智能科学技术著作丛书

强化学习原理及其应用

王雪松　朱美强　程玉虎 著

科学出版社

智能科学技术著作丛书

强化学习原理及其应用

王雪松　朱美强　程玉虎　著

科学出版社

北京

内 容 简 介

作为一类求解序贯优化决策问题的有效方法，强化学习在运筹学、计算科学和自动控制等领域得到广泛应用，业已成为机器学习领域最活跃的研究分支之一。

现阶段，强化学习研究的核心问题是如何解决维数灾难，提高学习效率。本书的主要内容正是针对上述问题展开的，分别从值函数逼近、直接策略搜索和基于谱方法的学习 3 个方面来阐述强化学习的理论、方法及其应用，共 13 章。第 1 章~第 2 章为强化学习概述和相关基础理论。第 3 章~第 5 章为基于值函数估计的强化学习方法，包括基于支持向量机、测地高斯基的强化学习和基于抽象状态的贝叶斯强化学习。第 6 章~第 9 章为直接策略搜索强化学习方法，包括基于增量最小二乘时间差分的 Actor-Critic 学习、融合经验数据的 Actor-Critic 强化学习、基于资格迹的折扣回报型增量自然 Actor-Critic 学习和基于参数探索的期望最大策略搜索。第 10 章~第 13 章为基于谱方法的强化学习研究，包括基于拉普拉斯特征映射的启发式策略选择、Dyna 规划和迁移研究。为便于应用本书阐述的算法，书后附有部分强化学习算法 MATLAB 源程序。

本书可供理工科高等院校计算机科学、信息科学、人工智能和自动化技术及相关专业的教师与研究生阅读，也可供自然科学和工程领域中的研究人员参考。

图书在版编目(CIP)数据

强化学习原理及其应用/王雪松，朱美强，程玉虎著. —北京：科学出版社，2014.6
(智能科学技术著作丛书)
ISBN 978-7-03-040640-8

Ⅰ.①强… Ⅱ.①王… ②朱… ③程… Ⅲ.①学习方法－研究
Ⅳ.① G791

中国版本图书馆 CIP 数据核字(2014) 第 098840 号

责任编辑：惠 雪／责任校对：宋玲玲
责任印制：肖 兴／封面设计：许 瑞

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

双青印刷厂 印刷
科学出版社发行 各地新华书店经销
*
2014 年 6 月第 一 版 开本：B5 (720 × 1000)
2014 年 6 月第一次印刷 印张：17
字数：320 000
定价：89.00 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》编委会

《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。

“人工智能”学科自 1956 年诞生的五十余年来，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，在理论方面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 Intlligence，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

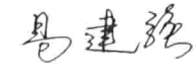
中国人工智能学会荣誉理事长

2005 年 12 月 18 日

序

作为一种重要的机器学习方法，强化学习不需要给定各种状态下的教师信号即可学习，对于求解复杂的优化决策问题具有广泛的应用前景。强化学习由控制理论、统计学和心理学等相关学科发展而来。经过多年的发展，强化学习目前已经成为一类求解序贯优化决策问题的有效方法。但大量研究结果仍然是针对小规模、离散状态和动作空间的问题，应用在大规模或连续状态和动作空间的优化决策问题中会出现“维数灾难”，导致学习效率不高。如何解决维数灾难，提高算法效率是现阶段强化学习面临的主要问题。该书的内容正是围绕着这一主要问题展开的，具有重要的学术价值。

该书是作者近年来在国家自然科学基金、教育部“新世纪优秀人才支持计划”、江苏省自然科学基金以及教育部博士学科点专项科研基金项目的资助下，取得的一系列关于强化学习研究成果的结晶，不仅是对已有研究成果的全面总结，也是对当前强化学习研究成果的重要补充。书中全面、系统地介绍了强化学习的基本概念、发展历史、分类及其部分主要算法，并重点围绕当前强化学习领域的热点问题展开研究，主要包括：基于值函数估计的强化学习方法、直接策略搜索强化学习方法和基于谱图理论的强化学习。此外，为理论联系实际和便于读者理解算法思想，书中还介绍了机器学习方法的若干典型应用，如倒立摆平衡控制、小船过河控制、电梯群控、机器人迷宫行走问题等。在阐述各种强化学习理论与核心技术时，均给出了研究的意义和必要性、算法思想、技术措施以及算法步骤；在阐述其应用时，均给出了应用背景、参数设置、算法对比结果等。该书学术思想新颖且内容范围广泛，写作结构清晰，逻辑性强，阐述严谨。相信该书的出版能进一步推动和促进强化学习领域的研究与发展。



