ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 3,May 2023,pp. 652－664
DOI:10. 16337/j. 1004-9037. 2023. 03. 014

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

一种全局供需感知的均值场多智能体强化学习订单分配算法

宋　旺，胡　祥，张玉辉，卫文江，周雅诗，康　傲

（华北电力大学控制与计算机工程学院，北京 102206）

摘　要：提出一种具备全局供需动态感知能力、基于均值场多智能体强化学习的网约车平台订单分配算法。该算法通过将多智能体强化学习与均值场理论相结合，提升了智能体在局部空间上相互之间的协作性；通过注入全局空间上供需的动态分布信息，提升了智能体对全局供需分布的感知和优化能力。本文构建了真实历史数据驱动的模拟器，用于算法的训练和评估。实验表明，在全天时段和高峰期时段两个不同场景下，本文提出的算法在网约车司机累计收益及订单应答率两个重要指标上均显著优于现有的订单分配算法。实验结果充分验证了本文提出算法的有效性。

关键词：多智能体强化学习；均值场；全局供需动态感知；网约车平台；订单分配

中图分类号：TP391　　**文献标志码：**A

Mean-Field Multi-agent Reinforcement Learning Order Dispatch Algorithm with Awareness of Global Supply-Demand Dynamics

SONG Wang, HU Xiang, ZHANG Yuhui, WEI Wenjiang, ZHOU Yashi, KANG Ao

(School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China)

Abstract: This paper proposes an order dispatch algorithm of online ride-hailing platform based on mean-field multi-agent reinforcement learning with the ability to globally perceive supply-demand dynamics. Our algorithm improves the collaboration between agents in the local area by integrating multi-agent reinforcement learning with mean-field theory, and enhances the ability of agents on perceiving and optimizing the global supply-demand gap across the global area by injecting the context about global supply-demand dynamics. Besides, we built a data-driven simulator for the training and evaluation of algorithms. Extensive experiments show that in two different scenarios of a whole day and rush hour, our algorithm significantly outperforms the existing order dispatch algorithms in terms of order response rate and accumulated drivers' income. The experimental results convincingly validate the effectiveness of our algorithm.

Key words: multi-agent reinforcement learning; mean-field; global perceive supply-demand dynamics; online ride-hailing platform; order dispatch

基金项目：国家自然科学基金(52078212)。
收稿日期：2022-05-19；**修订日期：**2022-08-20

引　言

随着4G网络技术、全球定位技术的迅猛发展,网约车服务正深刻地改变和重塑着人们的出行方式。它不但减少了乘客路边等待时间、提升了乘客的出行体验[1],还缓解了城市交通拥堵、优化了交通资源分配[2],正在成为一种重要的城市公共交通模式。

网约车服务的蓬勃发展催生了一批网约车平台,包括:滴滴出行、Uber等。而作为支撑平台业务的核心任务,订单分配是平台必须解决的关键问题。订单分配问题是指平台在收到乘客提交的出行订单后,如何将订单以最优的方式分配给空驶司机,从而减小订单需求和空车供给在时空上的分布差距,最大程度撮合供需双方交易的问题。订单分配问题解决得好坏不仅影响到乘客的用户体验,还影响到司机和平台的收益,甚至影响到用户和司机对平台的忠诚度。图1展示了订单分配任务及其复杂性。图中圆圈代表订单、三角形代表空驶司机。距离远收益高的订单可能会使司机驶往订单稀少的偏远地区,减少司机的长期收益;距离近收益低的订单也有可能使司机驶往订单多的热点地区,增加司机的长期收益。

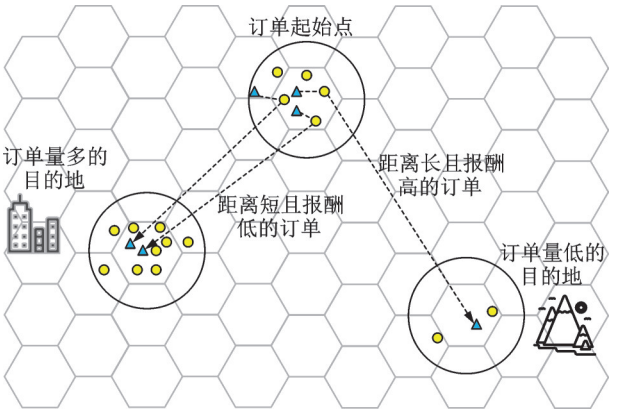


图1　订单分配任务及其复杂性示意图

Fig.1　Illustrative diagram about order dispatch task and its complexity

目前,解决订单分配问题的方式主要有两种,分别是基于组合优化的和基于强化学习的方法[3]。基于组合优化的方法将订单分配问题刻画为由订单和空驶车辆组成的二分图,从而将原问题转化为二分图的匹配问题[4-6]。但基于组合优化的方法往往存储和计算开销太大,无法满足平台对算法实时性的要求。得益于强化学习等研究领域的快速进步,基于强化学习的订单分配研究也得到了快速发展并取得了丰硕的成果。文献[7-11]提出了集中式控制架构的强化学习订单分配算法,此类算法能够在有效减少计算开销的同时提升网约车平台的司机累计收益和订单应答率,但是集中控制架构存在因中央控制器故障而引发系统整体瘫痪的“单点失效”风险[12]。因此,最近的研究中多采用基于分布式控制架构的多智能体强化学习[13]订单分配算法。其中,Li等[14]引入均值场强化学习[15]的思想来提升位于同一局部空间的智能体之间的协作能力,而Zhou等[16]通过提升订单和空驶车辆在整个空间的分布一致性来提升算法的整体效果。

本文提出了一种全局供需感知的均值场多智能体强化学习订单分配算法。该算法通过构建全面反映局部供需关系的均值特征,提升了智能体在局部区域的相互协作性;通过为策略网络增加旨在提高全局供需感知能力的KL散度[17]调节因子,强化了算法对供需分布全局一致性的优化能力,进而提升了平台的司机累计收益和订单应答率。

订单分配问题早期研究的思路主要遵循了“就近分配原则”,形成了在空间维度上订单就近分配给附近的空车、在时间维度上订单分配给应答最快空车的“抢单”分配模式[18-19]。“抢单”分配模式虽然简单并且易于实现,但是由于没有对分配过程在时间和空间上进行全局优化,所以无法满足平台对算法的性能要求。为了克服“抢单”模式的性能瓶颈,研究者们将订单分配问题的研究思路转向“派单”分配模式。“派单”模式初期普遍采用了组合优化的方法,如Gao等[5]构建了一个包含网约车净利润总额和乘客等待时间的函数,以此进行权重计算,并通过KM算法[20]进行司机和订单的匹配;文献[6]提出在每次

