用于交通信号灯控制的特征表示近似 Q 学习

李昱朔[1,2] 姚明海[2]

(浙江师范大学数理信息工程学院 浙江 金华 321004)[1] (浙江工业大学信息工程学院 杭州 310000)[2]

摘 要 强化学习通过与环境的交互来学习行为策略。强化学习方法是在线的增量学习,易于实现。文中提出了基于函数近似的强化学习算法,并将其用于自适应交通信号灯控制。基于表格的强化学习需要完全的状态表征,随着车道数和路口数的增加,计算复杂度呈指数增长,即使中小规模的交通网络也很难实现,从而不能应用于实际的交通信号灯控制。因此文中使用基于特征的状态表征来有效地解决维数灾难问题;通过简便的方法获取车流的拥塞等级以及红灯的时长,使用函数近似定义 Q 值,进而实现高效的自适应控制。在 GLD 上的仿真实验结果验证了该自适应控制方法的有效性和可行性。

关键词 自适应交通灯控制,强化学习,Q 学习

中图法分类号 TP181 **文献标识码** A

Q-learning with Feature-based Approximation for Traffic Light Control

LI Min-shuo[1,2] YAO Ming-hai[2]

(College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, Jinhua, Zhejiang 321004, China)[1]

(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310000, China)[2]

Abstract Reinforcement learning(RL) learns the policy through interaction with the environment. RL algorithms are online, incremental, and easy to implement. This paper proposed a Q-learning algorithm with function approximation for adaptive traffic light control (TLC). The application of table-based Q-learning to traffic signal control requires full-state representations and cannot be implemented, even in moderate-sized road networks, because the computational complexity exponentially grows in the numbers of lanes and junctions. This paper tackled the dimension disaster problem by effectively using feature-based state representations and used a broad characterization of the levels of congestion. The experiment results show that the proposed method is effective and feasible.

Keywords Adaptive traffic light control, Reinforcement learning, Q-learning

1 引言

随着城市交通量的不断增大,以及道路基础设施存在的局限性,要提高交通系统车流通过量,必然要考虑交叉口红绿灯的智能设计。由于面临的诸多难点,自适应交通控制仍是科研和工程难题,其中主要的两个难点是建模和优化。

首先,交通系统通常是分布式的多层复杂系统。如何既准确又便利地描述交通系统仍是一个未完全求解的问题。目前绝大多数控制系统通过预估交通系统的未来状态来预先给出合适的信号配置,这也加重了交通系统建模问题的重要性和难度。目前主要有两种方法来解决这个难题。一种是基于流模型的方法,其构建分析模型来描述在不同位置测量的宏观交通流的动态性[1],如 CTM(Cell Transmission Models)细胞传输模型,其由于简单且有效,常被考虑使用。然而,当交通场景比较复杂时,其建模成本和误差有待商酌。另一种是基于模拟的方法,使用人工智能以及模拟来预估未来交通状态[2]。通过交通流测量的历史记录,人工智能模块学习并重建宏观动态性;而模拟部分描述和重构微观个体的行为,为更好地描述宏观动态性增加了灵活性。然而,人工智能模块学习和模拟都很耗时,并且由于这些方法没有直接可用的理论分析工具,性能改善非常困难。

其次,如果已经建立了交通流的描述,如何确定最优的交通信号灯控制是另一个难题[3-4]。对于那些基于流模型的方法,可以使用数学的规划方法来求解给定的目标函数,同时要求满足从分析模型中提取约束条件。而对于那些基于人工智能和模拟的方法,可以基于所学的关于控制行为和行为效果的关系做因果逆向处理,然后使用尝试和测试方法,基于预估或模拟的行为效果来寻找最优或次优信号控制。

