(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114118851 B
(45)授权公告日 2024.07.16

(21)申请号 202111470404.3
(22)申请日 2021.12.03
(65)同一申请的已公布的文献号
申请公布号 CN 114118851 A
(43)申请公布日 2022.03.01
(73)专利权人 东南大学
地址 210000 江苏省南京市麒麟科创园智识路26号启迪城立业园04幢
(72)发明人 王帅 胡世杰 梅洛瑜
(74)专利代理机构 南京众联专利代理有限公司 32206
专利代理师 薛雨妍
(51)Int.Cl.
G06Q 10/0631(2023.01)
G06Q 10/083(2024.01)
G06N 3/092(2023.01)
(56)对比文件
CN 110110950 A,2019.08.09
CN 107230014 A,2017.10.03
审查员 武茹茹

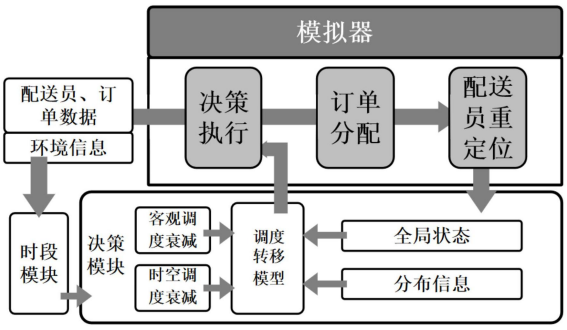
权利要求书2页 说明书6页 附图2页

(54)发明名称

一种基于强化学习的区域间供需平衡的方法

(57)摘要

本发明提供一种基于强化学习的区域间供需平衡的方法,建立基于强化学习的区域间供需平衡技术框架;该技术考虑到即时配送场景下区域层面供需不平衡的问题,通过选择合适的配送员进行调度,综合考虑供需平衡效率与配送员个人效益,实现整体的供需平衡。使用该技术可以较为高效地实现区域间供需平衡,并有效减少调度时配送员的利益损失。



CN 114118851 B

1.一种基于强化学习的区域间供需平衡的方法,其特征在于:

建立基于强化学习的区域间供需平衡的技术框架;其中所述技术框架由三个模块组成,分别包括以下:

模块一:模拟器模块,用于模拟配送员在即时配送中的常规流程以及为算法构建训练和评估的环境;通过平台中采集的配送员配送过程中的GPS信息,可以获得配送员的分布,各个商家产生的订单记录,以及区域的地理信息,提取该部分数据的特征,用于驱动模拟器;

模块二:决策模块,采用基于Actor-Critic强化学习的调度转移模型,根据模拟器模块对该模块的输入,通过网络提取特征,作出基于该时间步初始状态信息的决策,对该时间步的配送员进行行为指导,对各区域的配送员进行再平衡;通过从即时配送平台提取商店位置信息、订单数据、配送员GPS数据;根据多种数据的分析与特征提取,获取不同区域不同时间的统计数据,实现模拟器驱动;获取模拟器中不同时态的状态信息,输入调度转移模型获取配送员调度决策,反馈至模拟器循环从而根据数据进行调度转移模型的训练;

模拟器设计可以根据输入数据初始化配送员和订单的分布,并模拟真实世界来分配订单和调度配送员;

首先获取统计数据以驱动模拟器运行,模拟器设计如下:

在模拟器中,我们构建了四个类,以便为模拟器中的操作创建对象,分别为:区域、网格、订单和配送员;以下将分别对四个类进行定义;

(1)Region:自身属性包含区域ID、区域内包含的网格ID、区域内配送员总数和区域内订单总数;主要操作包括获取区域包含的网格、计算所有网格内的配送员总数、计算所有网格内的订单总数;该类的主要作用为作为区域单位,为之后的调度转移动作设置好各个区域的边界;同时在reward计算时,以各区域作为单位计算各个区域内的订单与配送员数之比,并计算衡量算法优劣的相关的衡量指标;(2)Node:将区域网格离散化,更便于强化学习中状态空间和动作空间的构建;自身主要属性包含区域ID、网格ID、网格类型、配送员数、订单数和邻居网格;主要操作包括网格初始化、获取邻居网格、计算网格内配送员数量、生成网格内订单、移除网格内配送员、设置网格内配送员上下线、添加配送员、移除分派的订单;网格承担着各种操作的载体的作用,区域的数据计算同样也依赖于网格内部的计算;

