强化学习理论及应用

Reinforcement Learning Theory and Applications

◆张汝波 编著

哈尔滨工程大学出版社

强化学习理论及应用

Reinforcement Learning Theory and Applications

张汝波　编著
顾国昌　主审

哈尔滨工程大学出版社

图书在版编目(CIP)数据

强化学习理论及应用/张汝波编著.—哈尔滨:哈尔滨工程大学出版社,2001.4
ISBN 7-81073-142-4

Ⅰ.强... Ⅱ.张... Ⅲ.①学习系统-基本知识 ②智能控制-基本知识 Ⅳ.TP273

中国版本图书馆 CIP 数据核字(2001)第 14293 号

内 容 简 介

强化学习(Reinforcement Learning)一词来源于行为心理学,它把行为学习看作反复试验的过程。目前,在国际机器学习和智能控制研究领域中,强化学习是一个热点研究问题。本书全面系统地介绍了强化学习理论及其在智能控制中应用的最新成果。全书共十一章,主要介绍了强化学习的结构模型、瞬时差分方法、自适应启发评价学习、Q-学习,及在一些控制系统、智能机器人中的应用实例。

本书可作为高等院校和科研机构从事计算机应用,人工智能及智能控制等相关专业的教师、研究人员、研究生及高年级本科生参考使用。

哈尔滨工程大学出版社出版发行
哈尔滨市南通大街145号 哈工程大学11号楼
发行部电话:(0451)2519328 邮编:150001
新 华 书 店 经 销
哈尔滨工业大学印刷厂印刷

*

开本 850mm×1 168mm 1/32 印张 9.3125 字数 250 千字
2001 年 4 月第 1 版 2001 年 4 月第 1 次印刷
印数:1~1 000 册
定价:13.50 元

前　言

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、机器人及人工智能等领域有许多应用。但在传统的机器学习分类中很少提到强化学习。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、监督学习和强化学习。强化学习一词最初来源于行为心理学,它把行为学习看作反复试验的过程。观察生物(特别是人)适应环境的学习过程可以发现它有两个特点:一是人从来不是静止地被动等待,而是主动对环境做试探;二是从环境对试探动作的反馈信号看,多数情况下是评价性(奖或罚)的,而不是像监督学习那样给出正确答案。生物在行动-评价的环境中获得知识,改进行动方案以适应环境,达到预想的目的。具有上述特点的学习就是强化学习。

所谓强化学习,可以说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。强化学习不同于监督学习,主要表现在教师信号上。在强化学习中,由环境提供的强化信号对产生动作的好坏作一种评价(通常为标量信号),而不告诉强化学习系统如何去产生正确的动作。由于外部环境提供的信息很少,强化学习系统必须靠自身的经历进行学习。

目前,在国际机器学习和智能控制研究领域中,强化学习是一个热点研究问题,欧美等国都设立了国家级科研基金来支持这方面的研究。但在国内,对强化学习的研究还处于起步阶段,有关这方面的研究论文还不多见,特别是缺少系统介绍强化学习的研究专著。本书是在作者的博士论文及本人对多年科研工作总结的基础上,加上对所收集的国内外最新文献进行归纳、整理编写而成

的。本书全面、系统地介绍了强化学习理论及其在智能控制和智能机器人等领域应用中的最新成果，全书共十一章。

本书的第1章是绪论。详细介绍了强化学习的基本概念及强化学习的发展历史、强化学习的研究状况和强化学习研究存在的问题。

第2章介绍了强化学习的工作原理和强化学习系统的结构模型，以及输入模块、评价模块和策略模块的实现方法。

第3章介绍了与强化学习有关的其它知识，即简单介绍了马尔柯夫决策过程、动态规划及蒙特卡罗方法等。

第4章介绍了瞬时差分方法。重点介绍了瞬时差分方法的基本原理、瞬时差分方法的收敛性及瞬时差分法的 Worst－Case 分析。

第5章介绍了自适应启发评价学习系统的结构，及系统中自适应评价单元、动作搜索单元、随机动作选择单元的实现方法以及连续动作的强化学习问题。

第6章介绍了强化学习的一个重要学习方法——Q－学习。首先介绍了 Q－学习的基本原理及实现方法；讨论了 Q－学习的收敛性及收敛速度，然后介绍了 Q－学习的其它算法。

