强化学习原理

第一章读书笔记

姓名：石若川 学号：2111381 专业：智能科学与技术

1 导论

1.1 强化学习

1. 强化学习的概念

强化学习是学习做什么（即如何把当前情境映射为动作）才能使数值化的收益最大化。

2. 两个显著特征

试错搜索（trial-and-error search）与延迟收益（delayed reward）

3. 不完全可知的马尔科夫决策过程（incompletely-known Markov decision processes）

基本思想：在智能体为了实现目标而不断与环境产生交互的过程中，抓住智能体所面对的真实问题的主要方面。具备学习能力的智能体必须能够在某种程度上感知环境的状态，然后采取动作并影响环境状态。智能体必须同时拥有和环境状态相关的一个或多个明确的目标。

马尔可夫决策过程就包含三个方面：感知（sensation）、动作（action）和目标（goal）。

4. 强化学习与监督学习、无监督学习的区别

- 与监督学习

  监督学习利用带标注的数据集进行学习，目的在于让系统具备判断或泛化能力，并不适用从交互中学习的问题。强化学习为了在未知领域达到收益最大，必须从自己的经验中学习。（非泛化的）

- 与无监督学习

  无监督学习的目的在于寻找未标注数据集中的隐含结构，强化学习并不需要做到这一点。

- 此外，强化学习明确地考虑了目标导向的智能体与不确定的环境交互这整个问题。而很多其他方法都是只考虑子问题，而忽视了子问题在更大情境下的适用性。

5. 强化学习面临的挑战

智能体会偏爱过去它有效产生过收益的动作（利用 exploitation），但也需要尝试未选择过的动作（探索 exploration）。如何权衡利用与探索，是强化学习所面临的挑战。

6. 强方法与弱方法

基于一般原则的方法，例如搜索或学习，为弱方法（weak methods）。基于知识的方法为强方法（strong methods）。强化学习研究在追求更简单的人工智能普适原则。

1.2 示例

下棋的决策、石油控制的实时调整、新生羚羊的奔跑……以上例子均反映了智能体与环境的交互，在不确定的环境中智能体想要实现一个目标。智能体的动作会影响环境未来的状态，进而影响未来的决策。智能体可以通过经验来改进性能。

1.3 强化学习要素

除了智能体和环境之外，强化学习有四个核心要素：策略（policy）、收益信号（reward signal）、价值函数（value function）和（可选的）对环境建立的模型（model）。

1. 策略

策略定义了智能体在特定时间的行为方式，是环境状态到动作的映射。策略本身可以决定行为，因此策略是强化学习智能体的核心。一般来说，策略可能是环境所在的状态和智能体所采取的动作的随机函数。

2. 收益信号

收益信号定义了强化学习问题中的目标。智能体的唯一目标是最大化长期总收益。社会信号是改变策略的主要基础。如果策略选择的动作导致了低收益，那么可能会改变策略。从而在未来的这种情况下，选择一些其他的动作。一般来说，收益信号可能是环境状态和在此基础上采取的动作的随机函数。

3. 价值函数

收益信号表明了短时间内什么是好的，而价值函数则表示了长远来说什么是好的。一个状态的价值是一个智能体从这个状态开始，对将来累计的总收益的期望。某状态的即时收益可能很低，但它仍可能具有很高的价值，因为之后某一时刻可能会出现高收益的状态。

在制定和评估策略时，我们最关心的是价值。动作选择是基于对价值的判断作出的。我们寻求能带来最高价值而不是最高收益状态的动作，因为这些动作从长远来看，会为我们带来最大的累计收益。

收益是由环境直接决定的，而价值必须综合评估，并根据智能体在整个过程中观察到的收益序列重新估计。价值评估方法几乎是所有学强化学习算法中最重要的组成部分。

4. 对环境建立的模型

模型是对环境反应模式的模拟，它允许对外部环境的行为进行判断。环境模型会被用于规划。在外部环境发生变化之前，先考虑未来可能发生的各种情形，从而预先决定采取何种动作。使用环境模型和规划来解决强化学习问题的方法称为有模型的方法。而简单的无模型的方法则是直接的试错。

