基于深度强化学习的外卖即时配送实时优化

陈彦如，刘珂良，冉茂亮

（西南交通大学经济管理学院，四川 成都 610031）

摘　要：为了应对外卖配送任务在用餐高峰期运力紧张、订单延迟送达率高的挑战，提出一种基于深度强化学习（DRL）的外卖即时配送实时优化策略，以提升外卖平台长期客户服务水平。首先，充分考虑外卖配送中备餐时间、取送顺序、时间窗等约束，以最大化期望平均客户服务水平为目标，建立考虑随机需求的外卖即时配送问题的马尔可夫决策过程（MDP）模型；其次，设计一种结合近似策略优化（PPO）算法和插入启发式（IH）算法的外卖即时配送优化策略 PPO-IH。PPO-IH 使用融合注意力机制的选择策略网络对订单-骑手进行匹配，通过 PPO 算法对网络进行训练，并使用插入启发式算法更新骑手路径。最后，通过与贪婪策略（Greedy）、最小差值策略、分配启发式以及两种深度强化学习算法进行对比实验，结果表明。PPO-IH 分别在 71.5%、95.5%、87.5%、79.5%与 70.0%时段数据中表现更优，同时平均客户服务水平更高，平均每单配送时间更短、延迟送达率更低。此外，PPO-IH 在不同骑手数、不同订单密度以及不同订单时间窗场景下具有一定的有效性和泛化性。

关键词：外卖配送；实时优化；深度强化学习；马尔可夫决策过程；近似策略优化；注意力机制

中图分类号：TP181　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0069559

Real-time Optimization of Instant Meal Delivery Based on Deep Reinforcement Learning

CHEN Yanru, LIU Keliang, RAN Maoliang

(School of Economics and Management, Southwest Jiaotong University, Chengdu 610031, Sichuan, China)

【Abstract】To address the challenges of tight capacity and high delayed rate of meal delivery tasks during peak dining period, a real-time optimization policy based on Deep Reinforcement Learning (DRL) for instant meal delivery is proposed to improve the long-term customer service level of platforms. First, considering the constraints of meal preparation time, pickup and delivery sequence, and time window in meal delivery, the instant meal delivery problem with stochastic requests is modeled as a Markov Decision Process (MDP) to maximize the expected average customer service level. Second, the Proximity Policy Optimization (PPO) algorithm is combined with the Insertion Heuristic (IH) algorithm to design an instant meal delivery optimization policy, PPO-IH. A policy network with an integrated attention mechanism is employed by PPO-IH for matching orders to couriers, and the network is trained by the PPO algorithm. The courier routes are updated with an IH algorithm. Finally, through comparative experiments with the Greedy, minimum difference strategy, allocation heuristic, and two deep reinforcement learning algorithms, PPO-IH is shown to perform better in 71.5%, 95.5%, 87.5%, 79.5%, and 70.0% days with the given data, respectively. Additionally, PPO-IH achieves a higher average level of customer service, shorter average delivery times per order, and a lower rate of delayed deliveries. Furthermore, PPO-IH demonstrates certain effectiveness and generalization under different rider numbers, order densities, and order time window scenarios.

【Key words】meal delivery; real-time optimization; Deep Reinforcement Learning (DRL); Markov Decision Process (MDP); Proximal Policy Optimazation (PPO); attention mechanism

0　引言

近年来，外卖行业日趋成熟，用户规模日益庞大，客户对外卖配送的服务体验要求也越来越高。客户希望外卖订单下单后，平台能够尽快响应，并且尽早送达。因此，外卖配送具有极高的时效性要求。调查显示，约 71%的外卖订单在午餐与晚餐时段下达[1]，导致外卖配送在用餐时段上存在配送压力大、延迟订单多等问题与挑战。因此，设计智能派单策略，快速响应订单，合理规划不同时段有限运力，从而减少延迟订单量，提升整体客户服务体验，对提高外卖即时配送服务水平有着重要意义。

收稿日期：2024-03-13　　**修回日期：**2024-05-19

基金项目：国家自然科学基金（72371206）。

通信作者 E-mail：chenyanru@swjtu.cn

动态需求下的外卖即时配送问题具有订单动态到达、多车辆、一对一取送、强时效性及时间窗限制等特征，在本质上是随机动态取送货问题(SDPDP)[2]。随机动态取送货问题在现实生活中有着丰富的应用，例如网约车、外卖送餐、同城急送等场景。传统 SDPDP 研究在已知信息的基础上，根据即时成本或收益对路径进行优化，当观察到新的信息后，对路径进行重新创建或再优化[3-4]。然而，再优化的方法需要确定合适的再优化时点，算法运行时间长，优化结果易受新的随机信息影响，同时无法考虑未来信息对当前决策的影响，难以实现长期最优决策。因此，强化学习或深度强化学习方法开始应用到外卖配送问题中[5-6]，以期在当前决策中整合未来信息。然而，现有的深度强化学习解决外卖即时配送问题的研究存在一定的局限性。如文献[5-6]针对单个骑手训练深度 Q 网络(DQN)以实现骑手调度，但决策时未考虑系统内骑手间的相互影响，同时假设骑手容量为 1，无法同时携带多个餐品，不考虑备餐时间，对外卖配送问题进行简化，场景考虑简单。为考虑系统内所有骑手信息，文献[7]将多个骑手的路径信息作为双重深度 Q 网络(DDQN)的输入，将其表示为一个固定长度的向量。然而，骑手路径长度随时间与工作情况动态变化，部分时段易出现特征稀疏等问题，同时，其输入特征长度与价值网络输出层维度需固定等问题导致训练后的强化学习策略无法适应骑手数量不同或变化的情况，泛化性差。为解决以上问题，本文参考 seq2seq 模型的编码器结构设计了骑手信息编码器，可适应不同数量骑手的输入，同时在网络结构中引入注意力机制，使得在决策时能够充分考虑系统内所有骑手间的相互影响，从全局角度实现长期更优决策。

本文的主要贡献如下：

1)问题场景考虑更为完备。本文基于外卖配送时间，同时考虑多骑手、备餐时间、取送顺序、软时间窗等约束，对外卖即时配送问题进行详细描述，并以最大化平均客户服务体验为目标，建立基于路径的马尔可夫决策过程模型。

2)设计针对路径问题的强化学习方法。本文设计了结合近似策略优化(PPO)[8]算法与启发式算法的深度强化学习外卖即时配送优化策略，该策略结合强化学习的顺序决策能力与启发式方法的路径规划能力，有效地缩减了强化学习的动作空间，并具有一定的长期优化能力。

3)从保证策略泛化性与系统全局决策的角度设计网络结构。本文设计一种基于注意力机制的骑手选择策略网络，其编码器结构允许不同数量骑手的状态信息输入，可以适应骑手数量不同或变化的环境，提高了策略的泛化性。此外，注意力机制的加入使策略运行时能够捕捉系统内不同骑手间的相互影响，根据全局状态，实现全周期更优决策。

