doi:10.11835/j.issn.1000-582X.2025.10.007
引用格式:喻志成,赵俊鹏,刘永刚,等.基于ARMA车速预测的智能车交叉口强化学习决策研究[J].重庆大学学报,2025,48(10):68-80.

基于ARMA车速预测的智能车交叉口强化学习决策研究

喻志成[1],赵俊鹏[2],刘永刚[1],夏甫根[3],叶　明[4]
(1.重庆大学 高端装备机械传动全国重点实验室,重庆 400044;2.北京航天发射技术研究所,北京 100076;
3.成都壹为新能源汽车有限公司 成都 611730;4.重庆理工大学 车辆工程学院 重庆 400054)

摘要:为解决无信号交叉口的智能车决策控制问题,以双向单车道交叉口下两车合流工况为对象,采用强化学习方法开展研究,建立车辆状态空间到动作空间的映射。针对目前研究中环境车辆车速设置过于简单问题,以实际场景下采集的数据作为环境车辆的轨迹信息构建场景模型。基于自回归滑动平均模型对环境车辆的车速进行预测。结合智能车及预测的环境车辆车速时序信息建立先行让行决策模型计算本车参考车速,引入参考车速构建强化学习的奖励函数加速训练收敛速度。结果表明:所提出的强化学习模型具有较快收敛速度,训练得到的智能体在与不同驾驶风格的环境车辆博弈时能安全通过交叉口,为无信号交叉口智能车安全通行决策控制提供参考依据。

关键词:交叉口;自动驾驶;自回归滑动平均模型;强化学习
中图分类号:U471.15　　**文献标志码:**A　　**文章编号:**1000-582X(2025)10-068-13

Research on reinforcement learning-based autonomous vehicle decision-making at intersections using an ARMA speed forecasting model

YU Zhicheng[1], *ZHAO Junpeng*[2], *LIU Yonggang*[1], *XIA Pugeng*[3], *YE Ming*[4]
(1. State Key Laboratory of Mechanical Transmission for Advanced Equipment, Chongqing University, Chongqing 400044, P. R. China; 2. Beijing Aerospace Launch Technology Research Institute, Beijing 100076, P. R. China; 3. Chengdu Yiwei New Energy Vehicle Co., Ltd., Chengdu 611730, P. R. China; 4. Vehicle Engineering Institute, Chongqing University of Technology, Chongqing 400054, P. R. China)

Abstract: To address the challenge of autonomous vehicle decision-making and control at unsignalized intersections, this study investigates the merging behavior of two vehicles at a two-way single-lane intersection. Reinforcement learning is used to establish a mapping between the vehicle state space and action space for autonomous decision-making. To overcome the limitations of overly simplified speed settings in existing studies,

收稿日期:2020-12-28
基金项目:重庆市技术创新与应用发展专项重大项目(CSTB2023TIAD-STX0035);四川省科技计划项目(2019YFG0528)。Supported by Chongqing Municipal Technological Innovation and Application Development Special Program (CSTB2023TIAD-STX0035) and Sichuan Science and Technology Funding Project (2019YFG0528).
作者简介:刘永刚(1982—),男,教授,博士研究生导师,主要从事智能车决策与控制、车辆变速传动及智能控制、新能源汽车动力系统优化与控制方向研究,(E-mail)andyliuyg@cqu.edu.cn。

real-world trajectory data of surrounding vehicles are used to construct an environmental traffic model. The autoregressive moving average (ARMA) model is applied to predict the speeds of surrounding vehicles. By integrating the predicted speed profiles with the autonomous vehicle's motion parameters, a forward decision-making model is established to calculate reference speeds. These reference speeds are incorporated into the reinforcement learning reward function to accelerate training convergence. Experimental results show that the proposed model achieves rapid convergence, and the trained agent can safely navigate the intersection while interacting with surrounding vehicles exhibiting diverse driving behaviors. This work provides a reference framework for improving the safety and efficiency of autonomous vehicle decision-making at unsignalized intersections.

Keywords: intersections; autonomous vehicles; autoregressive moving average model (ARMA); reinforcement learning

随着自动驾驶科技的持续进步,大量具备初级自动驾驶辅助系统(advanced driver assistance system,ADAS)的车辆已逐步投入使用。然而,在复杂道路环境下实现可靠自主驾驶,是当前技术由低阶向高阶演进过程中面临的核心挑战。其中,无信号灯控制的交叉路口由于交通事故频发,已成为众多学者聚焦的研究重点,特别是在该场景下车辆的智能决策与控制方法方面[1]。

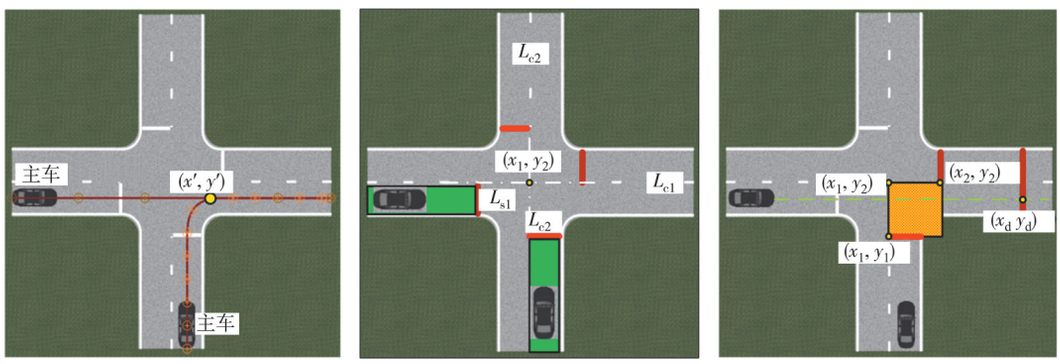
依据对周边车辆行为意图的预测结果制定规划策略,是一种常见且高效的解决思路[2-3]。Noh 通过预估轨迹划定自动驾驶车辆可能面临的危险范围,借助威胁评估、贝叶斯概率网络与时窗滤波相结合的方式,对风险进行量化分析并进行智能车的决策[4]。Ramyar 等[5]提出的 Takagi-Sugeno 数据驱动技术能够较准确估计驾驶员在交叉口的行为。这种技术使用少量的策略模型进行训练,减少计算复杂度。近年来,将隐马尔可夫(partially observable markov decision process, POMDP)方法运用于交叉口决策的研究逐渐增多[6-7]。Shu 等[8]提出了基于关键转向点(comprehensive transaction platform, CTP)的方法解决交叉口左转决策问题,基于 POMDP 的方法被运用于求解最优决策序列。Kye 等[9]针对无信号灯交叉路口的通行决策问题,提出一种基于意图感知的自动驾驶决策方案。该方法通过推断周边交通参与者的行为意图,将车辆的决策任务构建为部分可观测的马尔科夫决策过程。Hubmann 及其团队[10]提出一种可在线运行的 POMDP 架构,该框架能够适应多种驾驶场景下的自主决策任务。著名的 DeepMind 公司通过围棋等博弈游戏证明了深度强化学习在决策领域的可行性[11-12],在交叉路口智能车辆的决策与控制研究中,多种机器学习方法得到应用[13-14]。例如,Isele 等[15]应用深度强化学习(deep reinforcement learning,DRL)技术,训练出在任务完成效率与成功率等多个维度上超越传统启发式策略的模型,但其系统泛化性能存在不足。针对 DRL 在复杂环境中训练迭代量过大的问题,Qiao 等[16]引入了自动课程生成(automatically generated curriculum,AGC)机制,用于处理交叉路口的自主驾驶决策任务。通过对比 AGC 与随机序列生成的训练结果,发现 AGC 能够在大幅缩短训练周期的前提下,达到相当甚至更优的性能表现。

