维普资讯 http://www.cqvip.com

基于强化学习的多 Agent 系统*)

唐文彬　朱淼良
(浙江大学人工智能研究所　杭州 310027)

The Multi-Agent System Based on Reinforcement Learning

TANG Wen-Bin　ZHU Miao-Liang
(Institute of Artificial Intelligence, Zhejiang University, Hangzhou 310027)

Abstract　Reinforcement learning allows agent that has no knowledge of an environment to cooperate more efficacious each other. This paper presents an approach for developing multi-agent reinforcement learning systems based on equation principle. The experiment shows agent can produces the desired behavior under all kinds of situation.

Keywords　Agent, Reinforcement learning, EDP, Equation principle

1　引言

强化学习允许自主 Agent 能够在没有关于任务和环境的先验知识的条件下通过不断地反射学习提高自己完成任务的能力，但是强化学习需要大量的计算，也就意味着大量的时间的消耗，对于许多实时系统的计算能力是一个大的挑战。进一步，如果是在多个 Agent 组成的 Agent 社会中，每个 Agent 的行动都有可能导致环境状态的改变，使得普通的相对单个 Agent 的强化学习的计算量大大增加，况且在多 Agent 的社会中，多 Agent 合作完成任务，使任务能够更快更有效地完成，这个已经不是面向单个 Agent 考虑的强化学习可以解决的。因此，在多 Agent 存在的环境中，必须要有针对多 Agent 的强化学习。本文就是提出一种基于对等原理面向多 Agent 的强化学习方法。使得 Agent 不但可以利用其他 Agent 的学习经验，还可以相互合作更有效地完成任务。

2　强化学习基本概念

通过强化学习 Agent 能够更容易地适应环境的变化，当它到达目标状态的时候，Agent 得到奖赏，并且它所做出的合理的策略将被得到加强。强化学习的算法有很多种，Q-学习(Q-leaning)是最著名的强化学习之一。把环境看作是一个离散状态空间，可以用马尔可夫决策过程(Markov Decision Process, MDP)来形式化环境的状态变化。

通常，一个 MDP 可以用一个四元组来表示：$\langle S,A,R,T\rangle$ 其中：S 是环境状态集；A 是 Agent 的行为动作集；R 是奖赏函数 $R:S\times A\rightarrow R$；T 是状态转移函数：$T:S\times A\rightarrow \Pi(S)$。记 $R(s_1,a,s_2)$ 是 Agent 在环境状态 $s_1\in S$ 下采用动作 $a\in A$ 使环境状态迁移到 $s_2\in S$ 时所获得的奖赏值。$T(s_1,a,s_2)$ 是 Agent 在环境状态 $s_1\in S$ 下采用动作 $a\in A$ 使环境状态迁移到 $s_2\in S$ 的概率。

Q-学习在每一个状态下为了提供一个行为选择将搜索全部可能的状态以期得到最优的策略。因此，Agent 最优的策略为在每一个状态选用 Q-值最大的行为。强化学习发生在系统从环境反馈的信息中，表现的是对过去的行为的满意程度。系统利用这些过去的信息决定将来的行为。如果系统在时刻 t 的状态是 s_t，并且采用了行动规划 a_t，因此，系统在时刻 $t+1$ 迁移到状态 s_{t+1}，同时得到强化学习的值 r_{t+1}。策略 π 表明在某个状态下选择行为，需要考虑的是在时刻 t 采纳行为 a_t 的情况下，在时刻 $t+1$ 所能得到的强化学习的值，就是包括了所有子序列强化学习的值的折扣和，也就是折扣奖赏值的和 $Q^*(s,a)$。