(3)Order:自身属性包含起始网格、目的网格、下单时间、持续时间、价格、等待时间;主要操作为通过输入的数据对自身的订单信息逐一初始化;订单类仅仅是用于创建订单对象以便于订单分派的操作,关于订单自身,在本模拟器中没有额外的自身操作;

(4)Courier:自身属性包含配送员ID、是否在线、是否服务、接收的订单、当前所处网格;主要操作包括订单的接收、订单的配送以及自身在线与离线状态的转换;配送员类是算法调度转移的主要对象,在自身按照现实即时配送流程执行操作的同时,会接收到算法分派的调度转移指令

模块三:时段模块,根据订单数据,提取一天内不同时段的订单特征,根据订单的时段变化特点,对决策模块进行调度力度的指导。

2.根据权利要求1所述的一种基于强化学习的区域间供需平衡的方法,其特征在于:模拟器中,若不进行跨区域调度的决策,每个区域的配送员数量的变化仅受到配送员离线和上线的操作的影响;为模拟真实场景,配送员数量在每个时间步中进行变;空闲的配送员接

收到调度决策直接执行;配送员进行跨区域调度时,区域到区域间转移的过程必定是一个空跑状态,根据空跑时长我们可以得到概率D_T;根据空跑时长决定配送员调度的相应概率,移动距离越长,空跑时间越长,配送员被选择调度概率越低;正在服务的配送员在现有订单配送结束后执行决策,根据订单完成时长我们可以得到概率O_T;根据订单完成时长决定配送员调度的相应概率,订单完成时间越长,配送员被选择调度概率越低;未被选择进行调度的配送员,继续进行该区域的正常订单配送流程。

3.根据权利要求1所述的一种基于强化学习的区域间供需平衡的方法,其特征在于:所述调度转移模型的具体设计如下:

该模型方法主体框架为A2C算法框架,包括critic网络和actor网络,用于评估策略网络与输出调度转移策略;

该模型通过输入模拟器每个时间步下的状态,根据不同网格的不同状态下的配送员,其被分配的调度策略概率也不尽相同,如下所示;

$\pi=D_TO_TP(a_t(k))$

其中$P(a_t(k))$表示当前配送员执行各个动作的概率,D_T、O_T如前所述,分别为考虑配送员空跑距离与订单完成时长的相应动作概率衰减因子,π为考虑该部分因素后各个配送员的动作决策概率。

4.根据权利要求3所述的一种基于强化学习的区域间供需平衡的方法,其特征在于:对于每次动作的选择,其中存在着显著的客观的实际因素,我们需要进行一些处理;从地理因素上看,若两个区域并非相邻的,则对于这两个区域间的调度决策必然会导致配送员的长距离空跑,尤其当区域分别位于城市的两端时;因此,对于明显的地理不相邻的区域,我们设定其相互之间不会进行配送员的调度,得到N_{gi};

从获得的收益角度看,当考虑前述的两个个人因素后,进行区域调度转移所能获得的收益不及停留于当前区域,则不对该决策进行考虑,得到V_{gi};

综上,我们得到本算法中决策动作的选择概率如下所示;

$\pi=D_TO_TP(a_t(k))N_{gi}V_{gi}$;

通过模拟器状态输出,决策模型决策输入,根据现实数据进行模型的网络的训练与评估,使其能得到优异的配送员调度策略。

5.根据权利要求1所述的一种基于强化学习的区域间供需平衡的方法,其特征在于:所述时段模块的作用如下:根据初始的订单数据输入,提取订单数量一天内分布特征,由于一天内订单数量存在高峰与低谷,如午高峰与晚高峰;考虑到在订单数量高峰期,配送员整体供应紧张,对配送员进行调度会影响配送员效率与收益,时段模块在每一个时段前,根据该时段订单数量对决策模块进行调度力度的改变,在订单数少的时间段,多进行调度以为后续订单高峰期做好准备;在订单数高峰期,少进行调度,以避免调度产生的负面效益。

一种基于强化学习的区域间供需平衡的方法

技术领域

[0001] 本发明于涉及计算机计算技术领域,尤其涉及一种基于强化学习的区域间供需平衡的方法。

背景技术

[0002] 即时配送场景中,由于区域的划分,订单和配送员数量随时间进行动态变化,不同区域间的订单数量与配送员数量会出现不平衡现象,部分区域订单数过多而部分区域配送员过多,这导致了整体的配送效率的降低。

