文章编号：1000－8152(2000)05－0637－06

强化学习理论、算法及应用 *

张汝波　顾国昌　刘照德　王醒策
（哈尔滨工程大学计算机系·哈尔滨，150001）

摘要：强化学习(reinforcement learning)一词来自于行为心理学，这一理论把行为学习看成是反复试验的过程，从而把环境状态映射成相应的动作. 首先全面地介绍了强化学习理论的主要算法，即瞬时差分法、Q－学习算法及自适应启发评价算法；然后介绍了强化学习的应用情况；最后讨论了强化学习目前所要研究的问题.

关键词：强化学习；瞬时差分法；Q－学习；自适应启发评价；智能控制系统
文献标识码：A

Reinforcement Learning Theory, Algorithms and Its Application

ZHANG Rubo, GU Guochang, LIU Zhaode and WANG Xingce
(Department of computer science, Harbin Engineering University·Harbin, 150001, P. R. China)

Abstract: The term, reinforcement learning, comes from behavior psychology that takes behavior learning as trial and error, by which the states of environment are mapped into corresponding actions. First, the main algorithms, temporal difference, *Q*-learning and adaptive heuristic critic, are roundly introduced. Then, the application of reinforcement learning is presented. Finally, some present research projects of reinforcement learning are discussed.

Key words: reinforcement learning; temporal difference; Q-learning; adaptive heuristic critic; intelligent control system

1　引言(Introduction)

强化学习(reinforcement learning, 又称再励学习，评价学习)是一种重要的机器学习方法，在智能控制、机器人及分析预测等领域有许多应用. 但在传统的机器学习分类中没有提到过强化学习. 而在连接主义学习中，把学习算法分为三种类型，即非监督学习(unsupervised learning)、监督学习(supervised learning)和强化学习. 所谓强化学习就是智能系统从环境到行为映射的学习，以使奖励信号(强化信号)函数值最大，强化学习不同于连接主义学习中的监督学习，主要表现在教师信号上，强化学习中由环境提供的强化信号是对产生动作的好坏作一种评价(通常为标量信号)，而不是告诉强化学习系统 RLS(reinforcement learning system)如何去产生正确的动作. 由于外部环境提供的信息很少，RLS 必须靠自身的经历进行学习. 通过这种方式，RLS 在行动-评价的环境中获得知识，改进行动方案以适应环境.

本文全面地介绍了强化学习理论的主要算法及应用情况；讨论了强化学习进一步研究的问题.

2　强化学习的发展历史及研究现状(Developmental history and researchful status quo of reinforcement learning)

强化学习是人工智能领域中既崭新又古老的课题，其研究历史可粗略地划分为两个阶段：第一阶段是 50 年代至 60 年代，可以称为强化学习的形成阶段；第二阶段是 80 年代以后，可以称为强化学习的发展阶段.

在第一阶段，“强化”和“强化学习”这些术语由 Minsky 首次提出并出现在工程文献上[1]. 当时数学心理学家探索了各种计算模型以解释动物和人类的学习行为. 他们认为学习是随机进行的，并发展了所谓的随机学习模型[2]. Widrow, Hoff 和 Rosenblatt 这些神经网络先驱们，以及心理学家 Bush 和 Mosteller 等都研究过强化学习. 他们利用了“奖励”和“惩罚”这样的术语，但他们的研究系统越来越趋向于监督学习[2~4]. 在控制理论中，由 Waltz 和付京孙于 1965 年分别独立提出这一概念[5]. 在应用方面，最早的应用例子是 Samuel 的下棋程序，该程序采用类似值迭代、瞬时差分和 Q 学习的训练机制，来学习用线性函数表示的值函数[6]. Widrow 及其同事们在研究监督学习时，认识到监督学习和强化学习之间的不同，并于 1973 年 Widrow, Gupta 和 Maitra 改正了 Widrow-Hoff 监督学习规则(常称为 LMS 规则). 新规则可实现强化学习，即根据成功和失败的信号进行学习，代替原来的使用训练样本进行学习. 他们用“有评价的学习”一词代替“有教师的学习”[7]. Saridis 把强化控制系统的控制器看成一个随机自动机，首次系统地提出了采用强化学习来解决随机控制系统的学习控制问题的方法[8].

