智能科学技术著作丛书

增强学习与近似动态规划

徐 昕 著

科学出版社
www.sciencep.com

智能科学技术著作丛书

增强学习与近似动态规划

徐 昕 著

科 学 出 版 社

北 京

内 容 简 介

本书对增强学习与近似动态规划的理论、算法及应用进行了深入研究和论述。主要内容包括：求解 Markov 链学习预测问题的时域差值学习算法和理论，求解连续空间 Markov 决策问题的梯度增强学习算法以及进化-梯度混合增强学习算法，基于核的近似动态规划算法，增强学习在移动机器人导航与控制中的应用等。本书是作者在多个国家自然科学基金项目资助下取得的研究成果的总结，意在推动增强学习与近似动态规划理论与应用的发展，对于智能科学的前沿研究和智能学习系统的应用具有重要的科学意义。

本书可作为高等院校与科研院所中从事人工智能与智能信息处理、机器人与智能控制、智能决策支持系统等专业领域的研究和教学用书，也可作为自动化、计算机与管理学领域其他相关专业师生及科研人员的参考书。

图书在版编目(CIP)数据

增强学习与近似动态规划/徐昕著. —北京：科学出版社，2010
(智能科学技术著作丛书)
ISBN 978-7-03-027565-3

Ⅰ.①增… Ⅱ.①徐… Ⅲ.①机器学习-研究②动态规划-研究
Ⅳ.①TP181

中国版本图书馆 CIP 数据核字 (2010) 第 085567 号

责任编辑：张海娜 / 责任校对：赵燕珍
责任印制：赵 博 / 封面设计：耕者设计工作室

科学出版社出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

源海印刷有限责任公司印刷
科学出版社发行 各地新华书店经销

*

2010 年 5 月第 一 版 开本：B5 (720×1000)
2010 年 5 月第一次印刷 印张：14
印数：1—2 500 字数：267 000

定价：45.00 元

(如有印装质量问题，我社负责调换)

《智能科学技术著作丛书》序

“智能”是“信息”的精彩结晶，“智能科学技术”是“信息科学技术”的辉煌篇章，“智能化”是“信息化”发展的新动向、新阶段。

“智能科学技术”(intelligence science&technology，IST)是关于“广义智能”的理论方法和应用技术的综合性科学技术领域，其研究对象包括：

• “自然智能”(natural intelligence，NI)，包括“人的智能”(human intelligence，HI)及其他“生物智能”(biological intelligence，BI)。

• “人工智能”(artificial intelligence，AI)，包括“机器智能”(machine intelligence，MI)与“智能机器”(intelligent machine，IM)。

• “集成智能”(integrated intelligence，II)，即“人的智能”与“机器智能”人机互补的集成智能。

• “协同智能”(cooperative intelligence，CI)，指“个体智能”相互协调共生的群体协同智能。

• “分布智能”(distributed intelligence，DI)，如广域信息网、分散大系统的分布式智能。

1956年，“人工智能”学科诞生，50年来，在起伏、曲折的科学征途上不断前进、发展，从狭义人工智能走向广义人工智能，从个体人工智能到群体人工智能，从集中式人工智能到分布式人工智能，在理论方法研究和应用技术开发方面都取得了重大进展。如果说，当年“人工智能”学科的诞生是生物科学技术与信息科学技术、系统科学技术的一次成功的结合，那么，可以认为，现在“智能科学技术”领域的兴起是在信息化、网络化时代又一次新的多学科交融。

1981年，“中国人工智能学会”(Chinese Association for Artificial Intelligence，CAAI)正式成立，25年来，从艰苦创业到成长壮大，从学习跟踪到自主研发，团结我国广大学者，在“人工智能”的研究开发及应用方面取得了显著的进展，促进了“智能科学技术”的发展。在华夏文化与东方哲学影响下，我国智能科学技术的研究、开发及应用，在学术思想与科学方法上，具有综合性、整体性、协调性的特色，在理论方法研究与应用技术开发方面，取得了具有创新性、开拓性的成果。“智能化”已成为当前新技术、新产品的发展方向和显著标志。

