多智能体强化学习在城市交通网络信号控制方法中的应用综述*

杨文臣[1,2]，张 轮[2†]，Zhu Feng[3]

(1. 云南省交通规划设计研究院 陆地交通气象灾害防治技术国家工程实验室，昆明 650031；2. 同济大学 道路与交通工程教育部重点实验室，上海 201804；3. 南洋理工大学 土木与环境工程学院，新加坡 639798)

摘 要：交通信号控制系统在物理位置和控制逻辑上分散于动态变化的网络交通环境，将每个路口的交通信号控制器看做一个异质的智能体，非常适合采用无模型、自学习、数据驱动的多智能体强化学习(MARL)方法建模与描述。为了研究该方法的现状、存在问题及发展前景，系统跟踪了多智能体强化学习在国内外交通控制领域的具体应用，包括交通信号 MARL 控制概念模型、完全孤立的多智能体强化学习(MARL)的控制、部分状态合作的多智能体强化学习控制和动作联动的多智能体强化学习(MARL)控制，分析其技术特征和代际差异，讨论了多智体强化学习方法在交通信号控制中的研究动向，提出了发展网络交通信号多智能体强化学习集成控制的关键问题在于强化学习控制机理、联动协调性、交通状态特征抽取和多模式整合控制。

关键词：智能交通；交通控制；多智能体强化学习；闭环反馈；联动协调；数据驱动

中图分类号：TP181；U491.51 **文献标志码**：A **文章编号**：1001-3695(2018)06-1613-06

doi:10.3969/j.issn.1001-3695.2018.06.003

Multi-agent reinforcement learning based traffic signal control for integrated urban network: survey of state of art

Yang Wenchen[1,2], Zhang Lun[2†], Zhu Feng[3]

(1. *National Engineering Laboratory for Surface Transportation Weather Impacts Prevention, Broadvision Engineering Consultants, Kunming* 650031, *China*; 2. *Key Laboratory of Road & Traffic Engineering for Ministry of Education, Tongji University, Shanghai* 201804, *China*; 3. *School of Civil & Environmental Engineering, Nanyang Technological University, Singapore* 639798, *Singapore*)

Abstract: Urban traffic control (UTC) systems, geographical and logical distribution in dynamic changing traffic environments, are well suited for multi-agent reinforcement learning (MARL) approach because of their model free, self-learning, and data-driven features. To investigate the state-of-the-art, this paper comprehensively surveyed main chanllenges and recent trends, the MARL methods and techniques applied to UTC systems, including general framework of MARL for UTC, totally independent MARL, partially state-cooperation MARL, and joint-action MARL. By comparing key characteristics and differences of the leading MARL approaches, it discussed several future directions toward the successful deployment of MARL technology in traffic control systems, and addressed four critical issues in developing agent-based traffic control systems for integrated network as mechanism of RL traffic signal control, joint-action coordination, feature partitioning of traffic state and multi-model integrated control.

Key words: intelligent transportation; traffic control; multi-agent reinforcement learning(MARL); closed feedback; joint-coordinated cooperation; data driven

0 引言

计算机、通信和交通检测技术的变革式发展，促使城市交通信号控制系统的技术环境正从数据贫乏向数据丰富的时代演化发展[1]，数据驱动的区域交通控制被提出并在近年来取得了长足发展[2,3]。自 Thorpe[4] 于 1997 年首次将强化学习(reinforcement learning，RL)方法应用于交通信号最优化控制以来，多智能体强化学习(multi-agent reinforcement learning，MARL)在区域交通信号自适应控制领域迅速发展并已有实际应用[5~8]。从控制理论来看，MARL 控制可根据控制效果的反馈信息自主学习并优化策略知识，是一种真正的闭环反馈控制[9]；从控制范围来看，其可精确推理多个路口间的最优联合动作，丰富了区域交通协调控制的内容及形式[10]；从控制实时性来看，它没有复杂的模型优化模块，采用秒级的即时决策，可实时响应时变交通流的变化[11]；从系统可拓展性来看，分散式 MARL 控制具有统一的结构模型，可针对特定路网结构和交通流特性进行相应改造[5]；从系统兼容性来看，MARL 控制本身仅需要系统的输入和输出数据，对数据具体采集的技术和形式无要求[7]。

作为一种无模型、自学习的迭代型数据驱动方法，MARL 为实现闭环反馈的自适应控制提供了一种内涵式的解决方法。本文系统回顾了现有 MARL 方法在城市道路交通网络信号控制中的研究和应用，探讨了将 MARL 应用于大规模区域交通

收稿日期：2017-06-10；**修回日期**：2017-07-24 **基金项目**：云南省交通厅科技计划资助项目(云交科 2014(A)23)；国家“863”计划资助项目(2012AA112307)

作者简介：杨文臣(1985-)，男，云南昌宁人，博士，主要研究方向为智能交通控制、交通安全；张轮(1972-)，男(通信作者)，教授，博导，主要研究方向为智能交通运输系统(Lun_zhang@tongji.edu.cn)；Zhu Feng(1986-)，男，助理教授，博导，主要研究方向为智能交通控制、交通网络建模与优化.