世界各地的道路交叉口使用的是固定信号计时,即按周期循环的固定方式配置信号灯。虽然这种策略很容易实现,但它没有考虑交通流的动态性。强化学习(Reinforcement learning,RL)[5-7]是通过试错学习的方式与环境交互学习,Agent 对环境采取行为,环境反馈给 Agent 奖赏。如果行为执行后,环境反馈给 Agent 的奖励大,那么 Agent 下次选择该行为的概率就大,反之,则小。依照此学习方式,Agent 具有自寻优能力。将这种学习方式应用在交通信号自适应控制上是非常合理的。强化学习方法通过学习信号灯控制行为以及行为导致的交通流的变化来实现对复杂交通系统的内隐建

李昱朔(1974—),女,硕士,讲师,主要研究方向为机器学习,E-mail:lmshappy@zjnu.cn(通信作者);**姚明海**(1964—),男,博士,教授,主要研究方向为认知学习、机器学习。

模。其从学习过的输入-输出对中寻找最优或次优交通信号灯配置[8-9]。交通信号灯控制的主要难点在于:随着交通流状态和控制行为数目的增加,交通信号灯控制的复杂度呈指数增长。我们采用了基于特征表示的函数近似方法来表征值函数,通过合理设置参数,对交通道路的特征进行等级划分,从而大大降低了状态空间的维度。

2 交通信号灯控制问题表示

RL 算法中 Q-learning 的概念最早由 Watkins 于 1989 年提出来[5],随后 Watkins 和 Dayan 于 1992 年对其进行了收敛性证明[10]。Q-learning 算法是一种模型无关 RL 技术,即不需要环境的精确模型。Q-learning 算法通过与环境交互,采取试错学习的方式,学习一个状态-动作值函数(Q 值),每一步对给定的动作给予一个期望回报值(奖惩)。最优策略就是选择能使 Q 值达到最值的动作。

将 Q 学习应用于交通信号灯的控制,首先要求给出状态、行为和奖惩函数的定义。我们假设道路网络有 $m(m\geqslant 1)$ 个交叉口。每一个交叉口有多个路口,每条路口有多个车道。

1)状态

状态使用队列长度和延续时间向量来表示。某车道的延续时间是从该车道的信号灯变红开始计时。对于绿灯车道,延续时间为零。一个中央控制器从各个车道上的传感器接收状态信息,然后决定一个周期中哪个交通信号灯为绿。然后决策被传送回到各个交叉口。为简单起见,我们假设没有传播和反馈延迟。车道的延续时间计数器在绿灯信号下为零,信号变红时开始计时。如果交通网络共有 N 条车道,则在 t 时刻的状态是:

$$s_n=(q_1(n),\cdots,q_N(n),t_1(n),\cdots,t_N(n))^{\mathrm{T}}$$

其中,$q_i(n)$是车道 i 在时刻 n 的队列长度,$t_i(n)$是车道 i 在时刻 n 红灯信号的延续时长。

2)行为

行为 a_n 是这个交通网络 m 个交叉口的信号配置组合(是一个可行的交通信号灯组合)。形式为 $a_n=(a_1(n),\cdots,q_m(n))^{\mathrm{T}}$。其中 $a_i(n)$是交叉口 i 在时刻 t 的信号配置。我们只考虑可行的信号配置操作集,而不是所有可能的交通信号灯红绿组合(其随信号灯数量的增加呈指数增长)。因此,行动集 $A(s_n)=\{$在状态 S_n 可行的信号配置$\}$。

3)奖惩函数

这里的奖惩函数考虑两个组元。第一个组元是各个车道队列长度的总和,第二个组元是各个车道红灯延续时间的总和。我们的想法是调节交通流量减少队列长度,且为确保公平,不会使某车道的红灯延续时间很长;同时考虑了主车道要比其他车道有更高的优先级。本文通过在成本函数中为不同车道设置不同的权重,来实现主要道路交通的优先级。假设 L_p 表示应该给予更高优先级的车道的下标集合。阶段成本函数 $r(s,a)$定义为:

$$r(s_n,a_n)=r_1\times(r_2\sum_{i\in L_p}q_i(n)+s_2\sum_{i\notin L_p}q_i(n))+s_1\times(r_2\sum_{i\in L_p}t_i(n)+s_2\sum_{i\notin L_p}t_i(n)) \tag{1}$$

