DOI:10.13196/j.cims.2022.11.009

基于不确定需求的无人驾驶出租车优化调度

周晓婷[1],吴禄彬[1],章　宇[2],姜善成[1+]
(1.中山大学 智能工程学院,广东　深圳　518107;2.西南财经大学 工商管理学院,四川　成都　611130)

摘　要:为了减少乘客在高峰期打车难和出租车空载的情况,面对不确定的出行需求,提出一个无模型深度强化学习框架,以解决无人驾驶出租车调度问题。该框架使用马尔可夫决策模型进行建模,综合考虑了运营商收益与顾客等待成本,使用基于策略的深度强化学习算法——双延迟深度确定性策略梯度算法(TD3)对无人驾驶出租车进行调度,达到合理分配空闲车辆资源的目的。基于纽约市的真实出租车出行数据搭建了环境模拟器,通过在训练过程中加入不确定需求来增强算法鲁棒性。实验结果证明了该方法在求解不确定需求下的无人驾驶出租车调度问题时的有效性。

关键词:强化学习;无人驾驶出租车;车辆调度;策略梯度
中图分类号:TP301;U9　　**文献标识码**:A

Optimal repositioning of driverless taxi under uncertain demand

ZHOU Xiaoting[1], *WU Lubin*[1], *ZHANG Yu*[2], *JIANG Shancheng*[1+]
(1. School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518107, China;
2. School of Business Administration, Southwestern University of Finance and Economics, Chengdu 611130, China)

Abstract: To reduce the amount of empty taxies and make passengers more easily to take a taxi in peak hours, a model-free deep reinforcement learning framework was proposed to dispatch driverless taxi under uncertain demand. The framework comprehensively considered the benefit of service providers as well as the waiting cost of customers. A well-designed Twin Delayed Deep Deterministic policy gradient (TD3) algorithm was introduced to optimize the problem and allocate resources. The simulator was built based on real taxi trip data from New York. To improve the robustness of the algorithm, uncertain demands were added to the training process. The experimental results showed that the algorithm could make non-shortsighted and effective strategy under uncertain demand.

Keywords: reinforcement learning; driverless taxi; vehicle repositioning; policy gradient

0　引言

传统出租车在高峰时期总会出现乘客“打车难”与车辆空载这两种难以平衡的问题[1]。而且由于运营平台、司机、乘客的博弈,全局最优的调度策略往往不能被贯彻执行。随着物联网、通信技术、人工智能技术等发展,自动驾驶技术在不断成熟[2]。目前我国不少一线城市已经开展各类无人驾驶汽车的前期测试与探索活动,相信在不久的未来,共享出租车公司,如哈啰、百度等很可能搭建自动驾驶出租车队用于搭载乘客,以缓解当下出租车平台在高峰期所面临的各类问题。面对城市交通中乘客出行需求的不确定性,如何有效利用无人驾驶出租车可集中调度的特点来调度空闲的无人驾驶出租车,从而满足

收稿日期:2022-01-04;修订日期:2022-03-29。Received 04 Jan. 2022;accepted 29 Mar. 2022.
基金项目:国家重点研发计划资助项目(2020YFB1713800);国家自然科学基金资助项目(71901180,71801031);广东省基础与应用基础研究基金资助项目(2019A1515011962)。**Foundation items**: Project supported by the National Key Research and Development Program, China (No. 2020YFB1713800), the National Natural Science Foundation, China (No. 71901180, 71801031), and the Guangdong Provincial Basic and Applied Basic Research Foundation, China (No. 2019A1515011962).

未来的出行需求，对提高无人驾驶出租车服务水平具有重要意义。

车辆调度问题是车辆路径规划问题的一个子问题[3]，针对不同应用场景，国内外学者一直尝试运用现代运筹优化理论获取对应场景下的全局最优解[4-5]。目前从服务提供者角度来说，大多数运营商采用定价激励的策略进行车辆调度[6]。例如采用顾客加价、司机调度奖励、峰时定价等策略来引导司机去需求量高的地方[7]。但也有学者对此类实时动态定价的有效性提出质疑，KOOTI 等[8]根据优步收集的真实数据分析出，峰时定价策略并没有给车辆调度带来较大的积极影响。

研究者研究了大量基于模型的车辆调度算法。ZHANG 等[9]根据排队理论搭建了按需系统(Mobility on Demand，MOD)以调度出租车，他们通过求解线性规划模型找出一种最优的调度策略，并应用到纽约的出租车案例中。实验证明，该算法在满足需求的情况下有效减少了出租车队规模。KIM 等[10]为了最小化出租车调度成本，将多目标的出租车调度问题转化为一个网络流问题，通过最小费用最大流算法求解。用韩国首尔地区的真实出租车数据进行模拟研究，证明了算法的有效性。BOYACI 等[11]提出一种允许决策者权衡运营商和用户利益的多目标混合整数规划模型来解决共享汽车调度问题。MA 等[12]则研究了一种无人驾驶出租车系统，该系统通过提前获取乘客需求来搭建系统的时空网络，通过线性规划让系统在最低成本和最小计算量上作出最优的调度决策，通过案例表明，该系统可以有效降低汽车拥有率。上述方法都是基于严格的数学模型，当涉及变量过多或者维度过高时，这些数学模型不能很好地适应，且面对大规模问题，求解效率不佳。启发式优化算法能够全面有效地搜寻最优解，而且面对大规模问题能够保证效率，因此受到很多研究者的青睐。谢榕等[13]采用人工鱼群算法对出租车进行基于全局角度的智能调度，从而实现对出租车的合理调度。何胜学等[14]将蚁群算法与遗传算法结合，来求解出租车调度策略，并通过实验证明了算法的有效性。上述方法都是在乘客的需求是静态的假设下建模的，然而在现实场景中，若是仅根据当前的乘客需求进行调度，则不能很好地应对未来可能出现的供需不平衡的情况。

本文提出基于不确定需求的无模型强化学习方法来解决无人驾驶出租车调度问题。通过在强化学习训练中引入不确定需求，从而使训练出来的模型能更好地适应城市交通中乘客的不确定需求。在强化学习的无模型算法中，其学习代理并不依赖于模型的任何先验信息，无需用参数估计模型，而是直接与训练环境交互来更新控制策略。在实际使用中，直接调用训练好的模型就可以得到调度策略。因此，强化学习算法即使面对大规模问题也能高效地做出性能稳定的调度方案[15]。近年来，采用强化学习算法解决调度问题的研究有很多[16]，如陈勇等[17]、张景玲等[18]、黎声益等[19]、MAO 等[20]。其中 MAO 等[20]与本文研究最为接近，该文献将车辆调度算法与强化学习结合，运用深度强化学习方法 actor-critic 算法[21]来优化车辆调度，并通过实验证明该算法收敛于理论上界。然而，actor-critic 算法已被证实会过高估计动作值，即对动作价值函数的估计会有误差，这种误差累积的偏差会导致任意的坏状态被估计为高值，从而导致次优的策略更新，以致于策略网络无法收敛。