中国科学院自动化研究所

2014 年 2 月 18 日

前　言

学习是人类具有的一种重要智能行为，而机器学习是一门研究怎样用计算机来模拟或实现人类学习活动的学科。机器学习从很多学科吸收了成果和概念，包括人工智能、概率论与数理统计、哲学、信息论、生物学、认知科学和控制论等，是多门学科有机交叉的新颖研究方向。机器学习的研究不仅是人工智能领域的核心问题，而且已成为近年来计算机科学与技术领域中最活跃的研究分支之一。

机器学习可以分为有监督学习、无监督学习和强化学习三类。不同于有监督学习和无监督学习的学习方式，强化学习是模拟人和高等哺乳动物的学习机制，强调在与环境的交互中“试错与改进”，其最大的特点是不需要系统模型即可实现无导师的在线学习。经过多年的发展，强化学习已经成为一类求解序贯优化决策问题的有效方法，在运筹学、计算科学和自动控制等领域得到广泛应用。但大量研究结果仍然是针对小规模、离散状态和动作空间的问题，应用在大规模或连续状态和动作空间的优化决策问题中会出现“维数灾难”，导致学习效率不高，甚至难以保证算法的收敛性。怎样解决大规模和复杂应用中的维数灾难、提高强化学习的效率，已成为现阶段强化学习的核心问题。本书的主要内容正是针对强化学习的这一核心问题而展开的。

著者长期从事强化学习的研究工作，在国家自然科学基金、教育部“新世纪优秀人才支持计划”、江苏省自然科学基金，以及教育部博士学科点专项科研基金项目资助下，提出了一系列提高强化学习算法效率的方法，并将其成功地应用于许多复杂的实际问题中。著者的这些工作大大丰富了强化学习理论，提高了强化学习方法解决实际问题的能力，也为强化学习方法在其他领域的进一步应用奠定了技术基础，具有重要的理论意义和实际应用价值。

本书是著者在国内外本领域权威期刊以及有影响的国际会议论文集上所发表的十余篇学术论文的基础上进一步加工、深化而成的，是对已有研究成果的全面总结。本书围绕着克服维数灾难，分别从值函数逼近、直接策略搜索和基于谱方法的学习等 3 个方面来阐述强化学习理论、方法及其应用，共 13 章。第 1~2 章为强化学习概述和相关基础理论，主要介绍强化学习的基本情况、研究及应用现状和相关基础理论；第 3~5 章为基于值函数估计的强化学习方法及其应用，包括：基于半参数支持向量机、概率型支持向量机的强化学习方法，基于测地高斯基的策略迭代方法和基于抽象状态的贝叶斯强化学习方法；第 6~9 章为直接策略搜索强化学习方法及其应用，包括：基于增量最小二乘时间差分的 Actor-Critic 学习，融合经

验数据的 Actor-Critic 强化学习，基于资格迹的折扣回报型增量自然 Actor-Critic 学习和基于参数探索的期望最大策略搜索；第 10~13 章是对基于谱方法的强化学习进行研究，包括：基于谱图理论的强化学习基础，基于拉普拉斯特征映射的启发式策略选择、Dyna 规划，基于谱方法的强化学习基函数与子任务策略混合迁移算法。为便于应用本书阐述的算法，书后附有部分强化学习算法源程序。著者愿将这些研究成果与国内外同行一起分享，以推动该领域的进一步研究与发展。