其中,$r_i,s_i\geqslant 0$,$r_i+s_i=1$,$i=1,2$,且 $r_2>s_2$。这样,在 L_p 中的车道就被赋予了更高的成本,从而追求成本最优化的策略一定会为这些车道分配一个更高的优先级来最小化成本函数。

正如上文提到的,我们考虑的是无限水平折扣成本框架。折扣因子 γ 扮演着至关重要的作用。如果使用较小的 γ 对未来成本进行折扣,则对未来成本就给予较少重视。在实验中设置 $\gamma=0.9$。

3 基于函数近似的 Q 学习

传统的 Q 学习算法通过增量随机算法来获取最优配置策略。然而,这种方法需要一个 Q-表来存储所有可能的状态-行为对的 Q 值。虽然这种算法在小的状态和行动空间是有用的,但在考虑较大的交通网络比如网格或有多个交叉口的干道时,情况就会恶化,原因是如果状态和行为空间急剧增大,就会面临维数灾难问题。正是这一点使 Q 学习算法在中型和大型的交通网络上都不可实现。为了缓解维数灾难问题,本文将采用基于特征表示的函数近似方法。这些方法处理上述问题的思想是使计算复杂度可控。下面首先介绍基于特征的方法,并给出使用函数近似 Q 学习的交通信号灯控制方法。

3.1 基于特征表示的函数近似

Q 学习算法可以采用基于函数的近似,其思想是把 Q 值近似表示为:

$$Q\approx\sum_{i=1}^{d}(\sigma_{s,a}(i))^{\mathrm{T}}\theta_i,s\in S,a\in A(s) \tag{2}$$

其中,$\sigma_{s,a}$是 d 维的特征向量,用于描述状态-行为对(s,a)。相比于所有可能的状态-行为对(s,a)集合的尺寸,维数 d 大幅度降低。θ 是可调参数,其维数大小与 d 相同。

3.2 基于函数近似的 Q 学习

本节给出基于函数近似的交通信号灯控制 Q 学习算法。虽然表格型 Q 学习算法需要完整的状态信息,计算效率较低,但其变体如果使用函数近似,对 Q 值函数参数化,那么将大大减少计算的时间和空间,从而提供良好的性能。函数近似 Q 学习是表格型 Q 学习的改进版,其修改的是参数 θ,维度大小为 d。现在待求解的系统只有 d 个变量,而不是原来求解的有$|S\times A(S)|$个变量的系统。

假设 s_n,s_{n+1}分别表示在 n 和 $n+1$ 时刻的状态,θ_n 是对参数 θ 的第 n 次修改。算法使用如下的更新规则:

$$\theta_{n+1}=\theta_n+\alpha(n)\sigma_{s_n,a_n}\times(r(s_n,a_n)+\gamma\min_{a\in A(s_{n+1})}\theta_n^{\mathrm{T}}\sigma_{s_{n+1},a_n}-\theta_n^{\mathrm{T}}\sigma_{s_n,a_n}) \tag{3}$$

虽然算法修改的是参数 θ,但最终是根据 $Q(s,a)\approx\theta^{\mathrm{T}}\sigma_{s,a}$ 修改 Q 值函数,本质上是原始 Q 函数集的子空间。

基于每个车道等待队列的长度及红灯延续时长定义特征,我们选择的特征表示如下:

$$\sigma_{s_n,a_n}=(\sigma_{q_1(n)},\cdots,\sigma_{q_N(n)},\sigma_{t_1(n)},\cdots,\sigma_{t_N(n)},\sigma_{a_1(n)},\cdots,\sigma_{a_m(n)})^{\mathrm{T}}$$

其中:

$$\sigma_{q_i(n)}=\begin{cases}0, & \text{if } q_i(n)<L_1\\ 0.5, & \text{if } L_1\leqslant q_i(n)\leqslant L_2\\ 1, & \text{if } q_i(n)>L_2\end{cases}$$

$$\sigma_{t_i(n)}=\begin{cases}0, & \text{if } t_i(n)\leqslant T_1\\ 1, & \text{if } t_i(n)>T_1\end{cases}$$

