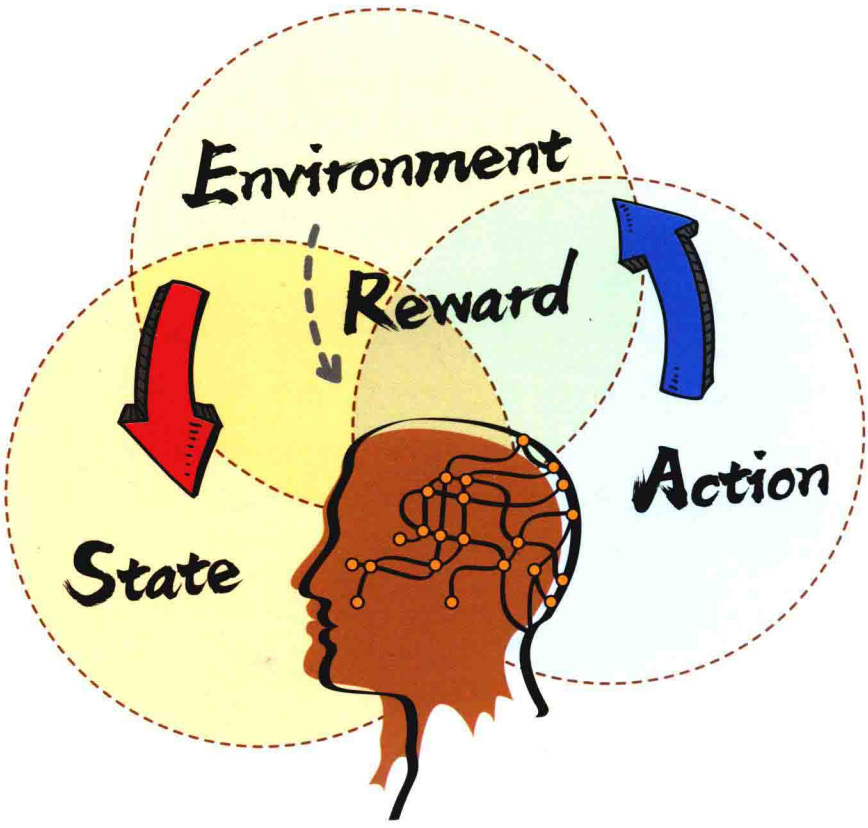
Broadview®
www.broadview.com.cn

博文视点AI系列



深入浅出强化学习

原理入门

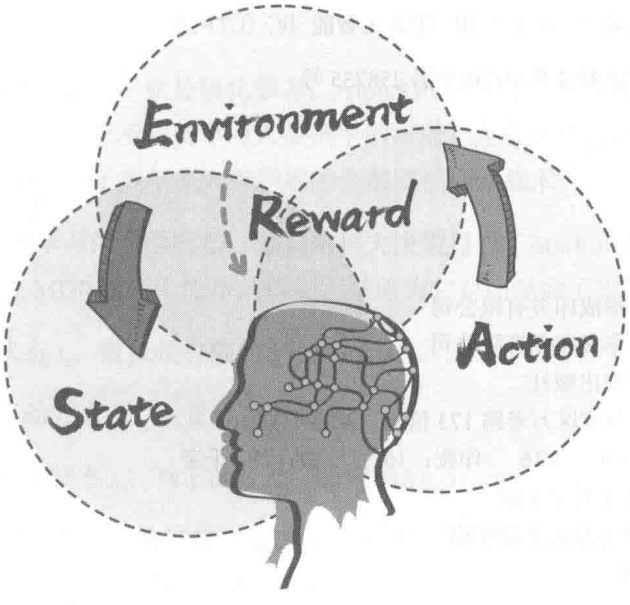
郭　宪　方勇纯　编著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

博文视点AI系列

深入浅出强化学习
原理入门

郭　宪　方勇纯　编著



電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书用通俗易懂的语言深入浅出地介绍了强化学习的基本原理，覆盖了传统的强化学习基本方法和当前炙手可热的深度强化学习方法。开篇从最基本的马尔科夫决策过程入手，将强化学习问题纳入到严谨的数学框架中，接着阐述了解决此类问题最基本的方法——动态规划方法，并从中总结出解决强化学习问题的基本思路：交互迭代策略评估和策略改善。基于这个思路，分别介绍了基于值函数的强化学习方法和基于直接策略搜索的强化学习方法。最后介绍了逆向强化学习方法和近年具有代表性、比较前沿的强化学习方法。