由于该问题的状态空间是连续的，本文采用一种基于状态空间连续的算法——双延迟深度确定性策略梯度算法(Twin Delayed Deep Deterministic policy gradient algorithm，TD3)[22]。该算法可以有效解决高估动作值的问题，从而得到最优的调度策略。为了更有效应对城市交通中乘客的不确定出行需求，本文将不确定需求与强化学习结合，在不确定需求环境下训练模型。通过神经网络捕捉到需求的随机性，模型能更好地应对需求变化的情况。最后，使用纽约市真实的出租车数据来模拟乘客需求，并将数据集划分为训练集和测试集来验证算法的合理性。实验证明，在需求不确定的情况下训练的模型在验证集和需求突变的情况下均表现较好，更具鲁棒性。

1 无人驾驶出租车调度问题的强化学习建模

根据马尔可夫决策对无人驾驶出租车调度问题建立模型。

首先定义无人驾驶出租车调度问题的服务区域和时间。假设无人驾驶出租车在特定一段时间内服务特定的区域。首先，假定服务 N 个离散区域，其中集合 $[N]=\{1,2,\ldots,N\}$ 代表区域索引从 1 到 N。然后假设服务时间是由离散的时间间隔 Δt 表示。因此，时间集合可以表示为 $[T]=[1,2,\cdots,$

$T]$。为了简化无人驾驶出租车调度问题,假设 Δt 足够小,所有无人驾驶出租车的调度动作都发生在时间间隔的初始处。比如对无人驾驶出租车搭载乘客来说,若在 t 时刻初始处,没有无人驾驶出租车搭载该乘客,则无人驾驶出租车最快能在 $t+1$ 时刻初始时搭载该乘客,而不会在 t 时刻和 $t+1$ 时刻之间搭载该乘客。进一步假设,乘客等待时间超过一定阈值后会取消订单并对运营商有取消订单的惩罚。所有无人驾驶出租车搭载乘客到指定地点后,无人驾驶出租车变为闲置车辆可以重新调度使用。

然后定义调度无人驾驶出租车的动作、顾客的需求、区域内无人驾驶出租车数量。x_{ijt} 表示在 t 时刻,从区域 $i \in [N]$ 被调度到区域 $j \in [N]$ 的无人驾驶出租车数量,其中区域 i 可以等于区域 j,代表无人驾驶出租车停留在原地;p_{ijt} 表示在 t 时刻,想从区域 i 到区域 j 的顾客需求数量;v_{it} 表示在 t 时刻,区域 i 的闲置无人驾驶出租车数量。一旦调度的无人驾驶出租车 x_{ijt} 确认后,就能进一步确定在 t 时刻无人驾驶出租车服务的从区域 i 到区域 j 的顾客数量,定义为 $y_{ijt}=\min(x_{ijt},p_{ijt})$。若 $x_{ijt}<p_{ijt}$,代表需求大于供应,会有部分顾客没有被服务,需要至少等待一个时刻才能被服务;若 $x_{ijt}\geq p_{ijt}$,代表需求小于或者等于供应,在 t 时刻,所有想从区域 i 到区域 j 的顾客需求都能被满足或者部分没有搭载乘客的空车从区域 i 被调度到区域 j 以满足未来需求。

最后定义无人驾驶出租车调度的成本。首先从顾客的角度来说,没有被服务的顾客需要等待一定时间才能搭载车辆。对于顾客来说会有一个等待时间成本。w_{ijt} 代表让一位想从区域 i 到区域 j 的顾客在 t 时刻等待了一个时间间隔 Δt 的成本。因此,在 t 时刻到 $t+1$ 时刻有乘车需求但没有车辆搭载的顾客等待时间成本定义为 $\sum_{i,j\in N}(p_{ijt}-y_{ijt})w_{ijt}$。从运营商角度,有调度空车成本与乘客取消成本。假设在 t 时刻从区域 i 调度空车到区域 j 的成本定义为 c_{ijt},在 t 时刻从 i 区域到 j 区域的顾客取消订单的数量定义为 n_{ijt},顾客取消订单的单位成本定义为 d_{ijt},所以在 t 时刻运营商成本为 $\sum_{i,j\in N}(x_{ijt}-y_{ijt})c_{ijt}+n_{ijt}d_{ijt}$。

根据以上假设和强化学习的要求,进一步搭建状态空间、动作空间和奖励函数。

(1)状态空间

状态空间的定义如式(1)所示,每个状态可以被定义为 s_t 由当前时刻 t;等待着的顾客需求 $P_t=\{p_{ijt}:\mathrm{i}\in[N],j\in[N]\}$;可用车辆 $V_t=\{v_{it}:\mathrm{i}\in[N]\}$ 三部分组成。

$$s_t=\{t,P_t,V_t:t\in[T],P_t\in\mathbb{R}_+^{n\times n},V_t\in\mathbb{R}_+^{n}\}。\tag{1}$$

(2)动作空间

动作空间的定义如式(2)所示。动作 a_t 是 x_{ijt} 组成,其中 $i,j\in[N]$。x_{ijt} 代表在 t 时刻从区域 i 调度到区域 j 的无人驾驶出租车,其中调度动作需要满足约束——从 i 区域调度到任意区域的无人驾驶出租车数量应该等于当前时刻 i 区域的空闲无人驾驶出租车数量。区域 i 可以等于区域 j,代表无人驾驶出租车没有被调度。

$$A(s_t)=a_t=\left\{x_{ijt}:\sum_{\forall \mathrm{j}\in \mathrm{N}}x_{ijt}=\nu_{it},i\in[N],\ j\in[N],x_{ijt}\in\mathbb{R}_+\right\}。\tag{2}$$

