(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114819617 A
(43)申请公布日 2022.07.29

(21)申请号 202210431112.7

(22)申请日 2022.04.22

(66)本国优先权数据
202210204157.0 2022.03.03 CN

(71)申请人 北京邮电大学
地址 100876 北京市海淀区西土城路10号

(72)发明人 左兴权 刘英卓 黄海 艾冠群

(74)专利代理机构 北京永创新实专利事务所 11121
专利代理师 易卜

(51)Int.Cl.
G06Q 10/06(2012.01)
G06Q 50/30(2012.01)

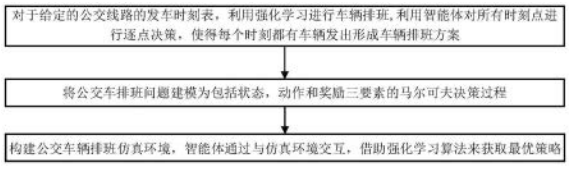
权利要求书2页 说明书7页 附图2页

(54)发明名称

一种基于强化学习的公交车辆排班方法

(57)摘要

本发明公开了一种基于强化学习的公交车辆排班方法,属于公交车辆排班领域,具体为:首先、将某条公交线路的车辆排班问题建模为马尔可夫决策过程,其中包括状态,动作、奖励三个要素。将发车时刻表中每个时刻点作为一个决策点,利用强化学习智能体选取当前可用车辆从该时刻点发出,从而覆盖该时刻点。构建公交车辆排班的仿真环境,通过智能体与仿真环境的交互,利用强化学习算法获得智能体的最优策略。然后,利用智能体对发车时刻表中每个时刻点按时间顺序逐点决策,由此得到公交车排班方案。本发明提供了一种公交车辆的在线调度方法,在发生交通拥堵,交通事故等不确定事件的情况下,能保证全部覆盖发车时刻表,且所用车辆数目少。



CN 114819617 A

1.一种基于强化学习的公交车辆排班方法,其特征在于:

(1)对于给定的公交线路的发车时刻表,利用强化学习进行车辆排班;将公交车辆排班问题建模为马尔可夫决策过程,发车时刻表中每个时刻点作为决策点;对于每个时刻点,由智能体根据控制点中车辆的信息,选取一个车辆在该时刻点从该控制点发出,从而覆盖该时刻点;利用智能体对发车时刻表中的所有时刻点进行逐点决策,使得时刻表中每个时刻都有车辆发出,最终形成车辆排班方案;

(2)车辆排班问题的马尔可夫决策过程模型包括状态、动作和奖励;

状态为智能体的输入,包括:当前时刻点所有可选车辆的信息;动作为智能体的输出,即:选取控制点中的某一车辆在该时刻点发出;奖励包括:车辆数目和车辆等待时间;

(3)构建公交车辆排班仿真环境,智能体通过与仿真环境的交互,借助强化学习算法来获取最优策略。

2.根据权利要求1所述的一种基于强化学习的公交车辆排班方法,其特征在于,所述排班仿真环境为车辆运营过程的模拟;

仿真环境中包含运营时间内任一时刻的车辆行驶信息,具体包括车辆位置、行驶方向、运行时间、工作时间和休息时间,这些信息作为智能体的输入,智能体根据输入产生输出的动作,即选择发出的车辆,动作作用于仿真环境来更新车辆行驶信息;通过智能体与环境的交互,实现车辆运营过程的模拟。

3.根据权利要求1所述的一种基于强化学习的公交车辆排班方法,其特征在于,所述马尔可夫决策过程,其特征在于:

1)状态:对于当前控制点内每个可供选择的车辆,为其构建一个车辆表示向量,该向量包含该车辆与排班相关的行驶信息;可供选择的车辆指当前时刻停靠在该控制点内的可发车的车辆以及还未发出的车辆;由所有可供选择车辆的表示向量构成的矩阵,作为状态;

2)动作:对于每个决策点,智能体的动作为选取一个可供选择的车辆从该时刻表中发出;动作空间中包括所有的可供选择的车辆,其中还未发出的车辆表示为一辆车;

3)奖励:包括主线奖励和支线奖励;根据车辆排班问题的优化目标来构造奖励函数,奖励需要考虑的因素包括:车辆数目,执行奇数行程的车辆数和时刻点覆盖情况。

4.根据权利要求3所述的一种基于强化学习的公交车辆排班方法,其特征在于,所述的奖励,采用主线奖励和支线奖励相结合的奖励方法,支线奖励在每步决策后给出,主线奖励在决策的最后一步给出;

主线奖励包括:1.未使用车辆数目的奖励N_u;2.车辆运行时间方差的惩罚σ_h;3.具有奇数行程的车辆数目惩罚N_o;4.短班车数目的惩罚N_s;5.执行满行程的车辆数目奖励N_f;

主线奖励为这五项的加权和:

