计算与反馈结构

Fig. 1 Architecture of ACD

ACD 方法包括启发式动态规划 (Heuristic dynamic programming, HDP)、对偶启发式规划 (Dual heuristic programming, DHP) 和全局对偶启发式规划 (Globalized dual heuristic programming, GDHP) 等[21]. 与传统的增强学习算法相比, ACD 等近似动态规划方法除了利用了增强学习的执行器–评判器学习控制结构外, 还强调利用对 MDP 模型的估计来设计高效的学习控制算法, 部分 ACD 算法如 DHP 还对 MDP 值函数的导数进行估计, 以提高算法的收敛性能. 在 HDP 与 DHP 中, 评价器网络的学习通常采用基于近似梯度的值函数逼近 TD 学习, 其输出层权值迭代公式分别如式 (31) 和式 (32) 所示:

$$W_{t+1} = W_t + \alpha_t(R_t + \gamma \widehat{V}(x_{t+1}) - \widehat{V}(x_t))\frac{\partial f_1}{\partial W} \tag{31}$$

$$W_{t+1} = W_t + \alpha_t\left(\frac{\partial R_t}{\partial x_t} + \gamma\frac{\partial \widehat{V}(x_{t+1})}{\partial x_t} - \frac{\partial \widehat{V}(x_t)}{\partial x_t}\right)\frac{\partial f_2}{\partial W} \tag{32}$$

其中, f_1, f_2 分别为 HDP 和 DHP 评价器神经网络的非线性映射函数, α_t 为学习因子.

在 DHP 的执行器网络中, 根据评价器输出的值函数导数估计信息进行行为策略参数的迭代:

$$\Delta \boldsymbol{w}_u = \frac{\partial \widehat{V}(x(t+1))}{\partial u(t)}\frac{\partial u(t)}{\partial \boldsymbol{w}_u} = \frac{\partial \widehat{V}(x(t+1))}{\partial x(t+1)}\frac{x(t+1)}{u(t)}\frac{\partial u(t)}{\partial \boldsymbol{w}_u} \tag{33}$$

上述采用神经网络的 ACD 近似动态规划方法仍然存在网络结构的设计与学习参数的优化问题, 学习过程中容易陷入局部极小值, 并且算法的收敛性理论仍然有待完善. 近年来的相关研究进展包括: 文献 [90] 提出了线性离散时间系统 H_∞ 控制的 Q- 学习算法, 通过估计行为值函数来逼近相应代数 Riccati 方程的解, 从而不需要已知系统动力学模型信息. 文献 [91] 考虑了离散时间线性二次零和博弈的最优策略问题, 设计了 ACD 控制器, 但需要已知系统动力学模型. 文献 [92] 给出了针对一般非线性系统的基于值迭代 HDP 算法的收敛性分析, 证明 HDP 算法将收敛到 HJB 方程的最优控制解.

由于时滞因素 (通常包括状态时滞和控制时滞等) 往往是各类系统或模型中所不可避免的问题, 所以基于 MDP 的动态时滞系统学习控制也开始得到学术界的关注. Wei 等[93] 针对一类同时具有状态和控制时滞的非线性系统提出了一种迭代近似动态规划算法, 该算法引入了一个时滞矩阵函数, 通过迭代算法对时滞非线性系统的最优控制律进行逼近, 并且可以结合神经网络实现对连续状态与动作空间的泛化. 近年来的其他相关研究进展包括: 对于存在饱和驱动器的时滞系统, Song 等[94] 给出了基于启发式动态规划的近似最优学习控制算法; Zhang 等[95] 提出了具有控制约束的离散时间仿射非线性系统的近似最优控制算法.

目前, 基于 ACD 的近似动态规划方法在复杂非线性系统的学习控制方面有很大的应用价值[91, 96], 并且取得了一些应用研究成果, 但在学习算法与理论分析方面还需要进一步加强.

7 增强学习与近似动态规划的应用

在智能机器人系统[97−103]、电力系统控制[104−107]、工业过程与汽车电子控制[108−111]、武器系统与通讯网络的学习控制[112−120]、大规模生产调度与管理[121−123] 等领域, RL 与 ADP 的应用正在不断得到推广和深入. 在许多模型复杂或者不确定的优化控制问题中, 增强学习都显示了相对传统控制方法的性能优势. 具体来说, RL 与 ADP 的应用领域主要包括如下几个方面:

1) 智能机器人系统的自主学习控制. 机器人的学习系统研究是近年来机器人学界的研究热点之一, 主要目标是克服机器人系统面临的不确定环境与非线性动力学特性, 实现高性能的机器人自主行为和运动控制. Hasegawa 等研究了结合模糊逻辑的两关节悬摆机器人的学习控制, 获得了较好的效果[97]. 近年来, 移动机器人的学习控制也得到了广泛研究[98−103], 例如, Meng 等提出了一种移动机器人自主避障的混合学习控制方法[100], Lin 等研究了基于自适应评价增强学习的移动机器人防滑控制[101], Juang 等研究了采用模糊增强学习的移动机器人墙壁跟随控制方法[102]. 在已有的研究工作中, 如何提高机器人学习控制器的泛化能力与学习效率

是重点关注的问题, 并且还有待进一步完善. 另外, 自主机器人系统的高效在线学习控制也是值得进一步研究的方向.