分配前对供需进行实时预测，通过预测的结果来指导算法进行全局优化。然而，大多数基于组合优化的方法不仅计算开销巨大，而且优化过程没有考虑对司机长期收益的影响，因而难以满足平台的应用需求。

为了解决这一问题，强化学习方法逐渐被用于订单分配当中。Xu等[7]将订单分配问题建模为马尔科夫决策问题，根据每一个订单的价格、起始位置等特征建立订单价值表，然后使用KM算法等匹配方法将司机和订单进行匹配。Wang等[9]利用DQN算法[21]框架，提出一种基于深度强化学习的订单分配方法。Wang等[10]提出了一种基于强化学习的动态二分图匹配方法，每个时刻根据二分图中的司机数、订单数、订单收益等特征来决定此时对二分图进行匹配或延迟到下个时刻再对其进行匹配，即动态地改变二分图匹配的时间窗口来适应不同的供需情况。基于强化学习的订单分配算法能够很好地应用于大规模的时空环境，但是以上算法都基于集中控制的结构，会产生“单点失效”问题。

近期，基于分布式控制的多智能体强化学习被用于订单分配算法当中。Delima等[22]提出了一种基于QMIX[23]的多智能体强化学习订单分配算法，但其仅在小规模应用场景取得了一定的效果；Li等[14]在多智能体强化学习的基础上引入均值场以提升局部智能体之间的协作性，但他们没有考虑订单和空驶车辆在地理空间上的全局分布一致性；Zhou等[16]通过KL散度在算法中加入对订单和空驶车辆全局分布一致性的考虑来提升算法的整体收益，但他们没有考虑局部智能体之间的协作性。

受到文献[14,16]研究的启发，本文提出了一种全局供需感知的均值场多智能体强化学习订单分配算法。与文献[14,16]的方法不同，本文提出的算法不仅同时考虑智能体的局部协作性和供需分布的全局一致性，而且运用特征工程方法构建了向量化的均值特征，从而增强了智能体对局部供需关系的感知能力，进一步提升了算法的性能。

1 问题描述

本文旨在解决网约车平台的订单分配问题，通过对空车供给和订单需求进行优化分配，实现司机累计收益与订单应答率的最大化。为了简化问题的描述，这里先作两个假设：(1)问题的离散时空假设，即：在时间维度上，用等长的时间片对时间进行离散化，在空间维度上，用六边形网格世界对城市地图进行离散化；(2)派单范围假设：为了确保将接乘时间限定在合理区间内，与大多数平台的运营情况一致，本文假设订单分配问题的派单范围为订单所在的网格和与其直接相邻的6个网格。

基于以上假设，在每一时间片t，订单分配过程可以描述为：(1)乘客提交出行订单；(2)平台根据所有空车和订单在网格空间上的离散分布情况，将出行订单分配给其派单范围内的空车；(3)空车司机去往订单起点接乘，经过时间Δt将乘客送达目的地后收到平台的奖励；(4)空车司机以上一个订单目的地为起点继续等待平台派单。如图2所示。

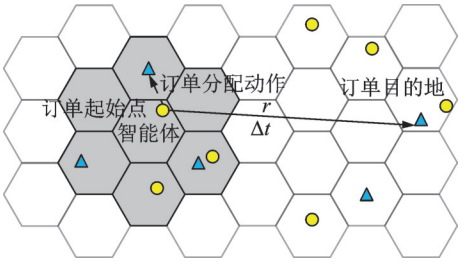


图2 订单分配过程示意图

Fig.2 Procedure of order dispatch

由于下一个时间步的订单分配策略仅依赖于上一个时间步分配决策后形成的订单、空车分布等状态信息，与上个时间步之前的状态信息无关。因此，订单分配过程是典型的马尔科夫决策过程，订单分配问题是一个典型的马尔科夫决策问题。由于整个订单分配的决策过程可以分解到各个网格，并且使用局部观测值可以有效降低状态信息在决策过程中带来的通信压力，因此进一步将订单分配问题建模为一个基于多智能体决策、局部可观测的马尔科夫决策问题[24]：$M=<N,\mathcal{S},\mathcal{O},\mathcal{A},\mathcal{P},\mathcal{R},\gamma>$。其中，$N$代表智能体的个数，$\mathcal{S}$代表全局状态空间，$\mathcal{O}$代表智能体的局部观测空间，$\mathcal{A}$代表智能体的动作空间，$\mathcal{P}$代表状态转移概率，$\mathcal{R}$代

表奖励函数,γ代表衰减因子。相关定义如下:

(1) 智能体:把地图上每一个六边形网格定义为一个智能体,故与网格数相同,智能体的数量为N,且不随时间发生改变。

(2) 状态空间$\mathcal{S}$:t时刻全局状态$\boldsymbol{s}_t\in\mathcal{S}$,是指$t$时刻空车和订单在全部网格上的数量分布情况以及一个代表当前系统时刻的One-hot编码,$\boldsymbol{s}_t\in\mathbf{R}^{N\times 2+T}$。

(3) 局部观测空间$\mathcal{O}$:$\mathcal{O}$是全局状态$\mathcal{S}$的1个子集。t时刻智能体$\mathbf{Agent}_i$的局部观测值$\boldsymbol{o}_t^i\in\mathcal{O}$,是指空车和订单在其派单范围网格上的数量分布情况,以及分别代表时间和空间的One-hot编码,$\boldsymbol{o}_t^i\in\mathbf{R}^{7\times 2+T+N}$。

(4) 联合动作空间$\mathcal{A}$:t时刻所有智能体的联合动作$\boldsymbol{a}_t\in\mathcal{A}$,$\mathcal{A}=\mathcal{A}_1\times\mathcal{A}_2\times\cdots\times\mathcal{A}_N$。其中,$\mathcal{A}_i$代表单个智能体$\mathrm{Agent}_i$的派单范围,即:每个网格智能体可以执行7个不同的派单动作,将所在网格内的订单派往包括所在网格在内的7个网格区域。智能体$\mathbf{Agent}_1$的动作空间如图3灰色区域所示。

(5) 状态转移概率$\mathcal{P}$:状态转移概率$\mathcal{P}$是t时刻所有智能体执行联合动作$\boldsymbol{a}_t$后全局状态$\boldsymbol{s}_t$转移为$\boldsymbol{s}_{t+1}$的概率,$p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,\boldsymbol{a}_t):\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]$。

(6) 奖励函数$\mathcal{R}$:联合奖励函数$\mathcal{R}$是N个智能体根据环境状态和联合动作计算的整体奖励,单个智能体的奖励函数$\mathcal{R}_i$则表示根据环境状态及该智能体所做的动作计算其应获得的奖励,$\mathcal{R}_i:\mathcal{S}\times\mathcal{A}\to\mathbf{R}$。