为了适时总结、交流、宣传我国学者在“智能科学技术”领域的研究开发及应用成果，中国人工智能学会与科学出版社合作编辑出版《智能科学技术著作丛书》。需要强调的是，这套丛书将优先出版那些有助于将科学技术转化为生产力以及对社会和国民经济建设有重大作用和应用前景的著作。

我们相信，有广大智能科学技术工作者的积极参与和大力支持，以及编委们的共同努力，《智能科学技术著作丛书》将为繁荣我国智能科学技术事业、增强自主创新能力、建设创新型国家做出应有的贡献。

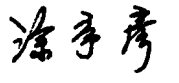
祝《智能科学技术著作丛书》出版，特赋贺诗一首：

智能科技领域广

人机集成智能强

群体智能协同好

智能创新更辉煌



中国人工智能学会荣誉理事长

2005 年 12 月 18 日

前　言

增强学习(reinforcement learning，RL)又称为强化学习或再励学习，它是近年来机器学习和智能控制领域的前沿和热点，与监督学习和无监督学习并列三大类机器学习方法之一。增强学习强调以不确定条件下序贯决策的优化为目标，是复杂系统自适应优化控制的一类重要方法，具有与运筹学、控制理论、机器人学等交叉综合的特点。特别是近十年来，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的关注热点。例如，美国国家科学基金会于2006年召开的近似动态规划论坛(NSF-ADP06)，IEEE分别于2007年和2009年召开的近似动态规划与增强学习专题国际研讨会(IEEE ADPRL'2007、IEEE ADPRL'2009)等。另外，IEEE计算智能学会于近年专门成立了近似动态规划与增强学习技术委员会(IEEE TC on ADPRL)。在以电梯调度、网络路由控制等为代表的大规模优化决策应用中，增强学习显示了相对传统监督学习和数学规划方法的优势。在智能机器人系统、复杂不确定系统的优化控制等领域，增强学习的应用也正在不断得到推广。

本书是作者多年从事增强学习与近似动态规划理论、算法与应用研究的成果总结，许多成果是近年来最新取得的研究成果，是一部系统探讨增强学习与近似动态规划的学术著作。

本书有以下几个特点：

(1) 新颖性和前沿性。本书深入论述了增强学习与近似动态规划的核心与前沿研究课题——大规模连续空间Markov决策过程的值函数与策略逼近问题，对近年来取得的研究进展进行了充分讨论。本书大多数理论、算法与实验结果都是作者近年来在研究工作中取得的成果。

(2) 多学科交叉。增强学习与近似动态规划的研究涉及机器学习、运筹学、智能控制、机器人学等多个学科领域，具有较强的学科交叉特点和较宽的学科覆盖面，对相关领域的学术创新起到了积极的促进作用。

(3) 理论与应用密切结合。本书在论述增强学习与近似动态规划理论和算法研究进展的同时，结合智能控制、机器人等领域的应用实例，在算法研究和理论分析的基础上，开展了大量的仿真和实验验证，有利于读者尽快把握理论和应用的结合点。

本书得到了国防科技大学贺汉根教授、胡德文教授的支持和鼓励，同时得到了国防科技大学无人车辆与机器学习项目组以及合作单位中南大学蔡自兴教授课题

组、吉林大学陈虹教授课题组的帮助和支持，作者指导的研究生张洪宇、张鹏程等协助参加了相关的实验工作，在此作者一并向他们表示感谢。自 2000 年以来，作者先后主持或参与多个国家自然科学基金项目，其中两个项目为国家自然科学基金重点项目（有关项目资助号分别为 90820302、60774076、60075020、60234030、60303012）。在此，特别向国家自然科学基金委员会致以衷心的感谢。本书的研究工作还得益于国际学术交流提供的良好学术氛围。国家自然科学基金和俄罗斯国家基础科学研究基金将本书的部分研究工作列为中俄国际合作项目，作者通过在俄罗斯科学院信息与自动化研究所的访问研究，以及与俄罗斯科学院 Adil Timofeev 教授的学术交流，进一步扩展了研究思路。在研究过程中，增强学习领域的奠基人之一、加拿大 Alberta 大学的 Sutton 教授以及 Littman 博士与作者多次通过互联网进行有关增强学习的学术讨论，使作者深受启发。国际人工智能研究基金（AI Access Foundation）、美国 Michigan 大学的增强学习学术资源库不仅及时提供了最新的研究资料，而且通过互联网建立了增强学习研究的活跃学术气氛。在此，作者向所有提供支持和帮助的国际同行表示由衷的感谢。

