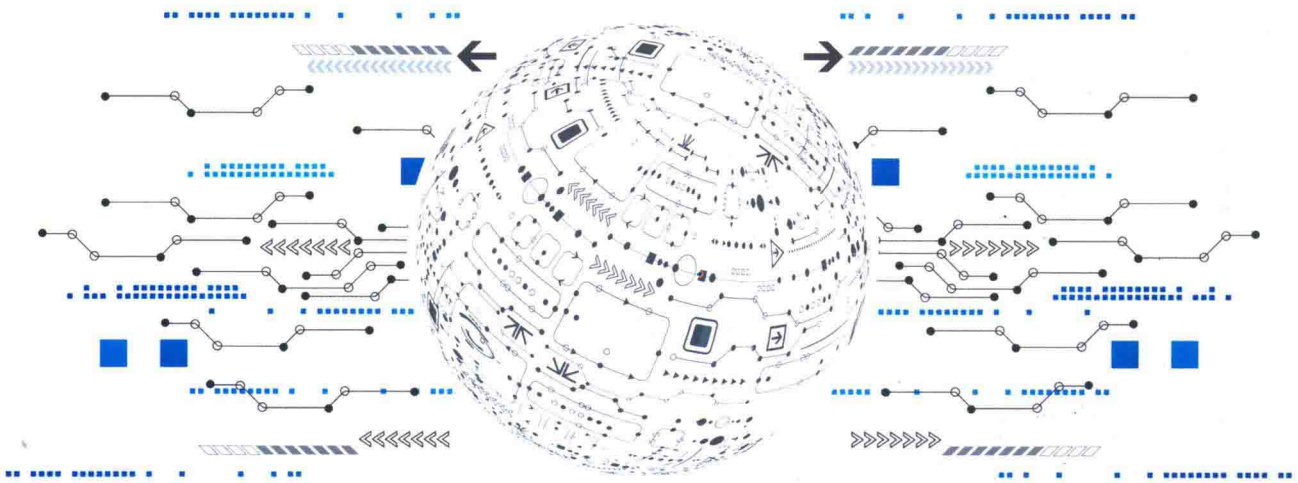
強化学習と深層学習 C言語によるシミュレーション

强化学习与深度学习

通过C语言模拟

[日] 小高 知宏 著　　张小猛 译



机械工业出版社
CHINA MACHINE PRESS

强化学习与深度学习：通过C语言模拟

[日] 小高　知宏　著
张小猛　译

机械工业出版社

強化学習と深層学習 C 言語によるシミュレーション，Ohmsha，1st edition，by 小高 知宏，ISBN：978－4－274－22114－9

Original Japanese Language edition KYOKA GAKUSHU TO SHINSO GAKUSHU CGENGO NI YORU SIMULATION by Tomohiro Odaka

Published by Ohmsha，Ltd.

北京市版权局著作权合同登记　图字 01－2018－5159 号。

图书在版编目（CIP）数据

强化学习与深度学习：通过 C 语言模拟/（日）小高　知宏著；张小猛译．—北京：机械工业出版社，2019.6

ISBN 978-7-111-62718-0

Ⅰ．①强…　Ⅱ．①小…　②张…　Ⅲ．①机器学习－研究　Ⅳ．①TP181

中国版本图书馆 CIP 数据核字（2019）第 090039 号

机械工业出版社（北京市百万庄大街 22 号　邮政编码 100037）

策划编辑：任　鑫　责任编辑：闾洪庆

责任校对：梁　静　封面设计：马精明

责任印制：李　昂

唐山三艺印刷有限公司印刷

2019 年 7 月第 1 版第 1 次印刷

184mm×240mm · 10.5 印张 · 231 千字

标准书号：ISBN 978-7-111-62718-0

定价：59.00 元

电话服务　　网络服务

客服电话：010－88361066　　机　工　官　网：www.cmpbook.com

010－88379833　　机　工　官　博：weibo.com/cmp1952

010－68326294　　金　　书　　网：www.golden－book.com

封底无防伪标均为盗版　　机工教育服务网：www.cmpedu.com

本书以深度学习和强化学习作为切入点，通过原理解析、算法步骤说明、代码实现、代码运行调试，对强化学习、深度学习以及深度强化学习进行了介绍和说明。本书共4章。第1章介绍了人工智能、机器学习、深度学习、强化学习的基本概念。第2章以Q学习为例，重点介绍了强化学习的原理、算法步骤、代码实现、代码运行调试。第3章先对深度学习的几种常见类型和原理进行介绍，然后给出了例程和调试方法。第4章以Q学习中运用神经网络为例，介绍了深度强化学习的基本原理和方法，同时也给出了例程和调试方法。