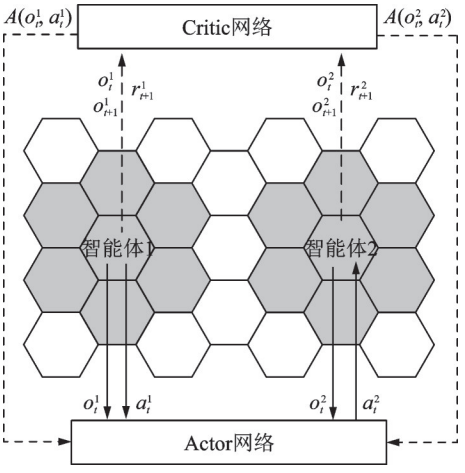


图3　POOD算法执行架构图

Fig.3　Framework of the POOD algorithm

2　本文方法

为了求解上述局部可观测马尔科夫决策问题,提出了一种新的多智能体强化学习订单分配算法。

2.1　基于局部观测的订单分配算法

结合多智能体强化学习和A2C算法,本文首先提出基于局部观测的订单分配算法框架(Partially observed order dispatch, POOD)。该算法框架假定所有智能体具备同质性,采用了共享的Actor-Critic模型[25]来进行训练和决策,如图3所示。由于位于地理边界的智能体在某些方向上没有邻居,该智能体的动作空间会受到约束,因此在算法中引入掩码机制来解决地理边界带来的动作空间约束问题。

在Actor-Critic模型中,Critic网络为智能体的状态价值函数,它的输入为智能体i的局部观测$\boldsymbol{o}_t^i$,输出为智能体所处状态的状态价值$V\left(\boldsymbol{o}_t^i\right)$,根据贝尔曼方程,Critic网络通过最小化下列损失函数来进行参数更新,即

$$L(\theta_v)=\left(V_{\theta_v}\left(\boldsymbol{o}_t^i\right)-V_{\text{target}}\left(\boldsymbol{o}_{t+1}^i;\ \theta_v',\pi\right)\right)^2 \tag{1}$$

$$V_{\text{target}}\left(\boldsymbol{o}_{t+1}^i;\ \theta_v',\pi\right)=\sum\nolimits_{a_t^i}\pi\left(a_t^i|\boldsymbol{o}_t^i\right)\left(\boldsymbol{r}_{t+1}^i+\gamma V_{\theta_v'}\left(\boldsymbol{o}_{t+1}^i\right)\right) \tag{2}$$

式中:θ_v为价值网络的网络参数;θ_v'为目标价值网络的网络参数;$\pi(\cdot)$为策略概率;$\boldsymbol{r}_{t+1}^i$为智能体i在$t+1$时刻所获得的即时奖励。

Actor网络为动作网络,它的输入同样为智能体i的局部观测$\boldsymbol{o}_t^i$,但是输出为此时该智能体选择每个动作的概率。

用$\boldsymbol{P}\left(\boldsymbol{o}_t^i\right)\in\mathbf{R}^7$代表Actor网络中输出的智能体$i$在$t$时刻所对应7个动作的概率，同时令智能体$i$在$t$时刻所对应7个动作的价值向量为$\boldsymbol{q}_{\text{valid}}\left(\boldsymbol{o}_t^i\right)=\boldsymbol{P}\left(\boldsymbol{o}_t^i\right)*\boldsymbol{G}_i$，这里$\boldsymbol{G}_i$为与智能体$i$位置有关的一个7维掩码向量。当智能体$i$处于地图边界，它的某个方向没有邻居时，$\boldsymbol{G}_i$向量中对应位置的掩码值为0，否则为1。于是智能体$i$向其第$k$个邻居分配订单的概率可以统一表示为

$$\pi_{\varnothing}\left(a_t^i=k|\boldsymbol{o}_t^i\right)=\frac{\left[\boldsymbol{q}_{\text{valid}}\left(\boldsymbol{o}_t^i\right)\right]_k}{\left\|\boldsymbol{q}_{\text{valid}}\left(\boldsymbol{o}_t^i\right)\right\|_1} \tag{3}$$

更新Actor网络的梯度公式为

$$\nabla_{\varnothing}J\left(\varnothing\right)=\nabla_{\varnothing}\log\pi_{\varnothing}\left(a_t^i=k|\boldsymbol{o}_t^i\right)A\left(\boldsymbol{o}_t^i,a_t^i\right) \tag{4}$$

式中：$\varnothing$为Actor网络的网络参数；$A\left(\boldsymbol{o}_t^i,a_t^i\right)$为每个动作的优势函数。本文用值函数的单步差分进行评估，即

$$A\left(\boldsymbol{o}_t^i,a_t^i\right)=r_{t+1}^i+\gamma V_{\theta_v^t}\left(\boldsymbol{o}_{t+1}^i\right)-V_{\theta_v}\left(\boldsymbol{o}_t^i\right) \tag{5}$$

2.2 基于均值场强化学习的订单分配算法

在订单分配过程中，因为有些智能体之间距离较近（不一定直接相邻），所以在它们之间的局部区域可能出现订单的重复分配、冲突分配等问题。解决这些问题的有效思路是促进局部智能体之间的协作，而POOD算法不具备这样的能力。

为了使算法能够促进智能体之间的局部协作性，借鉴Li等[14]的思想，通过构建能显式表达局部空间供需关系的均值特征，将POOD改进为一种基于均值场强化学习的订单分配算法（Mean-field actor-critic，MFAC）。

利用均值场多智能体强化学习提升智能体局部协作性的关键在于：在协同范围内构建全面反映协同目标的均值特征。这里协同范围不是指动作范围，是指一个智能体需要感知其他智能体的最远距离。如图4所示，智能体动作范围为半径为r的圆覆盖的灰色区域7个网格，而协同范围为半径为$2r$的圆覆盖的第1层和第2层邻居网格，r为相邻六边形网格中心点之间的距离。图4两个智能体不仅动作范围存在重合区域（图上深色区域），并且协同范围也存在重合区域。对于协同范围存在重合区域的智能体，如果它们的动作缺少协作性，就可能在协同范围重合区出现冲突派单、重复派单等低效派单动作。因此，如果要提高智能体的局部协作性，那么智能体就不仅要感知自己本地的环境状态，而且要感知协同范围内的其他智能体及它们执行动作后对环境形成平均影响。

