Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2017, 6(3), 91-96
Published Online August 2017 in Hans. http://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2017.63010

Hans汉斯

Adaptive Dynamic Programming Based on Hierarchical Learning

Qiao Lin, Minshuo Li

College of Mathematics Physics and Information Engineering, Zhejiang Normal University, Jinhua Zhejiang
Email: linqiao@zjnu.cn, lmshappy@zjnu.cn

Received: Jul. 21st, 2017; accepted: Aug. 1st, 2017; published: Aug. 4th, 2017

Abstract

This paper introduces an adaptive dynamic program method based on hierarchical learning. The motivations for this idea come from the levels of consciousness (LOC) model, which address the interdependence between consciousness and action in baby's development. The introduction of a multilevel goal representation into the adaptive critic is able to guide the system's decision-making to accomplish the long-term goal over time, mimicking certain levels of brain-like intelligence. The detailed system architecture, learning and adaption procedure are presented, and the learning and control capability of this approach is verified through light control in GLD (Green Light Domain).

Keywords

Cognitive, Adaptive Dynamic Programming, Neural Network, Hierarchical Learning

基于分层学习的自适应动态规划

林　巧，李旻朔

浙江师范大学，数理与信息工程学院，浙江　金华
Email: linqiao@zjnu.cn, lmshappy@zjnu.cn

收稿日期：2017年7月21日；录用日期：2017年8月1日；发布日期：2017年8月4日

摘　要

本文基于婴儿的认知发育模型LOC (Levels of Consciousness)提出了基于分层学习的自适应动态规划方法以改进学习和优化。根据LOC模型中感知的层次性以及工作目标的层次定义，为自适应动态规划设计了多层的目标网络结构及相应的分层学习方法。在自适应评价中引入多层的目标表征将引导系统做出好

文章引用: 林巧, 李旻朔. 基于分层学习的自适应动态规划[J]. 人工智能与机器人研究, 2017, 6(3): 91-96.
DOI: 10.12677/airr.2017.63010

的决策并最终实现目标。文中给出了分层自适应动态规划的系统结构、学习和自适应过程，并通过模拟系统GLD (Green Light Domain)，在自适应交通信号控制模拟实验上验证了该方法的学习和控制能力。

关键词

认知发育，自适应动态规划，神经网络，分层学习

1. 引言

理解大脑的智能，并开发出能模拟相应智能水平的自适应系统，一直是人类伟大的追求之一，也是未解的科学挑战[1]。随着脑研究和现代技术的发展，科学家和工程师们热切希望能找到一条有效的路径来构建高适应性和鲁棒性的复杂系统，且系统对非确定性和非结构化环境具有很好的容错能力。然而，虽然许多重要的基础研究以及工程应用取得了成功，但要实现真正人类脑的全智能机器仍是任重道远。一个基本的关键问题是如何设计智能系统，让其不断学习优化、学习预测以实现最终的目标。本文我们给出了基于分层学习的自适应动态规划来处理这个问题。

在近二十年来，工程和科学的不同研究分支都对机器智能研究广泛关注并取得了很多进展。其中自适应动态规划(ADP)被普遍认为是近似最优行为策略的唯一通用途径，且在许多复杂系统应用中展示了其具有巨大潜能来达到一定的智能水平，在某一程度上近似于真正的智能。简而言之，ADP 的关键思想是建立在 Bellman 等式[2]的基础上，依赖于与环境的不断交互最终得到最优。比如，给定一个系统的性能成本函数，动态规划的目标是选择控制序列 u 使成本函数最小，即公式(1)：

$$J^*\left(X\left(t\right)\right)=\min_{u(t)}\left\{U\left(X\left(t\right),u\left(t\right)\right)+\alpha J^*\left(X\left(t+1\right)\right)\right\} \tag{1}$$