总的来看,现有基于周边车辆状态预测的路径规划方法,往往先将碰撞等风险事件进行量化评估,再借助规则型策略完成车辆决策,然而这类策略适应性普遍有限。采用隐马尔可夫模型的方法虽在表达不确定性上具有优势,但存在计算复杂度高的问题。尽管可引入蒙特卡洛采样以缓解计算压力,该方法仍需要对行为空间做离散化处理,且离散粒度往往较粗。当前,深度强化学习在该领域的应用多聚焦于整个车辆队列通过交叉口的协调控制,针对单车之间交互博弈行为的研究仍较为欠缺。另外,很多研究为简化问题常假设周围车辆保持匀速行驶,这一设定与真实交通场景的动态特性存在显著差异。为此,本研究聚焦于无信号灯交叉口两车交互的博弈行为,利用真实道路采集的轨迹数据构建环境模型,旨在为无人驾驶决策机制提供更贴

近实际的理论支撑。在通行速度控制方面,引入强化学习方法,通过融合预测的他车速度信息构建礼让决策模型,推算参考车速提升学习效率,最终训练出能够适应不同驾驶风格、安全通过交叉口的智能决策体。

1 基于ARMA方法的车速预测

无信号交叉口的冲突情况较多,选取在双向单车道下智能车直行,环境车辆从位于智能车辆右手侧的路口驶入并右转的情况进行分析(如图1(a))所示)。这属于典型的交叉口合流情况,基于该工况验证所提方法的有效性。



(a) 博弈双方行驶意图 (b) 道路几何学模型 (c) 中心冲突区域

图1 环境模型

Fig.1 Environmental model

1.1 道路几何学模型

为便于后续对决策控制模型进行数学表述,先构建研究道路的几何模型。该道路为双向单车道设定,两车分别从2个方向入口驶入,合并进入同一车道。东西与南北方向车道中心线分别标识为L_{c1}与L_{c2},4处停止线依次标记为$L_{s1}-L_{s4}$(参见图1(b))。以车道中心线交点为几何模型原点(x_1, y_2),出口停止线及他车进入路口停止线与中心线L_{c1}、L_{c2}的交点分别为(x_2, y_2)与(x_1, y_1)(如图1(c))。这3个关键点共同界定一个中央冲突区(center crash area,CCA),多项研究与实际经验均指出,该区域为车辆碰撞高发段,具有显著的行车风险。

1.2 真实工况下的车辆轨迹数据

在无信号交叉口自动驾驶决策研究中,多数现有方法将他车运动轨迹假设为匀速行驶[8-9,14],这一设定与真实交通动态存在较大偏差。为提升研究的真实性,基于Open ITS数据平台提供的实际交通信息进行分析[17]。研究选取中国云南省昆明市文艺路与新迎路交叉口为观测点,于路口西南侧高层建筑架设摄像设备,进行连续7天全天候视频采集。最终,借助图像处理技术从视频序列中提取参与交互车辆的具体运动轨迹。

交叉口博弈车辆最后驶入同一路口的情况为合流工况。不同于其他典型工况,合流工况下车辆通过交叉口的先后顺序非常明显。两车安全通过交叉口并达到自身行驶意图的前提是在交叉口行进过程中决定自身的通行优先级,即先行通过交叉口还是让行通过交叉口。因此,根据直行车辆通过交叉路口的先后顺序将数据集分为2大类:直行车辆先行与直行车辆让行。现将合流工况第1~60组数据分类,剔除部分不合适数据如图2和图3所示。从图中曲线可以看出无论是直行车先行组还是让行组的数据,两车交互过程中速度波动均较为剧烈。通过对比同一类工况下先行车与让行车的速度及加速度曲线,不难发现先行通过交叉口的车辆其速度波动较后续通过交叉口的车辆更加剧烈。结合交叉口的工况考虑这是非常合理的现象,先行通过交叉口的车辆需要以相对另一辆车较大的速度通过。为了保证行车安全,驾驶员需要不断确认自车与对象车辆的状态。