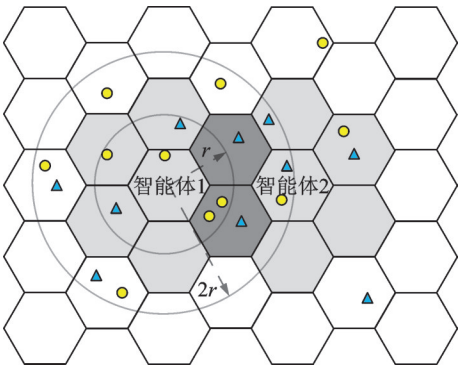


图4 智能体协同范围示意图

Fig.4 Demonstration of the collaboration range

由此可以构建能够反映智能体i协同范围内供需关系的均值特征，即有

$$\bar{o}_i=\frac{N_i^{\text{driver}}}{N_i^{\text{order}}} \tag{6}$$

式中N_i^{driver}、N_i^{order}分别为智能体i协同范围内的空车司机总数和订单总数。例如在图4中圆圈代表订单，三角形代表空车司机，对智能体1而言，$N_1^{\text{driver}}=8$，$N_1^{\text{order}}=9$，$\bar{o}_1=8/9$。

MFAC通过将均值特征$\bar{o}_i$加入到智能体的局部观测,来增强智能体对协同范围内供需关系的感知能力,进而促进智能体动作的局部协作性。加入均值特征后,Critic网络的损失函数为

$$L(\theta_v)=\left(V_{\theta_v}\left(\boldsymbol{o}_t^i,\bar{o}_t^i\right)-V_{\text{target}}\left(\boldsymbol{o}_{t+1}^i,\bar{o}_i^{t+1};\theta_v',\pi\right)\right)^2 \tag{7}$$

相应地,Actor网络通过如下梯度公式进行更新

$$\nabla_{\varnothing}J\left(\varnothing\right)=\nabla_{\varnothing}\log\pi_{\varnothing}\left(a_t^i=k|\boldsymbol{o}_t^i,\bar{o}_t^i\right)A\left(\boldsymbol{o}_t^i,\bar{o}_t^i,a_t^i\right) \tag{8}$$

$$A\left(\boldsymbol{o}_t^i,\bar{o}_t^i,a_t^i\right)=r_{t+1}^i+\gamma V_{\theta_v'}\left(\boldsymbol{o}_{t+1}^i,\bar{o}_{t+1}^i\right)-V_{\theta_v}\left(\boldsymbol{o}_t^i,\bar{o}_t^i\right) \tag{9}$$

2.3　基于多维向量均值场的订单分配算法

尽管MFAC构建了用于显式地表述本地供需关系的一维均值特征,提高了智能体在局部空间的相互协作性,但是该特征过于简化,不能全面准确地反映本地供需关系。例如,在智能体两层邻居范围内的司机数量为9、订单数量为10的情况下,均值特征的值为0.9,而当智能体两层邻居范围内的司机数量为90,订单数量为100时,均值特征的值仍为0.9,而智能体所处的这两种环境状态显然是不同的。由此可见,使用司机数量和订单数量的比值作为唯一的均值特征,虽然能够在一定程度表征智能体所在局部地区的供需情况,但是它的表达能力十分有限,无法区分那些供需比相同、但量级差别很大的局部环境。所以需要进一步增强均值特征对局部供需关系的表达能力,使它能够更好地适应大规模复杂场景下的订单分配任务。

基于上述原因,本文运用特征工程方法构建了全面反映局部供需关系的均值特征,将均值特征由一维均值升级到多维向量均值,提出了一种全新的、基于向量均值的多智能体强化学习订单分配算法(Vector mean field actor-critic, VMFAC)。

从不同角度设计了以下多维均值特征为

$$\bar{\boldsymbol{o}}_i^a=\left(\frac{N_i^{\text{driver}}}{N_i^{\text{order}}},\left|N_i^{\text{driver}}-N_i^{\text{order}}\right|\right) \tag{10}$$

$$\bar{\boldsymbol{o}}_i^b=\left(\frac{N_i^{\text{driver}}}{N_i^{\text{order}}},N_i^{\text{driver}}\right) \tag{11}$$

$$\bar{\boldsymbol{o}}_i^c=\left(\frac{N_i^{\text{driver}}}{N_i^{\text{order}}},N_i^{\text{order}}\right) \tag{12}$$

新的均值特征能够更准确地描述智能体所处局部环境中空车和订单的实际供需差距,有效地避免了对两种供需比相近但量级完全不同的局部环境状态不加区分的问题,进一步促进了智能体之间的局部协作性。此时,VMFAC算法的Critic网络损失函数为

$$L(\theta_v)=\left(V_{\theta_v}\left(\boldsymbol{o}_t^i,\bar{\boldsymbol{o}}_t^i\right)-V_{\text{target}}\left(\boldsymbol{o}_{t+1}^i,\bar{\boldsymbol{o}}_{\boldsymbol{t}+1}^i;\theta_v',\pi\right)\right)^2 \tag{13}$$

相应地,Actor网络的梯度更新公式为

$$\nabla_{\varnothing}J\left(\varnothing\right)=\nabla_{\varnothing}\log\pi_{\varnothing}\left(a_t^i=k|\boldsymbol{o}_t^i,\bar{\boldsymbol{o}}_t^i\right)A\left(\boldsymbol{o}_t^i,\bar{\boldsymbol{o}}_t^i,a_t^i\right) \tag{14}$$

$$A\left(\boldsymbol{o}_t^i,\bar{\boldsymbol{o}}_t^i,a_t^i\right)=r_{t+1}^i+\gamma V_{\theta_v'}\left(\boldsymbol{o}_{t+1}^i,\bar{\boldsymbol{o}}_{t+1}^i\right)-V_{\theta_v}\left(\boldsymbol{o}_t^i,\bar{\boldsymbol{o}}_t^i\right) \tag{15}$$

2.4　本文订单分配算法

虽然VMFAC能够有效地进一步促进智能体之间的局部协作性,但是它缺少对全局空间内供需分布一致性的直接考虑。如果算法能够更好地感知并优化全局供需分布的一致性,就可以有效地缩小全局的供需差距,从而提升算法在全局空间内的表现。因此,本文提出了一种全局供需感知的均值场多智能体强化学习订单分配算法(Vector mean field actor critic with Kullback-Leibler divergence, CORNFLAKE)。

