HZ BOOKS 华章教育
CRC Press Taylor & Francis Group
智能科学与技术丛书

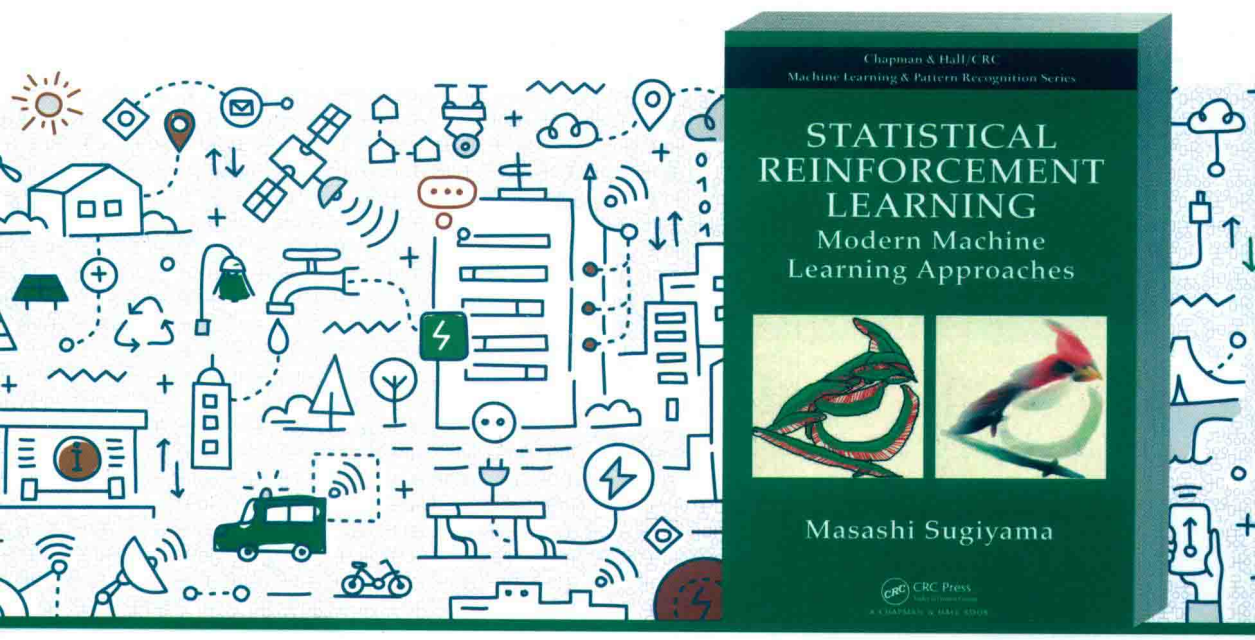
Statistical Reinforcement Learning

Modern Machine Learning Approaches

统计强化学习

现代机器学习方法

[日] 杉山将（Masashi Sugiyama） 东京大学 ◎ 著
高阳 南京大学 ◎ 等译



机械工业出版社
China Machine Press

智能科学与技术丛书

Statistical Reinforcement Learning

Modern Machine Learning Approaches

统计强化学习

现代机器学习方法

[日] 杉山将（Masashi Sugiyama） 东京大学 ◎ 著

高阳 南京大学 ◎ 等译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

统计强化学习：现代机器学习方法 /（日）杉山将著；高阳等译 . —北京：机械工业出版社，2019.4
（智能科学与技术丛书）
书名原文：Statistical Reinforcement Learning: Modern Machine Learning Approaches

ISBN 978-7-111-62245-1

Ⅰ. 统… Ⅱ. ①杉… ②高… Ⅲ. 机器学习 Ⅳ. TP181

中国版本图书馆 CIP 数据核字（2019）第 049678 号

本书版权登记号：图字 01-2017-2733

本书将统计学习和强化学习相结合，对强化学习函数估计中的基函数设计、样本重用以及策略搜索、模型估计等做了深入浅出的介绍。全书共 11 章，分为四部分：第一部分（第 1 章）介绍了强化学习的基本知识；第二部分（第 2 ～ 6 章）介绍了模型无关策略迭代的知识；第三部分（第 7 ～ 9 章）介绍了模型无关策略搜索的知识；第四部分（第 10 ～ 11 章）介绍了基于模型的强化学习。本书适合从事人工智能和机器学习研究和应用的专家学者、技术人员、研究生阅读。