在六七十年代，强化学习研究进展比较缓慢. 进入 80 年代以后，随着人们对人工神经网络的研究不断地取得进展，以及计算机技术的进步，人们对强化学习的研究又出现了高潮，逐渐成为机器学习研究中的活跃领域. Barto 和他的同事，

* 基金项目：黑龙江省自然科学基金资助项目.
收稿日期：1999－02－26；收修改稿日期：2000－01－10

在只用强化信号(而没有用到环境的状态信息)反馈的条件下,提出了联想奖惩算法(associated reward penalty, ARP)算法[9]. Barto 于1983年介绍了强化学习在实际控制系统中的应用情况,他采用了两个单元 ASE(associative search element)及ACE(adaptive critic element),构成了一个评价控制系统,经过反复学习,使倒摆维持较长的时间. 实际上这一思想就是强化学习中的 AHC(adaptive heuristic critic)算法的早期形式[10]. 之后,Sutton 于1984年,在他的博士论文中提出了AHC算法,比较系统的介绍了AHC思想. 文中采用两个神经元形式,对不同的算法进行了大量实验[11]. 另外,Sutton 于1988年在《Machine Learning》上发表了题为"Learning to Predict by the Methods of Temporal Differences"著名论文[12],可以说这是一篇经典之作. 文中提出了瞬时差分 TD(temporal differences)方法,解决了强化学习中根据时间序列进行预测的问题,并且在一些简化条件下证明了 TD 方法的收敛性. Dayan 对 TD 方法的收敛性作了进一步的证明[13]. 许多学者对 TD 法进行了分析和改进[14~18]. 在强化学习方法中,另一个比较著名的算法,就是 Watkins 等人提出了的 Q-Learning[19]. Watkins 对 Q-Learning 方法的收敛性进行了证明. Jing Peng 及 Wliiams 等人提出了多步的 Q-Learning 方法[20]; Szepesvar 在一定条件下证明了 Q-学习的收敛速度[21]. Werbos 等人通过将强化学习与最优控制理论和动态规划联系起来而进行了理论上的研究[22]. Singh 采用随机逼近的方法来解决最优控制问题[23],提出了替换资格迹(replacing eligiblity traces)计算方法并对替换迹进行了理论分析,证明了替换迹具有学习速度快而且也比较可靠的特点[24]. Schwartz 采用非折扣性能评价的方法来选择动作策略,通过试验验证了在某些情况下学习的效果优于 Q-Learning[25]. Mahadevan 采用平均强化值的方法,提出了 R-Learning 方法,并与 Q-Learning 进行比较,结果表明 R-Learning 学习效果较好[26]. Tadepali 及 OK Dokyeong 提供了基于模型及平均强化值的 H-Learning 方法,通过对自主导引车的试验研究,表明该算法收敛较快,也具有较好的鲁棒性[27]. 国际期刊《Machine Learning》分别在1992年和1996年出版了强化学习的专辑,着重登载数篇强化学习的理论研究论文[17, 18, 20, 24, 26, 28~30]. 《Robotics and Autonomous System》在1995年也出版了强化学习的专辑,主要介绍关于强化学习在智能机器人上的应用情况[31, 32].

从国内情况看,强化学习还处于起步阶段. 阎平凡在《信息与控制》1996年发表综述文章,介绍了强化学习的原理、主要算法及其在智能控制中的应用情况[33];并对基于可靠度最优的强化学习算法及在过程控制上的应用进行了研究[34]. 徐宁寿等采用强化学习方法对改进型广义预报控制器的设计参数进行了自学习寻优研究,并在液位控制实验系统上作了实时应用研究[35]. 杨璐采用强化学习中的 TD 法对经济领域的预测问题进行了研究[36]. 马莉、蔡自兴采用强化学习方法,对非线性系统控制问题进行了仿真实验[37];张汝波对基于强化学习的智能机器人避碰行为的学习方法进行了研究[38~40]. 蒋国飞将 Q-学习应用于倒摆控制系统[41],并通过对连续空间的离散化,证明了在满足一定条件下 Q-Learning 的收敛性[42].

3 强化学习的主要算法(Main algorithms of reinforcement learning)

3.1 瞬时差分方法(Temporal difference method)

最著名的用于解决时间信度分配问题的方法就是由 Sutton 于1988年提出的瞬时差分 TD(temporal difference method)方法[12]. 设观测数据为 $x_1, x_2, \cdots, x_m, z$,其中每个 x_i 是在时刻 t 得到的观测向量,z 是最终的结果,对每个观测-结果序列,相应的预测序列为 $p(1), p(2), p(3), \cdots, p(m)$,其中 $p(t)$ 都是 z 的估计.

假设 $p(t)$ 采用神经网络来实现,对每个观测,就决定了一个变化量 Δw_t. 一个完整的序列处理之后,w 按下式被修改:

$$W \leftarrow W + \sum_{t=1}^{m} \Delta w_t. \tag{1}$$

定义目标函数:

$$E_t = \frac{1}{2}[z - p(t)]^2. \tag{2}$$

监督学习的权值修改过程是:

$$\Delta w_t = -\eta \frac{\partial E_t}{\partial w_t} = \eta[z - p(t)] \frac{\partial p(t)}{\partial w_t} = \eta[z - p(t)] \nabla_w p(t). \tag{3}$$

其中,η 是学习率. 将 $z - p(t)$ 表示成时刻 t 之后的预测的变化之和,就是

$$z - p(t) = \sum_{k=t}^{m} [p(k+1) - p(k)].$$

这里 $p(m+1) = z$.