在CORNFLAKE算法中，本文通过KL散度来衡量订单和空车司机在空间上的全局一致性，t时刻KL散度的计算公式为

$$D_{KL}^{t}=\sum_{i=1}^{N}p_t(i)\log\frac{p_t(i)}{q_t(i)} \tag{16}$$

$$p_t(i)=\frac{t\text{时刻第}i\text{个网格中订单个数}}{t\text{时刻整个空间内的订单总数}} \tag{17}$$

$$q_t(i)=\frac{t\text{时刻第}i\text{个网格中空车个数}}{t\text{时刻整个空间内的空车总数}} \tag{18}$$

本文希望得到在t时刻执行后能够使$t+1$时刻的KL散度尽可能小，即全局供需分布尽可能一致的策略为

$$\pi^{*}=\arg_{\pi}\min D_{KL}^{t+1} \tag{19}$$

考虑到供需一致性对订单分配影响的前瞻性和全局性，将供需一致性约束施加到Critic网络，通过改善Critic对各网格订单热度的估值，间接地将供需一致性考虑因素传递至Actor网络的订单分配过程，施加一致性约束后Critic网络的损失函数为

$$L(\theta_v)=\left(V_{\theta_v}\left(\boldsymbol{o}_t^i,\bar{\boldsymbol{o}}_t^i\right)-V_{\text{target}}\left(\boldsymbol{o}_{t+1}^i,\bar{\boldsymbol{o}}_{t+1}^i;\theta_v',\pi\right)\right)^2+\lambda D_{KL}^{t+1} \tag{20}$$

式中λ决定了算法对全局供需一致性的感知程度。

3 数据处理及结果分析

本节首先介绍实验的前期准备工作，包括数据处理、模拟器构建等。然后通过实验验证提出的全局供需感知的均值场多智能体强化学习订单分配算法的有效性。最后分别对算法各改进阶段的有效性、算法在采用不同均值特征时的性能、以及全局供需一致性参数λ对模型的影响进行了分析。

3.1 实验前期准备

在本文中，通过对Lin等[26]提出的网约车空车调度模拟器进行改进，构建了适用于订单分配的模拟器。该模拟器将城市地图以六边形网格进行空间离散化，按10 min为时间步长进行时间离散化，将一天划分为144个时间步，采用历史数据回放的形式对订单产生、订单分配等过程进行了模拟。

本文所使用的数据为滴滴出行公司“盖亚”数据开放计划提供的公开数据集[27]，该数据集包含滴滴出行公司2016年11月份在成都市共计30天的订单历史数据以及相应的网约车轨迹数据。其中，订单历史数据包括：订单编号、计费开始时间、计费结束时间、上车位置经纬度、下车位置经纬度以及司机完成该订单获得的收益；轨迹数据包括：司机编号、订单编号以及网约车完成订单经过的空间坐标轨迹。本文使用了该数据集中的订单历史数据，并对数据的使用进行了合理的简化：(1)用计费开始时间模拟订单产生的时间，用乘客上车位置模拟订单产生的位置；(2)用订单完成时间模拟空车产生的时间，用乘客下车位置模拟空车产生的位置。对30天订单数据集的统计分析表明，该城市平均每天的订单量约为24万，每天各个时间段产生的订单数量具有明显的规律性和相似性：每天的变化规律基本一致，订单需求具有显著的高峰期，并且基本上出现在80～110时间步时段，约下午2点至7点时段。全天订单量随时间变化情况如图5所示。为了将订单历史数据注入模拟器，本文对原始数据作如下预处理：

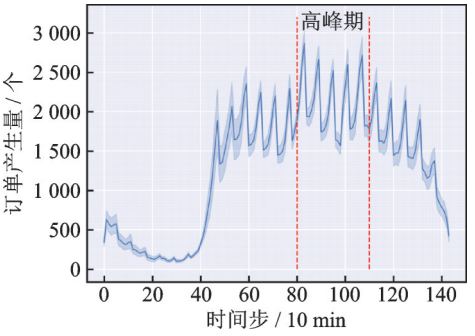


图5 全天订单量随时间变化情况

Fig.5 Order requests over time in a whole day

(1)地理区域网格化预处理:首先统计订单起止位置的经纬度,按照99.98%的位置覆盖率确定本研究的目标地理区域。然后将该区域划分成边长为1.2 km的六边形网格地图。网格地图包含504个六边形网格,网格按照地理位置排列成为21×24的二维阵列,每个网格设置单独的网格编号。

(2)订单数据离散化预处理:将全天24 h以10 min作为步长进行时间离散处理,全天离散化为144个时间步,每个时间步设置单独的时间编号。将原始订单数据的连续时空信息映射到由离散时间和网格地图组成的离散时空。离散化预处理后的订单数据如下:订单产生地网格编号、目的地网格编号,订单产生时所在的时间步编号、订单持续时间步数、以及完成订单获得的收益等。

实验的运行环境为Ubuntu SMP 16.04.1操作系统,Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40 GHz处理器,内存大小为128 GB。

参数设置如下:(1) 模型参数:Critic和Actor网络均为全连接神经网络,中间隐藏层的神经元个数均为128、64、32;均使用ReLU激活函数;(2) 训练参数:学习率α为0.001,衰减因子γ为0.95,批次大小为3 000,每回合Critic网络和Actor网络训练4000个批次,训练约60个回合,两个网络的损失函数均收敛到稳定值。其中,在参数选取阶段学习率α的取值范围为[0.000 1,0.1),衰减因子γ的取值范围为[0.8,1),最终的参数取值均为利用Optuna[28]调参获得的优化参数。

为了获得稳定的评估结果,本文使用10个不同的随机种子,按照历史同期均匀随机采样的方法,采样10个全天订单序列数据用于评估,并且取10次评估的均值作为最终结果。评价指标如下:(1)ADI(Accumulated drivers' income):平台中所有司机所获得的累计收益;(2)ORR(Order response rate):订单应答率,即网约车平台成功分配给司机的订单数量与平台接到的订单总数的比值。

本文使用如下基准方法与本文提出的方法在ADI和ORR两个指标上进行比较。(1)RAN(Random):随机的订单分配算法,不考虑任何额外因素,将订单随机分配给司机。(2)Revenue-based(REV):基于报酬的订单分配算法[14],在每个时间段,该算法把订单按照其价格进行从高到低的排序,优先将价格高的订单分配给空车司机。(3)Response-based(RES):基于订单应答率的分配算法[14],在每个时间段,该算法通过把订单按照其起点到终点的距离进行排序,优先将距离近的订单分配给空车司机。(4)DQN:根据Lin等[26]提出的方法构建的基于DQN的多智能体强化学习订单分配算法,有别于本文方法,它的输入为全局状态。(5)KL-Based:根据Zhou等[16]的思想构建的一个基于DQN的多智能体强化学习订单分配算法,它的输入为全局状态,通过KL散度考虑订单和空车在全局空间的分布一致性来做全局优化。(6)CORNFLAKE:本文提出的全局供需感知的均值场多智能体强化学习订单分配算法。

3.2　实验结果及分析

3.2.1　总体性能比较

与现有订单分配算法研究的比较方法一致,本文也没有直接比较各算法在评价指标上的性能,而是比较各算法相对于RAN算法在评价指标上的提升百分比。表1详细比较了CORNFLAKE和各个基准模型在全天时段和高峰期时段两个场景下的总体性能。从表1可以看出,由于增强了智能体局部协作性以及智能体对全局供需的感知和优化能力,尽管CORNFLAKE算法中智能体只拥有局部观测能力,但它在全天时段和高峰