控制的关键问题。

1 交通信号 MARL 控制基本概念

1.1 RL 控制标准模型

交通信号 RL 智能体的标准模型如图 1 所示。每个路口的交通信号机被抽象为一个智能体,控制对象为道路交通网络上的时变交通流。RL 智能体与被控对象在闭环系统中不断进行交互,通过观察交通环境的实时状态,提取信号控制所需的交通状态信息和反馈奖励信息,选择相应的行为动作并执行;进而跟踪评测所选择动作的控制效果,以累积回报收益最大化为目标,优化控制策略直至收敛到状态和动作的最优概率映射。因此,RL 智能体将控制系统的优化过程按照时间进程划分为状态相互联系的多个阶段,并在每个阶段根据当前状态进行最优决策,这是典型的马尔可夫决策过程(Markov decision process,MDP)[5]。

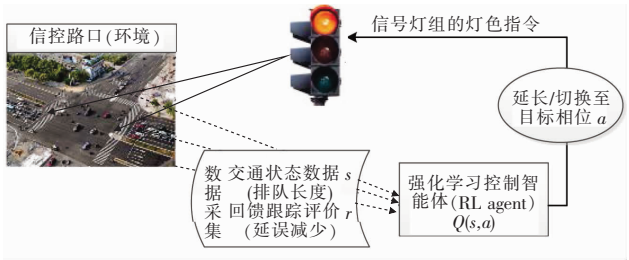

图 1 交通信号 RL 智能体的标准模型

相较于动态规划(dynamic programming)求解 MDP 需要系统状态转移概率和反馈函数模型,RL 将学习看做是试错过程,根据智能体自身的采样探索(状态 s-行动 a 评价 r),采用 RL 学习算法改进控制策略知识以适应随机变化的环境。典型的 Q 学习算法[12]如式(1)所示,智能体按照使各状态下行为动作获得的总累积回报值达到最大的原则选择动作。

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

其中:$Q(s,a)$ 为交通状态 s 下采取信号动作 a 的累积回报函数值,简称 Q 值;α 为学习率;γ 为折扣率;r_{t+1} 为 t 时刻在交通状态 s 下采取信号动作 a 的即时回报奖励值。

1.2 RL 控制优化技术

根据 RL 智能体学习频率及优化参数的不同,交通信号 RL 优化技术分为周期式和非周期式控制(cyclic or acyclic)两种类型。其主要技术特征如表 1 所示。

表 1 RL 控制优化技术的特征

类别	更新频率	优化参数	交通状态	信号动作	协调方式
周期式 RL 控制[13~15]	整周期	周期、绿信比、相位差	排队长度、相位持续时长等	参数调整的权值等	相位差
非周期式 RL 控制[10,16,17]	单位延长时间(1~3 s)	相位结构、相位顺序、相位时长	排队长度、相位持续时长等	相位编号	状态共享或动作联动

夏新海[16]和 Salkham 等人[18,19]提出了固定周期式 RL 控制方法。如图 2 所示,在相位结构和相位顺序固定的前提下,周期式 RL 控制以周期、绿信比和相位差作为控制方案的配时参数,每隔当前周期的整数倍时间间隔,采用 RL 算法对这些参数进行优化调整,以响应路口交通需求波动。这种优化技术的控制方案结构固定,配时参数更新具有滞后性,并通过相位差技术实现走廊方向的信号协调,是一种响应式(responsive)自适应交通控制。

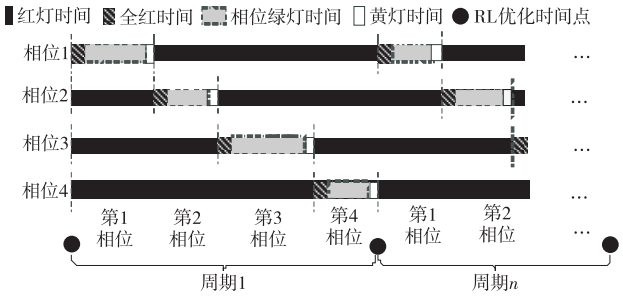

图2 固定周期式RL控制方法的相位时序示意图

Thorpe[4]和 El-Tantawy 等人[9,10]研究了非周期式 RL 控制方法。如图 3 所示,非周期式 RL 控制遵循感应信号控制的逻辑框架,在满足交通控制基本约束的前提下,根据时变交通流的波动,每隔单位延长时间,采用 RL 算法对相位结构、相位顺序或绿灯时长进行优化,以实时响应交通需求的变化。这种优化技术摒弃了传统协调控制中周期和相位差的概念,由实际交通流即时决策相位方案及相位时长,并通过多个路口信号灯的联动实现区域交通协调控制,以尽可能保证车队连续通行,是一种实时(real-time)自适应交通控制。

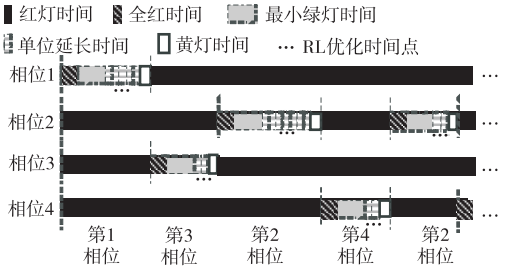

图3 非周期式RL控制方法的相位时序示意图

1.3 交通网络 MARL 控制

交通网络 MARL 控制是单路口 RL 控制向随机博弈(stochastic game,SG)环境下区域交通网络的扩展,以期通过多个路口 RL 智能体间的联动协调,逼近网络交通流的最优均衡策略[16]。由于区域内全部 RL 智能体同时学习和同时决策,每一个 RL 智能体都面临移动目标学习问题(moving-target learning problem)[5],即本地智能体的最优策略将随着区域内其他智能体策略的变化而变化。许多学者通过构建多路口间交通信号的联动协调机制,采用基于协调的 MARL 进行系统的分散决策与优化。根据智能体间交通状态和信号动作的协调水平,交通网络 MARL 控制可分为三类[7]:完全独立的多智能体强化学习控制(totally independent MARL)、部分状态合作的多智能体强化学习控制(partially state cooperation MARL)和动作联动的多智能体强化学习控制(joint-action MARL)。

