(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115330095 A
(43)申请公布日 2022.11.11

(21)申请号 202211256537.5

(22)申请日 2022.10.14

(71)申请人 青岛慧拓智能机器有限公司
地址 266000 山东省青岛市高新技术产业开发区火炬路100号盘谷创客空间D座206-1房间

(72)发明人 张晓彤 史磊石 张振良

(74)专利代理机构 北京中强智尚知识产权代理有限公司 11448
专利代理师 陈宇楠

(51)Int.Cl.
G06Q 10/04(2012.01)
G06Q 10/06(2012.01)
G06Q 50/02(2012.01)
G06K 9/62(2022.01)

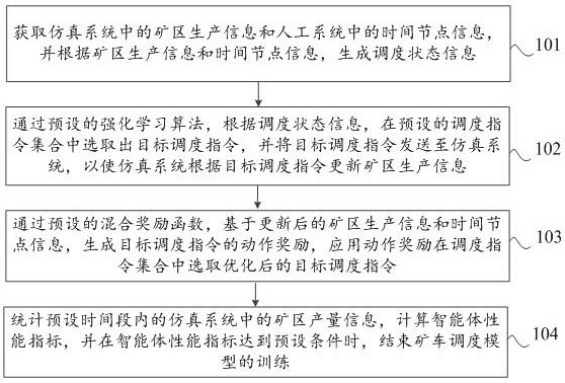
权利要求书4页 说明书17页 附图6页

(54)发明名称

矿车调度模型训练方法、装置、芯片、终端、设备及介质

(57)摘要

本发明公开了一种矿车调度模型训练方法、装置、芯片、终端、设备及介质，涉及车辆调度及智慧矿山技术领域。其中，所述方法应用于矿车调度模型训练装置中，该装置包括调度智能体和交互环境，交互环境包括仿真系统和人工系统，该方法包括：根据仿真系统中的矿区生产信息和人工系统中的时间节点信息，生成调度状态信息；根据调度状态信息，在调度指令集合中选取目标调度指令，将目标调度指令发送至仿真系统；通过混合奖励函数，基于矿区生产信息和时间节点信息，生成目标调度指令的动作奖励；根据矿区产量信息，计算智能体性能指标，在智能体性能指标达到预设条件时，结束矿车调度模型的训练。上述方法能够提高奖励获取的及时性，降低训练时间成本。



CN 115330095 A

1.一种矿车调度模型训练方法,应用于矿车调度模型训练装置中,其特征在于,所述矿车调度模型训练装置包括调度智能体和交互环境,所述交互环境包括仿真系统和人工系统,所述方法包括:

S1:获取所述仿真系统中的矿区生产信息和所述人工系统中的时间节点信息,并根据所述矿区生产信息和所述时间节点信息,生成调度状态信息;

S2:通过预设的强化学习算法,根据所述调度状态信息,在预设的调度指令集合中选取出目标调度指令,并将所述目标调度指令发送至所述仿真系统,以使所述仿真系统根据所述目标调度指令更新所述矿区生产信息;

S3:通过预设的混合奖励函数,基于所述更新后的矿区生产信息和所述时间节点信息,生成所述目标调度指令的动作奖励,应用所述动作奖励在所述调度指令集合中选取优化后的目标调度指令,重复执行步骤S1至步骤S3,不断选取优化后的目标调度指令,并将所述优化后的目标调度指令发送至所述仿真系统,以使所述仿真系统根据所述优化后的目标调度指令更新所述矿区生产信息;

S4:统计预设时间段内的所述仿真系统中的矿区产量信息,计算智能体性能指标,并在所述智能体性能指标达到预设条件时,结束矿车调度模型的训练。

2.根据权利要求1所述的方法,其特征在于,在所述获取所述仿真系统中的矿区生产信息和所述人工系统中的时间节点信息之前,所述方法还包括:

在所述仿真系统中模拟生成所述矿区生产信息,其中,所述矿区生产信息包括仿真路网信息、至少一个仿真装载设备、每个所述仿真装载设备的设备信息和设备状态、至少一个仿真卸载设备、每个所述仿真卸载设备的设备信息和设备状态、至少一个仿真矿车、每个所述仿真矿车的车辆信息和车辆状态、以及每个所述仿真装载设备和每个所述仿真卸载设备之间的行驶时间中的至少一种信息;

在所述仿真系统中的仿真矿车发送车辆调度请求时,根据所述矿区生产信息,在所述人工系统中生成针对所述仿真矿车的时间节点信息,其中,所述时间节点信息包括所述调度指令集合中的每个所述调度指令对应的行驶时间、每个所述调度指令对应的预期等待时间、所述仿真系统中每个仿真装载设备的剩余服务时间,以及仿真系统中每个仿真卸载设备的剩余服务时间中的至少一种信息。

3.根据权利要求1或2所述的方法,其特征在于,所述获取所述仿真系统中的矿区生产信息和所述人工系统中的时间节点信息,并根据所述矿区生产信息和所述时间节点信息,生成调度状态信息,包括:

在所述仿真系统中的仿真矿车发送车辆调度请求时,获取所述仿真系统中的矿区生产信息和所述人工系统中的时间节点信息,其中,所述车辆调度请求在所述仿真矿车的车辆状态更新为装载完成状态或卸载完成状态时发送;

根据所述矿区生产信息和所述人工系统中的时间节点信息,生成针对所述仿真矿车的调度状态信息,其中,所述调度状态信息包括所述仿真矿车的位置信息、所述调度指令集合中的每个调度指令的动作可用性信息、每个所述调度指令对应的行驶时间、每个所述调度指令对应的预期等待时间、所述仿真系统中每个仿真装载设备的剩余服务时间和故障信息,以及仿真系统中每个仿真卸载设备的剩余服务时间和故障信息中的至少一种信息。

4.根据权利要求3所述的方法,其特征在于,所述通过预设的强化学习算法,根据所述

调度状态信息，在预设的调度指令集合中选取出目标调度指令，并将所述目标调度指令发送至所述仿真系统，以使所述仿真系统根据所述目标调度指令更新所述矿区生产信息，包括：

针对所述调度指令集合中的每个调度指令，通过预设的价值函数，计算每个所述调度指令在所述调度状态信息下的价值数值，其中，所述调度指令由所述仿真矿车的出发地和目的地组成；

将数值最大的所述价值数值对应的调度指令确定为目标调度指令，并将所述目标调度指令发送至所述仿真系统中的仿真矿车中；

在所述仿真矿车执行完成所述目标调度指令时，更新所述矿区生产信息。

5.根据权利要求4所述的方法，其特征在于，所述应用所述动作奖励在所述调度指令集合中选取优化后的目标调度指令，包括：

根据所述目标调度指令的动作奖励，对所述价值函数进行更新，得到优化后的价值函数，应用所述优化后的价值函数在所述调度指令集合中选取优化后的目标调度指令。

6.根据权利要求2所述的方法，其特征在于，所述混合奖励函数由人工奖励函数和仿真奖励函数组成；则所述通过预设的混合奖励函数，基于所述更新后的矿区生产信息和所述时间节点信息，生成所述目标调度指令的动作奖励，包括：

在所述时间节点信息中，提取出所述目标调度指令对应的行驶时间和预期等待时间，并根据所述行驶时间和所述预期等待时间之和，得到所述目标调度指令对应的行程时间；

将所述目标调度指令对应的行程时间输入至所述人工奖励函数中，得到所述目标调度指令的人工奖励值；

根据所述更新后的矿区生产信息，计算所述目标调度指令从执行开始至执行完成之间的仿真装载设备利用率；

将所述仿真装载设备利用率输入至所述仿真奖励函数中，得到所述目标调度指令的仿真奖励值；

根据所述人工奖励值和所述仿真奖励值，得到所述目标调度指令的动作奖励。

7.根据权利要求6所述的方法，其特征在于，所述根据所述人工奖励值和所述仿真奖励值，得到所述目标调度指令的动作奖励，包括：

根据所述更新后的矿区生产信息，计算智能体性能指标，并根据所述智能体性能指标，确定人工奖励权重系数；

根据所述人工奖励权重系数，计算仿真奖励权重系数，其中，所述人工奖励权重系数和所述仿真奖励权重系数的和为预设值；

根据所述人工奖励权重系数与所述人工奖励值的乘积与所述仿真奖励权重系数与所述仿真奖励值的乘积的和值，得到所述目标调度指令的动作奖励。

8.根据权利要求7所述的方法，其特征在于，所述根据所述更新后的矿区生产信息，计算智能体性能指标，并根据所述智能体性能指标，确定人工奖励权重系数，包括：

在所述更新后的矿区生产信息中，提取出预设时间段内的矿区产量信息，并将所述预设时间段内的矿区产量信息划分为多个子时间段内的矿区产量信息；

对所述多个子时间段内的矿区产量信息进行线性拟合，得到性能指标拟合斜率，并计算所述多个子时间段内的矿区产量信息的平均值，得到性能指标均值；

判断所述性能指标拟合斜率是否小于预设的斜率阈值，并判断所述性能指标均值是否大于预设的性能指标阈值，其中，所述斜率阈值为负值；

若所述性能指标拟合斜率小于所述斜率阈值，则对所述人工奖励权重系数进行递增计算；

若所述性能指标均值小于等于所述性能指标阈值，则对所述人工奖励权重系数进行递增计算；

若所述性能指标均值大于所述性能指标阈值，则对所述人工奖励权重系数进行递减计算。

9.根据权利要求1所述的方法，其特征在于，所述统计预设时间段内的所述仿真系统中的矿区产量信息，计算智能体性能指标，并在所述智能体性能指标达到预设条件时，结束矿车调度模型的训练，包括：

在所述仿真系统中的矿区生产信息中，提取出预设时间段内的矿区产量信息，并将所述预设时间段内的矿区产量信息划分为多个子时间段内的矿区产量信息；

计算所述多个子时间段内的矿区产量信息的平均值，得到性能指标均值，计算所述多个子时间段的矿区产量信息与所述性能指标均值之间的偏差值；

当所述偏差值小于预设的偏差阈值时，判定矿车调度模型训练完成，并结束所述矿车调度模型的训练。

10.一种矿车调度模型训练装置，其特征在于，所述矿车调度模型训练装置包括调度智能体和交互环境，其中，所述交互环境包括仿真系统和人工系统，所述调度智能体包括：

状态模块，用于获取所述仿真系统中的矿区生产信息和所述人工系统中的时间节点信息，并根据所述矿区生产信息和所述时间节点信息，生成调度状态信息；

动作模块，用于通过预设的强化学习算法，根据所述调度状态信息，在预设的调度指令集合中选取出目标调度指令，并将所述目标调度指令发送至所述仿真系统，以使所述仿真系统根据所述目标调度指令更新所述矿区生产信息；