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：朱秀英　　责任校对：殷　虹
印　　刷：北京市兆成印刷有限责任公司　　版　　次：2019 年 5 月第 1 版第 1 次印刷
开　　本：185mm × 260mm　1/16　　印　　张：12.5
书　　号：ISBN 978-7-111-62245-1　　定　　价：79.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991　88379833　　投稿热线：（010）88379604
购书热线：（010）68326294　　读者信箱：hzjsj@hzbook.com

随着 Google 公司 DeepMind 团队提出 DQN 技术，并研发了 Alpha Go 击败围棋世界冠军李世石、柯洁等人，强化学习技术逐渐成为人工智能和机器学习技术的研究热点和显学，从技术的发展期进入了爆炸期，各种深度强化学习技术层出不穷。

然而，自 20 世纪五六十年代的技术萌芽，到八九十年代的理论奠定，再到如今的技术爆炸，强化学习技术经历了漫长的积淀过程。尽管深度强化学习技术是“通用人工智能”的代表性技术，但强化学习仍有诸多子领域有待发展，例如多智能体强化学习技术、逻辑强化学习技术、强化学习迁移技术等。当研究者跳出棋类博弈和机器人控制等经典智能任务，试图在军事、经济、金融等领域或实际工程中应用强化学习时，强化学习仍面临表示困难、收敛慢等诸多难题。这给强化学习技术带来了挑战和新的活力。

本书是为数不多的强化学习专业书籍，作者是日本知名的机器学习学者杉山将先生。在首届中国计算机学会国际人工智能会议(CCF-ICAI2018，济南)期间，译者和杉山将先生进行了交谈。他非常高兴中文版的诞生，并期待能对中国的读者有所帮助。本书更侧重于强化学习的基础，而非目前热门的深度强化学习技术。本书从模型无关策略迭代、模型无关策略搜索、模型相关强化学习三个技术路线角度，对强化学习函数估计中的基函数设计、样本重用以及策略搜索、模型估计等做了深入浅出的介绍。特别是本书结合了统计学习的诸多方法对强化学习的相关技术进行介绍，给人以耳目一新的感觉。

南京大学计算机科学与技术系推理与学习研究组董绍康、吴章凯、陈佳瑞、朱枝睿、张剑、刘艳芳、顾峥、秦铁鑫、季雯、董传奇、黄中豪等研究生参与了本书的部分翻译工作。在为期近一年的翻译过程中，虽然我们已经对译稿进行仔细校对，查阅了大量相关资料，使译文尽可能符合中文习惯和保持术语的一致性，但由于本书涉及的范围非常广泛，错误或不当之处仍难以完全避免，敬请各位读者和同行专家谅解，诚挚希望读者将相关意见、建议发送到电子邮箱 gaoy@nju.edu.cn 与我们联系。

本书适合从事人工智能和机器学习研究和应用的专家学者、技术人员、研究生阅读。最后，特别感谢机械工业出版社华章公司的朱秀英编辑，没有她的信任、耐心与支持，本书不可能顺利出版。

译　者

2018 年 12 月 15 日于南京

如果没有一个知识渊博的老师给出明确的指示，智能体如何从经验中学习呢？强化学习是机器学习中的一个子领域，它通过将通用奖赏信号与其过去的动作相关联来研究智能体如何学习最佳行为。该学科利用了行为心理学、经济学、控制论、运筹学和其他不同领域的关键思想，以模拟学习过程。在强化学习中，环境通常被建模为马尔可夫决策过程，其向智能体提供瞬时奖赏和状态信息。但是，智能体无法接触到环境结构的变化，需要学习如何选择适当的行为以最大限度地提高其整体奖赏。

