计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2021.20211023
58(12):2585-2603,2021

面向无人驾驶时空同步约束制导的安全强化学习

王金永[1,2] 黄志球[1,2] 杨德艳[3] Xiaowei Huang[4] 祝 义[3] 华高洋[1,2]
[1](南京航空航天大学计算机科学与技术学院 南京 211106)
[2](高安全系统的软件开发与验证技术工信部重点实验室(南京航空航天大学) 南京 211106)
[3](江苏师范大学计算机科学与技术学院 江苏徐州 221116)
[4](利物浦大学计算机科学系 英国利物浦 L69 3BX)
(jinyongw@nuaa.edu.cn)

Spatio-Clock Synchronous Constraint Guided Safe Reinforcement Learning for Autonomous Driving

Wang Jinyong[1,2], Huang Zhiqiu[1,2], Yang Deyan[3], Xiaowei Huang[4], Zhu Yi[3], and Hua Gaoyang[1,2]
[1](*College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing* 211106)
[2](*Key Laboratory of Safety-Critical Software (Nanjing University of Aeronautics and Astronautics), Ministry of Industry and Information Technology, Nanjing* 211106)
[3](*School of Computer Science and Technology, Jiangsu Normal University, Xuzhou, Jiangsu* 221116)
[4](*Department of Computer Science, University of Liverpool, Liverpool, UK* L69 3BX)

Abstract Autonomous driving systems integrate complex interactions between hardware and software. In order to ensure the safe and reliable operations, formal methods are used to provide rigorous guarantees to satisfy logical specifications and safety-critical requirements in the design stage. As a widely employed machine learning architecture, deep reinforcement learning (DRL) focuses on finding an optimal policy that maximizes a cumulative discounted reward by interacting with the environment, and has been applied to autonomous driving decision-making modules. However, black-box DRL-based autonomous driving systems cannot provide guarantees of safe operation and reward definition interpretability techniques for complex tasks, especially when they face unfamiliar situations and reason about a greater number of options. In order to address these problems, spatio-clock synchronous constraint is adopted to augment DRL safety and interpretability. Firstly, we propose a dedicated formal properties specification language for autonomous driving domain, i.e., spatio-clock synchronous constraint specification language, and present domain-specific knowledge requirements specification that is close to natural language to make the reward functions generation process more interpretable. Secondly, we present domain-specific spatio-clock synchronous automata to describe spatio-clock autonomous behaviors, i.e., controllers related to certain spatio- and clock-critical actions, and present safe state-action space transition systems to guarantee the safety of DRL optimal policy generation process. Thirdly, based on the formal specification and policy learning, we propose a formal spatio-clock synchronous constraint guided safe reinforcement learning method with the goal of

收稿日期:2021-10-10;**修回日期**:2021-11-16
基金项目:国家重点研发计划项目(2018YFB1003900);国家自然科学基金项目(61772270, 62077029)
This work was supported by the National Key Research and Development Program of China (2018YFB1003900) and the National Natural Science Foundation of China (61772270, 62077029).
通信作者:黄志球(zqhuang@nuaa.edu.cn)

easily understanding the safe reward function. Finally, we demonstrate the effectiveness of our proposed approach through an autonomous lane changing and overtaking case study in the highway scenario.

Key words spatio-clock synchronous constraint; formal specification; safe reinforcement learning; temporal difference; intelligent traffic simulation; autonomous driving safety

摘 要 无人驾驶系统综合了软件和硬件复杂的交互过程,在系统设计阶段,形式化方法可以保证系统满足逻辑规约和安全需求;在系统运行阶段,深度强化学习被广泛应用于无人驾驶系统决策中.然而,在面对没有经验的场景和复杂决策任务时,基于黑盒的深度强化学习系统并不能保证系统的安全性和复杂任务奖励函数设置的可解释性.为此提出了一种形式化时空同步约束制导的安全强化学习方法.首先,提出了一种形式化时空同步约束规约语言,接近自然语言的安全需求规约使奖励函数的设置更具有解释性.其次,展示了时空同步自动机和状态-动作空间迁移系统,保证强化学习的状态行为策略更加安全.然后,提出了结合形式化时空约束制导的安全强化学习方法.最后,通过无人驾驶汽车在高速场景变道超车的案例,验证所提方法的有效性.

关键词 时空同步约束;形式化规约;安全强化学习;时序差分;智能交通仿真;无人驾驶安全

中图法分类号 TP181

无人驾驶系统可以看作是自车车辆和环境在时间空间上的转移,其安全性一直是阻碍无人驾驶普及的首要问题.系统安全性(safety)可定义为坏的事情永远不要发生,无人驾驶安全指的是车辆之间和车辆与周围的环境之间不要发生碰撞.无人驾驶决策模块需要根据动态的时钟和空间信息做出正确安全的决策,提供安全的时空道路通行权(right-of-way),避免在某一时刻自车车辆与周围环境车辆的空间重叠.安全攸关的无人驾驶车辆处在复杂的时空运动轨迹中,在面对没有训练过的未知不确定场景时,系统设计者不仅需要采用强化学习(reinforcement learning, RL)的方法来训练车辆,使得车辆获得最优的策略和累计奖励,而且更加需要特定领域的形式化规约方法,为强化学习提供安全约束和保障.

无人驾驶决策的方法可以分为 2 类:1)基于规则(rule-based)的方法,比如有限状态机(finite state machine);2)数据驱动(data-driven)的方法,比如机器学习(machine learning)的方法.基于规则的决策系统可以根据事先定义好的规则输出事先规定好的动作,从而实现系统的安全与可解释性[1-4].尽管基于规则的方法可以为安全的系统操作提供保障,但是只有复杂的现实环境与待验证的模型相匹配才可以满足系统设计的需要[5].对于复杂的无人驾驶交通场景,系统状态转换较多,决策系统不能理解周围复杂的交通环境,而且对于未识别的系统状态无法给出较好的决策行为[6].基于数据驱动的方法,通过与环境交互大量的学习训练,可以理解学习周围的环境.对于深度强化学习来说,学习的过程受奖励函数的指导,产生从观测输入到动作输出的策略.但是基于学习的决策方法,较多依赖数据和黑盒训练,缺少学习过程的安全约束和决策结果的可解释性[7-12].因此本文主要关注无人驾驶系统设计与运行部署阶段的安全性与奖励函数设置的可解释性,具体研究无人驾驶车辆在时间空间同步约束系统中不发生时空重叠(no spatio-clock overlaps)以及形式化规约制导的安全强化学习方法(formal specification guided safeRL).

近期的研究工作显示无人驾驶系统已经广泛采用时间空间约束的形式化规约和强化学习相关技术.在无人驾驶领域空间建模方面,协同无人驾驶系统安全性需要采用专用的领域特定的空间逻辑来描述车辆的物理空间位置和逻辑空间关系[13-14].无人驾驶系统包含连续状态变化和离散状态迁移,无人车变道控制器可以用来保证空间位置不发生重叠[15-16].在无人驾驶领域时间建模方面,时间自动机被用来建模通信无人驾驶车辆之间的行为[17],时钟约束规约语言(clock constraint specification language, CCSL)被用来建模系统的安全约束[18-19].时空同步约束系统中的事件触发不仅受严格的时序和物理时间的约束,同时还要逻辑和物理空间关系的约束[20-21].无人驾驶决策系统的核心问题是如何安全高效地分配路权来解决可能的时空重叠.无人驾驶行为需要在满足

时间约束的同时满足空间关系的约束,反之亦然,这种方法可以保证在某一时刻和空间位置,只允许最多一辆车存在.时空安全攸关的事件包括安全卫士和不变式规约,需要满足逻辑时序和物理时钟关系,同时需要满足逻辑空间关系和物理空间位置.然而在无人驾驶领域,现有的时空同步约束规约方法和验证技术是不充分的.

无人驾驶系统是一类典型的信息物理系统(cyber-physical systems, CPS),高保真的物理模型开发成本很高并且难以验证.随着计算能力的提高和可用交通数据的增加,越来越多的数据驱动的机器学习技术可以用来解决无人系统的智能决策.基于强化学习的无人驾驶基本思想是智能体(agent)直接与环境(environment)之间不断试错,通过不断的交互,从环境中寻求最优策略(optimal policy),智能体会接收到环境发来的奖励(reward),强化学习的目标就是通过改进策略以期获得最大化累计奖励(maximize cumulative future rewards)[22-23].安全强化学习(safeRL)是指在学习训练和执行过程中,学习最优策略的过程需要满足严格的形式化安全约束[24-25].目前强化学习多数应用在模拟环境中,没有大规模被应用在实车训练和实践中,主要是用于2个原因:1)由于强化学习的状态空间和动作空间输入来自黑盒的卷积神经网络和全连接神经网络,缺乏可解释性成为深度强化学习部署到现实系统的主要障碍之一;2)由于面对新的没有训练过的交通场景,强化学习不能为安全攸关的无人驾驶系统提供严格的安全约束和负责任的规则约束.

针对上述面向无人驾驶安全的智能决策方法研究现状和存在的研究问题,本文主要关注无人驾驶领域专用的时空同步约束规约语言和时空同步行为描述自动机,并将形式化时空同步安全约束的强化学习用于无人驾驶决策模块,结合时空安全约束的强化学习来增强系统的安全性.主要研究贡献有5个方面:

1) 给出了时空同步约束系统的架构,并给出了无人驾驶领域相关的时空同步约束安全域和交通快照的定义;

2) 提出了面向无人驾驶领域的时空同步约束规约语言(spatio-clock synchronous constraint specification language, SCSL),该高层抽象规约语言可以处理时空同步约束的安全卫士条件和不变式,并且提出了强化学习奖励函数设置可解释性的方法;

3) 设计了无人驾驶领域的时空同步约束自动机(domain-specific spatio-clock synchronous constraint automata),系统设计者可以用该自动机描述无人驾驶行为和状态-动作空间迁移系统,这种方法为无人驾驶系统与环境之间的正确交互提供了安全保障和逻辑基础,为强化学习状态行为策略的生成提供了安全保障;

4) 提出了安全强化学习的方法框架,定义了时空同步约束奖励状态机,描述了非Markov性质的奖励函数结构,提供了任务目标和奖励函数之间的映射关系,指导非Markov性质奖励函数的设置;

5) 设计了无人驾驶汽车在高速公路变道超车的案例,在无人驾驶仿真环境SUMO中模拟实验展示本文所提方法的有效性.

1 研究基础

1.1 面向无人驾驶系统的强化学习架构

无人驾驶系统(autonomous driving systems)相当于强化学习系统中的智能体,集成了感知模块(perception module)、决策模块(decision-making module)、控制模块(control module)等智能功能,其中决策模块综合环境感知信息和自车(ego vehicle)信息,经过多次训练,利用强化学习可以实现模仿人类考驾照的类似人类驾驶行为,如图1所示.环境感知模块的主要功能是通过感知器和车辆通信获取强化学习的状态空间和动作空间输入信息,感知器主要包括传感器、激光、雷达和摄像机等,车辆通信主要包括车与车(vehicle to vehicle, V2V)、车辆与道路基础设施(vehicle to infrastructure, V2I)以及全球定位系统(global position system, GPS)等.决策模块是无人驾驶系统的软件模块,该模块提取出软件模块的输入信息,也就是高层抽象的场景理解信息,智能体通过形式化约束制导的安全强化学习,输出符合任务要求的最优策略;智能体软件模块经过学习与函数映射关系,输出可执行的行为决策信息给智能体执行模块,动作控制信息包括纵向运动的加速和减速等规划动作,以变道、超车、转向等横向运动规划信息.智能体的决策能力是实现车辆全自动化的基本要素,目前大量的研究采用强化学习来实施无人自主决策[26-29],综合周围交通环境和自车信息,产生安全通明安全的驾驶行为.在无人驾驶应用中,无人车做动作或者做决策,强化学习中的智能体就是无人车;而环境是与无人车交互的对象,无人车所处的真实交通物理世界就是环境,包括交互

过程中的道路信息、交通标志、交通规则、驾驶经验等规则和推理机制.动作空间是决策模块输出给执行模块的行为决策信息,奖励表示交通环境对智能体基于当前状态产生动作的打分,需要系统设计人员设计奖励函数,本文主要针对如何获得非 Markov 性质的安全可解释奖励函数的设置展开研究.

