强化学习中的迁移:方法和进展

王 皓,高 阳,陈兴国
(南京大学软件新技术国家重点实验室,江苏南京 210093)

摘 要: 传统机器学习方法认为不同的学习任务彼此无关,但事实上不同的学习任务常常相互关联.迁移学习试图利用任务之间的联系,利用过去的学习经验加速对于新任务的学习.机器学习各分支都已展开了对迁移学习的研究.本文综述了强化学习的迁移技术,依据认知心理学的理论将现有技术分为行为迁移和知识迁移两大类,并介绍、分析了各自的特点,并提出了一些开放性的问题.

关键词: 迁移学习; 强化学习; 知识; 行为; 认知心理学; 抽象; 泛化
中图分类号: TP18 **文献标识码**: A **文章编号**: 0372-2112 (2008) 12A-039-05

Transfer of Reinforcement Learning: The State of the Art

WANG Hao, GAO Yang, CHEN Xing-guo
(*State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing, Jiangsu* 210093, *China*)

Abstract: Conventional machine learning methods assume that different learning tasks are isolated, but actually they often relate each other. Transfer learning aims at exploiting such relationships and using historical learning experience to improve the learning on new tasks. Much work has been done regarding transfer learning in many sub-domains of machine learning. This paper surveys the transfer of reinforcement learning. According to theories of cognitive psychology, this paper classifies the transfer technologies of reinforcement learning into behavior transfer and knowledge transfer. This paper analyzes the state-of-the-art technologies and some open problems.

Key words: transfer learning; reinforcement learning; knowledge; behavior; cognitive psychology; abstraction; generalization

1 引言

人可以直接学习新的知识,也可以利用旧知识来帮助学习新知识.机器学习自诞生开始就试图模拟人的学习."直接学习新知识"是我们熟悉的传统机器学习问题,现有的方法通常假设学习任务是孤立的,在学习新任务的时候抛弃过去的学习经验和结果.直到上世纪 90 年代随着传统技术的成熟,机器学习才真正重视起"利用旧知识来帮助学习新知识"这一类被称为"迁移学习(transfer learning)"的问题.

尽管起步较晚,但所幸在此前长达两百多年的时间里,一大批优秀的心理学家已经展开了对迁移学习的研究,积累了不少理论(例如文献[1]).我们相信,认知心理学的理论有助于在机器学习领域中对迁移学习的研究.

目前在机器学习的很多子领域(例如神经网络、强化学习等)中,迁移学习的研究已取得了一定进展.由于看问题的角度不同,这些进展以不同的名称出现,例如"归纳迁移(inductive transfer)"、"终身学习(life-long learning)"等等.

本文综述了强化学习领域对迁移学习的研究.我们依据认知心理学的理论将强化学习的迁移分为两大类:行为迁移和知识迁移,并对现有方法做了系统的介绍和分析.

2 强化学习及其迁移

强化学习是一类根据环境反馈来学习的技术.强化学习 agent 辨别自身所处的状态,按照某种策略决定动作,并根据环境提供的奖赏来调整策略直至最优.

2.1 Markov 决策过程

MDP(Markov decision process)事实上已经成为了强化学习任务的标准描述.一个 MDP 可以表达为四元组 $\langle S, A, T, R\rangle$,其中 S 是环境状态的集合,A 是 agent 动作的集合,$T: S \times A \times S \to [0,1]$ 是状态转移概率,$R: S \times A \to \mathbf{R}$ 是瞬时奖赏.概率函数 $\pi: S \times A \to [0,1]$ 称为策略,agent 的学习目标是最优策略 π^*,该策略能使 agent

收稿日期:2008-05-07;修回日期:2008-11-23
基金项目:国家自然科学基金(No. 60775046);国家自然科学基金委创新研究群体科学基金(No. 60721002)

得到最多的回报 $W=\sum_{t=0}^{\infty}\gamma_{r_t}^{t}(0<\gamma<1)$.