1.4 局限性与适用范围

1. 强化学习的局限性

强化学习很依赖“状态”的概念，它既作为策略和价值函数的输入，又作为模型的输入与输出。

2. 进化算法与强化学习

进化算法不显式地计算价值函数，采取大量静态策略，每个策略在较长时间内与环境的一个独立实例进行交互，通过选择最多收益的策略及其变种来产生下一代的策略。

进化算法适用于以下情形：策略空间充分小；可以很好地结构化以周到好的策略；有充分的时间搜索；智能体不能精确感知环境状态的问题。

与强化学习相比，进化算法并不在与环境互动中学习的方法。进化算法忽视了所求策略是状态到动作的函数的事实，也没有注意到个体在生命周期中所经历的状态和动作。

1.5 扩展示例：井字棋

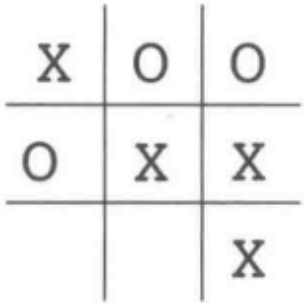

图 1.1: 井字棋

假设玩家存在破绽，有时候会输掉比赛。

- “极大极小” 算法：假设对手会按照某种特定方式玩游戏。这种情况下，玩家不会让游戏陷入可能会输的状态。
- 动态规划：需要输入对手在每种状态下下每一步棋的概率，但是这样的先验信息不可知。所以需要先学习对手走棋动作的模型，以一定的置信度，用动态规划的方式针对近似的对手模型计算最优解。
- 进化算法：直接在策略空间中搜索赢的概率高的策略，每种策略的获胜概率可以从与对手多次博弈中估算。利用爬山法、遗传算法、进化算法等进行搜索。
- 强化学习：建立数字表，每格表示一个游戏的可能状态。每个数字表示对获胜概率的最新估计（即价值）。获胜概率为 1 代表胜利，获胜概率为 0 代表无法取胜，其他状态的初始价值为 0.5。

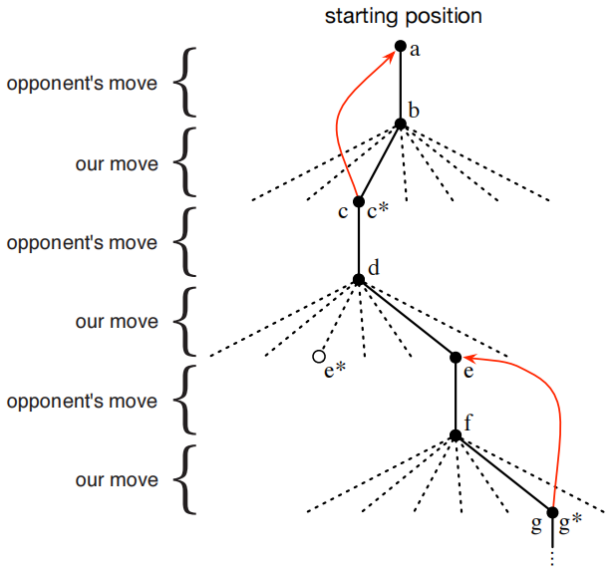

图 1.2: 走棋序列

在和对手多次对弈的过程中，大部分时候选择价值最高的动作（利用），偶尔从其他动作中随机

选择（探索），如图1.2。价值更新的过程通过下式更新

$$V(S_t) \leftarrow V(S_t) + \alpha\Big[V(S_{t+1}) - V(S_t)\Big]$$

其中，S_t 为做出动作前的状态，S_{t+1} 为做出动作后的状态，S_t 的价值用 $V(S_t)$ 表示，α 为一个小的正分数，称为步长参数，会影响学习速率。若步长参数睡时间适当减小，对于任意固定对手，该方法会收敛于最优策略下每个状态下的真正获胜概率。

进化算法只会考虑每局比赛的最后结果，而忽略博弈的过程。相反强化学习会对每个状态进行评估，利用博弈过程中的可用信息。

人工神经网络可以和强化学习相结合。神经网络为程序提供从经验中进行归纳的能力，因此在新的状态中，它根据保存的过去相识状态的信息来选择动作，由神经网络做出最后的决策。