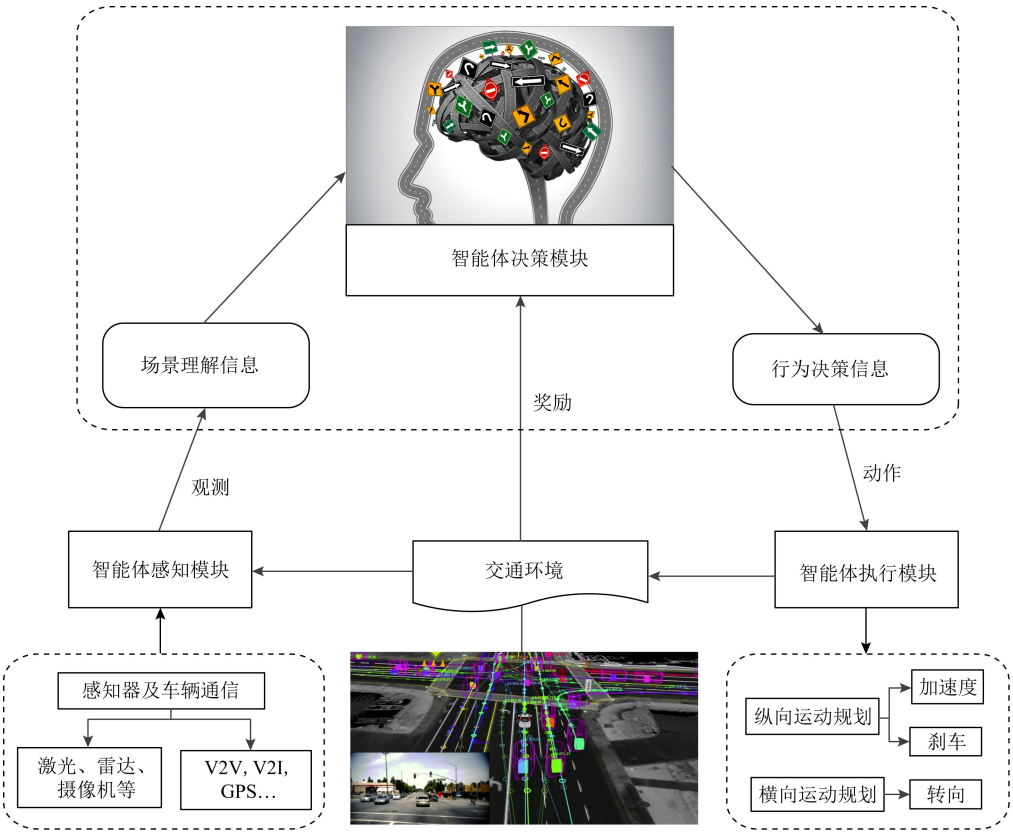


Fig. 1 The RL architecture for autonomous driving systems

图 1 无人驾驶系统的强化学习架构

强化学习的智能体可以建模为一个六元组 $\mathcal{M}_{\mathrm{MDP}}=\langle \mathcal{S},\mathcal{A},s_0,p_r,R,\gamma\rangle$,其中花体 $\mathcal{S}$ 是系统状态空间(state space),花体 $\mathcal{A}$ 是系统动作空间(action space),s_0 是初始状态(initial state),$p_r(\cdot|s,a)$ 表示系统在状态 s 下执行动作 a 的状态迁移概率或者分布(transition probability distribution),$R:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to\mathbb{R}$ 是奖励函数(reward function),$\gamma\in[0,1]$ 为折扣因子(discount factor),$\gamma=0$ 表示智能体只考虑眼前利益,不考虑长远利益;$\gamma=1$ 则表示累计奖励还偏重考虑远期价值;对于连续回合制(episode)决策系统,$\gamma\in(0,1)$ 可保证如果每一步奖励有界,则总回报(global return)也是有界的.智能体根据对环境状态的观测(observation),决定对环境采取的动作 a,其中 $a\in\mathcal{A}$;环境受智能体施加动作的影响,改变自己的状态 s,其中 $s\in\mathcal{S}$;智能体接收到环境基于动作发来的反馈奖励 R_t,记为 $R_t=R(s_t,a_t)$,从环境中寻求最优策略 π(policy).智能体决策系统会根据不同的观测决定采用不同的动作,策略是有关概率的集合或者分布,其元素是车辆根据观测进行下一步动作的函数对应关系.强化学习的学习对象是策略,学习目标是通过不断改进策略获得最大化累计回报 G_t(maximize cumulative return),使无人智能体获得尽可能多的来自环境的奖励,获得最优策略.在每一个时间步(timestep),无人驾驶智能体在当前状态根据观察来确定下一步的动作,这种映射策略关系表示为 $\pi:\mathcal{S}\to\mathcal{A}$.对于确定性策略,记作 $a=\pi(s)$,其中 a 表示动作,s 表示在策略 π 下的状态;而对于随机性策略,$\pi(a|s)\triangleq p_r\{\mathcal{A}_t=a|\mathcal{S}_t=s\}$,表示对某一状态 s 采取可能的动作 a 的概率.

1.2 双重深度 Q 学习网络

为了最大化累积奖励,定义当前时刻后的累积奖励作为回报,如果对未来的奖励信息简单求和,那么未来奖励信息的总和往往是无穷大.为了解决这一问题,引入折扣因子 γ,该参数体现了未来的奖励在当前时刻的价值奖励.从而定义回报奖励函数为 $G_t\triangleq R_t+\gamma R_{t+1}+\gamma^2R_{t+2}+\cdots=\sum_{k=0}^{\infty}\gamma^kR_{t+k}$,基于回

报奖励函数的定义,可以进一步定义状态价值函数(state value function)和动作价值函数(state-action value function).动作价值函数表示在状态 s_t 采取动作 a_t 后,采用策略 π 折扣回报的期望,该期望的结果标记为 $Q_\pi(s_t,a_t)$,定义为 $Q_\pi(s_t,a_t)\triangleq\mathbb{E}_\pi[G_t\mid S_t=s_t,\mathcal{A}_t=a_t]$.状态价值函数表示从状态 s 开始采用策略 π 的 Markov 链折扣回报 G_t 的期望,记为 $V_\pi(s_t)\triangleq\mathbb{E}_\pi[G_t\mid S_t=s_t]=\mathbb{E}_{\mathcal{A}_t\sim\pi(\cdot\mid s_t)}[Q_\pi(s_t,\mathcal{A}_t)]$.

贝尔曼方程式(1)和式(2)表示状态价值函数和动作价值函数之间的关系,也称为贝尔曼期望方程(Bellman expectation equations):

$$V_\pi(s)=\mathbb{E}_\pi\Big[\sum_{k=0}^{\infty}\gamma^k R_{t+k}\mid S_t=s\Big]=\sum_{a\in\mathcal{A}}\pi(a\mid s)Q_\pi(s,a)=\sum_{a\in\mathcal{A}}\pi(a\mid s)\times\Big(R(s,a)+\gamma\sum_{s'\in\mathcal{S}}p_r(s'\mid s,a)V_\pi(s')\Big),\tag{1}$$

$$Q_\pi(s_t,a_t)=R(s_t,a_t)+\gamma\sum_{s_{t+1}}p_r(s_{t+1}\mid s_t,a_t)V_\pi(s_{t+1}).\tag{2}$$

在多种策略函数 π 中选择最好的策略函数 $\pi^*=\arg\max_\pi Q_\pi(s_t,a_t)$,$\forall s_t\in\mathcal{S},a_t\in\mathcal{A}$,就可以得到最优动作价值函数 $Q^*(s_t,a_t)$,用最大化消除策略 π,表示为 $Q^*(s_t,a_t)=\max_\pi Q_\pi(s_t,a_t)$,$\forall s_t\in\mathcal{S},a_t\in\mathcal{A}$.同理,在多种策略函数中选择最好的策略函数 $\pi^*=\arg\max_\pi Q_\pi(s_t,a_t)$,$\forall s_t\in\mathcal{S},a_t\in\mathcal{A}$,就可以得到最优状态价值函数,表示为 $V^*(s_t)=\max_\pi V_\pi(s_t)$,$\forall s_t\in\mathcal{S}$.

强化学习的目标是学习最优策略 π^* 对应的最优动作价值函数 Q^*,当智能体尝试学习到最优的状态价值函数时,本文用 $Q_\pi^*(s,a)$ 表示,Q 学习(Q-learning)是一种常用的强化学习方法.Q 学习的目标是求得最优状态动作价值:

$$Q^*(s_t,a_t)=R_t+\gamma\sum_{s_{t+1}}p_r(s_{t+1}\mid s_t,a_t)\times\max_{a_{t+1}\in\mathcal{A}}Q^*(s_{t+1},a_{t+1}).\tag{3}$$

训练双重深度 Q 学习网络(double deep Q-learning network, DDQN)常用的算法是时序差分算法(tem-poral difference, TD).假设有一个模型 $Q(s,a;w)$,w 是模型训练的参数,给定一个四元组 (s_t,a_t,r_t,s_{t+1}),可推导出折扣回报为 $G_t=R_t+\gamma G_{t+1}$,根据式(3),当智能体执行动作 a_t 后,环境通过状态转移概率 $p_r(s_{t+1}\mid s_t,a_t)$ 可以计算出下一时刻的新状态 s_{t+1},而第 t 时刻的奖励最多只依赖于随机变量 S_t,A_t,S_{t+1},现在已经观测到了 s_t,a_t,s_{t+1},则可得变量 R_t 的观测值,记作 r_t,接着可以对期望做蒙特卡洛近似,可得:

$$Q^*(s_t,a_t)\approx r_t+\gamma\max_{a\in\mathcal{A}}Q^*(s_{t+1},a),\tag{4}$$

式(4)左边是总的动作价值函数预测,右边是执行一次动作 a,有了实际观测值 r_t 后,再用已经获得的 r_t 加上 $t+1$ 到回合结束的带折扣因子的最优动作价值函数估计.DDQN 在时刻 t 做出的预测值记为 $\hat{q}_t\triangleq Q(s_t,a_t;w)$,而式(4)右边是在时刻 $t+1$ 做出的预测,基于部分观测到的奖励 r_t,记为 $\hat{y}_t\triangleq r_t+\gamma Q(s_{t+1},a_{t+1};w_t)=r_t+\gamma\max_{a\in\mathcal{A}}Q(s_{t+1},a;w_t)$,$\hat{q}_t$ 和 $\hat{y}_t$ 都是对最优动作价值函数 $Q^*(s_t,a_t)$ 的估计,但是 TD 的目标 $\hat{y}_t$ 是在执行动作 a_t 后,基于部分事实,比估计值 $\hat{q}_t$ 更加可靠,所以可以用 $\hat{y}_t$ 对模型进行修正.定义损失函数(loss function)为

$$L(w)=\frac{1}{2}[Q(s_t,a_t;w)-\hat{y}_t]^2.$$

损失函数关于参数 w 的梯度为

$$\nabla_w L(w)\triangleq\frac{\partial L(w)}{\partial w}=[Q(s_t,a_t;w)-\hat{y}_t]\times\nabla_w Q(s_t,a_t;w).\tag{5}$$

为了让损失函数减小,由于目标是最小化目标函数,需要做一步策略梯度下降(policy gradient descent, PGD),沿着梯度的反方向更新参数,称为梯度下降 $w_{t+1}=w_t-\alpha\times\left.\frac{\partial L_t}{\partial w}\right|_{w=w_t}=w_t-\alpha\times(\hat{q}_t-\hat{y}_t)\mathrm{d}t$,做梯度下降也是为了让损失函数 $L(w)$ 不断减小,其中 $\mathrm{d}t$ 是动作价值函数的微分,记作 $\mathrm{d}t=\left.\frac{\partial Q(s_t,a_t;w)}{\partial w}\right|_{w=w_t}$,$\alpha$ 是学习率(learning rate),会影响梯度下降的收敛速度和神经网络的准确率,需要用户设置.下面给出 DDQN 求解最优动作价值的算法.

算法 1. DDQN 求解最优动作价值算法.

输入:观测到的环境状态 $s\in\mathcal{S}$ 和策略 π;

输出:2 个动作价值估计的平均值 $\frac{1}{2}(Q^0+Q^1)$.

① 初始化动作价值函数 $Q^i(s,a;w_i)$,$i\in\{0,1\}$;

② for each *episode*

/ * 每个回合时序差分学习 * /

③ 状态 s_t 和动作 a_t 采样,进行时序差分更新;

④ while (*episode*$<T$ 且 s_t 不是终止状态)

⑤ 选择最优动作 a^*,$a^*=$

$\arg\max_{a\in\mathcal{A}} Q^0(s_{t+1},a;w_0)$;

⑥ 进入到下一个状态 s_{t+1} 并且返回奖励 r_t;

⑦ 更新动作价值 $\hat{y}_t=r_t+\gamma Q^1(s_{t+1},a^*;w_1)$;

⑧ 计算 TD-error, $\delta_t=Q^0(s_t,a_t;w_0)-(r_t+\gamma Q^1(s_{t+1},a^*;w_1))$;

⑨ 更新 DDQN 参数 w_0 和 w_1;

⑩ return $\frac{1}{2}(Q^0+Q^1)$;

⑪ end while

⑫ 更新状态和动作 $s_t\leftarrow s_{t+1}, a_t\leftarrow a^*$;

⑬ end for

1.3 时钟约束规约语言和空间逻辑

时钟约束规约语言是基于时钟和时钟约束的形式化规约语言,时钟表示一个信号在每个离散时间点的触发状态;时钟约束描述信号时钟之间的逻辑关系.下面给出逻辑时钟、时钟调度和时钟历史的相关定义[30].

定义 1. 逻辑时钟.一个逻辑时钟 c 定义一个有限或者无限的时钟跳动组成的序列或者流,$c_i|_{i=1}^{\infty}\triangleq c_1c_2\cdots c_i\cdots, i\in\mathbb{N}^+$,$\mathbb{N}^+$ 为自然数(不包括 0)的集合.对于每一个 $c_i\in\{\text{tick},\text{idle}\}$,表示该时钟在时刻 i 的状态下,可以为触发状态 tick 或者非触发状态 idle.$\mathcal{C}$ 表示一个有限逻辑时钟的集合.