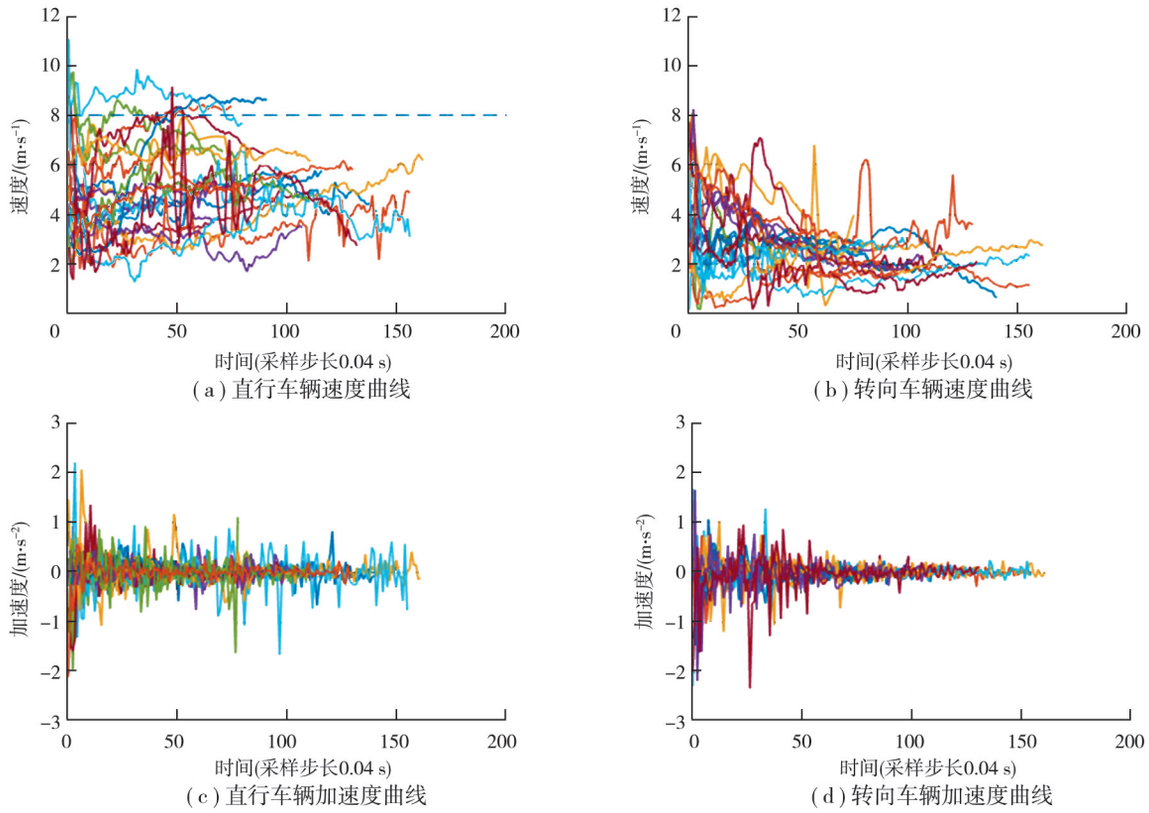


图 2　直行车先行工况

Fig.2　Scenario where the straight-moving vehicle has th right of way

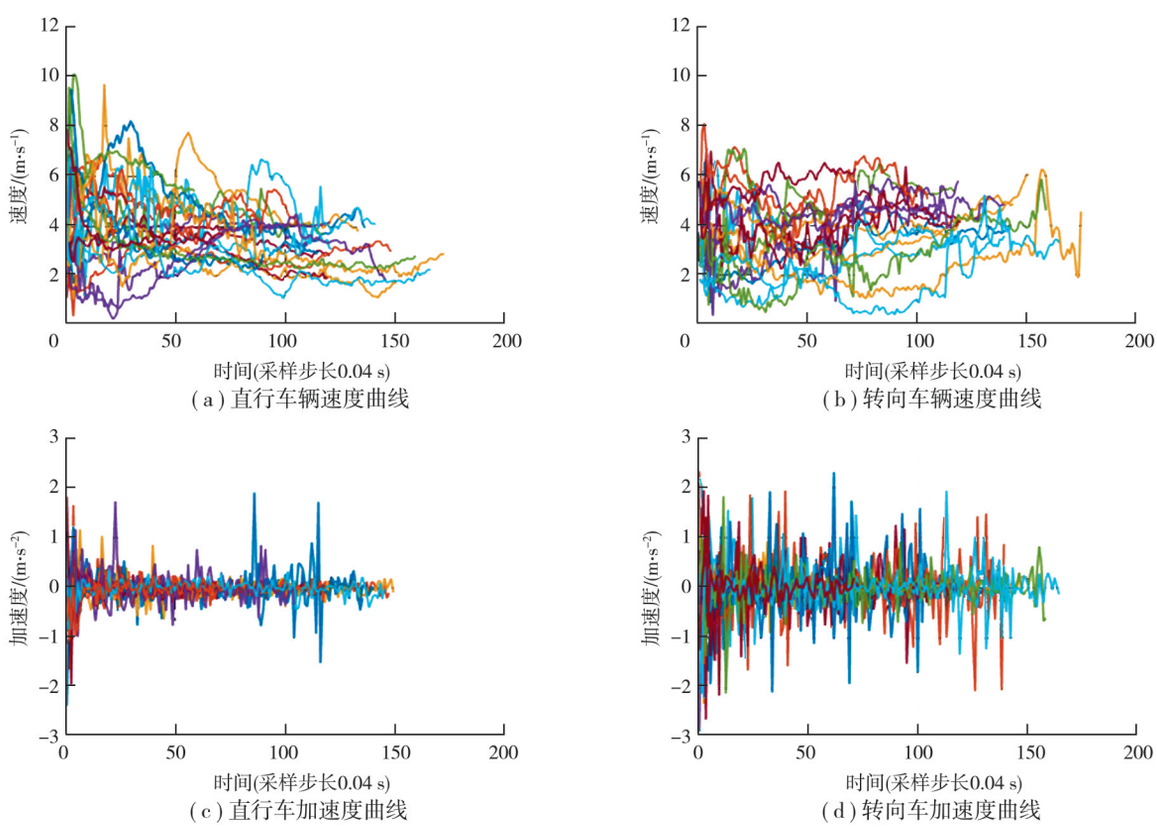


图 3　直行车让行工况

Fig.3　Scenario where the straight-moving vehicle yields

1.3 基于ARMA方法的环境车辆未来车速时序预测

本研究以直行通过交叉口的智能网联汽车(connected autonomous vehicle,CAV)为控制主体,环境车辆的转向轨迹则依据真实驾驶数据构建。考虑到制动响应延迟等实际因素,系统难以在极短时间内对突发的碰撞风险做出避让判断。为提高行车安全,决策需要具备一定前瞻力,要求对他车未来速度进行有效预估。当前车速预测已有多种技术路径,主要包括时间序列建模与机器学习方法等。机器学习方法通常依赖大规模样本训练,而时间序列模型则通过构建数学表征规避了这一限制。ARMA模型作为自回归(auto regressive, AR)与移动平均(moving average, MA)结合的经典方法,已被广泛应用于经济学等领域的趋势预测。本文采用自回归移动平均模型(autoregressive moving average model,ARMA)对环境车辆的未来速度进行预测。在该类模型中,自回归(AR)模型为基本形式之一,数学表达式如式(1)所示。式中:Y_t表示当前时刻t的序列值,$Y_{t-1},Y_{t-2},\cdots,Y_{t-p}$分别为序列在之前$p$个时刻的历史观测值;$\varepsilon_t$为独立同分布的随机误差序列,其期望与方差需满足式(2)所给出的关系。AR模型依据过去p个时间点的数据对当前值进行估计,而移动平均(MA)模型则侧重于对误差项的滑动过程进行建模,具体形式由式(3)表示。由此可将AR与MA组合,得到如式(4)所描述的ARMA模型整体结构。

$$Y_t=\beta_1 Y_{t-1}+\beta_2 Y_{t-2}+\cdots+\beta_p Y_{t-p}+\varepsilon_t, \tag{1}$$

$$E(\varepsilon_t)=D(\varepsilon_t)>0, \tag{2}$$

$$Y_t=\varepsilon_t+\alpha_1\varepsilon_{t-1}+\alpha_2\varepsilon_{t-2}+\cdots+\alpha_q\varepsilon_{t-q}, \tag{3}$$

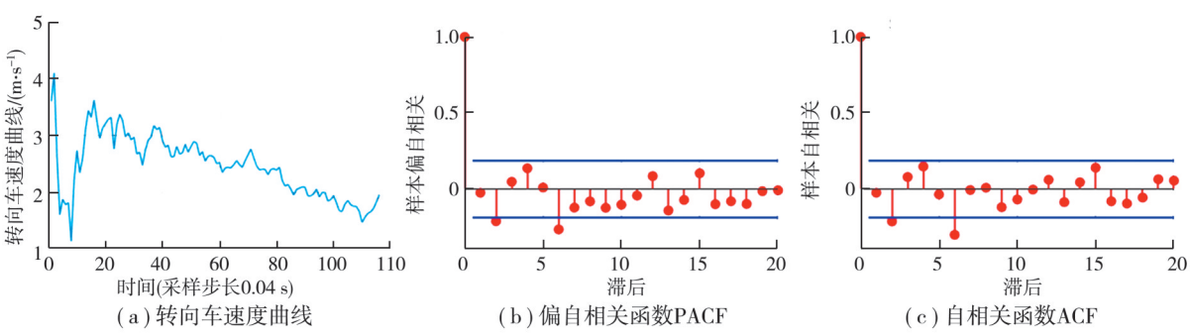
$$Y_t=\beta_0+\beta_1 Y_{t-1}+\beta_2 Y_{t-2}+\cdots+\beta_p Y_{t-p}+\varepsilon_t+\alpha_1\varepsilon_{t-1}+\alpha_2\varepsilon_{t-2}+\cdots+\alpha_q\varepsilon_{t-q}。 \tag{4}$$

以其中1组轨迹数据为例,其速度曲线如图4(a)所示,采用自相关函数ACF和偏自相关函数PACF法则确定ARMA的模型阶数p、q,建立基于ARMA方法的车速预测模型为排除伪回归现象,对序列进行平稳性检验。通过平稳性检验的数据具有如公式(5)所示的形式,其中,白噪声μ_t满足如公式(6)所示的特性。

$$Y_t=\mu_t, \tag{5}$$

$$E(\mu_t)=0。 \tag{6}$$

若原始数据不满足平稳性则对其取N阶差分直至其N阶差分满足平稳性。计算结果表明所选用的车速数据其二阶差分模型通过平稳性检验,随后对其二阶差分模型计算自相关函数和偏自相关函数,结果如图4(b)和(c)所示。