这样,(1)式可以写成:

$$W \leftarrow W + \sum_{t=1}^{m} \eta[z - p(t)] \nabla_w p(t) =$$

$$W + \sum_{t=1}^{m} \eta \sum_{k=t}^{m} [p(k+1) - p(k)] \nabla_w p(t) =$$

$$W + \sum_{k=1}^{m} \eta \sum_{t=1}^{k} [p(k+1) - p(k)] \nabla_w p(t) =$$

$$W + \sum_{t=1}^{m} \eta[p(t+1) - p(t)] \sum_{k=1}^{t} \nabla_w p(t),$$

即:

$$\Delta w_t = \eta[p(t+1) - p(t)] \sum_{k=1}^{t} \nabla_w p(t). \tag{4}$$

式中 η 为学习率. (4)式给出的过程称为TD(1)过程. 也可采用指数加权的形式,对以前所出现的观测向量的预测的调整赋以权重 $\lambda^k (0 \leqslant \lambda \leqslant 1)$:

$$\Delta w_t = \eta[p(t+1) - p(t)] \sum_{k=1}^{t} \lambda^{t-k} \nabla_w p(k), \tag{5}$$

称(5)式给出的过程为 TD(λ)算法.

3.2 Q-学习算法[19](Q-learning algorithm)

设环境是一个有限状态的离散马尔科夫过程,RLS 每步可在有限动作集合中选取某一动作,环境接受该动作后状态发生转移,同时给出评价 r. 环境状态以如下概率变化到

S_{t+1}:

$$\text{prob}[\ s = s_{s+1} / s_t, a_t\ ] = P[\ s_t, a_t, s_{t+1}\ ].$$

RLS 面临的任务是决定一个最优策略, 使得总的折扣奖励信号期望值最大. 在策略 π 的作用下, 状态 s_t 的值为

$$V^{\pi}(S_t) = r(\pi(S_t)) + \gamma \sum_{s_{t+1} \in S} P[\ s_t, a_t, s_{t+1}\ ]\ V^{\pi}(s_{t+1}). \quad (6)$$

动态规划理论保证至少有一个策略 π^* 使得

$$V^{\pi^*}(S_t) = \max_{a \in A}\{\ r(\pi(S_t)) + \gamma \sum_{s_{t+1} \in S} P[\ s_t, a_t, s_{t+1}\ ]\ V^{\pi^*}(s_t)\}. \quad (7)$$

Q -学习的思想是不去估计环境模型, 而是直接优化一个可迭代计算的 Q 函数, Watkin 定义此 Q 函数为在状态 s_t 时执行动作 a_t, 且此后按最优动作序列执行时的折扣累计强化值, 即

$$Q(s_t, a_t) = r_t + \gamma \max_{a_t}\{\ Q(s_{t+1}, a_t)\ |\ a_t \in A\}. \quad (8)$$

Watkin 证明了 Q -学习在一定条件下的收敛性. Q -学习可用多种神经网络来实现, 每一个网络的输出对应于一个动作的 Q 值, 即: $Q(s, a_i)$. 用神经网络实现 Q -learng 的关键是学习算法的确定. 根据 Q 函数的定义:

$$Q(s_{t+1}, a_t) = r_t + \gamma \max_{a \in A}\{\ Q(s_{t+1}, a)\}. \quad (9)$$

只有在得到最优策略的前提下上式才成立. 在学习阶段上式两边不成立, 误差信号为[40]:

$$\Delta Q = r_t + \gamma \max_{a \in A} Q(s_{t+1}, a) - Q(s_t, a). \quad (10)$$

其中, $Q(s_{t+1}, a_t)$ 表示下一状态所对应的 Q 值, 其中 ΔQ 通过调整网络的权值使误差尽可能小.

3.3　自适应启发评价算法(Adaptive heuristic critic algorithm)

AHC 强化学习系统的结构如图 1 所示. 系统主要有输入模块、随机动作选择模块、联想搜索网络 ASN(associative search network)及自适应评判网络 ACN(adaptive critic network)组成[11].