(3)奖励函数

奖励函数设置为成本的负数,如式(3),由顾客的等待成本、车辆的调度成本和顾客的取消成本组成。因此,调度系统的目标是作出令总成本最小的车辆调度策略,即作出令总收益最大的策略。$\pi_\theta(a_t|s_t)$ 代表策略网络,θ 代表策略网络的参数,策略网络输入状态 s_t,输出动作 a_t。可以根据式(4)更新策略网络。

$$r(s_t,a_t)=-\sum_{i,j\in N}[(x_{ijt}-y_{ijt})c_{ijt}+(p_{ijt}-y_{ijt})w_{ijt}+n_{ijt}d_{ijt}],\tag{3}$$

$$\theta^*=\mathrm{argmax}_\theta\sum_{t=1}^{T}\{E_{(s_t,a_t)\sim\pi_\vartheta(a_t|s_t)}[r(s_t,a_t)]\}。\tag{4}$$

总的来说,为避免维度诅咒,本文设置状态向量和动作向量都为连续变量。由于状态空间和动作空间都是连续的,采用更适用于连续动作空间的方法——双延迟深度确定性策略梯度算法 CTD3 算法。

2　无人驾驶出租车调度问题算法介绍

2.1　用于无人驾驶出租车调度的双延迟深度确定性策略梯度算法

TD3 算法是由深度确定性策略梯度算法(Deep

Deterministic Policy Gradient，DDPG)[23]进一步优化而来。DDPG 算法在处理连续动作空间的问题上有很好的表现效果，但是它通常对于超参数十分敏感，且会在训练时出现高估状态动作价值的问题。而 TD3 算法引入了两个目标动作价值网络来缓解高估的问题。

$\pi_\theta(a_t|s_t)$ 表示策略网络，输入状态就可以输出动作策略，其中 θ 表示策略网络的参数。$Q_\varphi(s_t,a_t)$ 表示动作动作价值网络，通过输入状态和动作，就可以输出评判该状态动作好坏的一个评价值，其中 φ 表示动作价值网络的参数。$Q_\varphi(s_t,a_t)$ 的含义是在当前状态采取动作 a_t，并一直使用动作 a_t 的策略到整个回合结束时奖励值之和的评估。TD3 算法通过动作价值网络来更新策略网络，动作价值网络越准确，策略网络采取的动作就越好。本文通过参考相关文献与多次实验设置了网络训练的超参数，超参数配置如表 1 所示。算法流程如图 1 所示。

表 1 算法参数设置

参数表示	说明
c	训练次数为 3 000 000
i	计数当前的训练次数
γ	折扣因子为 0.8
d	策略网络更新频率为 2
M	批处理大小为 218
τ	目标网络更新率为 0.005

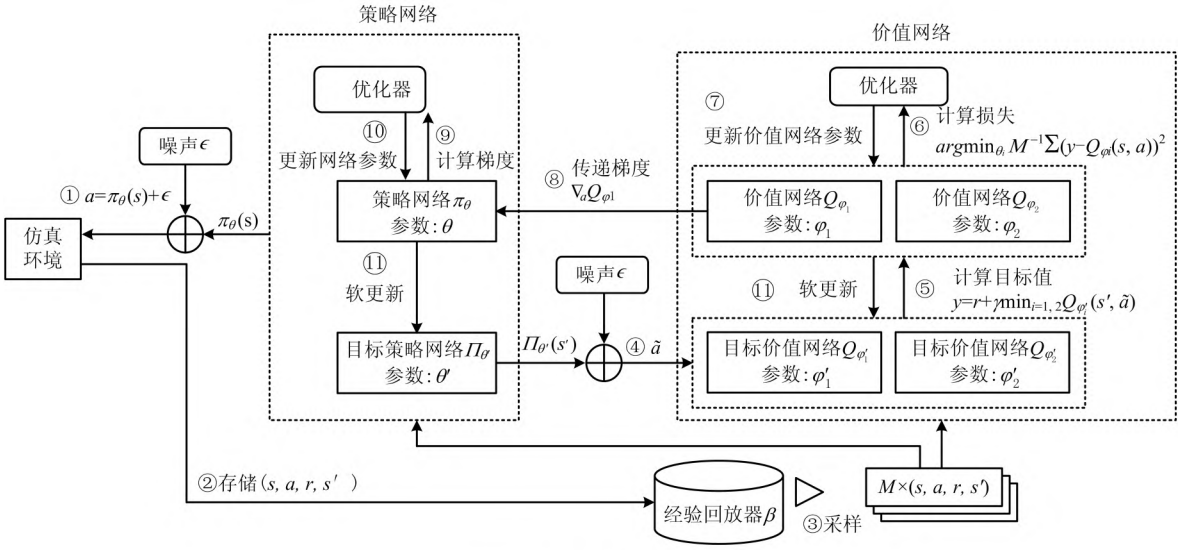

图1 TD3算法流程图

TD3 算法更新过程见算法 1。第 1～第 3 步初始化参数。第 5～第 12 步对主体进行训练，从而找到最优策略。第 5 步根据环境 s，策略网络输出动作，该动作添加噪声后得到 $a\leftarrow\pi_\theta(s)+\epsilon,\epsilon\sim N(0,0.2)$（添加噪声意味增加了探索），输入到环境里，得到该动作的奖励 r 与下一状态 s'。从而获取动作和状态序列(s,a,r,s')。第 6 步随机从经验回放器中采样批量的数据，通过 $\tilde{a}\leftarrow\Pi_{\theta'}(s')+\epsilon,\epsilon\sim clip(N(0,0.2),-0.5,0.5)$ 将下一状态输入到目标策略网络中并添加带有截断的噪声构造状态 s' 对应的动作 $\tilde{a}$，计算出两个目标动作价值网络值。其中两个目标动作价值网络的最小值与奖励一起构造出目标值 y。第 7 步，$Q_\varphi(s_t,a_t)$ 和目标值 y 通过均方误差构建损失函数。第 8～第 10 步，动作价值网络更新后，策略网络延迟更新（这里是价值网络更新两次，策略网络更新一次）。策略网络通过梯度上升的方式更新。目标网络参数则通过软更新(soft-update)的方式进行更新。软更新是一种常见的目标网络更新方式，该更新方式可以保证目标网络每次迭代都会更新。软更新利用当前网络参数与目标网络参数来更新网络，这样会使算法训练更稳定。

算法 1 TD3 算法。

1. 用随机网络参数 φ_1、φ_2、θ 来初始化动作价值网络 Q_{φ_1}、Q_{φ_2} 和策略网络 π_θ
2. 初始化目标网络参数 $\varphi'_1\leftarrow\varphi_1$，$\varphi'_2\leftarrow\varphi_2$，$\theta'\leftarrow\theta$
3. 初始化经验回放器 β
4. for $i=1$ to c：
5. 策略网络输出的动作，加上探索噪声后输入到环境里

