大规模强化学习

刘 全 傅启明 钟 珊 黄 蔚 著

科学出版社

大规模强化学习

刘 全 傅启明 钟 珊 黄 蔚 著

阜阳师范学院图书馆
藏书

科学出版社
北京

内 容 简 介

本书讨论大规模强化学习的理论及方法，介绍强化学习在大状态空间任务中的应用。该研究已成为近年来计算机科学与技术领域最活跃的研究分支之一。

全书共分六部分 21 章。第一部分是强化学习基础。第二部分是用于强化学习的值函数逼近方法。第三部分是最小二乘策略迭代方法。第四部分是模糊近似强化学习方法。第五部分是并行强化学习方法。第六部分是离策略强化学习方法。

本书可以作为高等院校计算机专业和自动控制专业研究生的教材，也可以作为相关领域科技工作者和工程技术人员的参考书。

图书在版编目 (CIP) 数据

大规模强化学习 / 刘全等著. —北京：科学出版社，2016.3
ISBN 978-7-03-047747-7

Ⅰ. ①大… Ⅱ. ①刘… Ⅲ. ①人工智能－研究 Ⅳ. ①TP18

中国版本图书馆 CIP 数据核字(2016)第 054637 号

责任编辑：王 哲 董素芹 / 责任校对：胡小洁
责任印制：张 倩 / 封面设计：迷底书装

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

文林印务有限公司 印刷

科学出版社发行 各地新华书店经销

*

2016 年 3 月第 一 版 开本：720×1 000 1/16
2016 年 3 月第一次印刷 印张：18 1/4
字数：352 000

定价：96.00 元

（如有印装质量问题，我社负责调换）

前　言

机器学习作为人工智能领域的研究热点和前沿，一直是智能科学和智能计算领域研究的核心，是实现机器智能的关键技术。在机器学习领域，根据与环境交互的特点，机器学习方法可以分为监督学习、无监督学习和强化学习。其中强化学习基于动物生理学和心理学的有关原理，采用人类和动物学习中的“试错”机制，强调从与环境的交互中学习，学习过程中仅需要获得评价性的反馈信号，以极大化累积奖赏为学习目标。在人工智能的早期研究中，受到心理学研究的影响，强化学习一度成为机器学习的研究热点之一。但由于强化学习问题本身的复杂性，在 20 世纪 80 年代，机器学习的研究工作和成果主要集中在监督学习和无监督学习。从 20 世纪 80 年代末开始，随着强化学习的数学基础研究取得突破性进展，对强化学习的研究和应用也日益开展起来，成为目前机器学习中富有挑战性和广泛应用前景的研究领域之一。

Machine Learning 分别在 1992 年和 1996 年出版了强化学习专辑，登载了数篇强化学习的理论研究论文，其中 Sutton 于 1992 年编辑的第一个专刊标志着强化学习发展成为机器学习领域的一个重要组成部分。*Robotics and Autonomous System* 在 1995 年也出版了强化学习专辑，主要介绍了强化学习在智能机器人领域的应用情况。美国国家科学基金会于 2006 年召开了近似动态规划论坛(NSFADP'06)。IEEE 从 2007 年开始每两年召开一次以“近似动态规划与强化学习”为主题的国际研讨会(IEEE ADPRL'2007、IEEE ADPRL'2009、IEEE ADPRL'2011、IEEE ADPRL'2013、IEEE ADPRL'2015)，到目前已召开 5 届。IEEE 计算机学会于近年专门成立了近似动态规划与强化学习的技术委员会(IEEE TC on ADPRL)。随着国内外对于强化学习理论和应用重视程度的不断提高，目前强化学习已经成为过程控制、作业调度、路径规划、Web 信息搜索、证券管理、期权定价等领域对目标行为优化的一种重要技术。

强化学习技术是一种介于监督学习和非监督学习之间的在线机器学习方法。由于其具有通过与环境交互并根据相关反馈进行学习的特性，使得该方法非常适合在线、实时预测及决策问题的处理。在大数据环境下，由于数据具有体量大、结构复杂、变化迅速等特点，难以将传统的机器学习方法直接用于大数据对象的求解，即使勉强将某些机器学习方法用于解决大数据的问题，通常也难以取得较为理想的结果。由于强化学习具有不需要监督信号且仅根据反馈信息就可以自学习的特性，使得其在大数据分析和处理方面受到国内外研究者的广泛关注。