第7章介绍了资格迹在强化学习的几个主要算法中，即在瞬时差分方法及 Q－学习算法中所起的作用。

第8章介绍了提高强化学习速度的方法，主要介绍了如何采用经验回放方法、环境模型方法及空间量化方法等来提高强化学习速度。

第9章介绍了强化学习控制系统，重点介绍了强化学习理论在非线性控制系统、过程控制系统和动态跟踪系统中的应用情况。

第10章介绍了强化学习控制系统在智能机器人的应用情况。主要以两种不同类型的智能机器人为例，详细介绍了采用不同强化学习方法使智能机器人具有了行为学习能力。

第11章介绍了强化学习理论在游戏、对弈及调度等领域中的

应用情况。

在本书的编写过程中得到了杨广铭、刘照德、王醒策、仲宇、戴元军等研究生在资料收集、整理等方面的帮助;哈尔滨工程大学计算机系顾国昌教授审阅了全书,提出了宝贵的意见,在此表示衷心的感谢。

书中的研究内容得到了黑龙江省自然科学基金的资助。

强化学习是一门较新的研究领域,有许多理论及应用问题还有待进一步研究解决。由于笔者的水平所限,书中难免存在一些不足和错误之处,欢迎读者批评指正。

作　者

2000 年 12 月

目　录

1 绪　论

1.1 学习的定义

学习是人类获取知识的主要形式，也是人类具有智能、提高智能水平的基本途径。建造具有类似人的智能机器是智能控制、人工智能研究的目标。要使机器具有一定的智能，一种方式是靠人事先编程来建立知识库和推理机制，这具有明显的局限性。我们希望机器具有从环境中学习的能力，即自动获取知识、积累经验、不断更新和扩充知识、改善知识性能的能力。

学习的概念在日常生活中使用极其广泛，其内涵虽然通俗但难于给出精确的定义。人们可以从不同的学科角度、根据不同的理解来表述学习。

维纳(Wiener)根据物种随时间变异的现象从生物进化的角度给出了学习的最一般的定义：具有生存能力的动物，是那些在它的个体的一生中，能被它所经历的环境所改造的动物；一个能繁殖的动物至少能够产生与它自己大致相似的动物，虽然这种动物不会完全相似到随时间的推移而不再发生变化的程度；如果这种变化是自我可遗传的，则就有了一种能受自然选择的基础；如果这种变化以某种有益的行为形式显现出来，则这种变化就会一代代地继续下去。这种从一代到下一代的变化形式就称为种族学习或系统发育学习，而特定个体中发生的行为变化或行为学习则称为个体发育学习[129]。

香农(Shannon)关于学习的定义考虑了所有可能的个体发育学习中的一个子集：假定一个有机体或一台机器处于某类环境中或者同该类环境有联系；存在一个对该环境是“成功”的量度或“自

适应"的量度,并且这种量度在时间上是局部的量度,即人们能在比该有机体生命期更短的时间内,测定这个成功的量度;如果对于所考虑的这类环境,这种局部的成功量度有随时间而改善的趋向,那么我们可以说,相对于所选择的成功量度,该有机体或机器正在为适应这类环境而学习着[42]。

奥斯古德(Osgood)从生理学角度表述了学习的定义:所谓学习是指在同类特征的重复环境中,有机体个体靠自己的自适应性,使自己的行为在竞争反应中不断地改变、增强,这类选择的变异是由个体的经验形成的。

特斯皮克(Tsypkin)把系统中的学习一词理解为一种过程,通过重复各输入信号并从外部校正该系统,从而使系统对于特定的输入信号具有特定的响应。而自学习就是不具有外来校正的学习,或不具备惩罚和奖励的学习[111]。

