一种改进的深度确定性策略梯度网络交通信号控制系统

刘利军[1,2]，王　州[1]，余　臻[1]
(1. 厦门大学航空航天学院，厦门 361101；2. 厦门大学深圳研究院，深圳 518057)

摘　要：交通信号系统控制着城市车辆运行秩序，其效率高低直接影响了社会经济的发展. 以十字路口的交通信号控制系统为研究对象，基于深度确定性策略梯度网络 DDPG 提出了一种改进算法. 结合交通环境的特点设计了特征增强和样本去重算法提高算法的性能. 通过对实际交通系统运行情况进行调研，基于 SUMO 仿真环境搭建了交叉路口交通仿真平台. 利用 FEPG 算法控制交通信号，实现了车辆的高效通行. 实验结果表明，该算法能够有效地降低车辆等待时间，减少车辆的污染排放.

关键词：交通信号控制；强化学习；样本去重；特征增强
中图分类号：TP181　　**文献标识码**：A　　DOI：10.19907/j.0490-6756.2021.043003

Improved deep deterministicpolicy gradient network traffic signal control system

LIU Li-Jun[1,2], *WANG Zhou*[1], *YU Zhen*[1]
(1. School of Aerospace Engineering, Xiamen University, Xiamen 361101, China;
2. Shenzhen Research Institute of Xiamen University, Shenzhen 518057, China)

Abstract: The traffic signal system controls the order of urban vehicles, and its efficiency directly affects the development of the social economy. This paper takes the traffic signal control system at the intersection as the research object, an improved algorithm is introduced based on the deep deterministic policy gradient (DDPG), in which the feature enhancement and sample deduplication algorithms combined with the characteristics of the traffic environment are designed to improve the performance of the algorithm. By analyzing the operation of the actual traffic system, a traffic simulation platform for intersections is set up based on the SUMO simulation environment. The FEPG algorithm is used to control traffic signals to achieve efficient vehicle traffic. Experimental results show that the algorithm can effectively reduce vehicle waiting time and reduce vehicle emissions.

Keywords: Traffic signal control; Reinforcement learning; Sample deduplication; Feature enhance

收稿日期：2020-03-24
基金项目：国家自然科学基金(61304110)；广东省自然科学基金(2018A030313124)；深圳市基础研究面上项目(JCYJ20190809163009630)；上海市自然科学基金(18ZR1443200)
作者简介：刘利军(1985－)，男，博士研究生，副教授，研究领域为非线性优化与控制，强化学习.
通讯作者：余臻. E-mail：yuzhen20@xmu.edu.cn

1 引　言

随着全球人口的快速增长,以及城市化进程的发展,专家预计 21 世纪的城市人口将急剧增加. 当务之急是城市有效地管理其基础设施以应对这一问题. 设计现代化城市时,一个关键的考虑因素是开发智能交通管理系统. 交通管理系统的主要目标是减少交通拥堵. 高效的城市交通管理能够节省时间和财务,并减少二氧化碳等大气污染物排放到大气中[1]. 已经有许多学者提出方案解决这个问题[2]. 主要道路的交叉路口一般是通过交通信号灯管理. 低效率的交通信号灯控制会导致许多浪费,增加车辆发生事故的风险[3]. 现有的交通灯控制信号按照固定程序不考虑实时交通流量,效率低. 因此,研究人员提出了许多改进方案,这些方案可以分为三类:第一类是预定控制程序,参考历史数据制定交通信号切换时间;第二类是利用传感器检测来往车辆,用以延长或缩短信号切换时间;第三类是自适应信号控制,根据交叉路口的当前状态自动切换[4]. 本论文对第三类控制方法展开研究,利用深度强化学习方法设计一种十字路口交通信号智能控制方法.