本书作者多年来一直从事强化学习的研究工作，在国家自然科学基金、国家博

士后基金、教育部科学研究重点项目、江苏省自然科学基金以及江苏省高校重点项目的资助下，提出了一整套大规模强化学习理论，解决了一系列强化学习方法的核心技术，并将这些理论和方法用于解决实际问题。

本书的主要内容曾发表于国内外权威期刊和学术会议上，部分内容已获省部、市级奖励。本书是在此基础上，经过进一步深化、加工而成的，是对已有研究成果的全面总结。全书共分六部分 21 章。第一部分是强化学习基础，包括第 1 章：强化学习概述；第 2 章：大规模或连续状态空间的强化学习。第二部分是用于强化学习的值函数逼近方法，包括第 3 章：梯度下降值函数逼近模型的改进；第 4 章：基于 LSSVR 的 Q-值函数分片逼近模型；第 5 章：基于 ANRBF 网络的 Q-V 值函数协同逼近模型；第 6 章：基于高斯过程的快速 Sarsa 算法；第 7 章：基于高斯过程的 Q 学习算法。第三部分是最小二乘策略迭代方法，包括第 8 章：最小二乘策略迭代算法；第 9 章：批量最小二乘策略迭代算法；第 10 章：自动批量最小二乘策略迭代算法；第 11 章：连续动作空间的批量最小二乘策略迭代算法。第四部分是模糊近似强化学习方法，包括第 12 章：一种基于双层模糊推理的 Sarsa(λ)算法；第 13 章：一种基于区间型二型模糊推理的 Sarsa(λ)算法；第 14 章：一种带有自适应基函数的模糊值迭代算法。第五部分是并行强化学习方法，包括第 15 章：基于状态空间分解和智能调度的并行强化学习；第 16 章：基于资格迹的并行时间信度分配强化学习算法；第 17 章：基于并行采样和学习经验复用的 E^3 算法。第六部分是离策略强化学习方法，包括第 18 章：基于线性函数逼近的离策略 Q(λ)算法；第 19 章：基于二阶 TD Error 的 Q(λ)算法；第 20 章：基于值函数迁移的快速 Q-Learning 算法；第 21 章：离策略带参贝叶斯强化学习算法。

本书总体设计、修改和审定由刘全完成，参加撰写的有傅启明、钟珊、黄蔚、杨旭东、肖飞、周鑫、穆翔、陈桂兴等，对以上作者付出的艰辛劳动表示感谢。本书的撰写参考了国内外有关研究成果，他们的丰硕成果和贡献是本书学术思想的重要来源，在此对涉及的专家和学者表示诚挚的谢意。本书也得到了苏州大学计算机科学与技术学院及软件形式化与自动推理学科组部分老师和学生的大力支持和协助，他们是凌兴宏、伏玉琛、朱斐、章晓芳、章宗长、陈冬火、鲁逊、周小科、王辉、金海东、王浩、于俊、孙洪坤、高龙、施梦宇、庄超、周谊成、尤树华、许丹、钱炜晟、章鹏、翟建伟、梁斌、徐进、许志鹏、朱海军、孙慈嘉、周倩等，在此一并表示感谢。

机器学习是一个快速发展、多学科交叉的研究方向，其理论及应用均存在大量亟待解决的问题。限于作者的水平，书中难免有不足之处，敬请读者指正。

作　者

2015 年 12 月

试读结束：需要全本请在线购买： www.ertongbook.com

目　录

前言

第 1 章　强化学习概述……1
1.1　简介……1
1.2　形式框架……3
1.2.1　马尔可夫决策过程……3
1.2.2　策略……6
1.2.3　回报……7
1.3　值函数……7
1.4　解决强化学习问题……9
1.4.1　动态规划：基于模型的解决技术……9
1.4.2　强化学习：模型无关的解决技术……16
1.5　本章小结……20
参考文献……21

第 2 章　大规模或连续状态空间的强化学习……23
2.1　简介……23
2.2　近似表示……24
2.2.1　带参数化值函数逼近……24
2.2.2　非参数化值函数逼近……28
2.3　值函数逼近求解方法……29
2.3.1　梯度下降方法……30
2.3.2　最小二乘回归……31
2.4　本章小结……31
参考文献……32

