计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200300021

基于改进深度强化学习方法的单交叉口信号控制

刘 志 曹诗鹏 沈 阳 杨 曦
浙江工业大学计算机科学与技术学院 杭州 310023
(lzhi@zjut.edu.cn)

摘 要 利用深度强化学习技术实现路口信号控制是智能交通领域的研究热点。现有研究大多利用强化学习来全面刻画交通状态以及设计有效强化学习算法以解决信号配时问题,但这些研究往往忽略了信号灯状态对动作选择的影响以及经验池中的数据采样效率,导致训练过程不稳定、迭代收敛较慢等问题。为此,文中在智能体模型设计方面,将信号灯状态纳入状态设计,并引入动作奖惩系数来调节智能体动作选择,以满足相位最小绿灯时间和最大绿灯时间的约束。同时,结合短期内交通流存在的时序相关性,文中采用优先级序列经验回放(Priority Sequence Experience Replay,PSER)的方式来更新经验池中序列样本的优先级,使得智能体获取与交通状况匹配度更高的前序相关样本,并通过双Q网络和竞争式Q网络来进一步提升DQN(Deep Q Network)算法的性能。最后,以杭州市萧山区市心中路和山阴路形成的单交叉口为例,在仿真平台SUMO(Simulation of Urban Mobility)上对算法进行验证,实验结果表明,提出的智能体模型优于无约束单一状态模型,在此基础上提出的算法能够有效缩短车辆平均等待时间和路口总排队长度,控制效果优于实际配时策略以及传统的DQN算法。

关键词:信号控制;动作奖惩系数;多指标系数加权;优先级序列经验回放;深度Q网络
中图法分类号 TP181

Signal Control of Single Intersection Based on Improved Deep Reinforcement Learning Method

LIU Zhi,CAO Shi-peng,SHEN Yang and YANG Xi
College of Computer Science and Technology,Zhejiang University of Technology,Hangzhou 310023,China

Abstract Using deep reinforcement learning technology to achieve signal control is a researches hot spot in the field of intelligent transportation. Existing researches mainly focus on the comprehensive description of traffic conditions based on reinforcement learning formulation and the design of effective reinforcement learning algorithms to solve the signal timing problem. However, the influence of signal state on action selection and the efficiency of data sampling in the experience pool are lack of considerations,which may result in unstable training process and slow convergence of the algorithm. This paper incorporates the signal state into the state design of the agent model,and introduces action reward and punishment coefficients to adjust the agent's action selection in order to meet the constraints of the minimum and maximum green light time. Meanwhile,considering the temporal correlation of short-term traffic flow,the PSER (Priority Sequence Experience Replay) method is used to update the priorities of sequence samples in the experience pool. It facilitates the agent to obtain the preorder correlation samples with higher matching degree corresponding to traffic conditions. Then the double deep Q network and dueling deep Q network are used to improve the performance of DQN (Deep Q Network) algorithm. Finally, taking the single intersection of Shixinzhong Road and Shanyin Road,Xiaoshan District,Hangzhou,as an example,the algorithm is verified on the simulation platform SUMO (Simulation of Urban Mobility). Experimental results show that the proposed agent model outperforms the unconstrained single-state agent models for traffic signal control problems,and the algorithm proposed in the paper can effectively reduce the average waiting time of vehicles and total queue length at the intersection. The general control performance is better than the actual signal timing strategy and the traditional DQN algorithm.

Keywords Signal control,Action reward and punishment coefficient,Weighted multi-index coefficient,Priority sequence experience replay,Deep Q Network

到稿日期:2020-03-03 返修日期:2020-05-10 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:浙江省公益技术研究计划项目(LGG20F030008);浙江省自然科学基金项目(LY20F030018)
This work was supported by the Public Welfare Technology Research Project of Zhejiang Province,China(LGG20F030008) and Natural Science Foundation of Zhejiang Province,China(LY20F030018).
通信作者:杨曦(xyang@zjut.edu.cn)

1　引言

随着汽车保有量的增长，城市道路的交通拥堵变得越来越严重，这导致了出行成本增加以及环境污染等一系列问题。交通信号控制是治理交通拥堵成本最低的方式，也是交通领域的研究热点。虽然传统控制方法取得了一定成绩，但没能考虑实际交通流状况，效率低下，不能充分发挥路口的通行能力，而强化学习采用马尔可夫决策过程，非常适合面向序列决策的信号控制问题，通过强化学习技术来优化城市道路信号控制方案正在成为重要的研究热点[1]。

利用强化学习方法对交叉口进行配时优化已经取得了一定的效果，但该方法通常采用手工提取特征的方式来简化状态空间假设，只能应对低维输入数据，一旦交通数据量增加，该方法的控制效果将会大打折扣[1-2]。随着大数据、云计算和5G通信技术的发展，能够采集和使用的数据种类和数据量都急剧增长。深度学习技术通过将底层特征组合抽象成高维特征，挖掘批量数据中隐藏的模式，来应对输入数据量较大的情况。近年来，越来越多的学者将两种技术相结合，即将深度强化学习技术应用到交通信号控制问题上，但是绝大多数研究依赖于交通流状态信息，忽略了信号灯状态对智能体动作选择的影响，并且主要采用DQN算法，利用循环神经网络[3]、卷积神经网络[2,4-7]和栈式自编码网络[8]等不同类型的网络来训练智能体，对算法自身的目标网络机制和经验回放机制的优化考虑得较少[4]，而且传统DQN算法采用的均匀采样会忽略经验池中样本数据的时序特性和智能体采样的效率问题[4-9]，容易导致训练过程不稳定，迭代收敛较慢等问题[2]。