近年来,很多研究者结合深度学习和强化学习技术来处理复杂的优化问题,例如 Atari 2600 游戏[5],围棋[6]等. 1997 年 Thorpe[7]首次将强化学习的方法应用到交通信号控制,大家开始意识到强化学习为解决非线性、不确定性的复杂路网问题提供了一种新的思路. 随着人工智能的快速发展,深度强化学习被应用到交通控制中. 2016 年 Li 等人将深度强化学习应用于交通信号控制中,降低了车辆 14%的等待时间[8]. 2018 年 Liang 等人[3]改进 Dueling DQN[9]和 DoubleDQN[10],将车辆等待时间降低了 25.7%. 本论文通过分析交通环境特点,设计了特征强化策略梯度算法(Feature Enhance DDPG,FEPG)算法并将其应用于交通信号控制系统中.

2 背　景

2.1 强化学习

强化学习的基本结构由智能体 Agent 和外界环境组成. Agent 通过执行动作,从环境获得下一状态和奖励值. 不断循环该过程,直到满足一定条件为止. 通常一个强化学习问题可以视为马尔可夫决策过程(Markov Decision Process, MDP)[11]. MDP 将强化学习任务定义为元组$<S,A,P,R,\gamma>$. 其中,S 是环境状态的集合;A 是 Agent 的动作空间集合;P 是状态之间的转移概率;R 是奖励函数;$\gamma\in[0,1]$是奖励值的折扣因子. Agent 从状态 s 采取动作 a 达到状态 s'并获得奖励 r 表示为(s,a,r_t,s'). 我们将 Agent 执行动作的方法称之为策略 π. 强化学习的目标是最大化累积奖励. 式(1)定义了累积奖励 R_t. 下标 t 表示 Agent 执行策略 π的第 t 步.

$$R_t=r_{t+1}+\gamma r_{t+2}+\gamma^2 r_{t+3}+\cdots=\sum_{k=0}^{\infty}\gamma^k r_{t+k+1} \tag{1}$$

以累积奖励为出发点,诞生了基于值的强化学习方法如 Q-learning[12]等. 利用神经网络拟合 Q 函数. 在策略 π下 Q 函数的定义为式(2).

$$Q^{\pi}(s,a)=E[R_t\mid s_t=s,a_t=a,\pi] \tag{2}$$

结合卷积神经网络(Convolutional Neural Networks, CNN)[13],循环神经网络(Recurrent Neural Network, RNN)[14]强化学习的应用场景变得更为丰富. 谷歌 DeepMind 工作室[15]于 2015 年提出了结合基于值的和基于策略方法优点的深度确定性策略梯度算法(Deep Deterministic Policy Gradient, DDPG). 本论文将改进 DDPG 网络利用特征增强算法和样本去重环节提升网络性能,最后,将算法应用于交通信号控制中.

2.2 交通环境模型

十字交叉路口交通环境十分复杂,且交通信号灯的控制十分重要. 本文利用 SUMO[16]仿真环境验证算法有效性. 道路模型的参数如表 1 所示. 车辆的污染排放模型可参考文献[17].

表 1　交通环境参数

Tab. 1　Traffic environment parameters

道路参数	取值	车辆参数	取值
入口车道数	3	车辆长度/m	4~10
出口车道数	3	启动加速/(m/s²)	0.8~2.6
车道长度/m	200	刹车加速/(m/s²)	4.0~4.5
限速/(km/h)	30	驾驶员熟练度	0.6~0.9

2.2.1 状态 s 描述　如图 1 所示十字路口状态模型. 路口的 4 个车道被划分为 4 个单元,每个单元的长度为 50 m. 单元格下方有白,灰色两个方格. 若某个单元内有 n 辆车,白色方格内数字则为 n,灰色方格代表该单元内车辆的平均速度,若 n 为 0 则平均速度取 0,否则取 n 辆车速度的平均值如式(3). 因此状态维度为 32,如表 2 所示.