增强学习与近似动态规划的理论与应用还处在快速发展阶段，相关研究不断推陈出新。由于作者水平有限，本书难免存在不足之处，敬请读者批评指正。作者将充分吸取读者意见和建议，结合自身的科研工作，不断修改完善本书内容，为推动智能科学与技术相关领域的发展贡献绵薄之力。

徐　昕

2010 年 3 月 26 日

于国防科技大学

目　录

第1章 绪 论

1.1 引 言

“失败乃成功之母”。在人类历史上，这句至理名言激励了许多仁人志士在挫折面前冷静反省，总结经验，最终通过不懈努力而取得成功。从失败中总结经验教训成为人类获取知识和技能的一个重要途径。有关学习心理学[1]的进一步研究表明，从挫折与失败中积累经验和知识不仅仅是人类学习的重要方式，在高等哺乳动物中也发现了大量的类似行为现象。在20世纪初有关动物学习心理学的研究中，这种基于“尝试与失败”(trial-and-error)或称为“试错法”的学习方式得到了以Thorndike为代表的学习心理学家的重视，并开展了大量的学习理论和动物学习实验的研究，形成了“行为主义”这一学习心理学的主要学派[2]。目前，大量的理论和实验结果已证明了“试错法”学习是高等动物获取直接经验的一种基本方式。

近年来，机器学习作为一个重要的研究热点和前沿，一直是智能科学和智能计算研究的核心，因为任何一个没有学习能力的计算系统都很难被认为是一个真正的智能计算系统。美国航空航天局JPL实验室的科学家在*Science*(2001年9月)上撰文指出：机器学习对科学研究的整个过程正起到越来越大的支持作用，该领域在今后的若干年内将取得稳定而快速的发展[3]。作为一个具有丰富学科背景的研究领域，机器学习与统计学、心理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增强学习(reinforcement learning，又称为强化学习或再励学习)[4]理论和方法的产生和发展。增强学习的一个基本特点是强调与环境的交互，利用评价性的反馈信号实现序贯决策的优化，因此与其他的机器学习方法如监督学习(supervised learning，又称有导师学习)和无监督学习(unsupervised learning，又称无导师学习)存在重要的区别。

目前，学术界通常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)。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[5](如ID-3、C4.5决策树学习、AQ系列算法等)、以反向传播(BP)算法为代表的监督式神经网络学习、基于实例的学习(instance-based learning)等。监督学习的应用领域包括模式分类、数据挖掘、基于神经网络

的辨识与控制以及专家系统等。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，在无监督学习系统中输入仅包括环境的状态信息，也不存在与环境的交互。

与监督学习和无监督学习不同，增强学习基于动物学习心理学的有关原理，采用了人类和动物学习中的“尝试与失败”机制，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(reward/reinforcement signal，也称为回报或增强信号)，以极大化未来的回报为学习目标，如图 1.1 所示。增强学习由于不需要给定各种状态下的教师信号，因此在求解先验信息较少的复杂优化决策问题中具有广泛的应用前景。在人工智能的早期研究中，由于受到学习心理学研究的影响，增强学习一度成为机器学习的研究热点之一。但由于增强学习问题本身的困难性和其他种种原因，在 20 世纪七八十年代，机器学习的研究工作和成果主要集中于监督学习和无监督学习，有关增强学习的研究则经历了一段类似于神经网络的“低谷”时期。到 20 世纪 80 年代末，增强学习的研究又重新得到了学术界的重视，并呈现了与运筹学、控制理论、机器人学等交叉综合的特点[1,6]。