为了解决这些问题，本文结合实际四岔路口的特点并遵循信号控制基本原则，引入动作奖惩系数来定义相位时长和动作选择之间的关系，使信号智能体更有效地感知环境状态，学习合理配时动作[1]，随后采用PSER的方法[10]来更新窗口内的序列样本优先级，提升智能体采样效益。最后以杭州市萧山区市心中路和山阴路形成的单交叉口为例，利用本文提出的3DQN_PSER(Double Dueling DQN with Priority Sequence Experience Replay)算法来实现单交叉口的信号控制。

2　相关工作

利用强化学习算法来实现交通信号控制，可以根据其算法属性分成两大类[11]。一类是基于概率策略的方法，主要代表是Actor-Critic方法，Moham等[12]提出将这种方法应用于城市路网，并分析了不同学习率下的平均旅行时间情况，实验结果表明该方法优于传统定时控制，但对于配时动作空间离散的情况，该方法并不适用。另一类是基于价值函数的方法，如Q-learning，SARSA算法等，Adulhai等[13]利用Q-learning算法进行两相位单交叉口在恒定比率交通流和可变交通流情况下的实验。Thrope等[14]采用SARSA算法来分析车辆数目表示、定长划分表示和变长划分表示这3种状态空间设计的信号控制效果。仿真结果表明SARSA算法优于定时控制以及传统的基于规则策略的配时方案。另外，EI-Tantawy等[15]总结了近年来强化学习在交通信号控制领域的发展情况，分析了强化学习智能体的3类核心要素构成的模型设计方案，通过仿真工具Paramics定量分析了强化学习技术中的各要素与信号控制之间的关系，结果表明合理设计强化学习模型要素是实现信号自适应控制的关键。这些传统的强化学习方法通常采用简化的手工提取状态特征，利用Q值表或简单线性函数的方式来估计Q值，由于实际交通环境复杂多变，当状态信息和动作空间增加时，容易造成“维数灾难”问题，目前多数研究通常是通过结合深度学习来抽取高维数据抽象特征，以实现降维，即深度强化学习。

近年来，有很多研究采用深度强化学习技术来实现单交叉口信号控制[4-8]，依据时变交通流的波动，每隔单位时间对相位结构、相位顺序和绿灯时长进行优化。Li等[8]采用稀疏栈式自编码网络来拟合不同状态下每个动作对应的值的输出，并以相序是否切换为动作方案，通过最小化样本误差函数值来寻找最优配时方案。Genders等[5]采用结合卷积神经网络的DQN算法来实现单交叉口信号控制，通过离散交通状态编码(Discrete Traffic State Code，DTSE)技术将各流向的进口道网格化，结合网格内的车辆位置和速度信息来设计状态张量，作为DQN算法的输入，从可选相位方案库中选择合适的动作，将决策点之间的累计车辆延误作为奖励函数。Wan等[6]提出了一种在贝尔曼迭代方程中嵌入动态折扣因子的深度强化学习算法来进行无周期约束的单交叉口智能控制。Gao等[4]采用基于卷积神经网络的DQN算法来提取原始实时流量数据特征，设计固定相序条件下的{0，1}动作调整方案，并通过经验回放和目标网络机制来提升算法的稳定性。Matthew等[7]提出将时间、信号灯状态和排队长度相结合的抽象位级表示法来刻画交通状态，随后通过规则检查模块来确定控制器可以选择的{0，1}动作，最终在两相位单交叉口上验证了深度学习对交通信号控制具有潜在优势。这些研究都是在理想的单交叉拓扑结构上进行仿真实验，设计状态空间时主要采用DTSE技术，未考虑信号灯状态的影响；配时动作设计主要基于固定相序和非固定相序模式，以两决策点之间的等待时间差值[4,9]、延误时间差值[5-6]、排队长度差值[8,16]作为奖励函数，智能体在经过大量迭代训练之后能够找到较优的配时方案，但是忽略了经验池中样本数据的时序特性和智能体采样的效率问题，容易造成训练过程不稳定和收敛速度较慢等问题。

3　3DQN_PSER算法框架

3DQN_PSER算法在处理信号控制问题时，其算法框架和具体网络参数如图1所示，主要分成3部分：主网络、目标网络和经验池。主网络部分主要是通过卷积神经网络来实现，首先经过3层卷积操作，将结果展平后与经全连接层1处理的信号灯状态信息连接，随后采用Dueling DQN技术[17]将全连接层2提取的抽象特征分流到两支路，即状态网络流$V(s)$和动作优势网络流$A(s,a)$，最终将两支路聚合起来得到每个动作对应的Q值。目标网络和主网络内部结构完全相同，用来辅助主网络的训练，避免强化学习输出的Q值反复震荡。此外，在采用Double DQN技术[18]时会借助主网络和目标网络来将信号智能体动作选择和值函数计算解耦，主网络选择动作，目标网络计算Q值。经验池主要用来存储有限

的训练样本，将每次和环境交互的奖励和状态更新情况记录下来，以便后期计算目标值。本文中的每条经验记录包含决策点前后状态 s 和s'、采取的动作 a、反馈的奖励 r、是否结束标志 $flag$、前后决策点配时动作所对应的奖惩系数 e 和e'、前后决策点的信号灯执行状态 φ 和φ'（由 0 和 1 组成，1 表示该流向车道为激活车道，即$\langle s,a,s',r,flag,e,e',\varphi,\varphi'\rangle$）。