强化学习还可以应用于部分状态被隐藏、智能体对不同状态感知不出区别或者没有环境模型情况。强化学习可以用于系统的总体控制，也可以用于细节控制。

1.6 总结

强化学习是一种对目标导向的学习与决策问题进行理解和自动处理的计算方法。它根据智能体通过与环境的直接互动来学习，而不需要监督或建模。

强化学习利用马尔科夫决策的框架，使用状态、动作和收益定义智能体与环境互动的过程。该框架反映了对因果关系、不确定性和显式目标存在性的认知。

价值和价值函数是强化学习的重要特征，对于策略空间的搜索十分重要，也是进化算法与强化学习不同之处。

1.7 强化学习的早期历史

强化学习的发展有两条历史主线，在交汇于现代强化学习之前它们是相互独立的。其中一条是源于源于动物学习心理学的试错法。这条主线贯穿了一些人工智能最早期的工作，并在二十世纪八十年代早期激发了强化学习的复兴。另一条主线则关注最优控制的问题，以及使用价值函数和动态规划的解决方案。这两条主线都与第三条关注时序差分方法的主线有一定程度的关联。在二十世纪八十年代末，这三条主线交汇在一起，产生了现代的强化学习领域。

1.7.1 第一条主线：最优控制

“最优控制”这一术语最早使用于二十世纪五十年代末，用来描述设计控制器的问题。其设计的目标是使得动态系统随时间变化的某种度量最小化或最大化。

- 20 世纪 50 年代中期，由 Richard Bellman 等人开发的针对这一问题的其中一种方法。该方法是 19 世纪 Hamilton 和 Jacobi 理论的延伸。该方法运用了动态系统状态和价值函数，或称为“最优回报函数”的概念。其定义了一个函数方程，称为贝尔曼方程。通过求解这个方程来解决最优控制问题的。这类方法被称为动态规划。
- 1957 年，Bellman 也提出了最优控制问题的离散随机版本，被称作马尔可夫决策过程。
- 1960 年，Ronald Howard 又设计出了 MDP 的策略迭代方法。

动态规划被普遍认为是解决一般随机最优控制问题的唯一可行方法。它遭受了所谓的维度灾难问题，这意味着它的计算需求随着状态变量的数量增加，呈指数级增长，但它仍然比一般方法更有效。但另一方面，最优控制和动态规划之间联系的认知过程却十分缓慢。这可能是由于学科之间的隔离以及它们的不同目标。另外，作为一种离线计算，动态规划主要依赖于精确的系统模型和贝尔曼方程的解析解。此外，动态规划的最简形态是延时间线反向推进的计算，使得人们很难看出它如何能够被进行前向计算的学习过程所利用。

- 1977 年，Witten 的工作被认为是学习和动态规划的思想结合。
- 1987 年，Werbos 明确论证了动态规划和学习方法之间的更紧密的相互关系，以及动态规划与理解神经和认知机制的相关性。
- 1989 年，Chris Watkins 用 MDP 形式对待强化学习，首次将动态规划方法与在线学习方法进行完全整合。
- 1996 年，Dimitri Bertsekas 和 John Tsitsiklis 创造的术语“神经动态规划”指的是动态规划和人工神经网络的结合。

1.7.2 第二条主线：试错学习

试错学习思想可以追溯到 19 世纪五十年代，Alexander Bain 对探索和实验学习方法的讨论，可以更具体的追溯到 1894 年，英国动物行为学家和心理学家 Conway Lloyd Morgan 使用这个术语来描述他对动物行为的观测实验。

- 1911 年，Edward Thorndike 提出了效应定律，描述了强化事件对选择行为倾向性的影响。他也许是第一个简洁明确表示出试错学习本质是学习原则的人。
- 1927 年，巴甫洛夫在条件反射著作中首次提出了“强化”一词。巴甫洛夫认为，强化就是动物行为模式的增强，它来源于动物受到增强剂的刺激后，与另一刺激或反应形成的短暂关系。后来一些心理学家扩展了“强化”一词的意义，也包括了弱化过程。同时，它还适用于对刺激事件的忽略或终止。
- 1948 年，图灵描述了一种“快乐-痛苦系统”的设计。试错思想在计算机中的应用最早出现于关于人工智能可能性的思考中。