第 3 章　梯度下降值函数逼近模型的改进……33
3.1　改进的梯度下降值函数逼近模型……33
3.1.1　势函数塑造奖赏机制……33
3.1.2　基于势函数塑造奖赏机制的值函数逼近模型……35
3.2　NRBF-GD-Sarsa(λ)算法……36

3.2.1　算法描述 …… 36
3.2.2　算法收敛性分析 …… 37
3.3　仿真实验 …… 39
3.3.1　实验描述 …… 39
3.3.2　实验设置 …… 40
3.3.3　实验分析 …… 41
3.4　本章小结 …… 43
参考文献 …… 44

第 4 章　基于 LSSVR 的 Q-值函数分片逼近模型 …… 45
4.1　LSSVR-Q-值函数分片逼近模型 …… 45
4.2　在线稀疏化样本池构建方法 …… 48
4.3　LSSVR-Q 算法 …… 49
4.4　仿真实验 …… 49
4.4.1　实验 1：Mountain Car 问题 …… 51
4.4.2　实验 2：DC Motor 问题 …… 54
4.5　本章小结 …… 57
参考文献 …… 58

第 5 章　基于 ANRBF 网络的 Q-V 值函数协同逼近模型 …… 59
5.1　Q-V 值函数协同机制 …… 59
5.2　Q-V 值函数协同逼近模型 …… 61
5.3　Q-V 值函数协同逼近算法 …… 63
5.3.1　QV(λ)算法 …… 63
5.3.2　算法收敛性分析 …… 65
5.4　仿真实验 …… 67
5.4.1　实验描述 …… 67
5.4.2　实验设置 …… 68
5.4.3　实验分析 …… 68
5.5　本章小结 …… 73
参考文献 …… 73

第 6 章　基于高斯过程的快速 Sarsa 算法 …… 75
6.1　新的值函数概率生成模型 …… 75
6.2　利用高斯过程对线性带参值函数建模 …… 77
6.3　FL-GPSarsa 算法 …… 78

6.4 仿真实验……81
6.4.1 带风的格子世界问题……81
6.4.2 Mountain Car 问题……84
6.5 本章小结……86
参考文献……87

第 7 章 基于高斯过程的 Q 学习算法……88
7.1 值迭代方法……88
7.2 用于值迭代的值函数概率生成模型……89
7.3 GP-QL 算法……90
7.4 仿真实验……93
7.4.1 实验 1：带悬崖的格子世界问题……93
7.4.2 实验 2：Mountain Car 问题……96
7.5 本章小结……97
参考文献……97

第 8 章 最小二乘策略迭代算法……99
8.1 马尔可夫决策过程……99
8.2 最小二乘策略迭代……100
8.2.1 投影贝尔曼等式的矩阵形式……100
8.2.2 最小二乘策略迭代……103
8.2.3 在线最小二乘策略迭代……104
8.3 本章小结……106
参考文献……106

第 9 章 批量最小二乘策略迭代算法……107
9.1 批量强化学习算法……107
9.2 批量最小二乘策略迭代算法……108
9.3 算法分析……111
9.3.1 收敛性分析……111
9.3.2 复杂度分析……113
9.4 仿真实验……114
9.4.1 实验描述……114
9.4.2 实验设置……115
9.4.3 实验分析……115
9.5 本章小结……120
参考文献……120

第 10 章　自动批量最小二乘策略迭代算法 …… 122
10.1　定点步长参数评估方法 …… 122
10.2　自动批量最小二乘策略迭代算法 …… 124
10.3　仿真实验 …… 125
10.3.1　实验描述 …… 125
10.3.2　实验分析 …… 125
10.4　本章小结 …… 130
参考文献 …… 130

第 11 章　连续动作空间的批量最小二乘策略迭代算法 …… 132
11.1　二值动作搜索 …… 132
11.2　快速特征选择 …… 133
11.3　连续动作空间的快速特征选择批量最小二乘策略迭代算法 …… 134
11.4　仿真实验 …… 136
11.4.1　实验描述 …… 136
11.4.2　实验设置 …… 136
11.4.3　实验分析 …… 136
11.5　本章小结 …… 140
参考文献 …… 141

