基于知识的Agent强化学习算法分析与研究

殷锋社
(陕西工业职业技术学院 陕西 咸阳 712000)

摘要：强化学习具有与环境交互的优势，笔者提出的基于知识的Q-学习算法(KBQL)就是利用Q-学习算法的这个特点，利用Agent的先验知识来缩小Agent学习的状态空间，以加速强化学习的收敛性，同时采用Agent的学习机制克服其知识的不精确性，从而提高学习算法的鲁棒性和适应性。

关键词：强化学习；KBQL；Agent；鲁棒性；适应性
中图分类号：TP3-01 文献标识码：A 文章编号：1674-6236(2011)11-0115-03

Analysis and research of Agent reinforcement learning algorithm based on knowledge

YIN Feng-she
(*Shanxi Polytechnic Institute*, *Xianyang* 712000, *China*)

Abstract: Reinforcement learning has the advantage of interacting with the environment, this paper presents a knowledge-based Q-learning algorithm (KBQL) Q is a learning algorithm using this feature, the use of Agent prior knowledge to narrow Agent learning state space, in order to accelerate the reinforcement learning Convergence, while using the learning mechanism Agent overcome inaccuracy of their knowledge, thereby enhancing the learning algorithm robustness and adaptability.

Key words: reinforcement learning; KBQL; Agnet; robustness; adaptability

学习是Agent适应复杂动态不确定环境的一项重要技能，在现有的各种学习算法中，强化学习是一种能与环境进行交互的、无需模型的在线学习算法，具有处理动态不确定性环境的优势，因而使其成为机器学习研究中的一个重要分支。

传统的强化学习算法研究没有考虑Agent的先验知识，尽管在形式上提供了一个统一的算法框架，但在实际应用中，这些没有启发知识的强化学习算法收敛速度都相当慢。另外，标准强化学习算法的收敛性是建立在可以任意遍历状态空间状态的前提下，但对于真实的物理环境(如机器人)，这种方式是不现实的。而且在实际应用中，Agent总可以获取各种形式的启发知识，因此将知识融入强化学习系统中，不仅可以改善强化学习算法的收敛性，而且还充分利用系统的资源(如专家知识等)。

1 强化学习

强化学习是学习如何把状态映射到动作使奖赏值达到最大的学习算法，Agent通过在环境中不断地感知和动作，来学习选择最优的动作以实现目标任务，强化学习坚实的理论基础和诱人的应用前景正逐渐受到各研究领域学者的广泛重视，不仅是研究智能学习的理论工具，同时又是实际应用的有效手段。下面对强化学习的基本原理及常用的基本算法进行介绍。

1.1 强化学习的基本原理[1]

强化学习系统的基本框图如图1所示，强化学习的基本原理是：如果Agent的某个动作导致环境正的奖赏（强化信号)，那么Agent以后产生这个动作的趋势便会加强；反之Agent产生这个动作的趋势减弱。

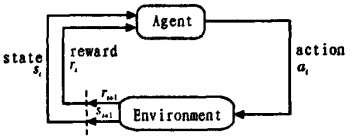

图1 强化学习基本框图

Fig. 1 Basic block diagram of the reinforcement learning

一般地，强化学习问题可以看成是一个Markov决策过程(Markov Deeision Proeesses, MDP)，其定义如下：

$$MDP=<S,A,R,P> \quad (1)$$

其中S是有限的离散状态空间，A是有限的离散动作空间；R是回报函数；p是状态转移函数，因此在已知状态转移概率函数P和回报函数R的环境模型知识下，可以采用动态规划技术求解最优策略。而强化学习着重研究在P函数和R函数未知的情况下，Agent如何获得最优策略[2]。