奖励模块，用于通过预设的混合奖励函数，基于所述更新后的矿区生产信息和所述时间节点信息，生成所述目标调度指令的动作奖励，应用所述动作奖励在所述调度指令集合中选取优化后的目标调度指令，重复执行所述状态模块、所述动作模块和所述奖励模块的步骤，不断选取优化后的目标调度指令，并将所述优化后的目标调度指令发送至所述仿真系统，以使所述仿真系统根据所述优化后的目标调度指令更新所述矿区生产信息；

评价模块，统计预设时间段内的所述仿真系统中的矿区产量信息，计算智能体性能指标，并在所述智能体性能指标达到预设条件时，结束矿车调度模型的训练。

11.一种芯片，其特征在于，所述芯片包括至少一个处理器和通信接口，所述通信接口和所述至少一个处理器耦合，所述至少一个处理器用于运行计算机程序或指令，以实现如权利要求1-9中任一项所述的矿车调度模型训练方法。

12.一种终端，其特征在于，所述终端包括如权利要求10所述的矿车调度模型训练装置。

13.一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，其特征在于，所述计算机程序被处理器执行时实现权利要求1至9中任一项所述的方法的步骤。

14.一种计算机可读存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现权利要求1至9中任一项所述的方法的步骤。

矿车调度模型训练方法、装置、芯片、终端、设备及介质

技术领域

[0001] 本发明涉及车辆调度及智慧矿山技术领域，尤其是涉及一种矿车调度模型训练方法、装置、芯片、终端、设备及介质。

背景技术

[0002] 在露天矿山的车辆调度场景中，基于强化学习的车辆调度模型可输出车辆调度策略，车辆调度策略可以使矿区车辆车辆有序的执行运输任务。其中，强化学习(Reinforcement Learning,RL)，又称再励学习、评价学习或增强学习，是机器学习的范式和方法论之一，用于描述和解决智能体(agent)在与环境的交互过程中通过学习策略以达成回报最大化或实现特定目标的问题。

[0003] 在现有技术中，基于强化学习的车辆调度模型在训练的过程中，智能体在下达调度指令后无法及时得到对应的环境奖励。通常来说，智能体需要下发几个调度指令甚至下发几十个调度指令，并在矿区车辆在抵达目的地或完成调度任务后，才能够得到信息反馈，这导致来自环境的奖励反馈十分稀疏。而为了得到环境奖励，在训练基于强化学习的车辆调度模型时往往需要付出极高的训练时间代价，这导致车辆调度模型的训练效率极低，训练成本很高。

发明内容

[0004] 有鉴于此，本申请提供了一种矿车调度模型训练方法、装置、芯片、终端、设备及介质，主要目的在于解决基于强化学习的车辆调度模型的训练效率低且训练成本高的技术问题。

[0005] 根据本发明的第一个方面，提供了一种矿车调度模型训练方法，应用于矿车调度模型训练装置中，所述矿车调度模型训练装置包括调度智能体和交互环境，所述交互环境包括仿真系统和人工系统，所述方法包括：

S1:获取所述仿真系统中的矿区生产信息和所述人工系统中的时间节点信息，并根据所述矿区生产信息和所述时间节点信息，生成调度状态信息；

S2:通过预设的强化学习算法，根据所述调度状态信息，在预设的调度指令集合中选取出目标调度指令，并将所述目标调度指令发送至所述仿真系统，以使所述仿真系统根据所述目标调度指令更新所述矿区生产信息；

S3:通过预设的混合奖励函数，基于所述更新后的矿区生产信息和所述时间节点信息，生成所述目标调度指令的动作奖励，应用所述动作奖励在所述调度指令集合中选取优化后的目标调度指令，重复执行步骤S1至步骤S3，不断选取优化后的目标调度指令，并将所述优化后的目标调度指令发送至所述仿真系统，以使所述仿真系统根据所述优化后的目标调度指令更新所述矿区生产信息；

S4:统计预设时间段内的所述仿真系统中的矿区产量信息，计算智能体性能指标，并在所述智能体性能指标达到预设条件时，结束矿车调度模型的训练。

[0006] 根据本发明的第二个方面，提供了一种矿车调度模型训练装置，所述矿车调度模型训练装置包括调度智能体和交互环境，其中，所述交互环境包括仿真系统和人工系统，所述调度智能体包括：

状态模块，用于获取所述仿真系统中的矿区生产信息和所述人工系统中的时间节点信息，并根据所述矿区生产信息和所述时间节点信息，生成调度状态信息；

动作模块，用于通过预设的强化学习算法，根据所述调度状态信息，在预设的调度指令集合中选取出目标调度指令，并将所述目标调度指令发送至所述仿真系统，以使所述仿真系统根据所述目标调度指令更新所述矿区生产信息；

奖励模块，用于通过预设的混合奖励函数，基于所述更新后的矿区生产信息和所述时间节点信息，生成所述目标调度指令的动作奖励，应用所述动作奖励在所述调度指令集合中选取优化后的目标调度指令，重复执行所述状态模块、所述动作模块和所述奖励模块的步骤，不断选取优化后的目标调度指令，并将所述优化后的目标调度指令发送至所述仿真系统，以使所述仿真系统根据所述优化后的目标调度指令更新所述矿区生产信息；

评价模块，统计预设时间段内的所述仿真系统中的矿区产量信息，计算智能体性能指标，并在所述智能体性能指标达到预设条件时，结束矿车调度模型的训练。

[0007] 根据本发明的第三个方面，提供了一种芯片，所述芯片包括至少一个处理器和通信接口，所述通信接口和所述至少一个处理器耦合，所述至少一个处理器用于运行计算机程序或指令，以实现上述矿车调度模型训练方法。

[0008] 根据本发明的第四个方面，提供了一种终端，所述终端包括上述矿车调度模型训练装置。

[0009] 根据本发明的第五个方面，提供了一种计算机设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述计算机程序被处理器执行时实现上述矿车调度模型训练方法。

[0010] 根据本发明的第六个方面，提供了一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现上述矿车调度模型训练方法。

[0011] 本发明的有益效果是：本发明提供的矿车调度模型训练方法可以应用于矿车调度模型训练装置中，所述矿车调度模型训练装置包括调度智能体和交互环境，其中，交互环境包括仿真系统和人工系统。上述矿车调度模型训练方法可以根据仿真系统中的矿区生产信息和人工系统中的时间节点信息生成调度状态信息，并可以基于调度状态信息选取出目标调度指令，以使仿真系统能够根据目标调度指令更新矿区生产信息，进而通过预设的混合奖励函数，基于更新后的的矿区生产信息和人工系统中的时间节点信息，生成目标调度指令的动作奖励，从而应用动作奖励不断地选取优化后的目标调度指令，直至智能体性能指标达到预设条件时，得到训练好的矿车调度模型。上述方法充分结合了仿真系统和人工系统的互补优势，使调度智能体能够同时从仿真系统和人工系统中学习调度策略，增加了行为奖励的密度，使基于强化学习的车辆调度模型可以及时的获取到与调度指令相关的行为奖励。上述方法极大的提高了获取奖励反馈的及时性，降低了车辆调度模型的训练时间成本，并提高了车辆调度模型的训练效率和模型性能。

[0012] 上述说明仅是本申请技术方案的概述，为了能够更清楚了解本申请的技术手段，而可依照说明书的内容予以实施，并且为了让本申请的上述和其它目的、特征和优点能够

更明显易懂,以下特举本申请的具体实施方式。

附图说明

[0013] 此处所说明的附图用来提供对本发明的进一步理解,构成本申请的一部分,本发明的示意性实施例及其说明用于解释本发明,并不构成对本发明的不当限定。在附图中:

图1示出了本发明实施例提供的一种矿车调度模型训练方法流程示意图;

图2示出了本发明实施例提供的一种矿车调度模型训练方法流程示意图;

图3示出了本发明实施例提供的一种矿车调度模型训练装置结构示意图;

图4示出了本发明实施例提供的一种矿车调度模型训练方法交互示意图;

图5示出了本发明实施例提供的一种矿车调度模型训练装置结构示意图;

图6示出了本发明实施例提供的一种矿车调度模型训练装置结构示意图;

图7示出了本发明实施例提供的一种计算机设备的结构示意图;

图8示出了本发明实施例提供的一种芯片的结构示意图;

图9示出了本发明实施例提供的一种终端的结构示意图;

图10示出了本发明实施例提供的一种计算机可读存储介质的结构示意图。

具体实施方式

[0014] 下文中将参考附图并结合实施例来详细说明本发明。需要说明的是,在不冲突的情况下,本申请中的实施例及实施例中的特征可以相互组合。

[0015] 在介绍本发明的各实施例之前,首先对强化学习中常见的概念进行说明,其中,各概念的定义列举如下:

智能体(Agent):强化学习的本体,作为学习者或决策者存在;

环境(Environment):智能体以外的一切,主要指状态;

状态(States):表示环境的数据,状态集是环境中所有可能的状态;

动作(Actions):智能体可以作出的动作,动作集是智能体可以作出的所有动作;

奖励(Rewards):智能体在执行一个动作后,获得的正负奖励信号;

策略(Policy):从状态到动作的映射,智能体基于某种状态选择某种动作的过程。

[0016] 强化学习的学习过程如下:1)智能体感知环境状态;2)智能体根据某种策略做出动作;3)动作作用于环境导致环境状态的改变;4)环境向智能体发出一个反馈信号,以使智能体优化自身的策略;5)不断重复上述过程,直至智能体性能在较长时间内无明显提升,认为模型训练结束。上述强化学习的目标是训练智能体寻找在连续时间序列里的最优策略,其中,最优策略是指使得长期累积奖励最大化的策略。

[0017] 目前,在基于强化学习的矿车调度模型中,调度智能体通常需要下发几个调度指令甚至下发几十个调度指令,并在矿区车辆在抵达目的地或完成调度任务后,才能够得到信息反馈,这导致来自环境的奖励反馈十分稀疏。为了得到环境奖励,在训练车辆调度模型时往往需要付出极高的训练时间代价,这导致车辆调度模型的训练效率极低,训练成本则很高,此外,这也导致基于强化学习的矿车调度模型在矿车调度问题上很难表现出较高的性能。

[0018] 实施例一

针对上述问题，在一个实施例中，如图1所示，提供了一种矿车调度模型训练方法，以该方法应用于矿车调度模型训练装置等计算机设备为例进行说明，其中，所述矿车调度模型训练装置包括相互交互的调度智能体和交互环境，所述交互环境包括仿真系统和人工系统。进一步的，上述方法可以包括以下步骤：