定义 2. 时钟调度.一个时钟调度 σ 反映了所有时钟在离散时间每个时刻的触发状态,是一个有限或者无限的序列 $\sigma(0)\sigma(1)\cdots\sigma(i)\cdots, i\in\mathbb{N}$,$\mathbb{N}$ 为 0 和正整数的集合.$\sigma(i)\in 2^c$ 表示在时刻 i 触发的时钟集合,形式化定义为

$$\sigma(i)=\begin{cases}\varnothing, & i=0,\\ \{c\,|\,c\in\mathcal{C}\wedge c(i)=\text{tick}\}, & i\geqslant 1.\end{cases}$$

定义 3. 时钟历史.对于给定的时钟调度 σ,时钟历史记录了在 σ 中每一个时钟截止到当前时刻的触发次数,形式化定义为

$$H_\sigma(c,i)=|\{j\,|\,j\in\mathbb{N}^+, j\leqslant i, c\in\sigma(j)\}|=$$
$$\begin{cases}H_\sigma(c,0)=0, & i=0,\\ H_\sigma(c,i+1)=H_\sigma(c,i), & \forall i\in\mathbb{N},\ c\notin\sigma(i),\\ H_\sigma(c,i+1)=H_\sigma(c,i)+1, & \forall i\in\mathbb{N},\ c\in\sigma(i).\end{cases}$$

基于上述定义,下面展示 6 种最基本的时钟关系的语法和语义,分别是优先关系(precedence)、互斥关系(exclusion)、子时钟关系(subclock)、同时关系(coincidence)、因果关系(causality)和交替关系(alternation).时钟关系包括逻辑时钟和二元连接操作符.时钟关系的语法定义如下,$Rel::=c_1\prec c_2\,|\,c_1\#c_2\,|\,c_1\subseteq c_2\,|\,c_1\equiv c_2\,|\,c_1\sim c_2\,|\,c_1\preccurlyeq c_2$,时钟关系的形式化语义见表 1.

Table 1 The Formal Semantics of Clock Relations

表 1 时钟关系的形式化语义

关系名称	符号	时钟关系语义
优先关系	$c_1\prec c_2$	$\forall i\in\mathbb{N}^+,(H_\sigma(c_1,i)=0\wedge H_\sigma(c_2,i)=0)\vee(H_\sigma(c_1,i)>H_\sigma(c_2,i))$
互斥关系	$c_1\#c_2$	$\forall i\in\mathbb{N}^+, c_1\notin\sigma(i)\vee c_2\notin\sigma(i)$
子时种关系	$c_1\subseteq c_2$	$\forall i\in\mathbb{N}^+, c_1\in\sigma(i)\rightarrow c_2\in\sigma(i)$
同时关系	$c_1\equiv c_2$	$\forall i\in\mathbb{N}^+, c_1\in\sigma(i)\leftrightarrow c_2\in\sigma(i)$
因果关系	$c_1\preccurlyeq c_2$	$\forall i\in\mathbb{N}^+, H_\sigma(c_1,i)\geqslant H_\sigma(c_2,i)$
交替关系	$c_1\sim c_2$	$\forall i\in\mathbb{N}^+, c_1(i)<c_2(i)\wedge c_2(i)<c_1(i+1)$

优先关系 $c_1\prec c_2$ 表示时钟 c_1 的触发快于时钟 c_2 的触发.互斥关系 $c_1\#c_2$ 表示若时钟 c_1 触发,则时钟 c_2 不触发,反之亦然.子时种关系 $c_1\subseteq c_2$ 表示在任意时刻,若时钟 c_1 触发.则时钟 c_2 必然触发.同时关系 $c_1\equiv c_2$ 表示时钟 c_1 和时钟 c_2 同时触发.因果关系 $c_1\preccurlyeq c_2$ 表示时钟 c_1 的触发不慢于时钟 c_2 的触发.交替关系 $c_1\sim c_2$ 表示时钟 c_1 和时钟 c_2 交替触发.

时空状态迁移表示空间关系随着时刻变化的而不断演化.区域连接演算(region connection calculus, RCC)主要用来描述空间区域间拓扑关系,该空间描述逻辑是 Randell 等人[31]在 Clarke[32]的空间演算逻辑公理的基础上提出的,将区域连接逻辑应用在无人驾驶系统时空同步规约中,可以用来描述无人驾驶车辆间的空间逻辑关系.

定义 4. 时空同步约束轨迹.对于空间安全攸关和时间安全攸关的无人系统,时空同步约束轨迹可以描述成九元组$(X,I,\mathcal{C},occ(I),app(I),safe(X),\eta,\partial,R)$,其中 X 表示空间区域的集合,对于每一个空间变量 $x_i, i\in\mathbb{N}^+$,有 $x_i\in X$;I 是无人驾驶车辆标识符集合,对于每一个车辆标识符 i,有 $i\in I$;$\mathcal{C}$ 表示时空动作相关的时钟集合,对于每一个逻辑时钟 $c\in\mathcal{C}$;$occ(I)$ 表示车辆占用的空间区域集合;$app(I)$ 表示车辆申请的空间集合;$safe(X)$ 表示无人驾驶车辆的安全区域集合;$\eta(\partial)$ 是后继(前序)操作符,对于每一个空间区域 x_i,其后继(前序)空间区域集合 $\eta(x_i)(\partial(x_i))$ 属于空间区域集合,也就是 $\eta(x_i)\subseteq P(X)$ $(\partial(x_i)\subseteq P(X))$,$P(X)$ 为集合 X 的幂集;R 表示区域连接演算 RCC-8,包括 8 种二元谓词$\{DC,EC,PO,EQ,TPP,NTPP,TPPi,NTPPi\}$.其中 $DC(x_1,x_2)$ 表示区域 x_1 与区域 x_2 相离

(disconnected);$EC(x_1,x_2)$表示区域 x_1 与区域 x_2 外部相切(externally connected);$PO(x_1,x_2)$表示区域 x_1 与区域 x_2 部分重叠(partially overlap);$EQ(x_1,x_2)$表示区域 x_1 与区域 x_2 相等(equal);$TPP(x_1,x_2)$表示区域 x_1 与区域 x_2 内包含相切(tangential proper part);$NTPP(x_1,x_2)$表示区域 x_1 与区域 x_2 非相切内包含(nontangential proper part);$TPPi(x_1,x_2)$,$NTPPi(x_1,x_2)$分别是 $TPP(x_1,x_2)$和($NTPP(x_1,x_2)$的逆操作.对于无人驾驶车辆来说,车辆占用区域不能相互包含,最危险的情况是外部相切,所以本文不考虑内切包含和非内切包含的空间关系,也就是不考虑相切内包含和非相切内包含空间关系以及逆关系{TPP,$NTPP$,$TPPi$,$NTPPi$},仅使用空间区域操作算子{DC,EC,PO,EQ}作为空间推理和表示占用空间的安全卫士条件.

无人驾驶车辆之间可以通信并且决定什么时间怎样做出决策,并且安全的无人驾驶系统需要具备可以解释的决策行为.这些决策可以帮助选择适合的行为并且能预测危险行为,无人系统决策涉及到安全攸关的时间和空间关系建模.系统行为可以描述为时钟关系随着空间变化而不断同步演化,文章给出了时空同步约束系统的抽象架构,如图 2 所示.假设在时刻 t_i 处,对于 2 种空间区域 x_1 和 x_2,存在一种 2 个空间区域相离的空间关系 $DC(x_1,x_2)$;在另外时刻 t_j 处,存在另外一种空间区域外部相切的空间关系 $EC(x_1,x_2)$;而在时刻 t_k 处,存在空间区域部分重叠的空间关系 $PO(x_1,x_2)$,两车占用的时空区域可能存在部分重叠的空间关系.对于无人驾驶车辆来说,空间区域相切和空间区域部分重叠是危险的空间位置关系,需要在决策模块提前避免发生碰撞的时空关系出现.无人驾驶系统状态和动作迁移不仅伴随着空间关系的变化,还伴随着严格的时钟关系的迁移.

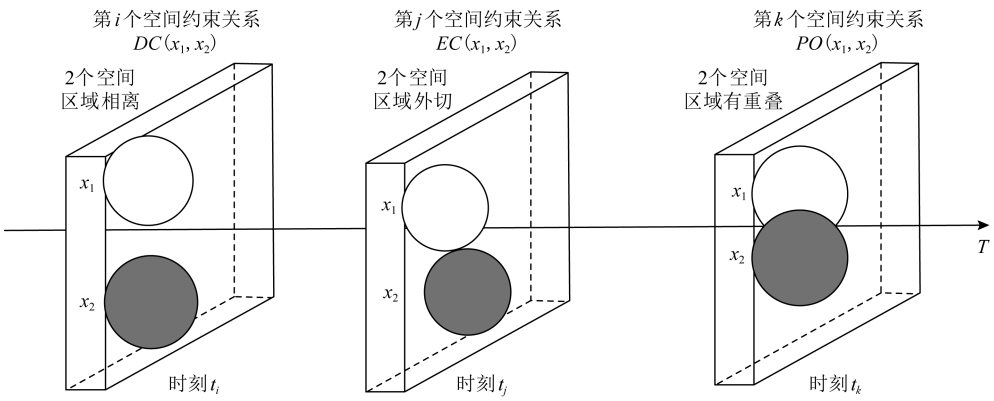


Fig. 2 An abstract architecture of spatio-clock synchronous constraint systems

图 2 时空同步约束系统的抽象架构

2 时空同步约束规约语言

为了安全决策和实时时空逻辑推理,本文给出无人驾驶交通快照、无人驾驶时空包络、和特定领域专用的时空同步约束规约语言的定义.

定义 5. 无人驾驶交通快照(traffic snapshot).无人驾驶交通快照可以描述为七元组 $TS=(res,clm,pos_x,pos_y,spd,acc,td)$,其中每一个元素分别代表无人驾驶系统相关的函数.

1) $res:\mathbb{N}\to P(\mathbb{L})$,车辆标识符 $I\in\mathbb{N}$,$res(I)\subseteq\mathbb{L}$ 表示车辆 I 占用的道路编号集合;

2) $clm:\mathbb{N}\to P(\mathbb{L})$,车辆标识符 $I\in\mathbb{N}$,$clm(I)\subseteq\mathbb{L}$ 表示车辆 I 申请占用的道路编号集合;

3) $pos_x:\mathbb{N}\to\mathbb{R}$,车辆标识符 $I\in\mathbb{N}$,$pos_x(I)$表示车辆在占用道路或者申请道路上的纵向位置信息;

4) $pos_y:\mathbb{N}\to\mathbb{R}$,车辆标识符 $I\in\mathbb{N}$,$pos_y(I)$表示车辆在占用道路或者申请道路上的横向位置信息;

5) $spd:\mathbb{N}\to\mathbb{R}$,车辆标识符 $I\in\mathbb{N}$,$spd(I)$表示车辆的纵向速度信息;

6) $acc:\mathbb{N}\to\mathbb{R}$,车辆标识符 $I\in\mathbb{N}$,$acc(I)$表示车辆的纵向加速度信息;

7) $td\in\{-1,0,1\}$表示车辆的驾驶方向,分别表示车辆右转、直行和左转.

所有交通快照组成的集合用$\mathbb{TS}$表示,$TS\in\mathbb{TS}$.

定义 6. 无人驾驶时空包络(spatio-clock envelope).

对于任意车辆 $I\in\mathbb{N}$ 和交通快照 TS,无人驾驶时空包络可定义为 $SCE=(pid,bd,L,I,V_x,V_y)_{clk}$,其中 $pid\in\mathbb{R}$ 是提前规定的最小距离(prescribed interval distance),包含了车辆的长度和规定的最小安全距离;$bd\in\mathbb{R}$ 表示刹车距离(braking distance),描述了车辆从开始刹车到最终停止的距离,对于自车车辆 ego 和刹车时间 t,定义刹车距离为

$$bd=spd(ego)\times t+\frac{1}{2}\times acc(ego)\times t^2.$$

加速度 $acc:\mathbb{N}\to\mathbb{R}$,安全距离(safety distance)是刹车距离加规定间隔距离 $sd=bd+pid$.$L=[l,n]\subseteq\mathbb{L}$ 表示时空安全包络包含的车道编号集合.

$V_x=[pos(I)-h,pos(I)+h]\subseteq\mathbb{R}$ 和 $V_y=[0,w]\subseteq\mathbb{R}$ 分别表示在时刻 clk 和位置 (pos_x,pos_y) 处,车辆所能观测到的纵向区域和横向区域.

定义 7. 时空同步约束规约语言.时空同步约束规约语言(spatio-clock synchronous constraint specification language, SCSL)的语法定义:

$$\phi_{\mathrm{SCSL}}::=\mathrm{true}\mid\neg\phi\mid\phi_1\wedge\phi_2\mid\exists v:\phi_1\mid u_{(l,t)}=v_{(l',t)}\mid re(A)_{(l,t)}\mid cl(A)_{(l,t)}\mid e_{(l,t)}(l',\delta)\mid send^c_{(l,t)}(m)Rget^c_{(l,t)}(m).$$

对于时空约束同步系统 $\mathcal{M}$,规约技术中命题公式解释为

1) $\mathcal{M}\models\mathrm{true}$ 表示系统在任何情况下都满足.

2) $\mathcal{M}\models\neg\phi\Leftrightarrow\mathrm{not}\ \mathcal{M}\models\phi$ 表示如果 $\neg\phi$ 满足,当且仅当 ϕ 不满足.

3) $\mathcal{M}\models\phi_1\wedge\phi_2\Leftrightarrow(\mathcal{M}\models\phi_1)\wedge(\mathcal{M}\models\phi_2)$ 表示 $\phi_1\wedge\phi_2$ 满足,当且仅当 ϕ_1 和 ϕ_2 同时满足.

