强化学习研究综述

陈学松[a,b]，杨宜民[a]
（广东工业大学 a. 自动化学院；b. 应用数学学院，广州 510006）

摘 要：在未知环境中，关于 agent 的学习行为是一个既充满挑战又有趣的问题，强化学习通过试探与环境交互获得策略的改进，其学习和在线学习的特点使其成为机器学习研究的一个重要分支。介绍了强化学习在理论、算法和应用研究三个方面最新的研究成果，首先介绍了强化学习的环境模型和其基本要素；其次介绍了强化学习算法的收敛性和泛化有关的理论研究问题；然后结合最近几年的研究成果，综述了折扣型回报指标和平均回报指标强化学习算法；最后列举了强化学习在非线性控制、机器人控制、人工智能问题求解、多 agent 系统问题等若干领域的成功应用和未来的发展方向。

关键词：强化学习；多智能体；马尔可夫决策过程
中图分类号：TP181 **文献标志码**：A **文章编号**：1001-3695(2010)08-2834-05
doi:10.3969/j.issn.1001-3695.2010.08.006

Reinforcement learning: survey of recent work

CHEN Xue-song[a,b], YANG Yi-min[a]
(*a. Faulty of Automation, b. Faculty of Applied Mathematics, Guangdong University of Technology, Guangzhou* 510006, *China*)

Abstract: The problem of agent learning to act in an unknown world is both challenging and interesting. Reinforcement learning has been successful at finding optimal control policies through trial-and-error interaction with dynamic environment. Its properties of self-improving and online learning make reinforcement learning become one of most important machine learning methods. The goal of this paper was to provide a comprehensive review of reinforcement learning about theory, algorithms and applications. First of all, this paper surveyed the foundation, model of environment of reinforcement learning. Discussed the convergence and generalization of the algorithms in the next. Then deeply discussed two representative selection of these algorithm, including discounted reward and average reward. Finally, provided some applications of reinforcement learning, and pointed out some challenges and problems of reinforcement learning.

Key words: reinforcement learning; multi-agent systems; Markov decision processes

机器学习是以知识的自动获取和产生为研究目标，是人工智能的核心问题之一。机器学习与统计学、心理学、机器人学等许多其他学科都有交叉。其中，学习心理学与机器学习的交叉综合直接促进了强化学习又称做增强学习或再励学习(reinforce learning，RL)理论与算法的产生和发展。所谓强化学习是一种以环境反馈作为输入的、特殊的、适应环境的机器学习方法，它的主要思想是与环境交互和试错，利用评价性的反馈信号实现决策的优化[1]。这也是自然界中人类或动物学习的基本途径。

根据反馈的不同，学习技术可以分为监督学习或称为有导师学习(supervised learning，RL)、无监督学习或称为无导师学习(unsupervised learning，UL)和强化学习三大类。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。在这种方法中，学习系统完成的是与环境没有交互的记忆和知识重组的功能。典型的监督学习方法包括归纳学习(ID-3、ID-5 决策树学习、AQ 系列算法等)，以 BP 算法为代表的监督式神经网络学习和基于实例的学习等。无监督学习方法主要包括各种自组织学习方法(聚类学习、自组织神经网络学习)等。强化学习通常包括两方面的含义[2]，一方面是将强化学习作为一类问题，即需要搜索智能系统的行为空间，以发现系统最优的行为；另外一方面是指解决这类问题的一种技术，即采用统计技术和动态规划方法来估计在某一环境状态下的行为效用函数值，从而通过行为效用函数来确定最优行为的技术。

近年来，强化学习技术在人工智能、机器学习和自动控制等领域中得到了广泛的研究和应用，并被认为是设计智能系统的核心技术之一[3~5]。随着强化学习算法和理论的深入，特别是强化学习的数学基础研究取得突破性进展之后，应用强化学习方法实现移动机器人行为对环境的自适应和控制器的优化成为机器人学领域研究和应用的热点之一[6]。

1 强化学习基础

描述一个智能系统面临的环境往往从以下五方面进行分析[5]：动态的(dynamic)还是静态的(static)；离散的(discrete)还是连续的(continuous)，状态完全可知(deterministic)还是状态部分可知(non-deterministic)，插曲式(Episodic) 还是非插曲式(non-episodic)，确定的(accessible) 还是不确定(inaccessible) 的。其中，所谓插曲式是指智能系统在每个学习的知识对下一个场景的学习是有用的；相反，非插曲式环境是指智能系

收稿日期：2010-01-11；**修回日期**：2010-03-08

作者简介：陈学松(1978-)，男，湖南益阳人，讲师，博士研究生，主要研究方向为人工智能与智能机器人等(chenxs@gdut.edu.cn)；杨宜民(1945-)，男，教授，博导，主要研究方向为智能机器人.