101、获取仿真系统中的矿区生产信息和人工系统中的时间节点信息，并根据矿区生产信息和时间节点信息，生成调度状态信息。

[0019] 其中，仿真系统可用于为调度智能体提供一个可交互的模拟矿山调度环境，并在与调度智能体交互过程中，为调度智能体提供所需的矿区生产信息。其中，矿区生产信息也称为环境信息，主要包括与模拟矿山调度环境相关的信息，包括仿真路网信息、仿真设备信息和仿真状态信息等等。例如，矿区生产信息可以包括仿真路网信息、仿真装载设备的设备信息和设备状态、仿真卸载设备的设备信息和设备状态、仿真矿车的车辆信息和车辆状态、以及仿真装载设备与仿真卸载设备之间的行驶时间等等。可以理解的是，仿真系统中的重要参数，如仿真装载设备与仿真卸载设备之间的行驶时间、仿真装载设备的装载时间、仿真卸载设备的卸载时间等，均可以采用真实的生产数据进行模拟分布，从而为调度智能体提供真实、稀疏的环境反馈。

[0020] 进一步的，人工系统可以基于仿真系统中的矿区生产信息进行统计计算，以得到与仿真系统中的仿真装载设备、仿真卸载设备、仿真矿车相关的各个时间节点信息。例如，时间节点信息可以包括仿真车辆采用调度指令集合中的每个调度指令的行驶时间和预期等待时间、每个仿真装载设备的剩余服务时间、以及每个仿真卸载设备的剩余服务时间等等。在本实施例中，人工系统可以理解为是一个“白盒”模型，通过人为模拟确定性进程来推断重要的时间节点，包括每个设备(仿真矿车、仿真装载设备、仿真卸载设备)的任务完成时间、仿真矿车抵达目的地时间等等。例如，假设仿真矿车的出发时间为t_1，预计行驶时间为δt，则仿真矿车抵达目的地的时间为$t_1+\delta t$，仿真装载设备完成上一任务的时间为t_a，则该仿真矿车完成装载的时间实际为$t_c=\max(t_1+\delta t, t_a)+t_2$，其中，$t_2$为仿真装载设备预计的装载时间，同理，卸载过程也与此类似。人工系统可以为调度智能体提供基于专家经验的即时、密集的人工奖励，从而提高基于强化学习的矿车调度模型的收敛速度。

[0021] 具体的，步骤101可以通过调度智能体中的状态模块实现，其中，状态模块可以通过观测仿真系统中的矿区生产信息，生成调度状态信息，并将生成的调度状态信息作为智能体调度决策的依据存储在状态模块中。调度状态信息可以包括多种重要的调度信息，例如，仿真矿车的当前位置信息和动作可用性信息、仿真矿车行驶至各个目的地的行驶时间和预期等待时间、每个仿真装载设备和每个仿真卸载设备的服务进程和设备可用性等等。在本实施例中，仿真矿车的调度状态信息中的时间节点信息可以在人工系统中读取。

[0022] 在本实施例中，状态模块可以在仿真系统中的仿真矿车发送车辆调度请求时，获取仿真系统中的矿区生产信息，并在人工系统中提取出与该仿真矿车对应的各个时间节点信息，进而根据上述矿区生产信息和时间节点信息，生成针对该仿真矿车的调度状态信息。其中，仿真矿车可以在车辆状态更新为装载完成状态或卸载完成状态时发送车辆调度请求，状态模块可以主动获取仿真矿车发出的调度请求，也可以被动接收仿真矿车发出的调度请求。

[0023] 本实施例提出人工系统和仿真系统组成的平行交互模型，可以由人工系统提供即

时、密集的人工奖励，由仿真系统提供真实、稀疏的环境反馈，两个系统始终保持与作业进程的平行一致性，以此实现了人工系统和仿真系统在学习机制上的优势互补。

[0024] 102、通过预设的强化学习算法，根据调度状态信息，在预设的调度指令集合中选取出目标调度指令，并将目标调度指令发送至仿真系统，以使仿真系统根据目标调度指令更新矿区生产信息。

[0025] 具体的，步骤102可以通过调度智能体中的动作模块实现。其中，动作模块可以接收状态模块输出的调度状态信息，并根据调度状态信息，在预设的调度指令集合中选取出基于该调度状态信息能够达到最大化价值的调度指令作为目标调度指令，然后将目标调度指令发送至仿真系统中的仿真车辆(即发送调度请求的仿真车辆)中，其中，最大化价值可以通过预设的价值函数计算得出。进一步的，仿真系统中的仿真车辆在接收到目标调度指令之后，可以根据目标调度指令从当前位置驶往目的地并执行装载或卸载任务，在执行完成目标调度指令执行后，仿真系统中的矿区生产信息会进行更新。

[0026] 在本实施例中，动作模块可用于维护调度指令集合$A=\{a_t \mid t \in R\}$，其中，t是时间点，R是实数域。具体的，调度指令集合A中所有可能的调度指令被分为成对的子集，成对子集是指从仿真系统中的多个装载点和卸载点分别选一个出发地i和一个目的地j，调度指令a_t仅由出发地和目的地组成，其中，t=(i，j)。进一步的，如果请求调度的仿真矿车位于仿真装载设备(假设一个装载点对应一个仿真装载设备)s处，调度智能体会将其分派到仿真卸载设备d处或原地等待，分别对应于调度指令$a_{s\to d}$和$a_{s\to\Phi}$，此时，不允许仿真矿车驶往装载区。相反，如果仿真矿车在仿真卸载设备d处请求调度，则调度智能体只允许仿真矿车前往仿真装载设备s或原地等待，分别对应于调度指令$a_{d\to s}$和$a_{d\to\Phi}$。因此，对于一个仿真矿车来说，调度指令集合中的调度指令总是部分有效的，基于此，可以定义一个有效的动作集a_{valid}和一个无效的动作集$a_{invalid}$。

[0027] 举例来说，假设仿真装载设备使用标识s，仿真卸载设备使用标识d，则多台仿真装载设备为s_1，s_2，s_3，…，多台仿真卸载设备为d_1，d_2，d_3，…。假设，仿真装载设备集合为$S=\{s_1, s_2, s_3, \cdots\}$，仿真卸载设备集合为 $D=\{d_1, d_2, d_3, \cdots\}$，那么，仿真矿车可以执行的调度指令集合为$\{a_{s1\to d1}, a_{s1\to d2}, a_{s1\to d3}, a_{s1\to\Phi}, a_{s2\to d1}, \cdots, a_{d1\to s1}, a_{d1\to s2}, a_{d1\to s3}, a_{d1\to\Phi}, \cdots\}$，即调度指令集合为仿真装载设备集合和仿真卸载设备集合的双向笛卡尔积。进一步的，如果请求调度车辆位于仿真装载设备 s_1处，那么仿真矿车可以执行的调度指令仅有$\{a_{s1\to d1}, a_{s1\to d2}, a_{s1\to d3}, a_{s1\to\Phi}\}$，而不以 s_1为起点的调度指令，或者以卸载点d为起点的调度指令，仿真矿车则无法执行，因此，动作集合总是部分有效的。

[0028] 103、通过预设的混合奖励函数，基于更新后的矿区生产信息和时间节点信息，生成目标调度指令的动作奖励，应用动作奖励在调度指令集合中选取优化后的目标调度指令。

[0029] 具体的，步骤103可以通过调度智能体中的奖励模块实现。其中，奖励模块可以为调度智能体提供与目标调度指令对应的动作奖励，应用动作奖励，可以在下一次的调度过程中，为仿真车辆选取出优化后的目标调度指令。进一步的，重复执行上述步骤101至步骤103，可以不断的选取出优化后的目标调度指令，并将优化后的目标调度指令发送至仿真系统，以使仿真系统可以根据优化后的目标调度指令更新仿真系统中的矿区生产信息，从而持续激励调度智能体学习更优的调度策略。

[0030] 在本实施例中，奖励模块可以通过预设的混合奖励函数，基于仿真系统中更新后的矿区生产信息和人工系统中的时间节点信息，生成目标调度指令的动作奖励。其中，混合奖励函数可以由预设的人工奖励函数和仿真奖励函数组成，人工奖励函数可以基于人工系统中的重要的时间节点信息设置，例如可以基于仿真矿车的行程时间设置，模拟奖励可以基于仿真系统中的重要的矿区生产信息设置，例如可以基于矿区的产量或设备闲置率设置。进一步的，可以将仿真系统的矿区生产信息输入到仿真奖励函数中，得到仿真奖励值，将人工系统的时间节点信息输入到人工奖励函数中，得到人工奖励值，最后根据仿真奖励值和人工奖励值，得到目标调度指令的动作奖励，该动作奖励可以在仿真矿车执行完成目标调度指令后计算得出。

[0031] 104、统计预设时间段内的仿真系统中的矿区产量信息，计算智能体性能指标，并在智能体性能指标达到预设条件时，结束矿车调度模型的训练。

[0032] 具体的，在矿车调度模型的训练过程中，可以基于仿真系统中的矿区生产信息，统计预设时间段内的仿真系统中的矿区产量信息，并根据预设时间段内的矿区产量信息，实时计算矿车调度模型的智能体性能指标，其中，智能体性能指标的定义和计算方式可以预先设定。例如，智能体性能指标可以为一段预设时间段内的装载总量或卸载总量等信息，也可以是一段预设时间段内的设备闲置率信息，还可以是产量确定时的生产成本信息等等。智能体性能指标的具体设定可以根据实际应用场景的需求进行设定，本申请对智能体性能指标的具体设定不做进一步的限定。进一步的，当智能体性能指标在较长时间内无明显提升时，可以认为模型训练结束，以此得到一个基于强化学习的矿车调度模型。在得到矿车调度模型之后，可以在矿车调度模型中输入一个真实的环境状态，然后，调度智能体可以根据输入的环境状态输出一个最优策略。体现在调度问题中，可以是在车辆完成当前任务时，向调度智能体发起调度请求，调度智能体可以读取当前的环境状态(车辆，路网等)，并为请求调度车辆分配一个最优目的地，以此完成车辆调度。

[0033] 本实施例提供的矿车调度模型训练方法，可以应用于矿车调度模型训练装置中，该矿车调度模型训练装置包括调度智能体和交互环境，其中，交互环境包括仿真系统和人工系统。上述方法可以根据仿真系统中的矿区生产信息和人工系统中的时间节点信息生成调度状态信息，并可以基于调度状态信息选取出目标调度指令，以使仿真系统能够根据目标调度指令更新矿区生产信息，进而通过预设的混合奖励函数，基于更新后的矿区生产信息和人工系统中的时间节点信息，生成目标调度指令的动作奖励，从而应用动作奖励不断地选取优化后的目标调度指令，直至智能体性能指标达到预设条件时，得到训练好的矿车调度模型。上述方法充分结合了仿真系统和人工系统的互补优势，使调度智能体能够同时从仿真系统和人工系统中学习调度策略，从而增加了行为奖励的密度，使基于强化学习的车辆调度模型可以及时获取到与调度指令相关的行为奖励。上述方法提高了获取奖励反馈的及时性，降低了车辆调度模型的训练时间成本，提高了车辆调度模型的训练效率和性能。