[0003] 因此,实现区域间的供需平衡问题能有效提高整体配送效率,大多现有工作无法应用于该问题:(1)即时配送场景中存在着区域的划分,配送员通常活动于固定的配送区域;(2)配送员能一次进行多个订单的配送;(3)区域间不平衡随时间动态变化。

发明内容

[0004] 为解决上述问题,本发明公开了一种基于强化学习的区域间供需平衡的方法,通过结合配送员的轨迹数据和订单报告的过程数据,建立一个模拟器来模拟配送员的调度转移环境,考虑到即时配送场景下区域层面供需不平衡的问题,通过选择合适的配送员进行调度,综合考虑供需平衡效率与配送员个人效益,实现整体的供需平衡。使用该技术可以较为高效地实现区域间供需平衡,并有效减少调度时配送员的利益损失。

[0005] 一种基于强化学习的区域间供需平衡的方法,考虑到即时配送场景下区域层面供需不平衡的问题,可以有效平衡订单数量和配送员数量,并提高配送员调度效率。

[0006] 建立基于强化学习的区域间供需平衡技术框架;其中所述技术框架由三个模块组成,分别包括以下:

[0007] 模块一:模拟器模块:

[0008] 用于模拟配送员在即时配送中的常规流程以及为算法构建训练和评估的环境;通过平台中采集的配送员配送过程中的GPS信息,可以获得配送员的分布,各个商家产生的订单记录,以及区域的地理信息,提取该部分数据的特征,用于驱动数据驱动模拟器;

[0009] 模块二:决策模块:

[0010] 采用基于Actor-Critic强化学习的调度转移模型,根据模拟器模块对该模块的输入,通过网络提取特征,作出基于该时间步初始状态信息的决策,对该时间步的配送员进行行为指导,对各区域的配送员进行再平衡;

[0011] 通过从即时配送平台提取商店位置信息、订单数据、配送员GPS数据;根据多种数据的分析与特征提取,获取不同区域不同时间的统计数据,实现模拟器驱动;获取模拟器中不同时态的状态信息,输入调度转移模型获取配送员调度决策,反馈至模拟器循环从而根据数据进行调度转移模型的训练。

[0012] 模块三:时段模块:

[0013] 根据订单数据,提取一天内不同时段的订单特征,根据订单的时段变化特点,对决

策模块进行调度力度的指导。

[0014] 进一步优选的,模拟器设计:

[0015] 强化学习的交互性在训练和评估中带来了复杂的困难,因此通过为环境构建一个模拟器,提供了一个强化学习算法训练和评估的环境。我们构建的模拟器可以根据输入数据初始化配送员和订单的分布,并模拟真实世界来分配订单和调度配送员。

[0016] 首先获取统计数据以驱动模拟器运行,模拟器设计如下:

[0017] 在模拟器中,我们构建了四个类,以便为模拟器中的操作创建对象,分别为:区域(Region)、网格(Node)、订单(Order)和配送员(Courier);以下将分别对四个类进行定义。

[0018] (1)Region:自身属性包含区域ID、区域内包含的网格ID、区域内配送员总数和区域内订单总数;主要操作包括获取区域包含的网格、计算所有网格内的配送员总数、计算所有网格内的订单总数;该类的主要作用为作为区域单位,为之后的调度转移等动作设置好各个区域的边界;同时在reward计算时,以各区域作为单位计算各个区域内的订单与配送员数之比,并计算衡量算法优劣的相关的衡量指标;

[0019] (2)Node:将区域网格离散化,更便于强化学习中状态空间和动作空间的构建;自身主要属性包含区域ID、网格ID、网格类型、配送员数、订单数和邻居网格;主要操作包括网格初始化、获取邻居网格、计算网格内配送员数量、生成网格内订单、移除网格内配送员、设置网格内配送员上下线、添加配送员、移除分派的订单;网格承担着各种操作的载体的作用,区域的数据计算同样也依赖于网格内部的计算;

[0020] (3)Order:自身属性包含起始网格、目的网格、下单时间、持续时间、价格、等待时间等;主要操作为通过输入的数据对自身的订单信息逐一初始化。订单类仅仅是用于创建订单对象以便于订单分派的操作,关于订单自身,在本模拟器中没有额外的自身操作;