萨里迪斯(Saridis)认为,如果一个系统能对一个过程或其环境的未知特征所固有的信息进行学习,并将得到的经验用于进一步估计、分类、决策或控制,从而使系统的品质得到改善,那就称此系统为学习系统。而学习系统将其得到的学习信息用于控制具有未知特征的过程,就成为学习控制系统。学习控制系统能通过与控制对象和环境的闭环交互作用,并根据过去获得的经验信息,逐步改进系统自身的未来性能[79]。

按照人工智能大师西蒙(H. Simon)的观点,学习就是系统在不断重复的工作中对本身能力的增强或改进,使得系统在下一次执行同样的任务时,会比现在做得更好或效率更高。

这些表述说明了具有学习能力的智能系统具有以下特点:

①有一定的自主性:学习系统的性能是自我改进的;

②是一种动态过程:学习系统的性能随时间而变,性能的改进是在与外界反复作用的过程中进行的;

③有记忆功能:学习系统需要积累经验,用以改进其性能;

④有性能反馈:学习系统需要明确它的当前性能与某个目标

性能之间的差距，施加改进操作。

在本书中，我们把具有学习能力的系统或机器统称为 Agent（智能体）。

1.2 连接主义学习的分类

在连接主义学习中，学习算法基本上可以分为三种类型，即非监督学习（Unsupervised Learning）、监督学习（Supervised Learning）和强化学习（Reinforcement Learning）。

非监督学习规则在生理学上就是巴甫洛夫（Pavlov）的条件反射原理，当我们用一个毫无意义的刺激信号（如铃的响声）同时伴有另一个刺激信号（如食物）反复加给动物的时候，经过一段时间的训练后，动物就会建立一种联想，当再接受到相似的刺激信号时，动物就会产生条件反射。这种类型的学习完全是开环的，在神经网络学习中，称之为相关规则，即神经网络中的 Hebb 学习规则。

监督学习规则是一种反馈学习规则，当输入信号作用于系统后，观察其输出，由教师提供理想的输出信号，所产生的误差信号反馈给系统来指导学习，在神经网络学习中，称之为最小误差学习规则，或称之为 δ 规则。

观察生物（特别是人）为适应环境的学习过程可以发现它有两个特点：一是人从来不是静止地被动地等待，而是主动对环境作试探；二是从环境对试探动作的反馈信号看，多数情况下是评价性（奖或罚）的，而不是像监督学习那样给出正确答案。生物在行动－评价的环境中获得知识，改进行动方案以适应环境达到预想的目的。具有上述特点的学习就是强化学习（或称再励学习、评价学习，简记为 RL）。

1.3 强化学习的基本概念

当我们考虑学习的本质时，可能最先想到的是通过与环境的交互来学习的。当一个婴儿玩耍时，挥动手臂四处环顾并没有外在的教师指导，但的确有一个与环境的直观联系。对这种联系的体验产生了大量的关于因果关系、动作的结果和怎样做才能达到目标的信息。在我们的一生中，这种交互无疑是获得环境及我们自身知识的主要来源。无论正在学习开车或交谈，我们都敏锐地感觉到环境对我们的行为产生响应，而且我们想办法影响行为产生的结果。从交互中学习几乎是所有学习和智能理论的基本思想。

强化学习是学习如何把状态映射到动作，并且使得用数字表示的奖励信号最大。学习者并未被告知采取什么动作，像大多数机器学习的形式一样，但是必须通过试验来查明哪个动作产生最大的奖励。最有趣和最有挑战性的是动作不仅影响当前奖励，而且还影响下一个状态以及整个后继状态序列的奖励。这两个特性——反复试验搜索和延迟奖励，是强化学习的两个最突出特点。

强化学习与监督学习、统计模式识别和人工神经网络不同。监督学习是从有知识的外部教师给出的例子中学习，这是一种很重要的学习，但它单独从交互中学习还不够。在交互问题中，Agent 要获得所有代表性的例子常常是不现实的。在未知领域中——这也是学习被认为最有益之处——Agent 必须能通过自身经历来学习。