表 2　状态向量表

Tab. 2　State vector table

路口	路口 1				路口 2				路口 1				路口 2			
车辆数	1	0	2	1	1	0	1	1	0	0	0	2	1	0	2	0
平均速度/(m/s)	9.7	0	6.3	0.1	7.4	0	3.4	0.5	0	0	0	2	11	0	8.5	0

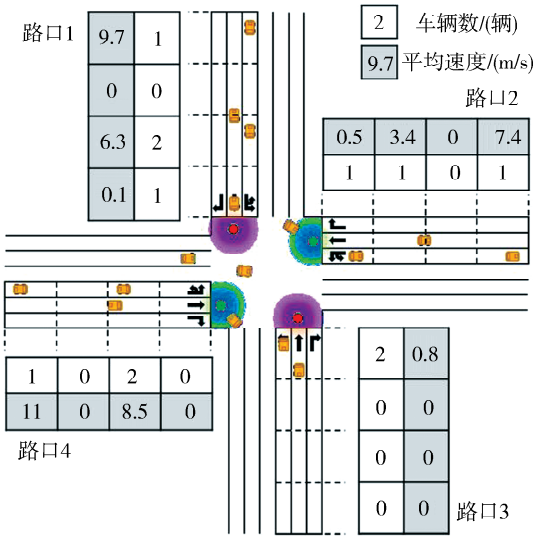

图 1　状态描述

Fig. 1　State description

$$\overline{v}=\frac{1}{n}\sum_{i=0}^{n}v_i \tag{3}$$

2.2.2 动作 a 描述　如表 3 所示，十字交叉路口的交通信号灯由 6 个信号相位组成. 相位按顺序从 1 号变化到 6 号不断循环. 每个相位由 4 个字母组成. 每个字母代表一个路口信号灯的状态. 其中：G 为绿灯，允许通行；R 为红灯，禁止通行；Y 为黄灯，注意通行. 例如相位 GRGR 表示此时 1、3 路口绿灯，2、4 路口红灯. 将 1 和 4 相位时间作为动作 a. 考虑到车辆行驶速度，其范围为 5～50 s. 其余相位设为安全的固定值.

表 3　交通信号相位

Tab. 3　Traffic signal phase

相位	值	持续时间/s	是否可控
1	GRGR	30	可控
2	YRYR	5	不可控
3	RRRR	10	不可控
4	RGRG	30	可控
5	RYRY	5	不可控
6	RRRR	10	不可控

2.2.3 奖励值 r 描述　奖励值在强化学习中的作用为提供网络训练方向，本论文目的为提高车辆在十字路口的通行效率，减少车辆等待时间. 每辆车的等待时间定义为其行驶速度为 0 时所持续的时间，记作 d，所有路口等待的所有车辆的累计等待时间记作 D. 在 t 时刻的累计等待时间为 D_t. 奖励值定义如式(4)所示.

$$r_t=k(D_t-D_{t+1})=k\left(\sum_{i=1}^{n}d_{t,i}-\sum_{j=1}^{m}d_{t+1,j}\right) \tag{4}$$

其中，n，m 为 t，$t+1$ 时刻路口等待的车辆数目；k 为常数；r 越大，代表减少的等待时间越多，反之越少. 算法的目的为尽量增大奖励值 r.

3　FEPG 算法设计

DDPG 网络由策略网络 A，价值评价网络 C 组成. 其中策略网络 A 分为 A-online 和 A-target 两个网络，其参数分别表示为 θ，θ'. 评价网络分为 C-online 和 C-target 两个网络，其参数分别表示为 φ，φ'. 本论文针对交叉路口环境的特点提出两点改进：(1) 交叉路口的状态维度为 32，其中可能存在对训练结果无贡献维度. 利用基于信息增益的特征增强算法，自动优化状态；(2) 在 DDPG 算法中，经验池是样本的唯一来源，保持经验池样本的多样性有利于网络收敛. 网络训练时按照余弦相似度算法随机丢弃样本. 本论文中将改进的 DDPG 网络框架简称为 FEPG(Feature Enhance DDPG)网络，如图 2 所示. FEPG 算法首先从预训练阶段开始，预训练阶段的主要目的是进行原始样本采集，并应用特征增强算法和样本去重算法. 特征值增强和样本去重算法的原理如下.