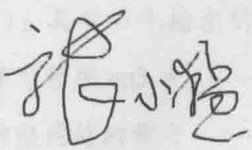
本书适合想要获得深度学习进阶知识、强化学习技术及其应用实践的学生、从业者，特别是立志从事AI相关行业的人士阅读参考。

译 者 序

最近计算机领域非常流行被称为“ABCD”的四门技术，其中A代表AI，即人工智能；B代表Block Chain，即区块链；C代表Cloud Computing，即云计算；D代表Big Data，即大数据。作为“ABCD”之首的人工智能，甚至被认为是一种在未来可以拯救人类的技术。毫无疑问，人工智能经历了几代人的研究和发展，已经进入了一个发展的快车道。

由于个人兴趣和工作原因，很幸运，我一直在从事“ABCD”四个领域的研究，同时也经常听到周围的朋友说想学习人工智能方面的知识和技术。确实，随着时代的发展，人工智能越来越需要普及，然而，很多人虽然想学，却不懂如何学，甚至不敢学。为什么呢？因为很多人工智能的教材都太过于高深了——介绍很多高等数学知识、统计学知识之后，才以这些为基础去深入讲解，这让很多许久没有接触过高等数学的人望而却步。

那么，有没有一本书是不需要从头理解数学体系，即可入手学习人工智能，尤其是细分领域的深度强化学习呢？我阅读过很多类型的人工智能方面的书籍，本书是较为循序渐进而无需过多数学知识的回顾就可以学习和掌握的。本书作者行文的基本思路是，首先介绍原理，接着介绍算法，然后介绍算法如何与具体编程语言进行匹配，最后通过整合后的完整代码进行运行和调试，为读者提供一个全面学习的通道。具体来说，本书首先对强化学习和深度学习的基础知识进行介绍，然后在此基础上，再对深度强化学习的原理和机制进行具体说明。同时，本书不仅仅是在概念上的说明，而且对具体算法用C语言进行了编码和实现，通过实际运行代码的方式去深入理解每一步的具体处理方法。极力推荐有兴趣的朋友购买阅读！



原书前言

近年来，被称为“深度学习”的机器学习方法在诸多领域取得了成功。深度学习诞生之初，在图像处理领域中为图像识别率取得历史性突破做出了非常大的贡献。随后，随着深度学习的不断发展，深度学习不局限于应用在图像处理领域，在各种各样的机器学习应用领域都取得了非常显著的成果。

在深度学习的成功案例中，有一个基于强化学习的深度学习技术应用方向。强化学习是单纯从一系列行动的结果进行行动知识学习的方法。在强化学习中引入深度学习的方法，一般我们称为深度强化学习。关于深度强化学习成功案例的应用报道非常多，例如，通过运用深度强化学习，计算机能够在汽车转向盘操控方面获得超越人类的技能；通过运用深度强化学习，可以制造出能够打败围棋世界冠军的 AI 围棋棋手等。

本书首先对强化学习和深度学习的基础知识进行介绍，然后在此基础上，再对深度强化学习的原理和机制进行具体说明。同时，本书不仅仅是在概念上的说明，而是对具体算法用 C 语言进行了编码和实现，通过实际运行代码的方式去深入理解每一步的具体处理方法。

最后，本书能够顺利成书，离不开作者在福井大学的教育科研活动中取得的经验。在此向福井大学的各位教职工和学生表示衷心的感谢。另外，借成书之际，也特别对 Ohmsha 出版社的各位编辑表示由衷的感谢。最后，我也要感谢支持我写作的家人们。