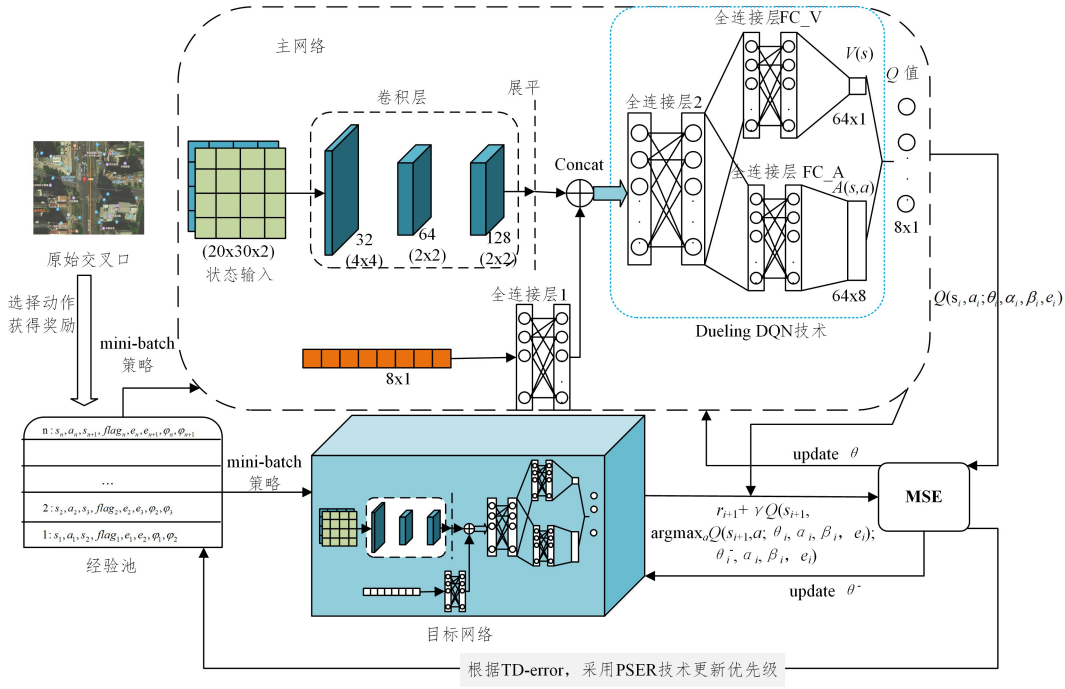


图 1 3DQN_PSER 算法的网络模型架构

Fig. 1 Network model architecture of 3DQN_PSER algorithm

4 基于深度强化学习的信号智能体设计

本文结合单交叉口信号控制的基本思想来设计深度强化学习的相关要素，通过单位延长时间来调整相位结构、顺序和绿灯时长，这里以杭州市萧山区市心中路-山阴路为例，根据交通控制的相关标准以及信号灯配置要求来设计信号智能体所处的环境状态、对应的配时动作和奖励函数。

4.1 状态表示

为了尽可能多地保留车辆通行信息，精确地描述交通状况，本文采用近些年较为流行的 DTSE 技术来对进口道进行网格化，并通过“拼接”的方式组成交通流状态信息输入[16]。对于有 H 个进口道的某典型四岔路口而言，其状态空间大小为 $H\times(L/c)\times Y$，其中，Y 表示刻画交通状态的指标数目，本文以车辆位置和对应的速度信息为指标，L 表示进口道检测器探测的区域长度，c 表示网格长度，如图 2(a)所示。

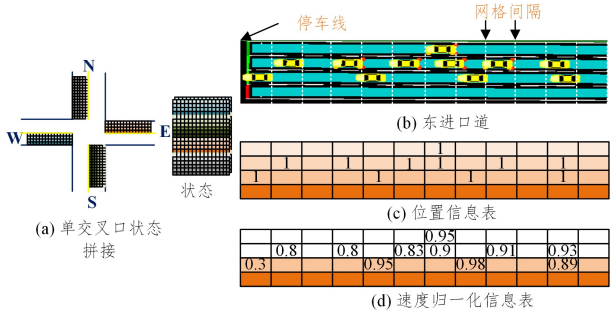


图 2 交通流状态设计图

Fig. 2 Design diagram of traffic flow situation

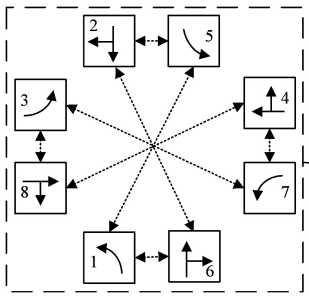
相比直接将交叉口图像信息进行输入[2]，这种方式能够压缩数据维度，去除冗余信息，从而加快训练速度。为了更清晰地描述这种交通流状态表示法，本文以东进口道为例，车辆分布情况如图 2(b)所示，得到其对应的二进制位置信息表和速度归一化信息表，如图 2(c)、图 2(d)所示。

除了路口当前的车辆运行和分布情况会反映交通状态之外，信号灯的当前状态也会影响交叉口的通行效率[14]，因此本文针对该典型路口的 8 个流向，设计一维二进制数组来表示信号灯状态，若信号灯当前状态 $\varphi=[1,0,0,0,1,0,0,0]$，根据对应位置对应流向的原则，可知流向 1 和流向 5 为绿灯信号，拥有车辆通行权。该部分信息作为 3DQN_PSER 算法模型的另一状态信息输入，首先通过全连接层 1 处理，随后与经过一系列处理后的交通流状态信息进行 Concat 连接，如图 1 所示，这种方式能够更加全面地刻画实时交通环境，挖掘路口交通环境的内部潜在特征。

