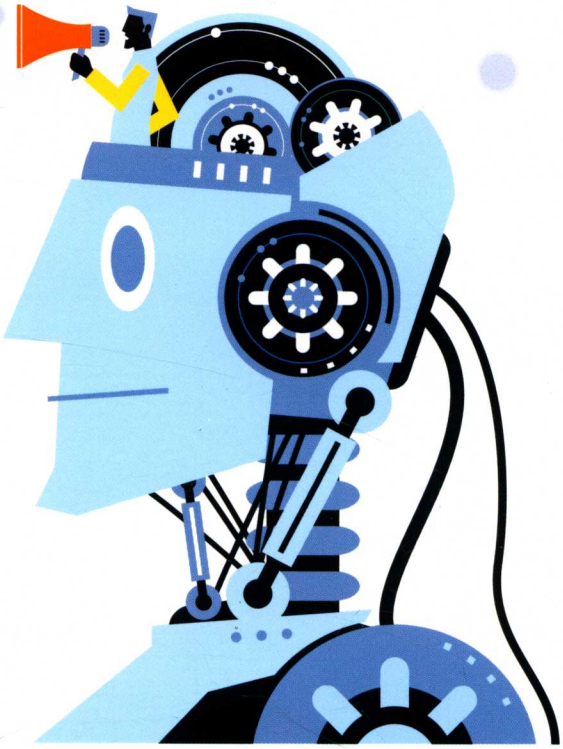
Alibaba Group 阿里巴巴集团 | 阿里技术丛书系列



强化学习实战

强化学习在阿里的技术演进和业务创新

笪庆 曾安祥 编著

全面展现强化学习在阿里

搜索场景 推荐场景 广告系统 的实践与创新

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

Alibaba Group 阿里巴巴集团 | 阿里技术丛书系列

强化学习实战

强化学习在阿里的技术演进和业务创新

笪庆 曾安祥 编著

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

内 容 简 介

本书汇集了阿里巴巴一线算法工程师在强化学习应用方面的经验和心得，覆盖了阿里巴巴集团多个事业部的多条业务线。书中系统地披露在互联网级别的应用上使用强化学习的技术细节，更包含了算法工程师对强化学习的深入理解、思考和创新。作为算法工程师，你将了解到强化学习在实际应用中的建模方法、常见的问题以及对应的解决思路，提高建模和解决业务问题的能力；对于强化学习方向的研究人员，你将了解到在游戏之外更多实际的强化学习问题，以及对应的解决方案，扩宽研究视野；对于机器学习爱好者，你将了解到阿里巴巴的一线机器学习算法工程师是如何发现问题、定义问题和解决问题的，激发研究兴趣以及提升专业素养。

本书适合算法工程师、强化学习方向的专业人员阅读，也可供机器学习爱好者参考。

图书在版编目（CIP）数据

强化学习实战：强化学习在阿里的技术演进和业务创新 / 笪庆，曾安祥编著. —北京：电子工业出版社，2018.10
（阿里技术丛书系列）
ISBN 978-7-121-33898-4

Ⅰ. ①强… Ⅱ. ①笪… ②曾… Ⅲ. ①机器学习Ⅳ. ①TP181

中国版本图书馆 CIP 数据核字(2018)第 064959 号

责任编辑：宋亚东
印　　刷：中国电影出版社印刷厂
装　　订：中国电影出版社印刷厂
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：14.75　字数：226 千字
版　　次：2018 年 10 月第 1 版
印　　次：2018 年 10 月第 1 次印刷
定　　价：89.00 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

当前的机器学习算法大致可以分为有监督学习、无监督学习和强化学习三类。强化学习和其他学习方法的不同之处在于：强化学习是智能系统从环境到行为映射的学习，以使奖励信号函数值最大。如果智能体的某个行为策略引发正的奖赏，那么智能体以后产生这个行为策略的趋势便会加强。强化学习是最接近自然界动物学习本质的一种学习范式。尽管强化学习从提出到现在差不多有半个世纪了，但是它的应用场景仍很有限，解决规模大一点的问题时会出现维数爆炸问题，难于计算，所以往往看到的例子都是相对简化的场景。

最近，强化学习因为与深度学习结合，解决海量数据的泛化问题，取得了令人瞩目的成果。在包括 DeepMind 自动学习玩 Atari 游戏，以及 AlphaGo 在围棋大赛中战胜世界冠军的背后，其强大武器之一就是深度强化学习技术。相对 DeepMind 和学术界看重强化学习的前沿研究，阿里巴巴则将重点放在推动强化学习的技术输出及商业应用上。

在阿里移动电商平台中，人机交互的便捷、碎片化使用的普遍性、页面切换的串行化、用户轨迹的可跟踪性等都要求系统能够对多变的用户行为，以及瞬息万变的外部环境进行完整建模。平台作为信息的载体，需要在与消费者的互动过程中，根据对消费者（环境）的理解，及时调整提供信息（商品、客服机器人的回答、路径选择等）的策略，从而最大化过程累积收益（消费者在平台上的使用体验）。基于监督学习方式的信息提供手段，缺少有效