[0021] (4)Courier:自身属性包含配送员ID、是否在线、是否服务、接收的订单、当前所处网格等;主要操作包括订单的接收、订单的配送以及自身在线与离线状态的转换。配送员类是算法调度转移的主要对象,在自身按照现实即时配送流程执行操作的同时,会接收到算法分派的调度转移指令。

[0022] 模拟器中配送员行为模式如下,若不进行跨区域调度的决策,每个区域的配送员数量的变化仅受到配送员离线和上线的操作的影响;为模拟真实场景,配送员数量在每个时间步中进行变化。

[0023] 空闲的配送员接收到调度决策直接执行;

[0024] 配送员进行跨区域调度时,区域到区域间转移的过程必定是一个空跑状态,根据空跑时长我们可以得到概率D_T;根据空跑时长决定配送员调度的相应概率,移动距离越长,空跑时间越长,配送员被选择调度概率越低;

[0025] 正在服务的配送员在现有订单配送结束后执行决策,根据订单完成时长我们可以得到概率O_T;根据订单完成时长决定配送员调度的相应概率,订单完成时间越长,配送员被选择调度概率越低;

[0026] 未被选择进行调度的配送员,继续进行该区域的正常订单配送流程。

[0027] 优选的,调度转移模型设计

[0028] 该模型方法主体框架为A2C算法框架,包括critic网络和actor网络,用于评估策略网络与输出调度转移策略;

[0029] 该模型通过输入模拟器每个时间步下的状态,根据不同网格的不同状态下的配送员,其被分配的调度策略概率也不尽相同,如下所示;

[0030] $\pi = D_T O_T P(a_t(k))$

[0031] 其中$P(a_t(k))$表示当前配送员执行各个动作的概率,D_T、O_T如前所述,分别为考虑配送员空跑距离与订单完成时长的相应动作概率衰减因子,π为考虑该部分因素后各个配送员的动作决策概率。

[0032] 同时,对于每次动作的选择,其中存在着显著的客观的实际因素,我们需要进行一些处理;从地理因素上看,若两个区域并非相邻的,则对于这两个区域间的调度决策必然会导致配送员的长距离空跑,尤其当区域分别位于城市的两端时;因此,对于明显的地理不相邻的区域,我们设定其相互之间不会进行配送员的调度,得到N_{g_i};

[0033] 从获得的收益角度看,当考虑前述的两个个人因素后,进行区域调度转移所能获得的收益不及停留于当前区域,则不对该决策进行考虑,得到V_{g_i};

[0034] 综上,我们得到本算法中决策动作的选择概率如下所示。

[0035] $\pi = D_T O_T P(a_t(k)) N_{g_i} V_{g_i}$;

[0036] 通过模拟器状态输出,决策模型决策输入,根据现实数据进行模型的网络的训练与评估,使其能得到优异的配送员调度策略。

[0037] 本发明的有益效果:该技术考虑到区域层面供需的动态变化和严格的时间约束,通过整体的配送员跨区域调度,综合考虑调度效率与配送员个人效益,实现整体的供需平衡;可以有效提高调度转移性能,并减少调度时配送员平均空跑时间。

[0038] 优选的,时段模块设计:

[0039] 根据初始的订单数据输入,提取订单数量一天内分布特征,由于一天内订单数量存在高峰与低谷,如午高峰与晚高峰。考虑到在订单数量高峰期,配送员整体供应紧张,对配送员进行调度会影响配送员效率与收益,时段模块在每一个时段前,根据该时段订单数量对决策模块进行调度力度的改变,在订单数少的时间段,多进行调度以为后续订单高峰期做好准备;在订单数高峰期,少进行调度,以避免调度产生的负面效益。

附图说明

[0040] 图1、本发明技术的技术框架图;

[0041] 图2、模拟器流程图;

[0042] 表1、本发明统计数据。

具体实施方式

[0043] 下面结合附图和具体实施方式,进一步阐明本发明,应理解下述具体实施方式仅用于说明本发明而不用于限制本发明的范围。需要说明的是,下面描述中使用的词语“前”、“后”、“左”、“右”、“上”和“下”指的是附图中的方向,词语“内”和“外”分别指的是朝向或远离特定部件几何中心的方向。

[0044] 实施例1:

[0045] 用于有商圈区域划分的大型城市即时配送系统;通过收集系统中记录的本设计所

需的数据,进行数据特征提取,进行模拟器驱动以训练评估跨区域配送员调度转移模型,最后可将其实际运用与即时配送平台上。

[0046] (一)基于强化学习的区域间供需平衡技术框架

[0047] 考虑到即时配送场景下区域层面供需不平衡的问题,该框架可以有效平衡订单数量和配送员数量,并提高配送员调度效率。

[0048] 系统框架如图1所示,由三个模块组成:首先是一个模拟器,用于模拟配送员在即时配送中的常规流程以及为算法构建训练和评估的环境。通过平台中采集的配送员配送过程中的GPS信息,可以获得配送员的分布,各个商家产生的订单记录,以及区域的地理信息等,提取该部分数据的特征,用于驱动我们设计的模拟器。

[0049] 决策模块采用基于Actor-Critic强化学习的调度转移模型,根据模拟器模块对该模块的输入,通过网络提取特征,作出基于该时间步初始状态信息的决策,对该时间步的配送员进行行为指导,对各区域的配送员进行再平衡。

[0050] 通过从即时配送平台提取商店位置信息、订单数据、配送员GPS数据;根据多种数据的分析与特征提取,获取不同区域不同时间的统计数据,实现模拟器驱动;获取模拟器中不同时态的状态信息,输入调度转移模型获取配送员调度决策,反馈至模拟器循环从而根据数据进行调度转移模型的训练。

[0051] 时段模块根据订单数据,提取一天内不同时段的订单特征,根据订单的时段变化特点,对决策模块进行调度力度的指导。

[0052] (二)模拟器设计

[0053] 强化学习的交互性在训练和评估中带来了复杂的困难,因此通过为环境构建一个模拟器,提供了一个强化学习算法训练和评估的环境。我们构建的模拟器可以根据输入数据初始化配送员和订单的分布,并模拟真实世界来分配订单和调度配送员,流程如图2所示。

[0054] 首先获取统计数据以驱动模拟器运行,数据如表1所示,

[0055] 模拟器设计如下:

[0056] 在模拟器中,我们构建了四个类,以便为模拟器中的操作创建对象,分别为:区域(Region)、网格(Node)、订单(Order)和配送员(Courier)。以下将分别对四个类进行定义。

[0057] (1)Region:自身属性包含区域ID、区域内包含的网格ID、区域内配送员总数和区域内订单总数。主要操作包括获取区域包含的网格、计算所有网格内的配送员总数、计算所有网格内的订单总数。该类的主要作用为作为区域单位,是本课题中主要的一个区域划分单元,为之后的调度转移等动作设置好各个区域的边界。同时在reward计算时,以各区域作为单位计算各个区域内的订单与配送员数之比,并计算衡量算法优劣的相关的衡量指标。

[0058] (2)Node:将区域网格离散化,更便于强化学习中状态空间和动作空间的构建。自身主要属性包含区域ID、网格ID、网格类型、配送员数、订单数和邻居网格等。主要操作包括网格初始化、获取邻居网格、计算网格内配送员数量、生成网格内订单、移除网格内配送员、设置网格内配送员上下线、添加配送员、移除分派的订单等。可见网格的属性与操作十分多,城市的网格划分作为一个常用的手段,选取合适的范围统一该范围内的同类信息,能有效的简化算法设计与网络训练的复杂度,同时又能保证算法应用于显示世界的有效性和合理性。因此对于作为模拟器中最最基础的单位,网格承担着各种操作的载体的作用,区域的

数据计算同样也依赖于网格内部的计算。

[0059] (3)Order:自身属性包含起始网格、目的网格、下单时间、持续时间、价格、等待时间等。主要操作为通过输入的数据对自身的订单信息逐一初始化。订单类仅仅是用于创建订单对象以便于订单分派的操作,关于订单自身,在本模拟器中没有额外的自身操作。

[0060] (4)Courier:自身属性包含配送员ID、是否在线、是否服务、接收的订单、当前所处网格等。主要操作包括订单的接收、订单的配送以及自身在线与离线状态的转换。配送员类是算法调度转移的主要对象,在自身按照现实即时配送流程执行操作的同时,会接收到算法分派的调度转移指令。