4.2 动作调整

信号智能体中，动作调整指相位方案和配时方案的调整。在该路口，遵循工程上划分相位的安全通行原则，列出在无冲突情况下的相位方案库，如图 3 所示，右转往往是驾驶员根据当时的交通状况所做的决定，此处不讨论，动作方案集合为 $A=\{1,2,3,4,5,6,7,8\}$，智能体在每个决策点都会从中选择一种相位方案，如果选择的动作和当前相位方案相同，则执行当前绿灯相位τ_g s，否则需要先执行过渡相位（黄灯相位）τ_y s，根据在决策点的不同选择来执行对应的相位方案[4]，在相位切换过程中，会涉及灯色状态的变化，具体的灯色状态转移关系如表 1 所列（其中，g 表示右转绿灯，G 表示直行或左转绿

灯，r 和 y 分别表示红灯和黄灯，且灯色状态序列按照“东南西北”这样的车道顺序构成）。



方案一	2，5
方案二	4，7
方案三	1，6
方案四	3，8
方案五	2，6
方案六	4，8
方案七	1，5
方案八	3，7

图 3　典型四岔路口的相位方案组合图

Fig. 3　Phase scheme combination diagram in typical four-intersection road

表 1　相位切换和信号灯变化情况

Table 1　Phase switching and signal change situation

方案	过渡相位灯色状态	切换后灯色状态
方案一	—	*grrrrrrrrrrrrrgGGGGG*
方案二	*grrrrrrrrrrrryyyyyy*	*gGGGgrrrrrrrrrrrrrr*
方案三	*yrrrrrrrrrrrryyyyyy*	*rrrrgGGGGGgrrrrrrrrr*
方案四	*yrrrrrrrrrrrrgyyyyy*	*rrrrrrrrrgGGGgrrrrr*
方案五	*yrrrrrrrrrrrrrrGGGGy*	*rrrrgGGGGrrrrrgGGGGr*
方案六	*grrrrrrrrrrrrryyyyyy*	*gGGrrrrrrrrgGGrrrrrrr*
方案七	*grrrrrrrrrrrrryyyyyG*	*grrrrrrrrGgrrrrrrrrG*
方案八	*yrrrrrrrrrrrrrgyyyyy*	*rrrGgrrrrrrrrGgrrrrr*

注：当前方案为方案 1，灯色状态为 *grrrrrrrrrrrrrgGGGGG*

在信号控制过程中，频繁切换相位可能会强行中断连续到达的交通流，从而削弱了交通系统的稳定性，对此，本文根据相位最小绿灯时间和最大绿灯时间进行约束，设计了动作奖惩系数 e，该参数最终会纳入 3DQN_PSER 算法的 Q 值的计算，使信号智能体更加合理地选择配时动作，动作奖惩系数 e 的表达式如下：

$$e=\begin{cases}\phi, & ph_1\leqslant G_{\min}\ \&\ ph_2\leqslant G_{\min}\\ \dfrac{ph_1}{G_{\min}}*\phi, & ph_1\leqslant G_{\min}\ \&\ G_{\min}<ph_2\leqslant G_{\max}\\ 0, & G_{\min}<ph_1\leqslant G_{\max}\ \&\ G_{\min}<ph_2\leqslant G_{\max}\\ \dfrac{G_{\max}-ph_2}{G_{\max}}*\phi, & G_{\min}<ph_1\leqslant G_{\max}\ \&\ ph_2>G_{\max}\\ -1*\phi, & G_{\max}<ph_1\ \&\ ph_2>G_{\max}\end{cases} \tag{1}$$

其中，$G_{\min}$ 和 $G_{\max}$ 分别表示相位最小绿灯时间和相位最大绿灯时间；Φ 表示奖惩尺度，用于衡量对智能体动作选择的影响程度，ph_1 和 ph_2 表示相位方案中两流向的绿灯时间，如图 3 所示，且式(1)的前提为 $ph_1<ph_2$。

信号智能体在进行动作选择时会采用强化学习中的探索和利用机制，本文采用 ε-greedy 策略，令选择探索的可能性为 ε_m，选择利用的可能性为 $1-\varepsilon_m$，随着训练步数的增加，ε_m 值越来越小，即选择利用最优的相位方案可能性会越来越大，其公式如下：

$$\varepsilon_m=\max\{\varepsilon_{\min},\varepsilon_{\text{init}}-\frac{m}{M}(\varepsilon_{\text{init}}-\varepsilon_{\min})\} \tag{2}$$

其中，m 表示当前的训练步数，M 表示训练的总步数，ε_m，$\varepsilon_{\text{init}}$，$\varepsilon_{\min}$ 分别表示当前的、初始的以及最终的 ε 值。

4.3　奖励函数

在强化学习处理交通信号控制问题中，奖励函数一般通过等待时间差值[4,9]、延误时间差值[5-6]、排队长度差值[8,16]等指标进行设计。本文采用多指标系数加权[2]的方案，从而更高效地引导智能体学习：