除了系统地介绍基本理论，书中还介绍了相应的数学基础和编程实例。因此，本书既适合零基础的人员入门学习、也适合相关科研人员作为研究参考。

图书在版编目（CIP）数据

深入浅出强化学习：原理入门 / 郭宪，方勇纯编著. —北京：电子工业出版社，2018.1
ISBN 978-7-121-32918-0

Ⅰ. ①深… Ⅱ. ①郭… ②方… Ⅲ. ①人工智能 Ⅳ. ①TP18

中国版本图书馆 CIP 数据核字(2017)第 258235 号

责任编辑：刘　皎
印　　刷：三河市华成印务有限公司
装　　订：三河市华成印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：720×1000　1/16　印张：16　字数：284 千字
版　　次：2018 年 1 月第 1 版
印　　次：2018 年 1 月第 1 次印刷
定　　价：79.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

推荐序一

强化学习是机器学习的一个重要分支，它试图解决决策优化的问题。所谓决策优化，是指面对特定状态（State，S），采取什么行动方案（Action，A），才能使收益最大（Reward，R）。很多问题都与决策优化有关，比如下棋、投资、课程安排、驾车，动作模仿等。

AlphaGo 的核心算法，就是强化学习。AlphaGo 不仅稳操胜券地战胜了当今世界所有人类高手，而且甚至不需要学习人类棋手的棋谱，完全靠自己摸索，就在短短几天内，发现并超越了一千多年来人类积累的全部围棋战略战术。

最简单的强化学习的数学模型，是马尔科夫决策过程（Markov Decision Process，MDP）。之所以说 MDP 是一个简单的模型，是因为它对问题做了很多限制。

1. 面对的状态 s_t，数量是有限的。

2. 采取的行动方案 a_t，数量也是有限的。

3. 对应于特定状态 s_t，当下的收益 r_t 是明确的。

4. 在某一个时刻 t，采取了行动方案 a_t，状态从当前的 s_t 转换成下一个状态 s_{t+1}。下一个状态有多种可能，记为 s_{t+1}^i，$i = 1 \ldots n$。

换句话说，面对局面 s_t，采取行动 a_t，下一个状态是 s_{t+1}^i，不是确定的，而是概率的，状态转换概率，记为 $P(s_{t+1}^i \mid s_t, a_t)$。但是状态转换只依赖于当前状态 s_t，而与先前的状态 $s_{t-1}, s_{t-2} \ldots$ 无关。

解决马尔科夫决策过程问题的常用的算法，是动态规划（Dynamic Programming）。

对马尔科夫决策过程的各项限制，不断放松，研究相应的算法，是强化学习的目标。例如对状态 s_t 放松限制：

1. 假如状态 s_t 的数量，虽然有限，但是数量巨大，如何降低动态规划算法的计算成本；

2. 假如状态 s_t 的数量是无限的，现有动态规划算法失效，如何改进算法；

3. 假如状态 s_t 的数量不仅是无限的，而且取值不是离散的，而是连续的，如何改进算法；

4. 假如状态 s_t 不能被完全观察到，只能被部分观察到，剩余部分被遮挡或缺失，如何改进算法；

5. 假如状态 s_t 完全不能被观察到，只能通过其他现象猜测潜在的状态，如何改进算法。

放松限制，就是提升问题难度。在很多情况下，强化学习的目标，不是寻找绝对的最优解，而是寻找相对满意的次优解。

强化学习的演进，有两个轴线：一个是不断挑战更难的问题，不断从次优解向最优解逼近；另一个是在不严重影响算法精度的前提下，不断降低算法的计算成本。

此书的叙述线索非常清晰，从最简单的解决马尔科夫决策过程的动态规划算法，一路讲解到最前沿的深度强化学习算法（Deep Q Network，DQN），单刀直入，全无枝枝蔓蔓之感。不仅解释数学原理，而且注重编程实践。同时，行文深入浅出，通俗易懂。

将本书与 Richard Sutton 和 Andrew Barto 合著的经典著作 *Reinforcement Learning: An Introduction, Second Edition* 相比，Sutton 和 Barto 在内容上更注重全面，覆盖了强化学习各个分支的研究成果；而本书更强调实用，是值得精读的教材。

邓侃

PhD of Robotics Institute, School of Computer Science, Carnegie Mellon University,