面,ϵ噪声服从正太分布 $a \leftarrow \pi_\theta(s)+\varepsilon, \varepsilon \sim N(0,0.2)$。环境根据输入的动作输出奖励 r 和下一阶段状态 s'。将元组序列(s, a, r, s')存储在经验回放器 β 中

6. 从经验回放器 β 中采样小批量 M 个元组序列(s, a, r, s'),并生成对应下一阶段有噪声的 $\tilde{a}$ 和目标动作价值网络值 y。

$$\tilde{a} \leftarrow \pi_{\theta'}(s')+\varepsilon, \varepsilon \sim clip(N(0,0.2),-0.5,0.5)$$

$$y \leftarrow r+\gamma min_{i=1,2} Q_{\varphi_i'}(s', \tilde{a})$$

7. 更新两个动作价值网络的参数 $\varphi_i \leftarrow \arg\min_{\theta_i} M^{-1}\sum(y-Q_{\varphi_i}(s,\tilde{a}))^2$

8. if $i \bmod d$:

9. 通过确定性策略梯度更新策略网络参数 θ :

$$\nabla_\theta J(\theta) = M^{-1}\sum_a^{\nabla} Q_{\varphi_1}(s,a)\big|_{a=\Pi_\theta(s)} \nabla_\theta \pi_\theta(s)$$

10. 软更新目标网络:

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

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

11. end if

12. end for

本研究采用前馈密集神经网络来构造策略网络和动作价值网络。因为在该问题的建模中,动作(调度的车辆)应该是非负的整数,且从区域 i 调度到任何区域的车辆总和应该等于区域 i 的空车数量。因此本文在策略网络中使用一个变换函数来完成约束,如图 2 与式(5)所示。通过式(5),将策略网络的输出 a_{ij} 转化为 x_{ij},从而满足约束。其中 a_{ij} 代表从 i 区域到 j 区域的策略网络输出值,v_i 代表 i 区域的闲置无人驾驶出租车。首先由于策略网络的激活函数是 tanh 激活函数,输出数据范围是$[-1,1]$,因此对输出的动作 a_{ij} 首先归一化到$[0,1]$之间,然后计算归一化后的 a_{ij} 与归一化后的所有从 i 区域调度到任何区域的策略网络输出值的比值,再乘上该区域的闲置车辆。最终得到满足约束条件的调度动作 x_{ij}。

$$F(a_{ij}) = x_{ij} = v_i \frac{0.5\times(a_{ij}+1)}{\sum_{\forall k\in N} 0.5\times(a_{ik}+1)}。\quad (5)$$

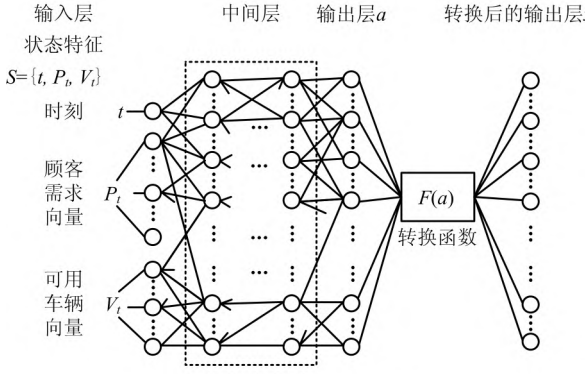

图2　策略网络的输出处理

2.2　用于验证 TD3 算法的混合整数规划模型描述

假设乘客需求和系统动力学的信息都是已知且确定的,以此为前提搭建混合整数规划模型求得无人驾驶出租车调度问题的奖励值理论上界。本文将整个调度问题视为求解静态的混合整数规划问题,该混合整数规划模型目标设置为成本最低来求解最优的调度策略。在后续实验中,将混合整数规划求得的理论上界与强化学习的结果进行比较,进而分析 TD3 网络训练过程的收敛效果。整数规划的定义如表 2 所示。

表 2　整数规划参数及其含义

参数	含义
ω_{ijt}	在 t 时刻,让一位想从区域 i 到区域 j 的顾客等待一个时间间隔的单位等待成本
c_{ijt}	在 t 时刻,调度一辆不载客的车辆从区域 i 到区域 j 的单位调度成本
d_{ijt}	在 t 时刻,一位本来打算从 i 区域到 j 区域的顾客取消订单的成本
λ_{ijt}	在 t 时刻,新增的从区域 i 到区域 j 的乘客需求数量
e_i	在 i 区域的初始化的空车数量
τ_{ijt}	在 t 时刻,车辆从区域 i 到区域 j 的行驶时间
$A_{it}=\{(j,t'):t'+\tau_{jit'}=t\}$	关于出发区域 j 与出发时间 t' 的变量对集合,满足该集合的出发车辆需要在 t 时刻到达 i 区域成为空闲车辆。
p_{ijt}	在 t 时刻,等待着的想从区域 i 到区域 j 的顾客需求
n_{ijt}	在 t 时刻,等待超过一定时间阈值后,从 i 区域到 j 区域取消订单的顾客需求数量
v_{it}	在 t 时刻,在区域 i 的闲置车辆
x_{ijt}	在 t 时刻,从区域 i 调度到区域 j 的无人驾驶出租车数量
y_{ijt}	在 t 时刻,搭载的从区域 i 到区域 j 的顾客数量

混合整数规划模型：

$$\min_{x,y,p}\sum_{i,j\in N}\sum_{t\in T}(p_{ijt}-y_{ijt})\omega_{ijt}+(x_{ijt}-y_{ijt})c_{ijt}+n_{ijt}\times \mathrm{d}_{ijt}。$$

s. t.

