加强学习 机器学习 人工智能 ③

计算机科学1998 Vol. 25 №. 3

加 强 学 习

A Summary on Reinforcement Learning

13-15 TP18

郭茂祖 陈 彬 王晓龙 洪家荣

(哈尔滨工业大学计算机系 哈尔滨150006)

摘 要 The model of reinforcement learning is presented in this paper. Then its development and studying status is discussed. And at last the application of reinforcement learning is pointed out.

关键词 Reinforcement learning, Markov decision process (MDP), Temporal difference (TD), Q-learning

加强学习(Reinforcement Learning,简称RL)是一种重要的机器学习方法,在机器人规划、分析预测等领域有许多应用。加强学习的任务即是寻找一条策略,为行为者(agent)在一给定状况下产生一个动作的规则。但在传统的机器学习分类中没有提到过加强学习,国内有关人工智能或机器学习的著作也很少提及。然而加强学习近几年发展很快,在著名期刊《Machine Learning》上已发表两次专刊。由Richard Sutton在1992年编辑的第一个专刊标志着加强学习发展成为机器学习领域的一个主要组成部分[1]。1996年由Leslie Pack Kaelbling编辑的第二个专刊占用了连续三期《Machine Learning》[2]。另外,在每年的国际会议《International Conference on Machine Learning》上,有关研究加强学习的论文也占较大比例,并吸引了许多新的研究者。

本文首先介绍加强学习的一般模型,然后讨论它的发展及目前研究状况,最后指出它的应用。

1 加强学习模型

在加强学习模型中,行为者与环境(外部世界)相互作用,并且同时接收到强化信号,该信号是由一决策与/或状态转换引起的惩罚或耗费。学习的任务是寻找一条合适的策略,即告知行为者在某一给定情况下选择哪种动作的一条规则。行为者与外部世界之间相互作用的最常用模型是一称作马尔可夫决策过程(MDP)的特殊随机过程。

在MDP中存在三个基本集合:在行为者看来的环境状态集合S,行为者的动作集合A以及强化信号的集合C。强化信号在下面也被称作立即耗费。本文限定讨论有限集S和A,并表示为自然数集N的子集。C⊂R的元素假定为有限的和非负的。一般地,行为者不必或不能在每一状态随意选择一个动作。非空集合A(i)⊆A表示在状态i容许选择的动作集合。

行为者与外部世界的相互作用划分为称作阶段或事件(Episodes)的序列。下面给出一个事件的模型:首先,行为者观察到一开始状态$i\in S$,并要选择和执行一个动作$a\in A(i)$。然后,出现一次状态转换:从状态i到后继状态$j\in S$。同时,行为者接收到一标量强化信号$r\in C$。立即耗费r表示在动作a下发生从i到j状态转换的效果。

MDP的本质在于:在任何时刻t,事件t的后继状态等于某一确定状态j的可能性仅仅依赖于该事件的开始状态i和动作a,但不另外依赖时间与过去事件。$P_t(i,a,j)$表示对某给定开始状态i和动作a,某事件的后继状态为j的概率。

类似地,在任何时刻t,事件t的立即耗费等于某一确定值$r\in C$的概率仅仅依赖于该事件的开始状态、动作和后继状态,但不另外依赖时间和过去事件。$P_t(i,a,j,r)$表示某事件在给定开始状态i、动作a和后继状态j下,该事件的立即耗费为r的概率。

郭茂祖 博士生,研究方向为机器学习、计算机图像处理。王晓龙 教授,博士生导师,研究方向为人工智能、自然语言理解、智能人机接口等。

在 MDP 中，对所有 $t>0$，t 的开始状态等于事件 t-1的后继状态，事件零的开始状态可由一概率分布给出。$P_0(i)$表示时标 $t=0$的事件开始状态为 i 的概率。假定对所有 $i\in S$，有 $P_0(i)>0$。

行为者的行为由一策略来确定，该策略为在任一给定状况下产生一动作的规则。在一般的策略模型中，行为者会随时考虑当前状态、时间以及过去的状态和动作来选择一个动作。而对于平稳策略，选择动作只要求行为者考虑当前状态。这种策略仅仅反映从状态到动作的映射，但证明是很有用的。

