《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。

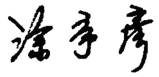
祝《智能科学技术著作丛书》出版，特赋贺诗一首：

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



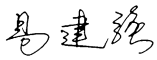
中国人工智能学会荣誉理事长

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

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；

(4) 强化信号 r 反馈给 Agent。

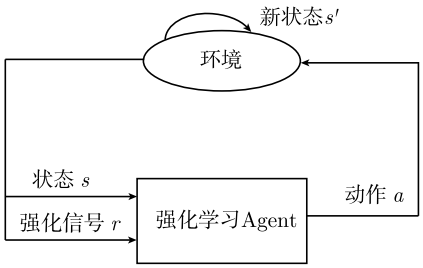


图 1-1 强化学习模型

强化学习具有如下特点[3]：

(1) 强化学习是一种弱的学习方式，体现为：Agent 通过与环境不断地试错交互来进行学习；强化信号可能是稀疏且合理延迟的；不要求 (或要求较少) 先验知识；Agent 在学习中所使用的反馈是一种数值回报形式，不要求有提供正确答案的教师，即环境返回的强化信号是 r，而不像监督学习中给出的教师信号 (s, a)；

(2) 强化学习是一种增量式学习，并可以在线使用；

(3) 强化学习可以应用于不确定性环境；

(4) 强化学习的体系结构是可扩展的。目前，强化学习系统已扩展至规划的合并、智能探索、监督学习和结构控制等领域。

1.1.2 强化学习基本要素

由强化学习模型可以看出，一个强化学习系统除了 Agent 和环境之外，主要还有 4 个基本元素：策略、值函数、回报函数和环境模型 (非必需)。这 4 个基本元素及其关系如图 1-2 所示[5]。强化学习系统所面临的环境由环境模型定义，但由于模型中状态转移概率函数和回报函数未知，Agent 只能够依赖于每次通过试错所获得的立即回报来选择策略。而在选择行为策略过程中，要考虑到环境模型的不确定性和目标的长远性，因此在策略和立即回报之间构造值函数 (即状态的效用函数)，用于策略的选择。

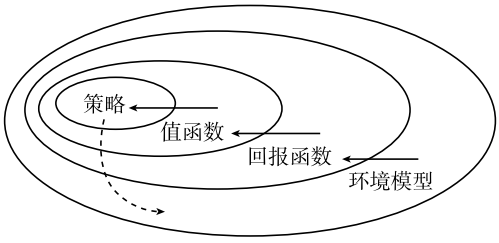


图 1-2 强化学习基本要素及其关系

1. 策略

策略 (policy) 定义了 Agent 在给定时刻的行为方式，直接决定了 Agent 的动作，是强化学习的核心。策略的定义如下：

定义 1.1(策略) Agent 在与环境交互过程中选择动作的方法称为策略 $\pi : S \times A \to [0,1]$，$S$ 为状态空间，A 为动作空间，$\pi(s,a)$ 表示在状态 s 下选择动作 a 的概率。策略的一个退化形式为 $\pi : S \to A$，称为确定性策略，表示在状态 s 下动作 $\pi(s)$ 的执行概率为 1，其他动作的概率执行均为 0。

关于任意状态所能选择的策略组成的集合 F，称为允许策略集合，$\pi \in F$。在允许策略集合中存在的使问题具有最优效果的策略 π^* 称为最优策略。策略与心理学中的刺激–反射 (stimulus-response) 规则相对应，在某些情况下策略可能是一个简单的函数或者查找表 (lookup table)；而在另一些情况下则可能需要大量的计算，例如搜索过程等。强化学习方法确定了 Agent 怎样根据经验改变其策略。

2. 回报函数

回报函数 (reward function) 定义了一个强化学习问题的目标，它将感知的环境状态 (或状态–动作对) 映射到一个强化信号 r，对产生的动作的好坏作一种评价。强化信号通常是一个标量，例如用正数表示奖赏，而用负数表示惩罚。强化学习的目的就是使 Agent 最终得到的总的回报值达到最大。回报函数往往是确定的、客观的，可以作为改变策略的标准。