收稿日期：2011-03-18 稿件编号：201103101
作者简介：殷锋社(1976—)，男，陕西乾县人，硕士研究生，副教授。研究方向：个性化仿真系统研究。

1.2 强化学习的基本算法

目前,强化学习主要有两大类算法:一类是值函数估计法,这是强化学习领域研究最为广泛的方法;另一类是策略空间直接搜索法,如遗传算法、遗传编程、模拟退火方法以及一些其他改进方法。基于值函数估计的常用算法主要有[5]:TD 算法、Q-学习算法、Saras 算法等。

1.2.1 TD 算法

TD(temporal difference)学习是强化学习技术中最主要的学习技术之一,TD 学习是蒙特卡罗思想和动态规划思想的结合,即一方面 TD 算法在不需要系统模型情况下可以直接从 Agent 经验中学习;另一方面 TD 算法和动态规划一样,利用估计的值函数进行迭代[6]。

最简单的 TD 算法为一步 TD 算法,即 TD(0)算法,这是一种自适应的策略迭代算法。但 TD(0)算法存在收敛慢的问题,其原因在于,TD(0)中 Agent 获得的瞬时奖赏值只修改相邻状态的值函数估计值。更有效的方法是 Agent 获得的瞬时奖赏值可以向后回退任意步,称为 TD(幻算法。TD(幻算法的收敛速度有很大程度的提高,算法迭代公式可用下式表示:

$$V(s)\leftarrow V(s)+\alpha[r_{t+1}+\gamma V(s_{t+1})-V(s)]e(s)$$

其中,$e(s)$定义为状态的资格迹(eligibilitytraees)。实际应用中 $e(s)$可以通过以下方法计算:

$$e(s)=\begin{cases}\gamma\lambda e(s)+1 & s \text{ 是当前状态}\\ \gamma\lambda e(s) & \text{其他}\end{cases} \tag{2}$$

上式说明,如果一个状态被多次访问,表明其对当前奖赏值的贡献最大,然后其值函数通过式(3)迭代修改。

1.2.2 Q 学习算法

Q-学习算法由 watkins 提出的一种模型无关的强化学习算法,也被称为离策略 TD 学习(off policy TD)[5],被誉为强化学习算法发展中的一个重要里程碑。Q-学习是对状态-动作对的值函数进行估计的学习策略,最简单的 Q 学习算法是单步 Q 学习,其 Q 值的修正公式如下:

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha[r_{t+1}+\gamma\max_a Q(s_{t+1},a)-Q(s_t,a_t)]$$

上式中参数 α 称为学习率(或学习步长),r 为折扣率。

在单步 Q-学习算法中,需要学习的状态-动作对值函数 0 值是对最优状态-动作对的值函数的近似,并与所遵循的策略无关。这一特点使算法的分析得到极大地简化,而且使收敛性的证明得以实现。

1.2.3 Sarsa 算法[6]

Sarsa 算法是 Rummery 和 Niranjan 于 1994 年提出的一种基于模型算法,最初被称为改进的 Q-学习算法。它仍然采用的是 Q 值迭代,但 sarsa 是一种在策略 Tn 学习(on-policy Tn)。

2 基于知识的 Q-学习算法

在前面介绍常用的 3 个强化学习算法中,其中 Q-学习算法与其他两种算法存在一个很大的不同之处,就是 TD 算法和 Saras 算法是在策略(on policy)学习方法,而 Q-学习算法是离策略 (off policy)学习方法。在策略学习方法中,要学习的最优值函数依赖于学习过程中当前所采取的策略,学习过程中选择的策略质量的好坏直接影响 Agent 要学习的最优策略。在离策略学习方法中,要学习的最优策略与在学习过程中采取的策略无关,这就使 Agnet 对学习策略的选择具有更大的控制力。故本节将要提出的基于知识的Q-学习(Knowledge-Based-Learning,KBQL)模型选用Q-学习算法来研究。