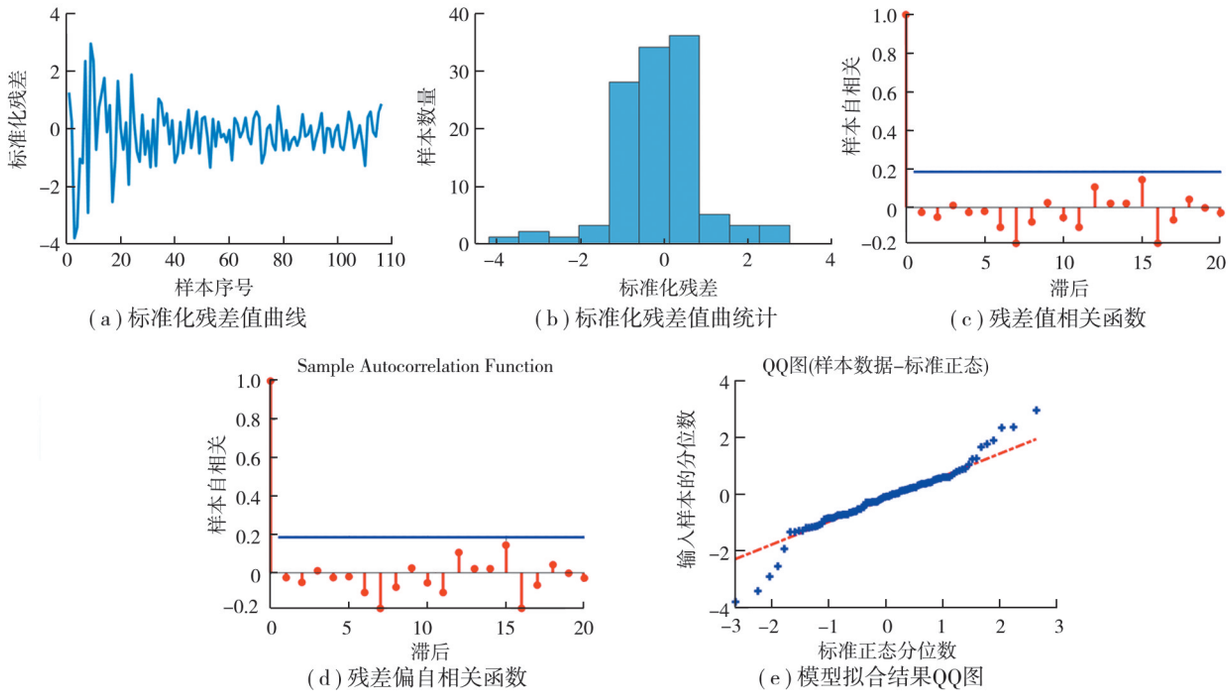


(a)转向车速度曲线 (b)偏自相关函数PACF (c)自相关函数ACF

图4 ARMA模型阶数仿真结果

Fig.4 Simulation results of ARMA model order

依据相关领域的研究经验总结[18],ARMA模型的模型阶数p、q均取6。对选取的轨迹速度曲线建立ARMA模型进行拟合,为了检验拟合的结果,需要对模型与原始数据进行残差分析,残差分析的结果如图5(a)~(e)所示。



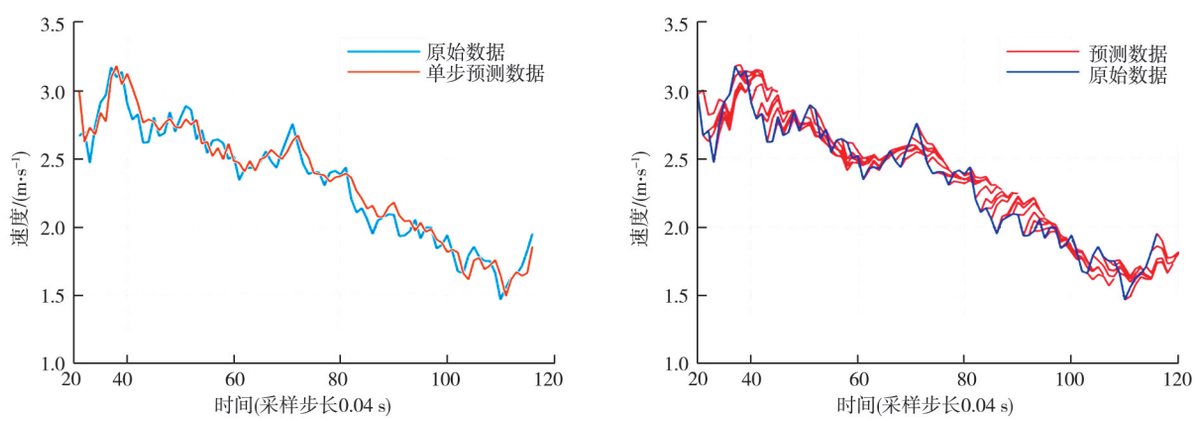
(a) 标准化残差值曲线　(b) 标准化残差值曲统计　(c) 残差值相关函数

(d) 残差偏自相关函数　(e) 模型拟合结果QQ图

图 5　ARMA模型拟合结果残差分析

Fig. 5　Residual analysis of ARMA model fitting results

最终,基于已构建的ARMA模型对车速时间序列开展预测分析。以前30个时间步(采样间隔0.04 s)的车速记录作为历史输入,自第31个时间步起执行滚动预测。分别设置了预测步长1和5两种工况,相应预测结果如图6所示。



(a) 单步预测结果　(b) 预测步长为5的多步预测结果

图 6　基于ARMA模型的车速预测结果

Fig.6　Speed prediction results based on ARMA model

预测结果表明,当预测步长为1时,模型在滚动预测过程中的均方误差为0.024 3 m/s;而预测步长为5时,对应的均方误差序列为[0.024 3,0.032 5,0.037 0,0.041 0,0.041 1] m/s。在整个预测时段原始车速平均值为2.331 m/s。所有预测步长下的误差均保持在约1%~2%,最大误差未超过2%,处于可接受范围,模型具有较好的预测性能。

2　基于强化学习的交叉口智能车速度决策研究

强化学习(reinforcement learning,RL)常用于解决复杂的决策控制问题。强化学习的基本单元可表示为

($\boldsymbol{S}$,$\boldsymbol{A}$,R,$\boldsymbol{S}'$),其中:$\boldsymbol{S}$表示当前时刻的状态;$\boldsymbol{S}'$表示状态$\boldsymbol{S}$经过动作$\boldsymbol{A}$之后所到达的新状态,而在这个过程中智能体获得了奖励R。所提出方法整体的决策流程如图7所示。

