强化学习概述

孙振坤
（苏州信息职业技术学院，江苏 苏州 215200）

摘要：近年来人工智能发展迅速，强化学习也受到了更多的关注，其应用研究已经遍步各行各业中。本文对强化学习进行了简要概述，介绍强化学习的基本理论，并讨论了相关的几种应用。

关键词：人工智能；强化学习

一、强化学习基本原理

强化学习（Reinforcement Learning，简称R L）作为机器学习领域另一个研究热点，已经广泛应用于工业制造、仿真模拟、机器人控制、优化与调度、游戏博弈等领域。R L的基本思想是通过智能体在某个环境中行动时，不断的采取动作与环境进行交互，使得智能体对这个未知环境的越来越适应，并且能从环境中不停的获得奖赏并逐渐实现奖赏最大化，从而找出了达到目标的最优策略。因此，R L方法主要就是通过交互及反馈的奖赏最终逐步找到解决问题的策略的过程。

RL学习系统一般包含智能体agent以及环境两个部分。智能体agent在环境中执行某种动作，从而与所处的环境进行交互，在执行动作后智能体将变成一个新的状态，同时还会得到环境在该动作下的奖励，直到智能体最终到达一个目标，此时不再选择动作，同时状态也不再发生改变。要想智能体最终能学习到一个解决问题的最优策略，只要使该过程反复进行，当奖赏最大化，策略也接近于最优化。

强化学习的决策过程是基于马尔可夫决策过程，通常假设强化学习任务满足马尔可夫性，一个马尔可夫决策过程首先是定义为一个四元组（S，A，P，R），S为agent所能达到的所有状态空间的集合；A表示agent所能选择的所有动作空间的集合，我们把每一个可能的动作记为a，从而可得$a \in A$；P为状态转移概率，也就是agent在一个状态s下，采取某个动作a后达到下一个状态s’的概率，记为P（s’|s，a）；其中s’为状态s的下一步状态；R为奖赏函数，是agent在状态s下采取动作a所达到下一个状态s’时所获得的立即奖赏，记为r（s，a）；通常情况下，该四元组中的P和R常常是未知的，需要agent不断在环境中进行探索，进行不断地试错。可以用图1来表示强化学习的过程。

图1　强化学习

比如，一个机器人学习捡垃圾。这个机器人就是一个agent，它的动作可能是走路可能是捡起东西，环境就是他活动的范围区域。它是直接走路还是捡起了垃圾，这种动作选择后，就是从一个状态转变到了另一个状态。当他学习到看到垃圾就要走过去捡起来，就会有相应的奖赏，就会持续的捡，直到把活动区域的垃圾全部捡完，就完成了强化学习。

二、强化学习和监督式学习的异同

强化学习和监督式学习的区别在于过程中是否有监督。监督式学习就好像在学习走路的小孩，旁边有家长在指点他，需要家长指点是因为只有家长才知道他怎样做是对的怎样做是错误的。而强化学习是没有人在旁边指导的，完全是通过自己的动作来得到结果，再通过是否有奖赏自己来调整自己的下一步动作，通过不停的调整，最终学会了在什么样的情况下选择什么样的动作，能得到最大化的奖赏。

两种学习方式的同共点都是在学习，学习输入和输出之间的映射关系，强化学习输出一个奖赏，可以用于判断刚才的动作是否是一个好的选择，但这个选择从长期来看未必是最好的，需要继续往下走很多步才能得到真正正确的反馈，而监督式学习输出的是关系，明强的指出输入和输出的对应关系。

三、强化学习的进一步研究模型

笔者在不断研究强化学习在生活工作中的应用情况，本节介绍两个可能可以应用强化学习实现智能化的实例。

（一）餐馆自动炒菜机

现在虽然已经有自动炒菜机，但调料的添加量还需要人工控制，口味完全还是手动。如果餐馆有智能炒菜机，能依据当地人的口味自己学习调整调料的用量，逐渐能做出适应当地人口味的菜肴，不但能大大减少厨师的工作量，还更智能化。假设动作action是调料的用量，可以用3克或1克为变化量，可做的选择如多放1克或少放一克等，中等量为厨师平时的用量。所给予的奖赏应该是该菜的点菜率或者销售量，只要卖掉一份rewards>0，否则reward=0。而当前状态就是该菜品当前已经卖掉多少份。从一开始调料由中等量随机选择不同的动作，进行口味的微调，经过长期训练，逐渐找到各调料最佳配方，也就是卖的最好的一种配方。

（二）足球守门员

足球运动员训练时，可使用机器人做守门员以训练。在这种环境下的agent就是机器人或者说是机器守门员，它的动作action就是他可以朝各个方向跳起的角度，rewards是在球踢来之后如果能弹回去就可以直接加一，如果没能挡住则直接减和，状态states为当前已经挡住球的个数。该模型主要是在不断学习迭代的过程中，根据球飞来的各种角度，学会改变自己的跳起角度，找到最合适的时机起跳，实现挡住球的数量的最大化。

四、结语

强化学习是一种无监督导师的在线学习方法，它一般通过迭代来减小后继状态估计之间的差异来完成迭代过程。强化学习是一个高深的学科，需要学习强化学习基础、马尔可夫决策、动态编程、蒙特卡洛搜索和时序差分学习以及深度学习基础编程，这也是今后强化学习研究的重点。

参考文献：

[1] 刘全，翟建伟等．深度强化学习综述．计算机学报，2018 41（1）：1-27.

[2] 徐松林．深度强化学习概述．电脑知识与技术，2019 15（3）：193-194.

基金项目：2018年江苏省高职院校教师专业带头人高端研修（编号：2018GRFX058）