$Q^*(s,a)$ 和最优策略 $\pi^*(s)$ 的求解表达式如下：

$$Q^*(s,a)=R(s,a)+\gamma\sum_{s'\in S}T(s,a,s')\max_{a'}Q^*(s',a')$$

$$\pi^*(s,a)=\arg\max_{a\in A}Q^*(s,a)$$

3　对等原理基本概念

在现实生活中，当我们无法通过通讯手段来了解其他人的时候，对他人行为的判断只有采用推测的方法。推测其他人常常使用虚拟条件句，如“如果我是他，我将……”，“现在这种情况下，我会……”等，也就是我们经常把自己作为其他人处的状态中的主体来判断。我们能够这样判断的依据是基于对等原理，也就是两个人处在对等的地位，环境是相同的，因此对待事物的态度和经验，以及判别事物的标准都大致相同，是同类型和同级别的主体。在判别准则相差不大的情况下，这种判别往往比较准确。进一步假如我们为了推测其他人的行为，我们也可以把自己放到那个主体所处的环境中，然后，推理者就可以把他自己在那种环境下将要采取的行为和即将面对的状态作为推理者对其他人在这个状态下的行为和状态变化的推理结果。对等原理在多 Agent 社会同样适用(我们这里考虑的是由完全相同的 Agent 构成的多 Agent 社会环境)，在同一个环境中，每个 Agent 的学习过程都是相似的，因此，在大量学习的作用下，可以相信 $Agent_i$ 和 $Agent_j$ 的学习效果是相当的，或者说是它们最后得到的表示学习成效的 Q-值的表是相似的。对等原理可以非严格地被形式化为：

$$\models T_2(s_1,\alpha)\rightarrow\beta \text{ iff } \models T_1(s_1,\alpha)\rightarrow\beta, T_1,T_2\in T$$

其中：$\alpha,\beta,s\in\Gamma$ 是基本命题符号，以及由逻辑连接符号$\neg$、$\wedge$、$\vee$、$\rightarrow$等生成的复合命题；T 是对等二元知识操作符。

则上式意思是如果 $Agent_1$ 在状态 s_1 的情况下由命题 α

*)本课题得到国防科工委“十五”攻关项目基金“智能机器人”资助。唐文彬　博士研究生，主要研究方向为分布式智能系统、智能机器人。朱淼良　教授，博士导师，主要研究方向为人工智能、计算机视觉、智能机器人。

维普资讯 http://www.cqvip.com

可以得到命题 β，那么 $Agent_2$ 在状态 s_1 的情况下也可以由命题 α 得到命题 β。进一步，记 $In(Agent,s)$ 表示 $Agent$ 的确处在状态 s 下，$Confirm(Agent,\phi,s)$ 表示 $Agent$ 在状态 s 下确信命题 ϕ，我们得到在实际运用过程中常用到的对等原理的变体：

$$Confirm(Agent_1,(In(Agent_1,s_1)\wedge(T_1(s_1,\alpha)\rightarrow\beta)),s_2)$$
$$\rightarrow$$
$$Confirm(Agent_1,(In(Agent_2,s_1)\wedge(T_2(s_1,\alpha)\rightarrow\beta)),s_2)$$

上式表示的是如果 $Agent_1$ 在状态 s_2 的情况下确信 $Agent_1$ 如果在状态 s_1，并且在状态 s_1 的情况下 $Agent_1$ 由命题 α 可以得到命题 β，那么可以得到 $Agent_1$ 在状态 s_2 的情况下确信在状态 s_1 下的 $Agent_2$ 也将由命题 α 得到命题 β。