MDP 一定有一个确定性的最优策略,并可能有一些非确定性的最优策略. 有几种途径可以学到最优策略 π^*,其中之一是学习每个状态在当前策略下的值函数 $V^{\pi}(s)=E_{\pi}\{W_t|\ s_t=s\}$,而 $\pi^*=\arg\max_{\pi}V^{\pi}$;另一种更常用的方法是学习状态-动作值函数:$Q^{\pi}(s,a)=E_{\pi}\{W_t|\ s_t=s,a_t=a\}$,而 $\pi^*(s)=\arg\max_a Q^*(s,a)$ 就是最优策略. 这里,$W_t=\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}$.

强化学习的经典算法包括 TD 学习、Q 学习、Sarsa 等[2].

2.2 传统强化学习方法的问题

当传统的强化学习的问题空间 $S\times A$ 变得庞大的时候,有两个严重的问题影响了强化学习的实用性. 其一是速率问题:为了寻找最优策略 π^*,强化学习算法必须搜索 $S\times A$ 空间,当 $S\times A$ 变得庞大时,这一过程非常耗时,甚至有些情况下 $S\times A$ 本身就是不可穷尽的,因此强化学习算法常常收敛较慢;其二是复用问题:无论是值函数 $V(s)$ 还是动作值函数 $Q(s,a)$ 或者是策略 π,强化学习的结果总是依赖于 $S\times A$ 的具体表示,这意味着只要问题略微改变,以前的学习结果就变得毫无用处. 但对于某些实际问题,由于训练代价较高,学习结果的可复用性是非常重要的.

上述两方面的原因激励了对强化学习进行迁移,因为本质上,迁移学习就是复用过去的学习经验和结果以加速对于新任务的学习.

2.3 迁移学习的分类

美国心理学家 Anderson 提出了 ACT(adaptive control of thought)理论[1],将认知分为过程性的和陈述性的,并把认知的迁移分为两个阶段:首先是过程性认知上升为陈述性认知,然后陈述性认知在任务间迁移,并在新任务中产生新的过程性认知. 撇开心理学的术语,具体到强化学习任务中,我们认为过程性认知对应于动态的行为(策略或动作),而陈述性认知则是 agent 所掌握的有关任务的静态知识(包括状态感知、V 值和 Q 值等). 强化学习值得我们注意的一个特征是,V 值和 Q 值与策略 π 紧密联系,以至于有时两者的划分并不那么严格,我们也可以将 V 和 Q 看成行为.

考虑到强化学习的特点(学习的目标是动态的策略,即行为),我们将强化学习中的迁移分为两类:行为迁移和知识迁移,如图 1 所示. 图中的粗箭头表示两种迁移的主要方向,但实现中可能经过虚线箭头所示的步骤.

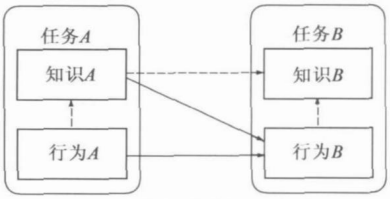

图1 迁移学习的两种方式

3 行为迁移

行为迁移通常意味着将先前学到的策略或者某些公共的"子过程"用于新任务的学习. 这一类技术侧重于挖掘、利用不同任务的解决方案之间的相似性.

3.1 直接的策略迁移

若状态空间 S 和动作空间 A 保持不变,则可以把过去任务的最优 Q 值当作新任务的初始 Q 值,通过这种方式在新任务中使用过去的经验[3].

这种方法有一个问题:旧的 Q 值固然包含两个任务间共性,但无疑也包含了一些干扰信息,不加区分的迁移会导致学习的质量降低(称为"负迁移"). 为此,文献[3]提出依靠经验观察,只迁移部分状态的 Q 值(那些可能有用且副作用不大的),以此平衡迁移效率和最优性损失. 进一步,文献[4]提出了一种在迁移过程中加强探索的动作选择策略:他们使用逐步递减的概率复用过去的策略,而用剩余的(递增)概率对新任务空间进行 ε 贪心探索.