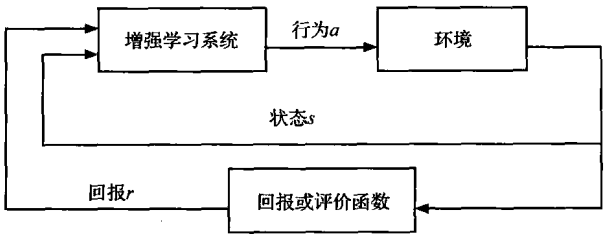


图 1.1 增强学习系统与环境的交互情况

目前，增强学习在理论和算法研究方面已取得了若干重要成果，并且显示了在求解复杂序贯(sequential)优化决策问题(通常建模为 Markov 决策问题)中的应用潜力[4,6~8]。但已有的理论研究结果仍然主要是针对小规模、离散状态空间问题，对大规模和连续空间的优化决策问题通常难以保证算法的收敛性，且存在学习效率不高的缺点。而现实世界的许多工程应用问题都具有大规模或连续的状态和决策空间，因此如何实现增强学习方法在大规模或连续状态和决策空间中的泛化(generalization)，提高增强学习在求解复杂问题时的学习效率，是决定增强学习方法能否得到广泛应用的关键。近十年来，有关近似动态规划(approximate dynamic programming，ADP)[9,10]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。在以电梯调度、网络路由控制等为代表的大规模优化决策应用中[4]，增强学习都显示了相对传统监督学习和数学规划方法的性能优势。在智能机器人系统、复杂不确定系统的优化控制器设计等领域，增强学习的应用也正在不断得到推广。

智能移动机器人是一类能够通过传感器感知环境和自身状态，实现在有障碍物的环境中面向目标的自主运动(navigation，也称为导航)，从而完成一定作业功能的机器人系统。目前，有关移动机器人导航方法已开展了许多研究，并获得了若干成功的应用[11~13]。由于移动机器人运动的环境多变，其导航控制方法涉及环境认知、优化决策、知识表示与获取等多项智能科学的关键问题，因此移动机器人的导航控制一直是机器人学和人工智能界的研究热点之一。在21世纪，移动机器人在工业、航天、建筑、服务业等领域的不断推广应用将对其导航控制技术提出越来越高的要求。

随着1997年美国的“Sojourner”火星探测机器人(图1.2)首次登上火星执行科学考察任务，利用移动机器人技术进行空间探测和开发成为21世纪全球各国开展科技和空间资源竞争的主要目标[14,15]。中国也针对这一趋势制定了以月球为近期目标的空间探测计划[16,17]。研究和发展中国的月球探测移动机器人技术，不但对于中国在激烈的空间技术和资源竞争中取得有利地位具有关键意义，同时对包括移动机器人导航控制在内的相关技术也具有巨大的促进作用。

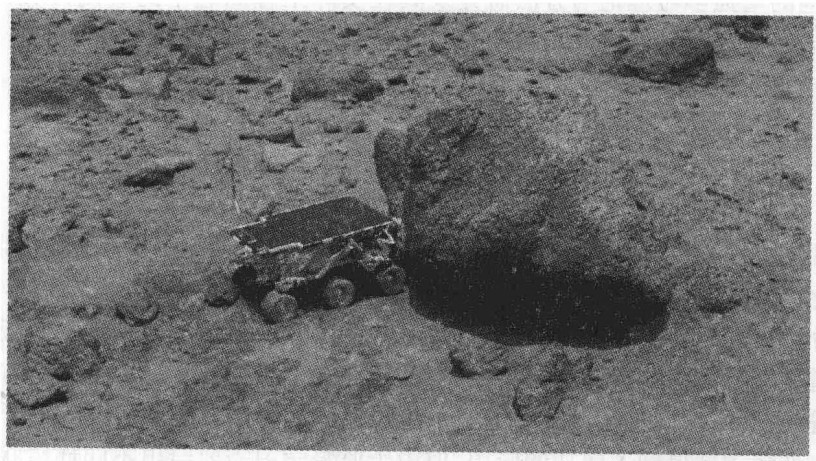


图1.2 美国的“Sojourner”火星探测机器人及其面临的复杂导航环境

移动机器人在月球和火星等外星球表面执行任务时，将面临未知环境的导航问题。在未知环境中的移动机器人导航控制技术成为月球和火星探测机器人的一项关键技术。在已有的移动机器人导航控制方法研究中，有关确定性环境中的导航方法已取得了大量的研究和应用成果。针对未知环境中的导航方法虽然也开展了一些研究，并提出了若干方法，但还有许多关键问题有待解决。这些关键问题主要包括未知环境中的移动机器人环境建模和定位、基于传感器的反应式(reactive)导航控制器的自适应和优化、环境理解、在线规划和决策等[18]。上述问题涉及人工智能和机器学习中的环境认知、知识表示和获取等有待进一步研究的重要领域。