的探索能力，造成其系统倾向给消费者推送曾经发生过行为的信息单元（商品、店铺或问题答案）。而强化学习作为一种有效的基于用户与系统交互过程建模和最大化过程累积收益的学习方法，在阿里一些具体的业务场景中进行了很好的实践并得到大规模应用。

- **在搜索场景中**，阿里巴巴对用户的浏览购买行为进行马尔可夫决策过程建模，在搜索实时学习和实时决策计算体系之上，实现了基于强化学习的排序策略决策模型，从而使得淘宝搜索的智能化进化至新的高度。双11桶测试效果表明，算法指标取得了近 20% 的大幅提升。
- **在推荐场景中**，阿里巴巴使用了深度强化学习与自适应在线学习，通过持续机器学习和模型优化建立决策引擎，对海量用户行为以及百亿级商品特征进行实时分析，帮助每一个用户迅速发现喜欢的商品，提高人和商品的配对效率，算法效果指标提升了 10%~20%。
- **在智能客服中**，如阿里小蜜这类的客服机器人，作为投放引擎的智能体，需要有决策能力。这个决策不是基于单一节点的直接收益来确定的，而是一个较为长期的人机交互的过程，把消费者与平台的互动看作一个马尔可夫决策过程，运用强化学习框架，建立一个消费者与系统互动的回路系统，而系统的决策是建立在最大化过程收益的基础上，达到一个系统与用户的动态平衡的。
- **在广告系统中**，如果广告主能够根据每一条流量的价值进行单独出价，广告主便可以在各自的高价值流量上提高出价，而在普通流量上降低出价，如此可以获得较好的**投资回报率（Return On Investment，ROI）**，与此同时，平台也能够提升广告与访客间的匹配效率。阿里巴巴实现了基于强化学习的智能调价技术，对于访问广告位的每一位访客，根据他们的当前状态去决定如何操作调价，给他们展现特定的广告，引导他们的状态向我们希望的方向上转移，双 11 期间实测表明，**点击率（Click-Through Rate，CTR）**、**每千次展示收入（Revenue Per Thousand，RPM）**和**成交金额（Gross Merchandise Volume，GMV）**均得到了大幅提升。

当然，强化学习在阿里巴巴内部的实践远不止于此，鉴于篇幅限制，本书只介绍了其中的一部分。未来深度强化学习的发展必定是理论探索和应用实践的双链路持续深入。希望本书能抛砖引玉，从技术和应用上帮助读者，共同推进深度强化学习的更大发展。

阿里巴巴研究员　青峰

2018 年 9 月于杭州

推荐序二

首先很欣慰地看到这本围绕强化学习应用的实践之作问世，经过几年在电商的大数据平台的持续积累，阿里巴巴的算法同学在决策智能方向迈出了坚实的一步。

回顾阿里巴巴电商搜索推荐技术的一路演进历程，有幸亲身经历了一个在大数据驱动下，学习和决策能力兼备的智能化体系的建立和发展。整本书围绕强化学习技术在搜索、推荐、广告、客服机器人等真实在线交互产品的实战经验进行了认真细致的论述，相信对从业者大有裨益，也期待更多优秀的工作应运而生。

本书大部分应用仍然是围绕着信息化系统来实验和论证的，信息化系统仍然具备了感知、匹配、选择、决策、反馈的完整闭环，而如何让强化学习技术给我们的日常生产生活中的决策问题带来价值，仍然有很长的一段路要走。本书第 12 章介绍的利用深度强化学习求解三维装箱问题，作为抛砖引玉，鼓励学者们积极探索强化学习理论在运筹优化方向的应用和探索，对于可以抽象为序列决策问题的运筹优化问题，基于传统组合优化方法的求解方式，往往会遇到响应时间长、数据利用率低等问题。第 12 章开启了如何利用数据驱动，将装箱问题建模成一个考虑如何按照顺序、位置、朝向摆放商品的序列决策问题，运用 DRL 方法优化物品的放入顺序，同时模型预测需要的时间在毫秒级左右，取代了启发式求解，在很大限度上降低了仓内库工的等待时间。

再比如，当前研发热情空前高涨的无人驾驶领域，在感知层面，随着智能传感器的升级换代，ADAS 的大量部署和数据的采集、算力的提升，感知本身在可见的将来不会是主要的瓶颈；而如何根据感知结果实现最优化控制，也就是决策算法将会是核心竞争力的体现。单存依赖深度学习建立的智能化系统失去了透明性和可解释性，仅仅依赖的是概率推理，也就是相关性，而非因果推断，而任何基于相关性作出的决策是很难保证稳定性和可靠性的。而因果推断的一个典型范例可以建立在基于强化学习的决策框架之上，它把一个决策问题当作是一个决策系统与它所处环境的博弈，这个系统需要连续做决策，优化的是长期累积收益。而众所周知的是，强化学习是一个基于 trial and error 的试错机制与环境交互，并基于收集到的数据不断改进自己的决策机制来最大化长期奖励，但是很难想象在实际无人驾驶场景中去做大量 trial，那样的代价是无法承受的。因此，我们需要思考构建一个物理环境的平行世界，来模拟路况的仿真环境，通过强化学习来做虚拟运行，获得最优的决策模型，并且还将产生大量的模拟数据，这对决策算法的成熟至关重要。很高兴也看到了本书中的第 5 章虚拟淘宝的研究，建立了一个与真实购物体系的平行宇宙，相信这样的工作对于去探索一个平台性电商的机制性研究都会有极大的参考价值。