在本书的撰写过程中，参考了大量的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源泉，在此对所涉及的专家和研究人员表示衷心的感谢。著者得到中国科学院自动化研究所博士生导师易建强研究员多方面的指导，易建强研究员在百忙之中不但仔细审阅了全部书稿，提出了许多非常中肯的建议和意见，而且欣然为本书作序，令著者深受鼓舞，在此向易建强研究员表示衷心的感谢！中国矿业大学的马小平教授、李明教授等为本书的撰写提供了许多有益的指导。除此之外，已毕业的硕士研究生冯焕婷、张依阳等在校期间为本书的研究成果付出了辛勤的汗水。在本书的撰写、编辑、修改及参考文献整理、图形绘制方面，硕士研究生张嘉睿、闫称等同学也付出颇多。同时，科学出版社的编辑惠雪等为本书的出版做了大量辛苦而细致的工作，在此一并表示感谢。

强化学习是一个快速发展、多学科交叉的新颖研究方向，其理论及应用均有大量的问题尚待进一步深入地研究。由于著者学识水平和可获得资料的限制，书中尚有不妥之处，敬请同行专家和读者批评指正。

著 者

2014 年 1 月于中国矿业大学

目　录

《智能科学技术著作丛书》序
序
前言
第 1 章　强化学习概述 ····· 1
1.1　强化学习模型及其基本要素 ····· 2
1.1.1　强化学习模型 ····· 2
1.1.2　强化学习基本要素 ····· 3
1.2　强化学习的发展历史 ····· 5
1.2.1　试错学习 ····· 5
1.2.2　动态规划与最优控制 ····· 6
1.2.3　时间差分学习 ····· 7
1.3　强化学习研究概述 ····· 7
1.3.1　分层强化学习研究现状 ····· 8
1.3.2　近似强化学习研究现状 ····· 10
1.3.3　启发式回报函数设计研究现状 ····· 15
1.3.4　探索和利用平衡研究现状 ····· 16
1.3.5　基于谱图理论的强化学习研究现状 ····· 17
1.4　强化学习方法的应用 ····· 19
1.4.1　自适应优化控制中的应用 ····· 19
1.4.2　调度管理中的应用 ····· 22
1.4.3　人工智能问题求解中的应用 ····· 22
1.5　本书主要内容及安排 ····· 23
参考文献 ····· 25
第 2 章　强化学习基础理论 ····· 41
2.1　马尔科夫决策过程概述 ····· 41
2.1.1　马尔科夫决策过程 ····· 41
2.1.2　策略和值函数 ····· 42
2.2　基于模型的动态规划方法 ····· 44
2.2.1　线性规划 ····· 45
2.2.2　策略迭代 ····· 45

2.2.3 值迭代 ····· 46
2.2.4 广义策略迭代 ····· 47
2.3 模型未知的强化学习 ····· 48
2.3.1 强化学习基础 ····· 48
2.3.2 蒙特卡罗法 ····· 49
2.3.3 时间差分 TD 法 ····· 54
2.3.4 Q 学习与 SARSA 学习 ····· 56
2.3.5 Dyna 学习框架 ····· 57
2.3.6 直接策略方法 ····· 59
2.3.7 Actor-Critic 学习 ····· 60
2.4 近似强化学习 ····· 61
2.4.1 带值函数逼近的 TD 学习 ····· 61
2.4.2 近似值迭代 ····· 63
2.4.3 近似策略迭代 ····· 65
2.4.4 最小二乘策略迭代 ····· 66
2.5 本章小结 ····· 68
参考文献 ····· 68
第 3 章 基于支持向量机的强化学习 ····· 71
3.1 支持向量机原理 ····· 71
3.1.1 机器学习 ····· 72
3.1.2 核学习 ····· 73
3.1.3 SVM 的思想 ····· 74
3.1.4 SVM 的重要概念 ····· 74
3.2 基于半参数支持向量机的强化学习 ····· 75
3.2.1 基于半参数回归模型的 Q 学习结构 ····· 76
3.2.2 半参数回归模型的学习 ····· 78
3.2.3 仿真研究 ····· 79
3.3 基于概率型支持向量机的强化学习 ····· 82
3.3.1 基于概率型支持向量机分类机的 Q 学习 ····· 82
3.3.2 概率型支持向量分类机 ····· 83
3.3.3 仿真研究 ····· 85
3.4 本章小结 ····· 88
参考文献 ····· 88
第 4 章 基于状态-动作图测地高斯基的策略迭代强化学习 ····· 90
4.1 强化学习中的基函数选择 ····· 90