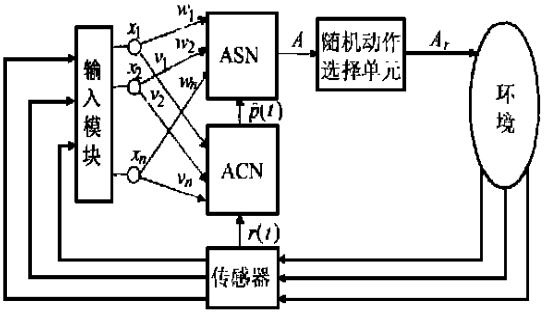

图 1　AHC强化学习系统的结构

Fig. 1　The frame of AHC reinforcement learning system

3.3.1　离散动作 AHC 算法的神经网络实现(The neural network implement of AHC algorithm for discrete actions)

所谓离散动作系统是指学习系统在每一时刻只选择同类动作的一个离散值. 设 RLS 的动作集合为 $A = \{a_1, a_2, \cdots, a_M\}$, ASN 网络的输入为环境的状态, 输出对应每个动作的指标值 $m(a_i)$. ACN 产生内部评价信号 $p(t)$, $p(t)$ 可以看成是目标函数的估计值. 按梯度法, 若使 $p(t)$ 的性能指标最大, 则权值调整为:

$$\Delta w_a(t) = \eta \frac{\partial p(t)}{\partial W_a(t)} = \eta \frac{\partial p(t)}{\partial m(a_i)} \frac{\partial m(a_i)}{\partial W_a(t)}. \quad (11)$$

若已知 $\frac{\partial p(t)}{\partial m(a_i)}$, 则可采用监督学习方式进行学习. 由于 $\frac{\partial p(t)}{\partial m(a_i)}$ 没有明确的关系式, 只能用近似的方法来得到[40]:

$$\frac{\partial p(t)}{\partial m(a_i)} \approx \eta[\ r(t) + \gamma p(t+1) - p(t)]\ [\ 1 - m(a_i)].$$

(12)

式中 $1-m(a_i)$ 表示动作 a_i 被选中后的概率之差; $[\ r(t) + \gamma p(t+1) - p(t)]$ 表示 TD 误差. 可用前向神经网络、自回归网络等来得到梯度 $\partial m(a_i) / \partial W_a(t)$.

ACN 的权值调整方法依据采用网络类型的不同而不同, 主要是根据如下目标函数来调整:

$$E_C = \frac{1}{2}[\ r(t) + \gamma p(t+1) - p(t)]^2. \quad (13)$$

故按梯度下降法, 权值调整为

$$\Delta w_c(t) = -\eta \frac{\partial E_c}{\partial W_c(t)} = \eta[\ r(t) + \gamma p(t+1) - p(t)]\ \frac{\partial p(t)}{\partial W_c(t)}. \quad (14)$$

$\partial p(t) / \partial W_c(t)$ 可采用前向网络、自回归神经网络、CMAC 和径向基函数网络等来实现.

3.3.2　连续动作的 AHC 算法(AHC algorithms for continuous actions)

上一节讨论的是离散动作 AHC 算法, 即随机动作选择单元是根据动作的概率分布, 每次选择一个动作 a_i, 且 a_i 只表示一个动作, 动作幅值不能由 ASN 决定, 而连续动作的 AHC 算法, 可以产生一个连续变化的动作. 连续动作的 AHC 算法的 ACN 的学习算法与离散的 AHC 算法相同.

ASN 网络的权值 W_a 可用梯度方法来调整:

$$\Delta w_a = \alpha \frac{\partial p(t)}{\partial W_a} = \alpha \frac{\partial p(t)}{\partial A} \frac{\partial A}{\partial W_a}, \quad (15)$$

其中 α 为学习率. 在估计梯度信息时, 利用 SRV(stochastic real valued) 方法[43]. 动作网络的输出并不直接作用于环境, 而是被当作动作的期望值, 实际的动作可围绕该值附近一定范围内搜索选择, 搜索的范围 $\sigma(t)$ 相当于一概率函数变量:

$$\sigma(t) = F[\ p(t)] = \frac{K}{1 + e^{p(t)}}. \quad (16)$$

一旦 $\sigma(t)$ 确定, 那么作用于环境的实际动作为:

$$a_k^r \sim N(a_k, \sigma(t)), \quad k = 1, 2, \cdots, M. \quad (17)$$

$N(\cdot)$ 为正态分布函数. 梯度信息由下式估计:

$$\frac{\partial p(t)}{\partial a_k} = a[\ r(t) + \gamma p(t+1) - p(t)]\ \cdot\ [\frac{a_k^r(t-1) - a_k(t-1)}{\sigma(t)}]. \quad (18)$$

σ 起到定标因子作用，$(a_k^{\gamma}-a_k)/\sigma$ 表示实际动作与期望动作的归一化偏差. 一旦 $\partial p(t)/\partial a_k$ 确定，就可利用式(15)修改网络的权值，$\partial A/\partial W_a$ 可利用多种神经网络来得到.