3. 值函数

回报函数表明眼前哪些是好的，是一种 “近视” 的表达信号，而值函数 (value function)(即状态的效用函数，又称评价函数) 则是 “远视” 的表征，它表达了从长远的角度来看哪些是好的。状态的值所表示的意义，大致来说，是从该状态起智能体所能积累的回报的总和。回报是环境给出的立即评价，而值函数则是随后一系列状态所对应的回报的累积。回报和值函数的联系是：没有回报就没有值函数，估计值函数的目的是为了获得更多的回报。举例来说，一个状态可能产生一个较低的立即回报，但是从长远看来可能会带来丰厚的回报和。因此，在选择行为时，通常会依据值函数做出决策而不是回报函数。选择那些能带来最大值函数的行为，而不是选择那些能带来最大回报的行为。但如何确定值函数要比确定回报函数困难得多，回报通常是由环境直接给出的，但值函数一般来说要进行估计。事实上，几乎所有强化学习算法的核心都是如何有效地估计值函数。

4. 环境模型

环境模型 (model of environment) 是某些强化学习系统的一个可选的组成部

分。环境模型就是模拟环境的行为方式。例如，给定一个状态和动作，模型可以预测下一个状态和回报。利用环境的模型，Agent 在做决策的同时可以考虑未来可能发生但尚未实际经历的情形，从而进行规划 (planning)。将模型和规划加入到强化学习系统是一个比较新的发展方向，它联系了强化学习与动态规划等其他基于模型或部分模型的方法，形成了一些较实用的新方法。

1.2 强化学习的发展历史

强化学习的发展主要包括 3 条主线：试错 (trial-and-error) 学习、动态规划与最优控制和时间差分 (temporal difference，TD) 学习 (图 1-3)。在经历了各自不同的进程之后，最终在 20 世纪 80 年代形成了现代强化学习的基本框架[6,8]。

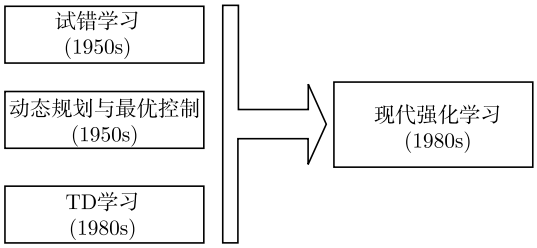


图 1-3 强化学习的发展主线

1.2.1 试错学习

第一条主线源于动物学习的心理学，通过试错达到学习的目的。这条主线贯穿于人工智能早期的研究工作中，也是使 20 世纪 80 年初期强化学习复苏的重要因素。最早简洁表述 “试错学习” 的是 Thorndike，他于 1911 年提出了 “响应定律”(law of effect)。Thorndike[9] 认为：“对于同一环境所做的几个响应，当那些伴随或紧跟着的响应使动物意愿得到满足且其他的条件相同时，对环境的联系将会被加强，所以，当环境重现时，这些响应重现的概率将更大。当那些同时或紧跟着的响应使动物的意愿受挫且在相同的其他条件下，与环境的联系将会削弱，所以，当环境重现时，它们出现的概率将越小。得到的满足程度越大，响应和环境的联系加强得越多。而不满足的程度越大，响应与环境的联系削弱得越多。”

从上述描述可以看出，Thorndike 思想的本质是：强调行为的结果有优劣之分并成为行为选择的依据，同时指出能够导致正回报的行为选择概率将增加，而导致负回报的行为选择概率则降低。Thorndike 的思想包含了试错的 2 个重要特点：选择性和联想性。选择性就是尝试学习不同动作并比较不同结果；联想性是指将可选择的动作与特定的状态联系在一起。进化学习中的自然选择具有选择性，但不具有

联想性；监督学习则仅具有联想性而不具有选择性。另外，“响应定律” 还反映了强化学习的另两个重要特性，即搜索和记忆。

虽然在心理学和其他学科上，响应定律都曾引起过非常多的争论，但是这些年来，由于其基本思想已被实验所证实，而且从直觉上看非常正确，所以该定律十分具有影响力。这是一种将搜索与记忆相结合的方法，它能够在多种试验动作中搜索，然后记住效果最好的动作。相对于监督学习原则来说，“响应定律” 是一种依靠选择的学习原则。