$r_m=w_1\times N_u-w_2\times\sigma_h-w_3\times N_o-w_4\times N_s+w_5\times N_f$

其中w_1、w_2、w_3、w_4、w_5分别正实数权值;

支线奖励包括:1.所选车是否为新车r_n;2.所选车的空闲时间在所有可选车辆中的排名r_k;3.车辆包含的行程数为偶数,则奖励r_o;4.所选车是否为长班车r_l;5.存在可用车辆时仍选择发新车,则惩罚r_e;6.车辆执行完行程后的等待时间r_w;

支线奖励为这六项的加权和:

$r_s=-w'_1\times r_n-w'_2\times r_k+w'_3\times r_o+w'_4\times r_l-w'_5\times r_e-w'_6\times r_w$

其中w'_1、w'_2、w'_3、w'_4、w'_5、w'_6为正实数权值。

5.根据权利要求1所述的一种基于强化学习的公交车辆排班方法，其特征在于，所述的强化学习算法包括基于值的强化学习算法，基于策略的强化学习算法和基于Actor-Critic的强化学习算法三类；

在符合问题的约束条件下，利用强化学习算法得到智能体的最优策略，利用不合理动作掩码和不合理动作惩罚两种方式来实现问题的约束条件。

一种基于强化学习的公交车辆排班方法

技术领域

[0001] 本发明属于公交车辆排班领域，特别涉及一种基于强化学习的公交车辆排班方法。

背景技术

[0002] 公交车辆排班是提高公交服务质量的关键一环。

[0003] 目前解决公交车排班问题使用的方法大多是基于精确算法和启发式算法，精确算法虽能 求出最优解，但是对于算力要求较高，无法很好的解决比较复杂的公交车辆排班问题；而启 发式算法虽能解决一些复杂的问题，且能获得一些不错的排班，但由于缺乏实时决策，导致 应对突发情况的能力较差，而且有时无法保证时刻点的全覆盖这一重要要求。

[0004] 由于现有方法都是每次产生一个完整排班方案的离线调度方法，因此无法在时刻点粒度 根据某个时刻点的实时信息进行实时决策。然而，在实际应用场景中，由于城市交通问题的 复杂性，交通拥堵，交通事故等不确定事件是一种很常见的现象，在遇到这样的现象时，原 有的车辆排班方法往往不能很好地应对。

[0005] 因此，亟需提出一种能进行实时决策，同时有良好的突发情况应对能力的公交车辆排班 方法。

发明内容

[0006] 针对上述问题，本发明提供一种基于强化学习的公交车辆排班方法，以实现实时决策， 同时能应对动态事件并提供保证时刻点全覆盖的优秀排班时刻表。

[0007] 所述的基于强化学习的公交车辆排班方法，具体步骤如下：

[0008] 步骤一、对于给定的公交线路的发车时刻表，利用强化学习进行车辆排班(调度)。将公 交车辆排班问题看作序列决策过程，发车时刻表中每个时刻点作为决策点。对于每个时刻点， 由智能体根据控制点中车辆的信息，选取一个车辆在该时刻点从该控制点发出，从而覆盖该 时刻点。利用智能体对发车时刻表中的所有时刻点进行逐点决策，使得时刻表中每个时刻都 有车辆发出，最终形成车辆排班方案。

[0009] 步骤二、将公交车辆排班问题建模为包括状态，动作和奖励三要素的马尔可夫决策过程。

[0010] 1)状态：对于当前控制点内每个可供选择的车辆，为其构建一个车辆表示向量，该向量 包含该车辆与排班相关的行驶信息。可供选择的车辆指当前时刻停靠在该控制点内的可发车 的车辆以及还未发出的车辆。由所有可供选择车辆的表示向量构成的矩阵，作为状态。为保 证强化学习算法的收敛速度，进一步从行空间和列空间两个角度对状态矩阵进行降维。

[0011] 对于行空间的降维过程为：首先，逐个判断各车辆是否符合条件约束。其中约束条件是 指：在当前发车时刻点，该车是否具备发车条件，即是否满足位置，休息时间和工作时间的 约束。然后，将符合条件约束的车辆划分为已发车车辆和未发车车辆，不符合条件约

束的车 单独分类;最后,在某个发车时刻点符合条件约束的未发车车辆可缩减为一行,而不符合条 件约束的车用来填充状态矩阵,保证矩阵的形状不变,因此最终缩减了状态矩阵的行数。

[0012] 对于列空间降维为:构造更精简的车辆表示向量,来减少状态矩阵的列数。通过分析车 辆行驶信息中与排班最相关的信息,挑选出剩余可行驶时间,休息时间,趟次数,剩余可工 作时间和车辆类型五项信息。

[0013] 2)动作:即智能体所选的车辆。

[0014] 对于符合约束条件且已发车的车辆,直接发出该车辆。

[0015] 对于符合条件约束且未发车的车辆,维护一个新车堆,当选择到该类车时,则从新车堆 中随机选出一辆车发出;