4 强化学习的应用(Application of reinforcement learning)

4.1 在游戏比赛中的应用(Application in game play)

游戏比赛在人工智能领域中始终是一个研究的问题，许多学者也正研究把强化学习理论应用到游戏比赛中. 在这方面，最早的应用例子是 Samuel 的下棋程序. 近来，Tesauro 把瞬时差分法应用于 Backgammon. Backgammon 大约有 10^{20} 个状态，Tesauro 采用三层 BP 神经网络把棋盘上的棋子位置与棋手的获胜概率联系起来. 通过训练取得在 40 盘比赛中负 1 盘的战绩[44, 45].

4.2 在控制系统中的应用(Application in control system)

强化学习在控制中的应用的典型实例，就是倒摆控制系统[46~49]. 倒摆控制是一个非线性不稳定系统，许多强化学习的文章都把这一控制系统作为验证各种强化学习算法的实验系统. 当倒摆保持平衡时，得到奖励，倒摆失败时，得到惩罚. 例如在 Barto 的实验系统中[10]，采用 ASE、ACE 两个神经元，通过多次的反复实验学习使得倒摆的平衡时间达到几十分钟. 强化学习另一个应用领域是在过程控制方面，采用强化学习方法不需要外部环境的数学模型，而是把控制系统的性能指标要求直接转化为一种评价指标，当系统性能指标满足要求时，所施控制动作得到奖励，否则，得到惩罚. 控制器通过自身的学习，最终得到最优的控制动作[50].

4.3 在机器人中的应用(Application in robot)

强化学习最适合、也是应用最多的，莫过于机器人领域. 近年来国际上兴起了把强化学习应用到智能机器人领域[51~57]. Hee Rak Beem 利用模糊逻辑和强化学习实现陆上移动机器人导航系统，可以完成避碰和到达指定目标点两种行为[58]. Winfried llg 采用强化学习来使六足昆虫机器人学会六条腿的协调动作[31]. Sebastian Thurn 采用神经网络结合强化学习方式使机器人通过学习能够到达室内环境中的目标[32]. 另外，强化学习也为多机器人群体行为的研究提供了一个新的途径.

4.4 在调度管理中的应用(Application in scheduling)

调度是一个随机优化控制问题的例子，具有很大的经济价值. Crites 和 Barto 将强化学习算法用于一个 4 个电梯、10 层楼的系统中[59]. 每一个电梯都有各自的位置、方向、速度和一系列表示乘客要离开的位置状态. 这个系统的状态集合将超过 10^{22} 个，用传统的动态规划方法(如值迭代方法)很难管理. 即使每回溯一个状态只要一秒钟，回溯集合中的所有状态便需约 1000 年的时间. Crites 和 Barto 采用平均的等待时间的平方做为电梯调度算法的性能. 用反传算法训练表示 Q 函数的神经网络. 与其它算法相比较，强化学习算法更加优越. 另外，强化学习在蜂窝电话系统中动态信道分配及机器调度问题上都有应用实例.

5 结论(Conclusion)

近年来，强化学习的理论及其应用研究正日益受到重视，关于强化学习的课题得到了美国国防部、美国国家科学基金及国家青年科学基金以及美国海军、空军研究办公室的资助[60~63]. 另外，德国、韩国、澳大利亚等国的学者都在开展有关强化学习的理论和应用研究[64~69]. 综合国内外的研究状况，笔者认为应从以下几个方面来进一步研究强化学习问题.

1) 系统地研究强化学习理论. 虽然国内外许多学者对强化学习理论进行了研究并取得了一定的成果，但笔者认为有关理论问题还未得到完全解决，还需要进行系统地研究. 例如：有关 AHC 算法的收敛性还没有得到证明，而 TD 算法及 Q-learning 的收敛性只是在一些简化的条件下得以证明，各种算法的计算复杂度问题目前还都没有解决.

2) 加强强化学习的应用研究. 强化学习的机理比较符合人及生物的学习过程，其思想与 Brooks 提出的行为主义思想是完全一致的. 虽然强化学习应用的范围比较广泛，但笔者认为强化学习比较适合应用于智能控制及智能机器人领域. 在智能控制方面，对于具有不确定模型的控制问题，一直是控制理论和控制工程实践中的难题，由于系统具有复杂的非线性和不确定性，使得基于数学模型的传统控制方法难于奏效. 随着控制系统复杂程度增加和控制技术的进展，学习控制正在成为一种切实可行的控制的手段. 强化学习为我们提供了一条有效的途径，采用强化学习方法可以构成一个实时学习控制系统. 在智能机器人方面，一方面可以采用强化学习实现智能机器人底层的基础控制；另一方面，也可以采用强化学习实现智能机器人的高层的行为学习，如机器人的路径规划、动作学习等. 从国内的研究状况看，强化学习的应用研究还不广泛，尤其是在实际系统中应用得更少. 因此，应加大这方面的研究力度.