(1)各车道在该决策点的排队长度之和 r_{queue}。

(2)相邻决策点之间的累计车辆等待时间之差 $r_{waitTime}$，若当前处于决策点 $t+1$，此时的累计等待时间 $r_{waitTime}=w_{t+1}-w_t$，如果 $r_{waitTime}<0$ 则表明这段时间路网比之前畅通，否则表明路网拥堵加重。

(3)各车道在该决策点的刹车数量之和 $r_{halting}$。

(4)当前决策点选择的动作是否会导致相位切换 r_{phase}，如果切换，则 $r_{phase}=1$，否则 $r_{phase}=0$。

综合以上因素，并结合各自对应的权重系数 k_1，k_2，k_3，k_4，加权得到最终的奖励：

$$r=k_1*r_{queue}+k_2*r_{waitTime}+k_3*r_{halting}+k_4*r_{phase} \tag{3}$$

5　基于优先级序列经验回放的 DQN 算法改进

DQN 算法的核心是引入了经验回放和目标网络机制，其中目标网络机制主要是为了稳定训练过程，Hasselt 等[18]提出的 Double DQN 算法就是对该机制进行优化，经验回放机制是为了打破数据之间的强关联性，以满足监督学习中的独立同分布条件，并采用小批量抽取(mini-batch 策略)训练的方式减少学习所需要的经验。为了进一步提高“有效”数据的利用效率和训练速率，Schaul 等[19]提出优先级经验回放(Priority Experience Replay，PER)的方法来更新经验池中的序列样本优先级，该方法是利用序列样本 i 的优先级 p_i 来计算其在 k 个样本中被采样的概率 $P(i)=\frac{p_i^{\sigma}}{\sum_k p_k^{\sigma}}$，从而计算得到其对应的重要性采样权重 $w_i=(\frac{1}{N}*\frac{1}{P(i)})\mu$，以对误差进行有效补偿[19]，指数 σ 和 μ 分别表示优先级使用程度和采样权重系数，取值范围为[0，1]，N 表示经验池容量的大小。

另外，在智能交通系统中，短时交通流预测是非常重要的研究方向，原因在于交通流数据是一种非平稳随机时间序列，能在连续的时间内呈现一定的规律性，并且具有明显的趋势性，简而言之，短期交通流具有时序相关性。对于序列决策的交通信号控制问题，交通状态相似的样本以及该时刻之前的某段连续样本数据对智能体此刻的训练都有较大的被采样隐藏价值。因此，本文采用 PSER 的方式来更新经验池中序列样本的优先级，具体方法如下。

令存放在经验池中的序列样本轨迹为：T_0，T_1，T_2，…，T_n，其对应的优先级为：p_0，p_1，p_2，…，p_n，由于本文引入信号灯状态 φ 以及配时动作奖惩系数 e，因此某时刻的 i 序列样本存储信息 T_i 为 $\{s_i,a_i,s_{i+1},r_i,flag_i,e_i,e_{i+1},\varphi_i,\varphi_{i+1}\}$，在 DQN 算法中，优先级通常指目标 Q 值和实际 Q 值之间的时序差分误差(Temporal-Difference error，TD-error)，这里根据采用的主网络参数 θ 和目标网络参数 θ^-，可以得到 TD-error 的计算公式：

$$\delta_i^{DQN}=r_{i+1}+\gamma\max_a Q(s_{i+1},a;\theta_i^-)-Q(s_i,a_i;\theta_i) \tag{4}$$

其中，$r_{i+1}+\gamma\max_a Q(s_{i+1},a;\theta_i^-)$ 表示目标 Q 值，$Q(s_i,a_i;\theta_i)$ 表示实际 Q 值。由于 DQN 算法在计算 TD-error 时往往采用相同的 Q 网络来计算值函数和选择动作，容易造成值函数过估

计，因此本文采用 Double DQN 技术来将动作选择和 Q 值计算分配到主网络和目标网络上[18]，从而得到 Double DQN 下的 TD-error：

$$\delta_i^{DDQN}=r_{i+1}+\gamma Q(s_{i+1},\arg\max_a Q(s_{i+1},a;\theta_i);\theta_i^-)-Q(s_i,a_i;\theta_i) \tag{5}$$

为了避免配时动作对短期内的交通状况影响过大，这里采用 Dueling DQN 技术[17]来将网络分成状态网络流 $V(s)$ 和动作优势网络流 $A(s,a)$，从而稳定训练过程，其对应的全连接层 FC_V 网络参数为 α，全连接层 FC_A 网络参数为 β，如图 1 所示。本文结合配时动作奖惩系数可以得到式(6)，结合 Double DQN 技术可以得到 3DQN 算法中值函数的近似优化目标公式(7)以及对应的 TD-error 计算公式(8)：

$$Q(s_i,a_i;\theta_i,\alpha_i,\beta_i,e_i)=Q(s_i,a_i;\theta_i,\alpha_i,\beta_i)+e_i \tag{6}$$

$$Y_i^{3DQN}=r_{i+1}+\gamma Q(s_{i+1},\arg\max_a Q(s_{i+1},a;\theta_i,\alpha_i,\beta_i,e_i);\theta^-{}_i,\alpha_i,\beta_i,e_i) \tag{7}$$

$$\delta^{3DQN}{}_i=Y^{3DQN}{}_i-Q(s_i,a_i;\theta_i,\alpha_i,\beta_i,e_i) \tag{8}$$