小高 知宏

2017 年 9 月

目　录

第1章

强化学习和深度学习

在本章中，我们重点讨论在人工智能中机器学习和强化学习所处的地位如何以及它们与深度学习的关系是怎样的。人工智能拥有各种各样的具体领域，而机器学习是这些领域中的一个，同时，强化学习和深度学习又是机器学习中的一个分支。而深度强化学习，则是在强化学习中引入深度学习方法的新型的机器学习方法。

1.1 机器学习和强化学习

首先，我们梳理一下人工智能和机器学习以及强化学习的关系。如图1.1所示，我们对它们之间的关系进行了描绘。人工智能当中存在着各种各样的研究领域，而机器学习是人工智能中的一个领域，机器学习与其他相关技术一起构成了整个人工智能研究领域。同时，作为人工智能其中一个领域的机器学习，也包含了各种各样的方法。其中，强化学习和深度学习是机器学习中的子领域。

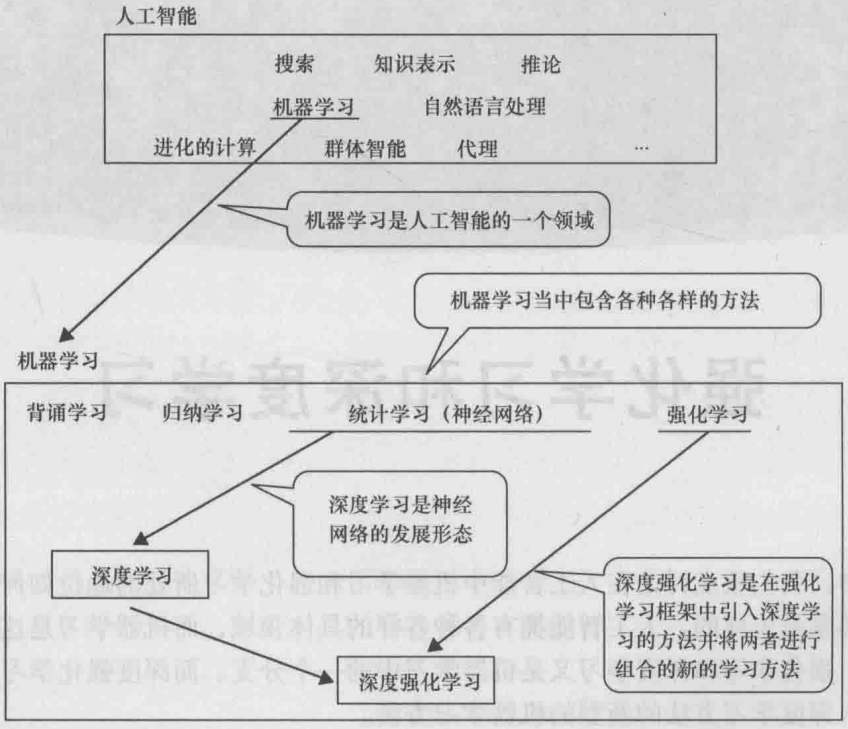


图1.1 人工智能、机器学习、强化学习以及深度强化学习的关系

近年来，在强化学习的基础上引入深度学习方法组合而成所谓的“深度强化学习”的方法被提上日程。深度强化学习是将普通的机器学习方法中的强化学习与深度学习进行结合，使得强化学习的学习能力大幅提升的学习方法。

接下来，我们来了解一下关于人工智能的研究是如何发展到深度强化学习领域的。

1.1.1 人工智能

人工智能（Artificial Intelligence，AI）是从生物和人类的智力活动中获取灵感，构筑出来能够制造出相关有用的智能软件的技术的学问。人工智能是以生物和人类的各种各样的智慧活动作为对象，开展和推进研究工作的。

例如，搜索（search）、知识表示（knowledge representation）、推论（inference reasoning）等智慧活动是人工智能研究早期的中心课题。在 20 世纪 50 年代以后，为了把这些研究成果用计算机程序进行实现，相关的计算机程序的实现方法也开始被人们大量地研究。

搜索的算法是，从大量的数据中找到目标数据的软件方法。开始研究搜索算法以来，从遍历每一种可能性的不遗漏搜索方法，到利用问题的性质来找出更多知识数据的方法，人们提出了各种各样的搜索方法。搜索的研究成果（如数据检索、汽车导航仪的目的地搜索、机器人的行动选择和游戏 AI 智能行动选择等）为面向大规模数据的高效处理软件的实现做出了很大的贡献（见图 1.2）。