其中，$X\left(t\right)$ 是系统的状态向量，u 是控制行为，U 是效用函数，是折扣因子。为了实践可行 ADP 使用函数来近似成本函数。比如，一种通用的近似器是使用 BP 算法的神经网络(NN)，其被研究者广泛研究，并被应用于许多不同的领域，包括工业控制、直升机控制、交通信号控制、电力控制等等[3] [4] [5] [6]。

本文，我们提出了基于分层学习的自适应动态规划来改进学习和优化。设计思路主要受启于生物系统的多阶段多层目标表征，比如一个生物系统在不同发育阶段有不同的目标，或者在同一个阶段有不同的目标。最明显的是在婴儿的认知发育过程中所体现的阶段性和层次性。Levels of Consciousness (LOC) 是关于幼儿认知发育过程的理论建模，基于 LOC 我们认为基于多层的强化信号表征有助于层次目标的形成与发育，并通过分层学习以及自顶向下或自底向上的方法来实现智能决策过程。

2. 层次自适应动态规划

LOC 认为婴儿的意识在三个维度上通过多个阶段进行发育，三个维度是：语义记忆、意识层次和工作记忆。语义记忆是客观信息的存储，比如事实和对象，语义记忆的发展促使意识层的发展，也就是联系语义记忆中的信息以完成工作记忆中的目标的能力。工作记忆是认知系统的行为部分，行为部分的发展促使婴儿完成更复杂的目标和任务。

Anderson 和 Bothell 关于认知结构的假设称为理性思维的适用性控制 ACT-R (Adaptive Control of

Thought-Rational)，通过分析比较，LOC 不同的发展阶段和组元可以映射为 ACT-R 中的不同部件[7] [8] [9]。从而我们可以基于 LOC 认知心理学模型和 ACT-R 认知计算模型，设计分层多模块的认知发育模型，为认知发育算法的设计提供基础的通用框架。

2.1. 系统层次结构

图 1 给出了概念图，使用分层目标网络来构建不同的目标层次。与传统的 actor-critic 设计相比，通过学习过程，表征多层目标。不同阶段的目标构建在分层结构内部，引导系统完成最终目标。高层目标为低层学习提供建议，即自顶向下学习；行为网络得到控制行为，评价网络评价当前控制行为，底层的目标通过与评价网络交互，修正上层的目标表征，即自底向上学习。通过这种方式，这个结构把自顶向下和自底向上学习联合在一起共同完成最终的目标。

2.2. 体系结构的设计与实现

系统的下层部分包括行为网络和评价网络，这与神经动态规划 NDP 相似。这一部分设计与 ADHDP 紧密联系，最主要的差别是该结构没有系统模型。ADHDP 中的系统模型用于预测系统未来的状态值，从而得到下一步的成本函数。而该结构只是简单记录前一时间的 J 值，一旦得到了当前 J 值，就可以计算得到时间差分 TD，并使用 TD 进行训练。这样，不需要额外的模型网络来预测系统未来的状态，也不用考虑在评价网络权值之上。

系统的上层部分是设计的关键内容，这部分包含了一系列层次化的目标网络来表征不同层次的目标，以加速学习。图中共有 L 个目标网络，每一层的目标网络的输入是当前系统的状态 $X(t)$、当前行为 $u(t)$、以及来自上层的目标输出 $s(t)$。主强化信号 $r(t)$是顶层目标网络的输入，其表示最后的目标状态，比如“成功”或“失败”。基于主强化信号，每层的目标网络都会输出一个目标信号 $s(t)$，这些目标信号构建了一个内部的值序列，用来表征当前控制行为的好坏。因此，需考虑内部目标信号作为不同层次的次目标信号，这种内部的目标表征为整个系统提供了一个丰富信息的学习目标，因此可以期望能引导系统改进学习性能。