(1) 特征增强算法：由于交通环境的状态量维度较高，其中很可能存在干扰收敛的无关维度. 特征增强算法的核心思想是通过样本的信息增益来筛选合适的特征. 将状态 s 表示为向量 $s=\{x_1,x_2,\ldots,x_{32}\}$. 特征筛选算法目的是降低不相关特征对训练的影响. 经验池在算法的预训练阶段将累积 n 组样本. 对于随机变量 X，可以求其信息熵：

$$H(X)=-\sum_{j=1}^{m}p(x_j)\log_2 p(x_j) \tag{5}$$

以及给定条件 Y 下的条件熵：

$$H(Y\mid X)=\sum_x p(x)H(Y\mid X=x) \tag{6}$$

信息增益则可以表示为

$$IG(Y\mid X)=H(Y)-H(Y\mid X) \tag{7}$$

信息增益越大说明随机变量 X 对于 Y 的贡献越大. 信息增益用于计算离散变量，因此将奖励值和状态离散化，状态的第 i 个维度的信息增益可以表示为

$$h_i=IG(R\mid X_i) \tag{8}$$

将特征因子 e_i 和特征加权状态 $\hat{s}$ 定义为

$$e_i=\begin{cases}1, h_i\geq\mu\\ 0, h_i<\mu\end{cases} \tag{9}$$

$$\hat{s}=\{\hat{x}_1,\hat{x}_2,\ldots,\hat{x}_{32}\}=\{e_1x_1,e_2x_2,\ldots,e_{32}x_{32}\} \tag{10}$$

其中，μ 取经验值 0.1. 通过 $\hat{s}$ 的定义可知信息增益小的特征维度对网络没有影响.

(2) 样本去重算法：样本去重的目的为保持样本的多样性，加快算法收敛. 新进加入的样本与样本池中的样本随机比对相似度. 相似度越高，样本被丢弃的可能性越高. 相似度利用样本间的余弦值表示，丢弃概率表达式如式(11)，其中 α 为丢弃系数.

$$p_{\text{discard}}=1-\alpha\cos(k)=1-\alpha\frac{s_i\cdot s_j}{||s_i||\cdot||s_j||} \tag{11}$$

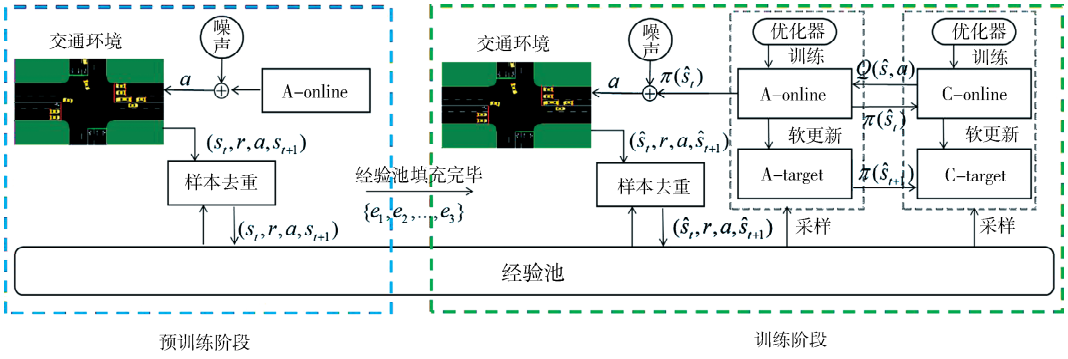

图 2　FEPG 算法结构

Fig. 2　FEPG algorithm structure

待数据采集完毕后接着进入训练阶段. 训练阶段中 4 个神经网络相互作用，迭代. 各个网络的作用和相互之间的关系如下.

A-online 网络负责输出动作，其输入为当前状态 $\hat{s}$，输出为动作 a. 其损失函数为

$$loss\_a=J(\theta)=-\frac{1}{m}\sum_{j=1}^{m}Q(\hat{s}_i,a_i\mid\theta) \tag{12}$$

A-target 网络用于动作采样，其输入为下一状态 $\hat{s}'$，输出为下一动作 a'. 网络参数一般采用软更新，其更新方程为

$$\theta'\leftarrow\tau\theta+(1-\tau)\theta' \tag{13}$$

C-online 网络负责拟合 Q 函数，其输入为当前状态 $\hat{s}$，动作 a，输出为 Q 值. y_i 为标签. 其损失函数为

$$loss\_c=J(\varphi)=\frac{1}{m}\sum_{j=1}^{m}(y_i-Q(\hat{s}_i,a_i\mid\varphi))^2 \tag{14}$$