强化学习算法是以优化预先指定的奖励函数为中心的，这些奖励函数类似于机器学习中的成本函数，而强化学习算法就是一种优化方法。由于某些算法特别容易受到奖励尺度和环境动力学(Environment Dynamics)的影响，我们更需要强调强化学习算法在现实任务中的适应性，就像成本优化(Cost-Optimization) 方法那样。在思考运用强化学习解决问题的时候，需要试图回答这样的问题：哪些设定使该研究有用？在研究社区中，我们必须使用公平的对比，以确保结果是可控的和可复现的。衷心地鼓励所有的业界同仁们带着好奇心、敬畏心，持续推动强化学习方向在实际应用领域的开花结果。

徐盈辉　阿里巴巴研究员，菜鸟人工智能部负责人

推荐序三

2018 年 7 月，在国际机器学习会议 ICML’18 上，“强化学习”占据 17 个 session，超越“深度学习”，成为唯一贯穿主会 3 天日程的主题；在国际人工智能联合大会 IJCAI’18 上，以强化学习为题的论文较上一年增长超过 50%；在国际智能体与多智能体会议 AAMAS’18 上，“学习”session 由上一年的 1 个增长为 4 个；国内，2018 年 8 月，在智能体及多智能体系统专题论坛上，数百人的会场座无虚席。种种迹象表明，强化学习近来已成为人工智能、机器学习中最受关注的研究方向之一。

然而，就在几年前还是另一番景象。2011 年我在导师周志华教授的指导下以演化计算理论基础为题取得博士学位，继而在周志华教授的指引下选择新的研究方向。强化学习希望赋予机器自主决策的能力，是富有挑战而在通向人工智能的道路上必不可少的一环，同时从技术上与我博士生期间的主要研究方向也有关联。切换到强化学习研究的想法，立即得到了周志华教授的肯定和支持。后续研究工作的开展，也得到了在这一方向上长期耕耘的南京大学高阳教授的支持和帮助。然而在几年前，寻找强化学习合作研究的学生时，我常常需要回答“强化学习在企业中有用吗”之类的问题，左思右想，最后只能尴尬的回应，“嗯，目前暂时可能用得很少”。其实，“用得很少”在当时已经是夸大的说法了，尤其是对于同学们最感兴趣的互联网企业。幸运的是，对“冷门”的强化学习，仍然有同学有兴趣合作，其中笪庆同学后来成为阿里强化学习技术应用的主力之一。

人工智能技术最终是面向应用的技术，“用得很少”对一个研究方向的发展无疑会产生严重的制约。所幸 2016 年，DeepMind 的 AlphaGo 系统借助强化学习技术达到的围棋水平超越人类职业选手，掀起了人工智能的新一轮热潮，也引发了对强化学习技术的广泛关注。然而，强化学习技术仍然很不成熟，在实际问题中应用面临很高的门槛，以至于最近有一些指责强化学习存在“泡沫”的声音。虚远大于实才会形成“泡沫”，而本书介绍的强化学习在阿里巴巴业务场景中的实践，就是强化学习可以切实落地的初步展示。其中，“虚拟淘宝”等工作也是我们与青峰、仁重团队合作，为解决强化学习落地过程中的障碍而进行的尝试。我们相信强化学习，这种被 DeepMind 认为是通向通用人工智能愿景的主要技术，在企业应用的支撑下会有更加蓬勃的发展生机，将会深刻地影响和改变人类社会。

俞扬

于南京大学

2018 年 9 月 15 日

推荐语

强化学习是关于智能体在与环境交互中学习序列决策策略的机器学习问题，将会在人工智能领域中发挥越来越重要的作用。由于学习难度高等原因，强化学习的应用大多局限于游戏等虚拟世界，在现实世界中的成功案例并不多见。阿里巴巴集团的同仁们将强化学习技术应用到搜索、推荐、广告、客服等业务上，并取得了很大的成功，令人钦佩，值得大家学习和借鉴。相信这部专著对关注强化学习技术的人都将大有裨益。

李航　今日头条人工智能实验室主任

强化学习，尤其是基于深度神经网络值函数的强化学习算法框架，在博弈等领域取得了举世瞩目的进展。然而，如何把这些基础算法框架应用到商业场景中，对问题建模和算法设计本身都提出了较高的要求和挑战。本书针对阿里的几个基础场景：搜索、推荐、广告，提出了一套基于深度学习和多智能体建模的通用强化学习算法框架，并针对每个场景提出了些新的设计和创新，并在阿里数据集上验证了算法效果，是行业内世界前沿的强化学习及应用参考书。

唐平中　清华大学交叉信息研究院副教授，
计算经济学研究室主任，中组部“千人计划”青年人才

目　录