第 12 章　一种基于双层模糊推理的 Sarsa(λ) 算法 …… 143
12.1　Q-值函数的计算和 FIS 的参数更新 …… 143
12.2　DFR-Sarsa(λ) 算法 …… 146
12.2.1　DFR-Sarsa(λ) 算法的学习过程 …… 146
12.2.2　算法收敛性分析 …… 147
12.3　仿真实验 …… 149
12.3.1　Mountain Car …… 149
12.3.2　平衡杆 …… 151
12.4　本章小结 …… 153
参考文献 …… 153

第 13 章　一种基于区间型二型模糊推理的 Sarsa(λ) 算法 …… 155
13.1　近似 Q-值函数的计算和参数的更新 …… 155
13.2　IT2FI-Sarsa(λ) 算法的学习过程 …… 157
13.3　算法收敛性分析 …… 158
13.4　仿真实验 …… 162

13.4.1 实验设置 …… 163
13.4.2 实验分析 …… 163
13.5 本章小结 …… 165
参考文献 …… 165

第 14 章 一种带有自适应基函数的模糊值迭代算法 …… 167
14.1 基函数的近似性能评价 …… 167
14.2 基函数的自适应细化更新方式 …… 169
14.3 ABF-QI 算法 …… 170
14.3.1 ABF-QI 算法的学习过程 …… 170
14.3.2 算法收敛性分析 …… 171
14.4 仿真实验 …… 172
14.4.1 问题描述与参数设置 …… 172
14.4.2 实验分析 …… 172
14.5 本章小结 …… 175
参考文献 …… 175

第 15 章 基于状态空间分解和智能调度的并行强化学习 …… 177
15.1 IS-SRL 和 IS-SPRL …… 177
15.1.1 子问题的学习过程 …… 177
15.1.2 IS-SPRL 的消息传递和调度 …… 180
15.1.3 学习步骤 …… 181
15.2 加权优先级调度算法 …… 183
15.3 收敛性分析 …… 186
15.3.1 模型和假设 …… 187
15.3.2 基于 IS-SRL 和 IS-SPRL 的 Q 学习算法的收敛性 …… 188
15.4 仿真实验 …… 190
15.4.1 不同调度算法的比较 …… 191
15.4.2 算法在不同参数下的性能比较 …… 191
15.4.3 不同算法的收敛速度的比较 …… 193
15.4.4 结果分析 …… 195
15.5 本章小结 …… 195
参考文献 …… 196

第 16 章 基于资格迹的并行时间信度分配强化学习算法 …… 198
16.1 资格迹与强化学习 …… 199

16.2 并行时间信度分配 …… 200
16.3 性能优化与系统容错 …… 203
16.3.1 状态迁移预测 …… 203
16.3.2 故障预防和恢复 …… 203
16.4 仿真实验 …… 204
16.5 本章小结 …… 206
参考文献 …… 207

第 17 章 基于并行采样和学习经验复用的 E^3 算法 …… 209
17.1 E^3 算法 …… 209
17.2 学习经验复用 …… 212
17.3 并行 E^3 算法 …… 212
17.4 系统容错 …… 215
17.5 仿真实验 …… 216
17.6 本章小结 …… 219
参考文献 …… 219

第 18 章 基于线性函数逼近的离策略 Q(λ) 算法 …… 221
18.1 离策略强化学习 …… 221
18.1.1 梯度下降法与线性函数逼近 …… 221
18.1.2 离策略强化学习算法 …… 224
18.2 GDOP-Q(λ) 算法 …… 226
18.2.1 GDOP-Q(λ) …… 226
18.2.2 收敛性分析 …… 227
18.3 仿真实验 …… 230
18.4 本章小结 …… 234
参考文献 …… 234

第 19 章 基于二阶 TD Error 的 Q(λ) 算法 …… 236
19.1 二阶 TD Error 快速 Q(λ) 算法 …… 236
19.1.1 二阶 TD Error …… 236
19.1.2 资格迹 …… 238
19.1.3 SOE-FQ(λ) …… 238
19.1.4 算法收敛性及时间复杂度分析 …… 239
19.2 仿真实验 …… 244
19.2.1 Random Walk 问题 …… 244