4) $\mathcal{M}\models\exists v:\phi_1\Leftrightarrow\exists\alpha:\mathcal{M}\oplus\{v\mapsto\alpha\}\models\phi_1$ 表示 $\exists v:\phi_1$ 满足,当且仅当变量 v 取值为 α 时,ϕ_1 满足.

5) $\mathcal{M}\models u_{(l,t)}=v_{(l',t)}\Leftrightarrow\upsilon(u_{(l,t)})=\upsilon(v_{(l',t)})$ 表示如果 $u_{(l,t)}=v_{(l',t)}$ 满足,当且仅当 u,v 在同一时刻 t 不同空间位置 l 和 l' 处的变量值相等.

6) $\mathcal{M}\models re(A)_{(l,t)}\Leftrightarrow res(\upsilon(A))_{(l,t)}=L$ 表示 $re(A)_{(l,t)}$ 满足,当且仅当标识为 A 的车辆在时刻 t 和空间位置 l 处占用了路段 L.

7) $\mathcal{M}\models cl(A)_{(l,t)}\Leftrightarrow clm(\upsilon(A))_{(l,t)}=L$ 表示 $cl(A)_{(l,t)}$ 满足,当且仅当标识为 A 的车辆在时刻 t 和空间位置 l 处申请了路段 L.

8) $\mathcal{M}\models e_{(l,t)}(l',\delta)\Leftrightarrow\mathcal{M}_{(l,t)}\xrightarrow{e}\mathcal{M}_{(l',t+\delta)}$ 表示 $e_{(l,t)}(l',\delta)$ 满足,当且仅当无人自主事件 e 在空间位置 l 处和时刻 t 发生,经过一段时间 δ 后,车辆从空间位置 l 移动到位置 l'.

9) $\mathcal{M}\models send^c_{(l,t)}(m)Rget^c_{(l,t)}(m)\Leftrightarrow$

① $send^c_{(l,t)}(m)\prec get^c_{(l,t)}(m)$, $\forall i\in\mathbb{N}^+$, $(H_\sigma(send^c_{(l,t)}(m),i)=0\wedge H_\sigma(get^c_{(l,t)}(m),i)=0)\vee(H_\sigma(send^c_{(l,t)}(m),i)>H_\sigma(get^c_{(l,t)}(m),i))$;

② $send^c_{(l,t)}(m)\ \sharp\ get^c_{(l,t)}(m)$, $\forall i\in\mathbb{N}^+$, $send^c_{(l,t)}(m)\notin\sigma(i)\vee get^c_{(l,t)}(m)\notin\sigma(i)$;

③ $send^c_{(l,t)}(m)\subseteq get^c_{(l,t)}(m)$, $\forall i\in\mathbb{N}^+$, $send^c_{(l,t)}(m)\in\sigma(i)\to get^c_{(l,t)}(m)\in\sigma(i)$;

④ $send^c_{(l,t)}(m)\equiv get^c_{(l,t)}(m)$, $\forall i\in\mathbb{N}^+$, $send^c_{(l,t)}(m)\in\sigma(i)=get^c_{(l,t)}(m)\in\sigma(i)$;

⑤ $send^c_{(l,t)}(m)\preccurlyeq get^c_{(l,t)}(m)$, $\forall i\in\mathbb{N}^+$, $H_\sigma(send^c_{(l,t)}(m),i)\geqslant H_\sigma(get^c_{(l,t)}(m),i)$;

⑥ $send^c_{(l,t)}(m)\sim get^c_{(l,t)}(m)$, $\forall i\in\mathbb{N}^+$, $send^c_{(l,t)}(m)(i)\prec get^c_{(l,t)}(m)(i)\wedge get^c_{(l,t)}(m)(i)\prec send^c_{(l,t)}(m)(i+1)$.

表示 $send^c_{(l,t)}(m)Rget^c_{(l,t)}(m)$ 满足,即在空间位置 l 处和时刻 t,消息 m 通过通道 c 的发送和接收满足关系 $R\subseteq Rel_{\mathrm{CCSL}}=\{\prec,\sharp,\subseteq,\equiv,\preccurlyeq,\sim\}$,也就是①优先关系:时钟 $send^c_{(l,t)}(m)$ 的触发快于时钟 $get^c_{(l,t)}(m)$ 的触发;②互斥关系:若时钟 $send^c_{(l,t)}(m)$ 的触发,则时钟 $get^c_{(l,t)}(m)$ 不触发,反之亦然;③子时种关系:在任意时刻,若时钟 $send^c_{(l,t)}(m)$ 触发,则时钟 $get^c_{(l,t)}(m)$ 必然触发;④同时关系:时钟 $send^c_{(l,t)}(m)$ 和时钟 $get^c_{(l,t)}(m)$ 同时触发;⑤因果关系:时钟 $send^c_{(l,t)}(m)$ 的触发不慢于时钟 $get^c_{(l,t)}(m)$ 的触发;⑥交替关系:时钟 $send^c_{(l,t)}(m)$ 和时钟 $get^c_{(l,t)}(m)$ 交替触发.

3 时空同步约束自动机和时空迁移条件

基于第 2 节提出的领域特定的时空规约语言,本节给出系统行为建模语言,即时空同步自动机(spatio-clock synchronous automata, SCSA),然后给出无人驾驶行为的转换条件,以及时空安全攸关转换系统的卫士条件和不变式.

定义 8. 时间空间同步约束自动机.该模型语法定义为七元组 $SCSA=(Q,q_0,\Sigma,Varb,E,Grd,Inv)$,其中:

1) Q 是系统状态的集合,每一个状态可以看作是系统控制图中的一个顶点;

2) q_0 是系统的初始状态;

3) Σ 是无人驾驶控制动作的集合;

4) $Varb$ 是系统变量的集合,有 4 类变量,分别是离散变量 $dVar$,连续变量 $ctVar$,时钟变量 $ckVar$,和空间位置变量 $sVar$;

5) $E\subseteq Q\times Grd\times\Sigma\times 2^{Varb}\times Q$ 是控制图中连接状态之间边的集合,E 代表迁移关系;

6) Grd 代表迁移卫士条件,Σ 代表迁移动作,2^{Varb} 需要重置的变量;

7) $Inv:Q\rightarrow\psi(Varb)$ 表示关于变量的不变式.

时空同步约束自动机中的状态 $Q=(n,v,clk,sp)$ 包括状态名称 n,变量名称 v,时钟以及时钟关系 clk,和空间位置及其关系 sp.对于迁移关系中的环境标签 λ,触发事件 $evt(\lambda)$,卫士条件 $grd(\lambda)$,和迁移动作 $act(\lambda)$,

$$(n,v,clk,sp)\xrightarrow{evt(\lambda)\wedge grd(\lambda)\wedge act(\lambda)}(n',v',clk',sp')$$

表示当观测到触发事件 $evt(\lambda)$,迁移卫士条件 $grd(\lambda)$ 满足,那么迁移动作 $act(\lambda)$ 将会被执行.

定义 9. 无人驾驶控制器动作.无人驾驶控制器动作(autonomous controller action, ACA)可以定义为 $ACA::=c(A,\psi_L)\mid wd\ c(A)\mid r(A)\mid wd\ r(A,\psi_L)\mid\tau$,其中 $\psi_L::=n\mid l_1\mid l_1+l_2\mid l_1-l_2, n\in\mathbb{N}, l_1, l_2\in LVar, c(A,l)$ 表示车辆 A 申请道路 l.动作 $wd\ c(A)$ 表示车辆 A 撤销申请的道路.动作 $r(A)$ 表示车辆 A 占用之前申请的道路.动作 $wd\ r(A,l)$ 表示车辆 A 撤销占用除了道路编号 l 以外的道路.τ 表示空动作,允许没有任何动作的交通迁移.

结合定义 8 和定义 9,给出时空同步自动机操作语义和控制器动作的例子,如图 3 所示.初始状态 Initial,状态 q_1 是碰撞检测状态 *ReserveOverlapCheck*,当接收到前方车辆同意超车 Agreement 信号,并且所有相关车辆和当前车辆都不存在可能碰撞时,即 $\forall c, roc(ego,c)==$false,并且申请道路在道路线标号范围内 $n+1\leqslant N$,状态迁移触发事件满足,卫士条件也满足,需要执行迁移动作,重置时钟变量 x,并且执行申请临近道路动作 $c(ego,n+1)$,接下来状态从 q_1 迁移到潜在碰撞检测状态 q_2,即申请道路空间重叠检测 *ClaimOverlapCheck*.在状态 q_2,如果车辆接收到拒绝信号 Decline,或者存在潜在碰撞的车辆,即满足 $\exists c, coc(ego,c)==$true,则需要重置时钟变量 x,并且执行撤销申请道路的动作 $wd\ c(ego)$.

下面给出时空同步迁移的定义.

定义 10. 时空同步迁移(spatio-clock synchronous transitions).对于任何车辆 $I\in\mathbb{N}$,经过时间 t 后,速度和纵向位置的演化迁移可以根据:

$$\begin{aligned}TS\xrightarrow{t}TS'\Leftrightarrow TS'=&(res,clm,pos'_x,pos_y,\\&spd',acc,td)\wedge\forall I\in\mathbb{N}:spd'=\\&spd(I)+acc(I)\times t\wedge pos'_x(I)=\\&pos_x(I)+spd(I)\times t+\frac{1}{2}acc(I)\times t^2.\end{aligned}\tag{6}$$

对于任何车辆 $I\in\mathbb{N}$ 和道路编号 $n\in\mathbb{L}$,车辆可以申请邻近道路 $n+1$ 或者 $n-1$,并且为了保证无人驾驶的安全性和负责任,车辆必须服从交通规则,例如在右侧行驶的国家(比如美国、中国和除了英国、澳大利亚以及其他英国殖民地国家以外的大部分国家),大部分国家只允许右侧行驶,但是在超车时需要申请左侧道路超车行驶,可以根据式(7)执行申请道路迁移和式(8)执行撤销申请道路的迁移:

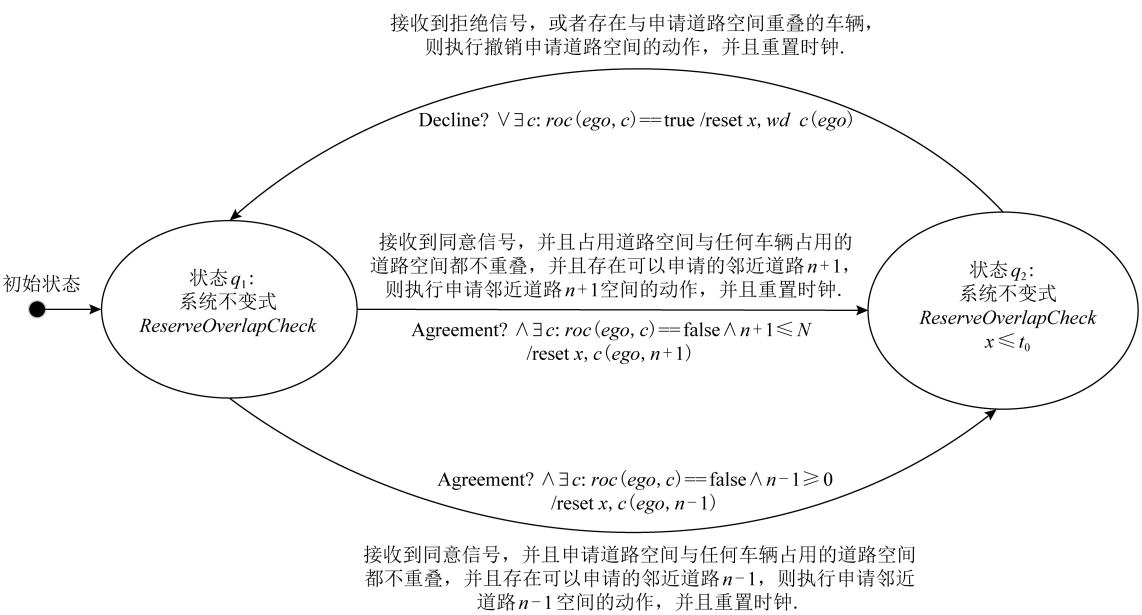


Fig. 3 An example of spatio-clock synchronous controller

图 3 时空同步自动机控制器例子

$$TS \xrightarrow{c(I,n)} TS' \Leftrightarrow TS' = (res, clm', pos_x, pos_y, spd, acc, td') \wedge (|clm(I)| = 0) \wedge (|res(I)| = 1) \wedge (td' = 1) \wedge (\{n+1, n-1\} \cap res(I) \neq \varnothing) \wedge (clm' = clm \oplus \{I \mapsto \{n\}\}), \quad (7)$$

$$TS \xrightarrow{wd\ c(I)} TS' \Leftrightarrow TS' = (res, clm', pos_x, pos_y, spd, acc, td') \wedge (td' = 0) \wedge (clm' = clm \oplus \{I \mapsto \varnothing\}). \quad (8)$$

重载算子$\oplus$含义如下,$\xi \oplus \{v \mapsto \alpha\}$可得$\{v \mapsto \alpha\}$中所有的关系,并且加上除去第一个元素在关系$\{v \mapsto \alpha\}$中,同时也在关系$\xi$中的关系后的所有剩余$\xi$中的关系.