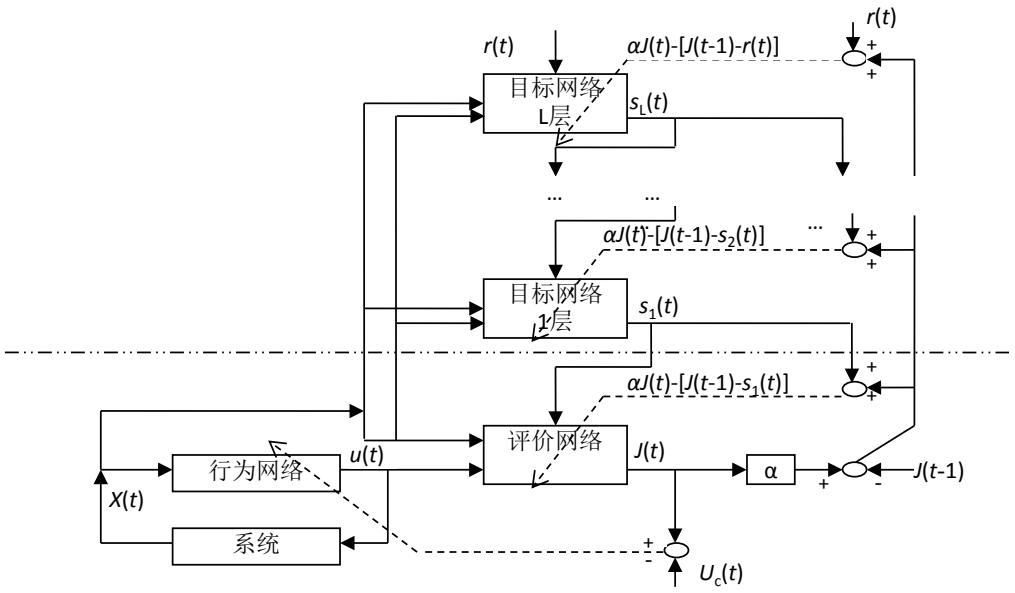

Figure 1. System hierarchical structure diagrams
图 1. 系统层次结构图

目标网络与评价网络直接交互，目标网络与行为网络间接交互。我们首先给出待优化的误差函数，用于修正第 m 层目标网络的参数，其定义见公式(2)和(3)：

$$e_{fm}(t)=\alpha J(t)-\left[J(t-1)-s_{m+1}(t)\right] \tag{2}$$

$$E_{fm}(t)=\frac{1}{2}e_{fm}^2(t) \tag{3}$$

顶层 L 的 $s_{m+1}(t)$由主强化信号 r 替代，m 层目标网络的输出 $s_m(t)$作为 $m-1$ 层目标网络的输入，同时也用于定义误差函数以调节本层目标网络的参数。最底层的输出 $s_1(t)$用于定义评价网络的误差函数，见公式(4)和(5)：

$$e_c(t)=\alpha J(t)-\left[J(t-1)-s_1(t)\right] \tag{4}$$

$$E_c(t)=\frac{1}{2}e_c^2(t) \tag{5}$$

在这个体系结构中，所有网络参数的学习和适应采用 BP 规则。我们的网络采用多层感知机 MLP，只有一个隐层，图 1 给出了行为网络、评价网络和目标网络的结构图。我们假定系统的状态 S 有 n 个变量，这样行为网络输入信号的个数为 n，而评价网络和目标网络输入信号的个数为 $n+2$。顶层目标网络没有更高层目标信号的输入，因此只有 $n+1$ 个输入。如果合理使用 BP (反向传播)规则，那么该学习方法可以泛化应用于任何函数近似。

行为网络的学习与文献[10] [11]相似，我们主要讨论目标网络和评价网络的学习，以展示它们之间如何通过交互改进学习。

我们首先定义评价网络和目标网络的输出，对于评价网络，根据公式(6)计算得到输出 $J(t)$。

$$J(t)=\sum_{i=1}^{N}w_{c_i}^{(2)}(t)\,p_i(t) \tag{6}$$

其中，$p_i(t)=\dfrac{1-\exp^{-q_i(t)}}{1+\exp^{-q_i(t)}},i=1,\cdots,N$；$q_i(t)=\sum_{j=1}^{n+2}w_{c_i}^{(1)}(t)\,x_j(t),i=1,\cdots,N$

q_i 是第 i 个隐层结点的输入，p_i 是第 i 个隐层结点的输出，$n+2$ 是输入的总个数，包括来自行为网络的 $u(t)$和最底层目标网络的 $s_1(t)$。