C-target 网络为 A-online 提供 Q 函数，其输入为下一状态 $\hat{s}'$，下一动作 a'，输出值用于计算标签值 y_i. 网络采用软更新，其更新方式表示为

$$\varphi'\leftarrow\tau\varphi+(1-\tau)\varphi' \tag{15}$$

式(12)和式(14)中 m 为训练 batch 大小，τ 一般取 0.01. 以上对各个网络的描述可以知道 target 网络应当分别和 online 网络具有相同的网络结构. 动作网络在输出时会添加均值为 0，方差按指数规律衰减的高斯噪声. 目的是鼓励智能体在前期探索尽可能大的状态空间. 根据以上对网络训练方法的分析，FEPG 算法流程如算法 1 所示.

算法 1　FEPG 算法流程

1) 初始化在线网络. 并赋予随机权重 θ，φ

2) 初始化目标网络 θ'，φ'，初始化经验池 R

3) for $j=1$ to M do

4) 随机初始化环境，获得初始状态 s

5) for $t=1$ to T do

6) 动作网络添加衰减的高斯噪声输出动作 a

7) Agent 执行动作 a 并获得样本存入经验池

8) 根据式(11)进行样本去重过程

9) if 经验池填充完毕

10) 根据式(8)和式(9)计算 $\{e_i\}$

11）if 开始训练

12）在经验池中采集 m 个样本并计算.

$$y_i = r_i + \gamma Q(\hat{s}_j', \pi(\hat{s}_i' \mid \theta') \mid \phi')$$

13）使用最小化方差更新在线价值网络：

$$Lc = \frac{1}{m}\sum_{j=1}^{m}(y_i - Q(\hat{s}', a_i \mid \phi))^2.$$

14）使用策略梯度更新动作网络：

$$Lc = \frac{1}{m}\sum_{j=1}^{m}(y_i - Q(\hat{s}', a_i \mid \theta)$$

15）采用软更新方式更新目标网络：

$$\theta' \leftarrow \tau\theta + (1-\tau)\theta'$$

$$\phi' \leftarrow \tau\phi + (1-\tau)\phi'$$

16）endfor

17）endfor

4　实　验

我们在上文所述的交通环境中验证算法有效性，并且对比定时控制 FTC（Fixed Timing Control），Pang 等人的 DDPG[18] 控制方法，Genders 等人的 DQN[19] 控制方法与本文的 FEPG 控制方法. 其中 FTC 为传统方法，DDPG 和 DQN 为新型的强化学习方法. 对比的核心指标为算法的收敛性能，车辆的平均等待时间以及车辆排放数据.

4.1　实验设计与 FTC 方法

如表 4 所示，本文设计了 4 组实验，分别是低车流密度低方差（Low Density Low Variance，LDLV），低车流密度高方差（Low Density High Variance，LDHV），高车流密度低方差（High Density Low Variance，HDLV），高车流密度高方差（High Density High Variance，HDHV）. 车流密度指所有路口车辆数据，方差指 4 个路口车流量值与平均值之差的平方和，方差越高表示路口的车流密度差异越大. 车流密度数据可见文献[20].

表 4　车流类型

Tab. 4　Traffic flow type

实验序号	实验名称	密度/(辆/h)	方差
1	LDLV	640	100
2	LDHV	640	400
3	HDLV	1 080	100
4	HDHV	1 080	400

一般交通信号灯的相位采用 FTC 方法控制. 实验测试不同环境下所有车辆的累积等待时间与 1 和 4 动作相位间隔的关系，每个间隔时间测试 2 000 s，统计结果如图 3 所示. 在图 3 中用绿色标记了交通累积等待时间较短的区间. 针对 4 组实验，将 LDLV 的固定间隔设置为 30 s，LDHV 的固定间隔设置为 20 s，HDLV 的固定间隔设置为 35 s，HDHV 的固定间隔设计为 20 s.