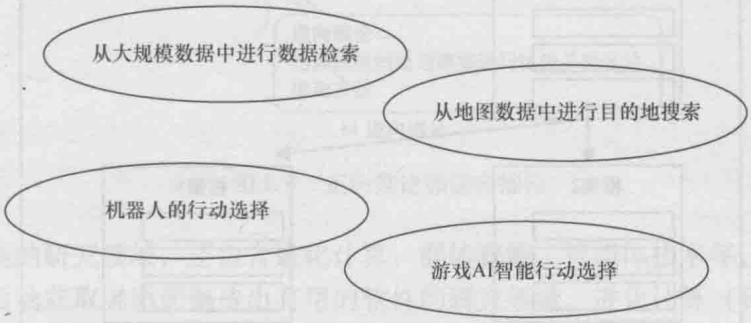


图 1.2　搜索——从大规模的数据中找出目的数据的软件方法

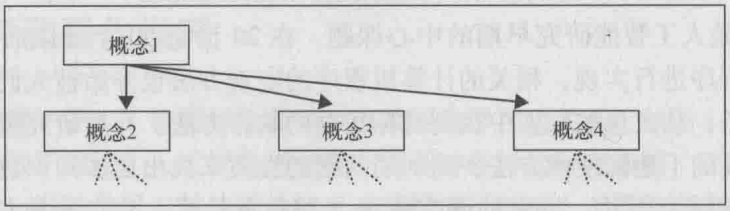
在人工智能领域中，将数据作为知识来利用的数据表现手法可以称为知识表示技术。知识表示与搜索和推论密切相关，针对各种各样的应用，人们提出了多种多样的知识表示方法。例如，以表达知识构成要素的概念之间的关系为目的的语义网络（semantic network）表示法[㊀]，框架（frame）表示法[㊁]，以知识为规则使得规则连锁处理简单化的生产系统（production system）等（见图 1.3）。

推论的算法是以已经存在的事实和知识为基础，生成出新知识的机制。推论有各种各样的形式。其中有以专家系统（expert system）[㊂]为例的正向推论。专家系统是推论系统的应用

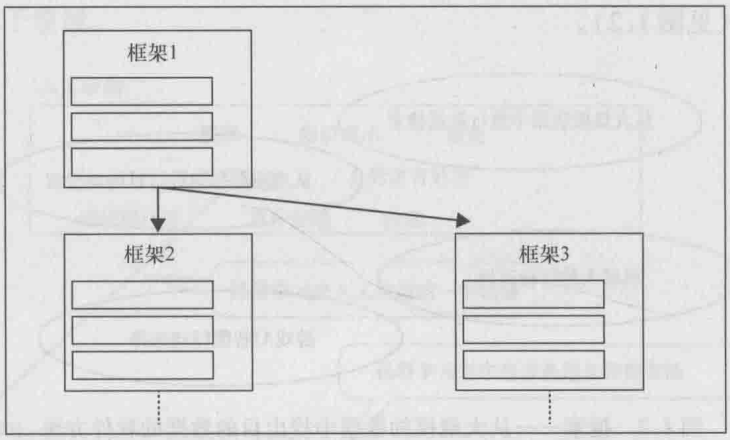
㊀ 语义网络表示法是一种以网络格式表达人类知识构造的形式，是人工智能程序运用的表示方式之一。由奎林（J. R. Quillian）于 1968 年提出。开始是作为人类联想记忆的一个明显公理模型提出，随后在 AI 中用于自然语言理解，表示命题信息。在专家系统中语义网络由 PROSPEUTOR 实现，用于描述物体概念与状态及其关系。它是由节点和节点之间的弧组成，节点表示概念（事件、事物），弧表示它们之间的关系。在数学上，语义网络是一个有向图，与逻辑表示法对应。——译者注