3) 提高强化学习速度的理论和方法研究. 虽然强化学习在理论及应用方面的研究取得了一定成绩，但真正应用到实际还有许多工作要做. 在强化学习中，环境给出的只是定性评价，正确的答案并不知道. 这样，作为目标函数的误差值及其梯度均未知，系统学习的难度势必增大，学习时间也会增长. 虽然目前已有许多提高强化学习速度的方法，如输入空间的量化方法、利用经验回放技术、利用动作模型、利用先验知识提高强化学习速度等. 但还需要从理论上来解决这一问题.

目前，强化学习在国际上是十分活跃的研究领域，在研究强化学习时应该注意以下几个问题：

1) 连续状态和连续动作问题. 通常研究的强化学习系统，其状态和动作都认为是有限的集合. 而在实际问题中，其状态和动作往往是连续的，而连续空间的强化学习问题，目前研究得还不够深入.

2) 非马尔可夫问题. 环境从一状态转移到另一状态不一定是马氏过程. 若环境是非马氏过程，一些算法的学习效果可能不好，甚至不收敛.

3) 探索(exploration)和功绩(exploitation)问题. 强化学习系统必须对这二者进行折衷处理，即获得知识和获得高回报之间进行折衷. 探索对学习来说是非常重要的，只有通过探

索才能确定最优策略，而过多的探索会降低系统的性能，甚至在某些情况下对学习产生不利的影响.

本文比较全面地综述了强化学习的发展状况及主要算法，指出了强化学习理论目前的研究方向及所需要注意的问题，希望本文有助于同行们对强化学习的研究.

参考文献(References)

[1] Minsky M L. Theory of neural analog reinforcement systems and its application to the brain model problem[D]. New Jersey, USA: Princeton University, 1954

[2] Bush R R & Mosteller F. Stochastic Models for Learning [M]. New York: Wiley, 1955

[3] Widrow B & Hoff M E. Adaptive switching circuits[A]. In: Anderson J A and Rosenfeld E. Neurocompating. Foundations of Research [M]. Cambridge, MA: The MIT Press, 1988, 126—134

[4] Rosenblatt F. Principles of Neurodymamics: Perceptrons and the Theory of Brain Mechanisms[M]. Washington DC: Spartan Books, 1961

[5] Waltz M D & Fu K S. A heuristic approach to reinforcement learning control systems[J]. IEEE Trans. Automatic Control, 1965, 10(3): 390—398

[6] Samuel A L. Some studies in machine learning using the game of checkers[J]. IBM Journal on Research and Development, 1967, 11: 601—617

[7] Widrow B, Gupta N K & Maitra S. Punish/reward: Learning with a critic in adaptive threshold system[J]. IEEE Trans. on Systems, Man, and Cybernetics, 1973, 3(5): 455—465

[8] Saridis G N. Self-Organizing Control of Stochastic System[M]. New York: Marcel Dekker, 1977, 319—332

[9] Barto A G, Sutton R S and Brouwer P S. Associative search network: a reinforcement learning associative memory[J]. Biological cybernetics, 1981, 40: 201—211

[10] Barto A G, Sutton S and Anderson C W. Neurallike adaptive elements that can solve difficult learning control problems[J]. IEEE Trans. on Systems, Man, and Cybernetics, 1983, 13(5): 834—846

[11] Sutton R S. Temporal credit assignment in reinforcement learning [D]. Amherst, MA: University of Massachusetts, 1984

[12] Sutton R S. Learning to predict by the methods of temproal difference [J]. Machine Learning, 1988, 3: 9—44

[13] Dayan P. The convergence of TD(λ) for general λ[J]. Machine Learning, 1992, 8: 341—362

[14] Wang Lichun and Denbigh P N. Monaural localization using combination of TD(λ) with back propagation[A]. IEEE Int. Conf. on Neural Network[C], San Francisco, USA, 1993, 187—190

[15] Cichosz P and Mulawka J J. Fast and efficient reinforcement learning with truncated temporal differences[A]. Proc. 12th Int. Conf. on Machine Learning[C], Morgan Kaufmann, San Francisco, USA, 1995, 99—107

[16] Cichosz P. Truncating temporal differences: on the efficient implementation of TD(λ) for reinforcement learning[J]. J. of Artificial Intelligence Research, 1996, 12: 287—318

[17] Badtke S J and Barto R G. Linear least-squares algorithms for temporal difference learning[J]. Machine Learning, 1996, 22: 33—57