[0016] 3)奖励:采用主线奖励和支线奖励结合的策略,使用公交车辆排班问题的目标来构造奖 励函数,目标包括使用的车辆数目,奇数行程数的车辆数目和时刻点覆盖情况三项。

[0017] 在奖励设计中,本发明采用了主线奖励和支线奖励相结合的奖励设置方法,支线奖励在 每步决策后给出,主线奖励在决策的最后一步给出。

[0018] 主线奖励包括:1.未使用车辆数目的奖励N_u;2.车辆运行时间方差的惩罚σ_h;3.具有奇数 行程的车辆数目惩罚N_o;4.短班车数目的惩罚N_s;5.执行满行程的车辆数目奖励N_f;

[0019] 主线奖励为这五项的加权和:

[0020] $r_m=w_1\times N_u-w_2\times\sigma_h-w_3\times N_o-w_4\times N_s+w_5\times N_f$

[0021] 其中w_1、w_2、w_3、w_4、w_5分别正实数权值。

[0022] 支线奖励包括:1.所选车是否为新车r_n;2.所选车的空闲时间在所有可选车辆中的排名r_k; 3.车辆包含的行程数为偶数,则奖励r_o;4.所选车是否为长班车r_l;5.存在可用车辆时仍选择 发新车,则惩罚r_e;6.车辆执行完行程后的等待时间r_w。

[0023] 支线奖励为这六项的加权和:

[0024] $r_s=-w'_1\times r_n-w'_2\times r_k+w'_3\times r_o+w'_4\times r_l-w'_5\times r_e-w'_6\times r_w$

[0025] 其中w'_1、w'_2、w'_3、w'_4、w'_5、w'_6为正实数权值。

[0026] 步骤三、构建公交车辆排班仿真环境,智能体通过与仿真环境的交互,借助强化学习算 法来获取最优策略。

[0027] 仿真环境为车辆运营过程的模拟。仿真环境中包含运营时间内任一时刻的车辆行驶信息, 具体包括车辆位置、行驶方向、运行时间、工作时间、休息时间等。这些信息作为智能体的 输入,智能体根据输入产生输出的动作(即选择发出的车辆),动作作用于仿真环境来更新车 辆行驶信息。通过智能体与环境的交互,实现车辆运营过程的模拟。

[0028] 智能体通过与仿真环境交互获取训练数据,并不断进行学习。所述的强化学习算法包括 基于值的强化学习算法,基于策略的强化学习算法、基于Actor-Critic的强化学习算法三类。在符合问题的约束条件下,利用强化学习算法得到智能体的最优策略。利用不合理动作掩码 (invalid action masking)和不合理动作惩罚(invalid action penalty)两种方式来实现问题的 约束条件。

[0029] 本发明与现有技术相比具有以下优点和有益效果:

[0030] 1)一种基于强化学习的公交车辆排班方法,相对传统的方法,本发明将公交车辆

排班问 题建模为马尔可夫决策过程,通过在线方法解决了离线方法无法应对交通拥堵,车辆故障等 不确定事件的问题,具有实时决策和实时应对突发情况调整的能力,且能提供保证时刻点全 覆盖的优秀排班时刻表。

[0031] 2)一种基于强化学习的公交车辆排班方法,通过对状态空间分别进行基于车辆分类降维 的行压缩和基于车辆表示向量精简化的列压缩,对整个状态空间进行了缩减,保证了强化学 习算法的收敛性。

[0032] 3)一种基于强化学习的公交车辆排班方法,采用主线奖励和支线奖励相结合的方法奖励 设置,使得公交车辆排班的多个目标能得到兼顾和平衡。

[0033] 4)一种基于强化学习的公交车辆排班方法,采用基于值的强化学习算法,基于策略的强 化学习算法和基于Actor-Critic的强化学习算法三类算法对于该问题分别进行解决。

附图说明

[0034] 图1为本发明一种基于强化学习的公交车辆排班方法的流程图;

[0035] 图2为本发明一种基于强化学习的公交车辆排班方法的结构框架示意图;

[0036] 图3为本发明一种基于强化学习的公交车辆排班方法中的状态矩阵构造图;

[0037] 图4为本发明所述的D3QN强化学习算法中的主网络架构图。

具体实施方式

[0038] 为了使本发明的目的、技术方案及优点更加清楚明白,以下参照附图并举实施例,对本 发明作进一步详细说明。

[0039] 本发明公开了一种基于强化学习的公交车辆排班方法,其包括:依据公交车辆排班问题 的运行逻辑构建出一个仿真环境;将公交车辆排班控制问题建模为一个马尔可夫决策过程模 型,并定义其中的状态,动作以及奖励函数;利用强化学习算法模型结合约束机制求解出最 优策略,利用最优策略进行公交车辆排班时车辆的实时选取。本发明改进了原有方法无法实 时决策,无法应对特殊情况,难以实现时刻点全覆盖等问题,并在车辆使用数量,奇数趟次 车辆数等主要指标上有一定的优势。