如果一组序列样本出现优先级皆很小甚至全为零的情况，将难以影响当前决策点先前的样本数据优先级，容易造成“优先级崩溃”以及智能体重采样[10]。为避免上述问题，PSER 算法首先通过 $p_i=|\delta_i^{3DQN}|+o$ 的方式来避免优先级为零，其中 o 是一个很小的常数，一般设成 0.0001，随后通过参数 η 来调整优先级更新速度，解决“优先级崩溃”问题，按照 $p_i\leftarrow\max(|\delta_i^{3DQN}|+o,\eta*p_i)$ 的方式对当前序列样本 i 进行优先级更新，最后以其为起始点，回放窗口 W 范围内连续的先前样本数据优先级，更新规则如下：

$$p_{i-1}=\max(\rho^1\ p_i,p_{i-1}) \tag{9}$$

$$p_{i-2}=\max(\rho^2\ p_i,p_{i-2}) \tag{10}$$

$$\vdots$$

$$p_{i-(W-1)}=\max(\rho^{W-1}\ p_i,p_{i-(W-1)}) \tag{11}$$

其中，ρ 为衰减系数，表示相邻决策点之间的优先级影响程度。

至此，由式(4)－式(11)可以看出 3DQN_PSER 算法计算和更新优先级的过程。另外，对于一个深度强化学习算法，网络参数的实时更新是智能体寻找最优配时方案的前提，本文使用均方差(Mean Square Error，MSE)作为损失函数，通过误差反向传递的方式来更新一轮网络模型参数，因此算法采用的损失函数为：

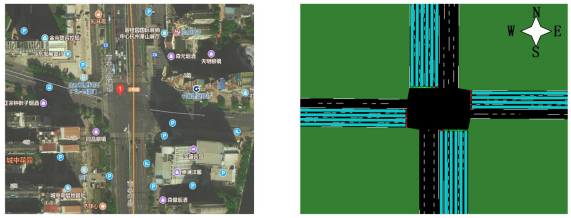
$$J=\frac{1}{B}\sum_{i=1}^{B}w_i(\delta_i^{3DQN})^2 \tag{12}$$

其中，B 表示抽取的批量样本数目。

6 案例分析和实验结果

6.1 案例分析和实验配置

为了测试本文算法，本实验利用仿真工具 SUMO 提供的 Traci 接口，将状态信息发送给用 python 进行编码的信号控制器，以杭州市萧山区市心中路-山阴路这个路口为实验场景，如图 4 所示，该路口是典型的四岔结构，其中南北向各有 6 个进口道与 4 个反向车道，东西向各有 4 个进口道和 2 个反向车道，且每个方位均有一个右转专用车道和左转专用车道来合理分配车流，具体的车道划分按照“东南西北”的原则，依次标注为 $L_0\cdots L_{19}$。此外，为了尽可能复现交通实际情况，本实验采用 2018-07-22 至 2018-07-28 这一周的流量数据集(流量记录超过 3.7 万条)，并对其进行预处理后作为实验的路由文件，对原始数据按照小时为单位进行统计，得到的流量分布情况如图 5 所示，本实验主要针对早晚高峰时段的交通量进行探讨，即7:00－9:00 和 18:00－20:00。



(a)实际路口结构图 (b)仿真路口结构图

图 4 实际单交叉口仿真图

Fig. 4 Simulation diagram of actual single intersection

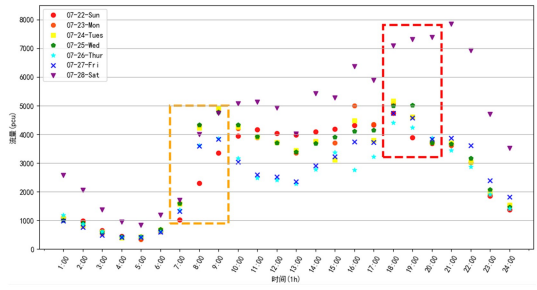


图 5 市心中路-山阴路的交通流情况

Fig. 5 Traffic flow situation from Shixinzhong Road to Shanyin Road

实验部分对比了传统 DQN 算法、实际配时策略(Base Case，BC)和结合 Schaul 等[19]提出的 PER 策略的 3DQN 算法[9]，即 3DQN_PER 算法。与本文相关的算法的参数如表 2 所列。每轮仿真运行 100 次迭代，其中每次迭代运行两小时的流量数据，迭代初始阶段会先进行两周期的路网缓冲，不纳入实验的最终结果。

表 2 实验参数设置

Table 2 Experimental parameter settings

参数	取值
网格长度 c/m	5
道路规定的上限速度 v_{max}/(m/s)	13.89
G_{min}，G_{max}	15，60
折扣系数 γ	0.98
衰减系数 ρ	0.65
学习率	0.001
批度大小 batch_size	64
状态矩阵大小	20 * 30 * 2
动作空间	8
回放窗口 W	10
优化器	AdamOptimizer
绿灯相位 τ_g，黄灯相位 τ_y	3，3
奖惩尺度 Φ	50
权重系数 k_1，k_2，k_3，k_4	－0.25，－0.25，－1，－0.25
实际配时方案 7:00－9:00	44，26，27，23
实际配时方案 18:00－20:00	68，37，40，35

6.2 实验结果

6.2.1 模型改进验证

为了探究融合信号灯状态和引入动作奖惩系数对智能体的影响，本文进行同等条件下的对比实验，将仅采用 DTSE 技术设计状态空间，不考虑动作奖惩系数对 Q 值的影响，并且