最早进行试错学习研究的可能是 Minsky、Farley 和 Clark 等于 1954 年开始的。Minsky 在他的博士论文中描述了一种叫作 SNARC(stochastic neural-analog reinforcement calculator) 的模拟机[10]，而 Farley 和 Clark 提出了另一个神经网络学习机。1961 年，Minsky 进一步讨论了几个与强化学习相关的主题，如信度分配问题 (credit assignment problem)，这个问题是强化学习必须涉及的，也是至今学者研究最多的难点[11]。后来，Waltz 和傅京逊于 1965 年，Mendel 于 1966 年在工程文献中较早地引用了 “强化” 和 “强化学习” 等概念[12]。1955 年，Farley 和 Clark 由 “试错学习” 转向泛化和模式识别的研究，即由强化学习转向监督学习，开始了强化学习和监督学习的混合研究状态[4]。Widrow 及其同事们在研究监督学习的时候，认识到监督学习和强化学习之间的不同。1973 年，Widrow、Gupta 和 Maitra 改正了 Widrow 和 Hoff 的监督学习规则 (常称为 LMS 规则) 得到新的学习规则。新规则可实现强化学习，即根据成功和失败的信号进行学习，代替了原来的使用训练样本进行学习的方法，他们用 “有评价的学习” 一词代替 “有教师的学习”[13]。另外，学习自动机对由试错学习发展起来的现代强化学习有着重要的影响，其中较为著名的有 1973 年 Tsetlin 的工作，以及 Barto 和 Anandan 发展的具有联想的学习自动机[4]。不过大部分早期的研究工作，主要是显示强化学习和监督学习的不同。

1.2.2 动态规划与最优控制

第二条主线是利用值函数和动态规划方法求解最优控制问题。最优控制一词在 20 世纪 50 年代末被用来描述为最小化动态系统行为的量度值而设计控制器的问题[14]。这种方法将动态系统的状态和值函数的概念用于定义函数方程 (现在通常被称为 Bellman 方程)。这类通过求解 Bellman 方程来解决最优控制问题的方法被称为动态规划。动态规划在过去的几十年中已取得了极大的发展，被广泛地认为是求解一般随机最优控制问题的唯一切实可行的方法。但是，动态规划存在 Bellman 所谓的 “维数灾难”(curse of dimensionality) 的问题，也就是说，动态规划的计算量需求随状态变量数目的增加而呈指数级增长。但较其他方法而言，动态规划仍是一个非常有效且应用广泛的方法。动态规划方法与强化学习密切相关，对于马尔科夫决策问题 (Markov decision process，MDP)，前者主要解决环境的状态转移概率和

回报函数模型的已知的决策问题，而后者主要处理状态转移概率和回报函数模型未知的情形。很多强化学习的方法与思想都来源于动态规划。

1.2.3 时间差分学习

在动物学习心理学中，与强化学习密切相关的另一个研究内容是时间差分 (temporal difference，TD) 学习，这同时也是强化学习的第三条发展主线。所谓时间差分，是指对同一个事件或变量在连续两个时刻观测的差值，这一概念来自于动物学习心理学中有关二次刺激的研究[15]。二次刺激是和食物、疼痛等第一刺激对应的刺激，两者有相同的属性。Minsky 是第一个意识到第二刺激可以应用于人工智能的学者。Samuel 于 1959 年第一次提出基于时间差分概念的学习并用于西洋跳棋[16]。Samuel 并没有和 Minsky 的工作有联系，其灵感来自于 Shannon 的建议：电脑可以通过评估函数来进行象棋游戏，并通过在线函数提高电脑水平。Minsky 于 1961 年在前者的基础上作了进一步的工作，并揭示了两者的关系。1972 年，Klopf 将试错学习与时间差分学习结合起来[4]。他对学习机用于大系统进行研究，并对局部强化学习产生兴趣，提出了学习系统的子系统可以对其他部分提供支持，并将其称为普遍强化。随后，Sutton、Barto 和 Friston 等学者作了更多、更为细致的研究[17,18]。1989 年，Watkins 将最优控制和时间差分学习结合，提出了现在广为使用的 Q 学习算法[19]。至今，仍有很多学者提出不同的 TD 改进算法，并有大量应用。

1.3 强化学习研究概述

20 世纪 90 年代，强化学习通过与运筹学、控制理论的交叉结合，在理论和算法方面取得若干突破性的研究成果，奠定了强化学习的理论基础，并在机器人控制、优化调度等序贯决策问题中取得成功的应用[1,2]。在经典强化学习中，回报具有延迟性的特点和通常采用的试错改进机制决定了智能体仅能依据学习中获取的稀疏回报来改进策略，而忽略了大量其他有用的信息和知识。同时，由于缺少有效的策略选择机制，致使智能体在状态空间中盲目地 “游荡”，给值函数的迭代带来不必要的开销。这两方面的原因使得强化学习算法需要对状态–动作序列进行 “无限” 或者 “足够” 多次的遍历才能收敛，所以在复杂问题中会出现维数灾难[4]。怎样解决大规模和复杂应用中的维数灾难，探索与利用的两难问题，已成为强化学习的核心问题。现阶段，解决这一难题的研究主流是对状态和动作空间进行抽象和泛化 (abstract and generalization)[20,21]。