2) 电力系统控制. 由于电力系统模型的复杂性和各种动态扰动的不确定性, 基于 MDP 的学习控制在电力系统控制中具有广泛的应用前景. Mohagheghi 等研究了多机 (Multi-machine) 电力系统的自适应评价设计学习控制器, 结合了模糊神经网络进行值函数与策略的逼近[105], 并且研究了具有比例–积分形式的学习控制器实现[104]. Park 等研究了电力系统中基于自适应评价设计的发电机学习控制, 分别采用多层前馈神经网络与径向基函数 (Radial basic function, RBF) 神经网络进行了近似最优值函数与策略的逼近[106]. 另外, 在电力传输过程中的 MDP 学习控制方法也取得了初步的研究成果[107].

3) 工业过程与汽车电子控制. 在复杂工业过程中, 基于 RL 与 ADP 的学习控制也取得了若干重要研究进展, 如采用在线监督器提高 GDHP 学习控制器容错能力的方法[112], 基于 ACD 的流加发酵重优化技术等[113]. 在汽车电子控制领域, Liu 等提出了汽车发动机燃空比学习控制的自适应评价学习算法[109], Shih 等研究了基于 ADP 的发动机废气再循环学习控制算法[108].

4) 大规模生产调度与控制. Boyan 提出了一种基于值函数学习和逼近的全局优化算法 — STAGE, 在一系列大规模优化问题的求解中, STAGE 算法的性能都超过了模拟退火算法 (SA)[124]. 采用具有贪心策略优化机制的 TD(λ) 算法和神经网络逼近器, Crites 等进行了电梯调度的优化[121], Zhang 等进行了生产中的 Job-Shop 问题的优化[122], 上述应用都取得了令人满意的结果, 显示了增强学习在优化和调度中广泛应用前景. 增强学习在优化调度中的其他应用还包括基于线性函数逼近 Q 学习算法的多处理机系统的负载平衡调度[123] 等.

5) 武器系统与通讯网络的控制. 在武器系统方面, 基于 MDP 的学习控制也对于克服非线性和外界扰动, 提高武器系统性能具有重要意义. Lin 研究了基于模糊基函数逼近的回转 (Back-to-turn, BTT) 弹头自学习导引控制方法[115], Bertsekas 等研究了导弹拦截系统的神经动态规划方法[114]. 在通讯网络方面, 主要的应用包括 CDMA 网络的呼叫容许控制[120] 和主动队列控制算法[119] 等.

8 发展趋势分析与展望

目前, 增强学习与近似动态规划理论和算法虽然已经取得了很多令人鼓舞的研究进展, 但对于实际工程中日益增加的复杂优化决策和学习控制问题, 仍然需要在理论和算法方面开展创新研究, 以进一步提高算法在大规模连续空间问题中的快速收敛与泛化性能, 并且在完善理论分析的同时, 推进增强学习和近似动态规划在实际系统中的广泛应用. 其中有待进一步研究解决的重要课题包括:

1) 基于高维 MDP 空间分解的学习控制

"分而治之" 是解决复杂问题的重要手段. 由于许多复杂优化决策问题具有高维状态与行为空间, 如何通过对高维问题的结构化分解来简化问题空间, 提高学习控制算法的效率是推动增强学习和近似动态规划方法广泛应用的重要技术途径. 利用对 MDP 问题空间的结构化分解或分层的思想是降低高维 MDP 计算复杂性的重要途径, 也是结构化或分层增强学习 (Hierarchical reinforcement learning, HRL) 逐渐得到广泛关注的主要原因[125−127]. 在不同 HRL 方法中, 任务分解和问题表达方式有所不同, 但其本质均可归结为划分任务并且抽象出系列子任务, 学习在不同层次上分别进行.

大多数 HRL 方法的抽象结构通常由专家直接设计确定, 不具有自学习和自适应能力, 在领域知识不完备或设计者经验不足时, 学习效率会受到不同程度的影响, 大规模问题的自动分层是解决该问题的途径之一, 是目前 HRL 领域研究的一个热点. Hengst 研究了 HEXQ 技术, 在一定条件下对离散空间问题进行了任务分层的自动学习[128]. 目前 HRL 的研究仍然主要局限于离散空间的表达, 且没有在实际应用问题中得到充分的应用验证. 针对这些不足, 今后 HRL 研究的发展趋势主要包括: 具有高效的连续空间逼近与泛化能力的结构化增强学习方法[129]、HRL 的自动分层理论与方法、基于部分感知马氏决策过程 (POMDP) 的 HRL 以及基于多 Agent 合作的结构化增强学习等, 同时结构化增强学习在实际大规模复杂问题中的应用也是重要的研究方向.

2) 动态系统学习控制的稳定性与鲁棒性理论

增强学习与近似动态规划的目标是实现 MDP 的近似最优控制, 可以采用基于仿真或者批量采样数据的离线学习控制模式或者与被控对象直接交互的在线学习控制模式. 对于离线学习控制模式, 算法的收敛性和近似最优性能是需要重点解决的问题, 但同时需要保证仿真数据或者批量采样数据的有效性. 对于在线学习控制模式, 除了学习算法的收敛性, 闭环学习控制系统的动态稳定性和鲁棒性也是有待解决的困难问题. 围绕上述问题, 近年来的有关研究进展包括: 采用对策论描述框架的 HDP 与 DHP 学习控制结构和算法[91]、基于树型支持向量机的自评价学习控制[130]. Abu-Khalaf 等在文献 [131] 中利用神经网络给出了 HJB 方程代价函数的逼近

解, 进而得到适用于饱和控制器的近似最优 (Nearly optimal) 受限状态反馈控制器. 文献 [132] 针对含饱和输入的系统, 借助 Hamilton-Jacobi-Isaacs (HJI) 方程构造策略迭代算法, 能够获得具有 H_∞ 次优性能的状态反馈控制器. 文献 [131] 针对文献 [132] 中考虑的问题更加系统地研究了如何综合分析复杂近似动态规划问题的闭式解. 在以上工作的基础上, 需要研究和建立针对一般非线性系统学习控制的近似动态规划理论[133], 其中的关键问题主要包括: 连续行为空间与连续时间 MDP 的增强学习与近似动态规划理论和算法、基于近似动态规划的学习控制系统稳定性分析、存在输入约束与外部干扰的非线性鲁棒学习控制器设计等.