[0061] 配送员行为模式如下。若不进行跨区域调度的决策,每个区域的配送员数量的变化仅受到配送员离线和上线的操作的影响;为模拟真实场景,配送员数量在每个时间步中进行变化。

[0062] 空闲的配送员接收到调度决策直接执行;

[0063] 配送员进行跨区域调度时,区域到区域间转移的过程必定是一个空跑状态,根据空跑时长我们可以得到概率D_T;根据空跑时长决定配送员调度的相应概率,移动距离越长,空跑时间越长,配送员被选择调度概率越低;

[0064] 正在服务的配送员在现有订单配送结束后执行决策,根据订单完成时长我们可以得到概率O_T;根据订单完成时长决定配送员调度的相应概率,订单完成时间越长,配送员被选择调度概率越低;

[0065] 未被选择进行调度的配送员,继续进行该区域的正常订单配送流程。

[0066] (三)调度转移模型设计

[0067] 设计的调度转移模型根据整体的配送员与订单分布,统筹的进行跨区域的配送员调度,具体设计如下:

[0068] 该模型方法主体框架为A2C算法框架,包括critic网络和actor网络,用于评估策略网络与输出调度转移策略;

[0069] 该模型通过输入模拟器每个时间步下的状态,根据不同网格的不同状态下的配送员,其被分配的调度策略概率也不尽相同,如下所示;

[0070] $\pi=D_TO_TP(a_t(k))$

[0071] 其中$P(a_t(k))$表示当前配送员执行各个动作的概率,D_T、O_T如前所述,分别为考虑配送员空跑距离与订单完成时长的相应动作概率衰减因子,π为考虑该部分因素后各个配送员的动作决策概率。

[0072] 同时,对于每次动作的选择,其中存在着显著的客观的实际因素,我们需要进行一些处理;从地理因素上看,若两个区域并非相邻的,则对于这两个区域间的调度决策必然会导致配送员的长距离空跑,尤其当区域分别位于城市的两端时;因此,对于明显的地理不相邻的区域,我们设定其相互之间不会进行配送员的调度,得到N_{g_i};

[0073] 从获得的收益角度看,当考虑前述的两个个人因素后,进行区域调度转移所能获得的收益不及停留于当前区域,则不对该决策进行考虑,得到V_{g_i};

[0074] 综上,我们得到本算法中决策动作的选择概率如下所示。

[0075] $\pi = D_T O_T P(a_t(k)) N_{g_i} V_{g_i}$;

[0076] 通过模拟器状态输出,决策模型决策输入,根据现实数据进行模型的网络的训练与评估,使其能得到优异的配送员调度策略。

[0077] (四)时段模块设计

[0078] 根据初始的订单数据输入,提取订单数量一天内分布特征,由于一天内订单数量存在高峰与低谷,如午高峰与晚高峰。考虑到在订单数量高峰期,配送员整体供应紧张,对配送员进行调度会影响配送员效率与收益,时段模块在每一个时段前,根据该时段订单数量对决策模块进行调度力度的改变,在订单数少的时间段,多进行调度以为后续订单高峰期做好准备;在订单数高峰期,少进行调度,以避免调度产生的负面效益。