[0040] 所述的基于强化学习的公交车辆排班方法,如图1所示,具体步骤如下:

[0041] 步骤一、对于给定的公交线路的发车时刻表,利用强化学习进行车辆排班(调度)。将公 交车辆排班问题建模为序列决策过程,发车时刻表中每个时刻点作为决策点。对于每个时刻 点,由智能体根据控制点中车辆的信息,选取一个车辆在该时刻点从该控制点发出,从而覆 盖该时刻点。利用智能体对发车时刻表中的所有时刻点进行逐点决策,使得时刻表中每个时 刻都有车辆发出,最终形成车辆排班方案。

[0042] 步骤二、将公交车辆排班问题建模为包括状态,动作和奖励三要素的马尔可夫决策过程。

[0043] 1)状态是对当前决策点排班所需信息的建模,包括:

[0044] 对于当前控制点内每个可供选择的车辆,为其构建一个车辆表示向量,该向量包含该车 辆与排班相关的行驶信息。可供选择的车辆指当前时刻停靠在该控制点内的可发车的车辆以 及还未发出的车辆。由所有可供选择车辆的表示向量构成的矩阵,作为状态。

为保证强化学 习算法的收敛速度,进一步从行空间和列空间两个角度对状态矩阵进行降维。

[0045] 对于行空间降维,首先,逐个判断各车辆是否符合条件约束。其中约束条件是指:在当 前发车时刻点,该车是否具备发车条件,即是否满足位置,休息时间和工作时间的约束。然 后,将符合条件约束的车辆划分为已发车车辆和未发车车辆,不符合条件约束的车单独分类; 经过车辆归类降维后,由于符合条件约束的新车的优先级相同,该类车可以缩减为一行;另 外,在某个发车时刻点符合条件约束的旧车数也远小于车辆总数,因此,这种降维方法可减 少状态矩阵的行数;

[0046] 对于列空间降维是通过构造更精简的车辆表示向量,来减少状态矩阵的列数。通过分析 车辆行驶信息中与排班最相关的信息,挑选出剩余可行驶时间,休息时间,趟次数,剩余可 工作时间,车辆类型五项信息。

[0047] 2)动作即智能体所选的车辆。

[0048] 对于符合约束条件且已发车的车辆,直接发出该车辆。

[0049] 对于符合条件约束且未发车的车辆,维护一个新车堆,当选择到该类车时,则从新车堆 中随机选出一辆车发出。

[0050] 3)奖励是强化学习智能体学习驱动力;

[0051] 使用公交车辆排班问题的目标来构造奖励函数,目标包括使用的车辆数目,奇数行程数 的车辆数目和时刻点覆盖情况三项;

[0052] 在奖励设计中,本发明采用了主线奖励和支线奖励相结合的奖励设置方法,主线奖励是 在决策序列结束后对于整个决策序列的评估奖励;支线奖励是在决策过程中每一步的评估奖 励,负责引导智能体更快达到主线奖励的目标。

[0053] 主线奖励考虑了整个排班方案的总车辆数、奇偶行程车辆数等评估指标,支线奖励考虑 了当前决策点所选车辆的合适程度。

[0054] 主线奖励包括:1.未使用车辆数目的奖励N_u;2.车辆运行时间方差的惩罚σ_h;3.具有奇数 行程的车辆数目惩罚N_o;4.短班车数目的惩罚N_s;5.执行满行程的车辆数目奖励N_f;

[0055] 主线奖励为这五项的加权和:

[0056] $r_m = w_1 \times N_u - w_2 \times \sigma_h - w_3 \times N_o - w_4 \times N_s + w_5 \times N_f$

[0057] 其中w_1、w_2、w_3、w_4、w_5分别正实数权值。

[0058] 支线奖励包括:1.所选车是否为新车r_n;2.所选车的空闲时间在所有可选车辆中的排名r_k; 3.车辆包含的行程数为偶数,则奖励r_o;4.所选车是否为长班车r_l;5.存在可用车辆时仍选择 发新车,则惩罚r_e;6.车辆执行完行程后的等待时间r_w。

[0059] 支线奖励为这六项的加权和:

[0060] $r_s = -w'_1 \times r_n - w'_2 \times r_k + w'_3 \times r_o + w'_4 \times r_l - w'_5 \times r_e - w'_6 \times r_w$

[0061] 其中w'_1、w'_2、w'_3、w'_4、w'_5、w'_6为正实数权值。

[0062] 步骤三、构建公交车辆排班仿真环境,智能体通过与仿真环境的交互,借助强化学习算 法来获取最优策略。

[0063] 仿真环境为车辆运营过程的模拟。仿真环境中包含运营时间内任一时刻的车辆行驶信息, 具体包括车辆位置、行驶方向、运行时间、工作时间、休息时间等。这些信息作为智能体的 输入,智能体根据输入产生输出的动作(即选择发出的车辆),动作作用于仿真环境