$$y_{ijt}=\min(x_{ijt},p_{ijt}); \tag{1}$$

$$p_{ijt+1}=p_{ijt}-y_{ijt}-n_{ijt}+\lambda_{ijt}; \tag{2}$$

$$v_{it}=\sum_{j\in[N]}x_{ijt}; \tag{3}$$

$$v_{i0}=e_i; \tag{4}$$

$$v_{it}=\sum_{j,t'\in A_{it}}x_{ijt'}; \tag{5}$$

$$p_{ij0}=\lambda_{ij0}; \tag{6}$$

$$x_{ijt}\in\mathbb{N}^+,\forall i,j\in N,t\in T。 \tag{7}$$

目标函数由乘客的等待成本、调度成本和顾客的取消成本组成，目标是最小化成本。约束条件式(1)规定，在任何时刻搭载的乘客数量要不等于等待的顾客数量(此时有足够多的车，供应大于等于需求)，要不等于调度的车辆(此时没有足够多的车，部分乘客需求无法满足，供应小于需求)。约束式(2)规定，下一时刻正在等待的乘客由上一时刻剩余的等待乘客和下一时刻新的乘客需求组成。约束式(3)规定，调度的车辆总和应该等于该区域的空闲车辆，当 $i=j$ 时，相当于车辆没有被调度。约束式(4)和式(6)代表每个区域的初始车辆和初始乘客都是已知的。约束式(5)表示，该区域的空闲车辆由在该时刻到达该区域的车辆组成，即在 t 时刻 i 区域的闲置车辆数量等于在 t' 时刻从 j 区域出发，经过 $\tau_{jit'}$ 时间行驶，在 t 时刻到达 i 区域的调度车辆动作的和。约束式(7)确保了调度车辆(决策变量)是非负的整数。

3 量化实验

3.1 实验设置

在模型训练之前，本文搭建了一个环境模拟器来模拟无人驾驶出租车的运营及调度过程。其中用户出行需求信息提取于真实的纽约市曼哈顿区域黄色出租车订单数据。假设所有出租车都是自动驾驶车辆，可以集中调度。因此，本文目标是利用强化学习 TD3 算法和该模拟器，来找出最优的无人驾驶出租车调度策略。

首先从 NYC TLC(taxi & limousine commission)获得了关于纽约市曼哈顿的地理坐标。该地图将纽约市曼哈顿区分为 64 个区域。然后从 NYC TLC 中获得了 2016 年 7 月黄色出租车在曼哈顿市的订单数据集。该数据集记录了乘客上车和下车的地点和时间、行驶距离、费用、费率类型、支付类型和司机报告的乘客数量等信息。

为减少模型验证的计算量同时不失其真实性，作了 3 种简化：①将无人驾驶出租车行驶区域划分为 8 个服务区，即把区域聚集成更大的区域，从而形成一个小的网络，如图 3 所示；②由于高峰时间段，供应与需求有着较大的差距。选取早高峰的 6 点～10 点的数据，时间间隔设定为 15 分钟；③假设每天每个区域的初始车辆分布是一样的。这 3 个简化有助于减少计算时间和计算量来验证所提方法。若有足够的计算能力，本文方法也可以推广到任何规模的网络和时间间隔。为了不失合理性，在仿真器中，结合当地的环境及相关政策，本文手动设置了其他参数，如旅行时间、等待成本、调度成本等，模拟无人驾驶出租车运营场景。

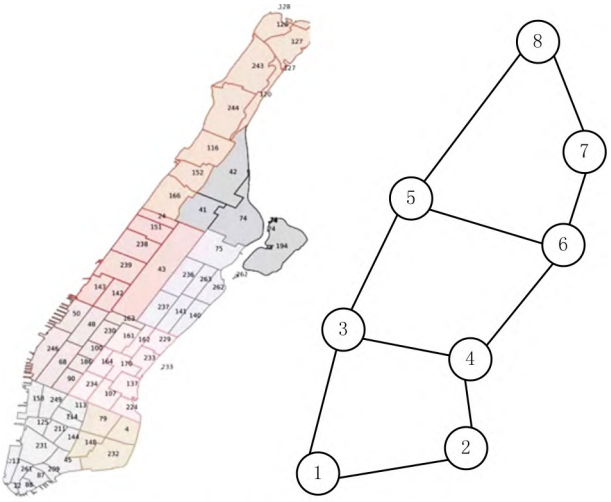

图3 左曼哈顿地图，右简化的曼哈顿区域

3.2 乘客需求确定仿真环境下的 TD3 架构部署与表现

本文的策略网络是由三层线性网络(大小为 256)和三层激活层(前两层为 relu 激活函数，最后一层为 tanh 激活函数)组成。动作价值网络由三层线性网络(大小为 256)和两层激活层(都为 relu 激活函数)组成。其次，为了与混合整数规划算法作对比，设定每天模拟器的乘客需求都是确定的，即每天每个时刻每个区域到另一个区域的需求都是确定的。因此，这种情况下，混合整数规划的目标函数值即为奖励函数值的理论上界。强化学习的训练过程是令奖励越大越好，此处设置的奖励值为成本的负数，即训练过程中成本会越来越小。在实验中，将

TD3 算法与强化学习的另一种算法深度确定性策略梯度算法(DDPG)进行比较。

实验总共训练了 300 万次,每 1 000 次进行验证,结果如图 4 所示。TD3 算法实验最终收敛在 -7.051×10^4,DDPG 算法最终收敛在 -7.403×10^4。利用 Gurobi 优化器求得混合整数规划的最优解为 -6.805×10^4。通过对比得知,TD3 算法与 DDPG 算法都收敛于整数规划理论最优值,但 TD3 算法比 DDPG 算法波动性更小、收敛更快且更接近于混合整数规划求得的理论上界。这是因为 TD3 算法在 DDPG 算法基础上有 3 个改进,首先采用两个动作价值网络更新学习的方式,可以有效抑制动作价值网络高估的问题;第二采用策略网络延迟更新的方法,使策略网络训练更加稳定;第三采用了目标网络平滑化的方法,通过计算目标动作价值网络值时动作添加噪声,从而让目标动作价值网络更新更准确和鲁棒。

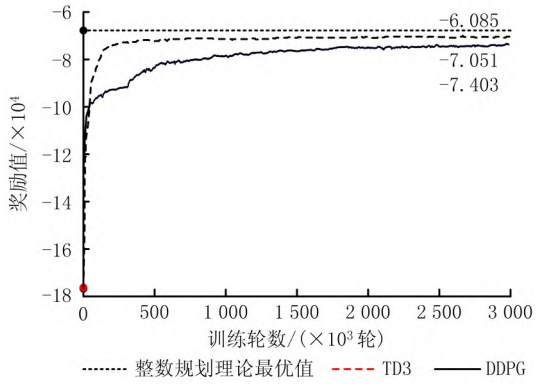

图4　需求确定情况下的训练情况

3.3　乘客需求不确定仿真环境下的 TD3 架构部署与表现

为了进一步测试 TD3 算法的实验表现,进一步允许乘客需求的随机性。用一个月的每个时刻每个区域的平均值作为乘客需求确定的情况,设为 D0,即 3.2 节中乘客需求确定下的仿真环境设置。接下来进一步为需求添加不确定性,将需求变为高斯分布,均值为一个月每个区域的需求均值,标准差设为 25%均值和 50%均值两种情况,表示为 D25 和 D50 的情况。通过这样的设置,得到 3 种需求环境:D0、D25、D50。