4.2 基于状态–动作图测地高斯基的策略迭代 ····· 91
4.2.1 MDP 的状态–动作空间图 ····· 92
4.2.2 状态–动作图上测地高斯核 ····· 93
4.2.3 基于状态–动作图测地高斯基的动作值函数逼近 ····· 94
4.3 算法步骤 ····· 95
4.4 仿真研究 ····· 96
4.5 本章小结 ····· 104
参考文献 ····· 104
第 5 章 基于抽象状态的贝叶斯强化学习电梯群组调度 ····· 106
5.1 电梯群组调度强化学习模型 ····· 107
5.2 基于抽象状态的贝叶斯强化学习电梯群组调度 ····· 108
5.2.1 状态空间抽象 ····· 109
5.2.2 强化学习系统的回报函数 ····· 110
5.2.3 贝叶斯网推断 ····· 110
5.2.4 状态–动作值函数的神经网络逼近 ····· 111
5.2.5 动作选择策略 ····· 112
5.3 仿真研究 ····· 112
5.4 本章小结 ····· 115
参考文献 ····· 115
第 6 章 基于增量最小二乘时间差分的 Actor-Critic 学习 ····· 117
6.1 策略梯度理论 ····· 118
6.2 基于常规梯度的增量式 Actor-Critic 学习 ····· 120
6.3 基于 iLSTD(λ) 的 Actor-Critic 学习 ····· 121
6.4 仿真研究 ····· 123
6.5 本章小结 ····· 126
参考文献 ····· 126
第 7 章 融合经验数据的 Actor-Critic 强化学习 ····· 128
7.1 增量式 Actor-Critic 学习算法的数据有效性改进 ····· 128
7.1.1 基于 RLSTD(λ) 或 iLSTD(λ) 的增量式 Actor-Critic 学习 ····· 130
7.1.2 算法步骤 ····· 132
7.1.3 仿真研究 ····· 133
7.2 基于自适应重要采样的 Actor-Critic 学习 ····· 140
7.2.1 基于最小二乘时间差分的 Actor-Critic 强化学习 ····· 141
7.2.2 基于重要采样的估计 ····· 143
7.2.3 基于自适应重要采样的估计 ····· 145

7.2.4　算法步骤 ····· 147
7.2.5　仿真研究 ····· 147
7.3　本章小结 ····· 150
参考文献 ····· 151
第 8 章　基于资格迹的折扣回报型增量自然 Actor-Critic 学习 ····· 153
8.1　自然梯度 ····· 154
8.2　自然策略梯度的估计方法 ····· 155
8.2.1　基于 Fisher 信息矩阵的自然策略梯度 ····· 155
8.2.2　基于兼容函数逼近器的自然策略梯度 ····· 156
8.2.3　自然策略梯度的仿真 ····· 157
8.2.4　自然策略梯度的特性 ····· 158
8.3　基于资格迹的折扣回报型增量自然 Actor-Critic 学习 ····· 158
8.4　仿真研究 ····· 161
8.5　本章小结 ····· 164
参考文献 ····· 165
第 9 章　基于参数探索的 EM 策略搜索 ····· 166
9.1　策略搜索强化学习方法分析 ····· 166
9.2　期望最大化策略搜索强化学习 ····· 167
9.3　基于参数探索的 EM 策略搜索学习 ····· 169
9.4　算法步骤 ····· 171
9.5　仿真研究 ····· 172
9.5.1　小球平衡问题 ····· 172
9.5.2　倒立摆平衡问题 ····· 175
9.6　本章小结 ····· 177
参考文献 ····· 178
第 10 章　基于谱图理论的强化学习基础 ····· 180
10.1　谱图理论与谱图分割 ····· 180
10.1.1　谱图理论与谱方法 ····· 180
10.1.2　谱图分割和谱聚类 ····· 181
10.2　基于谱图理论的流形和距离度量学习 ····· 183
10.2.1　流形学习概述 ····· 183
10.2.2　基于流形学习的度量学习 ····· 183
10.3　基于拉普拉斯特征映射法的强化学习 ····· 185
10.3.1　拉普拉斯特征映射法基础 ····· 185
10.3.2　基于拉普拉斯特征映射的强化学习 ····· 186