显然,很多情况下 S 和 A 保持不变的假定违反事实. 文献[5]放宽了这一限制,提出用映射 $\rho:\pi^*_{\text{past}}\to\pi^*_{\text{new}}$ 或映射组

$$\begin{cases}\rho_S:S_{\text{new}}\to S_{\text{past}},\\ \rho_A:A_{\text{past}}\to A_{\text{new}}\end{cases}$$

去适配过去任务的 π_{past},使之成为能够用于新任务中的策略. 到目前为止,所有的这些映射或映射组都是人工定义的.

3.2 option 的迁移

行为迁移的另一种思路是寻找并复用任务之间的公共"子过程",通常是宏动作(macro-action)或 option.

option 是对 MDP 中策略概念的推广,表示为三元组 $\langle I,\pi,\beta\rangle$,其中 $I\subseteq S$ 是起始状态集,$\pi:S\times A\to[0,1]$ 是策略,$\beta:S^+\to[0,1]$ 是终止概率,只有处在 I 中的状态时才可以执行 π,并由 β 决定是否终止执行.

Bernstein 假设任务 $\{X_i\}$ 服从某未知但固定的分布[6]. 他将过去任务中的一些非确定性最优策略 π_i^* 做简单平均,得到"混合策略" π_M 并构造"复用 option"为 $\langle S,\pi_M,n\rangle$,其中 $n\in N$ 表示 option 执行的步数."复用 option"用于迁移的基本原理在于取平均可以突出各个 π_i^* 之间一致性较强的部分,而这种一致性有助于反映各学习任务之间的共性. 但在另一方面,"复用 option"的问题是 π_i^* 的选择和 n 的设定依赖于人,并且会显著影响迁移的效果,更本质的不足在于它缺乏一个明确的目标来指导 option 的构造.

Pickett 等人则利用过去任务的确定性最优策略 $\{\pi_i^*\}_{i=1}^n$ 来构造 option[7]. 他们将 $\{\pi_i^*\}_{i=1}^n$ 的幂集 $\{\{\pi_{k_1}^*, \cdots \pi_{k_m}^*\} \mid 1 \leq k_1, \cdots, k_m \leq n\}$ 中的所有元素逐个进行严格匹配,抽出一致的部分做候选 option,随后对每一个候选的 option 进行评估. 显然,如果候选 option 本身的规模较大(指包含了较多的状态)并且产生它的集合包含的元素较多,那么就有理由认为它代表了任务之间的某种共性.

一种更具目的性的 option 构造方法是对过去的行为轨迹进行挖掘,去寻找子目标[8]. 直观地看,如果一个状态在成功的轨迹中被访问得多,而在失败的轨迹中被访问得少,那么它极有可能是原任务的一个重要子任务,因此可以将完成该子任务的策略构造成 option 在任务之间迁移. 监督学习中的多示例学习方法可以用在这里寻找子目标.

3.3 层次强化学习

与自底向上构造 option 的思路相反,迁移学习的另一种思路是主动将任务分解,构造任务和子任务的层次关系,再迁移用于完成子任务的策略. 层次强化学习(HRL)是这一类方法的有力工具,其实质是对动作的逐级抽象.

Dietterich 的 MAXQ 算法[9]是 HRL 的经典算法,MAXQ 利用 option 框架将 MDP 分解成若干个 SMDP,每个 SMDP 包含 3 个要素:状态抽象 B、动作(可被调用的子 SMDP) A 和终止预测 G. 我们可以自下而上学习各层 SMDP 的最优策略. 文献[10]进一步提出在任务树的每一个结点上存储完成任务的期望时间和奖赏,这样可以更有效地迁移任意子树.

3.4 MDP(SMDP)同态

Ravindran 等人使用代数方法对 SMDP 抽象进行了更深入的研究[11]. 他们将一个 SMDP M 同态映射到另一个 SMDP M',然后将解决 M 的策略提升为 M 解决的策略. Ravindran 等人的突出成果是:只要 SMDP 映射符合他们提出的同态定义,那么 M 的最优策略就一定能被提升为 M 的最优策略.