2 交通网络 MARL 控制方法

2.1 完全孤立的 MARL 控制

此方法假设路口处于静态随机的交通环境,即每个 RL 智能体的决策仅受路口本地状态和本地动作的影响,只需通过在式(1)的更新规则中增加智能体 i 的索引下标,将单智能体 RL 控制方法直接拓展并应用到多个路口即可,其基本形式[7,12]为

$$\begin{aligned}Q_i^k(s_i^k,a_i^k)&=Q_i^{k-1}(s_i^k,a_i^k)+\alpha^k[r_i^k(s_i^k,a_i^k,s_i^{k+1})+\\&\gamma\max_{a_i^{k+1}\in A_i}Q_i^{k-1}(s_i^{k+1},a_i^{k+1})-Q_i^{k-1}(s_i^k,a_i^k)]\\a_i^{k+1}&=\arg\max_{a_i^{k+1}\in A_i}Q_i^{k-1}(s_i^{k+1},a_i^{k+1})\end{aligned} \quad (2)$$

其中:k 为时间步;A_i 为智能体 i 的有限动作集合。既有研究中完全孤立的 MARL 控制技术特征如表 2 所示。

表 2　完全孤立的 MARL 控制技术特征

代表研究	学习算法	RL 控制器要素			选择策略	动作频率	路网规模	协调水平
		交通状态定义	信号动作定义	反馈函数				
Thorpe[4] (1997)	SARSA(λ)	车道级:到达车辆数,车辆与停车线间距、相位时长	相序固定	路网交通量的释放时间	ε-greedy	1 s	16(假设)	无
Wiering[20] (2000)	模型 Q 学习	车辆级:车辆排队位置及其等待时间	相序可变	相邻决策点之间的总延误	ε-greedy	1 s	6(假设)	无
Abdulhai 等人[9] (2003)	Q 学习	车道级:排队长度	相序可变	相邻决策点之间的总延误	softmax	1 s	1(假设)	无
EI-Tantawy 等人[13] (2011)	Q 学习、SARSA SARSA(λ)	车道级:排队长度,绿灯到达车辆数与红灯排队长队,累积延误	相序固定 相序可变	延误、累积延误、排队车辆数、停车次数	ε-greedy、softmax、ε-softmax	1 s	1(真实)	无
Jin 等人[14] (2015)	Q 学习、SARSA	车道级:车队到达时间间隔、车道占有率、相位时长	单位延长时间	相邻决策点总延误的差值	ε-greedy、softmax	1~4 s	1(假设)	无
马寿峰等人[15] (2002)	Q 学习	相位集:当前相位 ID、相位时长、绿灯到达车辆数等	相序固定	综合考虑绿灯通过量和红灯增加的排队	softmax	5 s	1(假设)	无

Thorpe[4] 首次将 SARSA 算法应用于交通信号控制,研究发现交通状态的定义是 SARSA 交通控制是否适用的关键。Wiering[20] 将基于模型的 RL 方法应用于小路网交通信号控制,提出了三种 RL 控制系统。仿真结果表明提出的 RL 控制系统均优于非自适应控制系统,综合考虑全部车辆状态转移 RL 控制的性能最好,但该方法存在交通状态数量呈指数级增长、每个路网节点需布设检测器、需实时计算状态转移概率等固有不足,使其难以应用到实际工程中。Abdulhai 等人[9] 将无模型 Q 学习应用于单路口交通信号控制,采用 Q 学习算法来智能决定相位切换时间及顺序。Oliveira 等人[21] 假定交通情景独立于智能体动作,提出了一种基于情景感知的强化学习算法 RL-CD,并将其应用于单路口交通信号控制。EI-Tantawy 等人[13] 具体定义了 RL 智能体五类核心要素的 15 种模型表征,采用 Paramics 交通仿真定量分析交通控制与强化学习各要素间的交互作用。研究发现科学设计 RL 模型要素是交通信号自学习控制的关键。Jin 等人[14] 提出了基于信号灯组的交通信号强化学习控制方法,每个流向的信号灯组为一个 RL 智能体,并采用 Q 学习算法根据交通流状态自动优化相位结构及顺序。

整体而言,对完全孤立的 MARL 控制,RL 智能体采用 Q、SARSA 或 R 学习等时间差分 RL 算法进行本地最优控制,不同学者分别定义其 RL 智能体五类结构要素。由于区域内多个 RL 智能体间同时学习和同时行动,而智能体间没有协调机制,每个智能体都面临着移动目标学习和唯一均衡对策的难题[5,7],系统无法收敛到均衡的联合策略,且孤立路口的交通状态不能有效描述过饱和的交通网络[8]。在其他研究方面,为平衡不同交通状态下控制收益和出行损失,有学者提出了多目标反馈的 RL 控制方法[22~24],这其中如何标准化设计多目标反馈函数是关键。为解决传统表格型 RL 控制系统的维度灾难(curse of dimensionality),自适应动态规划方法(approximate dynamic programming, ADP)[25] 被应用于交通信号控制。根据值函数逼近方式的不同,ADP 交通控制方法可分为基于智能计算的非线性逼近[26] 和带可调参数的数学逼近[27]。然而 RL、神经网络、逼近函数等多种模型或算法的深度耦合可能致使混合 MARL 控制模型的可解释性差。