3) MDP 学习控制的特征表示学习

MDP 学习控制的特征表示学习是指利用观测数据实现特征基函数的自动构造和优化选择, 提高值函数和策略逼近的精度. 值函数或策略逼近是实现增强学习在大规模或连续状态空间 MDP 中具有泛化能力的一个基本途径, 但其中一个有待解决的重要问题是如何有效的选择逼近器的特征基函数, 提高算法的泛化性能. 监督学习的统计学习理论虽然已经有大量的成果, 但由于 MDP 学习控制问题的特殊性, 仍然需要研究基于 MDP 状态观测数据的特征表示学习理论. KLSPI 学习控制算法[72] 采用的基于 ALD 的核稀疏化算法是实现基于核的特征构造的重要方法, 并且保证了基于核的近似动态规划算法的学习效率. 但核函数及其参数本身的选择仍然需要研究新的理论和算法[134]. Mahadevan 等研究了采用拉普拉斯框架的 Proto 值函数与表示策略迭代 (Representation policy iteration, RPI) 算法[135], 为基于流形的自动基函数构造提供了一种有效方法, 但针对大规模连续空间 MDP 值函数逼近问题的 Proto 值函数理论仍然有待完善. 今后, 如何进一步结合流形学习等理论成果研究增强学习与近似动态规划的基函数构造方法是值得关注的研究方向.

4) 引入模型与先验信息的动态系统学习控制

引入模型和先验信息是提高增强学习与近似动态规划方法学习效率的重要途径. 目前的研究工作主要包括结合模型信息的规划与学习混合控制结构、以及基于先验信息的回报函数设计. Sutton 提出的 Dyna 学习控制结构是有关规划与学习混合控制结构的早期工作, 强调利用模型的规划输出提高学习效率[9]. 近年来, 结合规划与学习的 MDP 学习控制方法开始得到更多的关注[136], 如文献 [137] 提出将基于样本的规划和基于模型的 RL 进行集成, 以降低规划的计算复杂性.

回报函数的设计对于增强学习算法的学习性能具有重要的影响, 设计与实现能够融合领域先验知识的 MDP 回报函数成为近年来的一个研究方向. 其中的一个研究热点是基于学习心理学的塑造 (Shaping) 行为学习理论, 研究增强学习中的回报 Shaping 理论和方法[138−142]. 在回报 Shaping 中, 一个重要概念就是回报时域 (Horizon)[139], 即执行某个行为后获得回报准确评价值估计的延迟, 显然如果回报函数设计得好, 就可以具有较小的回报时域, 加速学习收敛速度. 在自主机器人控制中, 有关学者已经对机器人行为学习的回报 Shaping 技术开展了初步研究, 包括回报势函数等[140]. 文献 [138] 在理论上证明了对于 MDP 的最优策略来说, 在引入回报势函数后仍然是新的 (具有不同回报函数的) MDP 的最优策略, 即回报势函数不改变策略的最优性. 文献 [139] 证明了引入回报势函数与采用先验知识初始化状态值函数是等价的, 从而进一步说明, 回报势函数方法类似于混合增强学习算法, 能够通过领域知识将增强学习的计算搜索过程集中到较优的策略空间上.

为实现回报函数的自动设计, 逆增强学习 (Inverse RL)[141] 在近年来也得到了关注. 逆增强学习的思想就是利用已有的经验数据来学习 MDP 的回报函数, 从而实现回报函数的自动设计. 目前逆增强学习的研究已经取得了若干研究成果, 如 Bayes 逆增强学习算法等[143], 但还有待深入研究和完善.

9 结论

作为 MDP 学习控制的主要理论和方法, 增强学习与近似动态规划的研究已经取得了若干重要的进展, 并且在一些实际系统中逐步得到推广应用, 目前仍然面临一系列理论和技术挑战. 可以预计, 随着新的 MDP 学习控制理论的建立和完善, 增强学习与近似动态规划今后将在更多的复杂优化决策与控制问题中得到应用.

References

1 Sklansky J. Learning systems for automatic control. *IEEE Transactions on Automatic Control*, 1966, **11**(1): 6−19

2 Fu K S. Learning control systems: review and outlook. *IEEE Transactions on Automatic Control*, 1970, **15**(2): 210−221

3 Fu K S. Learning control systems and intelligent control systems: an intersection of artifical intelligence and automatic control. *IEEE Transactions on Automatic Control*, 1971, **16**(1): 70−72

4 Saridis G N. Foundations of the theory of intelligent controls. In: Proceedings of the IEEE Workshop on Intelligent Control. New York, USA: IEEE, 1985. 23−28

5 Bristow D A, Tharayil M, Alleyne A G. A survey of iterative learning control a learning-based method for high-

performance tracking control. *IEEE Control Systems Magazine*, 2006, **26**(3): 96−114

6 Kaelbling L P, Littman M L, Moore A P. Reinforcement learning: a survey. *Journal of Artificial Intelligence Research*, 1996, **4**: 237−285

7 Bertsekas D P. *Dynamic Programming and Optimal Control (Volume 2)*. Belmont, MA: Athena Scientific, 1995

8 Puterman M L. *Markov Decision Processes*. New York, USA: Wiley, 1994

9 Sutton R, Barto A G. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998

10 Wang F Y, Zhang H G, Liu D R. Adaptive dynamic programming: an introduction. *IEEE Computational Intelligence Magazine*, 2009, **4**(2): 39−47