统在不同场景中学习的知识是无关的。在智能系统中[6]，如果状态的迁移是确定的，则可以确定下一个状态；否则在不确定的环境下，下一个状态依赖于某种概率分布。进一步，如果状态迁移的概率模型是稳定不变的，则为静态环境;否则为动态环境。显然，最复杂的一类问题是连续状态，部分可知、非插曲式、不确定的动态环境。

强化学习技术的基本原理是[2]：如果 agent 的某个动作导致环境正的奖赏，即为强化信号，则 agent 以后的每个动作的趋势便会加强；反之 agent 产生这个动作的趋势减弱。这与生理学中的条件反射原理是一致的。因此，强化学习的目标是学习一个行为策略，使得 agent 选择的动作能够获得环境最大的奖赏。在一个标准的强化学习框架结构中，它主要有四个要素,即策略(policy)、奖惩反馈(reward)、值函数(value function)和环境模型(model of environment)。在这四个要素中首先要解决的就是实时环境的数学模型。根据 agent 的个数和 agent 产生的动作状态可以建立三个数学模型：马氏决策过程(Markov decision processes，MDP)模型、矩阵对策(matrix games,MG)模型和随机对策(stochastic games,SG)模型。

2　强化学习理论

近两年,强化学习的研究都取得了丰硕的成果[7~16],对强化学习的研究主要是对强化学习的理论、强化学习的算法以及强化学习的应用三个方面。强化学习理论研究的主要内容是算法的收敛性和泛化有关的基础理论研究。

2.1　时序差分学习的收敛性

Sutton 等人[10]首次给出了时序差分(temporal difference，TD)学习的形式化描述和 TD(λ)学习算法,并且已经成为其他强化学习算法如 Q-学习算法的基础。针对 TD(λ)学习算法在求解平稳策略值函数预测时的收敛性,文献[11]证明了任意的表格型折扣回报 TD(λ)学习算法的概率收敛性;针对采用线性值函数逼近的 TD(λ)学习算法,文献[15]证明了平均意义下的收敛性;Tsitsiklis 等人[13]证明了线性 TD(λ)算法在概率 1 意义下的收敛性,并且给出了收敛解的逼近误差上界;针对 TD(λ)学习算法中的选取对学习性能的影响,文献[14]研究了 TD(λ)学习算法均方差与的函数关系,给出了一定假设下的表达式,并且通过实验进行了验证。

2.2　表格型强化学习的收敛性

用于求解 MDP 的学习控制问题的强化学习方法主要包括 Q-学习算法和 Sarsa 学习算法等。文献[15]证明了在学习因子满足随机逼近迭代算法条件下,并且满足 MDP 状态空间被充分遍历时,表格型 Q-学习算法以概率 1 收敛到 MDP 的最优值函数和最优策略。文献[16]进一步基于异步动态规划和随机逼近理论证明了 Q-学习算法的收敛性。Singh 等人[17]研究了表格型 Sarsa(0)学习算法的收敛性,证明了在两类学习策略条件下 Sarsa 学习算法的收敛性。采用值函数逼近器的强化学习控制算法,目前在收敛性分析理论方面还是比较缺乏。Chen Lei 等人[18]提出的残差梯度算法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。VAPS 算法虽然能够保证权值的收敛性,但无法保证策略的局部最优性[19]。Heger 等人[20]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,但是这种近似最优解的求解却相当困难。

2.3　强化学习的泛化方法

强化学习算法基本都是针对离散状态和行为空间的马氏决策过程,即状态的值函数或者行为的值函数采用表格的形式存储和迭代计算。但实际工程中的许多优化决策问题都具有大规模或连续的状态和行为空间,因此表格型强化学习也存在类似动态规划的维数灾难(curse of dimensionality)。为了克服这个困难,实现对连续状态和行为空间的马氏决策过程最优值函数和最优策略的逼近,必须研究强化学习的泛化方法,即利用有限的学习经验和记忆实现对一个大范围空间的有效知识获取和表示。目前提出的强化学习泛化方法主要包括以下三个方面：

a)值函数逼近方法的研究。随着神经网络的监督学习方法如反向传播算法的广泛研究和应用,将神经网络的函数逼近能力用于强化学习的值函数逼近逐渐开始得到学术界的重视。在时域差值学习的研究中,Duan 等人[21]利用递推最小二乘方法提出了 LS-TD(λ)算法;Boyan 等人[22]给出了直接求解稳态方程的 LS-TD(λ)算法;在神经网络作为值函数逼近器的研究中,文献[23]利用神经网络的时域差值学习实现了西洋棋的学习程序 TD-Gammon。

b)策略空间逼近方法。与值函数逼近方法不同,策略空间逼近方法通过神经网络等函数逼近器直接在马氏决策过程的策略空间搜索,但是存在如何估计策略梯度的困难。文献[24]提出了一种离散行为空间的策略梯度估计方法,目前尚未找到连续行为空间的策略梯度估计方法。

c)同时进行值函数和策略空间逼近的泛化方法。在此方法中,基本都采用了 Actor-Critic 的结构,即 Actor 网络实现对连续策略空间的逼近,Critic 网络实现对值函数的逼近。文献[25]研究了基于模糊系统的学习算法,提出了网络结构和参数同时在线调整的算法,当时该算法在线学习效率不高,系统不稳定。