表1　各方法相对于RAN方法在ORR和ADI上的性能比较

Table 1　Performance comparison in terms of ADI and ORR with respect to RAN　%

模型	全天		高峰期	
	ADI	ORR	ADI	ORR
REV	+0.70	+0.01	+0.90	+0.16
RES	−0.83	−0.15	−1.14	−0.38
DQN	+7.01	+2.64	+13.90	+4.17
KL-based	+14.05	+6.66	+25.09	+13.65
CORNFLAKE	+26.22	+14.42	+57.03	+37.72

期时段两个场景下的表现优于已有的基于全局观测的基准方法。

3.2.2 算法改进有效性分析

本文以POOD算法框架为基础，经过3次增量改进，形成最终的CORNFLAKE算法。为了验证每次改进的有效性，本文对改进各阶段算法的性能作了比较，如图6所示。可以看出，MFAC算法相较于POOD算法，在全天时段和高峰期时段两个场景下的ADI、ORR表现有小幅提升，说明引入均值特征对局部空间供需关系进行表征能够有效促进智能体的局部协作性，从而提升算法表现；VMFAC算法在全天时段和高峰期两个时段两个场景下的ADI、ORR表现均有大幅提升，说明采用多维向量均值特征能够更好地促进局部智能体之间的交互，从而提升算法效果；在VMFAC算法的基础上加入KL散度对全局供需信息进行感知和优化后，CORNFLAKE算法在全天的ADI、ORR表现有了小幅度的提升，说明利用KL散度增强订单和空驶司机在全局空间的分布一致性能够使算法具有更好的全局表现，但由于在城市实际环境中每时每刻有订单产生，不仅每个时刻订单量不同，而且订单在全局空间的分布不断变化。而由于城市中用户的生活习惯、上下班时间等因素的影响，不同时段订单分布的变化程度不同。当相邻时刻订单分布变化程度比较大时，相应地，订单调度算法中值函数的状态输入也会产生比较大的变化，这时如果值函数对剧烈变化的状态输入的估值也急剧变化，即：对每个网格地区的状态价值估计也产生相对比较大的波动，状态值估计的不稳定性会在后续的订单调度策略中得到扩散和强化。当一段连续时间内的订单调度策略一直都在剧烈变化时，就会使模型的动作输出处于一种相对不稳定的状态，无法十分有效地进行全局空间内供需分布一致性的优化。所以在CORNFLAKE算法加入KL散度直接对全局供需信息进行感知和优化后，算法在ADI和ORR的表现提升有限。

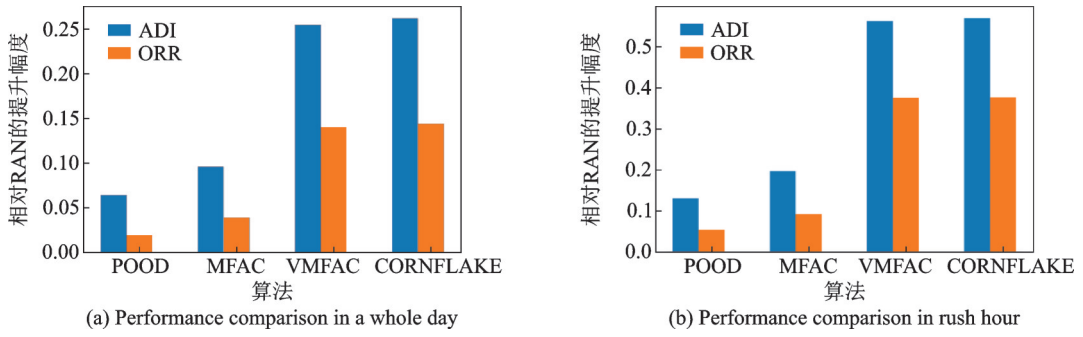


图6 全天时段和高峰期时段不同算法性能比较

Fig.6 Performance comparation of different algorithms in a whole day and rush hour

3.2.3 均值特征构建及性能分析

本文构建了多个不同的均值特征用于基于均值场强化学习的订单分配算法，并且比较和分析了各个均值特征对算法的性能影响。除了上面提到的3个多维均值特征 $\bar{\boldsymbol{o}}_i^a$、$\bar{\boldsymbol{o}}_i^b$、$\bar{\boldsymbol{o}}_i^c$ 外，还构建了以下4个一维均值特征

$$\bar{o}_i^d = N_i^{\text{driver}} \tag{21}$$

$$\bar{o}_i^e = N_i^{\text{order}} \tag{22}$$

$$\bar{o}_i^f = \frac{N_i^{\text{driver}}}{N_i^{\text{order}}} \tag{23}$$

$$\bar{o}_i^g = \left| N_i^{\text{driver}} - N_i^{\text{order}} \right| \tag{24}$$

实验结果如图7所示。实验结果表明,算法使用多维向量均值特征时的表现都要优于其使用标量均值特征时的表现,这说明多维向量均值特征能够更好地捕捉局部供需信息,从而促进局部智能体之间进行协作。并且,当均值特征取$\bar{\boldsymbol{o}}_i^c$时算法在全天时段和高峰期时段两个场景下的ADI、ORR均高于均值特征取其他值时的结果。

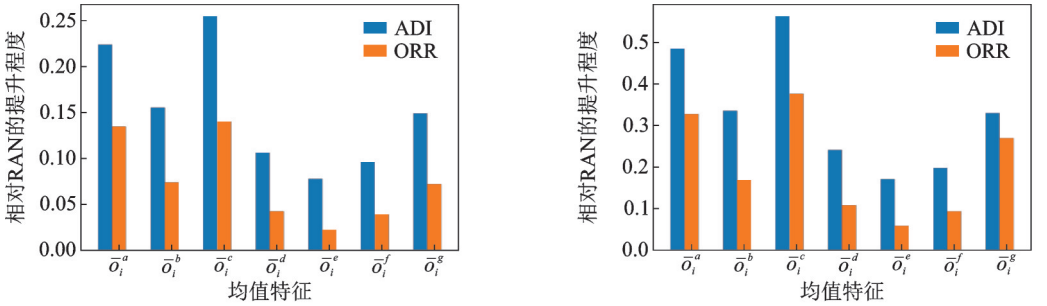


图7　不同均值特征对算法性能的影响

Fig.7　Impacts of mean-field features on the improved algorithms

3.2.4　参数λ对算法的影响

参数λ用于调节全局供需一致性条件对算法施加约束的强度。鉴于该条件对算法影响的前瞻性与全局性,需要在相对较长的时间跨度上,比如:本文采用60 min或6个时间步,评估参数λ对算法的影响;由于订单变化在供需动态平衡中的主动性和随机性,为了细化场景分析,需将进一步分析大时间跨度、不同的订单分布变化强度条件下,参数λ对算法的影响。