Masashi Sugiyama 教授的这本书从一个全新的现代角度讲述了强化学习算法。该书侧重于估算强化学习参数的统计特性，讲述了各种学习场景中的多种不同方法。这些算法分为不明确模拟环境动态的模型无关方法和为环境构建描述性过程模型的基于模型的方法。在每一个类别中，都有估计值函数的策略迭代算法，以及直接操纵策略参数的策略搜索算法。

针对每一个不同的强化学习场景，本书都详细列出了相关的优化问题。对每一种情况都进行了仔细的分析，重点是了解所得估计量和学习参数的统计特性。每一章都包含这些算法应用的示例，并对不同技术进行定量比较。这些例子来自各种实际问题，包括机器人运动控制和东方山水画。

总之，本书为强化学习算法引入了一种发人深省的统计处理方法，反映了作者在该领域的工作和研究状态，为快速发展的机器学习文献补充了最新的资料。初学者和经验丰富的研究人员都会发现此书是理解最新强化学习技术的重要来源。

Daniel D. Lee
美国宾夕法尼亚大学
工程和应用科学学院
GRASP 实验室

前　言

Statistical Reinforcement Learning: Modern Machine Learning Approaches

在即将到来的大数据时代，统计学与机器学习正成为数据挖掘不可或缺的工具。根据数据分析的类型，机器学习方法分为三类：

- **监督学习**：给定输入和输出的数据，监督学习的目标是分析输入、输出数据之间的关系。监督学习典型的任务包括回归(预测真实取值)、分类(预测类别)以及排序(预测顺序)。监督学习是最常用的数据分析工具，并且已经在统计学领域被研究了很长时间。监督学习在机器学习中近期的趋势是利用输入、输出数据的辅助信息来进一步改善预测的精度。例如，半监督学习利用额外的输入数据，迁移学习借用来自其他相似学习任务的数据，多任务学习同时解决多个相关学习任务。
- **无监督学习**：仅给定输入数据，无监督学习的目标是在数据中找到有用的东西。由于这种模糊的定义，无监督学习研究往往比监督学习更具特色。然而，由于其自动化以及廉价的特性，无监督学习被认为是数据挖掘中最重要的工具之一。无监督学习典型的任务包括聚类(根据数据的相似性进行数据分组)、密度估计(估计数据背后的概率分布)、异常检测(从数据中删除异常值)、数据可视化(将数据的维度降到1～3维)和盲源分离(从混合数据中提取原始源信号)。此外，无监督学习方法有时被用作监督学习中数据预处理的工具。
- **强化学习**：监督学习是一种合理的方法，但收集输入、输出数据通常过于昂贵。无监督学习的执行成本低廉，但往往是临时性的。强化学习介于监督学习和无监督学习之间——没有提供明确的监督(输出数据)，但我们仍然想学习数据背后的输入、输出关系。强化学习不是输出数据，而是利用奖赏来评估所预测的输出的有效性。提供诸如奖赏之类的隐性监督通常比提供明确监督更容易，成本更低，因此强化学习可以成为现代数据分析的重要方法。在强化学习的框架中也使用各种监督和无监督学习技术。

本书致力于从现代机器学习的角度介绍统计强化学习的基本概念和实用算法。还提供了各种图解说明示例——这些示例主要来自机器人领域，帮助读者理解强化学习技术的直观性和实用性。目标读者是计算机科学和应用统计学的研究生以及相关领域的研究人员和工程师。假设读者具备概率和统计学、线性代数以及初等微积分的基础

知识。

机器学习是一个快速发展的科学领域，希望本书能够帮助读者了解强化学习中的各种激动人心的话题，激发读者对机器学习的兴趣。请浏览我们的网站：http://www.ms.k.u-tokyo.ac.jp。