3　强化学习算法

按照学习系统与环境交互的类型,强化学习可以分为非联想强化学习(non-associative RL)和联想强化学习(associative RL)两大类。非联想强化学习系统仅从环境获得回报,而不区分环境的状态;联想强化学习系统则在获得回报的同时,具有环境的状态信息反馈,其结构类似于反馈控制系统。在非联想强化学习研究方面,主要用于一些理论问题的求解,如多臂赌机等。而联想强化学习按照获得的回报是否具有延迟可以分为即时回报联想强化学习和序贯决策强化学习两种类型。由于大量的实际问题都具有延迟回报的特点,用于求解延迟回报问题的序贯决策强化学习算法成为了研究的重点。在序贯决策强化学习算法研究中,采用了随机过程中的马尔可夫决策过程模型,根据 MDP 行为选择策略的平稳性和优化指标的不同,强化学习算法可以分为折扣型回报指标强化学习和平均回报强化学习。

3.1　折扣型回报指标强化学习算法

1)TD(λ)学习算法

时序差分学习方法在早期的强化学习和人工智能中占有

重要地位,并取得了一些成功的应用,但是一直没有建立统一的形式化体系和理论基础。Sutton 等人[10]首次提出了时序差分学习的形式化描述,证明了该算法在一定条件下的收敛性,从而为时序差分学习奠定了理论基础。

2)Q-学习算法

针对优化折扣回报指标的学习控制问题,Watkins[26]提出了表格型的 Q-学习算法,用于求解 MDP 的最优值函数和最优值策略。Peng 等人[27]提出了 $Q(\lambda)$ 算法,在该算法中结合了 Q-学习算法和 TD 学习算法中的适合度轨迹(eligibility traces),以进一步提高算法的收敛速度。为了进一步提高强化学习算法的学习效率,Jonsson 等人[28]基于自适应控制中模型辨识的思想,提出了具有在线模型估计的 Dyna-Q 学习算法,该方法在学习过程中对 MDP 的模型进行在线估计,虽然能显著提高效率,但必须以巨大的计算和存储量为代价。

3)Sarsa 学习算法

Singh 等人[29]提出了一种在线策略(on-policy)的 Q-学习算法,称为 Sarsa 学习算法。在 Q-学习算法中,学习系统的行为选择策略和值函数的迭代是相互独立的,而 Sarsa 学习算法则以严格的 TD 学习形式实现行为值函数的迭代,即行为选择策略与值函数迭代是一致的。Sarsa 学习算法在一些学习控制问题的应用中被验证具有优于 Q-学习算法的性能。

4)Actor-Critic 学习算法

上述三种学习算法具有一个共同特点是仅对 MDP 的值函数进行估计,行为选择策略则由值函数的估计完全确定。为了同时对值函数和策略进行估计,Barto 等人[30]提出了 Actor-Critic 学习算法,Critic 采用 TD 学习算法实现值函数的估计,Actor 则利用一种策略梯度估计方法进行梯度下降学习。Barto 等人只考虑了离散空间的情形,文献[25]进一步研究了求解连续行为空间 MDP 最优策略的 Actor-Critic 学习算法。此外,还有一种类型就是不对 MDP 的值函数进行估计,而只进行最优策略的估计,这一类强化学习被称为直接策略梯度估计算法。该算法存在梯度估计困难、学习效率低等缺点。

3.2　平均型回报指标强化学习算法

由于在某些实际问题中,优化目标更适合用平均回报指标来描述,平均回报指标强化学习算法逐渐得到了重视。目前该类型学习算法主要包括三种:

a)基于平均回报指标的时域差值学习算法。Preux 等人[31]将求解平均回报指标 MDP 策略评价问题的动态规划理论和方法应用于时域差值学习,提出了基于平均回报指标的时域差值学习算法。在该算法中,通过引入动态规划中相对值函数(relative value function,RVF)的概念,实现了在 MDP 模型未知时对平稳策略 MDP 的值函数估计。

b)R-学习算法。文献[32]提出了 R-学习算法,该算法通过对相对值函数的迭代和贪心的行为选择策略实现广义策略迭代过程。该文中的仿真研究表明了在某些情况下,R-学习算法的性能优于 Q-学习算法。

c)H-学习算法。文献[33]提出了 H-学习算法,该算法可以看做是一种基于在线模型估计的 R-学习算法。由于折扣型指标的强化学习算法在折扣因子趋向 1 时的性能与平均回报指标的性能类似,而在理论分析方面,折扣指标算法要远远比平均回报指标算法容易,因此,本文研究主要针对折扣型指标的强化学习及其应用。