存在于强化学习中而不存在于其它学习中的挑战是探索和获益之间的均衡问题。为了获得更多奖励，强化学习 Agent 必须优先选取以往曾带来最大奖励的动作。但是为了发现这种动作，它必须试验以往未选中的动作。Agent 利用已知的东西来获得奖励，也必须进行探索以便将来选择更好的动作。其中的矛盾是获

益和探索都不能单独进行而又不能引起任务失败。Agent 必须试探多种动作,注重那些看起来最好的动作。在随机任务中,每个动作必须执行多次以获得可靠的奖励估计期望值。获益－探索矛盾已经被数学家仔细研究了数十年。现在,我们可以看到,获益－探索均衡问题在通常定义的监督学习中从未出现过。

强化学习是从一个完整的、交互的目标搜寻 Agent 开始的。对所有的强化学习,Agent 都有明确的目标,能够感知环境的特征,选择影响环境的动作。另外,可以设想从一开始 Agent 就必须在复杂未知环境下进行工作。当强化学习包括规划时,它必须处理规划与实时动作选择的相互影响以及环境模型如何获得和改进的问题。

强化学习的研究趋势使人工智能与其它工程学科的联系越来越紧密。以前,人工智能被看作是与控制理论完全分离的,它与逻辑和符号有关,与数字无关。人工智能是大型的 LISP 程序,不是线性代数、微分方程,也不是统计学。最近几十年,这个观点逐渐被瓦解,人工智能和常规工程之间一直被忽视的领域现在则是最活跃的一块,包括像神经网络、智能控制以及强化学习等新领域。

1.4 强化学习的发展历史及国内外研究状况

强化学习是人工智能领域中既崭新又古老的课题,强化学习的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,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。几乎在同时,人工智能专家和控制论学者也试图独立研究随机学习模型。其基本工作是以利用确定性自动机作为固定的随机环境,学习系统操作模型。然后,利用随机自动机实现了模型的通用化。1960 年,“强化”和“强化学习”这些术语

由明斯基(Minsky)首次提出并出现在工程文献上。同时,在控制理论中沃尔特兹(Waltz)和付京孙于1965年分别独立提出这一概念[70],用以描述通过奖励和惩罚的手段进行学习的基本思想。这些学习是通过"试错"或"反复试验"(Trial and Error)的方式进行的。当一个行为带来正确(或错误)的结果时,这种行为就被加强(或削弱)。

早期的强化学习观点,出自于传统的心理学原则。即:斯朗迪克(Thorndike)的"响应定律"(Law of Effect):"对于同一环境所做的几个响应,当那些伴随或紧跟着的响应使动物意愿得到满足且其它的条件相同时,对环境的联系将被加强;当环境重现时,这些响应重现的概率会更大;当那些同时或紧跟着的响应使动物的意愿受挫且其它条件相同时,与环境的联系将被削弱;当环境重现时,它们出现的概率将减小;得到的满足的程度越大,响应和环境的联系加强得越多;不满足的程度越大,响应与环境的联系削弱得越多。"

虽然,在心理学和其它学科上,都对该定理做过非常多的讨论,但是,这些年来,由于其基本思想已被实验所证实,且从直观上看非常正确,所以它仍具有影响力。这是将搜索和记忆相结合的方法,在多种试验动作中搜索,记住效果最好的动作。相对于示教学习原则来说,"响应定律"是一种依靠选择的学习原则。

关于"响应定律"的最早研究报告见于1954年。当时,明斯基与克拉克(Clerk)都在这一年中发表了有关文章。在明斯基的博士论文里,描述了一种模拟机,叫做SNARC(Stochastic Neural - Analog Reinforcement Calculator),这种机器可以通过反复试验来学习。克拉克描述了神经网络学习机器,但这种机器没有泛化能力。在他们的论文中,讨论更多的是监督学习而不是强化学习。这两种学习方式之间的混淆从这时就开始了,而且一直持续到今天许多研究人员仍觉得自己在研究强化学习,事实上他们研究的是监督学习。就连现在的关于神经网络的书中也常将通过训练例子学习的网络,描述成反复试验学习系统。这是因为他们利用误