4 多 Agent 系统强化学习

在多 Agent 的系统中，Agent 不但表现为个体的行为，应该依照自己的规则行动，更重要的是它所应当表现出来的社会行为。多个 Agent 共同完成任务，可能会比单个 Agent 完成更有效率，或者是完成单个 Agent 所不能独立完成的任务。因此在 Agent 做出行动的决策之前，了解其它 Agent 的可能决策是相当地有用。由于每个 Agent 的动作都会带来环境状态的变化，Agent 成为了环境动态的一个因素，环境的变化很难用一个 MDP 来表示。因此，Agent 在多 Agent 的环境中做出决策的过程，是一个通过推测其它 Agent 而综合做出决策的过程，把它称为推测决策过程（Estimate Decision Process，EDP）。Agent 的决策过程分作三步：第一步是根据对等原理推测其他 Agent 的过程；第二步是选择自己的最优动作的过程；第三步是更新学习的过程。记第 i 个 Agent 为 $Agent_i$，则 $Agent_i$ 的 EDP 可以用一个五元组来表示：

$$\langle S^n,a_i,R,T^m,EA\rangle$$

其中：S^n 是 n 个 $Agent$ 联合环境状态集 $(s_1\times s_2\times\cdots\times s_n)$；$s_i\in S$ 是 $Agent_i$ 在当前环境中面临的状态；$a_i\in A$ 是 $Agent_i$ 的行为动作；R 是 $Agent_i$ 的奖赏函数 $R:S^n\times A\rightarrow R$；记 $A^n=(a_1\times a_2\times\cdots\times a_n)$ 为 n 个 $Agent$ 联合动作；T^m 是 n 个 $Agent$ 联合状态转移函数 $T^m:S^n\times A^n\rightarrow\Pi(S^n)$。

EA 是对其余 n-1 个 $Agent$ 的行为动作 $a_j\in A$，$j\in 1,2,\cdots,n,i\neq j$ 的预测函数。根据对等原理，$Agent_j$ 在 $Agent$ 联合环境状态为 $(s_1\times s_2\times\cdots\times s_n)$，自身状态为 $s_j\in$ 的情况下，通过 Q-值的计算和最优策略的选择得出的最优动作 $a_j\in A$，可以用假设 $Agent_i$ 在 $Agent$ 联合环境状态为 $(s_1\times s_2\times\cdots\times s_n)$，自身状态为 $s_j\in S$ 的情况下的情况下，通过 Q-值的计算和最优策略的选择，得出的最优动作是 $a'_j\in A$ 来代替。于是 $Agent_i$ 推测 n 个 $Agent$ 联合动作为 $EA^n=(a'_1\times a'_2\times\cdots\times a'_n)$。于是得到基于对等原理的多 Agent 强化学习折扣奖赏值的和 $Q^*(S^n,a)$ 和最优策略 $\pi^*(S^n)$ 的求解表达式如下：

$$Q^*(S^n,a)=R(S^n,a)+\gamma\sum_{j^{n'}}T(S^n,EA^n,S^{n'})\max_{a'}Q^*(S^{n'},a')$$

$$\pi^*(S^n,a)=\arg\max_{a\in A}Q^*(S^n,a)$$

强化学习过程如下：

1. 初始化所有的 Q-表 $Q^*(S^n,a)$
2. 通过输入和知识库系统判断当前的联合状态
3. 推测联合动作 $EA^n=(a'_1\times a'_2\times\cdots\times a'_n)$
4. 选择一个自身最优行为 $a_i\in A$
5. Agent 实现行为 $a_i\in A$，然后当环境的状态从 S^n 迁移到 $S^{n'}$，Agent 将得到一个奖赏值
6. 更新 Q-表的值
7. 进行到下一个时刻

5 仿真实验结果

为了能验证强化学习的有效性和功能的正确性，我们用多台计算机组成一个仿真平台。仿真实验在由多台 PC 微机，SGI 和 SUN 工作站组成的网络上完成，操作系统主要有 Windows NT 和 Unix 两种。环境中有两个机器人车辆，两个车辆的结构、推理方式等都完全相同，因此，完全适用对等原理的范围。因为推理过程需要进行大量的计算，我们设计了基于 OPS83 系统的 Rete 网结构和知识编码相结合的快速推理机 FRete，推理速度大大提高。图 1 所示是两个机器人车辆在仿真平台中的运行结果：从左到右图中分别是一个机器人车辆直线行进、一个机器人车辆避障行进、两个机器人车辆跟随行进、两个车辆相向行进和其中一辆超越行进的示意图。在跟随状态下，两个车辆保持相同的速度和相同的方向。在相向而行的情况下，车辆互相采取了规避的状态和动作。在超越的状态下，其中一个车辆降低速度并采取回避，另一个车辆加快速度并向另一方向行进。虽然这是在实验条件下进行的，没有考虑许多实际因素，但是也足以说明强化学习的有效性。