下面首先对 Q-学习算法的收敛性进行分析,在此基础提出的 KBQL 算法并对 Agent 内部的学习机制进行详细介绍。

2.1 Q-学习算法的收敛性分析[7]

Watkins 给出了在 Markova 决策环境下,Q-学习算法的收敛性的完整证明,下面介绍收敛定理。

Q-学习的值函数的修改迭代公式,则有以下定理:

$$\sum_{n=1}^{\infty}\alpha(n)=\infty \text{ 并且 } \sum_{n=1}^{\infty}\alpha^2(n)<\infty \tag{3}$$

当迭代步数 n 趋于无穷大时,假定所有的状态-动作对被无限地经常访问,那么,对所有的状态-动作对(s,a)由 Q-学习算法产生的 Q-值序列$\{Qn(s,a)\}$以概率-收敛于最优值 $Q(s,a)$。

定理 1 收敛定理:假设迭代公式中学习率参数 α 满足条件。

根据收敛定理可知,在 Q-学习中有一个附加要求:所有潜在有用的动作都应被测试,即要对所有允许的状态-动作对都应该经常探测(explore)足够的次数以满足收敛定理。因此,在给定动作集的情况下,状态集的大小是影响 Q-学习收敛速度的最重要因素。

2.2 KBQL 算法

标准 Q-学习算法的收敛速度依赖于要学习的状态空间大小,另外还与学习过程的动作选择机制有关。因此引入 Agent 的知识库来缩小要学习的状态空间,并将动作选择机制从强化学习模块中分离出来,提供一个更灵活的决策机制,利用 Agent 的知识来指导探测状态空间。它主要包括 3 个模块[8],即 RL 模块、决策控制模块和知识库 KB。

1)RL 模块 RL 模块里采用离策略的 Q-学习算法,但它与标准的 Q-学习算法模块有所不同,在标准 Q-学习算法中,其动作选择机制包含在算法模块中,而 RL 模块只实现状态-动作对的值函数 $Q(s,a)$的计算,即在每一个离散时刻 t,由决策控制模块确定一个动作 a_t,得到经验知识和训练样例$\{s_t,a_t,s_{t+1},r_{t+1}\}$,RL 模块根据此经验知识更新 Q 值。

2) 决策控制模块 决策控制模块主要实现 Q-学习算法的动作选择机制。利用 Agent 的知识来指导 Agent 的探测行为。当遇到 Agent 的知识无法指导的情况下,就启用标准 Q-学习算法中常用的 Soft-Max 动作选择策略。

3)知识库 KB Agent 的知识库既包括 Agent 的先验知识,也包括学习中 Agent 获得的知识,这些知识可以根据 Agent 当前环境的当前状态,为决策控制模块可以提供一个建议动作或一组动作集合。但这个动作是否具有可行性,知识库

中没有相关的知识，还需要 Agent 在学习过程验证，因此 Agent 内部必须要有一个学习机制，来不断修正知识库的知识。

2.3 KBQL 的学习机制

KBQL 的学习机制的基本思想就是[9]：在学习中不断修正由 Agent 的先验知识得到的错误决策。根据 Agent 的知识得到一个 sw 或 w 的动作，但执行这个动作致使 Agent 撞墙。把由 Agent 的知识直接得到一个错误决策（如撞墙）的状态直接称为直接异常状态，用异常标志以）s=true 表示。还有一些潜在的异常状态不能马上识别，是通过学习机制来识别的。

3 带有启发式回报函数的 assr(a 幻学习算法)

强化学习是解决 Markov 决策问题的主要方法，但当 Markov 决策过程具有大规模或连续的状态或行为空间时，强化学习面临两个主要的难题[10]，一个是存储与计算问题，第二个问题是泛化问题（generazition problem）。为了克服“维数灾难”带来的这两个问题，在强化学习中引入了值函数逼近（Valuefnctoin Approximation）器，它是解决强化学习在大规模和连续状态空间中的泛化问题的有效方法。但在基于值函数逼近的强化学习方法中，由于函数逼近器带来的误差或者函数逼近器本身的发散，会使强化学习算法的收敛性变差。因此，如能利用 Agent 的领域知识，引导 Agent 在有意义的状态空间进行搜索，将会改善学习算法的收敛性和学习效率。