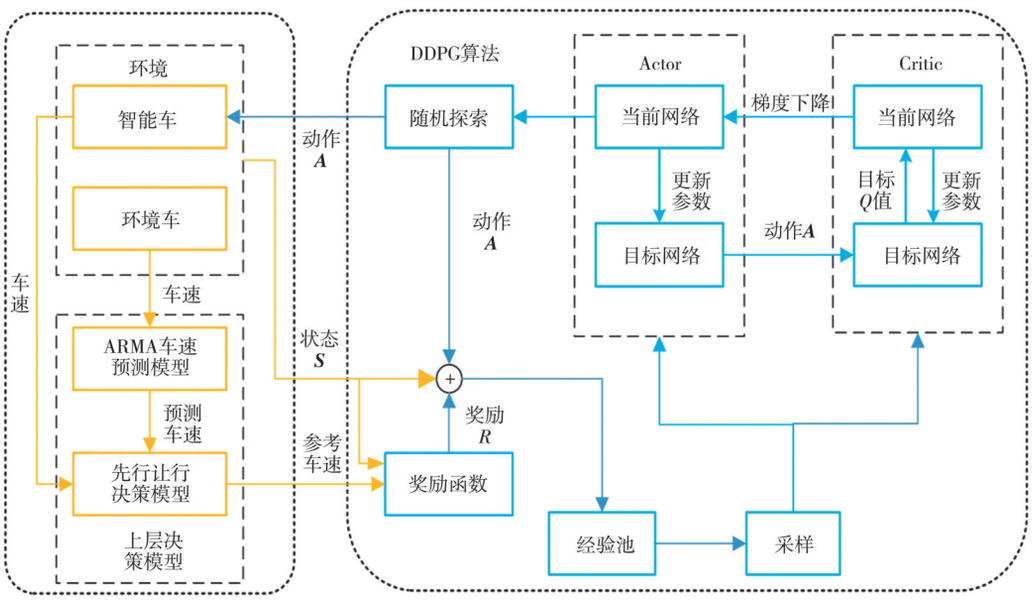


图7 基于ARMA车速预测的智能车交叉口强化学习决策流程图

Fig.7 Flow chart of reinforcement learning autonomous vehicle decision making at intersection based on ARMA speed forecasting model

2.1 状态空间

强化学习旨在训练一个能够依据环境观测信息做出合理决策的智能体(Agent)。无信号灯交叉口属于复杂交通环境,其状态信息涵盖目标车辆状态、道路结构、自车信息以及通过处理得到的车间距离等衍生指标。选用的状态空间序列的具体构成如式(7)所示

$$\text{observision}=(\boldsymbol{S}_{\text{ego}},\boldsymbol{S}_{\text{env}})^{\text{T}}, \tag{7}$$

$$\boldsymbol{S}_{\text{ego}}=\left(x_{\text{ego}},y_{\text{ego}},v_{\text{ego}}\right), \tag{8}$$

$$\boldsymbol{S}_{\text{env}}=\left(x_{\text{env}},y_{\text{env}},v_{\text{env}}\right), \tag{9}$$

式中:$\boldsymbol{S}_{\text{ego}}$与$\boldsymbol{S}_{\text{env}}$分别代表了智能车与环境车的状态序列;$x$,$y$,$v$则分别代表了横坐标,纵坐标以及速度信息。

2.2 动作空间

当前多数采用强化学习的交叉口自动驾驶决策研究中,智能体的输出常被定义为车辆的加速度指令。此类方式将决策与控制划分为2个层次:强化学习智能体作为上层决策单元输出目标加速度,底层控制器则根据该加速度计算出实际所需要的油门开度与制动压力信号。该方法对底层控制器的精度提出较高要求。为避免这一依赖,研究将强化学习智能体的输出直接设定为车辆动力学模型可执行的油门开度与制动压力指令,动作空间的具体定义如式(10)

$$\text{Action}=\left(\text{Throttle}_{\text{ego}},\text{Brakepress}_{\text{ego}}\right)^{\text{T}}。 \tag{10}$$

2.3 奖励函数

奖励函数的设计直接影响智能体策略的学习方向与收敛结果。在本研究关注的无信号交叉口场景中,保障车辆安全通过是首要目标。然而,若仅设置终点奖励与碰撞惩罚,可能导致智能体在训练过程中收敛至2种极端策略:一是过度保守,持续停车等待直至其他车完全通过;二是过于激进,持续加速并以过高速度强行通过路口。这2种行为均不属于安全高效的理想驾驶策略,因此,奖励函数需要加上对超出正常车速范围的惩罚,正常车速的阈值设定参考图2与图3中的实际交叉口通行中直行车辆车速曲线。奖励函数中抵达终

点、碰撞以及超速或者怠速的部分如式(11)~(13)

$$R_{\mathrm{G}}=\begin{cases}10\,000, & x_{\mathrm{ego}}>x_{\mathrm{d}}\cap x_{\mathrm{env}}>x_{\mathrm{d}},\\ 0, & x_{\mathrm{ego}}\leqslant x_{\mathrm{d}}\cup x_{\mathrm{env}}\leqslant x_{\mathrm{d}},\end{cases} \tag{11}$$

$$R_{\mathrm{C}}=\begin{cases}-10\,000, & \mathrm{collsion}=1,\\ 0, & \mathrm{collsion}=0,\end{cases} \tag{12}$$

$$R_{\mathrm{S}}=\begin{cases}-100, & v_{\mathrm{ego}}>v_{\mathrm{upper}}\cup v_{\mathrm{ego}}<v_{\mathrm{lower}},\\ 0, & v_{\mathrm{lower}}\leqslant v_{\mathrm{ego}}\leqslant v_{\mathrm{upper}},\end{cases} \tag{13}$$

式中：x_{d}为图1中建立的道路几何模型的终点横坐标；v_{upper}为速度上限值，依据中国相关法律法规和数据将速度的最大值设定为30 Km/h；v_{lower}为避免消极怠速策略所设定的速度下限值，结合数据设置为7.2 Km/h。除上述的3部分之外，在实际训练过程中，由于状态与动作空间的搜索范围过大，智能体探索效率降低，难以在合理时间内获取有效策略经验。针对该问题，所研究方法的一大贡献是将所建立的ARMA车速预测模型的预测结果引入构建先行让行决策模型。在两车通过合流点(x',y')之前决定是先行通过交叉口或者是让行通过交叉口，引导Agent采取更明确的加速或者减速策略。

$$\mathrm{TTC}_{\mathrm{ego}}=\frac{-v_{\mathrm{ego}}+\sqrt{v_{\mathrm{ego}}^2+2a_{\max}H_1}}{a_{\max}}, \tag{14}$$

$$H_1=x'-x_{\mathrm{ego}}, \tag{15}$$

$$\mathrm{TTC}_{\mathrm{env}}=\frac{H_2}{\overline{v_{\mathrm{pre}}}}, \tag{16}$$

$$H_2=\sqrt{\left(x'-x_{\mathrm{env}}\right)^2-\left(y'-y_{\mathrm{env}}\right)^2}, \tag{17}$$

$$\overline{v_{\mathrm{pre}}}=\frac{\sum_{i=1}^{l}v_{\mathrm{pre}}^i}{l}。 \tag{18}$$