[18] Robert E S and Warmuth K. On the worst-case analysis of temporal-difference learning algorithms[J]. Machine Learning, 1996, 22: 95—121

[19] Watkins J C H and Dayan P. Q-learing [J]. Machine Learning, 1992, 8: 279—292

[20] Jing Peng and Ronald J W. Increment multi-step Q-Learning[J]. Machine Learning, 1996, 22: 283—291

[21] Szepesvari C. The asymptotic convergence-rate of Q-learning[A]. Proceedings of Neural Information Processing Systems [C], Cambridge, MA: The MIT Press, 1997, 1064—1070

[22] Werbos P J. A menu of designs for reinforcement learning over time [A]. In: Miller T, Sutton R S, Werbos P J. Neural Networks for Control[M]. Cambridge, MA: The MIT Press, 1990, 25—44

[23] Singh S P. Reinforcement learning algorithms for average payoff Markovian decision processes[A]. Proc. 12th National Conf. on Artificial Intelligence[C], Menlo Park, CA, USA: AAAI Press, 1994, 202—207

[24] Singh S P. Reinforcement learning with replacing eligibility traces [J]. Machine Learning, 1996, 22: 159—195

[25] Schwartz A. A reinforcement learning method for maximizing undiscounted rewards [A]. Proc. 10th Int. Conf. Machine Learning [C], Morgan Kaufmann, San Mateo, CA, USA, 1993, 298—306

[26] Mahadevan S. Average reward reinforcement learning: foundations, algorithms and empirical results[J]. Machine Learing, 1996, 22: 159—195

[27] Tadepali P and OK D. Model-based average reward reinforcement learning[J]. Artificial Intelligence, 1998, 100: 177—224

[28] Williams R J. Simple statistical gradient-following algorithms for connectionist[J]. Machine Learning, 1992, 8: 229—256

[29] Tesauro G. Practical issues in temporal difference learning[J]. Machine Learning, 1992, 8: 257—277

[30] Sutton R S. The challenge of reinforcement learning[J]. Machine Learing, 1992, 8: 225—227

[31] Winfried Ilg and Karsten Berns. A learning architecture based on for adaptive control of the walking machine LAURON[J]. Robotics and Autonomous System, 1995, 15: 323—334

[32] Sebastian T and Mitchell T M. Lifelong robot learning[J]. Robotics and Autonomous System, 1995, 15: 25—46

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

[34] 俞星星, 阎平凡. 强化学习系统及其基于可靠度最优化的学习算法[J]. 信息与控制, 1997, 26(1): 332—339

[35] Xu Ningshou, Wu Zhanglei and Chen Liping. A learning modified generalized predictive controller[A]. 1991 IFAC Symposium on Intelligent Tuning and Adaptive Control (ITAC'91)[C], Singapore, 1991, 231—236