[0034] 实施例二

进一步的，作为上述实施例的具体实施方式的细化和扩展，为了完整说明本实施例的实施过程，如图2所示，还提供了一种矿车调度模型训练方法，以该方法应用于如图3所示的矿车调度模型训练装置为例进行说明，其中，该矿车调度模型训练装置包括相互交互的调度智能体和交互环境，该交互环境包括仿真系统和人工系统。其中，矿车调度模型训练

方法包括以下步骤：

201、在仿真系统中模拟生成矿区生产信息。

[0035] 具体的，在矿车调度模型的训练过程中，有效模拟露天矿山的生产场景中的动态随机特性，对于调度模型的训练和评估都是必要的。基于此，本实施例可以采用离散事件仿真器来模拟矿山的运输操作和矿区生产信息。其中，矿区生产信息可以包括仿真路网信息(如道路长度、宽度、坡度、限速等道路信息)、至少一个仿真装载设备(如电铲、挖机等)、每个仿真装载设备的设备信息(如型号、装载时间等信息)和设备状态(如未装载、正在装载、装载排队车数等信息)、至少一个仿真卸载设备、每个仿真卸载设备的设备信息(如型号、卸载时间等信息)和设备状态(如未卸载、正在卸载和卸载排队车数等信息)、至少一个仿真矿车、每个仿真矿车的车辆信息(如平均车速、装载量等信息)和车辆状态(如空载运输，重载运输，正在装载，正在卸载，排队等待装载，排队等待卸载等状态)以及每个仿真装载设备和每个仿真卸载设备之间的行驶时间等信息中的至少一种信息。

[0036] 在本实施例中，可以采用SimPy离散事件仿真器来模拟矿山的运输操作和矿区生产信息。其中，对于仿真系统中的重要参数，如行程时间、矿车的装载量、卸载设备的卸载时间等，可以采用真实的生产数据模拟分布。此外，仿真系统还可以模拟装置设备、卸载设备和矿车的随机故障事件，以便更真实的模拟出实际矿山的生产场景，以此为智能体提供真实、稀疏的环境反馈。

[0037] 进一步的，为了模拟车辆运输流程，本实施例还可以使用SimPy开发一个物料运输模拟器。其中，SimPy是一个基于过程的离散事件仿真框架。在这个框架中，仿真装载设备和仿真卸载设备可以被设计为是具有固定容量和排队效应的有限资源。在仿真矿车请求调度的时间点，所有仿真卸载设备和仿真装载设备的设备状态均作为状态向量传递给调度智能体。进一步的，为了增加模拟器的真实性，可以通过正太分布采样装载、卸载及运输时间，其参数来自矿山真实数据中学习。在本实施例中，物料运输模拟器可以有效应对露天矿的多样性和快速变化，为强化学习模型优化提供真实的行为奖励。

[0038] 202、在仿真系统中的仿真矿车发送车辆调度请求时，根据矿区生产信息，在人工系统中生成针对仿真矿车的时间节点信息。

[0039] 具体的，在仿真系统中的仿真矿车发送车辆调度请求时，可以根据矿区生产信息，在人工系统中生成针对该仿真矿车的时间节点信息，其中，时间节点信息可以包括调度指令集合中的每个调度指令对应的行驶时间、每个调度指令对应的预期等待时间、仿真系统中每个仿真装载设备的剩余服务时间，以及仿真系统中每个仿真卸载设备的剩余服务时间中的至少一种时间信息。本实施例通过计算仿真系统中每个设备的时间节点信息，可以人为模拟确定性进程来推断重要的时间节点，从而为调度智能体提供基于专家经验的即时、密集的人工奖励，以此提高基于强化学习的矿车调度模型的收敛速度。

[0040] 203、获取仿真系统中的矿区生产信息和人工系统中的时间节点信息，并根据矿区生产信息和时间节点信息，生成调度状态信息。

[0041] 具体的，在仿真系统中的仿真矿车发送车辆调度请求时，状态模块可以获取仿真系统中的矿区生产信息和人工系统中的时间节点信息，其中，车辆调度请求可以在仿真矿车的车辆状态更新为装载完成状态或卸载完成状态时发送。进一步的，状态模块可以根据矿区生产信息和人工系统中的时间节点信息，生成针对该请求调度的仿真矿车的调度状态

信息，其中，调度状态信息可以包括请求调度的仿真车辆的位置信息、调度指令集合中的每个调度指令的动作可用性信息、每个调度指令对应的行驶时间、每个调度指令对应的预期等待时间、仿真系统中每个仿真装载设备的剩余服务时间和故障信息，以及仿真系统中每个仿真卸载设备的剩余服务时间和故障信息中的至少一种信息。下面详细描述调度状态信息中的各个状态参数。

[0042] 1)仿真车辆的位置信息。具体的，仿真矿车在完成装载或卸载后请求调度的位置可以被状态模块所捕获，以便使调度智能体学习哪些调度指令是有效的，同时，可以根据当前的位置信息为仿真矿车指派最佳的调度目的地。

[0043] 2)每个调度指令对应的行驶时间。具体的，通过引入仿真车辆驶往每个目的地的行驶时间，可以使调度智能体能够掌握动作选择的时间效应。其中，每个调度动作对应一个行驶时间的数值，该数值表示到指定目的地的行驶时间，对于无效操作，该行驶时间的数值自动为0。此外，行驶时间的数值不包括等待装载的时间或等待卸载的时间。

[0044] 3)每个调度指令对应的预期等待时间。具体的，对于每个装载区和卸载区，状态模块可以计算仿真矿车驶往该目的地的预期等待时间，以使调度智能体能够学习前往各个目的地的等待时间成本。以仿真矿车前往装载区为例，等待时间主要源自两部分，一个是仿真装载设备正在装载的仿真矿车，另一个是在该仿真矿车之前抵达该装载区的仿真矿车。若仿真矿车在仿真装载设备完成这两类矿车的装载后抵达装载区，则等待时间为0，否则产生等待时间。等待时间从仿真矿车抵达仿真装载设备的时刻开始计算，独立于行驶时间。

[0045] 4)仿真装载设备和仿真卸载设备的剩余服务时间。具体的，对于每个仿真装载设备和每个仿真卸载设备，其是否在工作(装载或卸载)，以及如果在工作，剩余的服务时间是多少，是一项重要的状态信息。若仿真装载设备或仿真卸载设备没有为任何仿真矿车服务，则剩余服务时间为0。

[0046] 5)调度指令集合中的每个调度指令的动作可用性信息。具体的，动作可用性信息表示每个调度指令对于当前请求的仿真矿车来说是否有效，若动作有效，则取值为1，否则自动为0。

[0047] 6)仿真装载设备和仿真卸载设备的故障信息。具体的，露天矿山生产设备故障时常发生，状态模块可以及时获取仿真装载设备或仿真卸载设备的故障信息，以使调度智能体学习如何在动态环境下采取合适的行动。对于任何故障设备，该值取值为0，否则为1。

[0048] 本实施例通过在调度状态信息中引入预期等待时间、行驶时间等更具体的状态表示，可以保证强化学习算法训练的收敛性。

[0049] 204、针对调度指令集合中的每个调度指令，通过预设的价值函数，计算每个调度指令在调度状态信息下的价值数值。

[0050] 205、将数值最大的价值数值对应的调度指令确定为目标调度指令，将目标调度指令发送至仿真系统中发送车辆调度请求的仿真矿车中，以使仿真系统根据目标调度指令更新矿区生产信息。

[0051] 具体的，动作模块可以接收状态模块输出的调度状态信息，并根据调度状态信息，在预设的调度指令集合中选取出基于该调度状态信息能够达到最大化价值的调度指令作为目标调度指令，然后将目标调度指令发送至仿真系统中的仿真车辆(即发送调度请求的仿真车辆)中，进一步的，仿真系统中的仿真车辆在接收到目标调度指令之后，可以根据目

标调度指令从当前位置驶往目的地并执行装载或卸载任务，在执行完成目标调度指令执行后，仿真系统中的矿区生产信息会进行更新。在本实施例中，可以通过公式(1)所示的Q-learning 算法，在调度指令集合中选取目标调度指令，其中，公式(1)如下所示：

$$a_{target} = \arg\max_{a_t} Q^{\pi}(s_t, a_t) \quad (1)$$

其中，a_{target}为目标调度指令，s_t为调度状态信息，a_t为调度指令集合中的调度指令，$Q^{\pi}(s_t,a_t)$为调度智能体在调度状态信息s_t下采取调度指令a_t的价值函数，argmax $Q^{\pi}(s_t,a_t)$表达的是a_t的定义域的一个子集中的任一个调度指令a_t，均可使价值函数$Q^{\pi}(s_t,a_t)$取最大值。基于此，目标调度指令a_{target}可以在这个子集中选取，以使选取出的目标调度指令可以在调度状态s_t下，使价值函数$Q^{\pi}(s_t,a_t)$取最大值。

[0052] 206、通过预设的混合奖励函数，基于更新后的矿区生产信息和时间节点信息，生成目标调度指令的动作奖励。

[0053] 具体的，混合奖励函数可以由人工奖励函数和仿真奖励函数组成，基于此，目标调度指令的动作奖励可以通过以下方法得到：首先在人工系统中的时间节点信息中，提取出目标调度指令对应的行驶时间和预期等待时间，并根据行驶时间和预期等待时间之和，得到目标调度指令对应的行程时间，然后将目标调度指令对应的行程时间输入至人工奖励函数中，得到目标调度指令的人工奖励值，进而根据仿真系统中更新后的矿区生产信息，计算目标调度指令从执行开始至执行完成之间的仿真装载设备利用率，即计算目标调度指令至目标调度指令的下一个调度指令之间的仿真装载设备利用率，并将仿真装载设备利用率输入至仿真奖励函数中，得到目标调度指令的仿真奖励值，最后，根据人工奖励值和仿真奖励值，得到目标调度指令的动作奖励。

[0054] 在本实施例中，可以使用基于仿真矿车的行程驶时间w的人工奖励函数R_{art}来激励智能体，其中，人工奖励函数R_{art}如公式(2)所示：

$$R_{art}(s_t,a_t) = \begin{cases} e^{-\frac{w}{h}} & a_t \ is \ valid \\ 0 & else \end{cases} \quad (2)$$