1 相关工作

1.1 外卖配送路径优化

目前针对外卖配送路径优化的研究较为丰富，此类研究在已知的配送需求基础上，根据即时成本或收益对路径进行优化。针对随机需求的外卖配送问题，多数研究设计滚动时域的方法，将动态问题分解为多个静态子问题，再逐一优化。李桃迎等[4]考虑时间惩罚成本，设计“商家-客户”配对策略，引入 k-means 聚类算法对“商家-客户”进行聚类，在同一簇内设计遗传算法得到路径优化方案。张玉州等[3]设计了一种基于滚动时域控制的外卖配送问题模型，将生成的外卖订单按多个时间窗口划分，并使用最近邻域算法求解。余海燕等[9]提出了具有硬时间窗要求的生鲜配送问题，考虑合单配送的思想，设计了滚动时域延迟配送算法对问题进行求解。冯爱兰等[10]为解决“骑手困境”，研究了平台派单与骑手自主抢单相结合的外卖配送模式，并设计了自适应邻域搜索算法实现平台派单与骑手路径优化。然而，外卖配送任务时效性强，难以确定合适的滚动时域时间间隔，启发式算法的优化结果也会因新订单的出现受到干扰。因此，余海燕等[11]以订单到达为决策点，分别建立了以最小化平均每单配送距离和平均每单配送时间为目标的实时订单分配与路径优化模型，并设计了贪婪策略与最小差值策略求解两个模型。

1.2 考虑随机信息影响的外卖即时配送优化

以上研究均根据即时成本或收益对问题进行优化，未考虑未来随机信息对当前决策的影响，因此，一些学者尝试在决策中整合未来信息。STEEVER 等[12]提出一种新型的外卖商业框架模式，并建立了该模式的静态模型。针对动态问题，在插入启发式算法的基础上，提出一种基于骑手竞价的启发式方法，允许接受一些能够改善某些全局指标的次优分配方案，实验结果证明，相较于贪婪策略，该方法能获得更高的长期收益。ULMER 等[13]研究备餐时间随机的动态外卖即时配送问题，建立了马尔可夫决策过程(MDP)模型并设计了一种预期订单分配策略，该策略通过推迟非紧急订单的分配保证骑手分配空间的灵活，并利用成本函数近似方法影响骑

手路线规划，减少随机备餐时间对配送任务的影响。CHEN 等[14]提出了一种基于模仿学习的迭代匹配算法解决外卖配送问题，该算法包括迭代匹配启发式、机器学习模型与提供静态问题高质量解决方案的专家算法 3 个组件。在离线优化阶段，机器学习模型从专家算法提供的高质量解决方案中挖掘知识，在在线操作阶段，迭代匹配算法嵌入训练后的机器学习模型以完成订单-骑手的高效匹配。随着深度强化学习的发展，一些学者们希望利用深度强化学习的顺序决策能力，在动态的外卖即时配送问题中获得更高的长期收益。BOZANT 等[5]和 JAHANSHAHI 等[6]的研究均将外卖即时配送问题建模为 MDP，并训练 DQN 实现对骑手的调度与控制，然而，两个研究均对问题进行简化，求解难度大大降低，同时在深度强化学习算法训练时进行简化，仅将单个骑手视为智能体，将训练后的策略在测试阶段部署在系统内所有骑手上，分别对骑手进行控制。ZOU 等[7]提出了一种基于 DDQN 的强化学习调度策略，该策略使用神经网络为订单指派具体骑手，并在外卖配送环境模拟器上验证了策略的有效性，然而，所提出的策略泛化性差，无法在骑手数量不同的环境中运行。WANG 等[15]研究了考虑骑手行为的外卖配送问题，提出了一种基于订单推荐的在线深度强化学习框架，该框架包括基于 Actor-Critic 的订单推荐网络与骑手行为预测网络，骑手行为预测网络对骑手是否从订单推荐网络提供的订单集中抓取进行预测。进一步地，WANG 等[16]考虑骑手抢单时可能出现的冲突情况，设计了一种基于 Actor-Critic 的长短期记忆网络为单个骑手输出推荐订单列表，并提出了 3 种骑手排序规则应对骑手抢单冲突，实验结果证明，所提出的框架能够有效地提高抓单数量，并减少骑手抢单冲突。

综上所述，外卖即时配送问题存在时效性强、订单动态度高、时段性高峰等特点，求解难度大。现有研究在问题求解方法的设计上大多根据现有信息获得当前阶段解方案，缺乏对未来信息的考量，难以实现长期更优。深度强化学习结合深度学习强大的特征提取能力与强化学习的决策优化能力，通过持续采样和反馈学习，能够逐步地适应环境的动态变化，在解决复杂的决策问题方面展现了强大的潜力[17-18]。鉴于此，本文使用深度强化学习对外卖即时配送问题进行优化，寻求目标函数的长期更优。目前使用深度强化学习进行外卖即时配送实时优化的研究相对较少，并且存在问题特征考虑较少、求解规模小、未考虑系统内骑手间关联、泛化性差等不足。一些使用强化学习或深度强化学习方法研究网约车调度问题[19-20]、拼车问题[21]以及同日送达问题[22]的文献对本研究也有着重要的参考价值。对此，本文充分考虑外卖即时配送问题的动态性、时效性和时段高峰等特点，建立以最大化客户服务体验为目标的马尔可夫决策过程模型，同时设计了一种引入注意力机制的深度强化学习方法，通过集成的注意力机制捕捉骑手间的动态影响，实现全周期更优的决策。该框架使用近似策略优化算法对网络进行训练，实现对外卖即时配送问题的高效求解。

2 外卖即时配送的马尔可夫决策模型

外卖即时配送问题是一类顺序决策问题(SDP)，基于文献[23]，本文对该问题构建 MDP 模型。

2.1 问题描述

考虑某外卖平台某站点管理的区域 $G=(N,A)$，其中 $N=N_{\text{init}} \cup I \cup C$，表示骑手配送过程中可能经过的节点集合，包括骑手起点集合 N_{init}、餐厅节点集合 I 和客户节点集合 C。A 为 N 中所有节点两两间组成的弧的集合，每条弧 (n_i,n_j) 对应一个固定的旅行时间 $d(n_i,n_j)$。该站点有一组骑手 $V=\{1,2,\cdots,m\}$，在一段有限的时域 $T=[0,t_{\max}]$ 服务一组动态到达的订单 O，这些订单对应的餐厅来自该区域已知的一组餐厅 I，下单的用户来自该区域平台客户集合 C。

骑手 v 需从其起点 $n_{\text{init}}^v \in N$ 出发，开启配送任务。骑手在餐厅处取餐或客户处交付餐品均有一个固定的服务时间 $\bar{t}$。骑手交付当前分配的所有订单后，会在最后完成交付的节点等待，直到新的订单分配给骑手。当骑手正在前往某个节点 n 的过程中，如果既定路线计划发生更新，骑手会继续前往当前目标节点 n，而后再执行新的路线计划。在外卖即时配送任务中，骑手不断取餐与送餐，在同一时刻，骑手携带的餐品数量往往非常有限，因此本文忽略了订单的重量，同时不考虑骑手的容量限制。

一个已知的外卖订单 o 包含以下信息：订单对应餐厅 $r_o \in I$，下单时间 t_o，最早取餐时间 t_o^1，最晚柔性送达时间 t_o^2，下单用户 $c_o \in C$。骑手若在最早取餐时间 t_o^1 前到达餐厅，则需要等待至备餐完成才能取餐。骑手如果晚于最晚柔性送达时间 t_o^2 到达客户处交付，则会受到惩罚。本文的优化目标是最大化平均客户服务体验(R)，计算公式如式(1)所示：

$$R=\frac{1}{|O|}\sum_{o\in O}[t_o^2-a_o^{\text{real}}-(p-1)\times\max(0,a_o^{\text{real}}-t_o^2)] \tag{1}$$

式中：a_o^{real} 表示订单 o 的实际交付时间；p 为晚于最晚柔性送达时间送达受到的单位惩罚，p 的取值通常大于 1，表示在外卖配送过程中，客户对外卖订单的延迟送达接受度低。