4　强化学习应用

随着强化学习在算法和理论方面研究的深入,强化学习算法在实际的工程优化和控制中得到了广泛的应用。目前强化学习方法已经在非线性控制、机器人控制、人工智能问题求解、组合优化和调度、通信和数字信号处理、多智能体、模式识别和交通信号控制等领域取得了若干成功的应用。

1)强化学习在非线性控制中的应用

在强化学习的研究中,小车倒摆系统作为一种典型的非线性控制对象,成为强化学习应用和研究的目标之一。Zhang 等人[34]采用伪熵来改进 Q-学习算法并对小车倒摆系统进行了学习控制仿真;林芬等人[35]研究了基于偏信息学习的双层强化学习算法;蒋国飞等人[36]研究了基于神经网络 Q-学习的倒立摆学习控制;王雪松等人[37]以 Q-学习方法为例,提出的基于协同最小二乘 SVM 的强化学习,有效解决了强化学习系统连续状态空间的泛化问题。

2)人工智能中的复杂问题求解

各种复杂问题求解一直是人工智能研究的重要领域,早期的启发式搜索方法和基于符号表示的产生式系统在求解一定规模的复杂问题中取得了成功。但是这些方法在实现过程中都存在知识获取和表示的困难,如 IBM 的 Deep Blue 有大量参数和知识数据库,必须通过有关专家进行手工调整才能获得良好的性能。

强化学习算法与理论的研究为人工智能的复杂问题求解开辟了一条新的途径,强化学习的基于多步序列决策的知识表示和基于尝试与失败的学习机制能够有效地解决知识的表示和获取的问题。目前,强化学习在人工智能的复杂问题求解中已经取得了若干研究成果,其中有代表性的是 Tesauro[38]的 TD-Gammon 程序,该程序采用前馈神经网络作为值函数逼近器,通过自我学习对弈实现了专家级的 Back-Gammon 下棋程序。其他的相关工作包括 Thrun[39]研究的基于强化学习的国际象棋程序,并取得了一定的进展。

3)强化学习在优化与调度中的应用

基于 MDP 的强化学习算法将随机动态规划与学习心理学和时域差值原理相结合,利用学习来计算状态的评价函数,因而能够求解模型未知的优化和调度问题。采用基于函数逼近的强化学习算法来求解大规模的优化和调度问题是强化学习应用的一个重要方面。

Boyan[22]提出了一种基于值函数学习和逼近的全局优化算法,在一系列大规模优化问题的求解中,该算法的性能都超过了模拟退火算法。采用乐观的 TD(λ)算法和神经网络逼近器,Crites 等人[40]进行了电梯调度的优化,Zhang 等人[41]进行了生产中 Job-shop 问题的优化。上述应用都取得了令人满意的结果,显示了强化学习在优化和调度中广泛的应用前景。

4)强化学习在机器人控制中的应用

在基于行为的智能机器人控制系统中,机器人是否能够根据环境的变化进行有效的行为选择是提高机器人的自主性的关键问题[42]。要实现机器人的灵活和有效的行为选择能力,仅依靠设计者的经验和知识是很难获得对复杂和不确定环境的良好适应性的。为此,必须在机器人的规划与控制系统引入学习机制,使机器人能够在与环境的交互中不断增强行为的选择能力[43]。

机器人的学习系统研究是近年来机器人学的研究热点之一[44]。一些著名大学都建立了学习机器人实验室,比如,MIT 的 Learning and Intelligent Systems 实验室,该实验室的主要兴趣包括动态环境下机器人的学习行为、理解能力等的研究;CMU 的 Artificial Intelligence 实验室,该实验室主要研究规划、知识表示以及多机器人系统和博弈论;国内的中国科技大学 Multi-agent Systems 实验室,该实验室主要以蓝鹰仿真、蓝鹰四腿机器人为平台研究多智能体系统;广东工业大学建立了智能机器人研究室,以 RoboCup 中型组比赛为实时检验平台,以机器人足球比赛仿真组 2D 或 2D 比赛平台为仿真实验平台来研究多自主移动机器人的视觉子系统、决策子系统、通信子系统和运行控制子系统等。