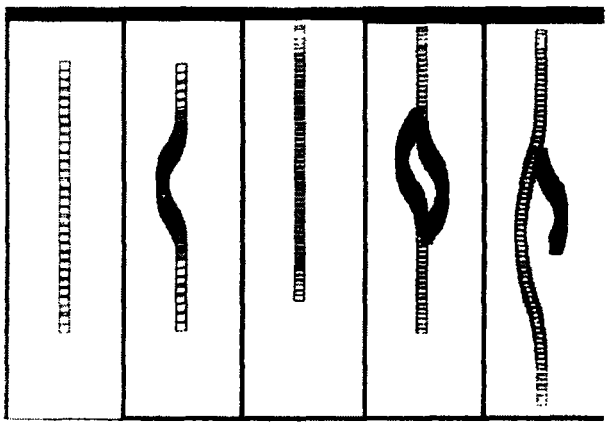

图 1 多机器人车辆仿真实验结果

总结 强化学习可以提高机器人车辆在未知环境中的自学习能力，多 Agent 的强化学习是对单 Agent 强化学习的扩展，并且由于更多 Agent 的加入，决策过程不再是 MDP。本文中提出了基于对等原理的 EDP 强化学习，可以有效地进行自学习，是对多 Agent 强化学习的有力尝试。但是随着 Agent 车辆数量的增加，计算量会急剧增加，因此下一步我们将改进推理算法，利用状态相关性，减少大量不必要的计算，提高车辆行动实时性。

参 考 文 献

1 Watkins C J C H. Learning from Delayed Rewards: [Ph. D. thesis]. Cambridge University, 1989

2 Watkins C J C H. Dayan P. Technical not: Q-learning. Machine Learning, 1992, 8: 279～292

3 Ohashi T, et al. State transition rate based reinforcement learning Systems, Man, and Cybernetics. In: 2000 IEEE Intl. Conf. Volume: 1, 2000. 236～241

4 Yamagnchi T, et al. Propagating learned behaviors from a virtual agent to a physical robot in reinforcement learning. In: Proc. IEEE Int. Conf. on Evolutionary Computation, 1996. 855～859

维普资讯 http://www.cqvip.com

普适计算综述

郑增威[1,2] 吴朝晖[1]
(浙江大学计算机科学与工程系 杭州 310027)[1] (浙江大学城市学院 杭州 310015)[2]

A Survey of Pervasive Computing

ZHENG Zeng-Wei[1,2] WU Zhao-Hui
(Department of Computer Science and Engineering, Zhejiang University, Hangzhou 310027)[1]
(City College, Zhejiang University, Hangzhou 310015)[2]
zhengzw@cs.zju.edu.cn

Abstract Pervasive computing has being now considered and noticed universally, which is the 21st century compute mode. This article first analyzes the origins of generating Mark Weiser's pervasive computing idea and the thrusts of researching it in the process of computing development. It then simply describes the pervasive computing projects of scientific research organizations of main developed countries, such as enterprise and universities, to realize Mark Weiser's idea. Next, it sketches a Layered Pervasive Computing (LPC) model and its related terms. Finally, the article closes with a discussion of great influence of pervasive computing technologies, and makes prospects of human being living environments such as economic, daily life, work, etc., according to the current development of computer hardware and software technologies and the evolvement of researching pervasive computing.