2.2　部分状态合作的 MARL 控制

部分状态合作的 MARL 控制通过智能体间的点对点通信,获得上/下游路口的交通数据,并以此拓展本地 RL 智能体的交通状态感知空间,构造了部分状态联合的 Q 值函数,提高其对动态随机环境的观察能力[7,12]。其基本形式如式(3)所示。

$$\begin{aligned} Q_i^k(\boldsymbol{s}_i^k,a_i^k) &= Q_i^{k-1}(\boldsymbol{s}_i^k,a_i^k)+\alpha^k[r_i^k(\boldsymbol{s}_i^k,a_i^k,\boldsymbol{s}_i^{k+1})+ \\ &\quad \gamma \max_{a_i^{k+1}\in A_i} Q_i^{k-1}(\boldsymbol{s}_i^{k+1},a_i^{k+1})-Q_i^{k-1}(\boldsymbol{s}_i^k,a_i^k)] \\ a_i^{k+1} &= \arg\max_{a_i^{k+1}\in A_i} Q_i^{k-1}(\boldsymbol{s}_i^{k+1},a_i^{k+1}) \end{aligned} \tag{3}$$

其中:$\boldsymbol{s}^k$ 为智能体 i 的联合状态向量,包括本地和邻近路口交通状态信息。既有研究中部分状态合作的 MARL 控制技术特征如表 3 所示。

表 3　部分状态合作的 MARL 控制技术特征

代表研究	学习算法	RL 控制器要素			选择策略	动作频率	路网规模	协调水平
		交通状态定义	信号动作定义	反馈函数				
Richter 等人[28] (2006)	改进 Q 学习	相位级:本地配时信息、到达车辆数、邻近交通强度	相序可变	进入交叉口的车辆数	softmax	周期	100 (假设)	邻近状态合作
Arel 等人[26] (2010)	Q 学习	车道级:车辆相对延误	相序可变	节省的延误时间	ε-greedy	20 s	5 (假设)	邻近状态合作
Salkham 等人[18] (2008)	改进 Q 学习	车道级:各相位排队车辆数、邻近回报值	相序固定	交叉口已释放车辆数和正在排队车辆的差	softmax	3 倍周期	64 (真实)	邻近共同奖励
Balaji 等人[18] (2010)	Q 学习	相位级:相位占有率、相位平均排队长度	绿灯时长调整系数	综合考虑当前及历史同期路口排队车辆数的变化率	–	周期	29 (真实)	邻近状态及 Q 值共享
Aziz 等人[29] (2013)	R 学习	车道级:相位相对排队长度、邻近最拥堵路口 ID	相位可变	基于状态划分的多目标反馈函数:排队、延误、通过量	ε-softmax	4 s	8 (假设)	邻近状态合作
Prabuchandran 等人[30] (2015)	Q 学习	车道级:各车道排队长度、相位时长	相序可变	综合考虑排队长度和相位时长的损失	ε-greedy UCB	10 s	9/20 (真实)	邻近反馈共享

Richter 等人[28] 将基于自然梯度的 Actor Critic 算法应用于交通信号控制,RL 智能体通过考虑邻近交叉口的交通强度拓展了状态空间,分别采用四种强化学习算法进行自学习交通控制。Salkham 等人[18] 通过邻近路口的反馈共享实现多个路口间交通信号的协调控制,每个智能体采用周期式优化的控制逻辑,根据繁忙和不繁忙的二元交通状态,每隔两个周期采用 Q 学习算法更新配时方案的绿信比。Balaji 等人[19] 提出一种共享邻近状态和策略知识的 MARL 控制方法。RL 智能体根据历史状态、实时交通状态和邻近路口协作信息,评估下一控制周期内本地路口交通需求及其变化趋势,综合考虑邻近路口的协作需求,采用加权法优化下一周期的相位绿灯时间,采用 Q 学习算法优化本地策略知识。Arel 等人[26] 研究了基于状态合作的 MARL 控制方法,研究采用 BP 神经网络存储 Q 值函数,实现在与环境交互过程中采用 Q 学习算法自学习最优控制策略。Aziz 等人[29] 首次将 R 学习应用于分散式交通信号控制,每个智能体共享邻近路口的状态信息,并采用基于交通状

态划分的反馈函数,实现多目标控制。Prabuchandran 等人[30]研究一种基于 Q 学习的分散式 RL 控制方法,每个智能体根据本地路口的交通数据和邻近路口的反馈信息,拓展本地反馈函数的结构,在与环境交互采样的过程中,采用即时决策控制的逻辑自主学习策略知识并优化相位顺序。

与完全孤立的 MARL 控制相比,部分状态合作的 MARL 控制可更准确地响应区域内交通状态模式的变化,系统控制效益平均提高 10% 以上,减少了大流量条件下排队上溯情况的发生[18,19,26,28~30]。但是这种控制方法仅考虑了上下游路口的简单状态标志信息,多路口之间没有动作联动,不能系统描述网络交通流的动态性[8];同时,研究发现周期式 RL 优化技术应用于多路口协调控制存在时效性差等问题[1,18]。在部分状态合作的 MARL 控制中,虽然 RL 智能体具有简单的协调能力,但仍面临两个挑战[5,7]:a)简单的状态合作并未考虑多智能体之间同时动作对策略的内在影响;b)多智能体仍采用贪心选择更新其值函数,系统可能收敛于非均衡的联合策略。