试读结束：需要全本请在线购买： www.ertongbook.com

19.2.2 Mountain Car 问题 …… 247
19.3 本章小结 …… 248
参考文献 …… 249

第 20 章 基于值函数迁移的快速 Q-Learning 算法 …… 251
20.1 自模拟度量与状态之间的距离 …… 252
20.2 基于值函数迁移的 Q-Learning 算法 …… 254
20.2.1 基于自模拟度量的值函数迁移 …… 254
20.2.2 VFT-Q-Learning …… 256
20.3 仿真实验 …… 257
20.3.1 问题描述 …… 257
20.3.2 实验设置 …… 258
20.3.3 实验分析 …… 258
20.4 本章小结 …… 262
参考文献 …… 263

第 21 章 离策略带参贝叶斯强化学习算法 …… 264
21.1 高斯过程 …… 264
21.2 基于高斯过程的离策略带参近似策略迭代算法 …… 265
21.2.1 基于高斯过程的值函数参数估计 …… 265
21.2.2 基于 VPI 的动作选择方法 …… 269
21.2.3 GP-OPPAPI …… 270
21.3 仿真实验 …… 273
21.4 本章小结 …… 275
参考文献 …… 276

第 1 章　强化学习概述

1.1　简　　介

通过与环境交互学习是人类获取知识的主要方法，也是人类提高智能水平的基本途径。人类智能研究的一个最核心问题就是构建具有类似人类智能的系统。该系统的一个主要特征就是能够适应未知环境，并逐渐增强其自身能力。

研究发现，生物进化过程中为适应环境而进行的学习主要有两个特点：一是生物从来不是静止地、被动地等待，而是主动地对环境进行试探；二是环境对试探动作产生的反馈是评价性的，生物根据对环境的评价来调整未来的行为。在人工智能领域中，将具有以上两个特点的学习称为强化学习(Reinforcement Learning, RL)，也可以称为增强学习或再励学习[1-3]。强化学习是从控制理论、统计学、心理学等相关学科发展而来的，最早可以追溯到巴普洛夫的条件反射实验。但直到 20 世纪 80 年代末、90 年代初强化学习技术才得到广泛的重视。由于强化学习具有自学习和在线学习的优点，它被认为是设计智能系统的核心技术之一。目前随着强化学习理论的不断发展和完善，强化学习技术越来越多地应用于工业控制、作业调度、生产管理等方面，并逐步成为机器学习领域的研究热点。

强化学习是一种交互式的学习方法，其主要特点为试错(trial-and-error)搜索和延迟回报(delay return)。学习过程是智能体(Agent)与环境不断交互并从环境的反馈信息中学习的过程。Agent 与环境交互的过程如下：①Agent 感知当前的环境状态(state)；②根据当前的状态和奖赏值(reward)(强化信号)，Agent 选择一个动作(action)并执行该动作；③当 Agent 所选择的动作作用于环境时，环境转移到新状态，并给出新的奖赏；④Agent 根据环境反馈的奖赏值，计算回报值(return)，并将回报值作为更新内部策略的依据。具体如图 1.1 所示。

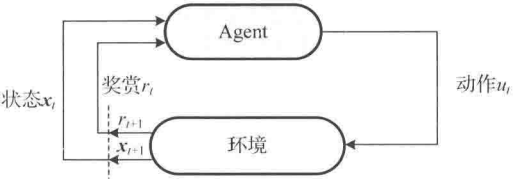

图 1.1　强化学习中 Agent 与环境交互过程

在这个过程中，并没有告诉 Agent 应该采取哪个动作，而是由 Agent 根据环境的反馈信息自己发现的。Agent 选择动作的原则是：尽量让 Agent 在以后的学习过程中从环境获得的正强化信号概率增大，即 Agent 应该使自己的动作受到环境奖励的概率增大，受到惩罚的概率减小。正是这样的学习特点，使得强化学习成为与监督学习(supervised learning)和非监督学习(unsupervised learning)并列的一种学习技术。

经典的强化学习算法，虽然在形式上提供了统一的框架，但在实际应用中存在以下几方面问题。

(1)“维数灾”，即学习参数个数随状态变量维数呈指数级增长的现象。经典的强化学习算法不具备较好的伸缩性。在一些大状态空间或连续状态空间的学习任务中，强化学习系统通常没有足够的资源和能力在有限的时间和空间内学习到一个合理的解决方案。“维数灾”问题严重限制了强化学习技术的广泛应用。