多移动机器人技术已经取得了许多可喜的进展,国际著名期刊,如《IEEE Transactions on Robotics and Automation》(后来更名为《IEEE Transactions on Robotics》)等陆续出版了几期有关多机器人的专辑[45~48]。最近,有关机器人的学习方法和理论的论文也不断出现,Nelson 等人[49]考虑了基于模糊逻辑和强化学习的智能机器人导航问题,Dieguez 等人[50]论述了全自主机器人的在线学习算法,Kyun 等人[6]用神经网络学习方法来控制全自主机器人,Samson 等人[51]分析了学习算法在机器人应用中的延迟问题。国内外有关强化学习算法的博士论文[5,52~57]进一步说明了强化学习在机器人控制中的应用越来越广泛和深入,但技术发展还是远远不能满足实用的要求。随着传感技术、智能技术和计算机技术等的不断发展提高,智能机器人一定能在生产和生活中扮演人的角色,更好地延伸人的手足功能和脑功能甚至社会功能,更大地提高生产率、减轻劳动强度、解放生产力,提高社会生活水平和人类的空间探索水平。

5)强化学习在多智能体中的应用

多智能体系统(multi-agent system,MAS)是由多个 agent 构成的系统,可泛指所有由多个自治或者半自治模块组成的系统。复杂问题往往需要多智能体通过协作来求解,如机器人足球比赛、风/光互补发电的能量管理系统等,然而多智能体的参与无疑又增加了问题求解的复杂性[58]。

多智能体学习并不是单个智能体学习的简单相加,多智能体强化学习过程是相当复杂的,直接依赖于多个智能体的交互。若一个 MAS 中包含多个执行不稳定策略且具有学习能力的 agent,那么每个 agent 将难以根据自己的行为确定迁移的目的状态,这个问题成为同时学习问题[59]。产生问题的原因主要是迁移的目的状态未必仅决定于自己的行为,而是常常会决定于其他 agent 的联合行为。典型的强化学习算法采用状态—动作来表示行为策略,因而不可避免地出现学习参数随状态变量维数呈指数级增长的现象[60],即维数灾难。目前,解决维数灾难问题的方法大致有四种[61]:状态聚类法、有限策略空间搜索法、值函数近似法和分层强化学习法。分层强化学习是通过在强化学习的基础上增加抽象机制,把整体任务分解为不同层次上的子任务,使每个子任务在规模较小的子空间中求解,并且求得的子任务策略可以复用,从而加快问题的求解速度[62]。分层强化学习中最主要的抽象方法是建立宏动作(macro),每个宏动作包含一个动作系列,可被系统或其他宏直接调用,从而形成了分层强化学习的控制机制[63]。

5　结束语

综上所述,尽管强化学习理论、算法和应用研究在国内外已经普遍开展,并且也已经取得了大量的研究成果,但是仍然有许多问题还亟待解决[64]。正如 Luis Nunes 等人明确提出多智能体强化学习的主要问题在于"When, why, how to exchange information?"。为了回答这个问题,理论分析[2]多智能体强化学习中的交互作用。在算法和理论方面,已经提出了多种表格型的强化学习算法,并建立了较为完善的收敛理论,但是对于连续、高维的马氏决策问题将面临维数灾难。目前已经提出的强化学习泛化方法如基于神经网络的强化学习方法等仍然存在学习效率不高、在理论上的收敛性难以保证等缺点[65],还有待进一步深入研究,以扩大强化学习在实际问题中的应用。在应用方面,实现强化学习在复杂、不确定系统中的优化控制问题,对于推动工业、航空、军事等各领域的发展有重要的意义,特别是对于多自主移动机器人系统来说,强化学习是实现具有自适应性、自学习能力的智能机器人的重要途径,为解决智能系统的知识获取这个瓶颈问题提供一个可行之法[66]。

参考文献:

[1] 高阳, 陈世福, 陆鑫. 强化学习研究综述[J]. 自动化学报, 2004,30(1):86-100.

[2] 周志华, 王钰. 机器学习及其应用[M]. 北京:清华大学出版社, 2007.

[3] 谭民, 王硕, 曹志强. 多机器人系统[M]. 北京: 清华大学出版社,2005.

[4] 原魁, 李园, 房立新. 多移动机器人系统研究发展近况[J]. 自动化学报, 2007,33(8):785-794.

[5] BOWLING M. Multi agent learning in the presence of agents with limitations[R]. Pittsburgh:Carnegie Mellon University, 2003.

[6] KYUN Y, OH S-Y. Hybrid control for autonomous mobile robotnavigation using neural network based behavior modules and environment classification[J]. Autonomous Robots, 2003,15(2):193-206.

[7] ARAI S, SYCARA K. Multi-agent reinforcement learning for planning and conflict resolution in a dynamic domain[C]//Proc of the 4th International Conference on Autonomous agents. 2000:104-105.

[8] VRANCY P, VERBEEK K, NOWE A. Decetralized learning in Markov games[J]. IEEE Trans on Systems, Man and Cyberne-tics Part B:Cybernetics, 2008, 38(4):976-981.

[9] LUCIAN B, ROBERT B, BART D S. A comprehension survey of multiagent reinforcement learning[J]. IEEE Trans on Systems, Man and Cybernetics Part C: Applications and Reviews, 2008, 68(2):156-172.