2.3 动作联动的 MARL 控制

为克服 MARL 控制的同时学习挑战和决策挑战,动作联动的 MARL 控制将式(1)中单智能体的状态和动作分别替换为动态随机环境下的联合状态和联合动作,并在每一个博弈对策阶段,估计均衡策略的值函数,实现多个智能体间的同时对策,通过如此反复迭代逼近最优策略[31],以此寻找随机环境下系统的唯一均衡[7,12]。其基本形式如式(4)所示。

$$\begin{aligned} Q_i^k(s^k,a^k) &= Q_i^{k-1}(s^k,a^k)+\alpha^k[r_i^k(s^k,a^k,s^{k+1})+\\ &\quad \gamma V_i^{k-1}(s^{k+1})-Q_i^{k-1}(s^k,a^k)]\\ V_i^{k-1}(s^{k+1}) &\in NE^i[(Q_1^k(s^{k+1}),\cdots,Q_n^k(s^{k+1})]\\ \pi^* &\in \arg\max_{\pi\in NE}\sum_{i\in N}\sum_{a\in A}\pi(a)\times Q_i^k(s^{k+1},a) \end{aligned} \tag{4}$$

其中:$s=\{s_1,\cdots,s_N\}$ 表示 N 个智能体的联合状态向量;$a=\{a_1,\cdots,a_N\}$ 表示其联合动作向量;A 为联合动作空间;$V_i(s)$ 为智能体 i 在联合状态 s 下的状态值函数;NE 为纳什均衡对策;$\pi(a)$ 为带有不确定性的混合策略,即 N 个智能体选择联合动作的概率。

在动作联动的 MARL 控制中,RL 智能体的 Q 值空间将随智能体数量、状态空间及动作空间的增加呈指数级增长,致使其值函数的存储和查询效率难以保证。因此,几乎所有动作联动的 MARL 算法的核心都是如何设计多智能体联合状态—联合动作的基础数据结构、精细协调机制和有效估计值函数,通过协调多个智能体的动作选择,达到唯一的系统均衡[5,7,8]。既有研究中动作联动的 MARL 控制技术特征如表 4 所示。

表4 动作联动的 MARL 控制的技术特征

代表研究	学习算法	RL 控制器要素			选择策略	动作频率	路网规模	协调水平
		交通状态定义	信号动作定义	反馈函数				
Kuyer 等人[32](2008)	模型 Q 学习	车辆级:车辆排队位置及其等待时间	相序可变	相邻决策点之间增加排队长度	ε-max-plus	1 s	15(假设)	邻近动作联动
EI-Tantawy 等人[10](2013)	Q 学习	车辆级:当前相位 ID,绿灯持续时长,各相位排队车辆数	相序可变	相邻决策点间累积延误的差值	ε-MARLIN	1 s	59(真实)	邻近动作联动
Zhu 等人[33](2015)	R 学习	车道级:当前相位 ID,最大排队长度相位 ID,路口拥堵水平	相序可变	相邻决策点间增加的排队车辆	ε-JTA	4 s	18(假设)	邻近动作联动
Medina 等人[34](2014)	Q 学习	相位级:当前相位 ID,各相位排队车辆数(离散化)	相序可变	综合考虑通过车辆数、红灯排队车辆数和系列惩罚规则	ε-max-plus	2 s	20(真实)	全局动作联动
夏新海[15](2013)	Q 学习	车辆级:进口道头车车辆位置,各流向最大排队长度	绿灯时长调整系数	通过车辆数和累积等待时间	ε-Nash 等	周期	1×9 3×3 等(假设)	全局动作联动

Kuyer 等人[32]拓展了 Wiering 的 RL 控制方法,提出了一种基于协作图的交通信号协调控制方法,并在邻近智能体之间相互交互本地状态信息,采用 max-plus 算法寻找最优联合动作。Max-plus 算法是一种直接协调方法,它需要在多智能体间循环地传递消息,联合动作的计算量大且算法可能陷入循环传递,需要设计算法强制终止的判定准则;加之,节点的信念定义是一个主观的近似效用而非准确的反馈激励,基于 max-plus 直接协调的 MARL 控制方法容易收敛到次优策略。

EI-Tantawy 等人[10]提出了一种基于 MARL 的区域交通信号一体化集成控制方法,每个智能体仅与其物理邻近的智能体进行两两对策,采用联合状态—联合动作的协调机制,在学习过程中多智能体间同时学习,以解决移动目标的问题,并遵循一般和博弈的 NSCP 算法,定制了邻近最好响应对策的 MARL 算法,在决策过程中协调智能体之间的动作选择,以解决唯一均衡决策问题。据笔者认知,这是迄今为止 MARL 在区域交通控制领域较为完备的应用研究。该系统正在加拿大伯灵顿的两个路口进行实地验证,但仍然存在控制参数未作综合分析、相位结构方案固定、没有过饱和交通流调控机制等问题,这限制了控制算法对时变交通流的响应能力以及复杂高饱和交通条件下的控制性能。

Zhu 等人[33]通过将区域交通信号的最优决策看做是最优联合动作的概率推理,采用概率图模型中的联合树推理来简化全联合动作概率推理的计算,提出一种基于联合树推理的网络交通信号协同联动控制方法。与 EI-Tantawy 仅考虑邻近路口定义目标函数不同,该方法综合考虑网络内全部路口定义目标函数。此算法在理论上实现了整体最好响应联合动作的推理,但随着纳入网络控制路口数量的增加,联合树推理的计算量急剧增大。因此,这种集中式推理算法较适合于小范围内干线交通信号的全联合最优动作的推理。

Medina 等人[34]研究了基于 MARL 的干线交通信号协调控制,通过将图论中双向消息循环传递的 max-plus 算法应用于干线内全部路口最优协调控制动作的推理,采用 Q 学习和 ADP 算法实现了干线上交通信号策略的自学习。Abdoos 等人[35]采用 tile coding 的线性逼近器提出一种基于函数逼近的 RL 交通控制器。系统采用两层分布式系统架构,网络层智能体采用全联合状态和全联合动作的方式实现智能体间的协调,并采用 tile coding 估计智能体的值函数以避免全体状态空间的维度灾难,而路口层智能体则负责执行 Q 学习算法,并受网络层智能体的主动控制。