11 Powell W B. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. New York: Wiley, 2007

12 Bertsekas D P, Tsitsiklis J N, Siklis J T. *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific, 1996

13 Liu D R. Approximate dynamic programming for adaptive control. *Acta Automatica Sinica*, 2005, **31**(1): 13−18

14 Lewis F L, Vrabie D. Reinforcement learning and adaptive dynamic programming for feedback control. *IEEE Circuits and Systems Magazine*, 2009, **9**(3): 32−50

15 Sutton R S, Barto A G, Williams R J. Reinforcement learning is direct adaptive optimal control. In: Proceedings of the American Control Conference. Waltham, MA: GTE Laboratories Inc., 1991. 2143−2146

16 Wang F Y, Jin N, Liu D R, Wei Q L. Adaptive dynamic programming for finite-horizon optimal control of discrete-time nonlinear systems with ε-error bound. *IEEE Transactions on Neural Networks*, 2011, **22**(1): 24−36

17 Wang F Y, Saridis G N. Suboptimal control for nonlinear stochastic systems. In: Proceedings of the 31st IEEE Conference on Decision and Control. Tucson, Arizona, USA: IEEE, 1992. 1856−1861

18 Saridis G N, Wang F Y. Suboptimal control for nonlinear stochastic systems. *Control Theory and Advanced Technology*, 1994, **10**(4): 847−871

19 Wang F Y, Saridis G N. On successive approximation of optimal control of stochastic dynamic systems. *Modeling Uncertainty: International Series in Operations Research and Management Science*. New York, NY: Springer, 2005. 333−358

20 Murray J J, Cox C J, Lendaris G G, Saeks R. Adaptive dynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 2002, 32(2): 140−153

21 Prokhorov D V, Wunsch D C II. Adaptive critic designs. *IEEE Transactions on Neural Networks*, 1997, **8**(5): 997−1007

22 Saridis G N. *Self-Organizing Control of Stochastic Systems*. New York: M. Dekker, 1977

23 Saridis G N [Author], Zheng Ying-Ping [Translator]. *Self-Organizing Control of Stochastic Systems*. Beijing: Science Press, 1984
(Saridis G N [著], 郑应平 [译]. 随机系统的自组织控制. 北京: 科学出版社, 1984)

24 Arimoto S, Kawamura S, Miyazaki F. Bettering operation of robots by learning. *Journal of Robotic Systems*, 1984, **1**(2): 123−140

25 Ahn H S, Chen Y Q, Moore K L. Iterative learning control: brief survey and categorization. *IEEE Transactions on System, Man, and Cybernetics Part C: Applications and Reviews*, 2007, **37**(6): 1099−1121

26 Wang Y Q, Gao F R, Doyle F J III. Survey on iterative learning control, repetitive control, and run-to-run control. *Journal of Process Control*, 2009, **19**(10): 1589−1600

27 Sun Ming-Xuan, Wang Dan-Wei, Chen Peng-Nian. Repetitive learning control for finite horizon nonlinear system. *Science China: Information Sciences*, 2010, **40**(3): 433−444
(孙明轩, 王郸维, 陈彭年. 有限区间非线性系统的重复学习控制. 中国科学: 信息科学, 2010, **40**(3): 433−444)

28 Saab S S. Selection of the learning gain matrix of an iterative learning control algorithm in presence of measurement noise. *IEEE Transactions on Automatic Control*, 2005, **50**(11): 1761−1774

29 Chen H F, Fang H T. Output tracking for nonlinear stochastic systems by iterative learning control. *IEEE Transactions on Automatic Control*, 2004, **49**(4): 583−588

30 Saab S S. A discrete-time stochastic learning control algorithm. *IEEE Transactions on Automatic Control*, 2001, **46**(6): 877−887

31 Chen H F. Almost sure convergence of iterative learning control for stochastic systems. *Science in China Series F: Information Sciences*, 2003, **46**(1): 69−79

32 Tan K K, Zhao S, Huang S, Lee T H, Tay A. A new repetitive control for LTI systems with input delay. *Journal of Process Control*, 2009, **19**(4): 711−716

33 Quan Q, Yang D, Cai K Y, Jiang J. Repetitive control by output error for a class of uncertain time-delay systems. *IET Control Theory and Applications*, 2009, **3**(9): 1283−1292

34 Pipeleers G, Demeulenaere B, Al-Bender F, De Schutter J, Swevers J. Optimal performance tradeoffs in repetitive control: experimental validation on an active air bearing setup. *IEEE Transactions on Control Systems Technology*, 2009, **17**(4): 970−979

35 Wu M, Zhou L, She J H. Design of observer-based H_∞ robust repetitive-control system. *IEEE Transactions on Automatic Control*, 2011, **56**(6): 1452−1457

36 Werbos P J. Neural networks for control and system identification. In: Proceedings of the 28th IEEE Conference on Decision and Control. Tampa, USA: IEEE, 1989. 260−265

37 Antsaklis P J. Neural networks for control systems. *IEEE Transactions on Neural Networks*, 1990, **1**(2): 242−244

38 Narendra K S, Parthasarathy K. Identification and control of dynamical systems using neural networks. *IEEE Transactions on Neural Networks*, 1990, **1**(1): 4−27

39 Liu G P. *Nonlinear Identification and Control: A Neural Network Approach*. New York: Springer-Verlag, 2001

40 Yu W. Nonlinear system identification using discrete-time recurrent neural networks with stable learning algorithms. *Information Sciences*, 2004, **158**: 131−147