通过上式可以计算智能车与环境车分别到达图1中合流点(x',y')所需要的时间。其中：智能车估算的是以所允许的最大加速度$a_{\max}$加速达到合流点的最短时间，最大加速度$a_{\max}$根据图2和图3的数据统计结果取2m/s^2；H_1为主车到达合流点的距离，H_2为环境车达到合流点的距离；v_{pre}^i是ARMA模型的预测结果$v_{\mathrm{pre}}^i\in[v_{\mathrm{pre}}^1,v_{\mathrm{pre}}^2,\cdots,v_{\mathrm{pre}}^{l-1},v_{\mathrm{pre}}^l]$；$l$是预测步长，其值为5。根据智能车和环境车到达合流点的时间决定交叉口通行的先后顺序。依据交叉口通行的先后顺序，公式(19)(20)分别计算出$\mathrm{TTC}_{\mathrm{ego}}<\mathrm{TTC}_{\mathrm{env}}$(先行)和$\mathrm{TTC}_{\mathrm{ego}}>\mathrm{TTC}_{\mathrm{env}}$(让行)时的参考车速。$T_{\mathrm{safe}}$是考虑了车辆刹车响应时间等因素在内的安全时距，取1.5 s。

$$v_{\mathrm{ref}}=\frac{S_1}{\mathrm{TTC}_{\mathrm{env}}-T_{\mathrm{safe}}},\ v_{\mathrm{ref}}<v_{\max}, \tag{19}$$

$$v_{\mathrm{ref}}=\frac{S_1}{\mathrm{TTC}_{\mathrm{env}}+T_{\mathrm{safe}}},\ v_{\mathrm{ref}}<v_{\max}, \tag{20}$$

将参考车速纳入强化学习的奖励函数中，如公式(21)所示。结合前式，得到强化学习训练过程奖励函数R_{reward}。

$$R_v=-\left(v_{\mathrm{ego}}-v_{\mathrm{ref}}\right)^2, \tag{21}$$

$$R_{\mathrm{reward}}=R_{\mathrm{G}}+R_{\mathrm{C}}+R_{\mathrm{S}}+R_v。 \tag{22}$$

2.4　训练算法

著名的DeepMind公司提出了深度确定性策略梯度算法(Deep-DPG，DDPG)，在高维动作空间或连续动作空间中都有很好表现，算法流程伪代码如表1所示。DDPG算法下的DRL拥有4个网络：Actor目标网络、Actor当前网络、Critic目标网络以及Critic当前网络。其中Actor当前网络负责策略网络参数θ^{μ}的迭代更新，

负责根据当前状态$\boldsymbol{S}_t$选择当前动作$\boldsymbol{a}_t$,用于和环境交互更新成为下一个状态$\boldsymbol{S}_{t+1}$并产生奖励值R。Actor目标网络负责根据经验回放池中采样的下一状态$\boldsymbol{S}_{t+1}$选择最优下一动作$\boldsymbol{a}_{t+1}$。网络参数$\theta^{\mu'}$定期从θ^{μ}复制。Critic当前网络负责价值网络参数θ^{Q}的迭代更新,负责计算当前$Q(\boldsymbol{S}_t,\boldsymbol{a}_t,\theta^{Q})$值。Critic目标网络负责计算目标$Q$值中的$Q'(\boldsymbol{S}_{t+1},\boldsymbol{a}_{t+1},\theta^{Q'})$部分,网络参数$\theta^{Q'}$定期从$\theta^{Q}$复制。因此,采用DDPG算法作为强化学习的训练算法。

$$Q=R+\gamma Q'\left(\boldsymbol{S}_{t+1},\boldsymbol{a}_{t+1},\theta^{Q'}\right)。\tag{23}$$

表 1 DDPG算法伪代码

Table 1 Pseudo-code of DDPG algorithm

算法: DDPG算法
输入:Actor当前网络,Actor目标网络,Critic当前网络,Critic目标网络,参数分别为θ^{μ},$\theta^{\mu'}$,θ^{Q},$\theta^{Q'}$,衰减因子γ,软更新系数τ,批量梯度下降的样本数m,目标Q网络参数更新频率C,最大迭代次数T,最大历元数M。随机噪声$\{\boldsymbol{N}\}$。
输出:最优Actor当前网络参数θ^{μ},Critic当前网络参数θ^{Q},随机初始化θ^{μ},θ^{Q},$\theta^{\mu'}=\theta^{\mu}$,$\theta^{Q'}=\theta^{Q}$以及经验回放的集合$D$。
for episode = 1,M do
初始化动作探索的随机噪声函数\mathcal{N}
接受初始状态$\boldsymbol{S}_1$
for i=1,T do
a) 在Actor当前网络基于状态$\boldsymbol{S}$得到动作$\boldsymbol{a}_t=\mu\left(\boldsymbol{S}_t\middle\vert\theta^{\mu}\right)+\boldsymbol{N}_t$;
b) 执行动作$\boldsymbol{a}_t$,得到新状态$\boldsymbol{S}_{t+1}$,奖励R,判断是否终止状态is_end;
c) 将$\left\{\boldsymbol{S}_t,\boldsymbol{a}_t,R,\boldsymbol{S}_{t+1},\text{is_end}\right\}$这个五元组存入经验回放集合$D$;
d) $\boldsymbol{S}_t=\boldsymbol{S}_{t+1}$;
e) 从经验回放集合D中采样m个样本$\left\{\boldsymbol{S}_j,\boldsymbol{a}_j,R_j,\boldsymbol{S}_{j+1},\text{is_end}_j\right\}$,$j$=1,2,…,$m$,计算当前目标$Q$值$y_i$: $y_i=\begin{cases}R_j, & \text{is\_end}_j=1,\\ R_j+\gamma Q'\left(\boldsymbol{S}_{j+1},\ \mu'\left(\boldsymbol{S}_{j+1}\middle\vert\theta^{\mu'}\right)\middle\vert\theta^{Q'}\right), & \text{is\_end}_j=0;\end{cases}$
f) 使用均方差损失函数$\frac{1}{m}\sum_{j=1}^{m}\left(y_i-Q\left(\boldsymbol{S}_j,\boldsymbol{a}_j\middle\vert\theta^{Q}\right)\right)^2$,通过神经网络的梯度反向传播更新Critic当前网络的所有参数θ^{Q};
g) 使用采样梯度策略更新Actor当前网络的所有参数θ^{μ}, $\nabla_{\theta^{\mu}}J\approx\frac{1}{m}\sum_{j=1}^{m}\nabla_{a}Q\left(\boldsymbol{S},\boldsymbol{a}\middle\vert\theta^{Q}\right)\Big\vert_{\boldsymbol{S}=\boldsymbol{S}_i,\boldsymbol{a}=\mu(\boldsymbol{S}_i)}\nabla_{\theta^{\mu}}\mu\left(\boldsymbol{S}\middle\vert\theta^{\mu}\right)\Big\vert_{\boldsymbol{S}_i}$;
h) 如果T%C=1,则更新Critic目标网络和Actor目标网络参数 $\theta^{Q'}\leftarrow\tau\theta^{Q}+(1-\tau)\theta^{Q'}$, $\theta^{\mu'}\leftarrow\tau\theta^{\mu}+(1-\tau)\theta^{\mu'}$,
end for
end for

3 建模与仿真分析

如图8所示,采用Matlab/Simulink与PreScan软件联合仿真,在PreScan中搭建虚拟交叉口驾驶场景模型,使用PreScan的Audi A8 2D车辆动力学模型。PreScan的车辆动力学模型与Simulink中的强化学习Agent通过虚拟总线传输数据。