(2)收敛速度慢。收敛速度慢与“维数灾”问题有着密切的关系。多数强化学习算法收敛到最优解的理论保障都是建立在“任意状态都能被无限次访问到”这个前提条件之上的。当问题环境比较复杂或出现“维数灾”问题时，Agent 的探索策略不能保证每个状态都能在有限的时间内被访问足够多的次数，因而 Agent 没有足够的经验能够在这些较少遇到的状态下作出正确的决策，这必然会导致算法的收敛速度较慢。这就使得强化学习在处理具有实时性要求的在线学习任务时，显得力不从心。

(3)探索(exploration)和利用(exploitation)平衡问题。在强化学习中，Agent 难以权衡长期和短期利益。一方面为了获得较高的奖赏，Agent 需要利用学到的经验在已经探索过的动作中贪心地选择一个获益最大的动作；另一方面，为了发现更好的策略，Agent 需要扩大探索范围，尝试以前没有或较少试过的动作。这样 Agent 就处于进退两难的境地。

(4)时间信度分配问题。由于强化学习具有延迟回报的特点，即环境反馈给系统的信息比较稀疏且具有较大的延时。所以当 Agent 收到一个奖赏信号时，决定先前的哪些行为应分配到相应的信度以及各自分配多少信度是比较困难的。例如，考虑在足球比赛游戏中选择参赛队伍的问题，假设选定的队伍在比赛的最后一秒以一分之差输掉了比赛，那么仅惩罚最后那一时刻的系统行为是不明智的。

针对强化学习中的“维数灾”问题，目前的解决方法主要包括状态聚类方法[4]、有限策略空间搜索方法[5]、值函数近似方法[6-8]、关系强化学习方法[9,10]、分层强化学习方法[11,12]等。

状态聚类方法通过把多个相似状态聚为单一状态而有效地缩减了状态空间，但缩减后的状态空间不具备马尔可夫(Markov)属性，导致强化学习系统的振荡周期很长，甚至无法收敛。有限策略空间搜索方法根据可观测的局部状态直接在有限的搜索空间中寻优，该方法经常陷入局部最优，求解质量得不到保证。值函数近似方法

使用一组特征基函数的组合来近似表示值函数，但所需的特征只有在具备问题先验知识的前提下才可以获取，并且也没有一种通用的近似方法能适用于所有的学习任务，如果该方法中的泛化偏置与问题不匹配，将导致收敛速度很慢，甚至不能正确收敛。关系强化学习方法用关系结构将强化学习泛化到关系表达的状态和动作上，通过使用一阶逻辑和决策树来学习决策，该方法的优点是可以将相似环境中的对象和已经学习到的知识泛化到不同的任务中。另外，使用关系表示也是一种比较自然的利用先验知识的方式，然而对于许多强化学习任务，很难给出一阶逻辑表示的先验知识。分层强化学习方法是通过在强化学习的基础上增加“抽象”机制，把整个任务分解为不同层次上的子任务，使每个子任务在规模较小的子问题空间中求解，并且求得的子任务策略可以复用，从而加快问题的求解速度，但是在复杂环境或未知环境中学习时，任务的层次结构很难事先确定。针对强化学习在实际应用中收敛速度慢的问题，已有很多研究对强化学习算法提出改进，从不同的角度来提高强化学习的收敛速度[13-15]，然而这些改进算法不可避免地增加了问题求解的复杂性。

针对强化学习探索和利用难以平衡的问题，目前的解决方案主要包括 ε 贪心、玻尔兹曼(Boltzmann)探索方法、最优初始值(Optimal Initial Value，OIV)方法、贝叶斯模型方法、置信区间估计方法、探索奖励方法等[16]。

在实际应用中，一个单独的学习 Agent 很难应对特征丰富的大状态空间问题，随着问题的状态空间在大小和维数上的不断增长，或者问题的复杂程度不断增加，一些已有的解决“维数灾”问题的方法越来越难以有效应用。因为像函数近似这类提高强化学习算法可扩展性的方法的能力是有限的，它们可以在一定程度上提高算法的扩展能力，却很难应付一些状态空间巨大的实际问题，如中国象棋(空间复杂度 10^{52})、围棋(空间复杂度 10^{160})[17]等。