围绕着泛化与抽象，分层强化学习 (hierarchical reinforcement learning, HRL)、近似强化学习 (approximate reinforcement learning，ARL)、关系强化学习 (relational reinforcement learning，RRL)、迁移强化学习 (transfer learning for reinforcement

learning) 等众多解决维数灾难的方案被提出[16-19]。从提高算法效率角度出发，启发式回报函数设计 (reward shaping，RS)、启发式策略选择等方法也被重点研究。总体说来，从知识的使用角度考虑，上述众多方法抛弃了智能体 “一无所知” 的假设，在考虑能与强化学习四要素 (模型、立即回报、值函数和策略) 有机融合的基础上，隐式或显式的利用先验或过程领域知识及相关模型[22,23]；从学习方式的角度出发，它们是各种监督学习、非监督学习和其他人工智能方法与强化学习的结合。

除此之外，基于贝叶斯理论的强化学习，多智能体、部分可观察 MDP(partially observable Markov decision process, POMPD) 领域的强化学习近年来也产生了丰富的研究成果[24-26]。本书重点综述分层强化学习、近似强化学习、启发式回报函数设计、探索与利用平衡问题和基于谱图理论的强化学习等方面的研究现状。

1.3.1 分层强化学习研究现状

根据经验，自然界和人类社会中的大部分复杂系统都具有层次结构，这种结构既为简单进化到复杂提供了可能，也为简化行为过程和描述方式奠定了基础[27]。同理，复杂系统的强化学习问题也可采用分层的方式得以化简。分层强化学习就采用了上述思想，利用分而治之原则，把一个复杂问题在不同层次上抽象为多个相对简单问题来求解，可有效地解决强化学习的维数灾难问题[28]。

分层强化学习 (HRL) 的核心思想是抽象分层，抽象机制允许强化学习系统忽略与当前子任务无关的细节。HRL 提出后引起了广大研究者的关注，近几年取得了显著进展，先后基于半马尔科夫决策过程 (semi-Markov decision process，SMDP) 提出了 Option、HAMs、MAXQ 和 HEXQ 等 4 类方法[28-30]。在标准的强化学习收敛条件下，Option、HAMs 收敛到最优策略解，MAXQ 收敛到递归最优解。某种意义上，HEXQ 可以视作为 Option 和 MAXQ 的自动分层版本。因此，对于确定性的最短路径问题，HEXQ 收敛到最优策略解；对于随机动作情况，则收敛到递归最优解。

在早期的研究中，HRL 中的子任务和分层结构都是预先定义的，不能满足动态未知环境的应用需求，所以任务的自动分层成为研究重点，HRL 的主要研究方向都以此展开。自动发现子任务的方法较多，根据其所用思想可归结为瓶颈和路标法、共用子空间法、多维状态法和马氏空间法等几类[30]。下面以基本的 4 类方法为主线对 HRL 进行综述。

1. 基于Option的研究现状

Option 方法将子任务抽象为一个 Option，学习过程中只在子任务完成时才决策，其他时刻按照子任务内部定义的动作执行，从而将单步动作拓展到多步情形，减少了学习的决策次数[29]。

四类 HRL 中，Option 方法由于简单实用，迁移容易而备受青睐，因此在该方法中的自动分层方法研究的最为丰富。根据任务分解的方式的不同，基于 Option 的自动分层方法又可分为瓶颈和路标法、聚类法和因子化 MDP 法 3 类。瓶颈法主要用于状态空间有明显分段或分区性质的任务；聚类法的应用范围则不局限于此；因子化 MDP 法适用于状态表达式可以因子化，并且因子化后各变量具有不相关性的任务。

1) 瓶颈和路标法

瓶颈和路标法的基本思想是寻找问题求解过程中的关键点，并将这些关键点作为子目标对任务进行分解[30]。寻找子目标的途径主要有状态访问频率法、值函数梯度法和基于图论的方法 3 类。

根据任务瓶颈的特性，成功路径中高信号梯度和访问频率的状态都可能是子目标。基于这种思想，文献 [31，32] 分别提出了值函数梯度法和状态访问频率法。值函数梯度法对于奖励信号延迟的强化学习失效。状态访问频率法存在将目标状态很近的非瓶颈点也作为子目标的缺点，因此需要对环境进行额外的探索来区分“重要” 和 “不重要” 的状态。为了解决此问题，文献 [33，34] 分别引入了时间度量和落差变化率来辅助寻找子目标。