来更新车 辆行驶信息。通过智能体与环境的交互,实现车辆运营过程的模拟。

[0064] 在符合问题约束的条件下,利用强化学习算法求解最优策略,本发明利用的强化学习算 法包括基于值的强化学习算法,基于策略的强化学习算法和基于Actor-Critic的强化学习算法 三类。

[0065] 而为了保证智能体所选动作符合约束条件,本发明利用不合理动作掩码(invalid action masking)和不合理动作惩罚(invalid action penalty)两种方式来实现问题的约束条件。

[0066] 实施例:

[0067] 如图2所示,是一种基于强化学习的公交排班方法的结构框架,包括环境模型S1,马尔 可夫决策过程模型S2和强化学习算法模型S3三部分;其中环境模型S1根据强化学习算法 模型S3所执行的动作action,返回更新后的状态state和对应的奖励reward;而动作,状态和 奖励三要素则组成了马尔科夫决策过程模型S2的主要部分。在训练和学习过程中,以马尔可 夫决策过程S2为基本的学习框架,强化学习算法模型S3通过不断与环境模型S1进行交互, 获取训练数据,并利用这些训练数据对网络进行训练,训练完成后即可对训练场景进行排班 结果评估。

[0068] 环境模型S1:对公交车排班方法逻辑进行建模,构建环境模型。

[0069] 马尔可夫决策过程模型S2:将公交车辆排班问题建模为一个马尔可夫决策过程,并定义 其中的状态,动作和奖励。

[0070] 强化学习算法模型S3:利用D3QN强化学习方法求解最优策略,并利用最优策略进行车 辆选择。

[0071] 本实施例中,上述的环境模型S1包括:

[0072] 环境用于与强化学习智能体进行交互的公交车辆排班环境逻辑,在本发明中,采用了gym 库的环境模型架构进行实现,需要实现以下功能,包括:

[0073] 公交车辆排班有关信息的保存,包括每辆车的行驶信息(是否开始运行,是否正在运行, 运行方向,当前行程已运行时间,总行驶时间,休息时间,行程数,出发时间,当前时间, 车辆类型),发车时间表,一天不同小时内单行程运行时间和当前时刻点等等。

[0074] 根据智能体所输出的动作,通过模拟车辆运行过程,对公交车辆排班有关的信息进行更 新。例如,在某个决策点,选择车辆V_1作为发车车辆,则当前决策步需要模拟从当前决策点 到下一个决策点之间这段时间整个环境的变化,包括每辆车的运行信息的更新和当前时刻点 等等。

[0075] 根据智能体所输出的动作,产生针对该动作的反馈。通过对智能体所做动作对于最终优 化目标的作用,对智能体进行奖励或惩罚,例如,如果选择的车辆是旧车而且该车已经休息 了很长时间,则会给智能体一个正反馈。

[0076] 本实施例中,上述的马尔可夫决策过程S2模型包括:

[0077] 马尔可夫决策过程模型中的三要素:状态,动作和奖励。

[0078] 状态要素S21包括状态矩阵,如图3所示,在状态设计中,由于车辆数目较多,会造成 强化学习训练困难等问题,例如,假如初始阶段共有50辆车,而每辆车对应的车辆向量为一 个10维向量,则整个状态矩阵的大小为500维,因此需要对状态空间进行降维,本发明采用 对于行空间和列空间分别进行降维的方法。

[0079] 针对行空间降维，提出一种依据车辆类型进行状态降维的方法，通过将车辆分为符合条 件约束的已运行过的车，符合条件约束的新车和不符合条件约束的车三类，其中，符合条件 约束的新车可以归为一行，符合条件约束的已运行的车也不会太多，最终本发明将原来的50 辆车使用该方法降维到了16辆；

[0080] 针对列空间降维，通过分析列空间对于决策最相关的信息，从原来的10维信息(是否开 始运行，是否正在运行，运行方向，当前行程已运行时间，总行驶时间，休息时间，行程数， 出发时间，当前时间，车辆类型)中挑选并总结出了剩余可行驶时间，休息时间，趟次数， 剩余可工作时间，车辆类型五项信息，对状态矩阵的列空间进行了降维。

[0081] 动作要素S22包括：

[0082] 在动作设计中，对于符合条件约束的已发车车辆，可直接发出；对于符合条件约束的未 发车车辆，本发明需要维护一个新车堆，当选择到该类车时，则从新车堆中随机选出一辆车。

[0083] 奖励要素S23包括：

[0084] 在奖励设计中，采用了主线奖励和支线奖励相结合的奖励设置方法。

[0085] 主线奖励包括：1.未使用车辆数目的奖励N_u；2.车辆运行时间方差的惩罚σ_h；3.具有奇数 行程的车辆数目惩罚N_o；4.短班车数目的惩罚N_s；5.执行满行程的车辆数目奖励N_f；