41 Goethals I, Pelckmans K, Suykens J A K, De Moor B. Identification of MIMO Hammerstein models using least squares support vector machines. *Automatica*, 2005, **41**(7): 1263−1272

42 Martinez-Ramon M, Rojo-Alvarez J L, Camps-Valls G, Munoz-Mari J, Navia-Vazquez A, Soria-Olivas E, Figueiras-Vidal A R. Support vector machines for nonlinear kernel ARMA system identification. *IEEE Transactions on Neural Networks*, 2006, **17**(6): 1617−1622

43 Wang X D, Ye M Y. Nonlinear dynamic system identification using least squares support vector machine regression. In: Proceedings of 2004 International Conference on Machine Learning and Cybernetics. Shanghai, China: IEEE, 2004. 941−945

44 Goethals I, Pelckmans K, Suykens J A K, De Moor B. Subspace identification of Hammerstein systems using least squares support vector machines. *IEEE Transactions on Automatic Control*, 2005, **50**(10): 1509−1519

45 Du J Y, Wang M. Nonlinear dead zone system identification based on support vector machine. In: Proceedings of the 6th International Symposium on Neural Networks. Wuhan, China: Springer, 2009. 235−243

46 Al-Ghanim A. An unsupervised learning neural algorithm for identifying process behavior on control charts and a comparison with supervised learning approaches. *Computers and Industrial Engineering*, 1997, **32**(3): 627−639

47 Le Tallec Y. Robust, Risk-Sensitive, and Data-Driven Control of Markov Decision Processes [Ph. D. dissertation], Massachusetts Institute of Technology, USA, 2007

48 Lee J M, Lee J H. Approximate dynamic programming-based approaches for input-output data-driven control of nonlinear processes. *Automatica*, 2005, **41**(7): 1281−1288

49 Sutton R S. Learning to predict by the methods of temporal differences. *Machine Learning*, 1988, **3**(1): 9−44

50 Seymour B, O′Doherty J P, Dayan P, Koltzenburg M, Jones A K, Dolan R J, Friston K J, Frackowiak R S. Temporal difference models describe higher-order learning in humans. *Nature*, 2004, **429**(6992): 664−667

51 Xu X. A sparse kernel-based least-squares temporal difference algorithm for reinforcement learning. In: Proceedings of 2006 International Conference on Natural Computation. Yantai, China: Springer, 2006. 47−56

52 Watkins C J C H, Dayan P. Q-Learning. *Machine Learning*, 1992, **8**(3−4): 279−292

53 Singh S P, Jaakkola T, Littman M L, Szepesvári C. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 2000, **38**(3): 287−308

54 Baird L. Residual algorithms: reinforcement learning with function approximation. In: Proceedings of the 12th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufman Publishers, 1995. 30−37

55 Xu X, He H G. Residual-gradient-based neural reinforcement learning for the optimal control of an acrobot. In: Proceedings of the IEEE International Symposium on Intelligent Control. Vancouver, Canada: IEEE, 2002. 758−763

56 Tsitsiklis J N, Van Roy B. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 1997, **42**(5): 674−690

57 Boyan J A. Technical update: least-squares temporal difference learning. *Machine Learning*, 2002, **49**(2-3): 233−246

58 Bradtke S J, Barto A G. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 1996, **22**(1−3): 33−57

59 Xu X, He H G, Hu D W. Efficient reinforcement learning using recursive least-squares methods. *Journal of Artificial Intelligence Research*, 2002, **16**: 259−292

60 Xu X, Xie T, Hu D W, Lu X C. Kernel least-squares temporal difference learning. *International Journal of Information Technology*, 2005, **11**(9): 54−63

61 Engel Y, Mannor S, Meir R. The kernel recursive least-squares algorithm. *IEEE Transactions on Signal Processing*, 2004, **52**(8): 2275−2285

62 Geramifard A, Bowling M, Sutton R S. Incremental least-squares temporal difference learning. In: Proceedings of the 21st Association for the Advancement of Artificial Intelligence (AAAI) on Artificial Intelligence. Boston, Massachusetts, USA: AAAI Press, 2006. 356−361

63 Johns J, Petrik M, Mahadevan S. Hybrid least-squares algorithms for approximate policy evaluation. *Machine Learning*, 2009, **76**(2−3): 243−256

64 Sutton R S, Maei H R, Precup D, Bhatnagar S, Silver D, Szepesvári C, Wiewiora E. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In: Proceedings of the 26th International Conference on Machine Learning. Montreal, Canada: ACM, 2009. 993−1000

65 Xu Xin, He Han-Gen. A gradient algorithm for neural-network-based reinforcement learning. *Chinese Journal of Computers*, 2003, **26**(2): 227−233
(徐昕, 贺汉根. 神经网络增强学习的梯度算法研究. 计算机学报, 2003, **26**(2): 227−233)

66 Gao Yang, Chen Shi-Fu, Lu Xin. Research on reinforcement learning technology: a review. *Acta Automatica Sinica*, 2004, **30**(1): 86−100
(高阳, 陈世福, 陆鑫. 强化学习研究综述. 自动化学报, 2004, **30**(1): 86−100)

67 Heger M. The loss from imperfect value functions in expectation-based and minimax-based tasks. *Machine Learning*, 1996, **22**(1−3): 197−225

68 Schölkopf B, Smola A J. *Learning with Kernels*. Cambridge: MIT Press, 2002

69 Vapnik V N. *Statistical Learning Theory*. New York: Wiley-Interscience, 1998

70 Lanckriet G R G, Cristianini N, Bartlett P L, El Ghaoui L, Jordan M I. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 2004, **5**: 27−72