对于任何车辆$I \in \mathbb{N}$和道路编号$n \in \mathbb{L}$,根据式(9)车辆占用之前申请或者占用的道路编号,根据式(10)车辆也可以撤销占用除了道路编号n以外的道路:

$$TS \xrightarrow{r(I)} TS' \Leftrightarrow TS' = (res', clm', pos_x, pos_y, spd, acc, td) \wedge (clm' = clm \oplus \{I \mapsto \varnothing\}) \wedge (res' = res \oplus \{I \mapsto res(I) \bigcup clm(I)\}), \quad (9)$$

$$TS \xrightarrow{wd\ r(I,n)} TS' \Leftrightarrow TS' = (res', clm, pos_x, pos_y, spd, acc, td) \wedge (res' = res \oplus \{I \mapsto \{n\}\}) \wedge (n \in res(I) \wedge |res(I)| = 2). \quad (10)$$

对于任何车辆$I \in \mathbb{N}$和加速度$acc \in \mathbb{R}$,加速度可以根据式(11)进行演化:

$$TS \xrightarrow{acc(I,a)} TS' \Leftrightarrow TS' = (res, clm, pos_x, pos_y, spd, acc', td) \wedge (acc' = acc \oplus \{I \mapsto a\}). \quad (11)$$

4 时空同步约束制导的强化学习方法

4.1 时空同步约束系统的安全迁移条件

根据第3节介绍的时空同步自动机和迁移条件,本节主要介绍无人驾驶安全卫士条件和结合形式化时空同步约束的强化学习方法.提出满足系统安全需求的奖励函数设置方法,重点关注时间空间同步约束的卫士条件和不变式.

定义 11. 无人驾驶安全卫士(autonomous driving safety guards). 根据定义4,时空安全卫士条件(即不发生碰撞)表示为申请和占用时空区域不发生重叠,可以检验安全性条件来确保没有交通事故发生(前提是车辆之间的通信不会发生延迟,车辆的感知器和其他硬件不会发生故障等).

$$safetyOverlap(c_1, c_2) = (c_1.pos_x \leqslant c_2.pos_x \leqslant c_1.pos_x + c_1.sd) \vee (c_1.pos_y \leqslant c_2.pos_y \leqslant c_1.pos_y + c_1.sd_y) \vee (c_2.pos_x \leqslant c_1.pos_x \leqslant c_2.pos_x + c_2.sd) \vee (c_2.pos_y \leqslant c_1.pos_y \leqslant c_2.pos_y + c_2.sd_y). \quad (12)$$

$safetyOverlap(c_1, c_2)$为真说明车辆c_1和c_2占用道路空间存在相交的区域.这时安全卫士条件意味着任何两辆车辆之间不存在相交的时空区域,可表示为

$$safe = \forall c_1, c_2 : ((c_1 = c_2) \vee (res(c_1) \bigcap res(c_2) = \varnothing) \vee \neg safetyOverlap(c_1, c_2)). \quad (13)$$

或者根据时空同步约束语言SCSL,可以形式化规约为

$$safe = \forall c_1, c_2 : ((c_1 = c_2) \vee (res(c_1) \bigcap res(c_2) = \varnothing) \vee \neg (re(c_1)_{(l,t)} \wedge re(c_2)_{(l,t)})). \quad (14)$$

占用时空重叠检测(reservation overlap check)$roc(c_1, c_2)$用来阻止任意两辆车占用重叠的时空区域,表示为

$$roc(c_1, c_2) = \exists c_1, c_2 : ((c_1 \neq c_2) \wedge (res(c_1) \bigcap res(c_2) \neq \varnothing) \wedge safetyOverlap(c_1, c_2)). \quad (15)$$

根据时空同步约束语言SCSL,安全卫士约束可以形式化规约为

$$roc(c_1, c_2) = \exists c_1, c_2 : (re(c_1)_{(l,t)} \wedge re(c_2)_{(l,t)}). \quad (16)$$

申请时空重叠检测(claiming overlap check)$coc(c_1, c_2)$阻止任意两辆车辆占用或者申请时空区域重叠.如果车辆c_1想把申请信息转换为占用道路信息并且安全地进行换道,车辆c_1需要检查自身申请的时空区域与另外一辆车c_2申请或者占用的道路空间是否有重叠,参照:

$$coc(c_1, c_2) = \exists c_1, c_2 : ((c_1 \neq c_2) \wedge ((clm(c_1) \bigcap res(c_2) \neq \varnothing) \vee (clm(c_1) \bigcap clm(c_2) \neq \varnothing) \wedge safeOverlap(c_1, c_2)). \quad (17)$$

根据时空同步约束语言SCSL,安全卫士约束可以形式化规约为

$$coc(c_1, c_2) = \exists c_1, c_2 : (cl(c_1)_{(l,t)} \wedge (re(c_2)_{(l,t)} \vee cl(c_2)_{(l,t)})). \quad (18)$$

4.2 形式化约束制导的安全强化学习框架

强化学习无法从环境交互的经验中学习到奖励函数,奖励函数必须由人工编写,而编写奖励函数具有2个挑战:1)感知数据不能直接用于奖励函数的输入参数(比如像素等),系统需要对数据进行抽象来获得强化学习中的状态空间和动作空间;2)复杂的学习任务需要分解为多个子学习任务(sub-goal task),子任务奖励不仅与当前的状态和动作有关系,还与历史状态的状态和动作有关,本文称为非Markov性质的奖励,所以系统设计人员需要对子学习任务的时序关系进行严格的形式化约束,否则不能达到安全累计回报的最优值.为了解决上述问题,

需要构建一个面向特定领域的状态-动作空间词汇表,本文给出将智能体学习经验(s,a,s')与状态-动作空间词汇联系起来的命题标签函数的定义.

定义 12. 状态-动作空间标签函数.假定集合 $\mathcal{P}$ 是命题符号集合,定义标签函数为 $\mathcal{L}:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to 2^{\mathcal{P}}$,该函数定义了从学习经验$(s,a,r,s')$到状态赋值之间的映射,其中 $\mathcal{P}$ 是包含系统中的状态空间和动作空间的命题集合.

该标签函数标注了强化学习中状态和动作的系统迁移,首先本文给出无人驾驶系统相关的状态空间和动作空间的抽象描述.状态空间可以描述为

$$S=(num_{\text{lane}},num_{\text{veh}},id_{\text{veh}},rl_{id},cl_{id},vr,ar,v_{id},a_{id},pos_x,pos_y,dd_{id},l_{lid},w_{lid},pd_x,pd_y)\in\mathcal{S},$$

其中,num_{lane}是道路的数量,num_{veh}是安全区域内车辆的数量,id_{veh}表示车辆的标识符,rl_{id}表示车辆占用的道路编号集合,cl_{id}表示车辆申请的道路编号集合,vr 表示车辆的速度区间$[v_{\min},v_{\max}]$,ar 表示车辆的加速度区间$[a_{\min},a_{\max}]$,v_{id}表示车辆 id 的速度,a_{id}表示车辆的加速度,pos_x 表示车辆的纵向位置,pos_y 表示车辆的横向位置,dd_{id}表示车辆的驾驶方向,l_{lid}表示道路的长度,w_{lid}表示道路的宽度,pd_x 表示两车纵向最小间隔,pd_y 表示横向最小间隔.动作空间为

$$A=(\text{cl\_left},\text{cl\_right},wd\_cl,res,wd\_res,\text{keep\_lane},\text{brake},\text{hard\_brake},\text{acc},\text{maintain},roc,coc,\text{overtake})\in\mathcal{A},$$

其中,cl_left 表示申请左转,cl_right 表示申请右转,wd_cl 表示不满足申请条件时撤销申请,res 表示占用道路$(r(I))$,wd_res 表示撤销道路占用$(wd\ r(I,n))$,keep_lane 表示保持当前道路,brake 表示刹车减速,hard_brake 表示遇到紧急情况时紧急刹车,acc 表示加速行驶,maintain 表示保持当前速度行驶,roc 表示占用道路时空重叠检测$(roc(c_1,c_2))$,coc 表示申请道路时空重叠检测$(coc(c_1,c_2))$,overtake 表示超车.

本文提出安全强化学习的研究框架如图 4 所示.智能体观测到的环境状态 $\mathcal{S}_t$,深度强化学习智能体根据环境输入,经过卷积层(convolutional layers)和池化层(pooling layers)的处理,得到环境的状态特征,再经过全连接层(fully connected layers)对环境特征的处理,输出智能体的动作空间 $\mathcal{A}_t$,在安全攸关的无人驾驶系统中,系统设计者需要建立安全防护机制,保证车辆输出的动作是安全并且符合交通规则,此时加上形式化约束的安全动作防护模块,以此来确保智能体输出安全的动作 $\mathcal{A}'_t$,然后交通环境反馈给智能体相应的奖励函数 R_t,由于任务的复杂性,每一项子任务之间的奖励函数需要满足一定的时序约束,这种带形式化约束制导的奖励函数设置,可以保证智能体获得安全的累计回报总体最优.在本文中,目前主要关注形式化约束制导的奖励函数设置,首先给出状态空间和动作空间的约束以及标签迁移函数的定义,然后给出解决非 Markov 性质奖励函数设置的时空同步约束奖励状态机定义和计算最优安全动作价值的算法.

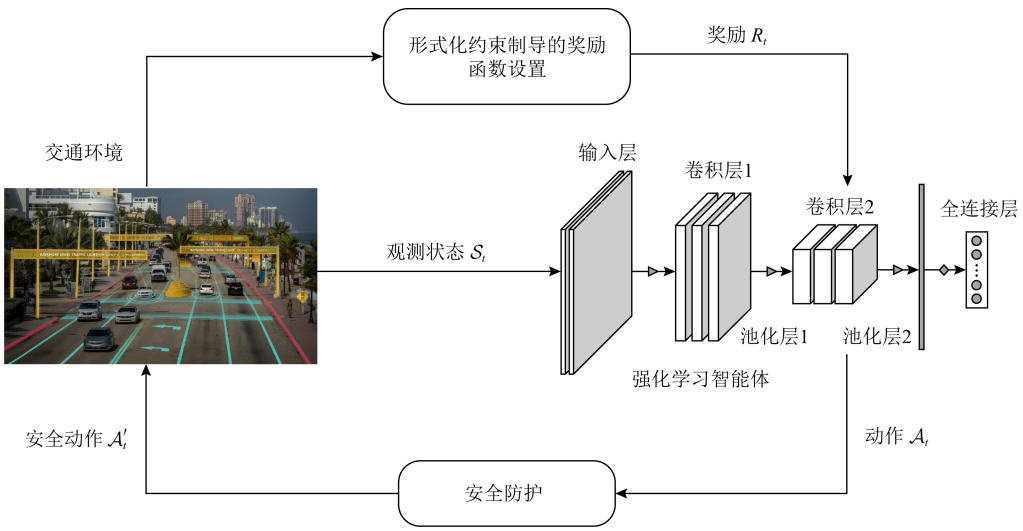


Fig. 4 The research framework ofsafeRL approach

图 4 安全强化学习方法的研究框架

为了保证安全累计回报的最大化,需要规定智能体获得奖励函数的时空同步约束条件,对于复杂的强化学习任务,各个子学习任务之间存在条件、循环和否定等时空约束关系,系统奖励不仅仅依赖Markov奖励决策过程,还受时空同步公式的约束,只有满足时空同步约束的动作才可以获得相应的奖励,记为 $r_t:(s_1,a_1)(s_2,a_2)\cdots(s_t,a_t)\models\phi$,也就是系统的历史经验 $\mathcal{H}_t$ 满足逻辑公式 ϕ(ϕ 是关于命题集合 $\mathcal{P}$ 的时空同步约束公式)时,系统才能获得奖励.$\mathcal{H}_t\models\phi$ 也可以理解为状态-动作空间迁移系统 $\mathcal{L}:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to 2^{\mathcal{P}}$ 满足公式 ϕ.下面给出时空同步约束的奖励函数状态机的定义,该状态机对应着完成相应任务时的形式化时空同步逻辑规约公式,从而解决了非Markov决策过程的奖励函数设置问题.

定义 13. 时空同步约束奖励状态机(spatio-clock synchronous constraint reward machine, SCSRM)[11]. 时空同步约束奖励状态机可以定义为九元组 $SCSRM\triangleq(\mathcal{S},\mathcal{A},\mathcal{P},\mathcal{L},U,u_0,\Sigma_{\mathcal{P}},\delta_u,\delta_r)$,给定非Markov决策过程中的元素 $\mathcal{S},\mathcal{A},\mathcal{P},\mathcal{L}$,命题符号集合 $\mathcal{P}$,系统状态集合 $\mathcal{S}$,以及动作集合 $\mathcal{A}$,状态-动作空间迁移系统 $\mathcal{L}:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to 2^{\mathcal{P}}$;$U$ 是该状态机中的状态集合;u_0 是该状态机的初始状态;$\Sigma_{\mathcal{P}}=2^{\mathcal{P}}$ 表示系统可以接收的输入状态和动作信息;状态机的状态转移函数 $\delta_u:U\times\Sigma_{\mathcal{P}}\to U$,表示在状态 $u_t\in U$ 时,满足标签迁移命题逻辑公式时,系统状态发生迁移,即 $u_{i+1}=\delta_u(u_i,\mathcal{L}(s_i,a_i,s_{i+1}))$;状态机的奖励函数 $\delta_r:U\times U\to[(S\times A)^+\times S\to\mathbb{R}]$ 表示在迁移关系中,时空同步约束状态机为非Markov奖励函数,即 $r_i=\delta_r(u_i,\mathcal{L}(s_i,a_i,s_{i+1}))=\delta_r(u_i,u_{i+1})$,根据定义,该奖励是从奖励状态机的初始状态到当前状态的满足逻辑公式 ϕ 的奖励之和.