3 种情况下的训练验证曲线图如图 4～图 6 所示。通过实验可以看出,TD3 算法在 D25、D50 两种不确定需求的情况下均可达到收敛。尽管需求随机性为 D50 的时候,奖励值波动较大,但仍然在 150 万轮之后趋于平稳。对比在 D0、D25、D50 三种环境的训练曲线,可以发现顾客需求不确定性越大,奖励值波动越大。这是符合规律的,因为顾客需求是式(3)奖励值其中的一个因变量。当顾客需求不确定性越大,奖励值波动也越大。但更关键的是,可以看到,在 3 种情况下训练的算法都可以达到收敛。因此可以得出结论:TD3 算法可以有效应对需求不确定环境下的无人驾驶出租车调度。图 7 给出了不同需求环境下训练出来的最优模型(即通过上述不同仿真环境训练得到的 D0-TD3、D25-TD3、D50-TD3 模型)分别在不同需求环境下的测试奖励值。对于 D25 与 D50 不确定需求环境的测试,随机采样符合 D25 与 D50 环境要求的 100 000 个需求样本,模型多次根据不同需求样本作出调度动作,最终统计出关于模型奖励值的箱型图,如图 7 所示。可以看到,在特定环境中训练出来的模型在该环境中测试结果最好。比如,D0-TD3 模型在 D0 环境中的测试结果比 D25-TD3 模型和 D50-TD3 模型更好。这是因为模型的训练环境和测试环境是一致的。但更值得留意的是,通过对比不同测试环境下模型的表现,发现在不确定需求环境中训练出来的模型(D25-TD3、D50-TD3)表现比确定环境中(D0-TD3)训练出来的模型鲁棒性更好。因此,实验可以证明,在训练中加入一定不确定性的需求,能使训练出来的模型面对不确定需求时表现得更鲁棒。为了进一步验证算法的鲁棒性,如图 8 所示,将不同环境下训练的模型在 2016 年 8 月真实数据上进行测试,可以看到不确定需求训练出来的模型在真实数据上测试结果更好,而且相对来说,D25-TD3 模型在 2016 年 8 月真实数据上的表现会比 D0-TD3 与 D50-TD3 模型表现更好。因此,在不确定需求 D25 环境训练出来的模型 D25-TD3 在真实场景中表现更好。

3.4　出行需求突变情况下的模型表现

在现实情况中,经常会遇到大型演出结束、景区闭园等乘客需求突然变化的情况,在这种情况下特别考验模型的能力。因此,进一步考虑实际需求与预期需求出现较大偏差的情况下模型的表现。如图 9～图 11 所示,将第 6 个时刻的乘客需求增加或者减少,虚线系列的线代表顾客需求曲线,实线系列的线代表的是调度的动作。图 9 代表在确定需求 D0 环境中训练出来的模型表现,可以看到,D0-TD3 模型应对需求特征的突然变化的情况下,模型的调

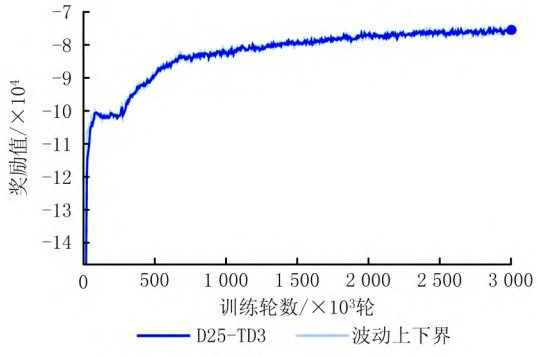

图5 需求随机D25的情况下的训练情况

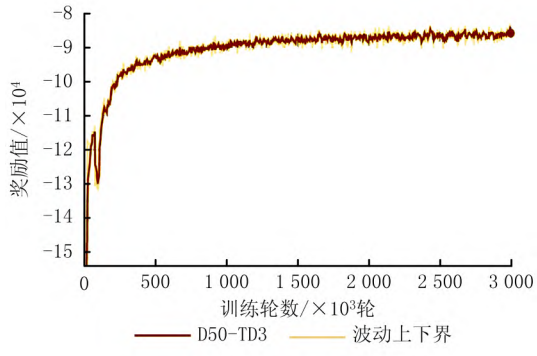

图6 需求随机D50的情况下的训练情况

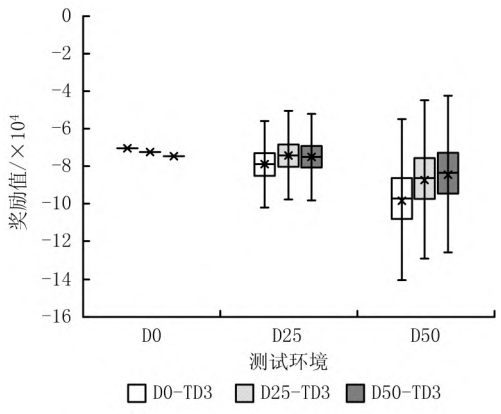

图7 交叉环境下三种模型的测试表现

度动作基本上没有改变，因此 D0-TD3 **模型面对需求突变的情况无法很好适应。而如图** 10 **和图** 11 **所示，**D25-TD3 **与** D50-TD3 **模型对需求突然变化的情况都作出了相应的调度变化。该实验进一步反映了在随机需求训练出来的模型，应对需求突然变化的情况表现得更鲁棒。这是因为深度强化学习算法的底层神经网络捕获了复杂的状态决策交互过程和相关的随机性，从而可以应对当前需求变化做出相应的调度控制。**