[0086] 主线奖励为这五项的加权和：

[0087] $r_m=w_1\times N_u-w_2\times\sigma_h-w_3\times N_o-w_4\times N_s+w_5\times N_f$

[0088] 其中w_1、w_2、w_3、w_4、w_5分别正实数权值。

[0089] 支线奖励包括：1.所选车是否为新车r_n；2.所选车的空闲时间在所有可选车辆中的排名r_k； 3.车辆包含的行程数为偶数，则奖励r_o；4.所选车是否为长班车r_l；5.存在可用车辆时仍选择 发新车，则惩罚r_e；6.车辆执行完行程后的等待时间r_w。

[0090] 支线奖励为这六项的加权和：

[0091] $r_s=-w'_1\times r_n-w'_2\times r_k+w'_3\times r_o+w'_4\times r_l-w'_5\times r_e-w'_6\times r_w$

[0092] 其中w'_1、w'_2、w'_3、w'_4、w'_5、w'_6为正实数权值。

[0093] 在实施例中，上述的强化学习算法模型S3具体可以分为基于值的强化学习方法，基于策 略的强化学习方法和基于Actor-Critic的强化学习方法。

[0094] 本发明尝试的强化学习方法主要包括基于值的强化学习方法中的DQN系列方法(Double DQN，Dueling DQN，D3QN，DRQN)等等，基于策略的强化学习方法中的TRPO，PPO等等，基于Actor-Critic强化学习方法中的DDPG，TD3，SAC等等。

[0095] 以D3QN强化学习算法为例求解最优策略，具体为：

[0096] 初始化replay buffer，容量为N，用来存储进行Q网络训练的样本。

[0097] 初始化主网络(main network)和目标网络(target network)两个网络，并随机初始化参 数。

[0098] 将获取的信息组成的状态矩阵s输入主网络，得到该状态下对应的Q值向量，并采用 epsilon-greedy的方法选择动作a，并从环境中获取奖励r，每一次这样的状态转移记作一个时 间步t，把每个时间步中获取的数据(s,a,s',r)放入replay buffer。

[0099] 从replay buffer中采样一个batch的数据，以目标网络计算的目标Q值作为标签，以主网 络计算的Q值为预测值，通过MSE计算损失，并通过Adam的方法进行参数更新。

[0100] 在实施例中,所述的D3QN强化学习算法中的主网络S31包括:

[0101] 以展平后的状态矩阵为输入,以Q值向量作为输出,构建出四层的网络结构。另外,还 需要在网络的最后一层加入约束添加模型,从而对不合理动作进行限制。具体的网络结构如 图4中间Neural Network部分所示。

[0102] 在实施例中,上述的约束添加模型S311包括:

[0103] 不合理动作掩码(invalid action masking)方法是给Q网络的输出加一层掩码,使得不合 理的动作对应的Q值被置为一个很小的负值以使该动作无法被选择。如图4右边Invalid Action Masking部分,A1,A2,A3,A4四个动作的值分别200,500,400,250,如果没有约束,应该选择对 应值最大的动作A2,但是加入了约束部分之后,由于动作A2不符合约束,因此被置为很小 的负值,因此,最终选择了动作A3。

[0104] 不合理动作惩罚(invalid action penalty)方法是对奖励进行处理,当强化学习智能体选 择到了不符合约束条件的车辆时,会反馈给智能体一个很小的负奖励,进而引导智能体在后 面规避这种情况。

[0105] 本发明提供的上述基于强化学习的公交车辆排班方法,通过强化学习算法智能体与公交 排班环境模型进行交互,获取各种各样的训练数据并存储在replay buffer中,强化学习智能 体通过这些数据进行学习,以做出更加优秀的决策。与现有技术相比,本发明的强化学习方 法可以进行实时决策,并可以处理一些特殊的场景,具有一定的自适应性。而且本方法不仅 可以用于公交车辆排班,也可以用于地铁排班,专车排班等一系列类似的排班问题。

[0106] 上所述仅为本发明的较佳实施例而已,并不用以限制本发明,凡在本发明的精神和原则 之内,所做的任何修改、等同替换、改进等,均应包含在本发明保护的范围之内。