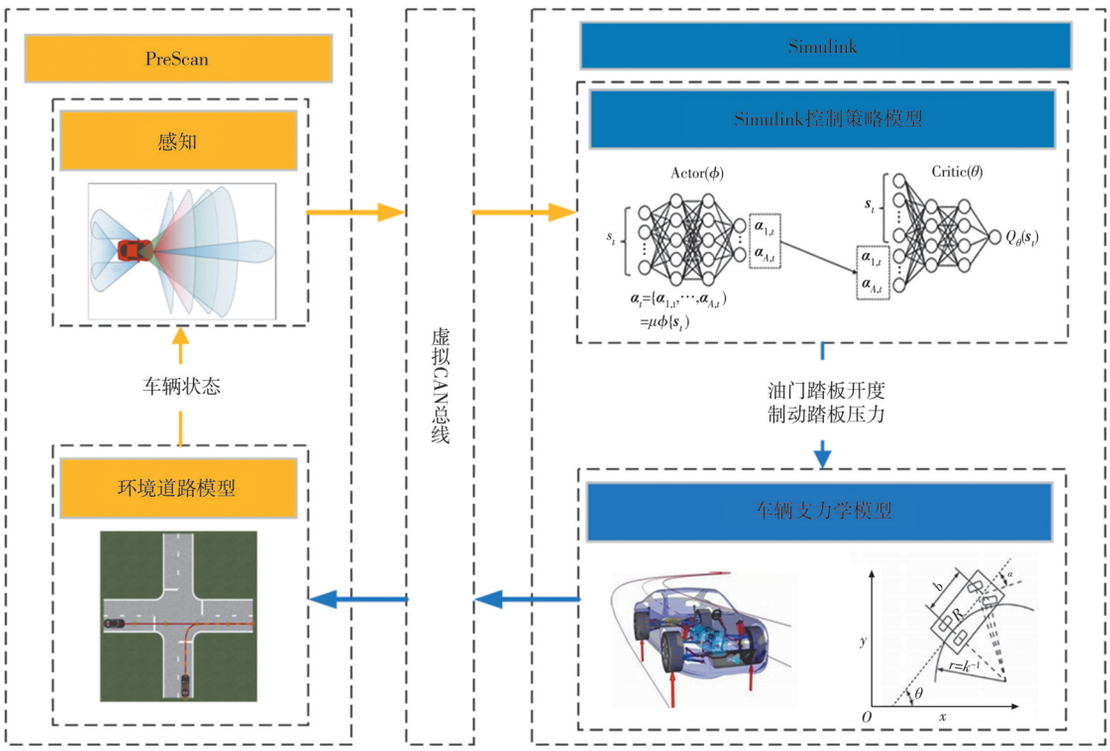


图 8　基于 PreScan 与 Simulink 的联合建模仿真

Fig.8　Joint modeling and simulation based on PreScan and Simulink

训练 Agent 的计算机 CPU 配置如下:intel i5-9600k3.7 GHz。训练过程中将环境车辆的轨迹设置分为 2 大类:激进类与保守类,分别对应于图 2 和图 3 中的转向车先行与转向车让行。车辆的轨迹数据来源于 Open ITS 数据平台中真实场景下采集到的数据[17]。训练过程中为环境车辆设置了 10 组不同轨迹,激进类与保守类的轨迹各 5 组,每次训练时环境车辆随机以其中一组轨迹作为预设轨迹行驶。智能车的初始速度在 3~6 m/s 范围内变化。将训练过程的最大历元数设置为 1 000,约花费 3 h 完成训练过程。训练过程中的奖励函数收敛曲线如图 9 所示。

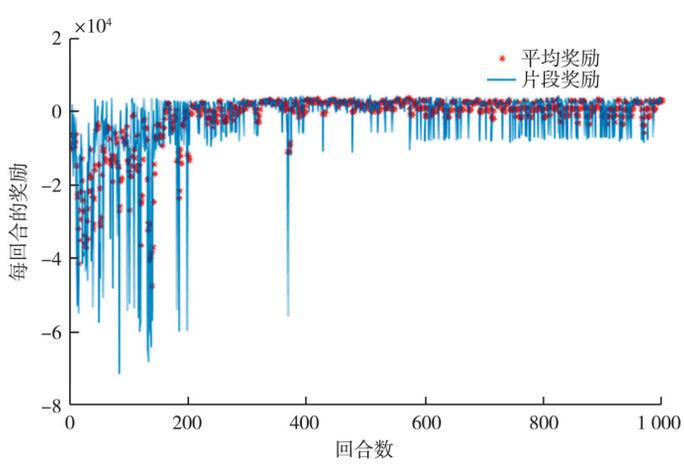
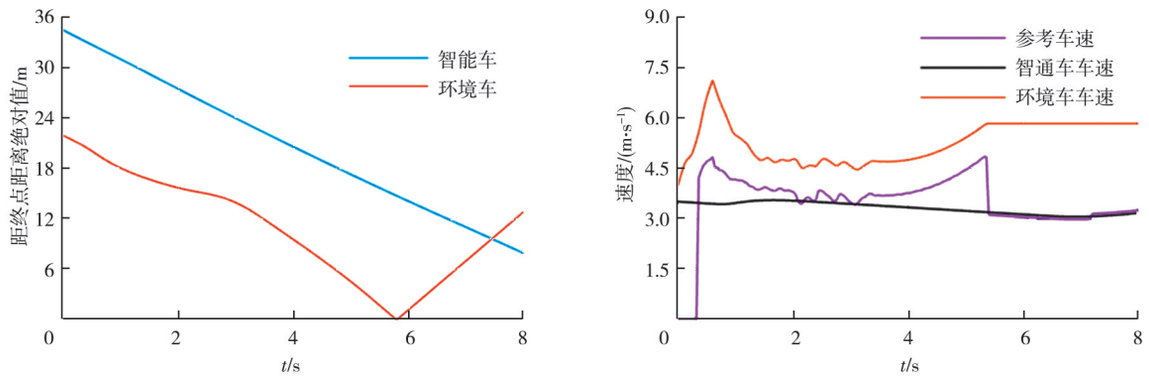


图 9　强化学习奖励函数收敛曲线

Fig.9　The convergence curve of reinforcement learning reward function

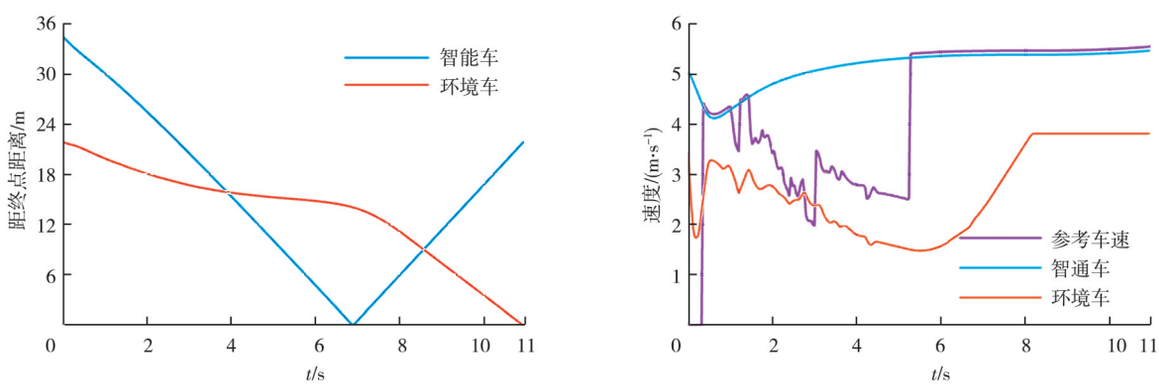
仿真结果显示在训练迭代次数到达约 200 次时,奖励函数趋向于收敛,整体的学习效率较高。为了验证最终的 Agent 效果,选取训练数据外的 2 组环境车速轨迹进行测试,环境车速曲线分别如图 10 和图 11 中橙色曲线所示。图 12 中黄色方框内是强化学习 Agent 控制的智能车,环境车辆以预先设置好的轨迹从右侧驶入交叉口。