Keywords Pervasive computing, Ubiquitous computing, Compute mode, LPC, Mobile computing, Distributed computing

1 引言

普适计算(Pervasive Computing)概念最早源自 Xerox PARC(Palo Alto Research Center)计算机科学实验室首席科学家 Mark Weiser[1]在 1988 年提出的"Ubiquitous Computing(缩写为 Ubicomp 或 UC)"思想,现在文献中又常以"Pervasive Computing"出现。其基本思想是为用户提供服务的普适计算技术将从用户意识中彻底消失,即用户和周围环境(无数大大小小的计算设备)在潜意识上进行交互,用户不会有意识地弄清楚服务来自周围何处的普适计算技术,就好比我们每天重复着开电灯、关电灯动作,却不会有意识地问自己电来自何方发电厂一样。从人的认知成长过程和计算的发展过程相结合来看,Mark Weiser 的思想可以用图 1 来概略体现。十年后的今天,随着计算机技术的发展,特别是计算机硬件(如低功耗、小型化、微型化轻便硬件设备的出现)的进一步发展和计算机网络技术(如高带宽网络、无线移动网络的普及和组播技术的应用)的不断成熟,普适计算受到计算机科学界更多的关注和重视,合理地近似实现 Mark Weiser 思想成为可能。

Satyanarayanan 撰文[2,3]认为,从分布式计算、移动计算到普适计算的发展历程来看,研究普适计算的动力在于四个方面:智能空间的有效运用(effective use of smart spaces)、不可见性(invisibility)、本地或局部可伸缩性(localized scalability)和屏蔽非均衡条件(masking uneven conditioning)。通过将计算基础结构嵌入到建筑结构中,一个智能空间(Smart Space)可以将两个脱节的世界(指移动和固定)联系在一起。如果一个普适计算环境能够持续满足用户期望,很少使用户意识到它的存在,这要求用户在潜意识上与环境互相影响[4]。而考虑到智能空间的复杂度越来越高,用户个人计算空间和其外部环境相互影响强度会变大。对一个无线移动用户来说,意味着存在严重的带宽、能量和干扰问题。多用户的出现将使这个问题更为复杂化。因此,可伸缩性是普适计算的一个关键问题。而不可见性可能是普适计算中最难的问题,要解决该问题必须在用户接口、环境认知协调等技术领域获得巨大的进步。

从技术上,实现普适计算必须做到满足三个条件:首先,市场上大量出现可供购买的尺寸大小不一、种类繁多的显示设备和廉价、低能耗计算设备;其次,存在将所有计算设备(如嵌入式计算设备、辅助设备)联接在一起的网络;最后,研制出

郑增威 博士生,主要研究领域为人工智能、网络计算。吴朝晖 博士,教授,博士生导师,主要研究领域为人工智能、软件工程、智能 CAD 等。

5 Yamagnchi T, et al. Reinforcement learning for a real robot in a real environment. In: European Conf. on Artificial Intelligence, Aug. 1996. 694～698
6 Hailu G, Sommer G. Embedding knowledge in reinforcement learning. In: Proc. 8th Int. Conf. on Artificial Neural Networks, Sep. 1998. 1133～1138
7 Huber M. A hybrid architecture for hierarchical reinforcement learning. In: Proc. IEEE Int. Conf. on Robotics & Automation, April 2000. 3290～3295
8 Peng J, Bhanu B. Closed loop object recognition using reinforcement learning. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1998, 20(2): 139～154
9 Schwartz J T, Shirir M. A survey of motion planning and related geometric algorithm. Artif. Intell. J., 1988, 37: 157～169
10 Canny J F. The Complexity of Robot Motion Planning. Cambridge, MA: MIT Press, 1988
11 朱淼良,张新晖,吴春明,钱徽. 自主机器人自组织结构 IRASO 的仿真研究. 计算机研究与发展, 1999, 36(7): 776～782
12 王文杰,田家启,史忠植. 多主体系统对其它主体的研究. 计算机研究与发展, 1998, 35(11): 971～974