[10] SUTTON R S, BARTO A G. Reinforcement learning: an introduction [M]. Cambridge:MIT Press, 1998

[11] ZOU Bin, ZHANG Hai, XU Zong-ben. Learning from uniformly ergodic Markov chains[J]. Journal of Complexity, 2009,25(2): 188-200.

[12] XU Xin. Sequential anomaly detection based on temporal-difference learning: principles, models and case studies [J]. Applied Soft Computing, 2010,10(3):859-867.

[13] TSITSILKIS J N, ROY B van. An analysis of temporal difference learning with function approximation[J]. IEEE Trans on Automatic Control, 2007, 42(5):674-690.

[14] YU Hui-zhen, BERTSEKAS D. Convergence results for some temporal difference methods based on least squares[J]. IEEE Trans on Automatic Control, 2009, 54(7):1515-1531.

[15] JIANG Cheng-zhi, SHENG Zhao-han. Case-based reinforcement learning for dynamic inventory control in a multi-agent supply-chain system [J]. Expert Systems with Applications, 2009, 36(3): 6520-6526.

[16] AL-BATAH M S, MATISA N A, ZAMLI K Z, *et al*. Modified recursive least squares algorithm to train the hybrid multilayered perceptron (HMLP) network[J]. Applied Soft Computing, 2010,10(1): 236-244.

[17] SINGH S P, JAAKKOLA T, LITTMAN M L, et al. Convergence results for single-step on-policy reinforcement learning algorithms[J]. Machine Learning, 2000,38(3):287-308.

[18] CHEN Lei, HUANG Guang-bin, PUNG Hung-keng. Systemical convergence rate analysis of convex incremental feedforward neural networks[J]. Neurocomputing, 2009,72(10-12):2627-2635.

[19] PARK C. Convergence rates of generalization errors for margin-based classification[J]. Journal of Statistical Planning and Inference, 2009, 139(8):2543-2551.

[20] HEGER M. The loss from imperfect value function in expectation-based and minimax based tasks[J]. Machine Learning, 2006,22(1-3):197-225.

[21] DUAN Hua, SHAO Xiao-jian, Hou Wei-zhen, *et al*. An incremental learning algorithm for Lagrangian support vector machines[J]. Pattern Recognition Letters, 2009,30(15):1384-1391.

[22] BOYAN J. Least-squares temporal difference learning[C]//Proc of the 16th International Conference on Machine Learning. 1999:278-287.

[23] TESAURO G J. Temporal difference learning and TD-Gammon[J]. Communications of ACM, 1995,38(3):58-68.

[24] SUTTON R, McALLESTER D, SINGH S, *et al*. Policy gradient methods for reinforcement learning with function approximation[C]// Proc of Advances in Neural Information Processing Systems. Cambridge:MIT Press, 1999:1057-1063.

[25] LIN C T, LEE C S G. Reinforcement structure/parameter learning for neural-network based fuzzy logic control system[J]. IEEE Trans on Fuzzy System,2008, 2(1):46-63.

[26] WATKINS C. Learning from delayed rewards[D]. Cambridge:King's College, University of Cambridge, 1989.

[27] PENG Jing, WILLIAMS R J. Incremental multi-step Q-learning[J]. Machine Learning,1996,22(1-3):283-290.

[28] JONSSON A, BARTO A. Causal graph based decomposition of factored MDPs[J]. Journal of Machine Learning Research, 2006,7(11):2259-2301.

[29] SINGH S, JAAKKOLA T, LITTMAN M L, *et al*. Convergence results for single-step on-policy reinforcement-learning algorithms[J]. Machine Learning, 2000,38(3):287-308.

[30] BARTO A G, SUTTON R S, ANDERSON C W. Neuronlike adaptive elements that can solve difficult learning control problems[J]. IEEE Trans on Systems, Man and Cybernetics, 1983,13(5):834-846.

[31] PREUX P, GIRGIN S, LOTH M. Feature discovery in approximate dynamic programming[C]//Proc of IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning. 2009:109-116.

[32] AISSANI N, BELDJILALI B, TTENTESAUX D. Dynamic scheduling of maintenance tasks in the petroleum industry: a reinforcement approach[J]. Engineering Appliactions of Aritificial Intelligence, 2009,22(7):1089-1103.

[33] BIEHL M, RIEGLER P. On-line learning with a perceptron[J]. Europhysics Letters, 1994,28(7):525-530.

[34] ZHANG Ping, STEPHANE C. Uncertainty estimatewith pseudo-entropy in reinforcement learning[J]. Control Theory and Applications, 1998, 15(1):100-104.

[35] 林芬, 石川, 罗杰文, 等. 基于偏信息学习的双层强化学习算法[J]. 计算机研究与发展, 2008, 45(9):1455-1462.

[36] 蒋国飞,吴沧浦. 基于Q学习算法和BP神经网络的倒立摆控制[J]. 自动化学报,1998,24(5):662-666.