致谢

感谢合作者 Hirotaka Hachiya、Sethu Vijayakumar、Jan Peters、Jun Morimoto、Zhao Tingting、Ning Xie、Voot Tangkaratt、Tetsuro Morimura 和 Norikazu Sugimoto 激动人心的创意讨论。感谢 MEXT KAKENHI（17700142、18300057、20680007、23120004、23300069、25700022 和 26280054）、大川基金会、欧盟 Erasmus Mundus 奖学金、AOARD、SCAT、JST PRESTO 计划以及 FIRST 计划的支持。

Masashi Sugiyama
日本东京大学

作者简介

Statistical Reinforcement Learning: Modern Machine Learning Approaches

Masashi Sugiyama 于 1974 年出生于日本大阪。他分别于 1997 年、1999 年和 2001 年获得日本东京工业大学计算机科学学士、硕士和工程博士学位。2001 年，他被任命为东京工业大学的助理教授，并于 2003 年晋升为副教授，2014 年升任东京大学教授。

他获得了 Alexander von Humboldt 基金会研究奖学金，并于 2003 年至 2004 年在德国柏林弗劳恩霍夫研究所做研究。2006 年，他获得了欧洲委员会计划 Erasmus Mundus 奖学金，并在苏格兰爱丁堡大学做研究。他于 2007 年获得 IBM 颁发的学院奖，以表彰他对非平稳机器学习的贡献；2011 年获得日本信息处理协会颁发的 Nagao 特别研究员奖以及教育、文化、体育、科学和技术部部长颁发的青年科学家奖，以表彰他对机器学习密度比范型的贡献。

他的研究兴趣包括机器学习和数据挖掘的理论和算法，以及信号处理、图像处理和机器人控制等应用领域。他出版了《Density Ratio Estimation in Machine Learning》(剑桥大学出版社，2012)和《Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation》(麻省理工学院出版社，2012)。

第一部分

Statistical Reinforcement Learning: Modern Machine Learning Approaches

简　介

强化学习介绍

强化学习致力于控制一个计算机智能体，使之在未知环境中完成任务目标。

本章我们首先在 1.1 节中给出强化学习的非正规概览。然后在 1.2 节中给出强化学习的正式描述。最后在 1.3 节中对本书内容进行概述。

1.1 强化学习

图 1-1 中给出强化学习的框图。在一个未知环境中(如迷宫)，一个计算机智能体(如机器人)基于其自身的控制策略进行动作(如行走)。然后，其环境状态被更新(例如向前移动)，并且这个动作用某种“奖赏”进行评估(如赞美、不表态或责骂)。智能体通过这样与环境的交互，在没有明确指导的情况下，被训练完成某种特定的任务(如走出迷宫)。因此，智能体并不是被训练改善一个短期的性能(如梯度式地逼近一个迷宫的出口)，而是被训练优化长期的成果(如成功地走出迷宫)。

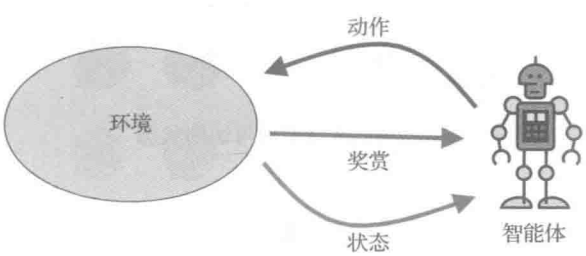


图 1-1 强化学习

强化学习问题包含不同的技术组件，如状态、动作、转移、奖赏、策略、值等。在进入数学细节之前(将在 1.2 节中给出)，我们将在这里通过解释强化学习问题，直观地理解这些概念。

让我们考察一个迷宫问题(图 1-2)，迷宫中有一个机器人，我们希望在不明确告诉它往哪个方向走的情况下，指导它到达目标。状态就是迷宫中机器人能够访问的位置。在图 1-3 的例子中，迷宫中有 21 个状态。动作就是机器人可以移动的方向。在图 1-4 的例子中，有 4 个动作，分别对应着向东、向西、向南、向北移动。状态和动