差信息来更新相关连接权值，这种错误是可以理解的，只是它忽略了行为的选择特性。而在根据“响应定律”的学习中，这种选择性又是最基本的特性。

虽然在明斯基的论文中提到过根据经验估算值函数进行学习的思想，但是对这一思想介绍最多的还是在塞缪尔(Samuel)的程序中，这一程序使用瞬时差分方法学习国际象棋的规则。塞缪尔的成果对强化学习方法有非常大的影响，对当今许多研究人员来说仍是具有挑战性的问题。明斯基的学位论文是对强化学习的早期研究，更有影响力的是他在 1960 年发表的论文——“Step Toward Artificial Intelligence”。这篇论文与塞缪尔的论文一样，对与现代强化学习有关的一些问题做了深入地讨论。尤其值得注意的是，他对一个复杂的强化学习系统中必须解决的计算问题的讨论，对今天的研究仍具有一定的意义，他把这一问题叫做信度分配(Credit Assignment)问题。明斯基将塞缪尔的“Chess player”算法所使用的值函数估算方法作为解决信度分配问题的重要途径进行了讨论。他指出，这与动物学习中的条件强化现象非常相似。

罗森布拉特(Rosenblatt)、威兆(Widrow)和哈弗(Hoff)这些神经网络先驱们，以及一些心理学家如布什(Bush)和莫斯台勒(Mosteller)，都在研究强化学习。并且利用了“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。这些系统适用于模式识别和感知学习，但不适用于与环境的直接交互学习。在 20 世纪六七十年代，强化学习被阴影所覆盖，并失去了主题。而这时，监督学习得到了广泛的研究，尤其是在模式识别领域[118]。但是，1963 年，安德瑞(Andreae)提出了一种叫做 STELLA 的强化学习机器，它用一个环境模型来促进学习。安德瑞明确地考虑到一个机器如何才能与它所接触的环境交互进行学习。随后发展起来的系统都有一个内部单元用以处理部分状态观测问题，这在强化学习中非常重要。虽然，安德瑞后来的工作中继续强调与环境的交互学习，但是他把重点放在了外部教师信号的作用

上。

在20世纪60年代,米锡尔(D. Michie)将研究重点放在强化学习上。他描述了一种简单的强化学习系统,叫做MENACE(Matchbox Educable Noughts and Crosses Engine),用来学习Tic-Tac-Toe游戏的玩法。系统中有一个火柴盒,盒子中装有一定数量的彩色珠子,用以表示游戏中所有的可能位置。珠子的颜色表示从当前位置出发的、可能的移动动作。在当前位置,通过在火柴盒中随机取出一个珠子,我们就可以确定MENACE的移动位置。当一次游戏结束时,可向火柴盒中加入(或取走)某种颜色的珠子,以强化(或惩罚)MENACE的决策。

1968年米锡尔描述了一种高级的Tic-Tac-Toc强化学习自动机,叫做GLEE(Game Learning Expectimaxing Engine)。它通过利用最大期望来估算值函数,实际上就是动态规划方法。米锡尔的Tic-Tac-Toe游戏Player程序讨论了怎样将一个大问题分解成为一系列独立的子问题,以使强化学习更加有效。

米锡尔和卡姆波(Chamber)提出了一种更先进的MENACE版本,其中引入了一个称为BOX的系统。该系统用来处理倒摆的学习问题,倒摆是连接在可移动小车上的,只有当小车到达铁轨的终点或倒摆倒下时,才产生错误信号。平衡学习就是以该信号为基础的,其引入是受到威兆和史密斯(Smith)倒摆平衡系统的启发。威兆和史密斯1964年提出的倒摆平衡系统,是通过监督学习方式来完成任务的。BOX系统虽然没有估算值函数,但巴特(Barto),苏顿(Sutton)和安德森(Anderson)都从中受到启发,在相应的倒摆平衡系统中使用了值函数的估算[14,67]。

威兆及其同事们在重点研究监督学习时,认识到监督学习和强化学习之间的不同,并于1973年由威兆、格博塔(Gupta)和麦卓(Maitra)改正了Widrow-Hoff监督学习规则(常称做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。他们将这种学习称为