在 MDP 中，状态与立即耗费因状态转换和立即耗费的随机特性而为随机变量，这些随机变量根本上依赖于行为者采用的策略。

2 加强学习发展

加强学习是学习从状态到动作的映射以极大化一标量奖赏或强化信号。加强学习的两个突出特征是试错(trial-and-error)搜索和延时奖赏。具体地说，它不象绝大多数机器学习那样，告知学习者采取何种动作，而是必须通过尝试动作揭示哪些动作产生最高奖赏；另一方面，动作不仅能影响立即奖赏，而且影响下一状况以及其后的所有奖赏。

加强学习在人工智能(AI)中是一个既新颖又古老的话题。Minsky 首先于1961年使用该术语[2]，并且 Waltz 与 Fu 于1965年在控制论中也提到它[4]。现在看来在机器学习研究中最早与加强学习相联系的是 Samuel 于1959年的棋盘比赛程序[5]，它利用瞬时-差分学习(temporal-difference learning)来管理延时奖赏，这一点类似现在的研究。当然，在心理学方面研究学习与强化已长达近一个世纪之久，而且这种研究已对人工智能工作产生了很大影响。

尽管如此，在本世纪六十年代后期及七十年代，对加强学习的研究仍然被淡忘了。直到八十年代早期，它才逐渐成为机器学习领域的一个比较活跃和得到认可的领域[6,7]。由 John Holland 于1975年[8]和1986年[9]最早提出的有关遗传算法及分类器系统研究对加强学习具有影响作用；学习自动机理论[10]同样有促进作用。之后，Chris Watkins[11]与 Paul Werbos[12]等通过将最优控制理论和动态规划与之联系，也促进了加强学习的理论研究。

3 加强学习研究状况

目前人们对于加强学习研究最多的有以下两类算法或技术：瞬时-差分学习与 Q-学习。关于它们的算法在文献[13]中已有论述，下面重点讨论它们的研究现状。

3.1 瞬时-差分学习

解决加强学习问题最常用的技术是基于动态规划，这是运筹学研究范畴[14,15]的一个概念，该技术基于以下思想：某状态效用(utility)的估计可以通过前视(looking ahead)以及使用后继状态效用的估计，这也就是 Sutton 于1988年[16]和 Watkins 于1989年[11]提出的瞬时差分(TD)技术的基础。然而，也有其它方法，它们基于直接优化某一策略[17]或者优化一值函数[18]。这些技术对考虑加强学习是很必要的。

目前进行的加强学习研究有些是理论的，有些是实验方面的。它们绝大多数利用联结网络的某些形式作为其学习方法的一部分。Williams 引入加强学习的类似于有导师联结学习中出现的梯度理论[19]，Williams 理论处理立即奖赏情况，而 Tesauro 则集中于延时奖赏[20]。Tesauro 分别利用瞬时-差分方法和有导师学习方法来下国际象棋，结果非常理想，他的瞬时-差分程序对于学习下棋比以前的世界冠军程序和专业人类选手都好得多。Millan 和 Torra[21]利用连续而非离散的状态与动作空间，这是首次将连续动作与时间差分学习相结合用于路径发现。

学习暂时-延缓(temporally-delayed)奖赏的问题逐渐容易理解。对于瞬时-差分算法(TD)的收敛性，已有完整的证明(Dayan & Sejnowksi，1994，Tsitsiklis，1994)[22,23]。Tsitsiklis 与 Van Roy 的文章为基于特征的表示提供了收敛结果[24]；此结果也是将加强学习扩展到大规模问题的关键。绝大多数 TD 方法的收敛结果依赖于基于 Markov 环境的假设；Schapire 与 Warmuth 一文证明了即使对于随机的非 Markov 环境，只要略微改变一下标准 TD(λ)算法，则其运行效果几乎与值函数的最好线性估计一样好[25]。

几乎所有 TD 算法的正式结果均使用期望无限范围的优化折扣模型，即提出的策略在极小化期望总折扣耗费方面是最优的。但采用期望值作为决策准则并不总是可靠的，为此 Heger 分别于1994年[26]年和1996年[27]提出在极小极大情形下的动态规划及加强学习，这时智能体(agent)应选择优化最坏可能结果的动作。作者通过分析实例发现极小极大准则也并不总是可靠的，在此基础上提出一种决策准则