夏新海[16]面向离散状态和连续状态问题的不同应用层次,以逼近最优化控制策略为目标,从直接值函数法到直接策略搜索法,从多人对策的均衡解到模型估计的近似解,再到梯度上升的最好响应解,分别设计了五种 MARL 控制算法。这是国内较为详细的 MARL 控制研究。但是众多算法参数的耦合增加了系统的不确定性,需要针对具体应用场景定制其策略梯度或是设计求解均衡解的启发式规则,进而可能将控制问题

复杂化;且在交通控制层面,笔者所提算法采用周期响应式优化调整逻辑,没有标准化设计五个 RL 智能体的结构要素。

值得注意的是,为实现区域交通网络的一体化集成控制,EI-Tantawy 和 Zhu 等人[10,33]根据邻近智能体间交互作用的假设,采用模块化分区技术构建邻近两两智能体间联合状态—联合动作的 Q 值函数,大幅精简了系统状态空间,并在历史交通数据的基础上估计邻近智能体的策略知识,采用启发式规则推理最优联合动作,以期通过邻近智能体彼此间的相互联动,倒逼智能体的行为动作对区域交通网络的影响,这为区域交通的一体化协调控制提供了新的思路。与多时段定时控制相比,此类动作联动的 MARL 控制虽然采用间接协调的机制,有效降低了路口延误达 50% 以上,大幅提升了网络级控制效益,但在 MARL 算法与交通信号控制交互作用机理、相位与配时参数一体化设计等方面的认识仍有待提升。

3　问题与挑战

交通信号 MARL 控制研究的演化发展如图 4 所示。

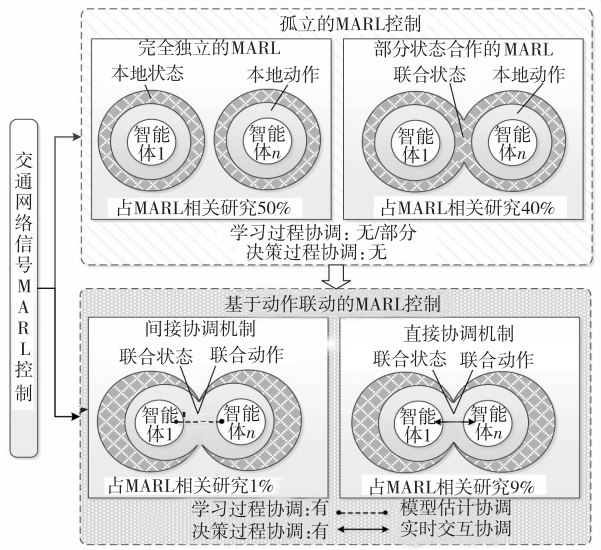

图4　基于MARL的交通信号控制研究的演化发展

绝大多数研究以假设的静态随机环境为对象,采用完全孤立[13~15,20,21]或部分状态合作[26~30]的协调机制进行本地路口的最优化控制,这制约了网络交通控制系统的整体效益。近年来,基于动作联动的 MARL 控制方法[10]发展迅速,其以联动协同的方式逼近全局最优的控制策略;同时,算法的验证也由假设的交通网络向现实的交通网络发展[32~34]。自马寿峰等人[15]于 2002 年首次将 Q 学习应用于单路口交通信号控制以来,绝大多数国内研究以单路口为对象,部分状态合作的 MARL、动作联动的 MARL 长期处于跟踪国外最新研究成果阶段[16,25,36~40]。整体而言,前两者将单智能体 RL 算法简单拓展并应用到多智能体系统,而动作联动的 MARL 则是随机环境下多人对策问题。MARL 交通控制的应用研究存在如下共性挑战:

a)交通控制与强化学习的交互作用机理。既有研究提出了许多 RL 控制方法,但没有系统地定量分析每一种 TD 学习算法的适用性。EI-Tantawy 等人对 RL 智能体五类要素的 16 种模型表征开展了仿真综合分析[13],但 RL 数据模型参数(状态离散等级、单位延长间隔等)对控制算法的影响仍有待深入挖掘。再者,既有 MARL 研究基本采用目标相位与当前相位是否一致作为相位切换的判定准则,频繁切换的相位可能强行中断连续到达的车流而削弱了交通系统的稳定性。因此,如何均衡 MARL 自治智能体的稳定性、灵活性及协调性,尚需要理论和实际工作的进一步检验。

b)多交叉口间联动协调机制。直接协调法与其他智能体间实时协商其最优动作的选择,代表性方法包括基于消息传递的集中式协作图法[41]和基于邻近的分布式直接策略搜索法[42]。这类方法通信需求和计算量大,并不适合大规模路网应用。间接协调法根据智能体交互过程中产生的历史数据,估计其他智能体动作选择的概率模型。其代表性方法有 NSCP 算法[43]。这类方法无须智能体间实时协商,适合大规模路网应用,但需要及时更新估计模型的知识。当前,如何在动态随机环境下高效和准确地实现多智能体的对策,仍是众多动作联动的 MARL 算法致力突破的关键[40]。

c)交通状态的特征抽取。既有 MARL 算法已采用邻近路口交互原则[44]等大幅精减系统的状态空间,但是联合学习单元的状态空间数量仍非常庞大。为克服维数灾难等难题,自适应动态规划方法被提出。但这类方法的好坏直接取决于逼近函数的配置和参数的选择,这给算法的明确物理意义和标准化设计带来了诸多挑战[25];同时,新的参数和模型的引入使得系统的推理决策更加复杂。笔者认为,不同层级(区域、干线、单点)的 MARL 控制问题对交通状态特征的要求不同,需要合理地抽取其数据模型的特征[45];否则,模型将被过度泛化或超拟合,数据反而将成为噪声。