71 Ormnoneit D, Sen S. Kernel-based reinforcement learning. *Machine Learning*, 2002, **49**(2−3): 161−178

72 Xu X, Hu D W, Lu X C. Kernel-based least squares policy iteration for reinforcement learning. *IEEE Transactions on Neural Networks*, 2007, **18**(4): 973−997

73 Williams R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 1992, **8**(3-4): 229−256

74 Baxter J, Bartlett P L. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 2001, **15**(1): 319−350

75 Wang Xue-Ning, Xu Xin, Wu Tao, He Han-Gen. The optimal reward baseline for policy-gradient reinforcement learning. *Chinese Journal of Computers*, 2005, **28**(6): 1021−1026 (王学宁, 徐昕, 吴涛, 贺汉根. 策略梯度强化学习中的最优回报基线. 计算机学报, 2005, **28**(6): 1021−1026)

76 Schraudolph N N, Yu J, Aberdeen D. Fast online policy gradient learning with SMD gain vector adaptation. In: Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2006. 1185−1192

77 Sutton R S, McAllester D, Singh S, Mansour Y. Policy gradient methods for reinforcement learning with function approximation. In: Advances in Neural Information Processing Systems 12. Cambridge, MA: MIT Press, 2000. 1057−1063

78 Lagoudakis M G, Parr R. Least-squares policy iteration. *Journal of Machine Learning Research*, 2003, **4**: 1107−1149

79 Ghavamzadeh M, Engel Y. Bayesian policy gradient algorithms. In: Advances in Neural Information Processing Systems 19. Cambridge, MA: MIT Press, 2007. 457−464

80 Xu X, Liu C M, Hu D W. Continuous-action reinforcement learning with fast policy search and adaptive basis function selection. *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, **15**(6): 1055−1070

81 Mahadevan S. Representation policy iteration. In: Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence. Edinburgh, Scotland: AUAI Press, 2005. 372−379

82 Barto A G, Sutton R S, Anderson C W. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on System, Man, and Cybernetics*, 1983, **13**(5): 834−846

83 Konda V R, Tsitsiklis J N. On actor-critic algorithms. *SIAM Journal of Control and Optimization*, 2001, **42**(4): 1143−1166

84 Prokhorov D V, Santiago R A, Wunsch II D C. Adaptive critic designs: a case study for neurocontrol. *Neural Networks*, 1995, **8**(9): 1367−1372

85 Saeks R, Cox C J, Neidhoefer J, Mays P R, Murray J J. Adaptive control of a hybrid electric vehicle. *IEEE Transactions on Intelligent Transportation Systems*, 2002, **3**(4): 213−234

86 Ferrari S, Stengel R. Online adaptive critic flight control. *Journal of Guidance, Control, and Dynamics*, 2004, **27**(5): 777−786

87 Mohagheghi S, del Valle Y, Venayagamoorthy G K, Harley R G. A proportional-integrator type adaptive critic design-based neurocontroller for a static compensator in a multimachine power system. *IEEE Transactions on Industrial Electronics*, 2007, **54**(1): 86−96

88 Lu C, Si J, Xie X R. Direct heuristic dynamic programming for damping oscillations in a large power system. *IEEE Transactions on System, Man, and Cybernetics, Part B: Cybernetics*, 2008, **38**(4): 1008−1013

89 Dalamagkidis K, Kolokotsa D, Kalaitzakis K, Stavrakakis G S. Reinforcement learning for energy conservation and comfort in buildings. *Building and Environment*, 2007, **42**(7): 2686−2698

90 Al-Tamimi A, Lewis F L, Abu-Khalaf M. Model-free Q-learning designs for linear discrete-time zero-sum games with application to H-infinity control. *Automatica*, 2007, **43**(3): 473−481

91 Al-Tamimi A, Abu-Khalaf M, Lewis F L. Adaptive critic designs for discrete-time zero-sum games with application to H_∞ control. *IEEE Transactions on System, Man, and Cybernetics, Part B: Cybernetics*, 2007, *Automatica*, 2007, **37**(1): 240−247

92 Al-Tamimi A, Lewis F L, Abu-Khalaf M. Discrete-time nonlinear HJB solution using approximate dynamic programming: convergence proof. *IEEE Transactions on System, Man, and Cybernetics, Part B: Cybernetics*, 2008, **38**(4): 943−949

93 Wei Q L, Zhang H G, Liu D R, Zhao Y. An optimal control scheme for a class of discrete-time nonlinear systems with time delays using adaptive dynamic programming. *Acta Automatica Sinica*, 2010, **36**(1): 121−129

94 Song R Z, Zhang H G, Luo Y H, Wei Q L. Optimal control laws for time-delay systems with saturating actuators based on heuristic dynamic programming. *Neurocomputing*, 2010, **73**(16−18): 3020−3027

95 Zhang H G, Luo Y H, Liu D R. Neural-network-based near-optimal control for a class of discrete-time affine nonlinear systems with control constraints. *IEEE Transactions on Neural Networks*, 2009, **20**(9): 1490−1503

96 Enns R, Si J. Apache helicopter stabilization using neural dynamic programming. *Journal of Guidance, Control, and Dynamics*, 2002, **25**(1): 19−25

97 Hasegawa Y, Fukuda T, Shimojima K. Self-scaling reinforcement learning for fuzzy logic controller-applications to motion control of two-link brachiation robot. *IEEE Transactions on Industrial Electronics*, 1999, **46**(6): 1123−1131