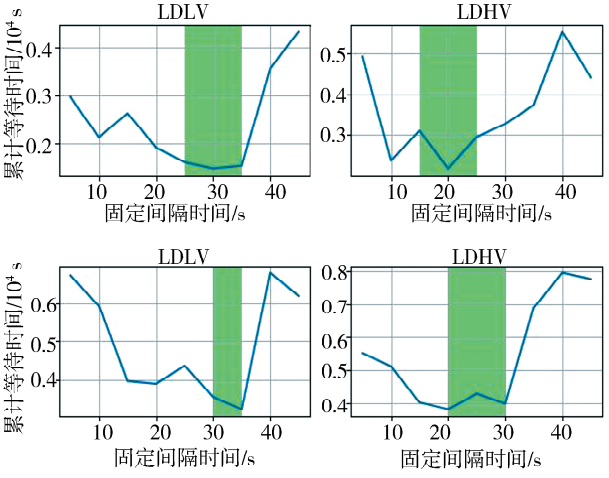

图 3　不同信号灯间隔的平均等待时间

Fig. 3　Average waiting time at different signal time intervals

4.2　实验对比和结果分析

根据上文对环境的描述，实验的状态维度为 32，动作维度为 1. 设计 FPPG 网络参数如表 5 所示. FEPG，DDPG，DQN 算法的经验池大小设为 1 万，Batchsize 为 100，折扣取 0.9. 训练 200 轮，每轮执行 200 步后训练结果如图 4 所示. 总体而言 FEPG 收敛最快性能最好，DDPG 性能其次. DDPG 控制算法中将每一个车道划分为多个单元，且需要保证每个单元内仅有一辆车，状态的维度为 480. 其中很可能包含大量无效维度，导致其收敛不稳定，特别是在 HDHV 环境中表现不佳. DQN 控制算法的输出动作值为离散变量. 实验中将输出动作按 5 s 一个间隔进行划分. 其输出量不连续无法精确控制，导致其效果最差.

表 5　网络结构参数

Tab. 5　Net structure parameters

网络名称	网络层数	激活函数	神经元数量
A-online	2	ReLU，tanh	80，30
A-target	2	ReLU，tanh	80，30
C-online	2	ReLU，linear	100，40
C-target	2	ReLU，linear	100，40

将训练完毕的网络在 4 个实验环境中进行测

试，测试时间为 8 000 s. 图 5 统计了 8 000 s 时间内不同算法每辆车的平均等待时间. LDLV 和 LDHV 统计了约 1 500 辆车，HDLV 和 HDHV 统计了约 2 500 辆车. 统计结果表明 FEPG 算法控制下的车辆等待时间平均比 BFI 降低了 23.5%，比 DQN 算法降低 9.8%，比 DDPG 算法降低了 7.6%. 结果表明 FEPG 算法学习到了如何高效的控制交通信号，证明了算法的有效性.