特征值对应 m 个交叉口的行为或信号设置。N 是交通网络中所有交叉口的车道总数。L_1 和 L_2 是队长的阈值,T_1

是红灯延续时长的阈值，参数 θ 的维数与特征 σ 的维数相同。该算法的优势是不需要为每个(s,a)对修改 Q 值，而是使用参数给出 Q 值的估计。对于固定的策略，函数近似的 Q 学习算法与 TD 算法是相似的，关于 TD 算法的有条件收敛性证明见文献[5]。

使用前面选择的特征近似表示 Q 值有很多优势。首先，不需要获取关于队长和延续时长的精确信息。使用车道上的 L_1 和 L_2 标志，可以估计交通流的低、中、高 3 个等级。同样，红灯延续时间也分为低于阈值 T_1 或高于阈值 T_1 两个等级。近似算法不需要精确的队长信息，这类信息常常难以获取。而在某时刻关于交通拥塞程度的特征(即低、中、高)是很容易得到的。

在实际应用中，因为需要的信息是用于判断交通拥塞的程度，所以只需在车道上放置两组传感器，一组对应 L_1，一组对应 L_2。如果 L_1 传感器没有检测到拥塞，可以推断是低等级；如果 L_1 传感器检测到拥塞，L_2 传感器没有检测到拥塞，可以推断是中等级；如果 L_2 传感器检测到拥塞，可以推断是高等级。延续时长的测量通常依赖于交叉口的计时器，同样也只需要给出等级信息，即高于或低于阈值 T_1。这些信息必须传给中央控制器，通过运行 TLC 算法得到信号配置。TLC 算法描述如算法 1 所示。

算法 1　TLC 算法

Initialize：

$\theta_i \leftarrow 0, \forall$ i；主干路集合 L_p；ε；迭代次数 iteration；

Loop：

n←0；

1. 获得当前状态 s 的特征表示 $\sigma_{q_i(n)}$ 和 $\sigma_{t_1(n)}$；

2. 按式(2)计算 Q 值，根据 ε-greedy 策略生成行为 a_n：

2.1 选择随机行为 a_{ran}；

2.2 选择使 Q 值最小的行为 a_{min}；

2.3 生成随机数 rand∈[0,1]；

2.4 if(rand＜ε) then $a_n \leftarrow a_{ran}$ else $a_n \leftarrow a_{min}$；

3. 执行 a_n，然后获取下一状态 s′的特征；

4. 根据式(1)计算回报 r；

5. 根据式(3)修改 θ；

n←n+1；

UNTIL n ＞ iteration

4　实验结果与分析

实验环境为开源仿真器 Green Light District(GLD)。仿真分两步进行：首先设置交通模拟参数和输入交通流，然后对每种算法在相同条件下运行并测试。在 GLD 实验平台上分别比较了固定配时方案(Fixed-time)、最长队列优先配时(Longest-queue)、神经网络(Neural-network)和近似 Q-learning 的方法，实验中选取的目标优化参数包括通过交叉口的车辆数、车辆平均等待时间、平均车辆排队长度。

在 GLD 仿真平台上对不同边缘节点设置不同车流量产生频率，设置参数如图 1 所示。交通流量的每分钟采集 1 次，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，以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。为了保证模拟测试的一致性，研究中设计东西南北直行车流都是相同的，东西南北左转车流量也遵循相同的变化，并且设置直行车流量大于左转车流量。模拟城市区域交通道路具体交通流量的设置如表 1 所列。