骑手的路线计划集合 $\theta=\{\theta_1,\theta_2,\cdots,\theta_m\}$ 包含了每位骑手的路线计划信息。骑手的路线计划包含一系列的已访问节点、计划访问节点与时间信息，包括到达节点、服务节点、离开节点的时间。骑手 v 在某个时刻的路线计划可表示为：

$$\theta_v=((n_{\text{init}}^v,d(n_{\text{init}}^v)),(n_1^v,a(n_1^v),s(n_1^v),d(n_1^v)),\cdots,(n_i^v,a(n_i^v),s(n_i^v),d(n_i^v)),\cdots,(n_h^v,a(n_h^v),s(n_h^v)) \tag{2}$$

式中：n_{init}^v 表示骑手 v 的出发地；$d(n_{\text{init}}^v)$ 表示骑手出发的时刻，通常为骑手接到第一个订单的时刻，路线计划最后一个节点的时间信息不包括出发时间，只有在路线计划发生更新，新的节点被插入至该节点后时，才会更新其出发时间；n_i^v 表示骑手路线计划中的第 i 个访问节点；$a(n_i^v)$ 表示骑手预计到达该节点的时刻；$s(n_i^v)$ 表示骑手对该节点进行服务的时刻；$d(n_i^v)$ 表示骑手离开该节点的时刻。设当前时刻为 t，$a(n_i^v)\leqslant t$ 的路线为已服务的历史路线计划，$a(n_i^v)$ 为订单实际送达时间 a_o^{real}，$a(n_i^v)>t$ 的路线表示暂定的路线计划，该部分会被实时更新。

基于式(1)可得单个骑手路线计划的预期平均客户体验，如式(3)所示：

$$r(\theta_v)=\sum_{o_i\in O_v}[t_{o_i}^2-a_{o_i}^v-(p-1)\times\max(0,a_{o_i}^v-t_{o_i}^2)] \tag{3}$$

式中：O_v 为骑手 v 路线计划中的订单集合；$t_{o_i}^2$ 为订单 o_i 的最晚柔性送达时间；$a_{o_i}^v$ 为预计访问订单 o_i 的客户点的时间。本文用一个例子对外卖即时配送优化问题进行说明，如图 1 所示，在 $t=5$ 时刻，骑手接到第一个订单 $o_1=$(I1,C1)，骑手从起点出发，此刻的计划路线可以表示为：

$$\theta_v=((n_{\text{init}}^v,5),(\text{I1},15,20,22),(\text{C1},35,35)) \tag{4}$$

在 $t=10$ 时刻，骑手接到新的订单 $o_2=$(I2,C2)，新的路线计划可表示为：

$$\theta_v=((n_{\text{init}}^v,5),(\text{I1},15,20,22),(\text{I2},30,30,32),(\text{C2},40,40,42),(\text{C1},50,50)) \tag{5}$$

图 1 中黑色实线箭头表示已完成配送的路径，虚线箭头表示因新订单插入而重新制定的配送路线，此时需更新各节点时间信息。

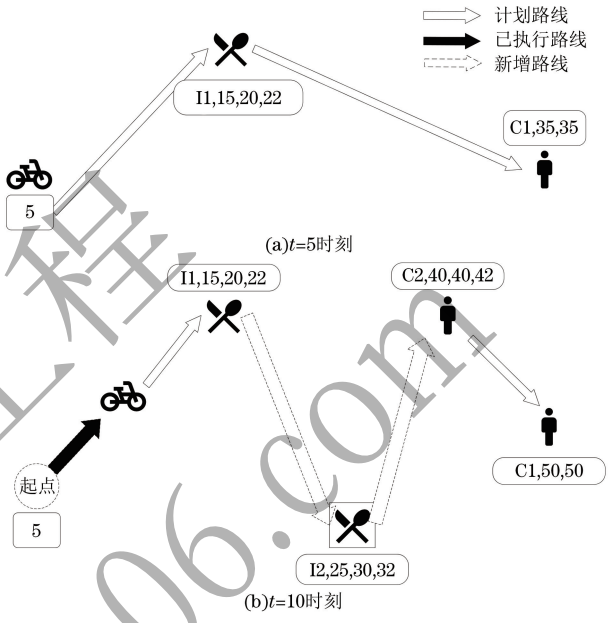

图 1 外卖配送问题示意图

Fig. 1 Schematic diagram of meal delivery problem

综上所述，本文研究的外卖即时配送问题具有需求随机、时效性、时段性高峰等多个特点，同时考虑取送约束、备餐时间、软时间窗等约束，求解较复杂。而现有外卖配送研究大多设计启发式算法根据已知需求与即时收益或成本进行路径优化，忽略了随机信息对优化的影响。因此，本文设计深度强化学习算法，对考虑不确定需求的外卖即时配送问题进行实时优化。

2.2 基于路径的马尔可夫决策过程模型

传统马尔可夫决策过程模型在对动态信息的捕捉与路线规划的表示方面不够理想。基于路径的马尔可夫决策过程模型扩展了传统马尔可夫决策过程模型的元素，以适应路径问题[21]。因此，本节针对外卖即时配送问题，构建了基于路径的马尔可夫决策过程模型。路径信息的加入会直接决定状态空间与动作空间，同时影响奖励计算与状态转移。模型各组成部分如下：

1)决策点。

本文的决策点为外卖配送请求发生的时刻。将第 k 个订单记为 o_k，其下单时间为 t_{o_k}，则第 k 个决策点的决策时间为 $t_k=t_{o_k}$。

2)状态。

状态包含了在决策点进行决策所需的必要信息。在本问题中，状态 S_k 包含了第 k 个订单的信息、所有骑手的信息及其路线计划的信息。因为骑

手的相关信息可以由路线计划获得，所以可以在状态空间中省略。状态的组成部分如下：

t_k：决策发生的时刻。

o_k：新的外卖配送订单请求及其详细信息，包括下单时间、餐厅位置、客户位置、备餐时间、最晚送达时间等。每个订单 o_k 可以表示为 $(t_{o_k},r_{o_k},c_{o_k},t^p_{o_k},t^d_{o_k})$。

O_k：决策点 k 已被分配的订单集合 $O_k=\{O^1_k,O^2_k,\cdots,O^m_k\}$。其中 O^v_k 表示骑手 v 在第 k 个决策点已被分配的订单集合，包括已交付订单、未取货订单、已取货未交付订单。

θ_k：决策点 k 的路线计划集合，$\theta_k=\{\theta^1_k,\theta^2_k,\cdots,\theta^m_k\}$，其中 θ^v_k 是骑手 v 在决策点的路线计划。在骑手未接收到任何订单的情况下，其路线计划信息仅仅包括出发点。

总的来说，本文将状态信息表示为一个四元组 $s_k=(t_k,o_k,O_k,\theta_k)$。在骑手 v 未接收到任何订单的情况下，O^v_k 为空，其路线计划信息也仅仅包括其出发点，出发时间也会随着第一个订单的接收而确定。