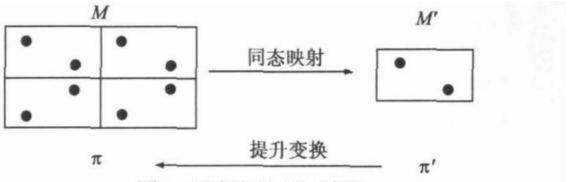

图2　同态方法的示意图

图 2 是同态方法的原理示意,在完成 M 的 4 个子任务时都可以使用 M 的策略,只不过为了完成下方的两个子任务需要对该策略做一些转换,Ravindran 等人称这种转换为"相对化". 从这个简单的示意图也可以看出,同态方法特别适用于具有某种对称性的问题.

对于复杂一些的情形,找同态映射是 NP 难问题. 现有的方法是提供给 agent 一些备选的映射,让 agent 去挑最合适的. Soni 等人利用新旧任务之间状态的关系来提供这样的备选映射[12].

4 知识迁移

不同于侧重解决方案的行为迁移,知识迁移技术注重对任务本身的理解,并试图学习解决问题的一般原理,因此知识迁移技术更多地涉及知识表示、规则提取等内容. 实际上,知识迁移的方法更符合 2.3 节所描述的现代心理学观点,即迁移学习过程不是简单的行为模仿,而是包含了抽象和理解在内的复杂过程.

4.1 值函数的迁移

一大类知识迁移方法的关注焦点是值函数 $V: S \to \mathbf{R}$. 传统的强化学习中值函数 V 强烈依赖于具体的任务,因此研究者们希望将 V 和具体任务分离.

Konidaris 等人提出了任务与论域分离的迁移学习框架. 在他们的框架中,agent 的每一个任务状态 s 对应一个论域描述 c,agent 学习一个论域值函数 $L: C \to \mathbf{R}$,并将它在任务间迁移.

Mahadevan 研究了全体值函数空间 $\mathbf{R}^S$ 的拓扑[13]. 他发现传统强化学习的 V 依赖于具体任务的根本原因是在迭代过程将 $\mathbf{R}^S$ 当作 $\mathbf{R}^{|S|}$ 处理,在策略迭代过程[2]使用 $\mathbf{R}^{|S|}$ 的欧式基,这实际上损失了 S 的结构信息. 于是,Mahadevan 提出直接使用 $\mathbf{R}^S$ 的抽象 Fourier 基. 这些新的正交基被称为"原值函数(proto-value function)",它们与具体的任务无关,而它们的线性组合却可以表达任意值函数,原理上,它们反映了论域的性质.

4.2 启发式信息的迁移

被迁移的知识还可能是引导 agent 完成任务的启发式信息. 这一类迁移学习方法通常利用监督学习方法从强化学习的结果中提取一些产生式规则.

Madden 等人的"渐进(progressive)强化学习"[14]使用 C4.5 决策树对 Q 学习过程进行日志记录,并学习日志记录得到一些概括的规则,这些规则在渐近复杂的任务间迁移. 每当 Q 学习的过程中遇到未探索过的状态,符号学习器就会根据当前规则给出一个动作建议. 显然,建立在过去经验上的规则会有助于当前的学习.

实际上,针对具体的学习任务,还有很多具体的迁移启发式信息的方式. 例如,文献[15]处理了一种动作集 A 比较庞大但真正有用的动作较少的任务. 他们将旧任务的最优策略涉及到的那些动作 $A^* \subseteq A$ 作为启发性知识迁移给新的任务.

4.3 关系强化学习

关系强化学习(RRL)是强化学习在新千年的重要发展之一,它通过用关系来表达状态和动作. 关系强化

学习将强化学习带到了一个新的层次,使用关系的语言,我们可以自然地描述、完成一系列相关的任务,解决学习的迁移问题.

Driessens 等人通过将目标和策略参数化,使原本只适用于解决单一任务(例如"将书放在桌子上")的策略可以解决一大类问题("将 X 放在 Y 上")[16]. 当然这些任务之间需要有相同的结构. 为了在不同结构的任务之间迁移,还可以对结构本身再次参数化,最终可能形成一个参数化的任务层次.