另外，在移动机器人的导航控制系统中，运动控制是一项重要技术。由于移动机器人机电系统动力学建模的复杂性，以及移动机器人运动学特性的非完整特性，使得高性能的运动控制器设计成为移动机器人应用的一个难题。智能控制方法具有对模型信息的依赖较少、能够实现控制器的自适应和优化的特点，因此针对不确定模型的智能运动控制方法的研究是移动机器人导航控制的一项重要研究课题[19~24]。

近年来，随着增强学习算法和理论研究的深入，应用增强学习与近似动态规划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国际机器人领域研究的热点[25~30]。

本书的研究工作是在国家自然科学基金项目“基于核的增强学习与近似动态规划方法研究”(60774076)、“高速公路车辆智能驾驶中的关键科学问题研究”(90820302)与“增强学习泛化方法研究及其在移动机器人导航中的应用”(60075020)等多项国家自然科学基金的支持下，研究用于求解大规模和连续空间优化决策问题的增强学习算法和理论，以及增强学习方法在移动机器人导航与控制中的应用。本书的研究成果对于推动求解大规模和连续状态与行为空间 Markov 决策问题的增强学习理论与方法研究及其在实际工程问题中的应用，以及利用机器学习方法提高移动机器人系统的自主导航和控制性能，都具有重要的科学意义和工程应用价值。

1.2 增强学习与近似动态规划的研究概况

增强学习的基本思想与动物学习心理学有关“试错法”学习的研究密切相关，即强调在与环境中的交互中学习，通过环境对不同行为的评价性反馈信号来改变行为选择策略以实现学习目标。来自环境的评价性反馈信号通常称为回报或增强信号，增强学习系统的目标就是极大化(或极小化)期望回报信号。虽然监督学习方法如神经网络反向传播(BP)算法[31]、决策树学习算法[5]等的研究取得了大量成果，并在许多领域得到了成功的应用，但由于监督学习需要给出不同环境状态下的教师信号，因此限制了监督学习在复杂优化控制问题中的应用。无监督学习虽然不需要教师信号，但仅能完成模式聚类等功能。由于增强学习方法能够通过与环境的交互获得评价性反馈信号，并且实现行为决策的优化，因此在求解复杂的优化控制问题中具有更为广泛的应用价值。

基于增强学习的上述特点，在早期的人工智能研究中曾一度将增强学习作为一个重要的研究方向，如 Minsky 有关增强学习的博士论文[32]、Samuel 的跳棋学习程序[33]等，但后来由于各种因素特别是求解增强学习问题的困难性，在 20 世纪七八十年代人工智能和机器学习的研究主要面向监督学习和无监督学习方法。进入 20 世纪 90 年代，增强学习在理论和算法上通过与其他学科如运筹学、控制理论

的交叉综合，取得了若干突破性的研究成果，并且在机器人控制、优化调度等许多复杂优化决策问题中获得了成功的应用[4,6]。

1.2.1 增强学习研究的相关学科背景

增强学习在算法和理论研究方面的一个重要特点就是体现了多学科的交叉综合。增强学习的研究与动物学习心理学、运筹学、进化计算、自适应控制、神经网络等学科领域都具有密切的联系。

1. *动物学习心理学*

有关动物学习心理学的研究为增强学习的算法和理论提供了思想和哲学基础。在动物学习心理学的研究中，关于动物“试误”(trial)型学习的思想最早由Thorndike于1914年提出[2]，该思想的实质是强调行为的结果有优劣之分并成为行为选择的依据。Thorndike称这种规律为“效应定律”(law of effect)，并指出效应定律描述了增强性事件对于动物行为选择趋势的影响，即能够导致正的回报的行为选择概率将增加，而能够导致负回报的行为选择概率将降低。文献[1]指出，效应定律包括了“试误”型学习的两个主要方面，即选择性和联想性。进化学习中的自然选择具有选择性，但不具有联想性；监督学习则仅具有联想性而不具有选择性。另外，“效应定律”反映了增强学习的另两个重要特性，即搜索和记忆。