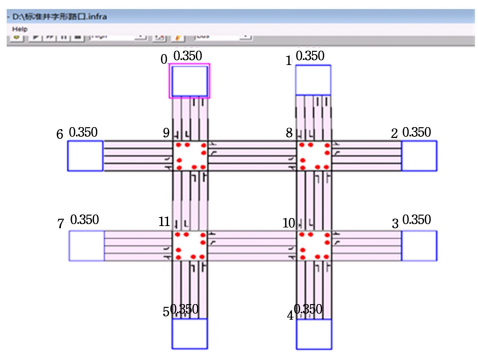

图 1　GLD 仿真平台中的交通网络图

表 1　各相位方向的交通流量设置

车流方向	时间段分布										平均车流量
	1	2	3	4	5	6	7	8	9	10	
东直行	1000	888	777	666	555	555	666	777	888	1000	777.2
东左转	300	250	200	150	100	100	150	200	250	300	200
西直行	1000	888	777	666	555	555	666	777	888	1000	777.2
西左转	300	250	200	150	100	100	150	200	250	300	200
南直行	1000	888	777	666	555	555	666	777	888	1000	777.2
南左转	300	250	200	150	100	100	150	200	250	300	200
北直行	1000	888	777	666	555	555	666	777	888	1000	777.2
北左转	300	250	200	150	100	100	150	200	250	300	200

在 GLD 仿真实验中，以 2000 个迭代周期为标准，分别对平均路口等待时间、通过交叉口车辆数、平均车辆排队长度进行测试，不同算法的实验结果如表 2 所列。

表 2　迭代周期为 2000 的评价参数比较

算法	周期	平均路口等待时间	通过交叉口车辆数	平均车辆排队长度
Fixed-time	2000	88.04	24710	56.4
Longest-queue	2000	79.14	27010	57.6
Neural-network	2000	92.17	27450	49.8
Q-learning	2000	44.63	29690	42.3

实验结果表明，在周期相同的情况下，城市 TLC 系统采用本文提出的特征表示近似 Q-learning 算法，能有效降低车辆排队长度和车辆在路口的等待时间，交叉口通过的车辆数也有了较大的增长。Q-learning 考虑了各个车道的拥塞状况，具备较强的适应性，缩短了平均车辆排队长度。Q-learning 同时考虑了红灯的时长，保证每个车道的等待时间不会太长，从而减少了平均路口等待时间。两者的结合使通过的车辆数有了明显的提高，从而提高了整个区域道路的通行能力。

(下转第 159 页)

ceedings of the 30th International Conference on Machine Learning (ICML-13). 2013:1139-1147.

[4] NAIR V, HINTON G E. Rectified Linear Units Improve Restricted Boltzmann Machines[C]// International Conference on Machine Learning. Omnipress, 2010:807-814.

[5] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[J]. Journal of Machine Learning Research, 2010, 9:249-256.

[6] DAHL G E, SAINATH T N, HINTON G E. Improving deep neural networks for LVCSR using rectified linear units and dropout[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013:8609-8613.

[7] GLOROT X, BORDES A, BENGIO Y. Deep Sparse Rectifier Neural Networks[J]. Journal of Machine Learning Research, 2010, 15:315-323.

[8] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer Science, 2012, 3(4):212-223.

[9] LE Q V, JIQUAN N, ADAM C, et al. On optimization methods for deep learning[C]// Proceedings of the 28th International Conference on International Conference on Machine Learning. 2011:265-272.

[10] DUCHI J, HAZAN E, SINGER Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization[J]. Journal of Machine Learning Research, 2010, 12(7):257-269.

[11] ZEILER M D. ADADELTA: An Adaptive Learning Rate Method[J]. arXiv: 1212.5701.

[12] TIELEMANT, HINTON G. RMSProp: Divide the gradient by a running average of its recent magnitude[R]. COURSERA: Neural Networks for Machine Learning. 2012.

[13] KINGMA D, BA J. Adam: A Method for Stochastic Optimization[J]. arXiv: 1412.6980.

[14] NESTEROV Y. A method of solving a convex programming problem with convergence rate O(1/k2)[J]. Soviet Mathematics Doklady, 1983(27):372-376.