下面给出计算带形式化约束制导的最优策略 Q 值更新计算算法,包括3层循环,第1层循环是关于训练过程回合数量(episodes T),行③是按照顺序选择奖励状态机里面的第 i 个子学习任务,u_p^i 表示子任务 i 的奖励状态机的第 p 个状态.第2层循环是在当前回合任务中迭代操作,当没有达到环境的终止状态时,进入第3层循环,利用safe_DDQN算法计算最优策略 Q 值,当智能体执行动作 a_t 时,从经验(s,a,s')中不断学习,对于给定的标签函数 $\mathcal{L}$,在状态 s' 处来自于 的命题或者事件成立,行⑧表示对于奖励状态机中的任何一个状态 u_j^i,在时空同步约束迁移标签函数作用下,系统迁移到状态 u_{j+1}^i,即 $u_{j+1}^i\leftarrow\delta_u(u_j^i,\mathcal{L}(s_i,a_i,s_{i+1}))$;行⑨是对应的非Markov奖励函数,具体奖励更新为 $r_i\leftarrow\delta_r(u_j^i,\mathcal{L}(s_i,a_i,s_{i+1}))=\delta_r(u_j^i,u_{j+1}^i)$;行⑬是时空同步约束制导的最优策略动作价值更新;行⑭是通过策略梯度下降来更新神经网络参数 w;行⑰是对时空同步约束奖励状态机、环境中的状态和动作进行更新操作.

算法 2. safe_DDQN求解最优动作价值算法.

输入:命题 $\mathcal{P}$、迁移 $\mathcal{L}$、SCSRM;

输出:形式化约束制导的动作价值估计.

① 初始化动作价值函数 $\hat{q}=Q(s,a;w)$;

② for $t=0$ to T do

/* 回合数内迭代寻找最优价值 */

③ 选择子学习任务 $i\leftarrow subtask(SCSRM,t)$;

④ 采样 $s_t,a_t,u_p^i\leftarrow u_0^i$;

⑤ while ($episode<T$ 且 s_t 不是终止状态)

⑥ 进入到下一个状态 s_{t+1} 并且返回奖励 r_t;

⑦ for $\hat{q}_j\in safe\_DDQN$ do

⑧ 更新时空同步约束奖励状态机状态;

⑨ 更新时空同步约束奖励状态机奖励;

⑩ if s_{t+1} 是终止状态 then

⑪ $\hat{q}_j\leftarrow\delta_r(u_j^i,u_{j+1}^i)(s_i,a_i,s_{i+1})$;

/* 当前价值为奖励 */

⑫ else

⑬ $\hat{q}_{j+1}\leftarrow\delta_r(u_j^i,u_{j+1}^i)(s_i,a_i,s_{i+1})+\gamma\max\limits_{a_{i+1}\in\mathcal{A}}\hat{q}_{j+1}(s_{t+1},a_{i+1};w_t)$;

/* 更新价值,执行新的动作 */

⑭ $w_{t+1}=w_t-\alpha(\hat{q}_j-\hat{q}_{j+1})\,\mathrm{d}t$;

/* 策略梯度下降更新 w */

⑮ end if

⑯ end for

⑰ $u_p^i\leftarrow\delta_u(u_p^i,\mathcal{L}(s_i,a_i,s_{i+1})),s_t\leftarrow s_{t+1},a_t\leftarrow a_{t+1}$; /* 状态机、状态和动作的更新 */

⑱ end while

⑲ end for

算法2在提高安全性的同时,比算法1多了一层for循环,算法嵌套的循环迭代次数增加了,因此算法复杂度提高为 $O(n^3)$,而算法1的算法复杂度是 $O(n^2)$,原因是为了满足非Markov性质的奖励,需要构建时空同步约束的奖励状态机,然后在形式化约束制导的奖励状态机里面选择子学习任务,用来在回合数规定内获得子学习任务满足形式化时空同步规约的最优动作价值,导致了算法复杂度的增加.同时,形式化时空同步约束制导的强化学习融合

了安全规则和先验经验等安全约束,智能体将不会去探索或者试错不正确的动作策略,在相对较少的训练次数内获得较好的奖励,所以加快了深度强化学习的效率,提高了算法的收敛速度.

对于无人驾驶系统,系统设计者希望车辆能够安全高效地行驶,同时车辆还需要遵守交通规则,所以考虑回报奖励函数时,需要综合考虑以下 3 个方面:1)车辆发生碰撞或者违反交通规则,假如奖励为 −2;2)如果车辆速度太慢,导致通行效率太低,假如速度小于 2 m/s,则奖励为 0;3)计算安全距离(safety distance),并设计奖励函数.同时需要限制训练回合的时间,因为在训练过程中,可能会出现车辆在某一个范围内转圈,需要限制回合时间,使得车辆进入后续回合的训练.基于回报奖励函数,可以设置回合时间,例如发生时空重叠或者违反交通规则;汽车速度小于最低速度;碰撞时间小于设定值等.

根据无人驾驶安全卫士条件的形式化描述,文章提出了基于时空同步约束的安全强化学习方法.该方法的主要思想是是指在强化学习训练和执行过程中,学习最优策略的过程需要满足形式化时空同步安全约束.系统很难及时预测前车的动作策略,车辆碰撞可能发生在自车与前车的距离小于安全距离时.基于这个原因,需要制定形式化安全距离约束来惩罚那些与前车距离小于安全距离的车辆.在定义 6 的基础上,再增加一个反应时间内行驶的距离 d_{re},跟驰车辆(following vehicle)的刹车距离[33] $d_{\mathrm{stop\_f}}$ 表示跟驰车辆在反应时间内以最大加速度和最小刹车加速度行驶的距离,计算为

$$d_{\mathrm{stop\_f}}=d_{\mathrm{re}}+d_{\mathrm{bra}}=v_{\mathrm{f}}t_{\mathrm{re}}+\frac{1}{2}a_{\mathrm{max,acc}}\times t_{\mathrm{re}}^{2}+\frac{(v_{\mathrm{f}}+t_{\mathrm{re}}a_{\mathrm{max,acc}})^{2}}{2a_{\mathrm{min,brake}}},\tag{19}$$

其中,v_{f} 表示跟驰车辆的速度,t_{re} 是车辆的反应时间,$a_{\mathrm{max,brake}}$ 是车辆紧急刹车(hard brake)的最大刹车加速度,$a_{\mathrm{max,acc}}$ 是后面跟驰车辆在反应时间内的最大加速度,$a_{\mathrm{min,brake}}$ 是反应时间过后前面车辆从开始刹车到车辆停止且前后两辆车没有发生碰撞的最小刹车加速度.

前面车辆(preceding vehicle)的刹车距离 $d_{\mathrm{stop\_p}}$ 表示车辆以初始速度 v_{p} 行驶,最大制动加速度 $a_{\mathrm{max,brake}}$ 刹车时车辆行驶的距离,计算为

$$d_{\mathrm{stop\_p}}=\frac{v_{\mathrm{p}}^{2}}{2a_{\mathrm{max,brake}}}.\tag{20}$$

车辆之间的最小距离 d_{min} 是后车刹车距离 $d_{\mathrm{stop\_f}}$ 与规定间隔 d_{pid} 之和,减去前车刹车距离 $d_{\mathrm{stop\_p}}$ 和两车间隔 d_{gap} 距离的和,车辆之间的距离必须满足式(21),才保证车辆不发生碰撞.

$$d_{\mathrm{min}}=(d_{\mathrm{stop\_f}}+d_{\mathrm{pid}})-(d_{\mathrm{stop\_p}}+d_{\mathrm{gap}})>0,\tag{21}$$

其中,d_{pid} 是规定的车辆之间的安全间隔距离,d_{gap} 是两车停止后的实际间隔距离.

违反安全距离的规定促使无人驾驶车辆在强化学习的过程中保持安全距离行驶.系统首先给出关于最小安全距离 d_{min} 的奖励反馈函数:

$$R_{\mathrm{distance}}=\begin{cases}\exp\left(\dfrac{-(d_{\mathrm{min}}-d_{\mathrm{pid}})^{2}}{10\times d_{\mathrm{pid}}}\right), & d_{\mathrm{min}}>0,\\ -2, & d_{\mathrm{min}}\leqslant 0.\end{cases}\tag{22}$$

利用安全强化学习训练车辆的过程中,车辆的速度不要超过允许的最大速度,同时希望车辆能够快速的到达目的地,所以系统设计人员需要设置关于训练车辆速度的奖励函数:

$$R_{\mathrm{speed}}=\begin{cases}\exp\left(\dfrac{-(v_{ego}-v_{\mathrm{max}})^{2}}{10}\right), & 0<v_{ego}\leqslant v_{\mathrm{max}},\\ -4, & v_{ego}>v_{\mathrm{max}},\\ -2, & v_{ego}\leqslant 0.\end{cases}\tag{23}$$

为了避免车辆偏离行驶车道撞向路边基础设施发生交通事故,本文设置车辆安全驾驶行为的正强化学习奖励函数,系统设计人员应鼓励车辆靠近路中间的道路行驶,用 $dis_{p_{y\_\mathrm{mid}}}$ 表示车辆的横向位置与道路中间横向位置的距离,为此设置车道保持的奖励函数:

$$R_{\mathrm{lane}}=\exp(-1.2\times dis_{p_{y\_\mathrm{mid}}}).\tag{24}$$

本文主要关注无人驾驶车辆强化学习智能体的安全性和通行效率,为此给出在时刻 t 的奖励函数:

$$R^{t}=R^{t}_{\mathrm{distance}}+R^{t}_{\mathrm{speed}}+R^{t}_{\mathrm{lane}}.\tag{25}$$

5 案例分析

5.1 案例需求分析

系统的需求描述如图 5 所示,当前道路上的自车车辆有超车意图时,首先需要打开转向灯,提示在时空安全包络内的周围车辆超车意图,此时有 3 件安全关键的事情需要确认:1)需要确认与同车道的车辆是否在安全距离内,也就是确认是否存在占用空间重叠;2)安全包络内的周围车辆观测到变道信号后,后面车辆不能加速;3)向前面车辆发送变道超车请求信息.前方车辆 preceding-vehicle 在接收到超车请求信号后,需要在一定时间(反应时间假设 2 s)内回复同意或者拒绝,如果自车车辆在反应时间内接收不到回复信息或者接收到拒绝信息,自车车辆

不能实施变道超车,需要保持原来的状态前进;如果与周围的申请道路车辆不满足安全距离约束,也就是存在申请道路的空间重叠,同样不能实施变道超车.需要同时满足在约定的反应时间内接收到同意信息和满足安全距离的情况下,才可以实施变道超车.超车动作可以分解成 2 次变道,首先是第 1 次变道到申请的道路,变道之后需要加速前进,在这个过程中,之前回复同意的车辆不能加速,当自车车辆的相对空间位置超过自车车辆前方车辆时,并且它们之间满足变道的安全距离时,自车车辆实施第 2 次变道,2 次变道超车的时间间隔不能超过超车规定时间.

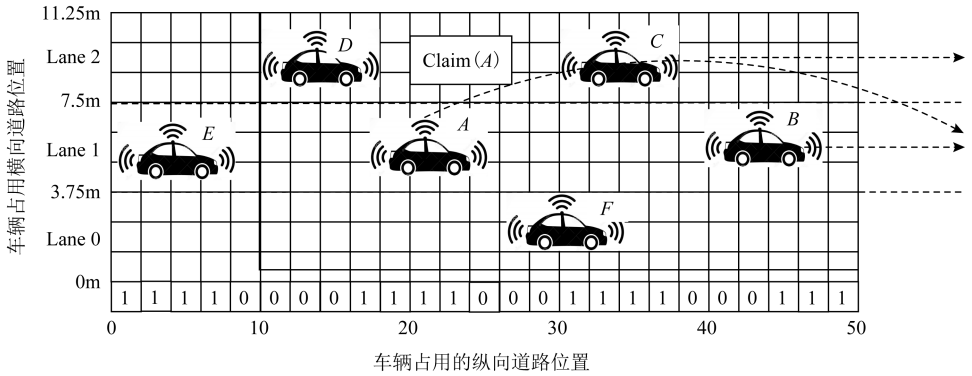


Fig. 5 An example of autonomous lane-change and overtaking in highway scenario

图 5 无人驾驶汽车高速场景下变道超车案例

建立时间空间同步自动机理论模型包含 3 个步骤:1)车辆左转信号 Turn_left 必须优先变道申请 Claiming_req,变道申请优先于互斥的时钟事件同意申请 Change_claim_agr 和不同意申请信号 No_change_claim_agr,该需求形式化规约为 $C_{\text{Turn\_left}} \prec C_{\text{Claiming\_req}} \prec (C_{\text{Change\_claim\_agr}} \sharp C_{\text{No\_change\_claim\_agr}})$.其次,当前面车辆回复同意超车信号后,占用道路碰撞检测 $roc(ego,c)$严格优先于占用道路动作 Reserving,而且占用道路动作又严格优先于前方车辆不能左转信号 No_turn_left,该需求可以形式化规约为 $roc(ego,c) \prec C_{\text{Reserving}} \prec C_{\text{No\_turn\_left}}$.2)申请邻近道路信号 $c(I,n)$与撤销申请信号交替发生 $wd\ c(I)$,该需求形式化规约为 $c(I,n) \sim wd\ c(I)$,同样地,占用道路事件 $r(I)$与撤销占用道路事件 $wd\ r(I,n)$交替发生,该需求形式化规约为 $r(I) \sim wd\ r(I,n)$.3)碰撞检测时钟 $roc(ego,c)$的发生不慢于申请变道的时钟 Claiming_req,2 个时钟满足因果关系,即 $roc(ego,c) \leqslant C_{\text{Claiming\_req}}$,同样地,潜在碰撞检测 $coc(ego,c)$触发互斥时钟占用之前申请的道路 Reserving 和存在潜在碰撞 Potentialcol,该需求可以形式化规约为 $coc(ego,c) \leqslant (C_{\text{Reserving}} \sharp C_{\text{Potentialcol}})$.