3)动作。

在每个决策点，外卖平台根据策略的指导从动作空间中选择一个动作。在外卖即时配送问题中，决策点的动作应该包括待分配订单-骑手匹配方案以及骑手的更新后的路线计划。本文将决策点 k 采取的动作定义为一个二元组 $x_k=(a_k,\theta^x_k)$。其中，θ^x_k 表示订单分配后获得的更新后的骑手路线计划集合，a_k 表示将订单分配给某骑手，即：

$$a_k=v,\ v\in V \tag{6}$$

4)奖励。

决策的奖励为路线计划更新前后引起的目标函数变化，目标函数的计算见式(3)。因此，模型的奖励设计为：

$$R(s_t,x_k)=r(\theta^x_k)-r(\theta_k) \tag{7}$$

5)随机信息与状态转移。

对外卖即时配送问题，状态转移分为两个部分：确定性转移与随机性转移。确定性转移为决策前状态 s_k 到决策后状态 s^x_k，如式(8)所示：

$$s^x_k=S^x(s_k,x_k) \tag{8}$$

在本问题中表现为订单分配并更新对应骑手的路线计划。随机性转移由随机信息 ω_{k+1} 决定，如式(9)所示：

$$s_{k+1}=S^\Omega(s^x_x,\omega_{k+1}) \tag{9}$$

在外卖即时配送问题中，ω_{k+1} 不仅确定了决策点 $k+1$ 待分配的订单的相关信息，也确定了下一个决策点的时刻 t_{k+1}，这影响着路线计划执行情况 θ^v_{k+1} 的更新与已分配订单交付情况信息 O^v_{k+1}。

综上所述，状态转移的更新函数可以总结为：

$$s_{k+1}=S^\Omega(S^x(s_k,x_k),\omega_{k+1}) \tag{10}$$

以上定义了在外卖即时配送问题中，基于路径的马尔可夫决策过程模型的各个组成部分。传统车辆路径问题的解是一组表示解方案的解向量，而在顺序决策问题中，解是一个策略 $\pi\in\Pi$，负责在各个决策点根据当前状态选择动作，外卖即时配送问题目标是寻找一个最优策略 π^*，以最大化期望总奖励：

$$\pi^*=\underset{\pi\in\Pi}{\operatorname{argmax}}E\left[\sum_{k=0}^{K}R(S_k,X^\pi(S_k)\mid S_0)\right] \tag{11}$$

式中：$X^\pi(S_k)$ 为在策略 π 的指导下，根据状态 S_k 选择的动作。

3 基于深度强化学习的外卖即时配送优化算法

3.1 插入启发式算法

如第 2.2 节对外卖即时配送问题的马尔可夫决策过程模型定义所示，提出的策略不仅需要确定新订单由哪一位骑手服务，还需要将订单规划至选定骑手的路线计划中，此外，在随机动态的外卖即时配送问题中，常要求决策迅速且准确。因此，本文考虑了一个插入启发式(IH)算法对骑手路径进行规划。本文设计的插入启发式算法类似于文献[11]提出的最小差值策略，但由于路线计划的定义有所不同，因此算法步骤也存在一定的差别。插入启发式算法如算法 1 所示。

算法 1 插入启发式算法

输入 骑手路线计划 θ^v_k 分配订单 o 的取餐点 r_o 与送餐点 d_o，时间 T，惩罚项系数 p

输出 最优路线计划 $\theta^{v,\text{best}}_k$ 与 best_score

1. 初始化 best_score=$-\infty$，最优路线计划 $\theta^{v,\text{best}}_k=\theta^v_k$；

2. 依据式(3)计算插入前路线计划的目标函数 score0；

3. 根据时间 T 确认路线计划的实施情况，确认取餐点 r_o 可插入位置集合 PSET；

4. for each position1$\in$PSET do：

5. 根据创建路线计划副本 $\theta^{v,1}_k$，将取餐点 r_o 根据 position1 插入至 $\theta^{v,1}_k$，依照先取后送约束，确认送餐点 d_o 的可能插入位置集合 DSET；

6. for each position2$\in$DSET do：

7. 根据 $\theta^{v,1}_k$ 创建路线计划副本 $\theta^{v,2}_k$，将送餐点 d_o 插入至 $\theta^{v,2}_k$。根据式(3)计算插入后路线计划目标函数 score1，根据式(7)计算插入前后目标函数变化值 Δscore；

8. if Δscore>best_score:
9. best_score←Δscore;
10. $\theta_k^{v,best} \leftarrow \theta_k^{v,2}$;
11. end
12. end

遍历骑手集合 V，依次运行插入启发式算法，选择具有最优 best_score 的骑手进行插入，即为贪婪策略。本文的深度强化学习方法将使用策略网络对订单-骑手进行匹配，并使用插入启发式算法对骑手路线进行规划。

3.2 状态聚合与特征设计

状态聚合通过将状态空间划分为若干个区域或“聚合体”来减少维度，使问题更易于处理。此外，状态聚合还能够减少噪声，提高稳定性与收敛性，是常见的解决状态空间高维的手段[22]。因此，本节将对第 2.2 节提出的马尔可夫决策过程模型的状态空间进行聚合，以设计网络的输入特征。本文将策略网络输入特征分为骑手特征与订单信息特征。

3.2.1 骑手特征

骑手特征主要从第 2.2 节定义马尔可夫决策过程模型中的 O_k^v 与 θ_k^v 中进行提取。提取的骑手特征具体如下：

$$s_{\text{courier}}=\{x_v,y_v,n_v,l_v,dw_v,dr_v,wt_v;\ v\in V\} \tag{12}$$

式中：x_v 与 y_v 分别表示骑手 v 当前所在位置信息；n_v 表示骑手 v 目前需要访问的节点个数；l_v 表示骑手 v 目前行程的预计总延迟送达时间；dw_v 表示骑手 v 使用插入启发式算法将待分配的订单插入到最佳位置，引起的因餐厅出餐时间导致的行程总等待时间变化值；dr_v 表示骑手 v 若使用插入启发式算法将待分配的订单插入到最佳位置，引起的路线计划目标函数值变化；wt_v 表示骑手 v 使用插入启发式算法将待分配的订单插入到最佳位置后，路线计划变化引起的响应餐厅订单时间的变化，这个概念来自文献[24]提出的公平度(Equity，E_{Equity})。

本文对这个概念进行改进，考虑外卖即时配送问题中，路线计划距当前时刻 t 更远的计划部分更容易出现订单节点的插入，因此对更远的节点的响应能力计算赋予更高的权重。参考第 1.1 节对路线计划的表达，公平度计算公式为：

$$E_{\text{Equity}}=\frac{\sum_{n\in\theta_{\text{valid}}^v}(1+\alpha)^{(a_n-t)}\left[\sum_{r\in I}d(r,n)\right]}{|\theta_{\text{valid}}^v|} \tag{13}$$

式中：θ_{valid}^v 为路线计划根据约束条件定义可更新的部分；α 是表征对未来节点响应能力重视程度的参数，根据实验测算，取值为 0.01；(a_n-t) 为节点 n 计划到达时间与当前时刻的时间差，越高的时间差会给予其计算更高的权重；$\sum_{r\in I}d(r,n)$ 为节点 n 到所有餐厅的旅行时间总和。

3.2.2 环境特征

令 s_{context} 为环境特征，其包含某决策时刻环境状态信息与订单状态信息，具体如式(14)所示：

$$s_{\text{context}}=\{t_k,r_x^k,r_y^k,r_h^k,c_x^k,c_y^k,u_k\} \tag{14}$$

式中：t_k 表示第 k 个决策点对应的时刻；r_x^k、r_y^k 表示订单 o_k 的餐厅位置的横纵坐标；r_h^k 表示订单 o_k 对应的餐厅热度信息，热度越高的餐厅有更多的客户下单；c_x^k、c_y^k 表示订单 o_k 对应的客户的位置的横纵坐标；u_k 表示订单 o_k 的紧急程度，该特征用订单餐厅到客户点的旅行时间 $d(r_k,c_k)$ 来表示。