㊁ 框架表示法是一种适应性强、概括性高、结构化良好、推理方式灵活，又能把陈述性知识与过程性知识相结合的知识表示方法。——译者注

㊂ 专家系统是一个智能计算机程序系统，其内部含有大量的某个领域专家水平的知识与经验，能够利用人类专家的知识和解决问题的方法来处理该领域问题。也就是说，专家系统是一个具有大量的专业知识与经验的程序系统，它应用人工智能技术和计算机技术，根据某领域一个或多个专家提供的知识和经验，进行推理和判断，模拟人类专家的决策过程，以便解决那些需要人类专家处理的复杂问题，简而言之，专家系统是一种模拟人类专家解决领域问题的计算机程序系统。——译者注



a)语义网络：概念之间的关系用网络的方式进行表示的知识表示方法



b)框架：在语义网络的基础上进行扩展，包含概念内部的结构的知识表示方法

```
IF   A      THEN     X
IF   B      THEN     Y
  ......
```

c)生产系统：用IF THEN的形式表示规则的知识表示方法

图1.3 知识表示的例子

实例之一。专家系统的推论方式是从已有的事实导出结论的正向推论方式。与之相对的还有逆向推论，逆向推论首先证明结论的正确性，然后由结论出发，逐级验证该结论的正确性，直至已知条件（见图1.4）。

作为正向推论的例子，例如医疗领域的医疗诊断专家系统，在医疗诊断中，通过对检查结果和既有的事实进行综合，运用知识推导出结论。与之相对应的，作为逆向推论的例子，例如定理证明专家系统，在定理的证明中，预先给予结论的证明对象，通过知识进行推论，推导出被认定为是事实的公理或已经被证明的定理。

搜索、知识表示、推论等技术同时也是机器学习（Machine Learning）和自然语言处理（Natural Language Processing）等应用技术得以实现的基础技术。机器学习是本书的主题，是以计算机程序获得知识为目的的人工智能技术。自然语言处理是用计算机程序处理英语、中文、日语等自然语言的技术。

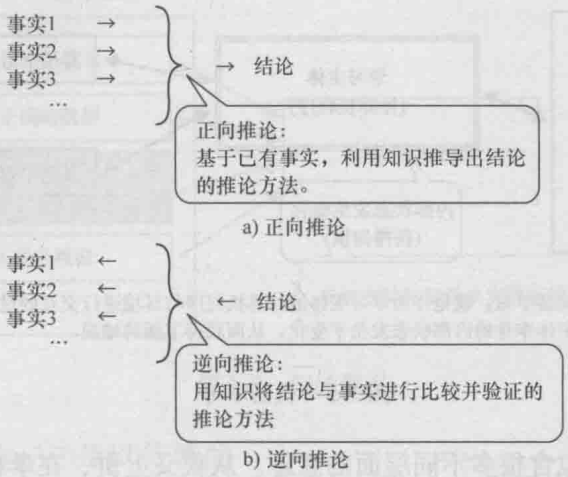


图 1.4　正向推论和逆向推论

在人工智能的研究领域，还包含进化计算、群体智能、智能体技术等，它们都是从生物和人类的智力活动获取灵感而制造出有用的软件的研究领域。进化计算（Evolutional Computing）[⊖]是受生物进化过程中“优胜劣汰”的自然选择机制和遗传信息的传递规律的影响，通过程序迭代模拟这一过程，把要解决的问题看作环境，把问题所有可能的解决方案组成一个“解决方案种群”，通过自然演化寻求“解决方案种群”的最优解。另外，群体智能（Swarm Intelligence）[⊖]，则是通过对鱼、鸟等生物群体所表现出来的智力行为进行研究和模拟，从而制造出智能软件。智能体技术（Agent Technology）是通过对生物和环境的交互方式进行建模，来创造出与环境相互作用的知识智能体的技术。

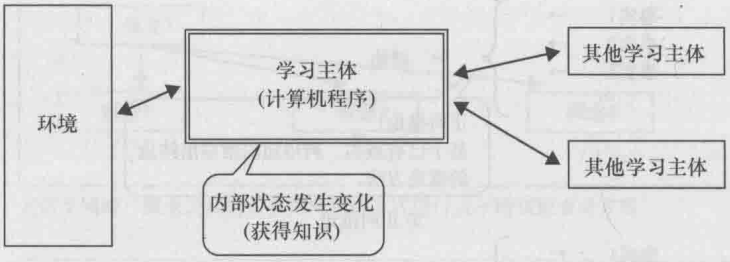
综上所述，在人工智能的研究领域中，通过模拟生物和人类的智力活动来制造智能软件的方法是多种多样的。而近年来，在这些方法中，机器学习技术是受到特别关注的。下一节，我们来探讨机器学习的基本概念。