作是定义强化学习问题的最基本元素。

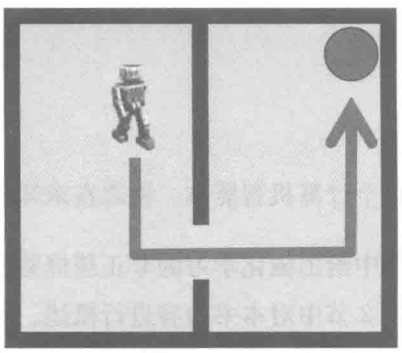


图 1-2　一个迷宫问题，我们希望指导这个机器人到达目标

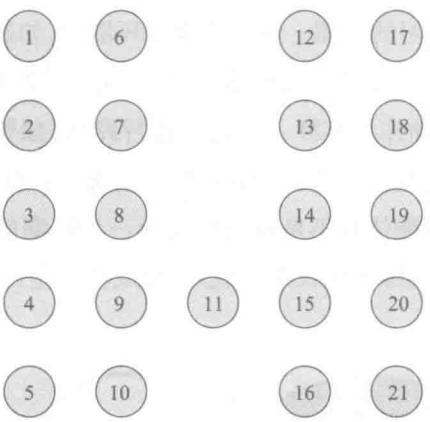


图 1-3　迷宫中的可访问状态

图 1-4　动作是机器人的可能移动

*转移*指定了在动作的作用下，状态之间是如何相互连接的(图 1-5)。因此，知道了转移就相当于知道了迷宫的地图。*奖赏*指定了在某个状态下，采用某个动作转移到

另一状态时机器人获得的收益/代价。在迷宫例子中，机器人到达目标时将收到一个正的奖赏。更具体来说，当在状态 12 采用动作“向东”转移到状态 17 时，或者当在状态 18 采用动作“向北”转移到状态 17 时，就会提供一个正的奖赏(图 1-6)。因此，直观地了解奖赏意味着知道目标状态的位置。为了强调在采取动作并转换到下一个状态后立即给予机器人智能体奖赏这一事实，我们称之为瞬时奖赏。

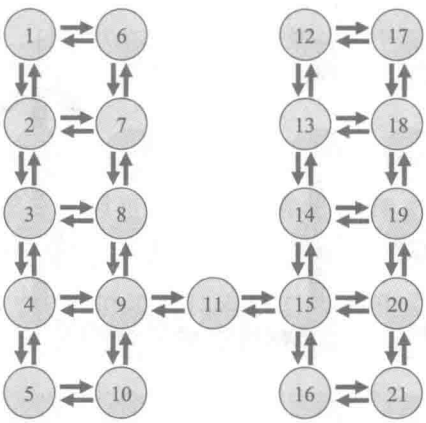


图 1-5　转移指定了在动作的作用下，状态之间是如何相互连接的。因此，知道了转移直觉上就相当于知道了迷宫的地图

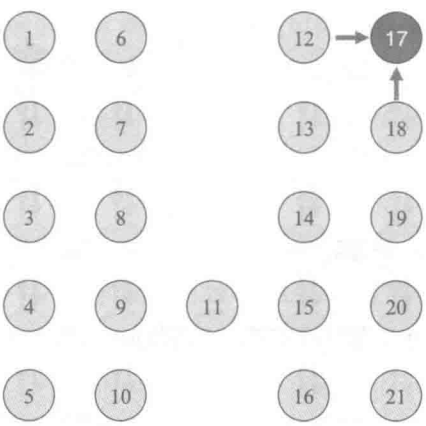


图 1-6　当机器人到达终点时就会得到正奖赏。因此奖赏指示了目标的位置

在上述设置中，强化学习的目标是寻找一个用于控制机器人的策略，该策略使机器人在长期运行时能够获得最大的奖赏量。这里，策略特指机器人在每个状态采用的动作(图 1-7)。通过策略，指定了从初始状态到终结状态机器人所采用的状态和动作序列。这样的序列被称为轨迹(还是参见图 1-7)。沿着一条轨迹的瞬时奖赏和被称为回报。在现实生活中，在一个较远未来所获得的奖赏通常会被打折，其原因是人们更