d)多模式交通整合控制。多数 MARL 研究[7,8]仅考虑了机动车,并未涉及公共交通、行人和非机动车等模式,且反馈激励仅以乘用车车辆数为建模的基本单元,这样就忽略了公共交通等大容量交通方式的综合效益。可考虑设计大容量公交优先等规则,采用多模式交通的综合效益权重,拓展反馈奖励的结构,以实现多模式交通的整合控制,这将更好地符合我国城市道路混合交通流的实际。

e)MARL 交通控制方法的应用边界的综合分析。当前研究采用交通仿真实验,从网络、干线和路口级三个层次对 MARL 交通控制效果进行评价。研究结果表明,MARL 控制效果与交叉口条件及流向组织、交通流条件、整体路网配置、邻近交叉口间距、状态离散等级等相关[46]。在实际应用时,交通管理者应结合路网特征及交通流特性等展开 MARL 效用的系统分析,综合考察 RL 智能体的灵活性、协调性和自适应性,以及其对网络交通流的影响等。

4　结束语

城市交通网络环境具有典型的动态性和随机性,在不能完全获取交通系统状态信息、不能完全理解系统内部机理、不能建立被控对象精确模型的条件下,基于 MARL 的交通信号控制方法仅利用控制过程的输入和输出数据,自主寻找隐含的控制知识,具有无模型、自学习、闭环反馈、联动协调等优点,是一种无模型纯数据驱动的交通控制方法,在理论上可解决基于模型的交通控制需要精确数学模型以及基于智能计算交通控制无自学习能力等固有不足,可为基于数据的区域交通控制提供一种可行之法。需要注意的是,国内外学者在突破交通信号 MARL 控制的先进算法以逼近理论最优解的同时,应重点关注其交通状态特征抽取、自稳定机制、多目标反馈、状态离散等一系列基础问题的认识与解析,切实推动 MARL 控制方法在工程实践中的应用。

参考文献:

[1] Hamilton A, Waterson B, Cherrett T, *et al*. The evolution of urban traffic control: changing policy and technology [J]. Transportation Planning and Technology, 2013, 36(1): 24-43.

[2] Zhang Junping, Wang Feiyue, Wang Kunfeng, *et al*. Data-driven intelligent transportation systems: a survey [J]. IEEE Trans on Intelligent Transportation Systems, 2011, 12(4): 1624-1639.

[3] Wu Xinkai, Liu H X. Using high-resolution event-based data for traffic modeling and control: an overview [J]. Transportation Research Part C: Emerging Technologies, 2014, 42(5): 28-43.

[4] Thorpe T L. Vehicle traffic light control using SARSA[R]. Colorado: Colorado State University, 1997.

[5] Bazzan A L C. Opportunities for multiagent systems and multiagent reinforcement learning in traffic control [J]. Autonomous Agents and Multi-Agent Systems, 2009, 18(3): 342-375.

[6] 陆化普, 孙智源, 屈闻聪. 大数据及其在城市智能交通系统中的应用综述 [J]. 交通运输系统工程与信息, 2015, 15(5): 45-52.

[7] El-Tantawy S, Abdulhai B. Towards multi-agent reinforcement learning for integrated network of optimal traffic controllers (MARLIN-OTC) [J]. Transportation Letters, 2010, 2(2): 89-110.

[8] Mannion P, Duggan J, Howley E. An experimental review of reinforcement learning algorithms for adaptive traffic signal control [M]//Autonomic Road Transport Support Systems. Berlin: Springer, 2016: 47-66.

[9] Abdulhai B, Karakoulas G J, Pringle R. Reinforcement learning for true adaptive traffic signal control [J]. Journal of Transportation Engineering, 2003, 129(3): 278-285.

[10] El-Tantawy S, Abdulhai B, Abdelgawad H. Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC): methodology and large-scale application on downtown toronto [J]. IEEE Trans on Intelligent Transportation Systems, 2013, 14(3): 1140-1150.

[11] Abdulhai B, Pringle P. Autonomous multiagent reinforcement learning-5GC urban traffic control [C]//Proc of the 82nd Annual Meeting of Transportation Research Board. 2003.

[12] Wiering M A, Van Otterlo M. Reinforcement learning: state-of-the-art [M]. New York: Springer-Verlag, 2012.

[13] El-Tantawy S, Abdulhai B. Comprehensive analysis of reinforcement learning methods and parameters for adaptive traffic signal control [C]// Proc of the 90th Transportation Research Board Annual Meeting. 2011.

[14] Jin Junchen, Ma Xiaoliang. Adaptive group-based signal control by reinforcement learning [J]. Transportation Research Procedia, 2015, 10(7): 207-216.

[15] 马寿峰, 李英, 刘豹. 一种基于 agent 的单路口交通信号学习控制方法[J]. 系统工程学报, 2002, 17(6): 526-530.

[16] 夏新海. 面向城市自适应交通信号控制的强化学习方法研究 [D]. 广州: 华南理工大学, 2013.

[17] Busoniu L, Babuska R, De Schutter B. A comprehensive survey of multi-agent reinforcement learning [J]. IEEE Trans on Systems, Man, and Cybernetics Part C: Applications and Reviews, 2008, 38(2): 156-172.

[18] Salkham A, Cunningham R, Garg A, *et al*. A collaborative reinforcement learning approach to urban traffic control optimization[C]//Proc of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. 2008: 560-566.