对于 m 层的目标网络，其输出定义见公式(7)~(10)。

$$s_m(t)=\frac{1-\exp^{-k_m(t)}}{1+\exp^{-k_m(t)}} \tag{7}$$

$$k_m(t)=\sum_{i=1}^{N}w_{f_{mi}}^{(2)}(t)\,y_{mi}(t) \tag{8}$$

$$y_{mi}(t)=\frac{1-\exp^{-z_{mi}(t)}}{1+\exp^{-z_{mi}(t)}},\ i=1,\cdots,N \tag{9}$$

$$z_{mi}(t)=\sum_{j=1}^{n+2}w_{f_{mi,j}}^{(1)}(t)\,x_j(t),\ i=1,\cdots,N \tag{10}$$

下标 m 对应 m 层目标网络，z_i 是第 i 个隐层结点的输入，y_i 是第 i 个隐层结点的输出，k 是输出结点的输入，N 是隐层结点的总个数，$n+2$ 是输入的总个数，包括来自行为网络的 $u(t)$和来自上层目标网络

的 $s_m(t)$，最高层的目标网络只有 $n+1$ 个输入，需要做相应的修改。

m 层目标网络的输出 $s_m(t)$是下一层$(m-1)$目标网络的输入，从而构建了一个互联的链路，一直到评价网络，在链中应用 BP 规则来调整目标网络的参数 w_{fm}。

3. 案例研究

交通信号灯是城市交通信号控制(Traffic Signal Control, TSC)系统的关键组成部分，对自适应城市 TSC 系统的优化研究目标是能自适应城市交通道路交通流量的实时变化，实现对交通流的最优控制。我们在开源交通仿真软件 GLD 中验证基于分层学习的自适应动态规划方法的性能。

在 GLD 实验平台上对基于四相位十字交通路口进行仿真实验。实验中选取的目标优化参数包括，车辆平均运行时间、车辆平均等待时间、车辆排队长度。交通流量的采集每 1 次/分钟，分为 10 个时间段，交通流量产生频率按照由大变小再变大的方式设计，模拟现实交通流的从早高峰到平峰，再到晚高峰的变化趋势。仿真平台 GLD 仿真实验以 1000 个周期为标准，分别对参数平均车辆排队长度 Average Waiting Queue Length (AWQL)、平均旅行时间 Average Travelling Waiting Time (ATWT)、到达目的地车辆数目、平均路口等待时间 Average Junction Waiting Time (AJWT)得出了相应的测试数据，见表 1。

由得出的实验数据表 1 可以看出，基于 HL-ADP 算法的单路口交通信号优化控制比通用的定时配置方案(Fixed-time)、最长队列优先(Longest-Queue)、遗传算法(GA)、神经网络方法(Neural Network)在 1000 个测试周期时间内取得比较好的效果。

在 1000 个模拟周期内，基于 HL-ADP 的城市 TSC 系统的平均路口等待时间(AJWT)的变化曲线是先上升再下降的。因为 HL-ADP 算法是不断与环境进行交互学习的，所以开始阶段会呈现 AJWT 上升很快的趋势，随着学习的进行，优化性能是不断提高的。实验结果表明：对于城市单路口，采用本文提出的基于 HL-ADP 的城市 TSC 系统算法，能有效降低车辆排队长度和车辆在路口的延误时间。