首先,通过对不同时刻订单分布变化情况的大时间尺度分析,如图8所示,可以很直观地发现在不同时间段订单在空间上的分布变化程度是不同的。t=26到t=32时段中订单分布变化较大;t=68到t=74时段中订单分布变化较小。

为了量化识别不同订单分布变化强度的应用场景,采用订单分布中心点的漂移量$\Delta_{t_1:t_2}$来大致地估计分布变化强度。以(x,y)表示订单起点所处网格在二维网格地理中的离散坐标,计算t_1时刻所有订单离散坐标的算术均值,即中心点坐标为$(\bar{x}_{t_1},\bar{y}_{t_1})$。同理,计算$t_2$时刻中心点坐标$(\bar{x}_{t_2},\bar{y}_{t_2})$,则订单分布中心点偏移量为

$$\Delta_{t_1:t_2}=\sqrt{\left(\bar{x}_{t_1}-\bar{x}_{t_2}\right)^2+\left(\bar{y}_{t_1}-\bar{y}_{t_2}\right)^2} \tag{25}$$

然后利用漂移量$\Delta_{t_1:t_2}$识别在全天订单分布变化程度最大和最小的10个大跨度时间段,研究在两个不同订单分布变化程度下,参数λ对算法的影响。

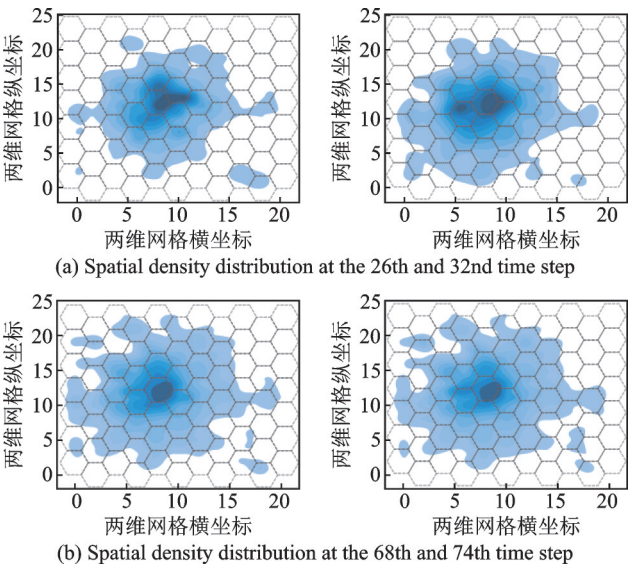
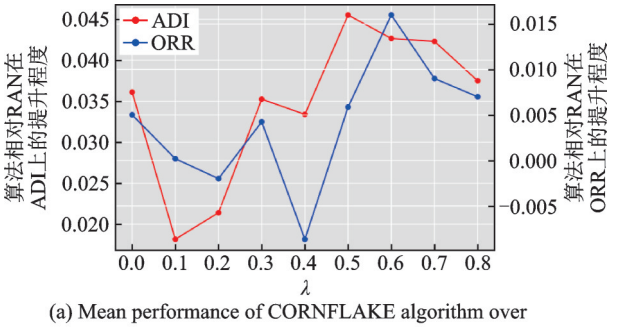


图8　订单需求空间分布变化的比较

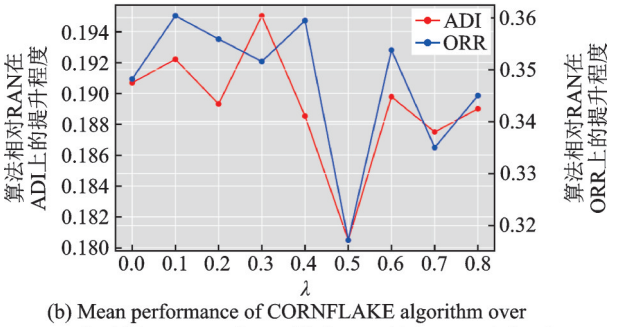
Fig.8　Comparison of variations in the spatial density distribution of order request

参数λ对CORNFLAKE算法性能的影响如图9所示。如图9(a)所示，当订单分布变化程度比较小时，λ为0.6时算法在ORR和ADI上取得较高值，这表明对于订单分布变化平缓、订单热点位置较稳定的场景，算法需要更加关注全局供需分布一致性。如图9(b)所示，当订单分布变化程度比较大时，λ为0.1时算法在ORR和ADI上取得较高值，这表明对于订单分布变化剧烈、订单热点位置不稳定的场景，算法不太需要考虑全局供需一致性的问题。

尽管真实应用场景下订单需求变化随时间可能时强时弱，但以上场景分析的意义在于：搜索得到的λ最优取值，不仅可以作为供需一致性条件的强度指标，而且也可以间接地作为需求热点分布的稳定性指标。比如，在前述总体性能比较中，搜索得到λ的最优取值为0.54，该值接近于订单分布变化平缓场景下λ的最优取值0.6。这表明全天大部分时间订单分布变化强度不大，需求热点分布相对稳定，算法可以更多地将供需一致性纳入考虑。



(a) Mean performance of CORNFLAKE algorithm over the 10 long-range time with the slightest variation in the order distribution versus the variation of parameter λ



(b) Mean performance of CORNFLAKE algorithm over the 10 long-range time with the most intense variation in the order distribution versus the variation of parameter λ

图9 参数λ对CORFLAKE算法性能的影响

Fig.9 Impacts of parameter λ on CORFLAKE algorithm

4 结束语

本文提出了一种全局供需感知的均值场多智能体强化学习订单分配算法CORNFLAKE，该算法首次将多维均值场理论应用在订单分配任务中。实验结果表明，无论全天时段还是高峰期时段，本文提出算法的性能均明显优于现有算法，证明了在多智能体强化学习订单分配算法的基础上，引入向量均值场以加强订单分配的局部协作性以及注入供需分布以提升订单分配全局一致性的改进思路的正确性和有效性。但由于在订单分布变化程度比较大时，CORNFLAKE算法没有考虑到值函数的不稳定性，所以该算法在以下3个方面还需进一步改进：(1)利用正则化等方法，在算法中加入对值函数稳定性的提升，以此来促进算法对全局供需一致性的提升作用；(2)在最大化司机累计收益和订单应答率的同时平衡各个司机之间的收入，提升司机收入的均衡性，确保不会出现过大收入差距；(3)将城市实时交通拥堵、天气状况、节假日等上下文信息纳入考虑范围，通过订单分配算法对网约车的调度作用，缓解城市在出行高峰期、恶劣天气以及节假日的出行难题。

参考文献：

[1] TIRACHINI A. Ride-hailing, travel behaviour and sustainable mobility: An international review[J]. Transportation, 2020, 47(4): 2011-2047.