4 结束语

笔者首先介绍了强化学习的基本原理及常用的 3 种强化学习算法，其中 Q-学习是一种离策略学习算法，要学习的最优策略与在学习过程中采取的策略无关，这就给了学习过程中的动作选择机制更大的灵活性。本章提出的基于知识的 Q-学习算法（KBQL）就是利用 Q-学习算法的这个特点，利用 Agent 的先验知识来缩小 Agent 学习的状态空间，以加速强化学习的收敛性，同时采用 Agent[11]的学习机制克服其知识的不精确性，从而提高学习算法的鲁棒性和适应性。还对基于值函数逼近的强化学习算法进行研究，详细分析了强化学习过程以及函数逼近器对强化学习的影响。基于 Agent 知识的学习算法不改变原有问题的最优策略，但可以利用这些知识来缩小要学习的状态空间或引导 Agent 向期望的状态空间进行搜索，从而改善强化学习算法的性能。

参考文献：

[1] 伍少成. Agent 的强化学习与通信技术研究及应用[D].广州：华南理工大学，2006.

[2] 胡山立，石纯一.Agent 的意图模型[J]. 软件学报，2000，11(7)：96-97.
HU Shan-li, SHI Chun-yi. Agent's intention model [J]. Journal of Software, 2000,11(7):96-97.

[3] 康小强，石纯一.一种理性 Agent 的 BDI 模[J]. 软件学报，1999，10(12)：268-274.
KANG Xiao-qiang, SHI Chun-yi. A rational Agnet the BDI model[J]. Software, 1999,10(12):268-274.

[4] 马光伟，徐晋晖，石纯一.Agent 思维状态模型[J]. 软件学报，1999，10(4)：42-48.
MA Guang-wei, XU Jin-hui, SHI Chun-yi. Agent mental state model[J]. Software,1999,10(4):42-48.

[5] Oldridge M.This15MYWORLD:TheLogieofanAgent·oriented-DAI testbed,ECAI-94,SPringer.

[6] 刘勇，蒲树祯，程代杰，等.BDI 模型信念特性研究[J].计算机研究与发展，2005，42(1)：54-59.
LIU Yang, PU Shu-zhen, CHENG Dai-jie, et al. BDI model characteristics of faith [J]. Computer Research and Development, 2005,42(1):54-59.

[7] 胡山立，石纯一.一个改进的理性 Agnet-BDI 模型[J].计算机研究与发展，2000，37(9)：125-129.
HU Shan-li, SHI Chun-yi. An improved rational Agnet a BDI model [J]. Computer Research and Development, 2000,37(9):125-129.

[8] 程显毅，夏德深.Agent 思维状态属性的研究[J].南京理工大学学报，2004，28(6)：34-38.
CHENG Xian-yi, XIA De-shen. Agent mental state of the property [J]. Nanjing University, 2004,28(6):34-38.

[9] 于江涛.多智能体模型、学习和协作研究与应用[D].浙江：浙江大学，2003.

[10] Russell S, Novig P. 人工智能：一种现代方法[M]. 姜哲，金奕江，张敏，等译.北京：人民邮电出版社，2010.

[11] 常承阳.基于多 Agent 的在线培训系统设计[J].现代电子技术，2009(18):110-112.
CHANG Cheng-yang.Design of online training system based on multi-Agent[J].Modern Electronics Technique,2009(18):110-112.

基于知识的Agent强化学习算法分析与研究

作者： 殷锋社， YIN Feng-she
作者单位： 陕西工业职业技术学院,陕西咸阳,712000
刊名： 电子设计工程
英文刊名： ELECTRONIC DESIGN ENGINEERING
年，卷(期)： 2011,19(11)