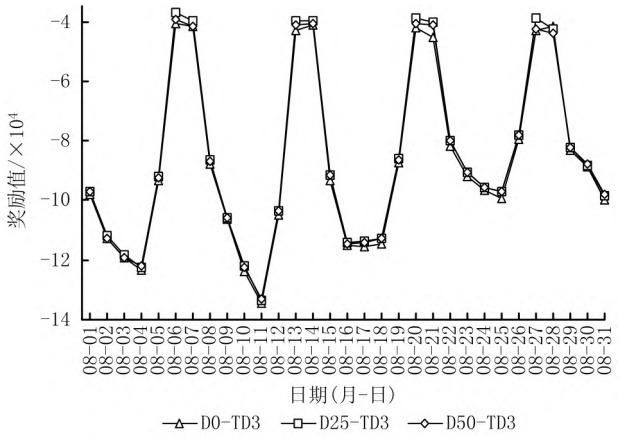

图8 不同需求下训练出来的模型在测试环境的结果

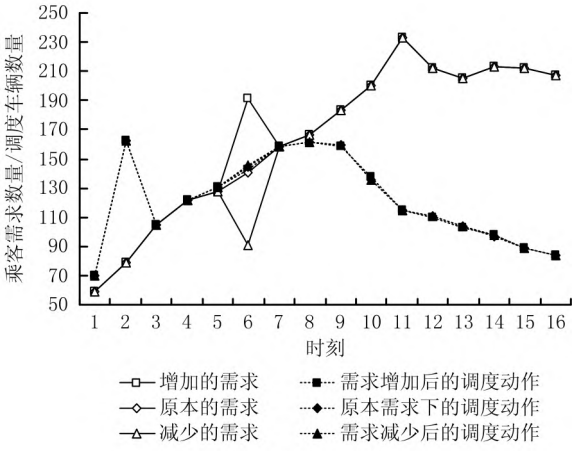

图9 D0-TD3模型在需求突变情况下的表现

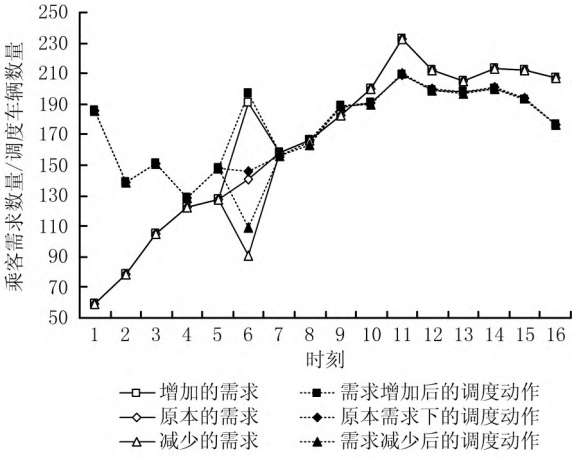

图10 D25-TD3模型在需求突增情况下的表现

4 结束语

本文提出一种用深度强化学习方法解决自动驾驶出租车调度问题。该方法基于双延迟深度确定性

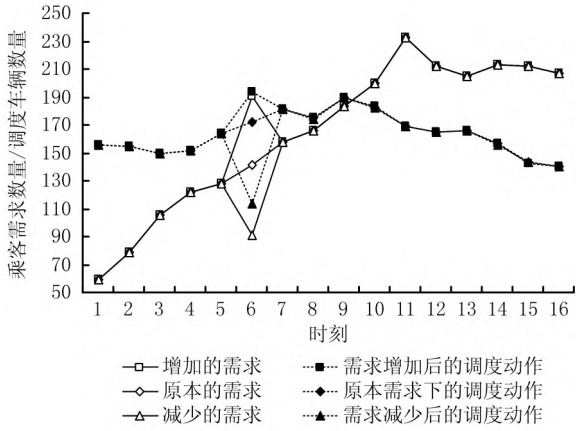

图11　D50-TD3模型在需求突增情况下的表现

策略梯度算法(TD3)框架,该框架由两个深度神经网络搭建。在实验中,首先对纽约市曼哈顿区域黄色出租车数据进行整理分析,然后假设系统动力学都是已知且确定的,因此可以通过混合整数规划得到奖励(总成本的负数)的理论上界。将双延迟深度确定性策略梯度算法应用在纽约市曼哈顿区域的黄色出租车的交通网络中。通过实验对比,在测试集上证实了 TD3 算法在需求不确定的情况下训练出来的模型的收敛性及有效性。同时,通过不确定交通需求和需求突变的情况来测试算法的鲁棒性,实验证明 TD3 算法能够有效应对需求不确定的情况。

本文还留下了很多有意思的值得拓展的研究。首先,本文实验建立在一个简化的交通网络上进行。由于不断增长的动作空间和状态空间,进行大规模的集中策略调度一直是一个挑战。未来可以尝试采用多智能体强化学习的方法,如 BOYALI[24] 将每个司机作为一个智能体,多个司机协同调度,从而可以有效提高调度系统运行的效率,SEOW[25] 采用多智能体模型,分布式调度出租车。其次本文实验中只考虑了单一模式的车辆,而在未来运营商可能由人类驾驶的车辆和无人驾驶出租车结合的车队组成[26],算法可以进一步结合两者的特点。除此之外,还可以进一步考虑拼车对调度策略的影响[27]。目前笔者的研究中没有考虑拼车系统,若能进一步考虑拼车系统,运营商就可以用更少的车辆满足更多的需求,进一步提高效率,节约能源,缓解交通拥堵。最后,目前只结合顾客的需求与现有的车辆进行调度,但可以参考更多的信息如交通情况等来参与决策,从而能利用更多的信息来进行优化调度。

参考文献:

[1] MACIEJEWSKI M, BISCHOFF J. Congestion effects of autonomous taxi fleets[J]. Transport, 2018, 33(4): 971-980.

[2] SAKHARE K V, TEWARI T, VYAS V. Review of vehicle detection systems in advanced driver assistant systems[J]. Archives of Computational Methods in Engineering, 2020, 27(2): 591-610.

[3] KUUTTI S, BOWDEN R, JIN Y, et al. A survey of deep learning applications to autonomous vehicle control[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(2): 712-733.

[4] JAVANSHOUR F, DIA H, DUNCAN G. Exploring the performance of autonomous mobility on-demand systems under demand uncertainty[J]. Transportmetrica A: Transport Science, 2019, 15(2): 698-721.

[5] CHEN Shengkai, FANG Shuiliang, TANG Renzhong. Demand forecasting based optimization of service configuration for cloud manufacturing[J]. Computer Integrated Manufacturing Systems, 2020,26(11):2944-2954(in Chinese).[陈晟恺,方水良,唐任仲. 基于需求预测的云制造服务租赁配置优化[J]. 计算机集成制造系统,2020,26(11):2944-2954.]

[6] WANG H, YANG H. Ridesourcing systems: A framework and review[J]. Transportation Research Part B: Methodological, 2019, 129: 122-155.

[7] BILLHARDT H, FERNÁNDEZ A, OSSOWSKI S, et al. Taxi dispatching strategies with compensations[J]. Expert Systems with Applications, 2019, 122: 173-182.