1.1.2　机器学习

机器学习是用机器即计算机程序进行学习来获得知识的技术。这里的学习，与生物或人类的学习一样，是学习主体计算机程序和外部环境进行交互的过程中，学习主体的内部状态变化，获得新的知识的过程，如图 1.5 所示。

⊖ 在计算机科学领域，进化计算是人工智能中的一个领域，进一步说是智能计算中涉及组合优化问题的一个子域。其算法是受生物进化过程中“优胜劣汰”的自然选择机制和遗传信息的传递规律的影响，通过程序迭代模拟这一过程，把要解决的问题看作环境，在一些可能的解组成的种群中，通过自然演化寻求最优解。——译者注

⊖ 群体智能源于对以蚂蚁、蜜蜂等为代表的社会性昆虫群体行为的研究。最早被用在细胞机器人系统的描述中。它的控制是分布式的，不存在中心控制。群体具有自组织性。——译者注



所谓机器学习，就是作为学习主体的计算机程序与环境进行交互的过程中，学习主体本身的内部状态发生了变化，从而获得了新的知识。

图1.5 机器学习

生物或人类的学习包含很多不同层面的意思。从狭义上讲，在学校学习和在家里看书自习都是典型的学习的例子。如果把学习这个词的意思再稍微拓宽一点，学习不仅仅是获取知识，获取体育技能和获取才艺也可以说是学习。再者，从每天的生活和经验中学习，使得人们能够更好地适应环境的过程也是学习。

同样地，机器学习也由各种各样的技术构成。如图1.6所示，在背诵学习（Rote Learning）中，比如人类为了记住年号和外语单词，把要学习的知识直接背诵并积累。背诵学习是一种简单的学习方法，比如，在中文输入法中的汉语拼音与汉字之间转换的转换可选项的学习上非常有实际应用价值。

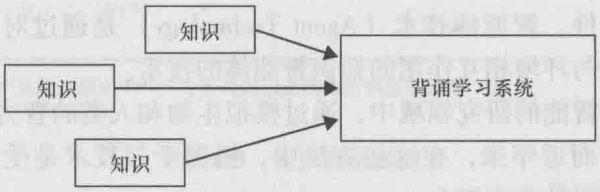


图1.6 背诵学习

归纳学习（Inductive Learning），是从个别例子中给予的大量数据中推导出新的知识的学习方法的总称（见图1.7）。一般来说，被给予的数据有正确的数据，也有不正确的数据。其中不正确的数据，我们通常形象地称为噪声（noise）。在“噪声”比较多的情况下，仅仅靠背诵学习并掌握事实数据是无法达到学习目的的。因此，以事实数据为基础，人们进一步提出了能够很好地说明事实数据知识的各种归纳学习方法。近年备受瞩目的、从大量的数据中找出规律性的大数据分析（Big Data Analysis）技术就是属于归纳学习的一种分析方法。

图1.1中所示的统计学习和强化学习是从具体方法的维度对机器学习进行分类的。统计

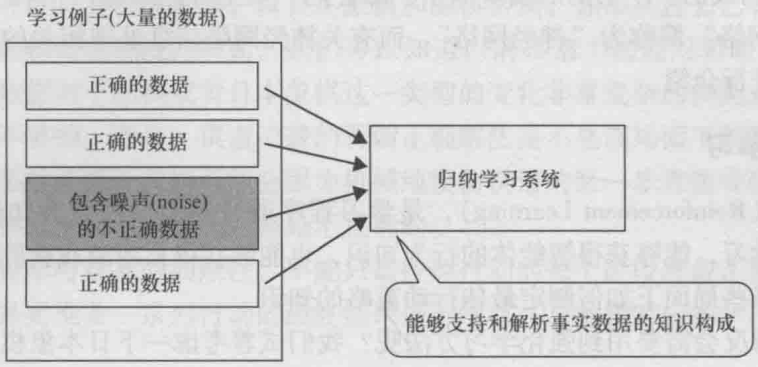


图 1.7　归纳学习

学习（Statistical Learning）[⊖]是以传统的统计学为基础的学习方法。在统计学习中，除了以前经典的统计学的方法之外，也包含了被称为人工神经网络（Artificial Neural Network，ANN）[⊜]的方法。人工神经网络是从生物的神经网络中受到启发，而被发明出来的统计学习方法（见图 1.8）。