的选择方法。

3.2 Q-学习

与瞬时-差分学习联系密切的是 Watkins 于1989年提出的 Q-学习[11]，它是目前最容易理解和广为使用的加强学习方法。由 Watkins 与 Dayan 首次提出 Q-学习收敛的一个完整证明[28]，是加强学习理论的里程碑。Lin 与 Singh 进一步扩展并推广 Q-学习及其它简单加强学习方法以便使其适用于更大及更难的任务。通过对学习方法进行系统比较，Lin 表明利用"由例子告知"和"再利用先前经验"的新方法可极大地加速学习[29]。Singh 扩展加强学习方法是通过实现从简单任务到较大的组合任务的转换[30]。Dayan 利用 Q-学习技术来扩展时间-差分学习方法的理论，并降低它们对马尔可夫环境假设的依赖性[21]。

对于 TD 与 Q-学习算法，人们经常抱怨其奖赏在状态空间内传播较慢，对此可引入跟踪机制。Singh 与 Sutton[32]为 TD 考虑一个新的跟踪机制，并证明它比标准机制具有某些理论及实验方面的优点。由 Peng 与 Williams[33]作的技术报告研究了利用由 Watkins 在 Q-学习中提到的跟踪。

最后，像其它学习一样，加强学习的一个关键问题是关于发现及利用偏置(bias)，偏置在加强学习中特别关键，因为它具有双重作用：一是用于进行合适的扩展，二是能指导初始探索以便收集有用的经验。Maclin 与 Shavlik[34]，允许人们以"建议"的形式对他们的加强学习系统提供偏置，将这种建议增加到值函数的神经网络表示中，并根据智能体的经验对建议进行调整。

其它加强学习工作，还有 Barto 等的动态规划[35]；Whitehead 与 Ballard 的有效感知[36]；Mahadevan 与 Connell 有关机器人的 Q-学习[37]；Booker[38]与 Grefensteete 等[39]有关遗传算法中的加强学习。

加强学习诱人之处部分原因在于它在某种意义上是整个 AI 问题的一个缩影，其任务即是某个自主学习智能体与其外界交互作用以完成某个目标。

国内对加强学习的研究较少，而且研究集中于应用方面。阎平凡最近对加强学习的原理及主要算法进行了介绍[31]。叶文曾于1991年将 TD 算法用于化工控制研究中。杨璐博士将 TD 算法与神经网络结合于时序实时建模，成功地应用到股票市场预测之中[40]。本文作者的研究是集中于加强学习基础理论方面的。

4 应用

机器人(robot)控制问题，像航海，杆平衡或者魔术游戏是经典的加强学习问题；但加强学习问题也出现于其它许多应用中。加强学习特别有趣的一类应用出现于符号学习中(Dietterich)[41]。Tesauro 的 TD-Gammon 系统[42]是一类符号级加强学习的一个例子，系统在开始就知道国际象棋的全部模型，因而原则上能够简单地计算最优模型。然而，这种计算是难以实现的，只能利用模型来产生经验，然后从经验中学习策略。最后得到了非常好的近似解，所用经验集中于游戏的绝大多数典型情况。符号级加强学习的另一个例子是 Zhang 与 Dietterich 的调度系统[43]。这时，把问题求解中有关学习搜索-控制规则的问题模型化为一加强学习问题，这种模型比典型的基于解释学习模型更合适[44]。

在未知环境中探索的问题是加强学习的一个关键问题。尽管容易理解 k-机翼敌机这类简单问题，但理解更一般环境的探索较难。Koenig 与 Simmons[45]考虑了具有目标的多状态环境中探索的特例，该文证明了即使找到目标一次的问题也可能是难以处理的，但是在表示方面的简单改进就能对问题的复杂度产生较大影响。

(参考文献共45篇略)

(上接第29页)

参 考 文 献

[1]姚天顺等，自然语言理解，清华大学出版社，1995

[2]熊学亮，情境理论模型评介，国外语言学，1993.4

[3]迟成英、麻志毅，文本理解与汉语文本结构分析，中文信息，1997.1

[4]Ma Zhiyi, Zhan Xuegong, Yao Tianshun, The Method of Knowledge Expression of Confirming Texts' Topics, ICCPOL97

[5]Forgas, J. P., Social Episodes, New York: Academic Press, 1979

[6]Ma Zhiyi, Yao Tianshun, Extracting Topics from Texts Based on Situations, PACLIC11, 1996

[7]姚天顺等，词汇语义驱动方法，机器翻译研究进展，电子工业出版社