[36] 杨璐, 洪家荣, 黄梯云. 用加强学习方法解决基于神经网络的时序实时建模问题[J]. 哈尔滨工业大学学报, 1996, 28(4): 136—139

[37] 马莉, 蔡自兴. 再励学习控制器结构与算法[J]. 模式识别与人

工智能, 1998, 11(1): 96－100
[38] 张汝波, 顾国昌, 张国印. 智能机器人行为学习方法研究[A]. 中国科协第三届青年学术年会论文集[C], 北京, 1998, 469－471
[39] 张汝波, 周宁, 顾国昌, 张国印. 基于强化学习的智能机器人避碰方法研究[J]. 机器人, 1999, 21(3): 204－209
[40] 张汝波, 强化学习研究及其在AUV导航系统中的应用[D]. 哈尔滨: 哈尔滨工程大学, 1999
[41] 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
[42] 蒋国飞, 高慧琪, 吴沧浦. Q学习算法中网格离散化方法的收敛性分析[J]. 控制理论与应用, 1999, 16(2): 194－198
[43] Gullapalli V. A stochastic reinforcement learning algorithm for learning real valued functions [J]. Neural Network, 1992, 3(3): 671－692
[44] Tesauro G J. TD-gammon, a self-teaching backgammon program, achieves master-level play [J]. Neural Computation, 1994, 6(2): 215－219
[45] Tesauro G J. Temporal difference learning and TD-gammon[J]. Communications of the ACM, 1995, 38(3): 58－68
[46] Anderson C W. Learning to control an inverted pendulum using neural network[J]. IEEE Control System Magazine, 1989, 30(4): 31－36
[47] Khan E. Reinforcement control with unsupervised learning [A]. Int. Joint Conference on Neural Network [C], Beijing, 1992, 88－93
[48] Berebji H R. Learning and tuning fuzzy logic controllers through reinforcements [J]. IEEE Trans. on Neural Networks, 1992, 3(5): 724－740
[49] Whitley D, Dominic S, Das R and Aanderson C W. Genetic reinforcement learning for neurocontrol problems [J]. Machine Learning, 1993, 13: 259－284
[50] Anderson C W and Hittle D C. Synthesis of reinforcement learning neural network and PI control applied to a simulated heating coil[J]. Artificial Intelligence Engineering, 1997, 11: 421－429
[51] Krose B J A and Van Dam J W M. Adaptive state space quantisation for reinforcement learning of collision-free navigation[A]. Proc. of the 1992 IEEE Int. Conference on Intelligent Robots and Systems [C], Raleigh, NC, USA, 1992, 1327－1332
[52] Millan J D R and Torras C. A reinforcement connectionist approach to robot path finding in nonlike environments[J]. Machine Learning, 1992, 8: 363－395
[53] Dillmann K B R and Zachmann U. Reinforcement learning for the control of an autonomous mobile robot[A]. Proc. of the 1992 IEEE Int. Conference on Intelligent Robots and Systems[C], Raleigh, NC, USA, 1992, 1808－1814
[54] Lin Longji. Self-improving reactive agent based on reinforcement learning, planning and teaching[J]. Machine Learning, 1992, 8: 293－321
[55] Pushkar P and Abdul S. Reinforcement learning of iterative behavior with multiple sensors[J]. Journal of Applied Intelligence, 1994, 4(5): 381－365
[56] Touzet C F. Neural reinforcement learning for behavior synthesis[J]. Robotics and Autonomous System, 1997, 22: 251－281
[57] Caironi PVC, Dorigo M. Training and delayed reinforcements in Q-learning agents [J]. Int. J. of Intelligent Systems, 1997, 12: 659－724
[58] Beom H B. A sensor-based navigation for a mobile robot using fuzzy logic and reinforcement learning[J]. IEEE Trans. on Systems, Man, and Cybernetics, 1995, 25(3): 464－477
[59] Crites R H and Barto A G. Improving elevator performance using reinforcement learning[A]. In: Touretzky D S, Mozer M C, and M E H. Advances in Neural Information Processing Systems[M]. Cambridge, MA: The MIT Press, 1995, 1017－1023
[60] Majal M. A comparative analysis of reinforcement learning methods [R]. Cambridge, MA: Massachusetts Institute of Technology, AD-A259893, 1991
[61] Lima p and Beard R. Using neural networks and dyna algorithm for integrated planning, reacting and learning in systems[R]. Troy, New York: Rensselaer Polytechnic Institute, NASA93－24743－1993
[62] Baird Ⅲ Leemon C. Learning with high dimension, continuous action [R]. Washington DC: Wright Laboratory, AD-A280844, 1993
[63] Zeng Dajun and Katia S. Using case-based reasoning as reinforcement learning framework for optimization in changing criteria[R]. Pittsburgh Pennsylvania: Carnegie Mellon University, AD-293602, 1995
[64] Anderson C W. Strategy learning with multilalyer connectionist representation[A]. Int. Conference on Machine Learning[C], Morgan Kaufmann, San Mateo, CA, USA, 1987, 103－114
[65] Mills P M and Zomaya A Y. Reinforcement learning using back propagation as building block[A]. IEEE Int. Conference on Neural Network[C], Singapore, 1991, 1554－1559
[66] Kokar M M and Reveliotis S A. Reinforcement learning: architectures and algorithms[J]. Int. of Intelligent Systems, 1993, 8: 875－894
[67] Moure A W and Atkson C G. Prioritized sweeping: reinforcement learning with less data and less time [J]. Machine Learning, 1993, 13: 103－130
[68] Pack K L. Associative reinforcement learning: function in K-DNF [J]. Machine Learning, 1994, 15: 279－293
[69] 郭茂祖, 陈彬, 王晓龙. 加强学习[J]. 计算机科学, 1993, 25(3): 13－15

本文作者简介

张汝波　1963年生. 分别于1984年, 1987年和1999年在哈尔滨船舶工程学院, 哈尔滨工程大学获学士, 硕士和博士学位, 现为哈尔滨工程大学计算机系副教授. 研究方向为机器学习, 计算智能, 智能控制及智能机器人.

顾国昌　1946年生. 1967年毕业于哈尔滨军事工程学院, 1987年在法国获博士学位. 现为哈尔滨工程大学计算机系教授, 博士生导师. 研究领域为智能机器人, 机器人体系结构, 行动决策和控制技术.

刘照德　1977年生. 1998年毕业于西安电子科技大学, 现为哈尔滨工程大学计算机系硕士研究生. 研究方向为智能控制及智能机器人.

王醒策　1978年生. 现为哈尔滨工程大学计算机系硕士研究生. 研究方向为智能控制及智能机器人.