所有的骑手特征与环境特征在输入至网络前均会通过 z-score 标准化。

3.3 外卖即时配送优化策略

3.3.1 动作空间削减与近似策略优化算法

对于外卖即时配送优化问题的序贯决策问题，理论上可以通过贝尔曼最优方程推导求出其解。然而，外卖即时配送问题状态空间过大，容易陷入“维度灾难”的问题。一方面，难以枚举外卖即时配送问题的所有可能状态，另一方面，外卖即时配送问题的决策不仅涉及订单-骑手匹配，也需要对骑手路线进行规划与更新，其动作空间庞大，并随着服务的订单增加而呈爆炸式增长，为网络输出层设计带来难题。强化学习算法擅长对动作价值进行评估，以选择长期更优的动作，而非对庞大的动作空间进行搜索。因此，本文提出了结合近似策略优化算法与插入启发式算法的深度强化学习策略 PPO-IH，深度强化学习算法对订单-骑手进行匹配，使用第 3.1 节提出的插入启发式算法对骑手路径进行规划，以在随机动态环境下获得长期更优奖励。

PPO 算法是一种基于 Actor-Critic 架构的深度强化学习算法，Actor 是一个策略网络，根据当前状态 S_k 输出动作分布 $\pi_\theta(\cdot|S_k)$，其中 θ 表示 Actor 网络的参数。Critic 网络是一个基于价值的网络，负责评估该状态下的预期回报 $V_\varphi(S_k)$，其中 φ 是 Critic 网络的参数。PPO 通过 Critic 网络引导 Actor 网络的更新，同时 Critic 网络也会根据采样数据学习与更新，提高价值评估的准确性。图 2 展示了 PPO 算法在外卖即时配送问题中的工作流程，PPO 算法的更新流程如算法 2 所示。通过查阅文献与反复实验，表 1 确定了 PPO 算法的部分参数。

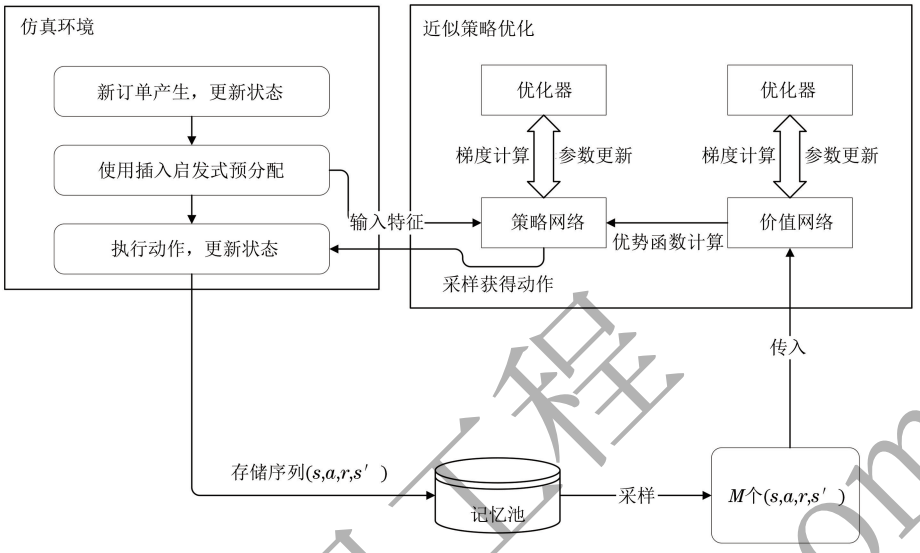

图 2 PPO 算法流程

Fig. 2 PPO algorithm procedure

算法 2 PPO 算法

输入 表 1 中关于算法的超参数

输出 训练后的策略网络 π_θ

1. 初始化策略网络参数 θ_0、价值网络参数 φ_0、记忆池 β；

2. for i=1 to Q do：

3. while not 达到终止状态：

4. 策略网络根据状态 S 采样动作，环境根据动作与随机信息生成即时奖励 r 和下个决策点状态 s′，将序列(s,a,r,s′)存入记忆池 β；

5. if 满足更新条件：

6. 从记忆池 β 中采样 m 个序列(s,a,r,s′)组成集合 D_k，使用价值网络与广义优势估计(GAE)计算各序列的优势函数 A_t；通过 Adam 优化器最大化 PPO 的目标函数：

$$\theta_{k+1}=\arg\max_{\theta}\frac{1}{|D_k|}\cdot\sum_{\tau\in D_k}\frac{1}{T}\sum_{t=0}^{T}\min\left(\frac{\pi_\theta(a_t\mid s_t)}{\pi_{\theta_{old}}(a_t\mid s_t)}A_t,\mathrm{clip}(r_t(\theta),1-\varepsilon,1+\varepsilon)A_t\right)$$

8. 通过 Adam 优化器最小化价值网络预测误差：

$$\varphi_{k+1}=\arg\min_{\varphi}\frac{1}{|D_k|}\sum_{\tau\in D_k}\frac{1}{T}\sum_{t=0}^{T}(V_\varphi(s_t)-R_t)^2$$

9. end

10. end

表 1 PPO 算法参数

Table 1 Parameters of the PPO algorithm

参数	说明
Q	采样幕数据，设置为 100 000
B	批处理大小，设置为 128
D	网络更新间隔，设置为 10
lr	学习率，设置为 1×10^{-4}
decay	学习率衰减系数 0.99，每采样 1 000 幕数据衰减一次
γ	优势函数估计参数折扣因子，设置为 0.9
λ	优势函数估计参数衰减参数，设置为 0.8
ε	裁剪系数，设置为 0.2

3.3.2 基于多头注意力的骑手状态编码器

图 3 展示了本文策略的网络结构。本文考虑文献[18]提出的编码器结构，对骑手状态信息进行提取，使用图 3 左侧结构的骑手状态编码器主要有以下 2 点原因：

1)这种编码器结构具有强大的灵活性与扩展性，能够适应不同数量骑手状态的输入。这使得在某个骑手数量确定的环境下训练的策略网络，能够在具有不同骑手数量的验证集上进行测试，提高了策略的通用性。

2)多头注意力机制的加入能够捕捉系统内所有可用骑手间的相互关系，使得网络能够在订单-骑手的匹配过程中从系统运力全局角度做出决策，而不是贪婪地进行选择。

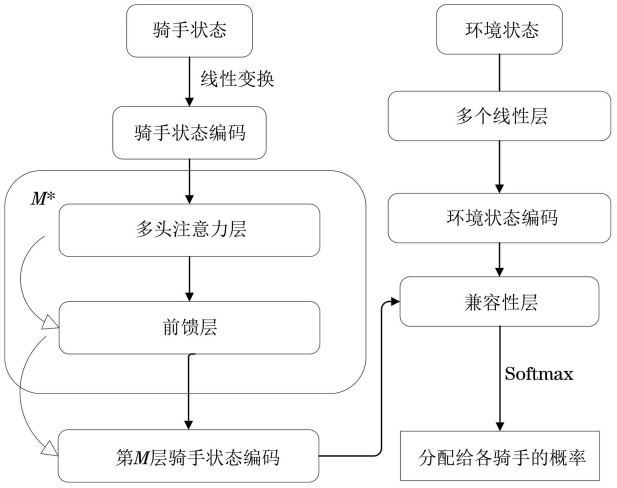

图 3 本文策略网络结构

Fig. 3 The network structure of the strategy in this paper

骑手状态首先进行线性变换，然后进入 M 个堆叠的注意力层。每个注意力层由一个多头注意力

层和前馈层组成，每个子层均加入了残差链接与批归一化(BN)。各子层不共享网络参数，各多头注意力层的头数为 8，隐藏层维度 $d_k=128$，前馈层的隐藏层维度为 256，注意力层层数 $M=4$。