基于图论的方法将状态空间的连接关系用图来表示，将任务分解问题转换为图的分割问题。相对于其他方法，该方法因有较好的理论基础而备受关注，提出了众多算法，其典型工作有：文献 [35] 将传统的最大流最小割算法用于任务分解提出了 Q-cut 算法；文献 [36, 37] 将谱图分割理论引入 HRL，得到了 L-cut 和级联分解算法 KCD；文献 [38] 将复杂图论中的重心 (betweeness) 用于子目标的判定，得到了一类基于重心的分割方法。

2) 聚类法

瓶颈法是针对单个状态的，而 Mannor 认为要完成一个复杂目标要经历几个中间阶段，每个阶段由多个状态组成，因此可把状态组成的聚类看作学习过程的中间阶段[39]。聚类的方法并不直接搜索和发现子目标，而是去寻找中间阶段，具有更好的适应性和鲁棒性。根据这种思路，基于人工免疫网络[30]、禁忌搜索[40]、K 聚类[41] 等技术提出了众多的任务分解方案。瓶颈法和聚类法都是对状态空间进行抽象，其本质是一样的，只是具体实现方式和适用范围不同而已。聚类方法中，两个聚类的边界状态可视为是瓶颈状态。

3) 因子化 MDP 法

当问题的状态可以用特征变量表示时，根据特征变量划分子任务是一种可行的候选策略，则 HEXQ 法[42] 就基于此，该方法的介绍和研究进展详见 1.3.1 节第 3 部分。

2. 基于MAXQ的研究现状

MAXQ 不直接将问题简化为单个 SMDP，而是以任务图的形式建立起可以同时学习的分层 SMDP[43]。该方法定义的递归最优策略不受子任务的上下文的限制，因此便于作为 “积木” 或 “构件” 应用于其他任务中。最初的 MAXQ 方法是针对于离散 MDP 环境的，Ghavaxnzadeh 和 Tang 等[44,45] 利用层次策略梯度方法将其扩展到连续环境的应用中。

对于 MAXQ 方法而言，自动产生任务图绝非易事，因此自动分层的 MAXQ 方法的研究成果不多，只有一些类 MAXQ 的分层方法。上面提到的的 HEXQ 就是其中一种。除此之外，文献 [46] 采用约束优化的思想，对采样序列进行搜索来产生任务层次；文献 [47, 48] 引入遗传算法与图模型来构造任务图。

3. 基于HEXQ的研究现状

HEXQ 将状态的表达方式因子化，通过排序每个因子分量的更新频率来获取子任务[42]。HEXQ 构造的任务分层中，每个状态变量为一层，每层包含一个简单的 MDP，该 MDP 与其他简单 MDP 通过瓶颈状态集连通。HEXQ 可以用于 Option 的生成，也能得到类似于 MAXQ 方法的任务层次。在最初的 HEXQ 算法中，变量的因子化排序具有一定启发性，算法的效率不稳定。为此，文献 [49，50] 利用贝叶斯网络构建因果关系图来解决该问题，分别提出了 VISA 算法和 HI-MAT 算法。这两种算法的区别在于建图的方式不一样。2008 年，文献 [51] 引入了偏序规划，进一步扩展了 HEXQ 的应用范围。

4. 基于HAMs的研究现状

HAMs 算法通过将每个子任务抽象为一个建立在 MDP 上的随机有限状态机，从而约束策略空间来简化决策过程，可应用于部分可观测领域[52]。HAMs 的任务分解实际上是一个构造有限状态机的过程，主要在策略空间上进行。

HAMs 中用有限状态机来表达 MDP 状态空间的区域策略，但有限状态机作为一种先验知识的表述方式，存在表述能力较差的不足。因此，文献 [53, 54] 引入高级编程语言 A_lisp 来替代有限状态机从而提出 PHAM 方法。文献 [55] 对 HAMs 方法进行深入分析，提出策略耦合 SMDP 的观点，并与 HEXQ 方法相结合，在考虑抽象性和层次性的情况下，提出新的分解方法。

1.3.2 近似强化学习研究现状

对于大规模或连续空间的 MDP 问题，智能体不可能遍历所有的状态–动作对，因此需要强化学习的值函数具有一定泛化能力，即利用有限的学习经验和记忆实现对一个大范围空间知识的有效获取和表示。强化学习中映射关系包括：状态空间到