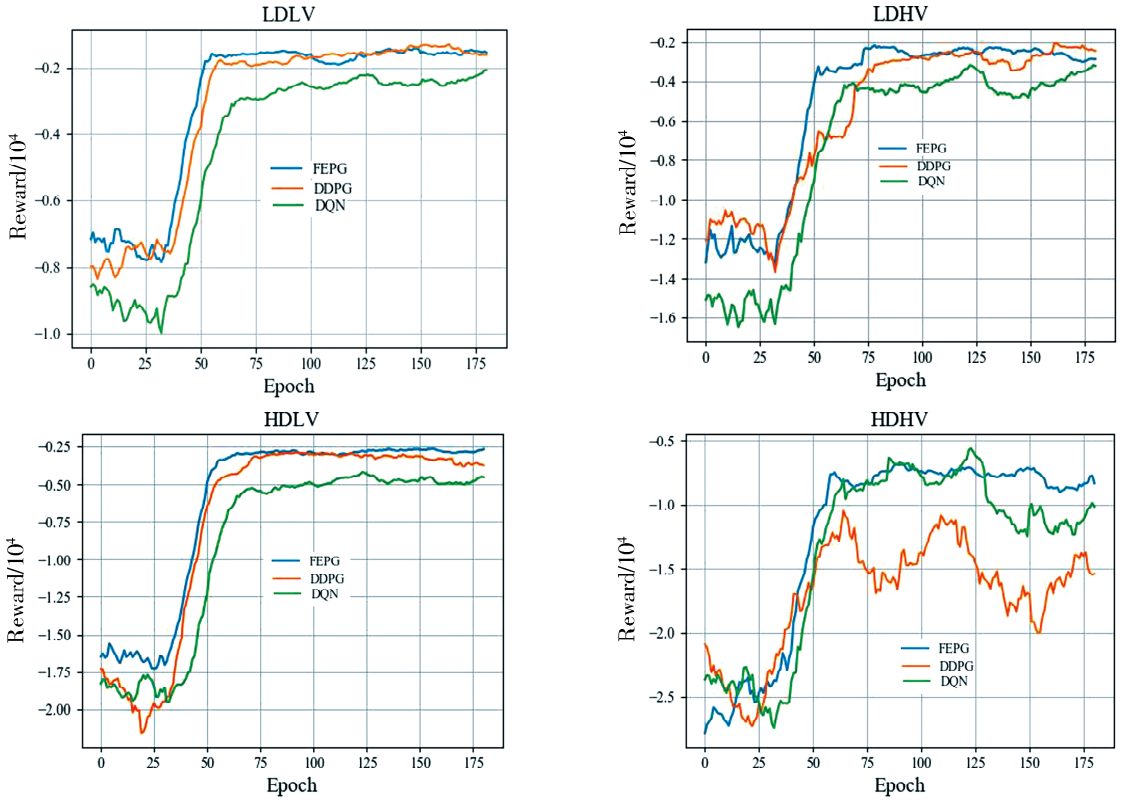

图 4　算法训练过程对比

Fig. 4　Comparison of training process

表 6　污染物排放对比

Tab. 6　Comparison of emissions

排放	FTC	DQN	DDPG	FEPG
CO/kg	8.168	7.132	6.913	6.797
CO_2/kg	261.7	231.8	210.2	214.9
HC/g	43.45	39.64	35.57	36.09
Noise/dB	56.53	53.27	52.64	51.59
NOx/g	114.5	100.8	102.6	94.15
PMx/g	5.804	4.975	4.789	4.778

以 HDLV 环境为例，表 6 是 FEPG 算法组对比其他算法在的污染排放情况. 通过 SUMO 提供的数据统计了 8 000 s 内 1 500 辆车的污染排放情况. 依次是一氧化碳、二氧化碳、碳氢化合物、汽车噪音、氮氧化合物、和颗粒物排放. 其中噪声为 8 000 s 内平均值，其余为累计值. 相比 FTC 算法，使用 FEPG 算法使污染排放比 FTC 平均降低了 19.7%，比 DQN 算法降低了 6.3%，比 DDPG 算法降低了 3.6%. 说明 FEPG 算法在提高通行效率的同时也降低了污染的排放.

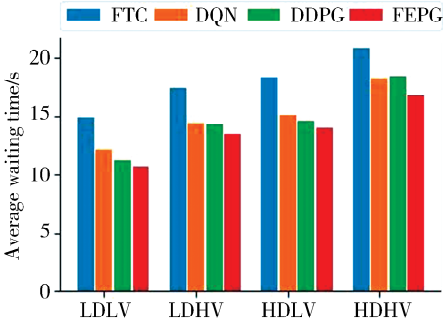

图 5　平均等待时间对比

Fig. 5　Comparisonof average waiting time

5　结　论

面对大城市日益严重的交通拥堵情况，本论文设计了一种强化学习算法控制十字路口交通信号灯. 基于 DDPG 算法结合特征增强算法和样本去重算法设计了 FEPG 算法. 改进算法相比文献

[18]的 DDPG 算法和文献[19]的 DQN 收敛的更快.以典型的十字路口交通信号灯为例,设计多组实验验证算法有效性. 实验结果表明 FEPG 网络在提高通行效率的同时也降低了汽车污染物排放. 高效状态特征选择可以提高强化学习成功率，交通环境中的数据多种多样,后续研究的一个方向是如何在更加复杂的交通网络中选取有效的环境特征.