之后，研究者利用试错思想研制了一系列电子机械设备。

- 1933 年，ThomasRoss 制造的一台机器，它能够穿越迷宫且通过开关设置记住路线。这是最早演示试错学习的电子机械设备。
- 1951 年，已经因为“机械乌龟”(1950) 成名的 W.Grey Walter 又制造了能够简单学习的版本。
- 1952 年，Claude Shannon 演示了一种名叫 Theseus 的迷宫老鼠，它利用试错法在迷宫中摸索，迷宫本身通过磁铁和继电器在地板上记录成功的路径。
- 1954 年，J.A.Deutsch 描述了一个以他的类似于基于模型的强化学习的行为理论 (1953) 为基础的解迷宫机器。
- 1954 年，Marvin Minsky 在他的博士论文中讨论了强化学习的计算方法，描述了他组装的台基于模拟信号的机器，他称其为“随机神经模拟强化计算器”，SNARCs(Stochastic Neural-Analog Reinforcement Calculators)，模拟可修改的大脑突触连接。

使用数字计算机，通过编程来进行各种类型的机器学习成为可能，其中一些也实现了试错学习。

- 1954 年 Farley 和 Clark 描述了一种通过试错学习的神经网络学习机器的数字化仿真程序。但他们很快就从试错学习转向推广性和模式识别，即从强化学习转向有监督学习。这时这些学习类型之间的关系开始出现混乱。许多研究人员认为自己在研究强化学习，但其实是在研究有监督学习。例如，Rosenblatt(1962) 和 Widrow 及 Hof(1960) 这样的神经网络先驱们显然是被强化学习所激励。虽然他们使用了“收益”和“惩罚”这样的语言，但他们所研究的系统是有监督的学习系统，适用于模式识别和感知学习。

- 1960 年，Minsky 在论文中讨论了几个关于试错学习的问题，包括预测、期望，以及他所称的“复杂强化学习系统中的基础性的功劳分配问题”：对于一项成功所涉及的许多项决策，如何为每项决策分配功劳?

- 1963 年，Andreae 开发了个叫作 STeLLA 的系统，它通过与环境的互动中的试错来学习。这个系统包括了关于环境的内部模型和后来开发的一个用来处理隐藏状态问题的“内心独白”模块(Andreae,1969)。Andreae 后来的工作 (1977) 虽然更强调从老师那儿学习，但仍然包括了很多反复试错，并且系统的目标之一就是产生创造性的新事件。这个工作的一个特性被称为“回流过程”，在 Andreae(1998) 中有详细描述，其提供了一个类似于我们前面提及的反向回溯更新的功劳分配机制。

- 1961 年和 1963 年，Donal Michie 描述了一个叫 MENACE(Matchbox Educable Naughts and Crosses Engine) 的简单试错学习系统，用来学习如何玩前述的井字棋游戏。该系统由对应于每个井字棋位置的火柴盒构成，每个火柴盒内含有许多彩色珠子，每一种不同颜色代表一种可能的移动方式。通过从当前游戏位置的。才和随机拿一个桌子就可以确定 MENACE 的移动当游戏结束时，我们会向曾经使用过的盒子里增加竹子或减少竹子一起来强化或惩罚 MENACE 的决策。

- 1968 年，Michie 和 Chambers 描述了另一种叫 GLEE(Game Learning Expectimaxing Engine) 的井字棋强化学习机和一个叫 BOXES 的强化学习控制器。他们采用 BOXES 使得一根杆子可以在一个可移动的小车上保持平衡，这一系统就是在失败信号的基础上工作的——当杆子倒下或车到达终点时，会有失败信号发出从而帮助系统学习。

- 1973 年，Widrow、Gupta 和 Maitra 修改了最小均方误差 (LMS, Least-Mean-Square) 算法，以建立一种强化学习规则，其可以从成功和失败信号中而不是从训练例子中学习。他们称这种学习形式为“选择性引导适应”，并将其描述为“向评论家学习”，而不是“向老师学习”。他们分析了这条规则，并展示了如何学会玩二十一点纸牌游戏。

对于自动学习机的研究对试错学习发展到现代强化学习有着直接的影响。这类方法用于解决非关联的、纯选择性的学习问题，又被称为 k 臂赌博机算法，即有 k 个控制杆的“单臂赌博机”算法。自动学习机是一种能够在这类问题中提高获得收益的概率的简单且无需大内存的机器。它源于 20 世纪 60 年代俄罗斯数学家、物理学家 M.L.Isetlin 以及他的同事们的工作。随机自动学习机是一种基于收益信号来更新动作概率的方法。相关的研究始于 William Estes 在 1950 年关于统计学习理论的研究，并被其他研究者推广，其中最著名的是心理学家 Robert Bush 和统计学家 Frederick Mosteller。