[8] KOOTI F, GRBOVIC M, AIELLO L M, et al. Analyzing Uber's ride-sharing economy[EB/OL]. [2021-04-04]. https://dl.acm.org/doi/pdf/10.1145/3041021.3054194.

[9] ZHANG R, PAVONE M. Control of robotic mobility-on-demand systems: a queueing-theoretical perspective[J]. The International Journal of Robotics Research, 2016, 35(1-3): 186-203.

[10] KIM B, KIM J, HUH S, et al. Multi-objective predictive taxi dispatch via network flow optimization[J]. IEEE Access, 2020, 8: 21437-21452.

[11] BOYACI B, ZOGRAFOS K G, GEROLIMINIS N. An optimization framework for the development of efficient one-way car-sharing systems[J]. European Journal of Operational Research, 2015, 240(3): 718-733.

[12] MA J, LI X, ZHOU F, et al. Designing optimal autonomous vehicle sharing and reservation systems: A linear programming approach[J]. Transportation Research Part C: Emerging Technologies, 2017, 84: 124-141.

[13] XIE Rong, PAN Wei, SHIBASAKI R. Intelligent taxi dispatching based on artificial fish swarm algorithm[J]. Systems Engineering—Theory & Practice, 2017, 37(11): 2938-2947(in Chinese).[谢　榕,潘　维,柴崎亮介. 基于人工鱼群算法的出租车智能调度[J]. 系统工程理论与实践,2017, 37(11): 2938-2947.]

[14] HE Shengxue, ZHAO Huiguang. Dispatching of taxipooling based on route optimization pattern[J]. Journal of Changsha University of Science & Technology: Natural Science, , 2018, 15(3): 14-20 (in Chinese). [何胜学, 赵惠光. 基于路径优化模式的出租车合乘调度[J]. 长沙理工大学学报: 自然科学版, 2018, 15(3): 14-20.]

[15] XIAO Pengfei, ZHANG Chaoyong, MENG Leilei, et al. Nonpremutation flow shop scheduling problem based on deep reinforcement learning[J]. Computer Integrated Manufacturing Systems, 2021, 27(1): 193-206(in Chinese). [肖鹏飞, 张超勇, 等. 基于深度强化学习的非置换流水车间调度问题[J]. 计算机集成制造系统, 2021, 2021, 27(1): 193-206.]

[16] KUUTTI S, BOWDEN R, JIN Y C, et al. A survey of deep learning applications to autonomous vehicle control[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(2): 712-733.

[17] CHEN Yong, WANG Haotian, YI Wenchao, et al. Algorithm of scheduling for multi-disturbance job-shop based on cellular automata and reinforcement learning[J]. Computer Integrated Manufacturing Systems, 2021, 27(12): 3536-3549 (in Chinese). [陈 勇, 王昊天, 易文超, 等. 基于元胞机与强化学习的多扰动车间调度算法[J]. 计算机集成制造系统, 2021, 27(12): 3536-3549.]

[18] ZHANG Jingling, FENG Qingbing, ZHAO Yanwei, et al. Hyper-heuristic for CVRP with reinforcement learning[J]. Computer Integrated Manufacturing Systems, 2020, 26(4): 1118-1129(in Chinese). [张景玲, 冯勤炳, 赵燕伟, 等. 基于强化学习的超启发算法求解有容量车辆路径问题[J]. 计算机集成制造系统, 2020, 26(4): 1118-1129.]

[19] LI Shengyi, MA Yumin, LIU Juan. Smart shop floor scheduling method for equipment load stabilization based on double deep q-learning[J/OL]. Computer Integrated Manufacturing Systems, 2021: 1-13[2021-04-30]. http://kns.cnki.net/kcms/detail/11.5946.tp.20210421.1622.018.html (in Chinese). [黎声益, 马玉敏, 刘 鹃. 基于双网络深度 Q 学习的面向设备负荷稳定的智能车间调度方法[J/OL]. 计算机集成制造系统, 2021: 1-13[2021-04-30]. http://kns.cnki.net/kcms/detail/11.5946.tp.20210421.1622.018.html.]

[20] MAO C, LIU Y L, SHEN Z J M. Dispatch of autonomous vehicles for taxi services: A deep reinforcement learning approach[J]. Transportation Research Part C: Emerging Technologies, 2020, 115: 102626.

[21] GRONDMAN I, BUSONIU L, LOPES G A D, et al. A survey of actor-critic reinforcement learning: Standard and natural policy gradients[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2012, 42(6): 1291-1307.

[22] FUJIMOTO S, HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[EB/OL]. [2021-04-30]. http://proceedings.mlr.press/v80/fujimoto18a/fujimoto18a.pdf.

[23] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[EB/OL]. [2021-04-30]. https://arxiv.org/pdf/1509.02971.pdf.

[24] BOYALI A, HASHIMOTO N, JOHN V, et al. Multi-agent reinforcement learning for autonomous on demand vehicles [C]//Proceedings of the IEEE Intelligent Vehicles Symposium (IV). Washington, D.C., USA: IEEE, 2019.

[25] SEOW K T, DANG N H, LEE D H. A collaborative multiagent taxi-dispatch system[J]. IEEE Transactions on Automation Science and Engineering, 2009, 7(3): 607-616.

[26] DUAN L, WEI Y, ZHANG J, et al. Centralized and decentralized autonomous dispatching strategy for dynamic autonomous taxi operation in hybrid request mode[J]. Transportation Research Part C: Emerging Technologies, 2020, 111: 397-420.

[27] LOKHANDWALA M, CAI H. Dynamic ride sharing using traditional taxis and shared autonomous taxis: A case study of NYC[J]. Transportation Research Part C: Emerging Technologies, 2018, 97: 45-60.

作者简介:

周晓婷(1999－),女,广东佛山人,硕士研究生,研究方向:智能优化算法、强化学习等,E-mail: zhouxt8@mail2.sysu.edu.cn;

吴禄彬(1997－),男,广东汕头人,硕士研究生,研究方向:鲁棒优化、决策理论与方法,E-mail: wulb7@mail2.sysu.edu.cn;

章 宇(1989－),男,四川内江人,教授,博士,博士生导师,研究方向:鲁棒优化方法及其在物流、供应链、交通、医疗管理中的应用,E-mail: y.zhang@swufe.edu.cn;

＋**姜善成**(1989－),男,辽宁丹东人,助理教授,博士,硕士生导师,研究方向:智能优化算法、深度学习等,通讯作者,E-mail: jiangshch3@mail.sysu.edu.cn。