4. 结语

在本文我们提出了基于分层学习的自适应动态规划方法来改进学习和优化。在自适应评价设计中增设了层次目标网络，给出了多阶段多层次的目标表征，通过目标网络与评价网络直接交互以及与行为网络的间接交互改进学习。我们使用一系列的相互连接的目标网络来表征层次目标，顶层的目标网络接受来自外界的主强化信号，主强化信号是学习系统的最终目标，中间层的目标网络给出关于当前行为的富信息目标表征。我们给出了详细的体系结构以及分层学习和自适应过程，在 GLD 上的模拟结果结构证明了该方法的有效性。

Table 1. The average of single crossing traffic simulation evaluation index
表 1. 单路口交通仿真评价指标参数平均值

算法	测试周期	平均车辆排队长度(AWQL)	平均旅行时间(ATWT) (cycle)	到达目的地车辆数量	平均路口等待时间(AJWT) (cycle)	预期仿真时间 (cycle)
Fixed-time	1000	29.66	209	33962	96	1000
Longest-Queue	1000	20.05	153	35603	74	1000
GA	1000	16.27	128	39008	75	1000
NN	1000	14.14	98	40058	70	1000
HL-ADP	1000	9.63	77	42360	58	1000

分层学习是理解大脑智能的关键内容，未来在这个领域会有大量的有趣的研究方向。本文给出了体系结构的设计与实现、学习过程和模拟实验。关于理论方面的研究也是很重要的，比如收敛性和稳定性的证明，这种理论分析可以加深理解该方法的本质。

基金项目

浙江省教育厅项目(项目编号：Y201328256)。

参考文献 (References)

[1] Werbos, P.J. (2009) Intelligence in the Brain: a Theory of How It Works and How It Build It. *Neural Networks*, **22**, 200-212. https://doi.org/10.1016/j.neunet.2009.03.012

[2] Bellman, R.E. (1957) Dynamic Programming. Princeton University Press, Princeton.

[3] Enns, R. and Si, J. (2004) Helicopter Flight Control Using Direct Neural Dynamic Programming, Handbook of Learning and Approximate Dynamic Programming. *IEEE Transcations on Neural Networks*, **14**, 535-559.

[4] Fu, J., He, H. and Zhou. X. (2011) Adaptive Learning and Control for Mimo System Based on Adaptive Dynamic Programming. *IEEE Transcations on Neural Networks*, **22**, 1133-1148. https://doi.org/10.1109/TNN.2011.2147797

[5] Hen, C.C. (2007) An Approximate Dynamic Programming Strategy for Responsive Traffic Signal Control. *Proceedings of the* 2007 *IEEE Symposium on Approximate Dynamic Programming and Reinforcement Learning*, Honolulu, 1-5 April 2007, 303-310.

[6] Zhang, H.G., Wei, Q.L. and Luo, Y.H. (2008) A Novel Infinite-Time Optimal Tracking Control Scheme for a Class of Discrete-Time Nonlinear Systems via the Greedy HDP Iteration Algorithm. *IEEE Transaction on System, Man and Cybernetics*, **38**, 937-942. https://doi.org/10.1109/TSMCB.2008.920269

[7] Zelazo, P.D. (2004) The Development of Conscious Control in Childhood. *TRENDS in Cognitive Sciences*, **8**, 12-17. https://doi.org/10.1016/j.tics.2003.11.001

[8] Lebiere, C. (1998) The Dynamics of Cognition: An ACT-R Model of Cognitive Arithmetic. Carnegie Mellon University, Pittsburgh.

[9] Ron, S. (2012) Memory Systems within a Cognitive Architecture. *New Ideas in Psychology*, **30**, 227-240. https://doi.org/10.1016/j.newideapsych.2011.11.003

[10] Prokhorov, D.V. and Wunsch, D.C. (1997) Adaptive Critic Designs. *IEEE Transactions on Neural Networks*, **8**, 997-1007. https://doi.org/10.1109/72.623201

[11] Werbos, P.J. (1992) Neural Control and Supervised Learning: An Overview and Evaluation, Handbook of Intelligent Control. Van Nostrand, New York.