3.3.3 基于缩放注意力的骑手选择

缩放注意力是其中一种特定形式的注意力机制。它通过调整注意力分数的比例来帮助稳定深度网络中梯度的流动，这是通过点积计算得出的注意力分数除以一个缩放因子来实现的。

订单状态特征通过多个连续的线性层获得订单状态编码。在兼容性层[23]中通过点积计算每个骑手状态编码 $\boldsymbol{h}$ 与环境状态编码 $\boldsymbol{k}_{\text{context}}$ 之间的相似度。点积结果通过 Tanh 函数的非线性变换与标量 C 的裁剪控制输出概率分布的熵值，最后经过 Softmax 层输出分配给各骑手的概率。计算公式如式(15)所示：

$$P=\text{Softmax}\left(C\cdot\text{Tanh}\left(\frac{\boldsymbol{h}\cdot\boldsymbol{k}_{\text{context}}^{\mathrm{T}}}{\sqrt{d_k}}\right)\right) \tag{15}$$

Critic 网络将第 M 层的骑手状态编码 $\boldsymbol{h}$ 与环境状态编码 $\boldsymbol{k}_{\text{context}}$ 拼接，经过多个线性层输出状态价值，进而引导 Actor 网络进行更新。

4 数值实验与结果分析

为了验证 PPO-IH 求解外卖即时配送问题的有效性，本文从多个角度进行数值分析。本文实验所用设备配置为 GeForce RTX1050TI 显卡，Intel® Core™ i7-8750H CPU @ 2.20 GHz，2.21 GHz 处理器，RAM 为 16 GB；软件运行环境为 Windows10 操作系统，Python 3.9，PyTorch 2.1.2 框架。

4.1 数据与参数设置

由于外卖即时配送问题没有标准数据集，因此本文参考文献[12]来生成实验数据。本文假设外卖订单于 10:30～14:00 下达，订单的到达服从泊松分布，不同时段订单密度不同，图 4 展示了模拟期间各时间段订单密度占峰值订单密度的比例，在峰值时期，每位骑手每小时期望服务订单数为 4 个。每个订单希望在 45 min 内完成交付，餐厅出餐时间因餐厅而异，服从 5～15 min 的均匀分布。

本文假设某个外卖站点负责一个 6 km×6 km 的区域，站点内有始终可用的 8 位骑手，骑手起点在区域中服从均匀分布，骑手速度为 25 km/h。为了反映道路距离与交通的影响，本文使用文献[22]提出的方法，将两个节点之间的欧氏距离乘以 1.5，近似为真实街道距离。区域内餐厅的数量为 20 个，订单和餐厅的空间分布遵循如下原则：在训练集中，餐厅与客户位置均在区域内服从均匀分布；测试集使用与训练数据相同的分布与参数设置，生成 200 天的模拟数据，用于衡量策略在训练过程中的表现；在验证集中，为考虑真实餐厅地理位置分布特征，对 Grubhub 与佐治亚理工学院提供的数据集(https://github.com/grubhub/mdrplib)中给出的餐厅地理位置数据进行抽样，确定 20 个餐厅的地理位置，客户位置仍然在区域内服从均匀分布，基于以上规则，生成了 200 天的模拟数据作为验证集。

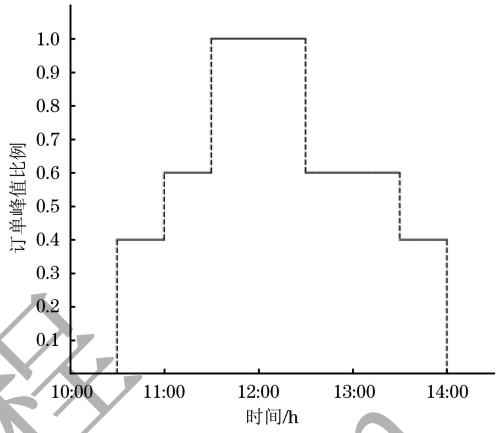

图 4 各时段订单密度比例

Fig.4 The proportion of order density in each time period

在训练过程中，Actor 网络会根据输出概率选择动作，而在测试期间，策略网络总是选择概率最大的骑手进行分配。

4.2 对比算法与对比指标设置

实验对比算法如下：

Greedy：使用第 3.1 节提出的插入启发式算法进行路线规划，在每个决策点，总是将订单分配给具有最优表现的骑手。这是一种短视的方法。Greedy 与 PPO-IH 均使用提出的插入启发式算法对路线进行规划，两者的性能差距能够直接反映 PPO-IH 在随机动态环境下的长期决策能力。

最小差值策略[11]：与 Greedy 相似，最小差值策略总是将订单分配给插入前后具有最小距离成本变动的骑手。

AH[13]：旨在解决多个订单与多个骑手的匹配问题，PPO-IH 与前两种对比算法不同，AH 使用滚动时域方法对订单进行积累，然后进行分配。本文设置滚动时域间隔为 5 min，并使用基于遗憾值的分配方法实现订单-骑手匹配，直到所有待分配订单均被分配。

DDQN[7]：使用 DDQN 将订单分配给具体骑手，为适应问题并保证对比公平性，除奖励函数与路径规划方法使用本文提出的奖励函数与插入启发式算法外，其他设置均不改变。

LAI[21]：使用 DQN 对插入启发式算法求解的目标函数系数进行调整，以指导不同时段的订单-骑手匹配与路径规划。为适应本文问题，使用 LAI 对插入前后的路径成本变化与客户服务水平变化进行调控，其余设置均不改变。

外卖即时配送问题涉及平台、客户、骑手、商家等多方利益，本实验使用如下对比指标：

平均客户服务水平(R)：见式(1)。

平均每单配送距离(f)[11]：骑手行驶总距离与交付订单数量比值。

平均每单配送时间(time)[11]：订单下单至交付所需时间的平均值。

订单延迟率(Delay)：延迟送达订单数量与订单总数的比值。

平均每幕数据运行时间(T)：策略在每幕数据上的平均运行时间，是算法复杂度与效率的体现。

此外，本文研究的外卖即时配送问题的需求具有较强的不确定性，除以上指标外，本文将在 200 天的验证集中，记录 PPO-IH 相对于各对比算法有更高的总奖励值的幕数(天数)百分比(H)，以衡量策略在随机需求环境下的稳定性。

4.3 数值实验与结果分析

图 5 为策略网络的训练曲线，展示了不同策略在训练过程中 200 天的测试集上的平均奖励。实线为 Greedy 在测试集上的平均客户服务水平，由于 Greedy 是一个确定性的策略，其曲线为固定水平线。实线表示 PPO-IH 在测试集上的平均客户服务水平，每采样 1 000 幕数据，PPO-IH 在测试集上进行测试，测试期间贪婪选择具有最大分配概率的骑手进行分配。通过观察可以得知，PPO-IH 在 50 000 幕数据后表现超过 Greedy，并且在此后效果进一步提升。PPO-IH Without Attention 的策略网络将M个注意力层更换为M个线性层，其余网络结构与参数均不变。PPO-IH Without Attention 在 50 000 幕数据已接近收敛，在训练末期效果也未能超越 Greedy，这表明注意力机制的加入使得策略网络能够学习系统中骑手复杂的依赖关系，以做出更优决策。