其余实验参数均保持一致，定义为无约束非单一状态模型设计。模型改进前后的性能情况可以通过奖励值的分布情况进行分析，如图 6 所示（图中每个点都代表一轮迭代的平均值），可以了解到无约束单一状态的 3DQN_PSER 算法在 100 轮训练过程中前期的奖励值跳跃明显，训练后期会在某范围内不停震荡，相比之下，本文的智能体模型在 100 轮迭代过程中收敛效果较好。

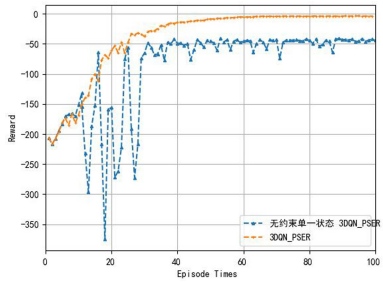


图 6　模型改进前后的奖励分布情况（2018-07-22 7:00－9:00）

Fig. 6　Reward distribution before and after model improvement (2018-07-22 7:00－9:00)

6.2.2　交通控制效果对比

为了探究 3DQN_PSER 算法的信号控制效果，本文以 2018 年 07 月 22 日晚高峰时段的实验数据进行算法对比实验。首先通过强化学习中的奖励值进行分析，如图 7 所示，DQN 算法在训练的过程中震荡较为明显，体现了强化学习是不断探索的过程。

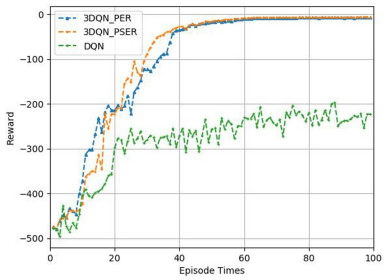


图 7　所有轮迭代的奖励分布情况（2018-07-22 18:00－20:00）

Fig. 7　Distribution of rewards for all iterations (2018-07-22 18:00－20:00)

通过对 3DQN_PER 和 3DQN_PSER 算法在训练过程中的奖励值分布情况进行分析发现，这两种算法均有收敛的趋势，说明都适用于交通信号控制问题。随后通过车辆平均等待时间和路口总排队长度这两个指标来分析由交叉口形成的“十”字路网区域的通行情况以及信号灯的使用效率，从而更清晰直观地反映信号控制的效果，如图 8、图 9 所示，DQN 算法在训练过程中振荡下降，最终控制效果不佳，3DQN_PER 算法和 3DQN_PSER 算法均在 60 次迭代之后趋于稳定，收敛的最终效果相近。另外，如表 3 所列，相比 3DQN_PER 算法和 3DQN_PSER 算法，BC 配时方案的控制效果存在明显的局限性。3DQN_PER 和 3DQN_PSER 算法在该路口形成的路网区域车辆平均等待时间方面相比 BC 控制策略减少了 57.69％左右，在排队长度方面，BC 控制策略的结果为 151.51 m，而本文采用的 3DQN_PSER 算法最终的收敛值为 38.89 m 左右，可以得到 3DQN_PSER 算法的排队长度缩减为 BC 控制策略的 25.6％。从以上结果来看，对经验回放部分的优化能够明显提升算法的信号控制效果，3DQN_PER 和 3DQN_PSER 算法明显优于 DQN 算法和 BC 控制策略，能够有效地提升该路口附近路网的通行能力。

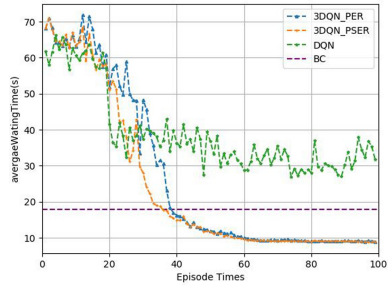


图 8　平均等待时间的分布情况（2018-07-22 18:00－20:00）

Fig. 8　Average waiting time distribution (2018-07-22 18:00－20:00)

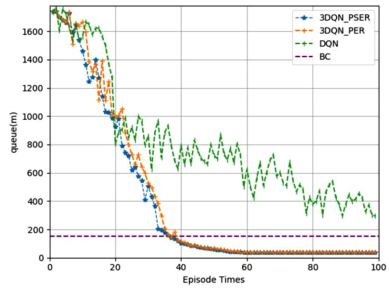


图 9　排队长度分布情况（2018-07-22 18:00－20:00）

Fig. 9　Queue length distribution (2018-07-22 18:00－20:00)

表 3　各算法的评价指标情况

Table 3　Evaluation indexes of each algorithm

算法	车辆平均等待时间/s	总排队长度/m
BC	17.94	152.51
DQN	31.69	294.87
3DQN_PER	7.86	39.78
3DQN_PSER	**7.59**	**38.89**

6.2.3　晚高峰交通控制效果

为了进一步探究算法的性能，本文选取了一周的晚高峰时段流量数据进行测试。实验结果如图 10 所示，相比 3DQN_PER 算法、DQN 算法和 BC 配时策略，3DQN_PSER 算法最终的车辆平均等待时间总是最低的，能有效提升交叉口信号灯的利用率，而实际配时策略 BC 所得到的车辆平均等待时间总是较长，说明实际路口的信号配时仍存在较大的优化空间，除 2018-07-22 对应的周日外，DQN 算法也能比实际配时方案 BC 取得更优的控制效果。

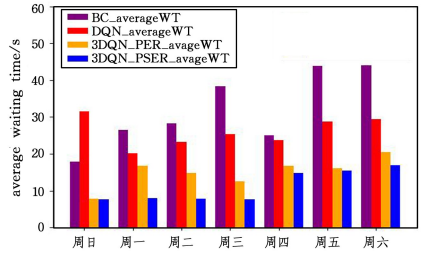


图 10　一周的平均等待时间情况（2018-07-22－2018-07-28 18:00－20:00）

Fig. 10　Average waiting time for a week (2018-07-22－2018-07-28 18:00－20:00)