其中，s_t为调度状态信息，a_t为目标调度指令，$R_{art}(s_t,a_t)$为目标调度指令的人工奖励值，a_t is valid指的是目标调度指令对于当前请求的仿真矿车是有效的，else指的是目标调度指令对于当前请求的仿真矿车是无效的，w为仿真矿车的行程驶时间，h为归一化因子，一般w取值需保证$w/h \in [0,2]$，e为自然常数，其值约为2.718，常作为指数函数的底数。在本实施例中，仿真矿车的行程驶时间w(取行驶时间与预期等待时间之和)由人工系统根据矿山环境实时计算。由于仿真矿车的行程驶时间w被认为为固定的，因此，可以预测一个调度指令将导致何种调度状态。人工奖励虽然很难捕获露天矿场景的动态性和随机性，但是，人工奖励不存在延迟问题，目标调度指令在执行后立即可以得到奖励，这种密集且及时的反馈确保了调度智能体在训练过程中的收敛性。

[0055] 进一步的，为了提高调度智能体的鲁棒性，本实施例还引入了一个完全基于仿真环境的仿真奖励函数来捕捉环境动态特性。在本实施例中，可以使用基于仿真装载设备利用率u的仿真奖励函数R_{sim}来激励智能体。其中，仿真装载设备利用率高代表调度系统可以

均衡地分配仿真矿车，以便使仿真矿车的运输能力尽可能与仿真装载设备的装载能力相匹配。其中，仿真奖励函数R_{sim}如公式(3)所示：

$$R_{sim}(s_t,a_t)=\begin{cases}e^{-\frac{u}{k}}-1 & a_t\ is\ valid\\ 0 & else\end{cases} \quad (3)$$

其中，s_t为调度状态信息，a_t为目标调度指令，$R_{sim}(s_t,a_t)$为目标调度指令的仿真奖励值，a_t is valid指的是目标调度指令对于当前请求的仿真矿车是有效的，else指的是目标调度指令对于当前请求的仿真矿车是无效的，u是目标调度指令a_t和下一个调度指令a_{t+1}之间这段时间内的仿真装载设备利用率，k为归一化因子，一般满足$k\in[1,2]$。仿真奖励反映了模拟环境的真实反馈，并且，该仿真奖励函数是稠密的，但是，仿真奖励与调度指令之间的因果关系并不明显，存在严重的奖励延迟。

[0056] 进一步的，混合奖励函数由人工奖励函数和仿真奖励函数组成，其中，人工奖励函数最小化行驶时间，仿真奖励函数最大化仿真环境设备利用率。人工系统提供即时密集的人工奖励，不存在奖励延迟问题，保证了调度智能体在训练过程中的收敛性；仿真系统提供真实可靠的仿真反馈，仿真奖励能够最大程度的还原仿真环境，使得调度智能体捕获环境动态特性，通过混合奖励函数，实现了人工奖励函数和仿真奖励函数在学习机制上的优势互补。

[0057] 进一步的，在一个可选的实施例中，目标调度指令的动作奖励可以通过以下方法计算：首先，根据仿真系统中更新后的矿区生产信息，计算智能体性能指标，并根据智能体性能指标，确定人工奖励权重系数，然后根据人工奖励权重系数，计算仿真奖励权重系数，其中，人工奖励权重系数和仿真奖励权重系数的和为预设值，最后，根据人工奖励权重系数与人工奖励值的乘积与仿真奖励权重系数与仿真奖励值的乘积的和值，得到目标调度指令的动作奖励。

[0058] 具体的，混合奖励函数综合了人工奖励函数和仿真奖励函数的优点。其中，人工奖励值基于领域知识提供密集且即时的反馈，以确保算法的收敛性，仿真奖励提高了算法在仿真环境中的鲁棒性。混合奖励方法使用自适应权重因子λ(与人工奖励权重系数为同一含义)，将人工奖励和仿真奖励与相结合。

[0059] 具体来说，在开始训练时，可以令人工奖励权重系数λ的初始值为1，在训练过程中，当智能体性能指标P提升到一定水平后(这里P可以选用仿真系统一个班次时间内的矿区产量信息，其中，矿区产量信息可以指一个班次内所有卸载点卸载物料的总吨数，一个工作班次大概8小时)，如智能体性能理论最大值P_{theor}的80%。λ开始逐渐降低，当λ低于0.5时，仿真奖励将占主导地位；一旦智能体性能下降到一定水平，如理论值60%，人工奖励权重系数λ将再次增加。在本实施例中，可以通过智能体性能指标P，得到当前状态下的人工奖励权重系数λ，进一步的，由于工奖励权重系数和仿真奖励权重系数的和为预设值(其中，预设值可以为任意值，本实施例以预设值为1进行举例说明)，因此，可以确定仿真奖励权重系数为1-λ(这里假设预设值为1)，进一步的，混合奖励函数R_{hyd}的计算公式如公式(4)所示：

$$R_{hyd}=\lambda*R_{art}+(1-\lambda)*R_{sim};(4)$$

其中，λ为人工奖励权重系数，1-λ为仿真奖励权重系数，R_{hyd}为混合奖励函数，R_{art}为人工奖励函数，R_{sim}为仿真奖励函数。

[0060] 进一步的，在一个可选的实施例中，可以通过以下方法对人工奖励权重系数进行实时调整：首先，在仿真系统中更新后的矿区生产信息中，提取出预设时间段内的矿区产量信息，并将预设时间段内的矿区产量信息划分为多个子时间段内的矿区产量信息，然后，对多个子时间段内的矿区产量信息进行线性拟合，得到性能指标拟合斜率，并计算多个子时间段内的矿区产量信息的平均值，得到性能指标均值，进而，判断性能指标拟合斜率是否小于预设的斜率阈值，并判断性能指标均值是否大于预设的性能指标阈值，其中，斜率阈值为负值，最后，若性能指标拟合斜率小于斜率阈值，则对人工奖励权重系数进行递增计算，若性能指标均值小于等于所述性能指标阈值，则对人工奖励权重系数进行递增计算，若性能指标均值大于性能指标阈值，则对人工奖励权重系数进行递减计算。在本实施例中，斜率阈值和性能指标阈值均可以根据实际场景设置在矿车调度模型训练装置的调度智能体中，例如，斜率阈值可以设置为小于等于-0.5，具体可以取值为-0.5，性能指标阈值可以设置为小于等于智能体性能理论最大值P_{theor}的80%，具体可以取值为智能体性能理论最大值P_{theor}的60%或70%。可以理解的是，人工奖励权重系数可以在模型训练的过程中实时更新，也可以按照一定的时间周期进行定期更新。

[0061] 举例来说，人工奖励权重系数λ的更新方式如下：

1)对最近N个轮次(即子时间段，时长可以预设，如8小时)内得到的智能体性能指标P_i (i=1,...,N)进行线性拟合，得到性能指标拟合斜率μ。

[0062] 2)当μ <－0.5时，通过公式(5)更新λ，其中，μ为负值表示智能体性能指标在下降，公式(5)的实质是对人工奖励权重系数λ进行递增计算。

[0063] $$\lambda = \lambda - 0.1*\mu \quad (5)$$

3)在更新λ的过程中，继续取最近N个轮次内智能体性能指标P_i (i=1,...,N)的均值M，其中，步骤3)与步骤1)中的最近N个轮次内智能体性能指标是相同的。

[0064] 4)若$M > 0.7*P_{theor}$，则通过公式(6)对人工奖励权重系数λ进行递减计算，其中，步骤4)和步骤2)中对于权重因子λ的更新是同步的。

[0065] $$\lambda = \lambda - 0.01 \quad (6)$$

5)若$M \leq 0.7*P_{theor}$，则通过公式(7)对人工奖励权重系数λ进行递增计算，其中，步骤5)和步骤2)中对于权重因子λ的更新也是同步的。

[0066] $$\lambda = \lambda + 0.01 \quad (7)$$

6)人工奖励权重因子λ在整个训练过程中将不断变化，直至智能体性能趋于稳定，最后，利用上述公式(4)，即可得到目标调度指令的动作奖励。

[0067] 可以理解的是，上述步骤1)至步骤5)仅作为举例说明，不作为对本实施例的限定，在实际应用过程中，可以任意替换公式中的数值或参数。

[0068] 在本实施例中，调度智能体能够在不同的学习阶段自动调整人工奖励权重因子。通过上述方法，当智能体性能水平较低时，可以自动提升基于专家经验的人工奖励权重，从而为智能体提供即时、密集的反馈，以加速调度智能体的收敛速度；当智能体性能达到一定水平后，可以自动提升基于仿真反馈的真实、稀疏的反馈，从而提升智能体在动态环境中的鲁棒性。

[0069] 207、根据目标调度指令的动作奖励，对价值函数进行更新，得到优化后的价值函数，应用优化后的价值函数在调度指令集合中选取优化后的目标调度指令。

[0070] 具体的，在得到目标调度指令的动作奖励之后，可以利用目标调度指令的动作奖励对当前调度状态下的价值函数进行更新，从而得到优化后的价值函数。应用优化后的价值函数，可以在下一次的调度过程中，选取出一个价值最大的调度指令作为优化后的目标调度指令，通过重复执行上述步骤201至步骤207，可以不断的选取出优化后的目标调度指令，并将优化后的目标调度指令发送至仿真系统，以使仿真系统可以根据优化后的目标调度指令更新仿真系统中的矿区生产信息，从而持续激励调度智能体学习更优的调度策略。

[0071] 在本实施例中，可以通过赋值公式(8)所示的Bellman方程对价值函数进行迭代更新，其中，公式(8)如下所示；

$$Q^{\pi}(s_t,a_t)=Q^{\pi}(s_t,a_t)+\alpha(r_t+\gamma\max_{a_t}Q^{\pi}(s_{t+1},a_t)-Q^{\pi}(s_t,a_t)) \quad (8)$$

其中，s_t为目标调度指令对应的调度状态信息，s_{t+1}为下一个目标调度指令对应的调度状态信息，a_t为目标调度指令，r_t为目标调度指令的动作奖励，$Q^{\pi}(s_t,a_t)$为调度智能体在调度状态信息s_t下采取调度指令a_t的价值函数，$Q^{\pi}(s_{t+1},a_t)$为调度智能体在下一个目标调度指令对应的调度状态信息s_{t+1}下采取调度指令a_t的价值函数，α为学习率，γ为衰减因子，max $Q^{\pi}(s_{t+1},a_t)$表达的是目标调度指令a_t使价值函数$Q^{\pi}(s_{t+1},a_t)$取到的最大值。

[0072] 208、统计预设时间段内的仿真系统中的矿区产量信息，计算智能体性能指标，并在智能体性能指标达到预设条件时，结束矿车调度模型的训练。

[0073] 具体的，在矿车调度模型的训练过程中，可以基于仿真系统中的矿区生产信息，统计预设时间段内的仿真系统中的矿区产量信息，并根据预设时间段内的矿区产量信息，实时计算矿车调度模型的智能体性能指标，其中，智能体性能指标的定义和计算方式可以预先设定。例如，智能体性能指标可以为一段预设时间段内的装载总量或卸载总量等信息，也可以是一段预设时间段内的设备闲置率信息，还可以是产量确定时的生产成本信息等等。进一步的，当智能体性能指标在较长时间内无明显提升时，可以认为模型训练结束，以此得到一个基于强化学习的矿车调度模型。在得到矿车调度模型之后，可以在矿车调度模型中输入一个真实的环境状态，然后，调度智能体可以根据输入的环境状态输出一个最优策略。体现在调度问题中时，可以是车辆完成当前任务时，向调度智能体发起调度请求，调度智能体可以读取当前的环境状态(车辆，路网等)，并为请求调度车辆分配一个最优目的地，以此完成车辆调度。