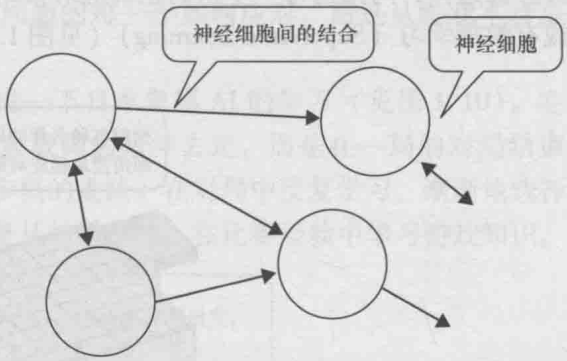


图 1.8　神经网络

生物的神经电路是由大量的神经细胞（neuron）相互连接而成的。机器学习领域的神经网络也一样，是通过模仿神经细胞，将人工神经细胞（artificial neu-

⊖ 统计学习理论是一种研究训练样本有限情况下的机器学习规律的学科。它可以看作是基于数据的机器学习问题的一个特例，即是在有限样本情况下的特例。统计学习理论从一些观测（训练）样本出发，试图得到一些目前不能通过原理进行分析得到的规律，并利用这些规律来分析客观对象，从而可以利用规律来对未来的数据进行较为准确的预测。例如，对某国未来几年人口数量进行预测，就需要先采集过去几年甚至几十年的人口数据，并对其变化规律做出统计学方面的分析和归纳，从而得到一个总体的预测模型，这样就可以对未来几年的人口总体走势做一个大概的估计和预测。——译者注

⊜ 人工神经网络是 20 世纪 80 年代以来人工智能领域兴起的研究热点。它从信息处理角度对人脑神经元网络进行抽象，建立某种简单模型，按不同的连接方式组成不同的网络。在工程与学术界也常直接简称为神经网络或类神经网络。神经网络是一种运算模型，由大量的节点（或称神经元）之间相互连接构成。每个节点代表一种特定的输出函数，称为激励函数（activation function）。每两个节点间的连接都代表一个对于通过该连接信号的加权值，称为权重，这相当于人工神经网络的记忆。网络的输出则根据网络的连接方式、权重值和激励函数的不同而不同。网络自身通常都是对自然界某种算法或者函数的逼近，也可能是对一种逻辑策略的表达。——译者注

ron）相互连接构成的。本书的后续章节中，都会把“人工神经细胞”简称为“神经细胞”，将“人工神经网络”简称为“神经网络”。而有关神经网络信息处理相关的具体内容，我们将在1.2节中进行介绍。

1.1.3 强化学习

强化学习（Reinforcement Learning），是学习程序通过经验学习行为知识的机器学习方法。通过强化学习，能够获得智能体的行为知识，也能够获得桌面游戏获胜策略的知识，甚至能够获得在某些局面下如何制定最佳行动策略的知识。

那么什么情况会需要用到强化学习方法呢？我们试着考虑一下日本象棋和围棋之类的游戏。为了取得棋类游戏的胜利，我们需要知道在各种各样的情况下，下一步应对的最佳策略。在这其中，获得每一步最佳策略的一种方法就是，请一个有经验的教师一步一步手把手地教。像这样，向知道每一步棋的正确应对策略的教师请教正确应对的学习方法叫作监督学习或有教师学习（Supervised Learning）（见图1.9）。