10.4　基于拉普拉斯特征映射的强化学习分析……190
10.5　本章小结……191
参考文献……191
第 11 章　基于拉普拉斯特征映射的启发式策略选择……194
11.1　探索和利用平衡问题概述……194
11.2　启发式策略选择原理……195
11.3　基于拉普拉斯特征映射的启发式策略选择……196
11.3.1　基本思想……196
11.3.2　基于拉普拉斯特征映射的启发式Q学习……197
11.4　算法步骤、计算复杂度和适用范围……202
11.4.1　算法主要步骤……202
11.4.2　计算复杂度……202
11.4.3　适用范围……203
11.5　仿真研究……203
11.5.1　5 房间格子世界……203
11.5.2　对称 4 房间格子世界……205
11.6　本章小结……206
参考文献……206
第 12 章　基于拉普拉斯特征映射的 Dyna 规划……208
12.1　强化学习在移动机器人自主导航中的应用研究概述……208
12.2　强化学习在井下救援机器人导航中的应用研究……209
12.3　基于拉普拉斯特征映射的 Dyna_Q 算法……210
12.3.1　Dyna_Q 的基本思想……210
12.3.2　基于谱图理论的优先级机制……211
12.3.3　算法步骤……212
12.3.4　计算复杂度分析和适用范围……212
12.4　仿真结果及分析……212
12.4.1　5 房间格子地图……213
12.4.2　对称 4 房间格子地图……213
12.4.3　9 房间格子地图……214
12.5　本章小结……215
参考文献……215
第 13 章　基于谱方法的强化学习迁移研究……217
13.1　基于谱图理论的强化学习迁移……217
13.1.1　强化学习迁移概述……217

13.1.2 基于谱图理论的强化学习迁移分析 ····· 219
13.2 基于谱图理论的 Option 自动生成研究 ····· 220
13.2.1 Option 原理 ····· 220
13.2.2 基于谱图分割的 Option 自动生成算法概述 ····· 221
13.2.3 虚拟值函数法 ····· 222
13.3 基于谱图理论的强化学习混合迁移方法 ····· 226
13.3.1 基函数的线性插值 ····· 226
13.3.2 迁移基函数的逼近能力 ····· 227
13.3.3 基函数与子任务策略的混合迁移 ····· 230
13.4 算法步骤和适用范围 ····· 231
13.4.1 算法步骤 ····· 231
13.4.2 适用范围 ····· 232
13.5 仿真实验与分析 ····· 232
13.5.1 地图不变迁移 ····· 233
13.5.2 地图比例放大迁移 ····· 233
13.5.3 实验结果统计分析 ····· 235
13.6 本章小结 ····· 237
参考文献 ····· 237
附录 ····· 240

第 1 章　强化学习概述

学习是人类智能的重要表现之一，人之所以能适应环境的变化并不断提高解决问题的能力，其原因在于人能通过学习积累经验，总结规律，以增长知识和才能，从而更好地改善自己的决策与行为。使计算机具有学习的能力，模拟或实现人类学习活动为目的的机器学习，是人工智能的一个重要研究领域，它的研究对于人工智能的进一步发展有着举足轻重的作用。机器学习 (machine learning) 一般定义为一个系统的自我改进的过程，以知识的自动获取和产生为研究目标[1]。机器学习的研究吸收了不同学科的成果和概念，包含了心理学、生理学、生物学、控制论、信息论、统计学以及人工智能在内的多种学科的交叉，具有很强的挑战性。