[0074] 在本实施例中，可以通过以下方法判断矿车调度模型是否训练完成：首先，在仿真系统中的矿区生产信息中提取出预设时间段内的矿区产量信息，并将该时间段内的矿区产量信息划分为多个子时间段内的矿区产量信息，然后计算多个子时间段内的矿区产量信息的平均值，得到性能指标均值，进而通过方差或标准差等指标数值，计算多个子时间段的矿区产量信息与性能指标均值之间的偏差值，最后，当该偏差值小于预设的偏差阈值时，可以认为智能体性能指标在较长时间内无明显提升，此时，可以结束矿车调度模型的训练。在本实施例中，预设时间段指的是从当前时间向前追溯的一段固定长度的时间，矿区产量信息指的是预设时间段内的装载总量或卸载总量，矿区产量信息也可以替换为预设时间段内的设备闲置率或生产成本信息等其他性能指标，子时间段可以理解为露天矿区场景中的一个轮次或者班次，通过统计最近一段时间多个轮次的矿区产量信息的偏差度值，即可计算出

智能体性能指标的变化范围，从而判定智能体性能指标在此时间段内是否有明显提升，并以此判定模型是否训练完成，其中，偏差度值可以用方差、标准差等统计数值进行表示。

[0075] 在本实施例中，矿车调度模型训练装置的结构图如图3所示，调度智能体与仿真环境之间的交互过程如图4所示。如图3和图4所示，当仿真系统中的仿真矿车完成装载任务或卸载任务时，可以向调度智能体发送调度请求，调度智能体中的状态模块可以基于该调度请求，获取仿真系统中的矿区生产信息o_t和人工系统中的时间节点信息，并基于矿区生产信息o_t和时间节点信息，生成调度状态信息s_t，然后，动作模块可以基于调度状态信息s_t，在预设的调度指令集合中选取一个调度指令a_t，并将该调度指令a_t发送至仿真系统中的仿真矿车中。进一步的，仿真矿车接收到调度指令a_t后，可以执行运输任务，此时，仿真矿车的车辆状态可能是空载状态或重载状态，当仿真矿车到达目的地之后，可能经过会排队等待后执行装载或卸载任务，也可能直接执行装载或卸载任务。任务执行完成后，仿真系统中的矿区生产信息会进行更新，并且，仿真系统和人工系统会向调度智能体的奖励模块反馈相应的信息，使调度智能体能够通过自适应权重调节模块确定人工奖励权重系数，并根据人工奖励权重系数和预设的人工奖励函数和仿真奖励函数，得到调度指令a_t的混合奖励值r_t，应用混合奖励值r_t，可以使调度智能体在接下来的调度过程中不断选取优化后的目标调度指令，并使得仿真系统能够不断通过优化后的目标调度指令更新矿区生产信息并反馈给调度智能体，以优化调度智能体的矿车调度策略。上述训练过程不断重复，直至矿车调度模型训练完成。

[0076] 本实施例提供的矿车调度模型训练方法，上述方法首先在仿真系统中模拟生成矿区生产信息，并根据矿区生产信息，在人工系统中生成针对仿真矿车的时间节点信息，然后根据仿真系统中的矿区生产信息和人工系统中的时间节点信息，生成调度状态信息，并通过价值函数，基于调度状态信息选取出目标调度指令，以使仿真系统能够根据目标调度指令更新矿区生产信息，进而通过预设的混合奖励函数，基于更新后的矿区生产信息和人工系统中的时间节点信息，生成目标调度指令的动作奖励，从而通过动作奖励更新价值函数，以及通过价值函数选取优化后的目标调度指令，不断重复上述训练过程，直至智能体性能指标达到预设条件时，得到训练好的矿车调度模型。上述方法通过人工系统提供密集的即时人工奖励，通过仿真系统提供真实可靠的仿真反馈，并通过混合奖励函数融合人工奖励函数和仿真奖励函数的优势，可以充分结合仿真系统和人工系统的互补优势，使调度智能体能够同时从仿真系统和人工系统中学习调度策略，极大的提高了获取奖励反馈的及时性，并降低了车辆调度模型的训练时间成本。并且，上述方法通过自适应权重因子对人工奖励函数和仿真奖励函数进行组合得到混合奖励函数，可以确保智能体能够快速收敛的同时提升其应对动态环境的鲁棒性。此外，上述方法通过在调度状态信息中引入设备预期等待时间、行驶时间等更具体的状态表示，可以保证强化学习算法训练的收敛性，同时，调度指令的设计也可以使模型的学习性及可解释性更强，从而方便调度人员的理解与维护。

[0077] 实施例三

进一步的，作为图1、图2所示方法的具体实现，本实施例提供了一种矿车调度模型训练装置，如图5所示，该装置包括：状态模块31、动作模块32、奖励模块33和评价模块34，其中：

状态模块31，可用于获取所述仿真系统中的矿区生产信息和所述人工系统中的时

间节点信息，并根据所述矿区生产信息和所述时间节点信息，生成调度状态信息；

动作模块32，可用于通过预设的强化学习算法，根据所述调度状态信息，在预设的调度指令集合中选取出目标调度指令，并将所述目标调度指令发送至所述仿真系统，以使所述仿真系统根据所述目标调度指令更新所述矿区生产信息；

奖励模块33，可用于通过预设的混合奖励函数，基于所述更新后的矿区生产信息和所述时间节点信息，生成所述目标调度指令的动作奖励，应用所述动作奖励在所述调度指令集合中选取优化后的目标调度指令，重复执行状态模块31、动作模块32和奖励模块33的步骤，不断选取优化后的目标调度指令，并将所述优化后的目标调度指令发送至所述仿真系统，以使所述仿真系统根据所述优化后的目标调度指令更新所述矿区生产信息；

评价模块34，可用于统计预设时间段内的所述仿真系统中的矿区产量信息，计算智能体性能指标，并在所述智能体性能指标达到预设条件时，结束矿车调度模型的训练。

[0078] 在具体的应用场景中，所述状态模块31，具体可用于在所述仿真系统中的仿真矿车发送车辆调度请求时，获取所述仿真系统中的矿区生产信息和所述人工系统中的时间节点信息，其中，所述车辆调度请求在所述仿真矿车的车辆状态更新为装载完成状态或卸载完成状态时发送；根据所述矿区生产信息和所述人工系统中的时间节点信息，生成针对所述仿真矿车的调度状态信息，其中，所述调度状态信息包括所述仿真矿车的位置信息、所述调度指令集合中的每个调度指令的动作可用性信息、每个所述调度指令对应的行驶时间、每个所述调度指令对应的预期等待时间、所述仿真系统中每个仿真装载设备的剩余服务时间和故障信息，以及仿真系统中每个仿真卸载设备的剩余服务时间和故障信息中的至少一种信息。

[0079] 在具体的应用场景中，所述动作模块32，具体可用于针对所述调度指令集合中的每个调度指令，通过预设的价值函数，计算每个所述调度指令在所述调度状态信息下的价值数值，其中，所述调度指令由所述仿真矿车的出发地和目的地组成；将数值最大的所述价值数值对应的调度指令确定为目标调度指令，并将所述目标调度指令发送至所述仿真系统中的仿真矿车中；在所述仿真矿车执行完成所述目标调度指令时，更新所述矿区生产信息。

[0080] 在具体的应用场景中，所述混合奖励函数由人工奖励函数和仿真奖励函数组成；所述奖励模块33，具体可用于在所述时间节点信息中，提取出所述目标调度指令对应的行驶时间和预期等待时间，并根据所述行驶时间和所述预期等待时间之和，得到所述目标调度指令对应的行程时间；将所述目标调度指令对应的行程时间输入至所述人工奖励函数中，得到所述目标调度指令的人工奖励值；根据所述更新后的矿区生产信息，计算所述目标调度指令从执行开始至执行完成之间的仿真装载设备利用率；将所述仿真装载设备利用率输入至所述仿真奖励函数中，得到所述目标调度指令的仿真奖励值；根据所述人工奖励值和所述仿真奖励值，得到所述目标调度指令的动作奖励。

[0081] 在具体的应用场景中，所述奖励模块33，具体可用于根据所述更新后的矿区生产信息，计算智能体性能指标，并根据所述智能体性能指标，确定人工奖励权重系数；根据所述人工奖励权重系数，计算仿真奖励权重系数，其中，所述人工奖励权重系数和所述仿真奖励权重系数的和为预设值；根据所述人工奖励权重系数与所述人工奖励值的乘积与所述仿真奖励权重系数与所述仿真奖励值的乘积的和值，得到所述目标调度指令的动作奖励。

[0082] 在具体的应用场景中，所述奖励模块33，具体可用于在所述更新后的矿区生产信

息中，提取出预设时间段内的矿区产量信息，并将所述预设时间段内的矿区产量信息划分为多个子时间段内的矿区产量信息；对所述多个子时间段内的矿区产量信息进行线性拟合，得到性能指标拟合斜率，并计算所述多个子时间段内的矿区产量信息的平均值，得到性能指标均值；判断所述性能指标拟合斜率是否小于预设的斜率阈值，并判断所述性能指标均值是否大于预设的性能指标阈值，其中，所述斜率阈值为负值；若所述性能指标拟合斜率小于所述斜率阈值，则对所述人工奖励权重系数进行递增计算；若所述性能指标均值小于等于所述性能指标阈值，则对所述人工奖励权重系数进行递增计算；若所述性能指标均值大于所述性能指标阈值，则对所述人工奖励权重系数进行递减计算。

[0083] 在具体的应用场景中，所述评价模块34，具体可用于在所述仿真系统中的矿区生产信息中，提取出预设时间段内的矿区产量信息，并将所述预设时间段内的矿区产量信息划分为多个子时间段内的矿区产量信息；计算所述多个子时间段内的矿区产量信息的平均值，得到性能指标均值，并计算所述多个子时间段的矿区产量信息与所述性能指标均值之间的偏差值；当所述偏差值小于预设的偏差阈值时，判定矿车调度模型训练完成，并结束所述矿车调度模型的训练。