[37] 王雪松, 田西兰, 程玉虎, 等. 基于协同最小二乘支持向量机的Q学习[J].自动化学报, 2009, 35(2):214-219.

[38] TESAURO G J. Temporal difference learning and TD-Gammon[C]. Communications of ACM, 1995,38(3):58-68.

[39] THRUN S. Learning to play the game of chess[C]//Advances in Neural Information Processing Systems. Cambridge:MIT Press, 1995: 1069-1076.

[40] CRITES R H, BARTO A G. Elevator group control using multiple reinforcement learning agents[J]. Machine Learning, 1998,10(2): 235-262.

[41] ZHANG Wei, DIETTERICH T G. Value function approximation and job-shop scheduling[C]//Proc of Workshop on Value Function Approximation. 1995:95-206.

[42] 秦志斌, 钱徽, 朱淼良. 自主移动机器人混合式体系结构的一种multi-agent 实现方法[J]. 机器人, 2006, 28(5):478-482.

[43] 杨洋, 陈小平. 动态不确定环境下的决策:一种分层决策模型[J]. 计算机科学, 2005, 32(1):151-154.

[44] 李 ,潘启树, 洪炳镕. 一种基于案例推理的多agent强化学习方法研究[J]. 机器人, 2009, 31(4):320-326.

[45] ALAMI R, CHATILA R, ASAMA H. Distributed autonomous robotic systems 6[M]. London:Springer-Verlag, 2007.

[46] ARKIN R C, BEKEY G A. Special issue on robot colonies[J]. Autonomous Robots, 1997, 4(1):1-153.

[47] BALCH T, PARKER L E. Special issue on heterogeneous multi-robot systems[J]. Autonomous Robots, 2000, 8(3):207-383.

[48] ARAI T, PAGELLO E, PARKER L. Editorial:advances in multi-robot systems[J]. IEEE Trans on Robotics and Automation, 2002, 18(5):655-661.

[49] NELSON H C, YUNG. An intelligent mobile vehicle navigator based on fuzzy logic and reinforcement learning[J]. IEEE Trans on Systems, Man and Cybernetics, Part B: Cybernetics, 1999, 29(2):314-321.

[50] DIEGUEZ A R, SANZ R, LOPEZ J. Deliberative on-line localpath-planning for autonomous mobile robots [J]. Journal of Intelligent and Robotic System, 2003, 37(1):1-19.

[51] SAMSON C. Control of chained system application to path following and time-varying point-stabilization of mobile robots [J]. IEEE Trans on Automatic Control, 1995, 40(1):64-76.

[52] 杨东勇. 多机器人协作学习与进化方法[D]. 杭州:浙江大学, 2004.

[53] 王醒策. 基于强化学习和群智能方法的多机器人协作协调研究[D]. 哈尔滨:哈尔滨工程大学,2005

[54] 祖丽楠. 多机器人系统自主协作控制与强化学习研究[D]. 长春:吉林大学,2007.

(下转第 2844 页)

[36] HERTZBERG C. A framework for sparse, non-linear least squares problems on manifolds[D]. Bremen, Germany: University of Bremen,2008.

[37] ARRAS K. The CAS robot navigation toolbox[EB/OL]. (2004-02). http://www.cas.kth.se/toolbox/.

[38] ARRAS K O, GRZONKA S, LUBER G M, *et al*. An efficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities[C]//Proc of IEEE International Conference on Robotics and Automation. 2008.

[39] ARRAS K O. Feature-based robot navigation in known and unknown environments[D]. Lausanne: Swiss Federal Institute of Technology Lausanne, 2003.

[40] ZHANG Sen, XIE Li-hua, ADAMS M D. An efficient data association approach to simultaneous localization and map building[J]. International Journal of Robotics Research, 2004,24(1):1493-1498.

[41] MARTINEZ J, CALWAY A. Efficiently increasing map density in visual SLAM using planar features with adaptive measurement[C]// Proc of British Machine Vision Conference. 2009.

[42] JIE Ming, HUANG Xian-lin, LU Hong-qian. Autonomous navigation method of lunar lander using multi-scale optical flow[J]. Chinese Journal of Sensors and Actuators, 2007, 20(11):2508-2512.

[43] CHEN Chu-song, HSIEH W T, CHEN J H. Panoramic appearance-based recognition of video contents using matching graphs[J]. IEEE Trans on Systems, Man and Cybernetics, 2004, 34(1):179-199.

[44] BRAJE W L, LEGGE G E, KERSTEN D. Invariant recognition of natural objects in the presence of shadows[J]. Perception, 2000, 29(4):383-398.

[45] SANTOS-VICTOR J, SANDINI G, CUROTTO F, *et al*. Divergent stereo for robot navigation: learning from bees[C]//Proc of IEEE CS Conference on Computer Vision and Pattern Recognition. 1993:434-439.