5.2 SUMO 仿真平台及实验设计分析

结合强化学习的 SUMO 仿真平台是一个开源、微观和连续的交通仿真软件包,支持动态路由的生成和输入,拥有可视化的图形界面,可以对时间和空间进行良好的定义和模拟,利用 TraCI 接口(traffic control interface)实现用 Python 语言进行模型开发与仿真,所以本文选择了 SUMO 作为仿真验证工

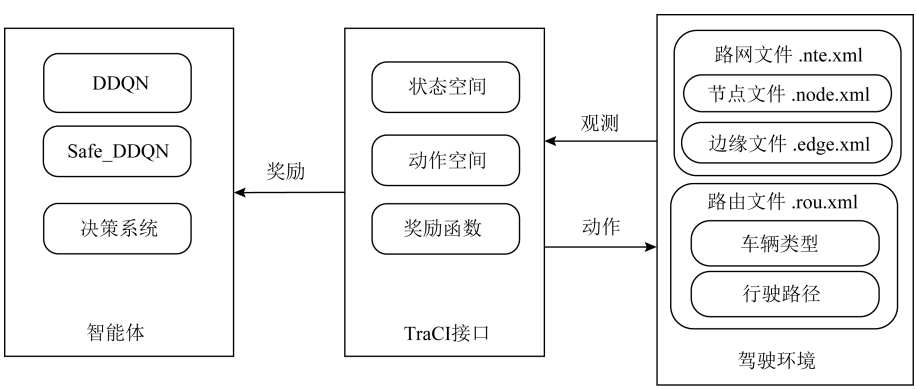


Fig. 6 Combine RL and SUMO simulation platform

图 6 部署强化学习算法的 SUMO 仿真平台

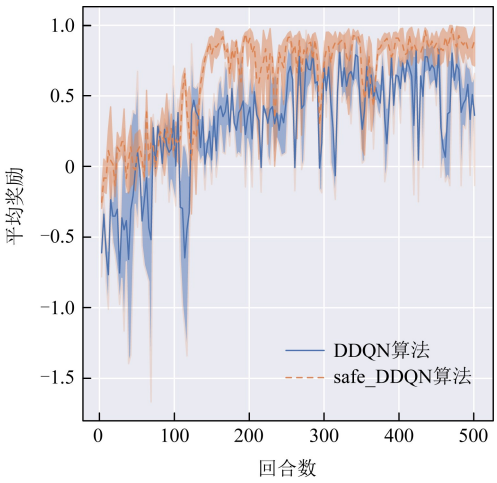
具,如图 6 所示系统由 3 个部分组成:1)驾驶环境,该模块路网文件(.net.xml)和路由文件(.rou.xml),路网文件描述了交通道路信息,包括节点文件(.node.xml)和连接边文件(.edge.xml);路由文件定义车辆行为、行驶路径和车流信息.2)TraCI 接口,它作为仿真平台和强化学习算法之间的接口,将仿真网络里的观测状态信息传递给算法,而后又将算法输出的高层动作传回给目标车辆智能体.3)智能体模块,借助 TensorFlow 库开发的 DDQN 算法和改进的 safe_DDQN 算法,实现智能体的智能安全决策.这里需要注意,DDQN 和 safe_DDQN 算法只输出高层控制动作,而对目标车辆的底层控制由 SUMO 模拟器实现.对于道路信息,SUMO 还可以通过 netconvert 程序将很多的第三方的路网文件转化为 SUMO 可读的文件,例如可以选定真实物理世界的 osm 格式的地图文件,直接生成一个可以运行仿真的 SUMO 路网文件;车辆的行为主要靠无人驾驶的感知模块获得,所以只要实际无人车系统的车辆感知系统满足道路运行测试条件的话,就可以将文中的方法部署在实际的无人驾驶车辆.

在 SUMO 中生成案例中的交通仿真场景如图 7 所示,交通场景中包含 3 条道路,车辆信息由路由文件生成,图 7 中的红色车辆是需要强化学习训练的自车车辆,黄色车辆是交通驾驶环境视窗中的其他周围车辆.

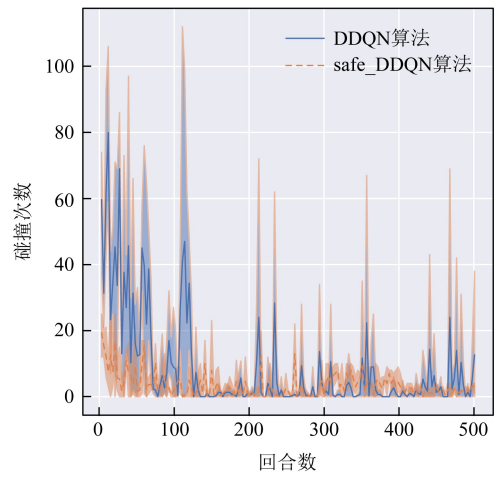
Fig. 7 Overtaking scenario screenshot in SUMO platform

图 7 SUMO 仿真超车场景快照

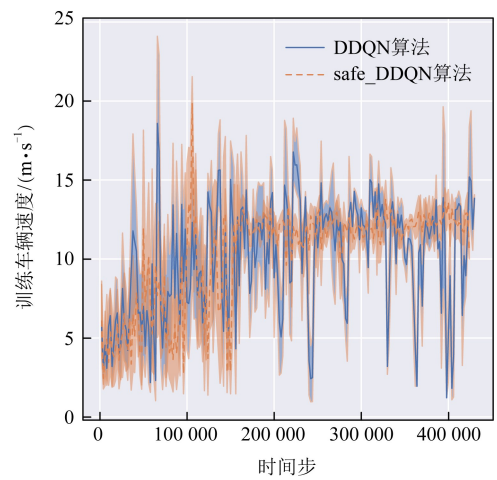
本文对 2 个算法分别进行了 500 回合的深度强化学习训练,其中每一回合表示自车车辆完成了一次在规定道路上的驾驶行为,并对仿真结果进行了比较.图 8(a)显示了算法在每一回合训练中获得的平均奖励,首先可以看到随着训练回合数的增加,2 个算法的平均奖励都在上升,但是 safe_DDQN 的平均奖励相比于普通的 DDQN 更高而且在较短的时间内趋于稳定,这也意味着本文提出的算法 safe_DDQN 更容易收敛.图 8(b)显示了算法指导的智能车在每一回合训练中发生的总碰撞次数,在 2 个算法中,智能车的碰撞次数随着训练回合数的增加而不断减少,说明智能车通过探索强化学习学得了正确的驾驶动作,减少了碰撞次数的发生,但是普通的



(a) 平均奖励值比较



(b) 总体碰撞次数的比较



(c) 车辆速度的比较

Fig. 8 Experimental results and analyses

图 8 实验结果对比分析

DDQN 算法不够稳定,在 400 回合后,仍有较高频率的碰撞发生,而 safe_DDNQ 此时已经几乎不会发生碰撞.在实验中设置每一个回合进行 1 000 时

间步(timesteps)的状态-动作空间采样,对应算法 1 行③和算法 2 行④,根据算法描述,有时候遇到终止状态仿真会提前终止回合的迭代,500 回合仿真大约有 40 万时间步,为了方便观察,每 20 个时间步选取一个数据,并绘制了如图 8(c)所示的智能体车辆的速度变化图.从图 8 中可以看到,safe_DDQN 算法指导的智能车,在 20 万时间步后,速度就趋于平稳,且始终维持在目标车速附近,而普通的 DDQN 算法指导的小车,速度波动始终很大,这不仅会影响乘车舒适度,还会带来安全隐患.由此可见,本文提出的 safe_DDQN 算法相比于普通的 DDQN 算法,在算法收敛速度、算法稳定性,以及其指导的智能车所带来的驾驶安全性、乘车舒适性方面都有较大的提高.

6 相关工作

首先介绍关于时间和空间形式化约束规约的相关工作.Chouhan 等人[34]利用形式化时间自动机模型和统计模型检测方法,为启发式无人驾驶交叉路口管理提供了相关的建模和验证技术.Cuer 等人[35]提出了从自然语言到形式化状态机的建模架构,并且不断删掉潜在的自然语言模糊性和不一致性.Briola 等人[36]用本体来描述空间关系,并且用 Prolog 来规约可解释的规则信息,方便系统设计者重复使用形式化的领域空间知识.多道路空间逻辑(multi-lane spatial logic)被用来建模无人驾驶领域的道路知识,包括乡村道路的应用[37]、城市道路的应用[13]和高速场景的应用[38].无人驾驶时间行为建模方面,文章[39]设计了随机混合时空规约语言和自动机,并提出了无人驾驶车辆发生碰撞时刻和位置的预测方法.Tan 等人[40]展示了空间时序事件建模的方法,该方法规约事件的触发不仅根据严格的时序和物理时间,而且还受空间关系的制约.上述研究关注了空间或者时间的建模,但是对于时空同步约束的自动驾驶系统来说,仍然缺乏领域相关的形式化规约语言,增强系统需求和任务的严格描述,减少自然语言的二义性,同时用来增强系统决策的安全性.

然后介绍结合安全约束的强化学习方法.对于安全攸关的无人驾驶系统,为了增强系统决策的安全性,结合形式化安全约束的强化学习技术得到越来越多的关注和研究.文献[6,41]结合可解释的基于规则的策略和黑盒的强化学习来实现系统的安全性和鲁棒性.纪守领等人[42]回顾了机器学习中的可解释性问题,分析了可解释性机器学习的安全性面临的挑战和研究方向.Krasowski 等人[43]通过扩展强化学习安全层来限制动作空间,利用基于集合的方法来增强强化学习的安全保障.Wachi 等人[44]提出了一种带约束 Markov 决策过程的强化学习方法,该方法通过扩展安全区域来不断学习安全约束.Gao 等人[45]设计了一个结合形式化线性时序规约(linear temporal logic, LTL)的奖励回报函数,可以更准确地估计损失函数的梯度和改善训练过程的稳定性.Wolf 等人[46]指出仅仅对专家知识进行建模是比较困难的,提出结合语义状态和交通规则来增强强化学习的安全性.Wang 等人[47]在无人驾驶任务决策方法中结合基于规则约束和强化学习的方法,实现了安全高效的变道行为.Chen 等人[48]提出了一种基于经验指导的深度学习多行动者-评论家算法,从优秀经验中学习指导网络,并对动作价值函数进行更新指导.Garcıa 和 Fernández[49]给出了安全强化学习的定义,并给出在系统学习和部署的过程中增加安全约束的相关文献综述.

7 总结与展望

针对基于强化学习的无人驾驶系统决策安全性问题,本文首先给出了时空同步轨迹的介绍,指出无人驾驶系统是时间空间安全攸关的系统,并给出了无人驾驶特定领域的时空同步约束规约语言.其次,无人驾驶的安全性主要是关心系统的时空是否重叠,本文基于时空同步约束语言和自动机的定义,展示了无人驾驶控制器和占用/申请(撤销占用/申请)空间的动作,规约了安全状态迁移的条件,给出了如何检测占用时空重叠和申请时空重叠的检测标准.然后,针对学习任务非 Markov 性质的奖励函数,本文给出系统状态-动作空间迁移系统的安全迁移条件,提出形式化时空同步约束奖励状态机,来提高获得强化学习非 Markov 性质奖励函数的安全性,进而来提高形式化约束制导的奖励函数设置的可解释性.最后通过无人驾驶高速场景下变道超车的案例,在 SUMO 仿真平台上验证所提方法的有效性.

在将来的研究工作中,首先需要优化强化学习安全约束的设计与训练,考虑交通规则对无人驾驶系统的影响,建立遵守交通规则的负责任的和安全的无人驾驶决策系统.其次,需要改进强化学习奖励回报函数的设计与训练,提升学习的效率和学习策

略的可解释性.最后,目前实验验证是在模拟的环境下进行,而且驾驶场景比较简单,在今后的工作中,需要在其他交通场景比如交叉路口和环岛等加强仿真实验,同时还需要在实际无人驾驶系统中部署结合安全约束的强化学习研究与应用,逐步推进安全强化学习在无人驾驶系统中的推广与实际应用.

作者贡献声明:王金永负责阅读文献和初稿写作,黄志球提供研究基金和科学问题,杨德艳提出算法思路和文献整理,Xiaowei Huang 提供实验方案和 safeAI 论文思路,祝义负责形式化理论与论文审查,华高洋负责实验设计与结果分析.