[15] NESTEROV Y. Introductory lectures on convex optimization: a basic course[M]. Applied Optimization Kluwer Academic, 1998: 119-120.

[16] GOODFELLOW I, BENGIO Y, COURVILLE A. 深度学习[M]. 符天凡等，译. 北京：人民邮电出版社，2017:240-249.

[17] FLETCHER R, POWELL M J D. A Rapidly Convergent Descent Method for Minimization[J]. Computer Journal, 1963, 6(6):163-168.

[18] MILLER S J. The Method of Least Squares[D]. Brown University, 2006:1-7.

[19] 朱志鹏，喻芳，曾青霞，等. 基于深度学习与偏最小二乘的分析方法及其医学应用[J]. 江西中医药大学报，2017，29(3)：94-97.

[20] SEBER G A F, LEE A J. Linear regression analysis[M]. Wiley, 2012.

[21] CRISTIANINI N, SHAWER-TAYLOR J. An introduction to support vector machines and other kernel-based learning methods[M]. Cambridge University Press, 2000.

[22] KENNY V, NATHAL M, SALDANA S. Heuristic algorithms [C]// ChE 345. Spring, 2014.

[23] EISELT H A, SANDBLOM C L. Heuristic Algorithms[C]// Integer Programming and Network Models. 2000:229-258.

(上接第 145 页)

结束语　基于无线传感网络设计道路交通管理系统的最大挑战是实现最优的信号灯控制以尽可能提高交通率并最小化道路拥塞。强化学习给出了求解该问题的有效方法。本文为交通信号灯的自适应控制设计了基于特征表示的 Q 学习算法，Q 学习不需要构建系统模型，能实时地适应交通状况。基于特征表示的函数近似可以有效地控制大规模交通网络的计算复杂度。仿真结果表明本文提出的 Q-learning 交通信号灯控制方法优于其他几个常用的算法。

未来的工作是在实际交通环境中进行实验，验证本文算法的优势；同时综合考虑多项信号控制指标，使得奖惩函数的设计符合实际的道路状况。

参 考 文 献

[1] ADAM I, WAHAB A, YAAKOP M, et al. Adaptive fuzzy logic traffic light management system[C]// 2014 4th International Conference on Engineering Technology and Technopreneuship (ICE2T). IEEE, 2014:340-343.

[2] COOLS S B, GERSHENSON C, D'HOOGHE B. Self-Organizing Traffic Lights: A Realistic Simulation[J]. Advances in Applied Self-Organizing Systems, 2016, 17(4):45-55.

[3] KAUR T, AGRAWAL S. Adaptive Traffic Lights Based on Hybrid of Neural Network and Genetic Algorithm for Reduced Traffic Congestion[C]// Recent Advances in Engineering and Computational Sciences (RAECS). 2014:1-5.

[4] SRINIVASAN D, CHOY M C, CHEU R L. Neural Networks for Real-Time Traffic Signal Control[J]. IEEE Transactions on Intelligent Transportation Systems, 2006, 7(3):261-272.

[5] SUTTON R S, BARTO A G. Introduction to reinforcement learning [J]. IEEE Transactions on Neural Networks, 1992, 8 (3/4):225-227.

[6] 高阳，陈世福，陆鑫. 强化学习研究综述[J]. 自动化学报，2004，30(1)：86-100.

[7] 刘忠，李海红，刘全. 强化学习算法研究[J]. 计算机工程与设计，2008，29(22)：5805-5809.

[8] SALKHAM A, CUNNINGHAM R, GARG A, et al. A Collaborative Reinforcement Learning Approach to Urban Traffic Control Optimization[C]. IEEE/WIC/ACM International Conferent on Web Intelligence and Intelligent Agent Technology. 2008: 560-566.

[9] XIE Y C. Development and evaluation of an arterial adaptive traffic signal control system using reinforcement learning[OL]. http://holl.hardle.net/1969.1/ETD-TAMU-2480.

[10] WATKINS C, DAYAN P. Q-learning [J]. Machine Learning, 1992, 8(3/4):279-292.