结束语 针对现有的深度强化学习方法在实现信号控制时容易忽略信号灯状态对智能体动作选择的影响以及经验池中的数据采样效率，本文提出了一种改进的深度强化学习算法来实现单交叉口的信号控制。本文首先在模型构建中结合交通流状态和信号灯状态来作为算法的输入，并引入了动作奖惩系数的概念来约束信号配时，还采用多指标系数加权进行奖励设计。随后在此模型的基础之上，采用 PSER 策略更新经验池中序列样本优先级，最终将 3DQN_PSER 算法应用在真实的单交叉口环境中，取得了良好的控制效果，优于实际配时策略 BC 以及传统 DQN 算法。目前的研究只限于单交叉口的信号控制问题，对于复杂的大规模城市路网而言，如何有效进行多交叉口的信号协调控制[20]，是下一步研究的重点。

参 考 文 献

[1] HUO Y S. A Summary of Traffic Signal Control Method Based on Reinforcement Learning[C]// The 12th Annual Conference of China Intelligent Transportation. 2017:858-865.

[2] SUN H, CHEN C L, LIU Q, et al. Traffic Signal Control Method Based on Deep Reinforcement Learning [J]. Computer Science, 2020, 47(2):169-174.

[3] ZENG J, HU J, ZHANG Y. Adaptive Traffic Signal Control with Deep Recurrent Q-learning[C]// IEEE Intelligent Vehicles Symposium. 2018:1215-1220.

[4] GAO J, SHEN Y, LIU J, et al. Adaptive Traffic Signal Control: Deep Reinforcement Learning Algorithm with Experience Replay and Target Network [J]. arXiv:1705.02755, 2017.

[5] GENDERS W, RAZAVI S. Using a Deep Reinforcement Learning Agent for Traffic Signal Control [J]. arXiv: 1611.01142, 2016.

[6] WAN C H, HWANG M C. Value-based deep reinforcement learning for adaptive isolated intersection signal control [J]. IET Intelligent Transport System, 2018, 12(9):1005-1010.

[7] MATTHEW M, LIPING F, GUANGGYUAN P. Adaptive Traffic Signal Control with Deep Reinforcement Learning- An Exploratory Investigation [C]// Transportation Research Board 97th Annual Meeting. 2019:18-33.

[8] LI L, LYU Y, WANG F Y, et al. Traffic Signal Timing via Deep Reinforcement Learning [J]. IEEE/CAA Journal of Automatic Sinica, 2016, 3(3):247-254.

[9] LIANG X, DU X, WANG G, et al. A Deep Reinforcement Learning Network for Traffic Light Cycle Control [J]. IEEE Transactions on Vehicular Technology, 2019, 68(2):1243-1253.

[10] BRITAIN M, BERTRAM J, YANG X, et al. Prioritized Sequence Experience Replay [J]. arXiv:1905.12726, 2019.

[11] YAU K, QADIR J, KHOO H, et al. A Survey on Reinforcement Learning Models and Algorithms for Traffic Signal Control [J]. ACM Computing Surveys, 2017, 50(3):1-38.

[12] ASLANI M, SEIPEL S, SAADI M, et al. Traffic signal optimization through discrete and continuous reinforcement learning with robustness analysis in downtown Tehhran [J]. Advanced Engineering Informatics, 2018, 38:639-655.

[13] ADULHAI B, PRINGLE R, KARAKOULAS G. Reinforcement learning for true adaptive traffic signal control [J]. Journal of Transportation Engineering, 2003, 129(3):278-285.

[14] THROPE T L, ANDERSON C W. Traffic light control using SARSA with three state representations [R]. Technical Report, IBM Corporation, 1996.

[15] EI-TANTAWY S, ABDULHAI B, ABDELGA-WAD H. Design of Reinforcement Learning Parameters for Seamless Application of Adaptive Traffic Signal Control [J]. Journal of Intelligent Transportation Systems, 2014, 18(3):227-245.

[16] LAI J H. Traffic Signal Control based on Double Deep Q-learning Network with Dueling Architecture [J]. Computer Science, 2019, 46(S2):117-121.

[17] WANG Z, SCHAUL T, HESSEL M, et al. Dueling Network Architectures for Deep Reinforcement Learning[C]// Proceeding of the 33rd International Conference on Machine Learning. 2016:1995-2003.

[18] VAN HASSELT H, GUEZ A, SILVER D. Deep Reinforcement Learning with Double Q-learning [C]// Association for the Advance of Artificial Intelligence. 2016:2094-2100.

[19] SCHAUL T, QUAN J, ANTONOGLOU I, et al. Prioritized Experience Replay [C]// Proceedings of the 4th International Conference on Learning Representations. 2016:322-355.

[20] FOERSTER J N, ASSAEL Y M, DE FREITAS N, et al. Learning to Communicate with Deep Multi-Agent Reinforcement Learning [C]// 29th Neural Information Processing Systems. 2016:10-22.

LIU Zhi, born in 1969, Ph.D, professor, is a member of China Computer Federation. Her main research interests include intelligent transportation and image processing.

YANG Xi, born in 1982, Ph.D, associate professor. His main research interests include control and optimization theory, intelligent transportation systems.