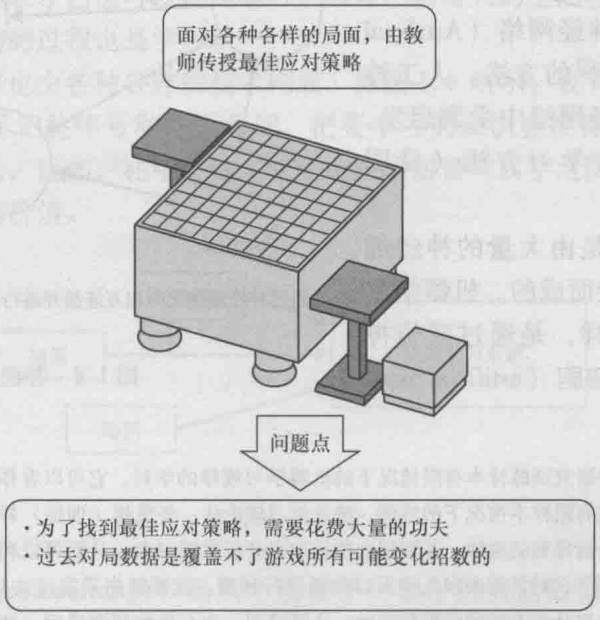


图1.9 监督学习（获取游戏知识的举例）

通过监督学习获得最佳应对策略的学习方法效率非常不错，但是，遗憾的是这个学习方法也是有局限性的。这个局限性在于，如果某个游戏局面的可能性非常多，获得最佳策略所需数据集的准备就显得非常的困难。

从游戏知识获得的例子来说，由于游戏中的局面的变化非常多，为每一个局面都准备一

个最佳应对策略并不是非常容易的。以日本象棋和围棋为例，如果从过去已有对局的棋谱数据来看，只要以前棋谱出现过的局面，我们可以知道以前棋谱上的应对策略是什么。然而，过去的这些棋谱数据对于围棋或者日本象棋这一类型的变化非常复杂的棋类来说，可供学习的数据量是远远不够的。而且，棋谱记载的所谓正确解法是不是该局面下的最佳策略也不得而知。甚至，在某些情况下我们可能会因为机械地模仿棋谱的某一步着法导致棋局失利，在这种情况下，在棋谱上记录的着法明显是不正确的。

为了摆脱监督学习自身的局限性，不能只是按照行动的每个阶段准备正确和不正确的解法进行学习，而是要准备一系列行动的最终结果进行学习。与此相对应的学习方法就是强化学习。

强化学习，就是从一系列的行动结果来开展的学习，是一种基于经验的学习方法。在我们上面讨论的下棋的例子中，学习下棋的方式稍微做一些改变，强化学习与监督学习不同，它不是像“监督学习”那样让教师传授各个局面的每一步如何应对，而是从游戏本身最终胜负的结果出发，来展开游戏知识的学习。

举个例子，我们以强化学习的视角来考虑一下日本象棋 AI 的学习（见图 1.10）。在这样的情况下，不是每个局面如何应对都紧随着教师的指导去走，而是在一局的对局结束时刻，从输赢的结果出发来评价对局中间每一步棋的走法。在对局中反复学习，渐渐地选择出每一步比较好的走法。换言之，强化学习就是从结果出发，在比赛经验中学习游戏知识。

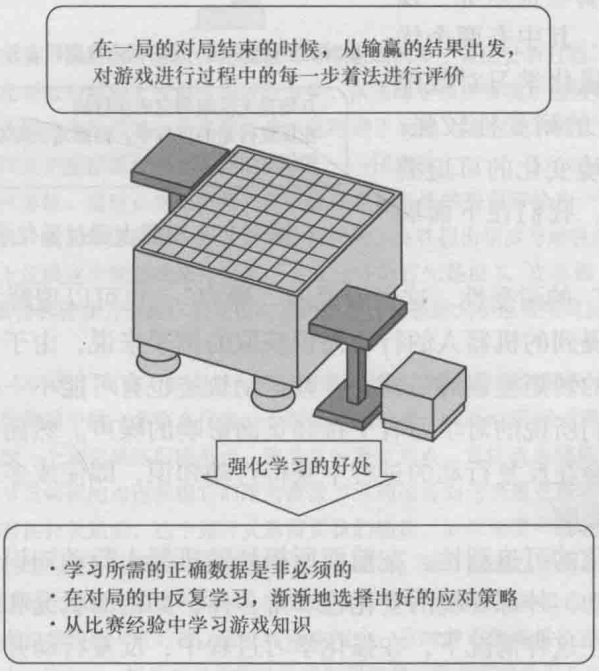


图 1.10　通过强化学习获得游戏知识