[0084] 在具体的应用场景中，如图6所示，本装置还包括仿真模块35，所述仿真模块35可用于在所述仿真系统中模拟生成所述矿区生产信息，其中，所述矿区生产信息包括仿真路网信息、至少一个仿真装载设备、每个所述仿真装载设备的设备信息和设备状态、至少一个仿真卸载设备、每个所述仿真卸载设备的设备信息和设备状态、至少一个仿真矿车、每个所述仿真矿车的车辆信息和车辆状态、以及每个所述仿真装载设备和每个所述仿真卸载设备之间的行驶时间中的至少一种信息；在所述仿真系统中的仿真矿车发送车辆调度请求时，根据所述矿区生产信息，在所述人工系统中生成针对所述仿真矿车的时间节点信息，其中，所述时间节点信息包括所述调度指令集合中的每个所述调度指令对应的行驶时间、每个所述调度指令对应的预期等待时间、所述仿真系统中每个仿真装载设备的剩余服务时间，以及仿真系统中每个仿真卸载设备的剩余服务时间中的至少一种信息。

[0085] 在具体的应用场景中，如图6所示，本装置还包括更新模块36，所述更新模块36可用于根据所述目标调度指令的动作奖励，对所述价值函数进行更新，得到优化后的价值函数，应用所述优化后的价值函数在所述调度指令集合中选取优化后的目标调度指令。

[0086] 需要说明的是，本实施例提供的一种矿车调度模型训练装置所涉及各功能单元的其它相应描述，可以参考实施例一和实施例二中的对应描述，在此不再赘述。

[0087] 实施例四

本发明实施例还提供了一种计算机设备的实体结构图，如图7所示，该计算机设备包括：处理器41、存储器42、及存储在存储器42上并可在处理器上运行的计算机程序，其中存储器42和处理器41均设置在总线43上所述处理器41执行所述程序时实现实施例一和实施例二所述的方法步骤。实施例一和实施例二中已经对矿车调度模型训练方法进行了详细的描述，在此不再赘述。

[0088] 实施例五

图8为本发明实施例提供的一种芯片的结构示意图，如图8所示，芯片500包括一个或两个以上(包括两个)处理器510和通信接口530。所述通信接口530和所述至少一个处理器510耦合，所述至少一个处理器510用于运行计算机程序或指令，以实现如实施例一和实

施例二所述的方法步骤。

[0089] 优选地，存储器540存储了如下的元素：可执行模块或者数据结构，或者他们的子集，或者他们的扩展集。

[0090] 本发明实施例中，存储器540可以包括只读存储器和随机存取存储器，并向处理器510提供指令和数据。存储器540的一部分还可以包括非易失性随机存取存储器(non-volatile random access memory，NVRAM)。

[0091] 本发明实施例中，存储器540、通信接口530以及存储器540通过总线系统520 耦合在一起。其中，总线系统520除包括数据总线之外，还可以包括电源总线、控制总线和状态信号总线等。为了便于描述，在图10中将各种总线都标为总线系统520。

[0092] 上述本申请实施例描述的方法可以应用于处理器510中，或者由处理器510实现。处理器510可能是一种集成电路芯片，具有信号的处理能力。在实现过程中，上述方法的各步骤可以通过处理器510中的硬件的集成逻辑电路或者软件形式的指令完成。上述的处理器510可以是通用处理器(例如，微处理器或常规处理器)、数字信号处理器(digital signal processing，DSP)、专用集成电路(application specific integrated circuit，ASIC)、现成可编程门阵列(field-programmable gate array，FPGA)或者其他可编程逻辑器件、分立门、晶体管逻辑器件或分立硬件组件，处理器510可以实现或者执行本发明实施例中的公开的各方法、步骤及逻辑框图。

[0093] 实施例六

图9为本发明实施例提供的一种终端的结构示意图，如图9所示，终端600包括上述矿车调度模型训练装置100。

[0094] 上述终端600可以通过矿车调度模型训练装置100执行上述实施例一和实施例二所描述的方法。可以理解，终端600对矿车调度模型训练装置100进行控制的实现方式，可以根据实际应用场景设定，本申请实施例不作具体限定。

[0095] 所述终端600包括但不限于：车辆、车载终端、车载控制器、车载模块、车载模组、车载部件、车载芯片、车载单元、车载雷达或车载摄像头等其他传感器，车辆可通过该车载终端、车载控制器、车载模块、车载模组、车载部件、车载芯片、车载单元、车载雷达或摄像头，实施本申请提供的方法。本申请中的车辆包括乘用车和商用车，商用车的常见车型包括但不限于：皮卡、微卡、轻卡、微客，自卸车、载货车、牵引车、挂车、专用车和矿用车辆等。矿用车辆包括但不限于矿卡、宽体车、铰接车、挖机、电铲、推土机等。本申请对智能车的类型不作进一步限定，任何一种车型均在本申请的保护范围内。

[0096] 本发明实施例中的终端作为一种执行非电变量的控制或调整系统，充分结合了仿真系统和人工系统的互补优势，使调度智能体能够同时从仿真系统和人工系统中学习调度策略，增加了行为奖励的密度，使基于强化学习的车辆调度模型可以及时的获取到与调度指令相关的行为奖励，极大的提高了获取奖励反馈的及时性，降低了车辆调度模型的训练时间成本，并提高了车辆调度模型的训练效率和模型性能。

[0097] 实施例七

基于上述如图1和图2所示方法，相应的，本发明实施例还提供了一种计算机可读存储介质，如图10所示，存储器720上存储有计算机程序，该计算机程序位于程序代码空间730，该程序731被处理器710执行时实现实施例一和实施例二所述的方法步骤。实施例一和

实施例二中已经对矿车调度模型训练方法进行了详细的描述，在此不再赘述。

[0098] 上述实施例中描述的方法可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。计算机可读介质可以包括计算机存储介质和通信介质，还可以包括任何可以将计算机程序从一个地方传送到另一个地方的介质。存储介质可以是可由计算机访问的任何目标介质。

[0099] 作为一种可能的设计，计算机可读介质可以包括紧凑型光盘只读储存器(compact disc read-only memory，CD-ROM)、RAM、ROM、EEPROM或其它光盘存储器；计算机可读介质可以包括磁盘存储器或其它磁盘存储设备。而且，任何连接线也可以被适当地称为计算机可读介质。例如，如果使用同轴电缆、光纤电缆、双绞线、DSL或无线技术(如红外，无线电和微波)从网站、服务器或其它远程源传输软件，则同轴电缆、光纤电缆、双绞线、DSL或诸如红外、无线电和微波之类的无线技术包括在介质的定义中。如本文所使用的磁盘和光盘包括光盘(CD)，激光盘，光盘，数字通用光盘(digital versatile disc，DVD)，软盘和蓝光盘，其中磁盘通常以磁性方式再现数据，而光盘利用激光光学地再现数据。

[0100] 显然，本领域的技术人员应该明白，上述的本发明的各模块或各步骤可以用通用的计算装置来实现，它们可以集中在单个的计算装置上，或者分布在多个计算装置所组成的网络上，可选地，它们可以用计算装置可执行的程序代码来实现，从而，可以将它们存储在存储装置中由计算装置来执行，并且在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤，或者将它们分别制作成各个集成电路模块，或者将它们中的多个模块或步骤制作成单个集成电路模块来实现。这样，本发明不限制于任何特定的硬件和软件结合。

[0101] 以上所述仅为本发明的优选实施例而已，并不用于限制本发明，对于本领域的技术人员来说，本发明可以有各种更改和变化。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包括在本发明的保护范围之内。

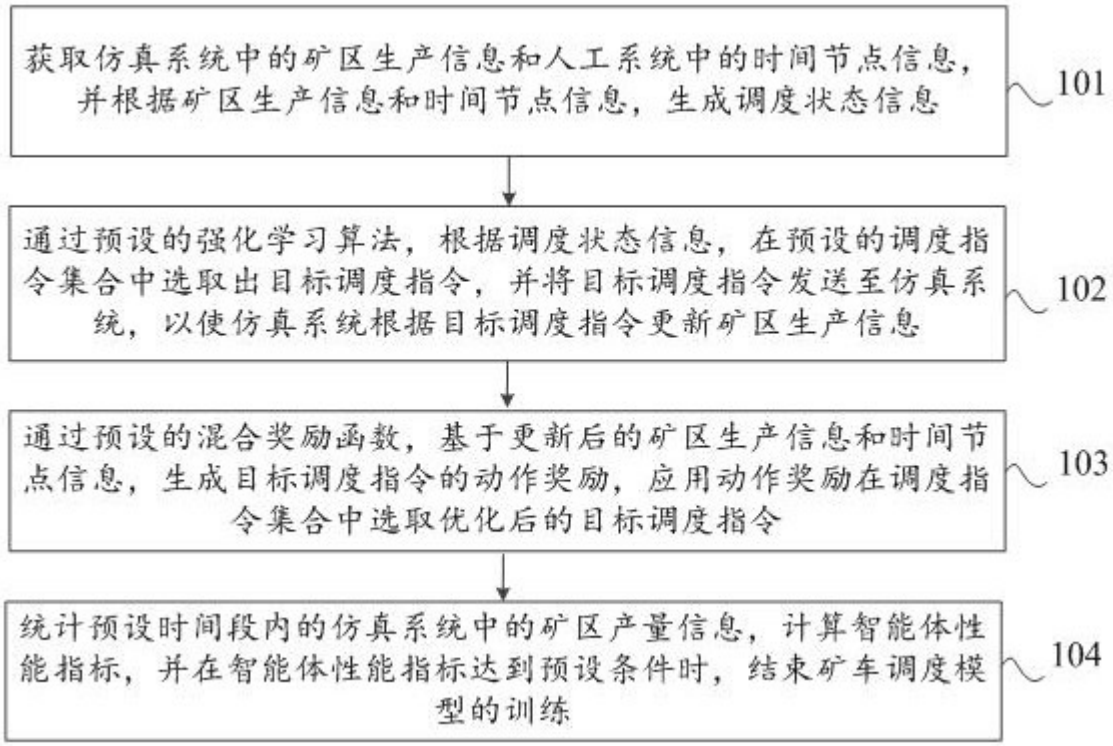

获取仿真系统中的矿区生产信息和人工系统中的时间节点信息，并根据矿区生产信息和时间节点信息，生成调度状态信息
101
通过预设的强化学习算法，根据调度状态信息，在预设的调度指令集合中选取出目标调度指令，并将目标调度指令发送至仿真系统，以使仿真系统根据目标调度指令更新矿区生产信息
102
通过预设的混合奖励函数，基于更新后的矿区生产信息和时间节点信息，生成目标调度指令的动作奖励，应用动作奖励在调度指令集合中选取优化后的目标调度指令
103
统计预设时间段内的仿真系统中的矿区产量信息，计算智能体性能指标，并在智能体性能指标达到预设条件时，结束矿车调度模型的训练
104