前 Oracle 主任架构师、前百度网页搜索部高级总监、

北京大数医达科技有限公司创始人

推荐序二

强化学习又称为增强学习或再励学习（Reinforcement learning），是 AlphaGo、AlphaGo Zero 等人工智能软件的核心技术。近年来，随着高性能计算、大数据和深度学习技术的突飞猛进，强化学习算法及其应用也得到更为广泛的关注和更加快速的发展。尤其是强化学习与深度学习相结合而发展起来的深度强化学习技术已经取得若干突破性进展。AlphaGo 与人类顶级棋手之间的对弈，使得深度强化学习技术在学术界和工业界得到了更为广泛的关注。强化学习不仅在计算机博弈中取得巨大成功，而且在机器人控制、汽车智能驾驶、人机对话、过程优化决策与控制等领域，也被认为是实现高级人工智能最有潜力的方法。

本人在多年从事强化学习与近似动态规划理论和应用的研究过程中，力求不断提升强化学习算法的快速收敛性和泛化性能，并且将强化学习新理论和新算法应用于移动机器人和自主驾驶车辆等领域，为智能移动机器人和自主驾驶车辆在复杂、不确定条件下的自主优化决策和自学习控制提供高效的技术手段。今后，随着相关理论和技术的不断进步，强化学习技术在智能机器人和自主驾驶车辆、复杂生产过程的优化决策与控制、天空与海洋无人系统等领域的应用将很快会有新的突破。

强化学习的思想从 20 世纪初便被提出来了，经过将近一个世纪的发展，强化学习与心理学、运筹学、智能控制、优化理论、计算智能、认知科学等学科有着密切的联系，是一个典型的多学科交叉领域。来自不同学科的概念和思想使得初学者学习和了解强化学习存在较大的困难。郭宪博士和方勇纯教授的这本《深入浅出强化学习：

推荐序四

AlphaGo 的诞生掀起了（深度）强化学习技术的一轮热潮，该方向已成为人工智能领域最热门的方向之一，由于其通用性而备受各个应用领域推崇，从端对端控制、机器人手臂控制，到推荐系统、自然语言对话系统等。（深度）强化学习也被 OpenAI 等公司认为是实现通用人工智能的重要途径。

然而目前强化学习中文资料相对零散，缺少兼具系统性和前沿性的强化学习教学及科研资料。郭博士的《深入浅出强化学习：原理入门》这本书恰好填补了这一空白。本书根据郭博士在知乎的强化学习专栏内容整理而成，条分缕析、通俗易懂，既对强化学习基础知识做了全方面“深入浅出”的讲述，又涵盖了深度强化学习领域一系列最新的前沿技术。因此它无论是对强化学习的入门者，还是强化学习领域研究人员和工程师，都是一本很好的推荐读物，相信不同的读者都会从中获益。

郝建业

天津大学副教授、天津市青年千人、天津大学“北洋青年学者”

推荐序五

受行为主义心理学研究启发，在机器学习领域中产生了一种交互式学习方法的分支，这便是强化学习，又称为增强学习。强化学习模拟的是人类的一种学习方式，在执行某个动作或决策后根据执行效果来获得奖励，通过不断与环境的交互进行学习，最终达到目标。强化学习概念早在上世纪就已经提出，在计算机领域，第一个增强学习问题是利用奖惩手段学习迷宫策略。然而，直到 2016 年 AlphaGo 对决李世石一战成名后，强化学习的概念才真正广为人知。强化学习主要应用于众多带有交互性和决策性问题，比如博弈、游戏、机器人、人机对话等，这些问题是常用的监督学习和非监督学习方法无法很好处理的。

本人一直从事移动机器人、机器视觉和机器学习领域的研究，以及人工智能课程的教学。此前，为了解决人形机器人斜坡稳定行走问题，在查阅深度学习相关资料的过程中，在网上偶然看到郭宪博士开辟的强化学习专栏，读后很有收获。现在他将专栏文章整理编著成书，重新按知识层次进行编排和补充，对于读者学习更有帮助。

本书覆盖了强化学习最基本的概念和算法。在基于值函数的强化学习方法中，介绍了蒙特卡罗法、时间差分法和值函数逼近法。在基于直接策略搜索的强化学习方法中，介绍了策略梯度法、置信域策略法、确定性策略搜索法和引导策略搜索。在强化学习的前沿部分，介绍了逆向强化学习、深度强化学习和 PILCO 等。除了深度学习算法本身，书中还对涉及的基础知识，如概率学基础、马尔科夫决策过程、线性方程组的数值求解方法、函数逼近方法、信息论中熵和相对熵的概念等也做了详细的说明。