参 考 文 献

[1] Mirchevska B, Pek C, Werling M, et al. High-level decision making for safe and reasonable autonomous lane changing using reinforcement learning [C] //Proc of the 21st Int Conf on Intelligent Transportation Systems (ITSC). Piscataway, NJ: IEEE, 2018: 2156-2162

[2] Zhou Zhijie, Cao You, Hu Changhua, et al. The interpretability of rule-based modeling approach and its development [J]. Acta Automatica Sinica, 2021, 47(6): 1201-1216 (in Chinese) (周志杰, 曹友, 胡昌华, 等. 基于规则的建模方法的可解释性及其发展[J]. 自动化学报, 2021, 47(6): 1201-1216)

[3] Du Jin, Zheng Qinghua, Li Haifei, et al. The research of mining association rules between personality and behavior of learner under web-based learning environment [C] //Proc of the Int Conf on Web-Based Learning. Berlin: Springer, 2005: 406-417

[4] Zheng Qinghua, Liu Jun, Zeng Hongwei, et al. Knowledge forest: A novel model to organize knowledge fragments [J]. arXiv preprint, arXiv:1912.06825, 2019

[5] Fulton N, Platzer A. Safe reinforcement learning via formal methods: Toward safe control through proof and learning [C] //Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2018, 32(1)

[6] Likmeta A, Metelli A M, Tirinzoni A, et al. Combining reinforcement learning with rule-based controllers for transparent and general decision-making in autonomous driving [J]. Robotics and Autonomous Systems, 2020, 131: 103568. DOI: 10.1016/j.robot.2020.103568

[7] Hammond L, Abate A, Gutierrez J, et al. Multi-agent reinforcement learning with temporal logic specifications [C] //Proc of the 20th Int Conf on Autonomous Agents and MultiAgent Systems. New York: ACM, 2021: 583-592

[8] Li Xiao, Serlin Z, Yang Guang, et al. A formal methods approach to interpretable reinforcement learning for robotic planning [J]. Science Robotics, 2019, 4(37). DOI: 10.1126/scirobotics.aay6276

[9] ToroIcarte R, Klassen T Q, Valenzano R, et al. Teaching multiple tasks to an RL agent using LTL [C] //Proc of the 17th Int Conf on Autonomous Agents and MultiAgent Systems. New York: ACM, 2018: 452-461

[10] Icarte R T, Klassen T, Valenzano R, et al. Using reward machines for high-level task specification and decomposition in reinforcement learning [C] //Proc of the Int Conf on Machine Learning. New York: ACM, 2018: 2107-2116

[11] Camacho A, Icarte R T, Klassen T Q, et al. LTL and beyond: formal languages for reward function specification in reinforcement learning [C] //Proc of the 28th Int Joint Conf on Artificial Intelligence (IJCAI). San Francisco: Morgan Kaufmann, 2019, 19: 6065-6073

[12] Somenzi F, Trivedi A. Reinforcement learning and formal requirements [C] //Proc of the Int Workshop on Numerical Software Verification. Berlin: Springer, 2019: 26-41

[13] Schwammberger M. An abstract model for proving safety of autonomous urban traffic [J]. Theoretical Computer Science, 2018, 744: 143-169. DOI: 10.1016/j.tcs.2018.05.028

[14] Yu Jinke, Petnga L. Space-based collision avoidance framework for autonomous vehicles [J]. Procedia Computer Science, 2018, 140: 37-45. DOI: 10.1016/j.procs.2018.10.290

[15] Bochmann Gv, Hilscher M, Linker S, Olderog ER. Synthesizing and verifying controllers for multi-lane traffic maneuvers [J]. Formal Aspects of Computing, 2017, 29(4): 583-600

[16] Kamali M, Linker S, Fisher M. Modular verification of vehicle platooning with respect to decisions, space and time [C] //Proc of the Int Workshop on Formal Techniques for Safety-Critical Systems. Berlin: Springer, 2018: 18-36

[17] Arcile J, Devillers R, Klaudel H. VerifCar: A framework for modeling and model checking communicating autonomous vehicles [J]. Autonomous Agents and Multi-agent Systems, 2019, 33(3): 353-381

[18] Huang Li, Kang E Y. Formal verification of safety & security related timing constraints for a cooperative automotive system [C] //Proc of the 22nd Int Conf on Fundamental Approaches to Software Engineering. Berlin: Springer, 2019: 210-227

[19] Wang Jinyong, HuangZhiqiu, Huang Xiaowei, et al. Multiclock constraint system modelling and verification for ensuring cooperative autonomous driving safety [J]. Journal of Advanced Transportation, 2020. https://doi.org/10.1155/2020/8830752

[20] Ciancia V, Gilmore S, Grilletti G, et al. Spatio-temporal model checking of vehicular movement in public transport systems [J]. International Journal on Software Tools for Technology Transfer, 2018, 20(3): 289-311

[21] Zhang Yuanrui, Mallet F, Chen Yixiang. A verification framework for spatio-temporal consistency language with CCSL as a specification language [J]. Frontiers of Computer Science, 2020, 14(1): 105-129

[22] Sallab A E L, Abdou M, Perot E, et al. Deep reinforcement learning framework for autonomous driving [J]. Electronic Imaging, 2017, 2017(19): 70-76

[23] Baheri A, Nageshrao S, Tseng H E, et al. Deep reinforcement learning with enhanced safety for autonomous highway driving [C] //Proc of 2020 IEEE Intelligent Vehicles Symp (IV). Piscataway, NJ: IEEE, 2020: 1550-1555.

[24] Gaon M, Brafman R. Reinforcement learning with non-markovian rewards[C] //Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020, 34(4): 3980-3987

[25] Rong Jikun, Luan Nan. Safe reinforcement learning with policy-guided planning for autonomous driving [C] //Proc of 2020 IEEE Int Conf on Mechatronics and Automation (ICMA). Piscataway, NJ: IEEE, 2020: 320-326

[26] Kiran B R, Sobh I, Talpaert V, et al. Deep reinforcement learning for autonomous driving: A survey [J]. IEEE Transactions on Intelligent Transportation Systems, 2021. DOI: 10.1109/TITS.2021.3054625

[27] Shalev-Shwartz S, Shammah S, Shashua A. On a formal model of safe and scalable self-driving cars [J]. arXiv preprint, arXiv:1708.06374, 2017

[28] Qiao Zhiqian, Muelling K, Dolan J M, et al. Automatically generated curriculum based reinforcement learning for autonomous vehicles in urban environment [C] //Proc of 2018 IEEE Intelligent Vehicles Symp (IV). Piscataway, NJ: IEEE, 2018: 1233-1238

[29] Li Meng, Li Zhibin, Xu Chengcheng, et al. Deep reinforcement learning-based vehicle driving strategy to reduce crash risks in traffic oscillations [J]. Transportation Research Record, 2020, 2674(10): 42-54

[30] Mallet F, De Simone R. Correctness issues on MARTE/CCSL constraints [J]. Science of Computer Programming, 2015, 106: 78-92. DOI: 10.1016/j.scico.2015.03.001

[31] Randell D A, Cui Zhan, Cohn A G. A spatial logic based on regions and connection [C] //Proc of the 3rd Int Conf Principles of Knowledge Representation and Reasoning. San Francisco CA: Morgan Kaufmann, 1992: 165-176

[32] Clarke, Bowman L. A calculus of individuals based on "connection" [J]. Notre Dame Journal of Formal Logic, 1981, 22(3): 204-218

[33] Althoff M, Lösch R. Can automated road vehicles harmonize with traffic flow while guaranteeing a safe distance [C] //Proc of 2016 IEEE 19th Int Conf on Intelligent Transportation Systems (ITSC). Piscataway, NJ: IEEE, 2016: 485-491

[34] Chouhan A P, Banda G. Formalverification of heuristic autonomous intersection management using statistical model checking [J]. Sensors, 2020, 20(16): 4506. DOI: 10.3390/s20164506

[35] Cuer R, Piétrac L, Niel E, et al. A formal framework for the safe design of the autonomous driving supervision [J]. Reliability Engineering & System Safety, 2018, 174: 29-40

[36] Briola D, Mascardi V, Gioseffi M. OntoScene, A logic-based scene interpreter: implementation and application in the rock art domain [J]. Theory and Practice of Logic Programming, 2020, 20(4): 456-511

[37] Hilscher M, Linker S, Olderog E R. Proving safety of traffic manoeuvres on country roads [C] //Proc of the Conf on Theories of Programming and Formal Methods. Berlin: Springer, 2013: 196-212

[38] Hilscher M, Linker S, Olderog E R, et al. An abstract model for proving safety of multi-lane traffic manoeuvres [C] //Proc of Int Conf on Formal Engineering Methods. Berlin: Springer, 2011: 404-419

[39] Wang Jinyong, Huang Zhiqiu, Huang Xiaowei, et al. An accident prediction architecture based on spatio-clock stochastic and hybrid model for autonomous driving safety [J]. Concurrency and Computation: Practice and Experience, 2021: e6550. https://doi.org/10.1002/cpe.6550

[40] Tan Ying, Vuran M C, Goddard S. Spatio-temporal event model for cyber-physical systems [C] //Proc of the 29th IEEE Int Conf on Distributed Computing Systems Workshops. Piscataway, NJ: IEEE, 2009: 44-50

[41] Talamini J, Bartoli A, De Lorenzo A, et al. On the impact of the rules on autonomous drive learning [J]. Applied Sciences, 2020, 10(7): 1-14

[42] Ji Shouling, Li Jinfeng, Du Tianyu, et al. Survey on techniques, applications and security of machine learning interpretability [J]. Journal of Computer Research and Development, 2019, 56(10): 2071-2096 (in Chinese)
(纪守领, 李进锋, 杜天宇, 等. 机器学习模型可解释性方法、应用与安全研究综述[J]. 计算机研究与发展, 2019, 56(10): 2071-2096)

[43] Krasowski H, Wang Xiao, Althoff M. Safe reinforcement learning for autonomous lane changing using set-based prediction [C] //Proc of the 23rd IEEE Int Conf on Intelligent Transportation Systems (ITSC). Piscataway, NJ: IEEE, 2020: 1-7

[44] Wachi A, Sui Yanan. Safe reinforcement learning in constrained markov decision processes [C] //Proc of Int Conf on Machine Learning. New York: ACM, 2020: 9797-9806

[45] Gao Qitong, Hajinezhad D, Zhang Yan, et al. Reduced variance deep reinforcement learning with temporal logic specifications [C] //Proc of the 10th ACM/IEEE Int Conf on Cyber-Physical Systems. New York: ACM, 2019: 237-248

[46] Wolf P, Kurzer K, Wingert T, et al. Adaptive behavior generation for autonomous driving using deep reinforcement learning with compact semantic states [C] //Proc of 2018 IEEE Intelligent Vehicles Symp (IV). Piscataway, NJ: IEEE, 2018: 993-1000

[47] Wang Junjie, Zhang Qichao, Zhao Dongbin, et al. Lane change decision-making through deep reinforcement learning with rule-based constraints [C] //Proc of 2019 Int Joint Conf on Neural Networks (IJCNN). Piscataway, NJ: IEEE, 2019: 1-6

[48] Chen Hongming, Liu Quan, Yan Yan, et al. An experience-guided deep deterministic actor-critic algorithm with multi-actor [J]. Journal of Computer Research and Development, 2019, 56(8): 1708 (in Chinese)
(陈红名, 刘全, 闫岩, 等. 基于经验指导的深度确定性多行动者-评论家算法[J]. 计算机研究与发展, 2019, 56(8): 1708)

[49] Garcıa J, Fernández F. A comprehensive survey on safe reinforcement learning [J]. Journal of Machine Learning Research, 2015, 16(1): 1437-1480

Wang Jinyong, born in 1983. PhD candidate. Student member of CCF. His main research interests include formal specification and verification, safe artificial intelligence, and autonomous driving safety.

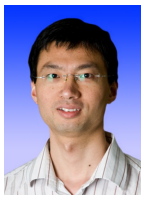
王金永,1983 年生.博士研究生.CCF 学生会员.主要研究方向为形式化规约与验证、安全人工智能和无人驾驶安全.

Huang Zhiqiu, born in 1965. PhD, professor, PhD supervisor. Distinguished member of CCF. His main research interests include system safety, formal methods, software engineering, and artificial intelligence.

黄志球,1965 年生.博士,教授,博士生导师.CCF 杰出会员.主要研究方向为系统安全、形式化方法、软件工程和人工智能.

Yang Deyan, born in 1982. PhD, associate professor, master supervisor. Her main research interests include statistical learning, game theory, and intelligent decision making.

杨德艳,1982 年生,博士,副教授,硕士生导师.主要研究方向为统计学习、博弈论和智能决策理论.



Xiaowei Huang, born in 1980. PhD, professor, PhD supervisor. His main research interests include artificial intelligence safety, formal methods, and autonomous robotics.

Xiaowei Huang,1980 年生.博士,教授,博士生导师.主要研究方向为人工智能的安全性,形式化方法和无人自主机器人.

Zhu Yi, born in 1976. PhD, professor, Master supervisor. Senior member of CCF. His main research interests include formal methods, software reliability, and intelligent software development.

祝　义,1976 年生.博士,教授,硕士生导师.CCF 高级会员.主要研究方向为形式化方法、软件可靠性和智能软件开发.

Hua Gaoyang, born in 1998. Master candidate. Student member of CCF. His main research interests include deep reinforcement learning and autonomous driving safety.

华高洋,1998 年生.硕士研究生.CCF 学生会员.主要研究方向为深度强化学习和无人驾驶安全.