[0079] 本发明方案所公开的技术手段不仅限于上述实施方式所公开的技术手段,还包括由以上技术特征任意组合所组成的技术方案。

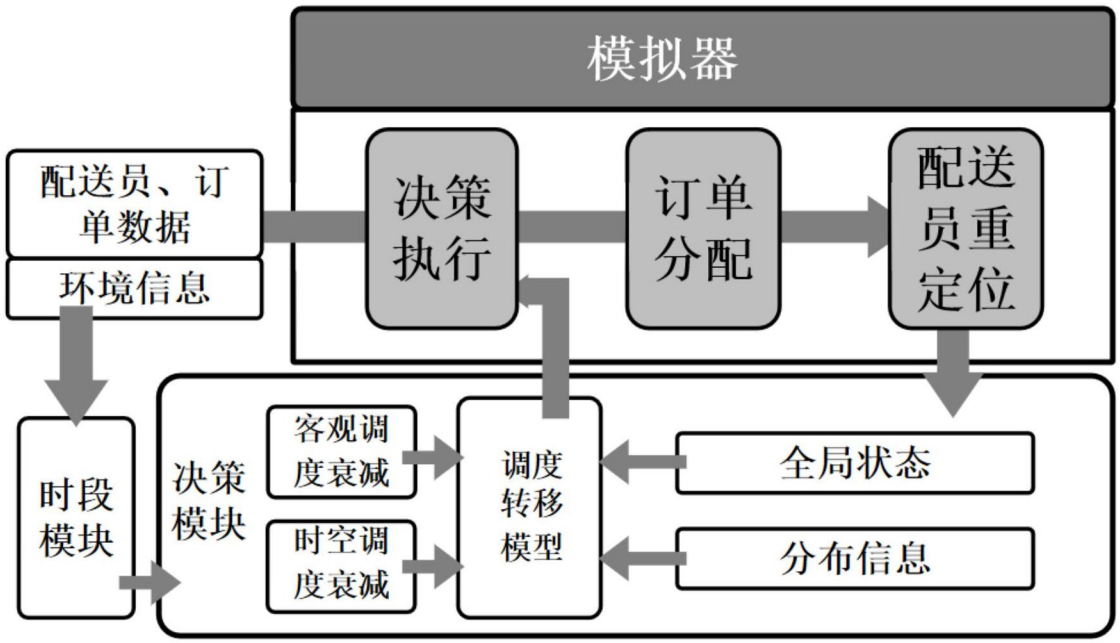

模拟器
配送员、订单数据
环境信息
决策执行
订单分配
配送员重定位
时段模块
决策模块
客观调度衰减
时空调度衰减
调度转移模型
全局状态
分布信息

图1

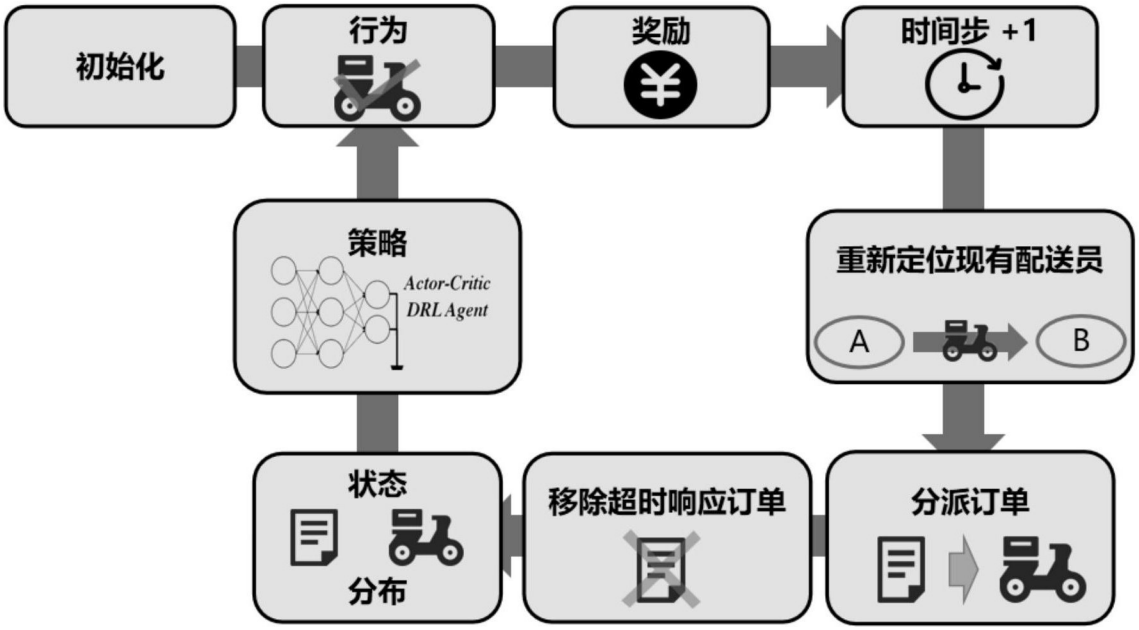

初始化
行为
奖励
时间步 +1
策略
Actor-Critic
DRL Agent
重新定位现有配送员
A
B
状态
分布
移除超时响应订单
分派订单

图2

数据名	意义
mapped_matrix_int	划分后各网格 id
mapped_matrix_type	各网格类型
region_matrix_int	各网格所属区域 id
real_orders	订单数据
courier_dist_time	各时间步配送员数量特征
courier_location_mat	各时间步配送员分布特征
probability	订单采样概率
onoff_courier_location_mat	各时间步配送员上下线变化特征

表 1