[46] KIM D, NEVATIA R. Symbolic navigation with a generic map[C]// Proc of IEEE Workshop on Vision for Robots. 1995:136-145.

[47] KIM D, NEVATIA R. Recognition and localization of generic objects for indoor navigation using functionality[J]. Image and Vision Computing, 1998, 16(11):729-743.

[48] FORSYTH D A, PONCE J. Computer vision: a mordern approach [M]. New Jersey: Prentice Hall,2002.

[49] MAURER M, BEHRINGER R, THOMANEK F. A compact vision system for road vehicle guidance[C]//Proc of the 13th International Conference on Pattern Recognition. Washington DC: IEEE Computer Society, 1996:313-317.

[50] POMERLEAU D, JOCHEM T. Image processor drives across America [J]. Photonics Spectra, 1996, 11(2):80-85.

[51] ZHANG Peng-fei, HE Ke-zhong, OUYANG Zheng-zhu, *et al*. Multi-functional intelligent outdoor mobile robot testbed-THMR-V[J]. Robot, 2002, 24(2):97-101.

[52] SUN Zhen-ping, AN Xiang-jing, HE Han-gen. CITAVT-IV: an autonomous land vehicle navigated by machine vision[J]. Robot, 2002, 24(2):115-121.

[53] YANG Ming. Overview and prospects of the study on driverless vehicles[J]. Journal of Harbin Institute of Technology, 2006, 38 (sup):1259-1262.

[54] OHYA A, KOSAKA A, KAK A C. Vision-based navigation by mobile robots with obstacle avoidance by single-camera vision and ultrasonic sensing[C]//Proc of IEEE/RSJ International Conference on Intelligent Robots and Systems. 1997:704-711.

[55] PETERSON K, ZIGLAR J, RYBSKI P. Fast feature detection and stochastic parameter estimation of road shape using multiple LIDAR [C]//Proc of IEEE/RSJ International Conference on Intelligent Robots and Systems. 2008:612-619.

[56] NAVARRO-SERMENT L E, MERTZ C, VANDAPEL N, *et al*. LADAR-based pedestrian detection and tracking[C]//Proc of the 1st Workshop on Human Detection from Mobile Robot Platforms. 2008.

[57] KALLIE C S, SCHRATER P R, LEGGE G E. Variability in stepping direction explains the veering behavior of blind walkers[J]. Journal of Experimental Psychology: Human Perception and Performance, 2007, 33(1):183-200.

(上接第 2838 页)

[55] 杨文. 多智能体系统一致性问题研究[D]. 上海:上海交通大学, 2009.

[56] 徐昕. 增强学习及其在移动机器人导航与控制中的应用研究[D]. 长沙:国防科学技术大学,2002.

[57] 陈春林. 基于强化学习的移动机器人自主学习及导航控制[D]. 合肥:中国科技大学,2006.

[58] 苏畅, 高阳,陈世福,等. SMDP 环境下自主生成 options 的算法研究[J]. 模式识别与人工智能, 2005, 18(6):679-675.

[59] SYAFIIE S, TADEO F, MARTINEZ E. Model-free learning control of neutralization processes using reinforcement learning [J]. Engineering Applications of Artificial Intelligence, 2007, 20(6):762-782.

[60] OLFATI-SABER R. Flocking for multi-agent dynamic systems: algorithms and theory[J]. IEEE Trans on Automation Control, 2006, 51(3):401-420.

[61] PIAO S, HONG B. Fast reinforcement learning approach to cooperative behavior acquisition in multi-agents system[C]//Proc of IEEE/RSJ International Conference on Intelligent Robots and Systems. 2002:871-875.

[62] DEMIRIS J, BIRK A. Interdisciplinary approaches to robot learning: World scientific series in robotics and intelligent systems[M]. [S. l.]:World Scientific Publishing Co, 2000.

[63] CHERUBINI A, GIANNONE F, LOCCHI L, *et al*. Policy gradient learning for a humanoid soccer robot[J]. Robotics and Autonomous Systems, 2009,57(8):808-818.

[64] KONDO T, LTO K. A reinforcement learning with evolutionary state recruitment strategy for autonomous mobile robots control[J]. Robotics and Autonomous Systems, 2004, 46(2): 111-124.

[65] ARAI S, SYCARA K. Multi-agent reinforcement learning for planning and conflict resolution in a dynamic domain[C]//Proc of the 4th International Conference on Autonomous Agents. New York: ACM, 2000:104-105.

[66] PALLOTTINO L, SCORDIO V G, BICCHI A. Decentralized cooperative policy for conflict resolution in multivehicle systems[J]. IEEE Trans on Robotics, 2007, 23(6):1170-1183.

[67] 张汝波, 顾国昌. 强化学习理论、算法及应用[J]. 控制理论与应用, 2000, 17(5):637-642.