(a) 两车到终点距离　　(b) 两车车速与参考车速对比

图10　环境车辆速度较激进场景下的仿真验证效果

Fig.10　Simulation results under the circumstance that the vehicle speed is more aggressive



(a) 两车到终点距离　　(b) 两车车速与参考车速对比

图11　环境车辆速度较保守场景下的仿真验证效果

Fig.11　Simulation results under the circumstance that the vehicle speed is more conservative

(a) 环境车辆速度较激进场景　　(b) 环境车辆速度较保守场景

图12　2种验证场景下的PreScan仿真动画

Fig.12　PreScan simulation animation in two verification scenarios

图10展示了环境车辆较为激进场景下Agent的决策控制效果。该场景下智能车以较低初始速度驶入交叉路口,而环境车在初始速度高于智能车的前提下在驶入交叉路口不久后激进加速到一个较高车速。在判断无法先于环境车辆通过交叉口时,上层的先行/让行决策模块计算出的参考车速略低于环境车的预测车速。Agent在选择让行通过交叉口的同时车速维持在不低于所设置的下限范围,较为平稳安全通过交叉口。由于预测车速的效果较好,参考车速的变化趋势与环境车的实际车速波动趋势较为一致。基于ARMA的车速预测模型需要一定历史时序信息作为依据来推断后续的车速信息。同时前段路程距离交叉口中心区域较

远，进行速度预测的必要性不大。因此，图 10 与图 11 中仿真的开始阶段参考车速为 0 m/s。为了获取较好的预测效果，在联合仿真时选择从第 11 步（仿真步长 0.04 s）开始对后续的时序信息做预测。

图 11 是环境车辆较为保守场景下 Agent 的决策控制效果，该场景下智能车初始速度较大且环境车辆的驾驶风格偏向保守。在仿真初期参考车速为 0 m/s 的情况下，Agent 无法估计环境车的未来信息，无法得出先行还是让行的决策，进而选择保守的减速策略。在 ARMA 的车速预测模型给出未来车速的预测信息之后，上层的先行/让行决策模块计算出智能车能够先于对方通过交叉口。因此，Agent 在经历一段时间的减速后立刻加速，先于环境车辆通过交叉口。在行驶过程中智能车的加速趋势逐渐趋于平缓，最终以未超过所设置上限范围的车速安全通过交叉口。

4　结论

针对无信号交叉口的智能车决策控制问题，以双向单车道交叉口下两车合流工况为研究对象，提出基于 ARMA 车速预测的交叉口强化学习决策方法。相较于其他的交叉口决策研究，控制车辆的博弈对象轨迹信息来源于真实场景下所采集的数据，因此，所提出的方法在强化学习方法运用于实际交叉口自动驾驶决策方面更具参考价值。基于自回归滑动平均模型对环境车辆的车速进行预测，结合智能车以及所预测的环境车辆车速时序信息建立先行让行决策模型计算本车参考车速，引入参考车速构建强化学习的奖励函数加速训练收敛速度，仿真结果表明所提出的强化学习模型能够快速收敛。选取与训练过程不同的数据作为轨迹信息验证所训练智能体的决策控制效果。结果表明训练得到的智能体在与不同驾驶风格的环境车辆博弈时能够安全通过交叉口，证明所提方法在无信号交叉口的智能车决策控制问题方面的有效性。

参考文献

[1] Shirazi M S, Morris B T. Looking at intersections: a survey of intersection monitoring, behavior and safety analysis of recent studies[J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 18(1): 4-24.

[2] Hult R, Zanon M, Gros S, et al. Optimal coordination of automated vehicles at intersections: theory and experiments[J]. IEEE Transactions on Control Systems Technology, 2018, 27(6): 2510-2525.

[3] Ma L, Xue J, Kawabata K, et al. Efficient sampling-based motion planning for on-road autonomous driving[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(4): 1961-1976.

[4] Noh S. Decision-making framework for autonomous driving at road intersections: safeguarding against collision, overly conservative behavior, and violation vehicles[J]. IEEE Transactions on Industrial Electronics, 2018, 66(4): 3275-3286.

[5] Ramyar S, Homaifar A, Anzagira A, et al. Fuzzy modeling of drivers' actions at intersections[C]//2016 World Automation Congress (WAC). Stockholm, Sweden: IEEE, 2016: 1-6.

[6] Bouton M, Cosgun A, Kochenderfer M J. Belief state planning for autonomously navigating urban intersections[C]//2017 IEEE Intelligent Vehicles Symposium. Hangzhou，China: IEEE, 2017: 825-830.

[7] Hubmann C, Quetschlich N, Schulz J, et al. A POMDP maneuver planner for occlusions in urban scenarios[C]//2019 30th IEEE Intelligent Vehicles Symposium. Paris, France: IEEE, 2019: 2172-2179.

[8] Shu K, Yu H, Chen X, et al. Autonomous driving at intersections: a critical-turning-point approach for left turns[C]//2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). Rhodes, Greece: IEEE, 2020:1-6.

[9] Kye D K, Kim S W, Seo S W. Decision making for automated driving at unsignalized intersection[C]//2015 15th International Conference on Control, Automation and Systems (ICCAS). Busan, South Korea：IEEE, 2015: 522-525.

[10] Hubmann C, Schulz J, Becker M, et al. Automated driving in uncertain environments: planning with interaction and uncertain maneuver prediction[J]. IEEE Transactions on Intelligent Vehicles, 2018, 3(1): 5-17.

[11] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning.[J/OL]. (2014-12-19)[2020-7-29]. https://arxiv.org/pdf/1312.05602.

[12] Silver D, Huang A, Maddison C J, et al. Mastering the game of go with deep neural networks and tree search [J]. Nature, 2016, 529(7587): 484-489.

[13] Zhou M, Yu Y, Qu X. Development of an efficient driving strategy for connected and automated vehicles at signalized intersections: a reinforcement learning approach[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 21(1): 433-443.

[14] Chen W L, Lee K H, Hsiung P A. Intersection crossing for autonomous vehicles based on deep reinforcement learning[C]// 2019 IEEE International Conference on Consumer Electronics. Piscataway: IEEE, 2019: 1-2.

[15] Isele D, Rahimi R, Cosgun A, et al. Navigating occluded intersections with autonomous vehicles using deep reinforcement learning[C]//2018 IEEE International Conference on Robotics and Automation. Brishane, Australia: IEEE, 2018: 2034-2039.

[16] Qiao Z, Muelling K, Dolan J, et al. Automatically generated curriculum based reinforcement learning for autonomous vehicles in urban environment[C]//2018 IEEE Intelligent Vehicles Symposium. Changshu, China: IEEE, 2018: 1233-1238.

[17] 刘淼淼.基于危险感知的无信号交叉口直行驾驶决策模型研究[EB/OL]. (2014-12-19)[2020-08-24]. http://www.openits.cn/openPaper/614.jhtml.
Liu M M. Research on straight-through driving decision model at no-signal intersections based on hazard perception[EB/OL]. (2014-12-19)[2020-08-24]. http://www.openits.cn/openPaper/614.jhtml. (in Chinese)

[18] Tsay R S, Tiao G C. Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA models[J]. Journal of the American Statistical Association, 1984, 79(385): 84-96.

（编辑 侯 湘）