尽管关系强化学习有非常令人期待的前景,Taylor 等人仍声称"不是所有的任务都能很容易地表达成关系问题"[17]. 他们是在研究 3 维的 Mountain Car 问题时得出的这一观点.

5 迁移学习的其他维度和发展方向

我们将迁移学习划分为"行为迁移"和"知识迁移",并介绍了相关的研究进展. 实际上,我们还可以按其他的标准对迁移学习方法进行划分. 例如,按照迁移的内容是否经过抽象可以将迁移过程分为"低通路(low-road)迁移"和"高通路(high-road)迁移". 低通路迁移主要迁移一些高度联系过的知识和技能,而高通路迁移则是迁移抽象的概念和方法. 大致地,我们可以将上述内容总结成图 3,这张图也反映了 10 余年来迁移学习技术发展的两个主要方向.

图3 迁移学习的二维分类

结合图 3 和第 3、4 节的描述可以看出,从低通路迁移到高通路迁移的发展过程反映了对被迁移信息的理解越来越深入;另一方面,从行为迁移到知识迁移的发展过程则反映了对任务本身的理解的深入.

6 一些开放性话题

对强化学习迁移的研究才刚刚起步,还有很多开放性的问题需要得到解答. 这里我们列出几点在我们看来比较重要的问题.

6.1 任务的相似度度量

并不是所有的迁移都能促进新任务的学习,相反,不恰当的迁移会导致在新任务上学习质量的降低[3]. 我们需要仔细辨别新旧任务的相似点和相似程度,在此基础上选择是否迁移以及迁移什么. 更进一步,我们还希望 agent 在面临新任务的时候能够自动地解决这些问题,而不是依赖于人工的映射或任务分解.

6.2 多 agent 迁移学习

传统机器学习在学习任务上的扩展带来了迁移学习问题,另一方面,传统方法在 agent 数上的扩展则带来了多 agent 学习(MAL)问题,因此我们可以认为迁移学习和 MAL 是互补的. 实际上在现实世界中,最普遍存在的学习是多人协作完成多个学习任务,例如球队的队员要相互配合战胜一个又一个对手. 这种"多 agent 的迁移学习"给我们带来了更大的挑战.

6.3 迁移学习的应用

迁移学习在多个实际问题中得到了应用. 迁移学习的许多结果都针对机器人足球赛(RoboCup)或运用在其中(例如文献[18]);文献[10]将他们的研究成果用在了实时策略游戏中;DARPA 将迁移学习的研究结果应于 CALO 计划,该计划的目标是开发一个基于 AI 技术的办公室助手,能够辅助管理人的日常工作. 我们还可以去发掘迁移学习更多的应用.

7 结束语

迁移学习是一个既老又新的研究方向,而强化学习则是一机器学习中颇具前景的技术领域. 本文扼要介绍了强化学习领域中的迁移学习研究进展. 我们借鉴了成熟的认知心理学理论,将强化学习的迁移划分为两大类:行为迁移和知识迁移,前者侧重解决方案的模仿和复用,而后者偏重于对问题本身的理解和抽象.

我们认为,迁移学习与多 agent 学习可以看作互补问题,前者是传统学习在学习任务方向的扩展,而后者是传统学习在 agent 方向的扩展,多 agent 迁移学习将是发展的趋势. 我们最后简单介绍了迁移学习的应用. 限于篇幅,不少话题难以展开,但我们相信,这篇文章会给对该研究方向感兴趣的人带来帮助.

参考文献:

[1] Anderson J R. Cognitive Psychology and Its Applications (third edition) [M]. New York: Freeman, 1990.

[2] Sutton R S, Barto A G. Reinforcement Learning [M]. Cambridge: MIT Press, 1998.

[3] Bowling M, Veloso M. Reusing learned policies between similar problems [A]. Proceedings of AI * IA-98 Workshop on New Trends in Robotics [C]. Berlin, Germany: Springer Verlag. 1998.