本书非常适合科技人员、高等学校师生和感兴趣人员作为入门强化学习的读物，也可作为相关研究和教学的参考书。

本书内容深入浅出、文字简单明了，采用了丰富的实例，让读者易读、易懂。同时配有习题和代码详解，能有效提升读者对理论知识的理解，帮助读者运用理论解决实际问题。建议读者跟随书中的示例和代码（https://github.com/gxnk/reinforcement-learning-code）来实现和验证相关强化学习算法，并可同时关注作者的知乎专栏（https://zhuanlan.zhihu.com/sharerl）以便更好地互动和探讨相关细节。

陈白帆

中南大学副教授 湖南省自兴人工智能研究院副院长

前　言

2017 年 5 月，AlphaGo 击败世界围棋冠军柯洁，标志着人工智能进入一个新的阶段。AlphaGo 背后的核心算法——深度强化学习——成为继深度学习之后广泛受人关注的前沿热点。与深度学习相比，深度强化学习具有更宽泛的应用背景，可应用于机器人、游戏、自然语言处理、图像处理、视频处理等领域。深度强化学习算法被认为是最有可能实现通用人工智能计算的方法。不过，由于深度强化学习算法融合了深度学习、统计、信息学、运筹学、概率论、优化等多个学科的内容，因此强化学习的入门门槛比较高，并且，到目前为止，市面上没有一本零基础全面介绍强化学习算法的书籍。

本书是笔者在南开大学计算机与控制工程学院做博士后期间，每周在课题组内讲解强化学习知识的讲义合集。在学习强化学习基本理论的时候，我深深地感受到强化学习理论中的很多概念和公式都很难理解。经过大量资料和文献的查阅并终于理解一个全新的概念时，内心涌现的那种喜悦和兴奋，鼓动着我将这些知识分享给大家。为此，我在知乎开辟了《强化学习知识大讲堂》专栏，并基本保持了每周一次更新的速度。该专栏得到大家的关注，很多知友反映受益良多，本书的雏形正是来源于此。在成书时，考虑到书的逻辑性和完整性，又添加了很多数学基础和实例讲解。希望本书能帮助更多的人入门强化学习，开启自己的人工智能之旅。

在写作过程中，博士后合作导师方勇纯教授给了大量的建议，包括书的整体结构、每一章的讲述方式，甚至每个标题的选择。写作后，方老师细致地审阅了全文，给出

了详细的批注，并多次当面指导书稿的修改。正是因为方老师的耐心指导与辛勤付出，本书才得以顺利完成。

同时，非常感谢组内的研究生丁杰、朱威和赵铭慧三位同学，通过与他们的交流，我学会了如何更明晰地讲解一个概念。本书的很多讲解方式都是在与他们的交流中产生的。

本书在写作过程中参考了很多文献资料，这些文献资料是无数科研工作者们日日夜夜奋斗的成果。本书对这些成果进行加工并形成了一套自成体系的原理入门教程。可以说没有这些科研工作者们的丰硕成果就没有今天蓬勃发展的人工智能，也就没有这本书，在此对这些科学工作者们表示由衷的敬意。

本书前六章的内容及组织思路很大部分参考了 David Silver 的网络课程，同时参考了强化学习鼻祖 Richard S. Sutton 等人所著的 *Reinforcement Learning: An Introduction*，在此向 Silver 和 Sutton 致敬。

本书第 8 章介绍了置信域强化学习算法，主要参考了 John Shulman 的博士论文，在此向 John Shulman 博士及其导师 Pieter Abbeel 致敬。第 10 章主要介绍了 Sergey Levine 博士的工作，在此对其表示感谢。在强化学习前沿部分，本书介绍了最近一年该领域很优秀的研究工作，如 Donoghue 的组合策略梯度和 Qlearning 方法，Tamar 的值迭代网络，Deisenroth 的 PILCO 方法和 McAllister 的 PILCO 扩展方法，在此对这些作者表示感谢。当然，本书还介绍了很多其他科研工作者的工作，在此对他们一并致谢。