参考文献：

[1] Mousavi S S, Schukat M, Howley E. Traffic light control using deep policy-gradient and value-function-based reinforcement learning [J]. IET Intell Transp Sy, 2017, 11: 417.

[2] Li L, Wen D. Parallel systems for traffic control: a rethinking [J]. IEEE T Intell Transp, 2015, 17: 1179.

[3] Liang X, Yan T, Lee J, *et al*. A distributed intersection management protocol for safety, efficiency, and driver's comfort[J]. IEEE Internet Things, 2018, 5: 1924.

[4] El-Tantawy S, Abdulhai B, Abdelgawad H. Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC): methodology and large-scale application on downtown Toronto [J]. IEEE T Intell Transp, 2013, 14: 1140.

[5] Mnih V, Kavukcuoglu K, Silver D, *et al*. Playing atari with deep reinforcement learning [J]. arXiv preprint arXiv, 2013: 1312.5602.

[6] 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: 484.

[7] Thorpe T L. Vehicle traffic light control using SARSA[C/OL]. (1997-02-03)[2020-01-05]. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=56B59D9E927D79D3A9E65FB67B7FEE32?doi=10.1.1.56.6788&rep=rep1&type=pdf.

[8] Li L, Lü Y, Wang F Y. Traffic signal timing via deep reinforcement learning[J]. IEEE/CAA J Autom Sinica, 2016(3): 247.

[9] Wang Z, Schaul T, Hessel M, *et al*. Dueling network architectures for deep reinforcement learning [J]. arXiv preprint arXiv, 2015: 1511.06581.

[10] Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning [J]. arXiv preprint arXiv, 2015: 1509.06461.

[11] Wang X N, He H G, Xu X. Reinforcement learning algorithm for partially observable Markov decision processes [J]. Control Decis, 2004, 19: 1263.

[12] Ohnishi S, Uchibe E, Yamaguchi Y, *et al*. Constrained deep Q-Learning gradually approaching ordinary Q-Learning [J]. Front Neurorobotics, 2019, 13: 103.

[13] 池涛，王洋，陈明. 多层局部感知卷积神经网络的高光谱图像分类[J]. 四川大学学报：自然科学版，2020，57：103.

[14] Cho K, Van Merriënboer B, Gulcehre C, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation [J]. arXiv preprint arXiv, 2014: 1406.1078.

[15] Lillicrap T P, Hunt J J, Pritzel A, *et al*. Continuous control with deep reinforcement learning [J]. arXiv preprint arXiv, 2015: 1509.02971.

[16] Krajzewicz D. Traffic simulation with SUMO-simulation of urban mobility [C]//Fundamentals of Traffic Simulation. NewYork, NY: Springer, 2010.

[17] Keller M. Handbook of emission factors for road transport (HBEFA) [EB/OL]. (2010-08-12)[2020-01-06]. https://www.hbefa.net/d/index.html.

[18] Pang H, Gao W. Deep deterministic policy gradient for traffic signal control of single intersection [C]//proceedings of the 2019 Chinese Control And Decision Conference (CCDC). [S.l.]: IEEE, 2019.

[19] Genders W, Razavi S. Using a deep reinforcement learning agent for traffic signal control [J]. arXiv preprint arXiv, 2016: 1611.01142.

[20] 北京交通发展研究院. 2019 北京市交通发展年度报告[R]. 北京：北京交通发展研究院，2020.

引用本文格式：

中　文：刘利军，王州，余臻. 一种改进的深度确定性策略梯度网络交通信号控制系统[J]. 四川大学学报：自然科学版，2021，58：043003.

英　文：Liu L J, Wang Z, Yu Z. Improved deep deterministic policy gradient network traffic signal control system [J]. J Sichuan Univ: Nat Sci Ed, 2021, 58: 043003.