98 Dong D Y, Chen C L, Chu J, Tarn T J. Robust quantum-inspired reinforcement learning for robot navigation. *IEEE-ASME Transactions on Mechatronics*, 2012, **17**(1): 86−97

99 Xu Xin. *Reinforcement Learning and Approximate Dynamic Programming*. Beijing: Science Press, 2010 (徐昕. 增强学习与近似动态规划. 北京: 科学出版社, 2010)

100 Meng J E, Chang D. Obstacle avoidance of a mobile robot using hybrid learning approach. *IEEE Transactions on Industrial Electronics*, 2005, **52**(3): 898−905

101 Lin W S, Chang L H, Yang P C. Adaptive critic anti-slip control of wheeled autonomous robot. *IET Control Theory and Applications*, 2007, **1**(1): 51−57

102 Juang C F, Hsu C H. Reinforcement ant optimized fuzzy controller for mobile-robot wall-following control. *IEEE Transactions on Industrial Electronics*, 2009, **56**(10): 3931−3940

103 Chen C L, Li H X, Dong D Y. Hybrid control for robot navigation — A hierarchical Q-learning algorithm. *IEEE Robotics and Automation Magazine*, 2008, **15**(2): 37−47

104 Mohagheghi S, del Valle Y, Venayagamoorthy G K, Harley R G. A proportional-integrator type adaptive critic design-based neurocontroller for a static compensator in a multimachine power system. *IEEE Transactions on Industrial Electronics*, 2007, **54**(1): 86−96

105 Mohagheghi S, Venayagamoorthy G K, Harley R G. Adaptive critic design based neuro-fuzzy controller for a static compensator in a multimachine power system. *IEEE Transactions on Power Systems*, 2006, **21**(4): 1744−1754

106 Park J W, Harley R G, Venayagamoorthy G K. Adaptive-critic-based optimal neurocontrol for synchronous generators in a power system using MLP/RBF neural networks. *IEEE Transactions on Industry Applications*, 2003, **39**(5): 1529−1540

107 Ray S, Venayagamoorthy G K, Watanabe E H. A computational approach to optimal damping controller design for a GCSC. *IEEE Transactions on Power Delivery*, 2008, **23**(3): 1673−1681

108 Shih P, Kaul B C, Jagannathan S, Drallmeier J A. Reinforcement-learning-based dual-control methodology for complex nonlinear discrete-time systems with application to spark engine EGR operation. *IEEE Transactions on Neural Networks*, 2008, **19**(8): 1369−1388

109 Liu D R, Javaherian H, Kovalenko O, Huang T. Adaptive critic learning techniques for engine torque and air-fuel ratio control. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2008, **38**(4): 988−993

110 Padhi R, Balakrishnan S N. Optimal management of beaver population using a reduced-order distributed parameter model and single network adaptive critics. *IEEE Transactions on Control Systems Technology*, 2006, **14**(4): 628−640

111 Hwang K S, Chao H J. Adaptive reinforcement learning system for linearization control. *IEEE Transactions on Industrial Electronics*, 2000, **47**(5): 1185−1188

112 Yen G G, DeLima P G. Improving the performance of globalized dual heuristic programming for fault tolerant control through an online learning supervisor. *IEEE Transactions on Automation Science and Engineering*, 2005, **2**(2): 121−131

113 Iyer M S, Wunsch D C II. Dynamic re-optimization of a fed-batch fermentor using adaptive critic designs. *IEEE Transactions on Neural Networks*, 2001, **12**(6): 1433−1444

114 Bertsekas D P, Homer M L, Logan D A, Patek S D, Sandell N R. Missile defense and interceptor allocation by neuro-dynamic programming. *IEEE Transactions on System, Man, and Cybernetics, Part A: Systems and Humans*, 2000, **30**(1): 42−51

115 Lin C K. Adaptive critic autopilot design of bank-to-turn missiles using fuzzy basis function networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2005, **35**(2): 197−207

116 Fakhrazari A, Boroushaki M. Adaptive critic-based neuro-fuzzy controller for the steam generator water level. *IEEE Transactions on Nuclear Science*, 2008, **55**(3): 1678−1685

117 Galstyan A, Czajkowski K, Lerman K. Resource allocation in the grid using reinforcement learning. In: Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems. New York, USA: IEEE, 2004. 1314−1315

118 Venayagamoorthy G K, Zha W. Comparison of nonuniform optimal quantizer designs for speech coding with adaptive critics and particle swarm. *IEEE Transactions on Industry Applications*, 2007, **43**(1): 238−244

119 Zhang Yan-Bing, Hang Da-Ming, Ma Zheng-Xin, Cao Zhi-Gang. A robust active queue management algorithm based on reinforcement learning. *Journal of Software*, 2004, **15**(7): 1090−1098
(张雁冰, 杭大明, 马正新, 曹志刚. 基于再励学习的主动队列管理算法. 软件学报, 2004, **15**(7): 1090−1098)

120 Liu D R, Zhang Y, Zhang H G. A self-learning call admission control scheme for CDMA cellular networks. *IEEE Transactions on Neural Networks*, 2005, **16**(5): 1219−1228

121 Crites R H, Barto A G. Elevator group control using multiple reinforcement learning agents. *Machine Learning*, 1998, **33**(2−3): 235−262

122 Zhang W, Dietterich T G. High-performance job-shop scheduling with a time-delay TD-λ network. In: Advances in Neural Information Processing Systems 8. Cambridge, MA: MIT Press, 1996. 1024−1030

123 Schaerf A, Shoham Y, Tennenholtz M. Adaptive load balancing: a study in multi-agent learning. *Journal of Artificial Intelligence Research*, 1995, **2**: 475−500