本书阐述的主要是前人提出的强化学习算法的基本理论，并没有介绍笔者个人的工作，但在此仍然要感谢目前我负责的两项基金的支持：国家自然科学基金青年基金（61603200）和中国博士后基金面上项目（2016M601256）。这两个项目都和强化学习有关，本书也可看成是这两个项目的前期调研和积累。关于更多笔者个人的工作，留待以后再与大家分享。

由于个人水平有限，书稿中难免有错误，欢迎各位同行和读者批评指正。我的个人邮箱是 guoxiansia@163.com，如有疑问，欢迎咨询。

最后，感谢我的家人，感谢我的爱人王凯女士，感谢她长时间对我的理解和支持，没有她的帮助，我一无所有，一事无成。这本书献给她。

郭宪

2017 年 11 月

目　录

绪论

1.1 这是一本什么书

这是一本人人都可以读懂的书。唐代大诗人白居易写诗定稿的标准是“老妪能解”，也就是说只有连市井中的老妇人都能听懂的诗才是好诗。本书力求做到这一点。不过，真正做到“老妪能解”的程度还是有困难的。因为强化学习是集数学、工程学、计算机科学、心理学、神经科学于一身的交叉学科。力图将这门“深奥”的学科讲明白，是写作本书的目的。

本书讲的是强化学习算法，什么是强化学习算法呢，它离我们有多远？2016 年和 2017 年最具影响力的 AlphaGo 大胜世界围棋冠军李世石和柯洁事件，其核心算法就用到了强化学习算法。相信很多人想了解或者转行研究强化学习算法或多或少都跟这两场赛事有联系。如今，强化学习继深度学习之后，成为学术界和工业界追捧的热点。从目前的形式看，强化学习正在各行各业开花结果，前途一片大好。然而，强化学习的入门却很难，明明知道它是一座“金山”，可是由于总不能入门，只能望“金山”而兴叹了。另外，市面上关于强化学习的中文书并不多，即便有，翻开几页出现的各种专业术语，一下就把人搞懵了。本来下定决心要啃下这块硬骨头的，可是啃了几天发现，越啃越痛苦，连牙都咯掉了，肉渣还没吃到。本书下决心不给大家吃骨头，只给肉，因此本书与其他教科书有以下几个方面的不同。

第一，本书的语言风格偏口语化。因为本书的写作目的是让大家尽快入门强化学

习。众所周知，学一门新的课程，最快的入门方式就是请私人家教进行一对一的训练。然而，由于各种原因，这种方式并非对每个人都现实可行。而本书，正希望通过这种口语化的方式与读者交流，尽量实现一对一的训练效果。读者们可以将这本书想象成自己的私人家教。

第二，本书不会将数学基础作为单独的章节列出来，而是在强化学习算法中用到哪些数学，就在那个章节里介绍。这样，就算是没有多少数学基础的读者也可以学习；而对于那些有数学基础的读者，通过将数学与具体的强化学习算法相结合，可以提升数学的应用能力。

第三，本书的每部分都包括理论讲解，代码讲解和直观解释三项内容。强化学习算法是应用性很强的算法，大部分读者学习强化学习算法的目的是用来解决实际问题的。一边学理论，一边写代码，可以使读者在学习的过程中，同步提升理论研究和解决问题两方面的能力。

第四，本书涵盖的内容相当丰富，几乎会涉及强化学习算法的各个方面。从最基础的强化学习算法到目前最前沿的强化学习算法都会有所涉猎。所以，本书可以说是“完全”教程。当然了，这里所谓的“完全”也只是相对的。因为，强化学习算法当前正处于快速发展中，每个月都会有新的突破。但是，强化学习的基本思想是不会那么快变化的，最新的突破都是基于这些基本的思想而来。所以，读完了本书，你再继续读最新的论文，就不会再有如读天书的感觉了。或者说，读完了本书你就可以参与到构建能改变世界的伟大算法中了。

我们再回到刚才的问题：什么是强化学习算法？

要回答这个问题，必须先回答强化学习可以解决什么问题，强化学习如何解决这些问题。

1.2 强化学习可以解决什么问题

如图 1.1 所示是强化学习算法的成功案例。其中的 A 图为典型的非线性二级摆系统。该系统由一个台车（黑体矩形表示）和两个摆（红色摆杆）组成，可控制的输入为台车的左右运动，该系统的目的是让两级摆稳定在竖直位置。两级摆问题是非线性系统的经典问题，在控制系统理论中，解决该问题的基本思路是先对两级摆系统建立精确的动力学模型，然后基于模型和各种非线性的理论设计控制方法。一般来说，这