经济学领域对于强化学习的研究热潮，始于 1973 年 Bush 和 Mosteller 的学习理论在一系列经典经济模型中的应用。这项研究的目的在于探索比起传统的理想经济主体，行为更像真人的人工智能体。该项研究又扩展到对博弈论语境中的强化学习的研究。强化学习和博弈论的结合是一个和应用于井字棋、跳棋和其他娱乐游戏的强化学习有很大不同的主题。

John Holland 基于选择原理提出了一个自适应系统的一般理论。他的早期工作主要关注试错方法的非关联形式，主要涉及进化方法和 k 臂赌博机。他在 1976 年提出并在 1986 年完善了分类器系统，包含关联和价值函数的真正的强化学习系统。Holland 的分类器系统的一个关键部分是用于功劳分配的“救火队算法”，它与我们在井字棋的案例和时序差分算法有很深的关联。另一个关键部分是遗传算法，一种用来演化出有效表示方式的进化算法。

在人工智能领域的强化学习中的试错方法的复兴中，最关键的人是 Harry Klopf。Klopf 意识到当研究者们仅仅关注有监督学习时，他们丢失了适应性行为的关键部分。根据 Klopf 的说法，丢失的是行为享乐的特点，即从环境中获得成就感控制环境使其趋向于理想的结局而远离不理想的结局。这是试错学习不可缺少的思想。Klopf 的想法对于 Sutton 等人的影响尤为深刻，Sutton 等人因为研究其思想，才重视有监督学习和强化学习的区别。

1.7.3 第三条主线：时序差分

时序差分是由时序上连续地对同一个量的估计驱动的，例如下赢井字棋的概率。这条主线比起其他两条更微小、更不显著，但是却对强化学习领域有很重要的影响。时序差分学习的概念部分源于动物学习心理学，特别是次级强化物的概念。次级强化物指的是一种与初级强化物 (例如食物或疼痛等) 配对并产生相似的强化属性的刺激物。

- 1954 年，Minsky 可能是第一个认识到该心理学的规律对人工智能学习系统很重要的人。
- 1959 年，Arthur Samuel 首次提出并实现了一个包含时序差分思想的学习算法，这个算法是他著名的跳棋程序的一部分。
- 1972 年，Klopf 将试错学习与时序差分学习的一个重要部分相结合。Klopf 的研究兴趣在于能够推广到大规模系统中的学习方法，因此他受局部强化的思想所启发，即一个学习系统的各部分可以相互强化。他发展了“广义强化”的概念，即每一个组件 (字面上指每一个神经元) 将其所有的输入视为强化项：将兴奋的输入视为奖励项，将抑制的输入视为惩罚项。
- 1978 年，Sutton 进一步探索了 Klopf 的想法，尤其是和动物学习理论的联系。他将由变化导致的学习规则用短期的连续预测表达。他和 Barto 优化了这些想法并基于时序差分学习建立了一个经典条件反射的心理学模型。
- 1981 年，Sutton 等人提出了一种方法用来在试错学习中使用时序差分学习，即“行动器-评判器”(actor-critic) 架构，并将这种方法应用于 Michie 和 Chambers 的平衡杆问题。Sutton(1984) 在他的博士论文中详细地研究了这个方法，并在 Anderson(1986) 的博士论文中进一步引入了反向传播的神经网络。大约在同一时间，Holland(1986) 将时序差分的思想通过他的救火队算法应用到他的分类器系统。
- 1988 年，Sutton 将时序差分学习从控制中分离出来，将其视作一个一般的预测方法。那篇论文同时介绍了 $TD(\lambda)$ 算法并证明了它的一些收敛性质。
- 1989 年，ChrisWatkins 提出的 Q 学习将时序差分学习和最优控制完全结合在了一起。这项工作拓展并整合了强化学习研究的全部三条主线的早期工作。PaulWerbos 自 1977 年以来证明了试错学习和动态规划的收敛性，也对这项整合做出了贡献。
- 1992 年，Gerry Tesauro 的西洋双陆棋程序 TD-Gammon 取得巨大成功，使这个领域受到了更多的关注。