直观上，通过增加计算、存储和网络宽带等资源可以从根本上提升算法的收敛速度和扩展能力。这就需要依赖于分布式的并行计算机体系结构。因此，并行强化学习方法[18,19]被提出，并成为强化学习研究的一个重要分支。该方法通过多个独立学习并共享信息的 Agent 来并行强化学习过程。其主要出发点是多个 Agent 能够用比单个 Agent 少得多的时间来完整地探索复杂的状态空间，并且这些 Agent 可以分布在不同的计算节点上，以便充分发挥并行体系结构的优势。

1.2　形 式 框 架

1.2.1　马尔可夫决策过程

强化学习问题可以利用马尔可夫决策过程(Markov Decision Processes, MDP)[20]框架来形式化地描述。

MDP 包含环境的状态空间 $\boldsymbol{X}$、Agent 的动作空间 U、环境的迁移函数 f 以及奖赏函数 ρ 等 4 个部分，即 $\langle \boldsymbol{X}, U, f, \rho \rangle$。

(1) 状态。

环境的状态集 $\boldsymbol{X}$ 定义为一个有穷集合 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$。这里，$N$ 为状态空间大小，即 $|\boldsymbol{X}| = N$。

(2) 动作。

Agent 的动作集 U 定义为一个有穷集合 $\{u_1, u_2, \cdots, u_M\}$。这里，$M$ 为状态空间大小，即 $|U| = M$。动作用来控制系统的状态。

(3) 迁移函数及奖赏函数。

在离散的时间步 k，对状态 $\boldsymbol{x}_k$ 采取动作 u_k，状态迁移到下一状态 $\boldsymbol{x}_{k+1}$，并得到奖赏。其迁移通过迁移函数得到，奖赏通过奖赏函数得到。根据迁移情况，可以分为确定环境迁移和随机环境迁移。

1) 确定环境迁移

确定环境迁移就是对状态 $\boldsymbol{x}$ 采取动作 u 后，根据迁移函数，迁移到确定的下一状态。其迁移函数 $f: \boldsymbol{X} \times U \to \boldsymbol{X}$

$$\boldsymbol{x}_{k+1} = f(\boldsymbol{x}_k, u_k)$$

同时，根据奖赏函数 $\rho: \boldsymbol{X} \times U \to \mathbb{R}$ 有

$$r_{k+1} = \rho(\boldsymbol{x}_k, u_k)$$

这里假设 $\|\rho\|_\infty = \sup_{x,u} |\rho(\boldsymbol{x}, u)|$ 是有穷的。奖赏是对采取动作 u_k，状态从 $\boldsymbol{x}_k$ 迁移到 $\boldsymbol{x}_{k+1}$ 得到的立即效果的评价，而不是对长期效果的评价。

例 1.1　确定环境环保机器人 MDP。环保机器人具有收集空易拉罐和返回基站充电两方面功能。考虑确定环境，如图 1.2 所示。

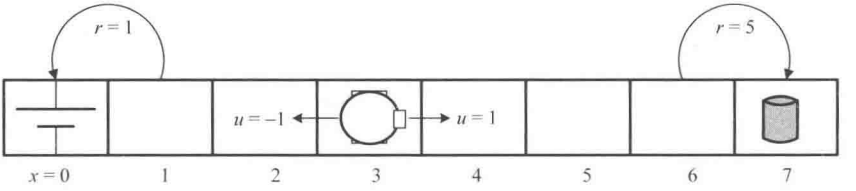

图 1.2　确定环境环保机器人问题

在该问题中，状态 x 描述机器人的位置。为了简化问题，将状态空间离散化为 8 个不同的状态，分别表示为 0～7：$X = \{0,1,2,3,4,5,6,7\}$。动作 u 描述机器人的运动方向，为了简化问题，机器人只能向左 ($u = -1$) 和向右 ($u = 1$) 移动，其离散动作空间为 $U = \{-1, 1\}$。状态 0 和 7 为吸收状态 (absorbing state)，即一旦机器人到达这两个状态，就不会再离开。对应的迁移函数是

$$f(x,u)=\begin{cases}x+u, & 1\leqslant x\leqslant 6\\ x, & x=0\text{ 或 }x=7\end{cases}$$