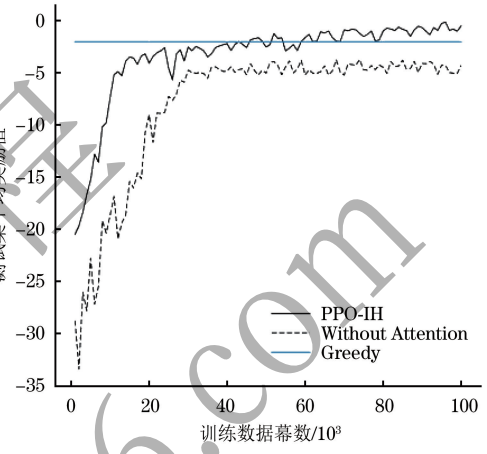

图 5 策略网络的训练曲线

Fig. 5 The training curve of the strategy network

表 2 展示了本文提出的 PPO-IH 与 5 种对比算法在 200 天验证集上的性能对比。相对 Greedy，PPO-IH 能够提供更高的平均客户服务水平、更短的平均每单配送时间和更低的订单延迟送达率，同时在更多的天数中有着更优的表现，这展示了 PPO-IH 在随机需求问题下的稳定性与优越性。通过观察可以发现，PPO-IH 在订单-骑手匹配过程中，通过牺牲平均每单配送距离，对系统内骑手进行隐式调度，合理规划不同时段运力，最终获得更优的长期平均客户服务水平。而在深度强化学习对比算法中，DDQN 各项指标均劣于 Greedy，LAI 通过调控对距离成本与客户服务水平的重视程度，在平均客户服务水平等指标上均略微优于 Greedy，但距离本文提出的 PPO-IH 仍有明显差距，其性能可能受限于 DQN、特征表示、奖励设计等多种因素。

表 2 各算法求解结果对比

Table 2 Comparison of the solution results of each algorithm

算法	R	time/min	f/km	Delay/%	H/%	T/s
PPO-IH	0.99	37.95	3.73	27.8	—	4.42
Greedy	−0.85	39.11	3.51	31.3	71.5	3.05
AH	−3.27	41.05	3.56	37.8	95.5	6.01
最小差值策略	−1.56	40.28	3.29	34.7	87.5	3.67
DDQN	−1.21	39.87	3.54	31.9	79.5	3.87
LAI	−0.74	38.98	3.48	30.8	70.0	3.32

进一步地，将 Greedy 在 200 天模拟数据上的表现升序排列，将 200 天数据分为对运力挑战低、高两部分，对两个部分数据上 PPO-IH 与 Greedy 算法各项指标进行分析，结果如表 3 所示。

表 3 不同运力挑战下算法性能对比

Table 3 Comparison of algorithms performance under different capacity challenges

运力挑战	算法	R	time/min	f/km	Delay/%	H/%
低	PPO-IH	6.38	35.15	3.90	19.4	59.0
	Greedy	5.82	36.02	3.66	22.3	
高	PPO-IH	−4.45	40.71	3.48	36.1	79.0
	Greedy	−6.91	41.95	3.31	39.9	

从表 3 可以看出，在运力挑战更高的数据集中，相对 Greedy，PPO-IH 在平均客户服务水平、平均每单配送时间、违约率和 PPO-IH 更优比例 4 个指标上均有明显的优势，这说明 PPO-IH 能够有效地应对运力挑战高的场景，减少订单延迟率，提高客户服务体验。同时，在运力挑战更高的场景下，PPO-IH 相对 Greedy 的平均每单配送距离的劣势也会更小。图 6 展示了两种运力挑战情况下的各幕数据的平均客户服务体验箱线图，在两种运力挑战下，PPO-IH 的客户服务体验平均值、中位数、上下四分位数均优于 Greedy。

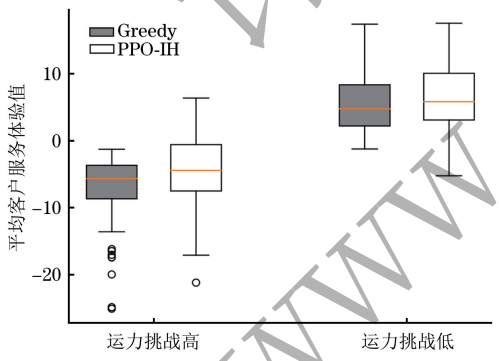

图 6 不同运力挑战下算法性能箱线图

Fig.6 Box plot of algorithms performance under different capacity challenges

为了验证 PPO-IH 在与训练数据参数、分布相异的数据上的性能变化，保持其他参数设置与训练数据相同，针对可用骑手数量、订单时间窗设计、订单密度的变化检验策略性能。其中，定制时间窗根据订单的餐厅节点与客户节点所需的旅行时间确定，表达式如下：

$$t_o^d = t_0 + \min\{30, u_t \times 1.3\} \tag{16}$$

式中：u_t 为订单紧急程度。订单密度以第 4.1 节描述为例，观察在峰值时刻，每位骑手每小时期望平均服务订单数量为 3.5、4、4.5 个的数据集中 PPO-IH 与 Greedy 性能的变化。实验结果如表 4 所示。

从表 4 可以看出，在各参数设置下，PPO-IH 相对 Greedy 均保持着更好的表现，但是相对于与训练数据同分布或相同参数的数据，表现仍有一定下滑。总体来说，当参数变化引起配送压力变大时，PPO-IH 的下滑趋势更小，展现了所提出的算法在运力不足的环境下的优势。在订单密度为 3.5 时，Greedy 在更多的天数中相对于 PPO-IH 有着更好的表现，但是其平均客户服务体验值略低于 PPO-IH。

表 4 泛化性实验结果

Table 4 Generalization experiment results

参数	参数值	R		H/%
		PPO-IH	Greedy	
骑手数	6	−24.50	−27.30	62.5
	7	−9.80	−11.20	68.5
	8	0.99	−0.85	71.5
	9	6.81	6.10	64.5
	10	8.56	8.98	56.0
时间窗	45	0.99	−0.85	67.5
	60	22.15	21.34	55.5
	定制	2.07	0.62	65.0
订单密度	3.5	6.15	6.05	45.5
	4.0	0.99	−0.85	67.5
	4.5	−3.25	−4.49	59.5

相比 Greedy 在各个时段均贪婪地选择最优骑手进行分配，PPO-IH 如何在不同时段做出决策来获得更高的长期收益是令人感兴趣的。然而深度学习可解释性弱，绘制不同策略在各决策点产生的奖励是了解 PPO-IH 决策逻辑的一个方法。图 7 为 PPO-IH 与 Greedy 在各决策点活跃订单(已分配但未交付)的预期平均客户服务水平差值图，纵坐标为决策点处 PPO-IH 活跃订单预期平均客户服务水平减去 Greedy 活跃订单预期平均客户服务水平。

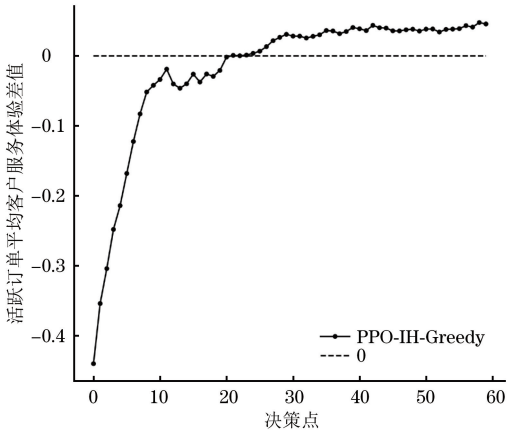

图 7 各决策点活跃订单平均服务水平差值图

Fig.7 Graph of average service level differences for active orders at each decision points