在机器学习范畴，依据从系统中获得反馈的不同，机器学习可以分为监督学习、无监督学习和强化学习三大类[2]。

监督学习 (supervised learning)，也称有导师的学习。这种学习方式需要外界存在一个 “教师”，它可对给定的一组输入提供应有的输出结果，而这种已知的输入-输出数据称为训练样本集，学习的目的是减少系统产生的实际输出和期望输出之间的误差，所产生的误差反馈给系统以指导学习。例如，在神经网络学习中，使用的是最小误差学习规则。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以 BP 算法为代表的监督式神经网络学习、归纳学习和基于实例的学习等。

无监督学习 (unsupervised learning)，又称无导师学习。它是指系统在不存在外部教师指导的情形下来构建其内部表征。这种类型的学习完全是开环的，例如在自组织特征映射神经网络中，网络的权值调节不受任何外来教师指导，但在网络内部能对基性能进行自适应调节。无监督学习中，系统的输入仅包含环境的状态信息，而不存在与环境的交互。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等。

研究者发现，生物进化过程中为适应环境而进行的学习有两个特点：一是人从来不是静止的被动地等待，而是主动地对环境作试探；二是环境对试探动作产生的反馈是评价性的，生物根据环境的评价来调整以后的行为，是一种从环境状态到行为映射的学习，具有以上特点的学习就是强化学习 (reinforcement learning)，或称再励学习、增强学习[3,4]。

这里需要指出的是，强化学习是一种与监督学习、无监督学习对等的学习模

式，而不是一种具体的计算方法，如神经网络、模糊推理、遗传算法等，但是这些计算方法可以与强化学习相结合。作为一种重要的机器学习方法，强化学习因不需要给定各种状态下的教师信号，则对于求解复杂的优化决策问题具有广泛的应用前景。

1.1　强化学习模型及其基本要素

1.1.1　强化学习模型

强化学习要解决的是这样的问题：一个能够感知环境的自治智能体 (Agent)，如何通过学习选择能够达到目标的最优动作，即强化学习 Agent 的任务就是学习从环境到动作的映射。强化学习不同于连接主义学习中的监督学习，主要表现在教师信号上，强化学习中由环境提供的强化信号是对 Agent 所产生动作的好坏作一种评价 (通常为标量信号)，而不是告诉 Agent 如何去产生正确的动作。由于外部环境提供了很少的信息，Agent 必须靠自身的经历进行学习。通过这种方式，Agent 在行动–评价的环境中获得知识，改进行动方案以适应环境。

Agent 为适应环境而采取的学习如果具有如下特征，则称为强化学习。

(1) Agent 不是静止的、被动的等待，而是主动对环境做出试探[4]；

(2) 环境对试探动作反馈的信息是评价性的 (好或坏)；

(3) Agent 在行动–评价的环境中获得知识，改进行动方案以适应环境，达到预期目的。

强化学习把学习看做是试探的过程，标准的 Agent 强化学习模型如图 1-1 所示[3,4]。在图 1-1 中，强化学习 Agent 接收环境状态的输入 s，根据内部的推理机制，输出相应的行为动作 a。环境在动作 a 的作用下，变迁到新的状态 s'，同时产生一个强化信号 (立即回报)r(奖励或惩罚) 反馈给 Agent，Agent 根据强化信号和环境当前状态选择下一个动作，选择的原则是使受到正的回报的概率增大。选择的动作不仅影响立即回报值，而且影响下一时刻的状态及最终强化值。在学习过程中，强化学习技术的基本原理是：如果系统某个动作导致环境正的回报，那么系统以后产生这个动作的趋势便会加强，反之系统产生这个动作的趋势便减弱。这和生理学中的条件反射原理是接近的。

可以看出，Agent 在与环境进行交互时，在每一时刻循环发生如下事件序列：

(1) Agent 感知当前的环境状态 s；

(2) 针对当前的状态和强化信号值，Agent 选择一个动作 a 执行；

(3) 当 Agent 所选择的动作作用于环境时，环境发生变化，即环境状态转移至新状态 s' 并给出强化信号 r；