[19] Balaji P G, German X, Srinivasan D. Urban traffic signal control using reinforcement learning agents [J]. IET Intelligent Transport Systems, 2010, 4(3): 177-188.

[20] Wiering M. Multi-agent reinforcement learning for traffic light control [C]// Proc of the 17th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann, 2000: 1151-1158.

[21] Oliveira D, Bazzan A L C, Da Silva B C, *et al*. Reinforcement learning based control of traffic lights in non-stationary environments: a case study in a microscopic simulator [C]//Proc of the 4th European Workshop on Multi-Agent Systems. 2006.

[22] Duan Houli, Li Zhiheng, Zhang Yi. Multiobjective reinforcement learning for traffic signal control using vehicular Ad hoc network [J/OL]. EURASIP Journal on Advances in Signal Processing, 2010. http://doi.org/10.1155/2010/724035.

[23] Brys T, Pham T T, Taylor M E. Distributed learning and multi-objectivity in traffic light control [J]. Connection Science, 2014, 26(1): 65-83.

[24] Khami M A, Gomaa W. Adaptive multi-objective reinforcement learning with hybrid exploration for traffic signal control based on cooperative multi-agent framework [J]. Engineering Applications of Artificial Intelligence, 2014, 29(3): 134-151.

[25] 赵冬斌, 刘德荣, 易建强. 基于自适应动态规划的城市交通信号优化控制方法综述[J]. 自动化学报, 2009, 35(6): 676-681.

[26] Arel I, Liu C, Urbanik T, *et al*. Reinforcement learning-based multi-agent system for network traffic signal control [J]. IET Intelligent Transport Systems, 2010, 4(2): 128-135.

[27] Prashanth L, Bhatnagar S. Reinforcement learning with function approximation for traffic signal control [J]. IEEE Trans on Intelligent Transportation Systems, 2011, 12(2): 412-421.

[28] Richter S, Aberdeen D, Yu Jin. Natural actor-critic for road traffic optimisation [C]//Advances in Neural Information Processing Systems. Piscataway, NJ: IEEE Press, 2006: 1169-1176.

[29] Aziz H M, Zhu Feng, Ukkusuri S V. Reinforcement learning-based signal control using R-Markov average reward technique accounting for neighborhood congestion information sharing [C]//Proc of the 92nd Annual Meeting Transportation Research Board. 2013.

[30] Prabuchandran K J, Bhatnagar S. Decentralized learning for traffic signal control [C]//Proc of the 7th International Conference on Communication Systems and Networks. Piscataway, NJ: IEEE Press, 2015: 1-6.

[31] Bazzan A L C, De Oliveira D, Da Silva B C. Learning in groups of traffic signals [J]. Engineering Applications of Artificial Intelligence, 2010, 23(4): 560-568.

[32] Kuyer L, Whiteson S, Bakker B, *et al*. Multiagent reinforcement learning for urban traffic control using coordination graphs [M]//Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2008: 656-671.

[33] Zhu Feng, Aziz H M A, Qian Xinwu, *et al*. A junction-tree based learning algorithm to optimize network wide traffic control: a coordinated multi-agent framework [J]. Transportation Research Part C: Emerging Technologies, 2015, 58(9): 487-501.

[34] Medina J C, Benekohal R F. Corridor-based coordination of learning agents for traffic signal control by enhancing max-plus algorithm [C]// Proc of the 93rd Transportation Research Board Annual Meeting. 2014.

[35] Abdoos, M, Mozayani, N, Bazzan, A. Hierarchical control of traffic signals using Q-learning with tile coding [J]. Applied Intelligence, 2014, 40(2): 201-213.

[36] 何兆成, 佘锡伟, 杨文臣, 等. 结合 Q 学习和模糊逻辑的单路口交通信号自学习控制方法 [J]. 计算机应用研究, 2011, 28(1): 199-202.

[37] 龙琼, 胡列格, 张谨帆, 等. 考虑交通管理策略的交叉口信号控制多目标优化 [J]. 中南大学学报: 自然科学版, 2014, 45(7): 2503-2508.

[38] 伦立宝. 基于强化学习的城市交通信号控制方法研究 [D]. 西安: 西安电子科技大学, 2013.

[39] 聂建强, 徐大林. 基于模糊 Q 学习的分布式自适应交通信号控制 [J]. 计算机技术与发展, 2013, 23(3): 171-174.

[40] 陈学松, 杨宜民. 强化学习研究综述[J]. 计算机应用研究, 2010, 27(8): 2834-2838, 2844.

[41] Kok J R, Vlassis N. Collaborative multiagent reinforcement learning by payoff propagation [J]. Journal of Machine Learning Research, 2006, 7(1): 1789-1828.

[42] Yagan D, Tham C K. Coordinated reinforcement learning for decentralized optimal control [C]//Proc of IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning. Piscataway, NJ: IEEE Press, 2007: 296-302.

[43] Weinberg M, Rosenschein J S. Best-response multiagent learning in non-stationary environments [C]// Proc of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems. New York: IEEE Computer Society, 2004: 506-513.

[44] Ono N, Fukumoto K. Multi-agent reinforcement learning: a modular approach [C]//Proc of the 2nd International Conference on Multi-Agent Systems. 1996: 252-258.

[45] Jeffrey G. Reinforcement learning for adaptive traffic signal control [EB/OL]. [2016-12-25]. http://cs229.stanford.edu/proj2015/369_report.pdf.

[46] Abdelgawad H, Abdulhai B, El-Tantawy S, *et al*. Assessment of self-learning adaptive traffic signal control on congested urban areas: independent versus coordinated perspectives [J]. Canadian Journal of Civil Engineering, 2015, 42(6): 353-366.