由图 7 可知，相较于短视的 Greedy，PPO-IH 在模拟初期倾向选择次优订单-骑手分配方案，对系统内各骑手进行调度，以应对未来订单高峰期时的需求，这说明 PPO-IH 在随机需求下的外卖即时配送问题中具有一定的全局决策能力。

5 结束语

本文将考虑随机需求的外卖即时配送问题建模为基于路径的马尔可夫决策过程，并设计 PPO-IH 对订单-骑手进行匹配与路径规划。PPO-IH 融合了注意力机制，能够识别与提取外卖配送系统内骑手间相互影响，从全局角度进行决策。实验结果表明，PPO-IH 在除平均每单配送距离外的所有指标上均超越其他对比算法，并展示了一定的泛化性。本文考虑的随机动态因素只有随时间产生的订单需求，并没有考虑如受意外事件影响的旅行时间、交付时间、备餐时间等，这些因素会直接影响骑手的调度与路线的规划，也会涉及已分配订单的转派等。另外，本文仅从客户角度出发，最大化平均客户服务体验，而外卖即时配送问题涉及骑手、客户、商家、平台等多个主体，如何在动态的配送过程中平衡各方利益[25]，特别是利用深度强化学习解决这一问题也是未来的研究方向之一。

参考文献

[1] 华经情报网. 2022 年中国网上外卖行业分析[EB/OL]. [2024-02-10]. https://www.huaon.com/.
Huajing Intelligence Network. Analysis of China's online food delivery industry in 2022[EB/OL]. [2024-02-10]. https://www.huaon.com/. (in Chinese)

[2] HILDEBRANDT F D, THOMAS B W, ULMER M W. Opportunities for reinforcement learning in stochastic dynamic vehicle routing[J]. Computers & Operations Research, 2023, 150: 106071.

[3] 张玉州, 叶亮, 郑军帅. 基于滚动时域控制的动态外卖配送问题优化[J]. 计算机技术与发展, 2019, 29(10): 83-88, 94.
ZHANG Y Z, YE L, ZHENG J S. Optimization of dynamic takeaway distribution problem based on receding horizon control[J]. Computer Technology and Development, 2019, 29(10): 83-88, 94. (in Chinese)

[4] 李桃迎, 吕晓宁, 李峰, 等. 考虑动态需求的外卖配送路径优化模型及算法[J]. 控制与决策, 2019, 34(2): 406-413.
LI T Y, LYU X N, LI F, et al. Routing optimization model and algorithm for takeout distribution with multiple fuzzy variables under dynamics demand[J]. Control and Decision, 2019, 34(2): 406-413. (in Chinese)

[5] BOZANTA A, CEVIK M, KAVAKLIOGLU C, et al. Courier routing and assignment for food delivery service using reinforcement learning[J]. Computers & Industrial Engineering, 2022, 164: 107871.

[6] JAHANSHAHI H, BOZANTA A, CEVIK M, et al. A deep reinforcement learning approach for the meal delivery problem[J]. Knowledge-Based Systems, 2022, 243: 108489.

[7] ZOU G Y, TANG J F, YILMAZ L, et al. Online food ordering delivery strategies based on deep reinforcement learning[J]. Applied Intelligence, 2022, 52(6): 6853-6865.

[8] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[EB/OL]. [2024-02-10]. https://arxiv.org/abs/1707.06347v2.

[9] 余海燕, 唐婉倩, 吴腾宇. 带硬时间窗的 O2O 生鲜外卖即时配送路径优化[J]. 系统管理学报, 2021, 30(3): 584-591.
YU H Y, TANG W Q, WU T Y. Vehicle routing problem with hard time windows for instant delivery of O2O fresh takeout orders[J]. Journal of Systems & Management, 2021, 30(3): 584-591. (in Chinese)

[10] 冯爱兰, 周映雪, 龚艳茹, 等. 抢派结合模式下外卖配送问题研究[J]. 控制与决策, 2024, 39(9): 3135-3142.
FENG A L, ZHOU Y X, GONG Y R, et al. Research on takeout distribution based on combination mode of order dispatching and grabbing[J]. Control and Decision, 2024, 39(9): 3135-3142. (in Chinese)

[11] 余海燕, 蒋仁莲. 基于众包平台的外卖实时配送订单分配与路径优化研究[J]. 工业工程与管理, 2022, 27(2): 146-152.
YU H Y, JIANG R L. Study on the real-time order allocation and routing problem of takeout food distribution on crowdsourcing platform[J]. Industrial Engineering and Management, 2022, 27(2): 146-152. (in Chinese)

[12] STEEVER Z, KARWAN M, MURRAY C. Dynamic courier routing for a food delivery service[J]. Computers & Operations Research, 2019, 107: 173-188.

[13] ULMER M W, THOMAS B W, CAMPBELL A M, et al. The restaurant meal delivery problem: dynamic pickup and delivery with deadlines and random ready times[J]. Transportation Science, 2021, 55(1): 75-100.

[14] CHEN J F, WANG L, REN H, et al. An imitation learning-enhanced iterated matching algorithm for on-demand food delivery[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(10): 18603-18619.

[15] WANG X, WANG L, DONG C X, et al. An online deep reinforcement learning-based order recommendation framework for rider-centered food delivery system[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(5): 5640-5654.

[16] WANG X, WANG L, DONG C X, et al. Reinforcement learning-based dynamic order recommendation for on-demand food delivery[J]. Tsinghua Science and Technology, 2023, 29(2): 356-367.

[17] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[EB/OL]. [2024-02-10]. https://arxiv.org/abs/1312.5602v1.

[18] KOOL W, VAN HOOF H, WELLING M. Attention, learn to solve routing problems![EB/OL]. [2024-02-10]. https://arxiv.org/abs/1803.08475v3.

[19] XU Z, LI Z X, GUAN Q W, 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, USA: ACM Press, 2018: 905-913.

[20] 黄晓辉, 杨凯铭, 凌嘉壕. 基于共享注意力的多智能体强化学习订单派送[J]. 计算机应用, 2023, 43(5): 1620-1624.
HUANG X H, YANG K M, LING J H. Order dispatching by multi-agent reinforcement learning based on shared

attention[J]. Journal of Computer Applications, 2023, 43(5): 1620-1624. (in Chinese)
[21] WEI C, WANG Y H, YAN X D, et al. Look-ahead insertion policy for a shared-taxi system based on reinforcement learning[J]. IEEE Access, 2018, 6: 5716-5726.
[22] CHEN X W, ULMER M W, THOMAS B W. Deep Q-learning for same-day delivery with vehicles and drones[J]. European Journal of Operational Research, 2022, 298(3): 939-952.
[23] ULMER M W, GOODSON J C, MATTFELD D C, et al. On modeling stochastic dynamic vehicle routing problems[J]. EURO Journal on Transportation and Logistics, 2020, 9(2): 100008.
[24] BATTA R, LEJEUNE M, PRASAD S. Public facility location using dispersion, population, and equity criteria[J]. European Journal of Operational Research, 2014, 234(3): 819-829.
[25] 熊浩, 郭昊颖, 鄢慧丽, 等. 外卖配送路径多目标实时优化研究[J]. 工业工程, 2023, 26(1): 98-107.
XIONG H, GUO H Y, YAN H L, et al. Multi-objective real-time optimization study of takeaway vehicle routes problem[J]. Industrial Engineering Journal, 2023, 26(1): 98-107. (in Chinese)

编辑 索书志