[2] QIN Zhiwei, ZHU Hongtu, YE Jieping. Reinforcement learning for ridesharing: A survey[C]//Proceedings of 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). New York: IEEE, 2021: 2447-2454.

[3] QIN Zhiwei, TANG Xiaocheng, JIAO Yan, et al. Ride-hailing order dispatching at DIDI via reinforcement learning[J]. INFORMS Journal on Applied Analytics, 2020, 50(5): 272-286.

[4] ZHANG Lingyu, HU Tao, MIN Yue, et al. A taxi order dispatch model based on combinatorial optimization[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2017: 2151-2159.

[5] GAO Guoju, XIAO MingJu, ZHAO Zhenhua. Optimal multi-taxi dispatch for mobile taxi-hailing systems[C]//Proceedings of the International Conference on Parallel Processing. Los Alamitos, CA: IEEE, 2016: 294-303.

[6] LIU Yifang, SKINNER W, XIANG Chongyuan. Globally-optimized realtime supply-demand matching in on-demand ridesharing[C]//Proceedings of The World Wide Web Conference. New York: ACM, 2019: 3034-3040.

[7] XU Zhe, LI Zhixin, GUAN Qingwen, et al. Large-scale order dispatch in on-demand ride-hailing platforms: A learning and planning approach[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2018: 905-913.

[8] HOLLER J, VUORIO R, QIN Zhiwei, et al. Deep reinforcement learning for multi-driver vehicle dispatching and repositioning problem[C]//Proceedings of 2019 IEEE International Conference on Data Mining (ICDM). New York: IEEE, 2019: 1090-1095.

[9] WANG Zhaodong, QIN Zhiwei, TANG Xiaocheng, et al. Deep reinforcement learning with knowledge transfer for online rides order dispatching[C]//Proceedings of 2018 IEEE International Conference on Data Mining (ICDM). New York: IEEE, 2018: 617-626.

[10] WANG Yansheng, TONG Yongxin, LONG Cheng, et al. Adaptive dynamic bipartite graph matching: A reinforcement learning approach[C]//Proceedings of 2019 IEEE 35th International Conference on Data Engineering (ICDE). New York: IEEE, 2019: 1478-1489.

[11] JINDAL I, QIN Zhiwei T, CHEN Xuewen, et al. Optimizing taxi carpool policies via reinforcement learning and spatio-temporal mining[C]//Proceedings of 2018 IEEE International Conference on Big Data (Big Data). Piscataway, NJ: IEEE, 2018: 1417-1426.

[12] LYNCH G S. Single point of failure: The 10 essential laws of supply chain risk management[M]. New Jersey: John Wiley and Sons. 2009.

[13] ZHANG Kaiqing, YANG Zhuoran, BAŞAR T. Multi-agent reinforcement learning: A selective overview of theories and algorithms[J]. Handbook of Reinforcement Learning and Control, 2021, 325(1): 321-384.

[14] LI Minne, QIN Zhiwei, JIAO Yan, et al. Efficient ridesharing order dispatching with mean field multi-agent reinforcement learning[C]//Proceedings of the World Wide Web Conference. New York: ACM,2019: 983-994.

[15] YANG Yaodong, LUO Rui, LI Minne, et al. Mean field multi-agent reinforcement learning[C]//Proceedings of Machine Learning Research. San Diego, CA: PMLR, 2018: 5571-5580.

[16] ZHOU Ming, JIN Jiarui, ZHANG Weinan, et al. Multi-agent reinforcement learning for order-dispatching via order-vehicle distribution matching[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. New York: ACM, 2019: 2645-2653.

[17] KULLBACK S, LEIBLER R A. On information and sufficiency[J]. The Annals of Mathematical Statistics, 1951, 22(1): 79-86.

[18] MIAO Fei, HAN Shuo, LIN Shan, et al. Taxi dispatch with real-time sensing data in metropolitan areas: A receding horizon control approach[J]. IEEE Transactions on Automation Science and Engineering, 2016, 13(2): 463-478.

[19] LEE D H, WANG Hao, CHEU R L, et al. Taxi dispatch system based on current demands and real-time traffic conditions[J]. Transportation Research Record, 2004, 1882(1): 193-200.

[20] MUNKRES J. Algorithms for the assignment and transportation problems[J]. Journal of the Society for Industrial and Applied Mathematics, 1957, 5(1): 32-38.

[21] FAN Jianqing, WANG Zhaoran, XIE Yuchen, et al. A theoretical analysis of deep Q-learning[C]//Proceedings of Learning for Dynamics and Control.[S.l.]: PMLR, 2020: 486-489.

[22] DE LIMA O, SHAH H, CHU T S, et al. Efficient ridesharing dispatch using multi-agent reinforcement learning[EB/OL]. (2021-03-15)[2022-04-19]. https://arxiv.org/abs/2006.10897.

[23] RASHID T, SAMVELYAN M, SCHROEDER C, et al. QMIX: Monotonic value function factorisation for deep multi-agent

reinforcement learning[C]//Proceedings of International Conference on Machine Learning. San Diego, CA: PMLR, 2018: 4295-4304.

[24] SPAAN M T J. Partially observable Markov decision processes[M]. Berlin, Heidelberg: Springer, 2012: 387-414.

[25] 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.

[26] LIN Kaixiang, ZHAO Renyu, XU Zhe, et al. Efficient large-scale fleet management via multi-agent deep reinforcement learning[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2018: 1774-1783.

[27] Didi Co. Data source: Didi Chuxing GAIA initiative[EB/OL]. (2020-07-13)[2022-04-19]. https://gaia.didichuxing.com.

[28] AKIBA T, SANO S, YANASE T, et al. Optuna: A next-generation hyperparameter optimization framework[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2019: 2623-263.

作者简介:

宋旺(1997-),男,硕士研究生,研究方向:强化学习,E-mail:791382245@qq.com。

胡祥(1976-),**通信作者**,男,博士,研究方向:人工智能、时空大数据与城市计算等,E-mail:colin_fox@ncepu.edu.cn。

张玉辉(1999-),女,硕士研究生,研究方向:强化学习、电网优化调度等,E-mail:zhangyuhui18@qq.com。

卫文江(1999-),男,硕士研究生,研究方向:强化学习、电网优化调度等,E-mail:wwj19990725@163.com。

周雅诗(1999-),女,硕士研究生,研究方向:智能交通,E-mail:2395960139@qq.com。

康傲(1999-),男,硕士研究生,研究方向:智能交通和强化学习,E-mail:kangao4455@163.com。

(编辑:刘彦东)