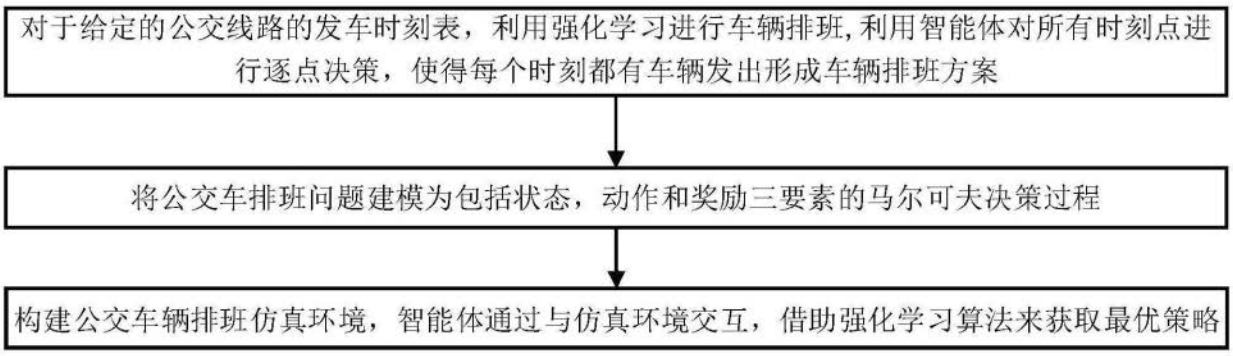

对于给定的公交线路的发车时刻表，利用强化学习进行车辆排班,利用智能体对所有时刻点进行逐点决策，使得每个时刻都有车辆发出形成车辆排班方案
将公交车排班问题建模为包括状态，动作和奖励三要素的马尔可夫决策过程
构建公交车辆排班仿真环境，智能体通过与仿真环境交互，借助强化学习算法来获取最优策略

图1

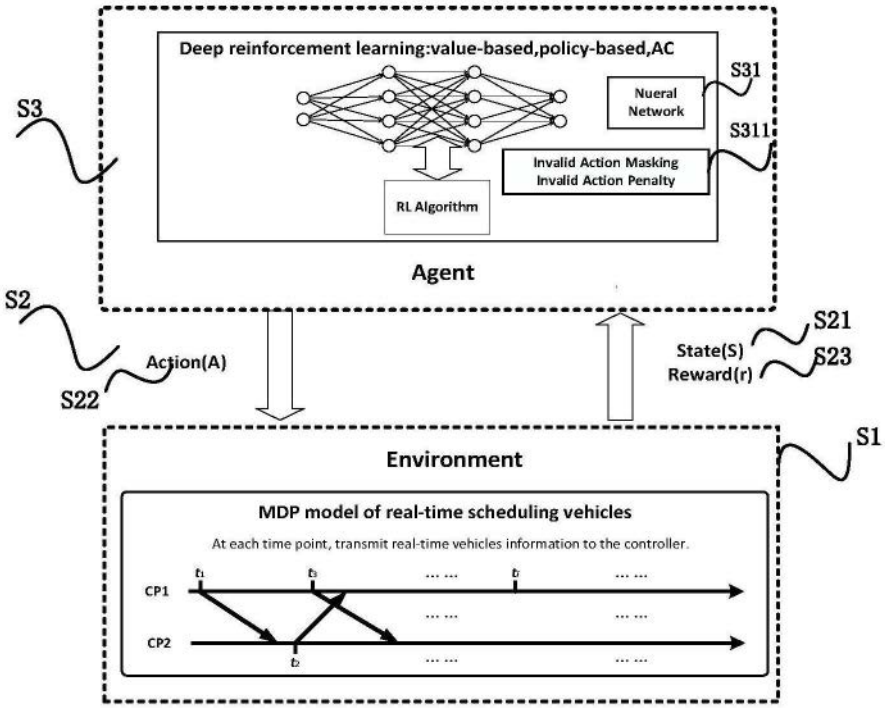

Deep reinforcement learning:value-based,policy-based,AC
Nueral Network
S31
S311
Invalid Action Masking
Invalid Action Penalty
RL Algorithm
S3
Agent
S2
Action(A)
S22
S21
State(S)
Reward(r)
S23
S1
Environment
MDP model of real-time scheduling vehicles
At each time point, transmit real-time vehicles information to the controller.
CP1
CP2