到达状态 7，机器人可以捡到一个易拉罐，并得到 5 的奖赏；到达状态 0，机器人充电，并得到 1 的奖赏；其他情况，奖赏均为 0。特别是当机器人到达吸收状态后，无论采取什么动作，只能得到 0 的奖赏。对应的奖赏函数是

$$\rho(x,u)=\begin{cases}5, & x=6,u=1\\ 1, & x=1,u=-1\\ 0, & \text{其他}\end{cases}$$

2）随机环境迁移

随机环境迁移就是对状态 $\boldsymbol{x}$ 采取动作 u 后，根据迁移函数，迁移到的下一状态是不确定的，而是一个随机变量。其迁移函数 $\tilde{f}:\boldsymbol{X}\times U\times\boldsymbol{X}\to[0,\infty)$。在状态 $\boldsymbol{x}_k$ 中，采取动作 u_k 后，下一状态 $\boldsymbol{x}_{k+1}$ 的概率在区间 $\boldsymbol{X}_{k+1}\subseteq\boldsymbol{X}$ 中，即

$$P(\boldsymbol{x}_{k+1}\in\boldsymbol{X}_{k+1}\,|\,\boldsymbol{x}_k,u_k)=\int_{\boldsymbol{X}_{k+1}}\tilde{f}(\boldsymbol{x}_k,u_k,\boldsymbol{x}')\mathrm{d}\boldsymbol{x}'$$

这里为了表明动作出现的次序，定义了一个离散的全局时间变量，$k=1,2,\cdots$。这样 $\boldsymbol{x}_k$ 表示在时间 k 时的状态 $\boldsymbol{x}$，而 $\boldsymbol{x}_{k+1}$ 表示在时间 $k+1$ 时的状态 $\boldsymbol{x}$。任意 $\boldsymbol{x}$ 和 u，$\tilde{f}(\boldsymbol{x},u,\cdot)$ 为一个带“·”的有效概率密度函数，这里“·”代表随机变量 $\boldsymbol{x}_{k+1}$。由于奖赏与迁移相关，迁移不再由目前的状态和所采取的动作决定，而奖赏函数也必然依赖于下一个状态，即 $\tilde{\rho}:\boldsymbol{X}\times U\times\boldsymbol{X}\to\mathbb{R}$。当状态迁移到 $\boldsymbol{x}_{k+1}$ 后，获得奖赏 r_{k+1}，即

$$r_{k+1}=\tilde{\rho}(\boldsymbol{x}_k,u_k,\boldsymbol{x}_{k+1})$$

这里假设 $\|\tilde{\rho}\|_\infty=\sup_{\boldsymbol{x},u,\boldsymbol{x}'}|\tilde{\rho}(\boldsymbol{x},u,\boldsymbol{x}')|$ 是有穷的。

当状态空间为离散时，迁移函数可以由 $\bar{f}:\boldsymbol{X}\times U\times\boldsymbol{X}\to[0,1]$ 给出，在状态 $\boldsymbol{x}_k$ 中，采取动作 u_k 后，到达下一状态 $\boldsymbol{x}'$ 的概率为

$$P(\boldsymbol{x}_{k+1}=\boldsymbol{x}'\,|\,\boldsymbol{x}_k,u_k)=\bar{f}(\boldsymbol{x}_k,u_k,\boldsymbol{x}')$$

对于任意 $\boldsymbol{x}$ 和 u，函数 $\bar{f}$ 必须满足 $\sum_{\boldsymbol{x}'}\bar{f}(\boldsymbol{x}_k,u_k,\boldsymbol{x}')=1$。

例 1.2　随机环境环保机器人 MDP。重新考虑例 1.1 的环保机器人问题。假设由于地面的问题，采取某一动作后，状态迁移不再确定。当采取某一动作试图向某一方向移动时，机器人成功移动的概率为 0.80，保持原地不动的概率为 0.15，移动到相反方向的概率为 0.05，如图 1.3 所示。

迁移函数 $\bar{f}$ 可以通过表 1.1 给出。

试读结束：需要全本请在线购买：　www.ertongbook.com