124 Boyan J, Moore A W. Learning evaluation functions to improve optimization by local search. *Journal of Machine Learning Research*, 2001, **1**: 77−112

125 Ghavamzadeh M, Mahadevan S. Hierarchical average reward reinforcement learning. *Journal of Machine Learning Research*, 2007, **8**: 2629−2669

126 Barto A G, Mahadevan S. Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems — Theory and Applications*, 2003, **13**(1−2): 41−77

127 Shen Jing. Research on Hierarchical Reinforcement Learning [Ph. D. dissertation], Harbin Engineering University, China, 2006
(沈晶. 分层强化学习方法研究 [博士学位论文], 哈尔滨工程大学, 中国, 2006)

128 Hengst B. Discovering Hierarchy in Reinforcement Learning [Ph. D. dissertation]. University of New South Wales, Australia, 2003

129 Xu X, Liu C M, Yang S X, Hu D W. Hierarchical approximate policy iteration with binary-tree state space decomposition. *IEEE Transactions on Neural Networks*, 2011, **22**(12): 1863−1877

130 Deb A K, Jayadeva G M, Chandra S. SVM-based tree-type neural networks as a critic in adaptive critic designs for control. *IEEE Transactions on Neural Networks*, 2007, **18**(4): 1016−1030

131 Abu-Khalaf M, Lewis F L. Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach. *Automatica*, 2005, **41**(5): 779−791

132 Abu-Khalaf M, Lewis F L, Huang J. Policy iterations on the Hamilton-Jacobi-Isaacs equation for H_∞ state feedback control with input saturation. *IEEE Transactions on Automatic Control*, 2006, **51**(12): 1989−1995

133 Abu-Khalaf M, Lewis F L, Huang J. Neurodynamic programming and zero-sum games for constrained control systems. *IEEE Transactions on Neural Networks*, 2008, **19**(7): 1243−1252

134 Ong C S, Smola A J, Williamson R C. Learning the kernel with hyperkernels. *Journal of Machine Learning Research*, 2005, **6**: 1043−1071

135 Mahadevan S, Maggioni M. Proto-value functions: a Laplacian framework for learning representation and control in Markov decision processes. *Journal of Machine Learning Research*, 2007, **8**: 2169−2231

136 Sutton R S, Szepesvári C, Geramifard A, Bowling M. Dyna-style planning with linear function approximation and prioritized sweeping. In: Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence. Helsinki, Finland: AUAI Press, 2008. 528−536

137 Walsh T J, Goschin S, Littman M L. Integrating sample-based planning and model-based reinforcement learning. In: Proceedings of the 24th AAAI Conference on Artificial Intelligence. Georgia, USA: AAAI Press, 2010. 612−617

138 Ng A Y, Harada D, Russell S. Policy invariance under reward transformations: theory and application to reward shaping. In: Proceedings of the 16th International Conference on Machine Learning. Bled, Slovenia: Morgan Kaufmann, 1999. 278−287

139 Wiewiora E. Potential-based shaping and Q-value initialization are equivalent. *Journal of Artificial Intelligent Research*, 2003, **19**(1): 205−208

140 Laud A, DeJong G. Reinforcement learning and shaping: encouraging intended behaviors. In: Proceedings of the 19th International Conference on Machine Learning. Sydney, Australia: Morgan Kaufmann, 2002. 355−362

141 Ng A Y, Russell S J. Algorithms for inverse reinforcement learning. In: Proceedings of the 17th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann, 2000. 663−670

142 Saksida L M, Raymond S M, Touretsky D S. Shaping robot behavior using principles from instrumental conditioning. *Robotics and Autonomous Systems*, 1998, 22(3-4): 231−249

143 Ramachandran D, Amir E. Bayesian inverse reinforcement learning. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. Hyderabad, India: AAAI Press, 2007. 2586−2591

徐 昕 国防科学技术大学机电工程与自动化学院教授. 主要研究方向为增强学习与智能学习系统、智能机器人和学习控制. 本文通信作者.
E-mail: xuxin_mail@263.net
(**XU Xin** Professor at the College of Mechatronic Engineering and Automation, National University of Defense Technology. His research interest covers reinforcement learning and intelligent learning system, intelligent robot, and learning control. Corresponding author of this paper.)

沈 栋 博士, 中国科学院自动化研究所博士后. 主要研究方向为迭代学习控制, 随机系统的控制与优化.
E-mail: dong.shen@ia.ac.cn
(**SHEN Dong** Ph. D., postdoctoral fellow at the Institute of Automation, Chinese Academy of Sciences. His research interest covers iterative learning control, control and optimization of stochastic systems.)

高岩青 美国亚利桑那大学系统工程系研究学者及兼职助理教授. 主要研究方向为机器人学, 动态系统控制与优化.
E-mail: yanqing@email.arizona.edu
(**GAO Yan-Qing** Research scientist and adjunct assistant professor in the Department of Systems and Industrial Engineering, University of Arizona, USA. Her research interest covers robotics, control and optimization of dynamic systems.)

王 凯 国防科学技术大学机电工程与自动化学院, 军事计算实验与平行系统技术研究中心及中国科学院自动化研究所复杂系统智能控制与管理国家重点实验室博士研究生. 主要研究方向为复杂系统的智能控制与管理和超级计算.
E-mail: kai.wang_nudt@hotmail.com
(**WANG Kai** Ph. D. candidate at the Center for Military Computational Experiments and Parallel Systems Technology, College of Mechatronics Engineering and Automation, National University of Defense Technology, and State Key Laboratory of Management and Control for Complex Systems, Chinese Academy of Sciences. His research interest covers intelligent control and management of complex systems and super computing.)