图2

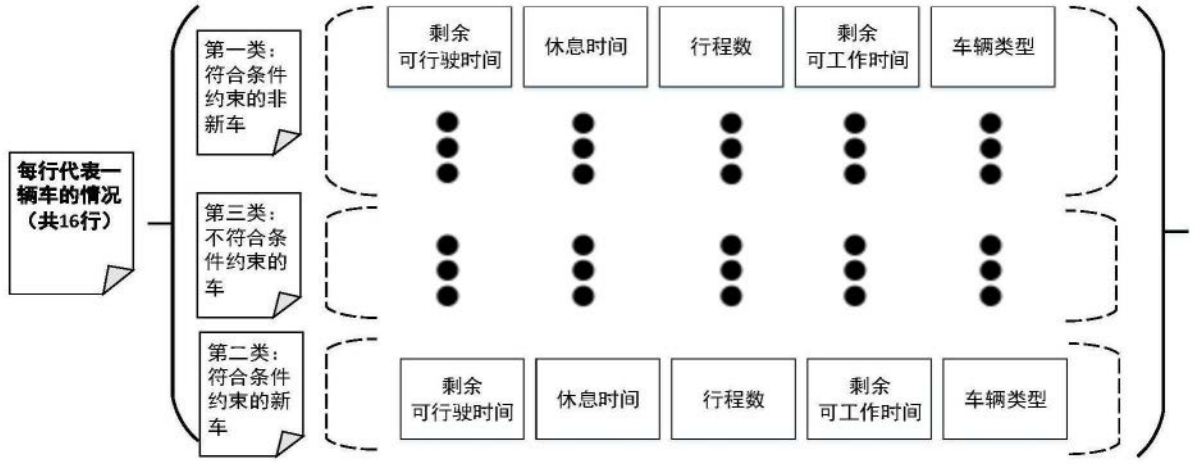

每行代表一辆车的情况（共16行）
第一类：符合条件约束的非新车
第三类：不符合条件约束的车
第二类：符合条件约束的新车
剩余可行驶时间
休息时间
行程数
剩余可工作时间
车辆类型
剩余可行驶时间
休息时间
行程数
剩余可工作时间
车辆类型

图3

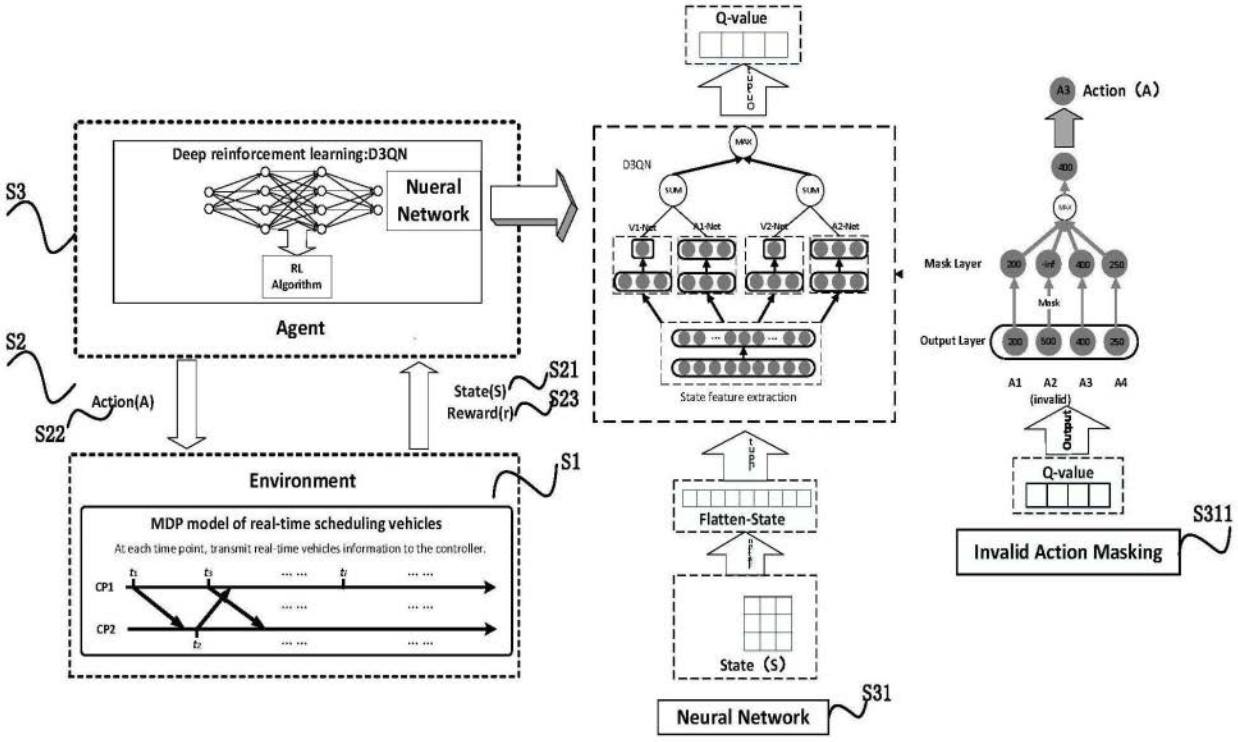

Deep reinforcement learning:D3QN
Nueral Network
RL Algorithm
Agent
S3
S2
Action(A)
S22
State(S)
Reward(r)
S21
S23
Environment
S1
MDP model of real-time scheduling vehicles
At each time point, transmit real-time vehicles information to the controller.
CP1
CP2
Q-value
MAX
D3QN
SUM
SUM
V1-Net
A1-Net
V2-Net
A2-Net
State feature extraction
Flatten-State
State（S）
Neural Network
S31
A3 Action（A）
400
MAX
Mask Layer
200
-inf
400
250
Mask
Output Layer
200
500
400
250
A1
A2
A3
A4
(invalid)
Output
Q-value
Invalid Action Masking
S311

图4