在动物学习心理学中与增强学习密切相关的另一个研究内容是时域差值(temporal-difference，或称为时间差分)理论[4,34]。所谓时域差值是指对同一个事件或变量在连续两个时刻观测的差值，这一概念来自于学习心理学中有关“次要增强器”(secondary reinforcers)的研究。在动物学习心理学中，次要增强器伴随主要增强信号如食物等的刺激，并且产生类似于主要增强信号的行为增强作用[1]。在早期的增强学习研究中，时域差值学习方法是一个重要研究内容，如Samuel[33]的跳棋学习程序中就采用了时域差值学习的思想。在近十年来的增强学习算法和理论研究中，时域差值学习理论和算法也同样具有基础性的地位。

2. *运筹学*

运筹学是与增强学习紧密联系的另一个学科。运筹学中有关Markov决策过程[35](Markov decision process，MDP)和动态规划的算法和理论为增强学习的研究提供了数学模型和算法理论基础，其中主要包括Bellman的最优性原理和Bellman方程、值迭代、策略迭代等动态规划算法[35,36]。动态规划和增强学习方法的联系由Minsky[32]在分析Sameul的跳棋学习程序时首先提出，并逐渐得到了普遍重视。许多增强学习算法如Q-学习算法[37]等都可以看做无模型的自适应动态规划算法。增强学习和动态规划两个学科的交叉综合成为推动增强学习算法和理论

研究的重要因素。近年来，求解大规模状态空间的动态规划方法如值函数逼近方法[38]等在增强学习领域也得到了广泛的重视。

3. 进化计算

进化计算(evolutionary computation)[39,40]是基于生物界的自然选择和基因遗传原理实现的一类优化算法，并被广泛应用于求解机器学习问题。目前，进化计算在算法和理论上已取得了大量研究成果，形成了遗传算法[39]、进化策略[40]和进化规划[41]三个主要的分支，并且在组合优化、自动程序设计、机器学习等领域[42]获得了成功的应用。虽然早期的进化计算与增强学习的研究相互独立，但随着研究的深入，进化计算方法在求解增强学习问题中的应用逐步得到重视。对于利用评价性反馈的增强学习问题，进化计算方法能够通过将回报信号映射为个体的适应度进行求解。在应用进化计算方法求解增强学习问题时，一个关键问题是如何对延迟回报进行时间信用分配(temporal credit assignment)。Holland 的分类器系统(classifier system)[43]对上述问题进行了开拓性地研究，在该算法中体现了时域差值学习的思想。近年来，求解增强学习问题的进化增强学习方法成为一个重要的研究课题。文献[44]对进化增强学习算法进行了深入研究。如何综合利用两种方法的优点实现多策略的高效增强学习系统是一个值得研究的课题。

4. 自适应控制

自适应控制是控制理论的一个重要分支，研究模型未知或不确定对象的控制问题。在自适应控制中，按照是否对模型进行在线估计，可以分为直接自适应控制方法和间接自适应控制方法两类。其中，直接自适应控制方法不建立对象的显式估计模型，而直接通过调节控制器参数实现闭环自适应控制；间接自适应控制方法则基于对象模型的在线辨识，对控制器的参数进行调节。文献[45]对增强学习作为一类直接自适应最优控制方法的特性进行了分析和研究，指出了增强学习与自适应控制理论的联系。与动态规划不同，增强学习不需要 Markov 决策过程的状态转移模型，而直接根据与环境的交互信息实现 Markov 决策过程的优化控制。在自适应控制中得到普遍关注的辨识与控制的关系类似于增强学习中行为探索(exploration)和利用(exploitation)的关系。行为探索是指不采用当前策略的随机化行为搜索，与自适应控制的辨识信号输入相对应；行为利用是指采用当前策略进行行为选择的优化，对应自适应控制的控制器参数优化设计。

5. 神经网络

神经网络的研究起源于对人类大脑的神经生理学和神经心理学的研究，目前已取得了丰富的研究成果，其中包括多种神经网络的结构模型和学习算法。在神