[4] Fernández F, Veloso M. Probabilistic policy reuse in a reinforcement learning agent [A]. Proceedings of the Fifth International Conference on Autonomous Agents and Multi-Agent Systems [C]. New York: ACM, 2006.

[5] Fernández F, Veloso M. Policy reuse for transfer learning across tasks with different state and action spaces[A]. Proceedings of The ICML-06 Workshop on Structural Knowledge Transfer for Machine Learning[C]. New York: ACM, 2006.

[6] Bernstein D S. Reusing old policies to accelerate learning on new MDPs[R]. Amherst: Amherst College, University of Massachusetts, 1999.

[7] Pickett M, Barto A G. PolicyBlocks: an algorithm for creating useful macro-actions in reinforcement learning[A]. Proceedings of the Nineteenth International Conference on Machine Learning [C]. San Francisco: Morgan Kaufmann, 2002. 506 - 513.

[8] Mcgovern A, Barto A G. Automatic discovery of subgoals in reinforcement learning using diverse density[A]. Proceedings of the Eighteenth International Conference on Machine Learning[C]. San Francisco: Morgan Kaufmann, 2001. 361 - 368.

[9] Dietterich T G. Hierarchical reinforcement learning with the MAXQ value function decomposition[J]. Journal of Artificial Intelligence Research, 2000, 13(2): 227 - 303.

[10] Mehta N, Natarajan S, Tadepalli P, A Fern. Transfer in variable-reward hierarchical reinforcement learning[A]. Proceedings of the NIPS-05 Workshop on Inductive Transfer[C]. Cambridge: MIT Press, 2005. 360 - 366.

[11] Ravindran B, Barto A G. SMDP homomorphisms: an algebraic approach to abstraction in semi-Markov decision processes [A]. Proceeding of the Eighteenth International Joint Conference on Artificial Intelligence [C]. San Francisco: Morgan Kaufmann, 2003.

[12] Soni V, Singh S. Using homomorphisms to transfer options across continuous reinforcement learning domains[A]. Proceedings of the Twenty-first National Conference on Machine Learning[C]. Boston: AAAI Press, 2006.

[13] Mahadevan S. Proto-value functions: developmental reinforcement learning[A]. Proceedings of the Twenty-second International Conference on Machine Learning [C]. New York: ACM, 2005.

[14] Madden M G, Howley T. Transfer of experience between reinforcement learning environments with progressive difficulty [J]. Artificial Intelligence Review, 2004, 21(3): 375 - 398.

[15] Sherstove A A, Stone P. Improving action selection in MDP's via knowledge transfer[A]. Proceedings of the Twentieth National Conference on Artificial Intelligence [C]. New York: ACM, 2005.

[16] Driessens K, Ramon J, Croonenborghs T. Transfer learning for reinforcement learning through goal and policy parameterization[A]. Proceedings of the ICML-06 Workshop on Structural Knowledge Transfer for Machine Learning[C]. New York: ACM, 2006.

[17] Taylor M E, Kuhlmann G, Stone P. Autonomous transfer for reinforcement learning [A]. Proceedings of the Seventeenth International Conference on Autonomous Agents and Multi-Agent Systems[C]. Estoril, Portugal: IFAAMAS, 2008.

[18] Torrey L, Shavlik J, Walker T, Maclin R. Skill acquisition via transfer learning and advice taking [A]. Proceedings of the Seventeenth European Conference on Machine Learning[C]. Berlin, Germany: Springer, 2006. 425 - 436.

作者简介:

王　皓　男,1983 年 1 月生于江苏扬州,现为南京大学计算机系硕士研究生,主要研究方向为强化学习、agent 理论、多 agent 系统.
E-mail: leafwanghao@gmail.com.

高　阳　男,1972 年出生,2000 年 3 月获南京大学计算机系博士学位,现为南京大学计算机系副教授,主要研究方向为分布式人工智能、机器学习,已发表论文 50 余篇.

陈兴国　男,1984 年 2 月出生,现为南京大学计算机系硕士研究生,主要研究方向为强化学习.