图1

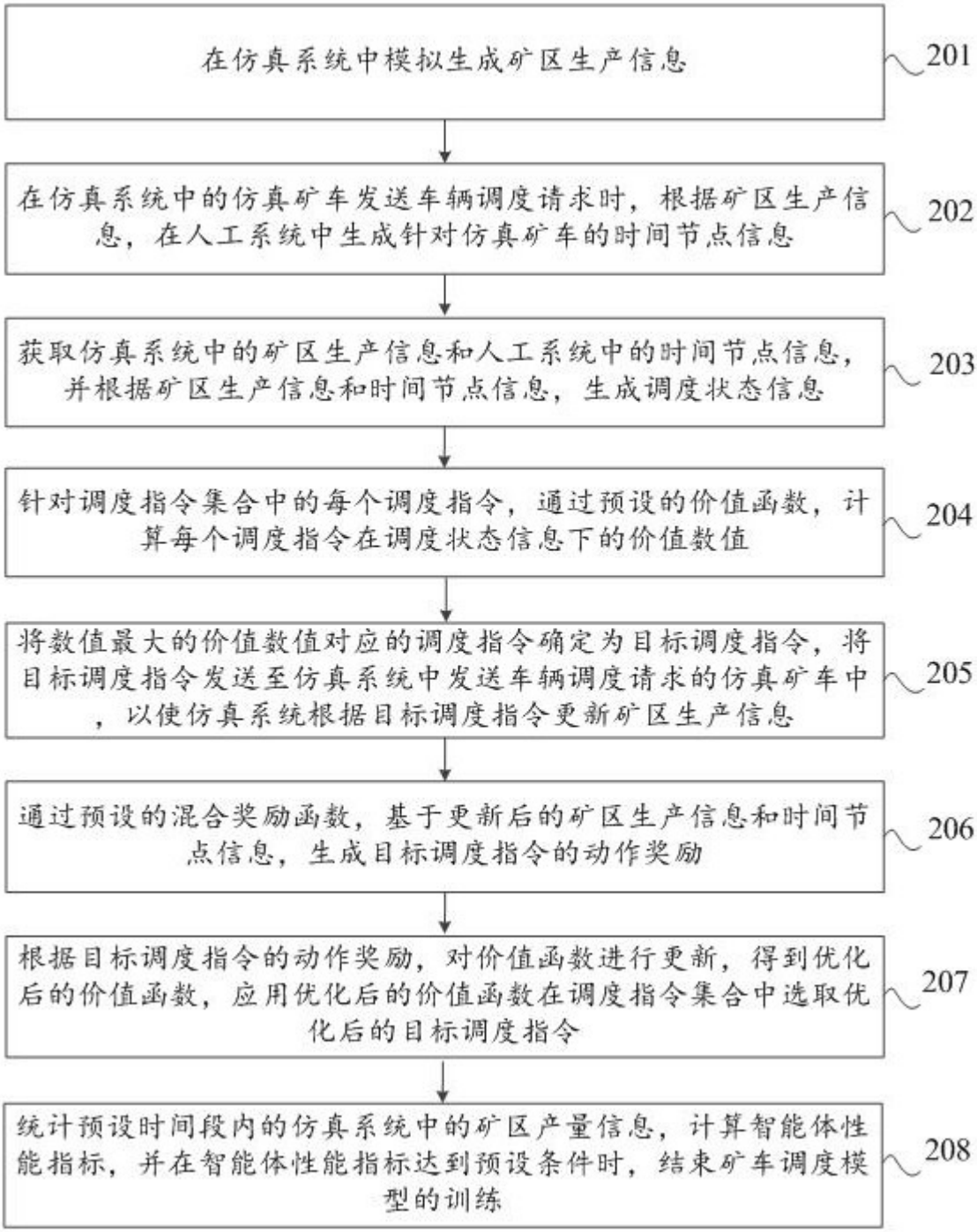

在仿真系统中模拟生成矿区生产信息
201
在仿真系统中的仿真矿车发送车辆调度请求时，根据矿区生产信息，在人工系统中生成针对仿真矿车的时间节点信息
202
获取仿真系统中的矿区生产信息和人工系统中的时间节点信息，并根据矿区生产信息和时间节点信息，生成调度状态信息
203
针对调度指令集合中的每个调度指令，通过预设的价值函数，计算每个调度指令在调度状态信息下的价值数值
204
将数值最大的价值数值对应的调度指令确定为目标调度指令，将目标调度指令发送至仿真系统中发送车辆调度请求的仿真矿车中，以使仿真系统根据目标调度指令更新矿区生产信息
205
通过预设的混合奖励函数，基于更新后的矿区生产信息和时间节点信息，生成目标调度指令的动作奖励
206
根据目标调度指令的动作奖励，对价值函数进行更新，得到优化后的价值函数，应用优化后的价值函数在调度指令集合中选取优化后的目标调度指令
207
统计预设时间段内的仿真系统中的矿区产量信息，计算智能体性能指标，并在智能体性能指标达到预设条件时，结束矿车调度模型的训练
208

图2

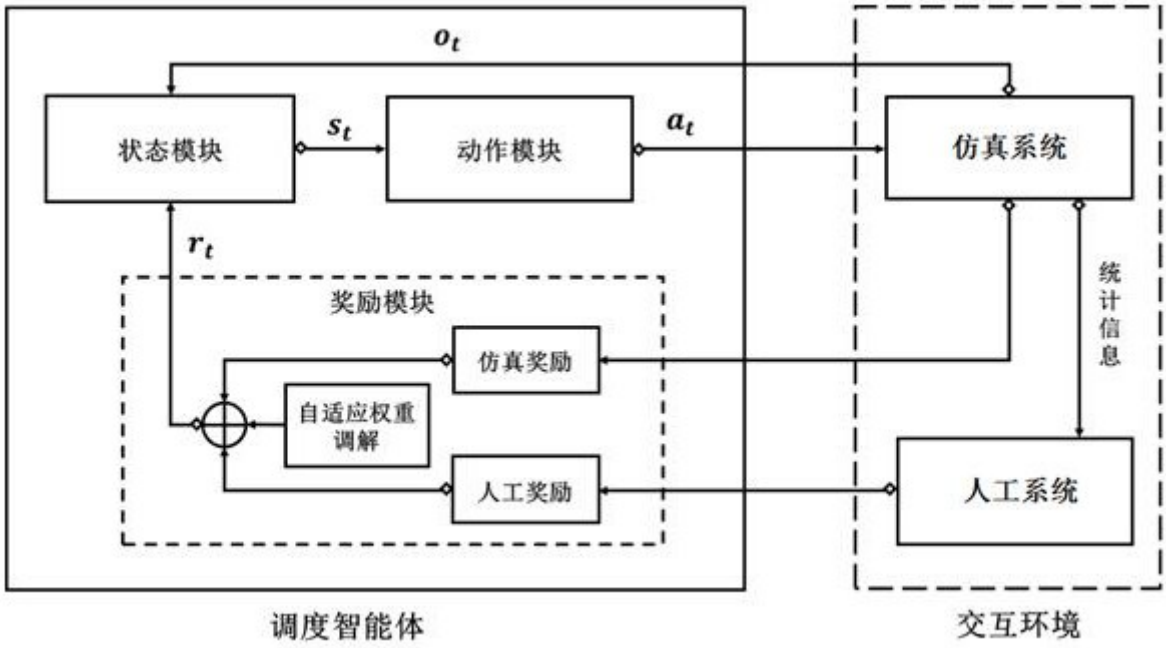

o_t
状态模块
s_t
动作模块
a_t
仿真系统
r_t
奖励模块
仿真奖励
自适应权重调解
人工奖励
统计信息
人工系统
调度智能体
交互环境

图3

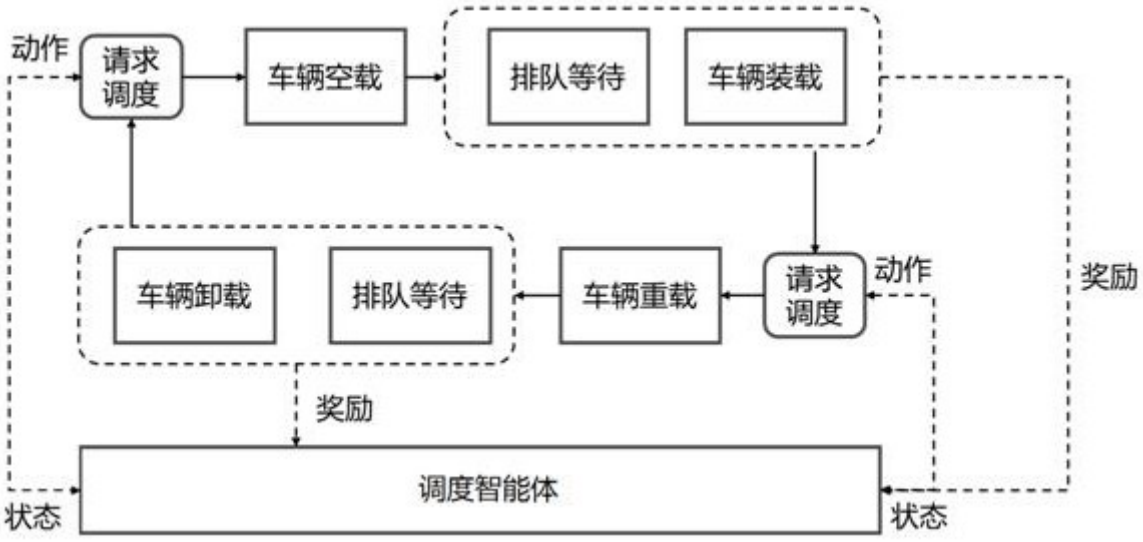

动作
请求调度
车辆空载
排队等待
车辆装载
车辆卸载
排队等待
车辆重载
请求调度
动作
奖励
奖励
调度智能体
状态
状态

图4

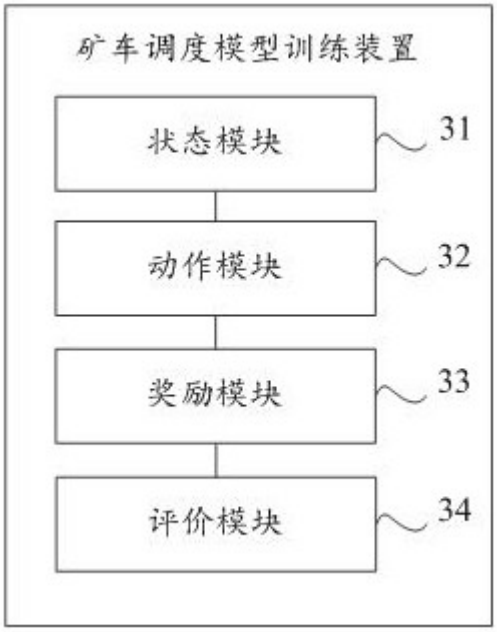

矿车调度模型训练装置
状态模块
31
动作模块
32
奖励模块
33
评价模块
34

图5

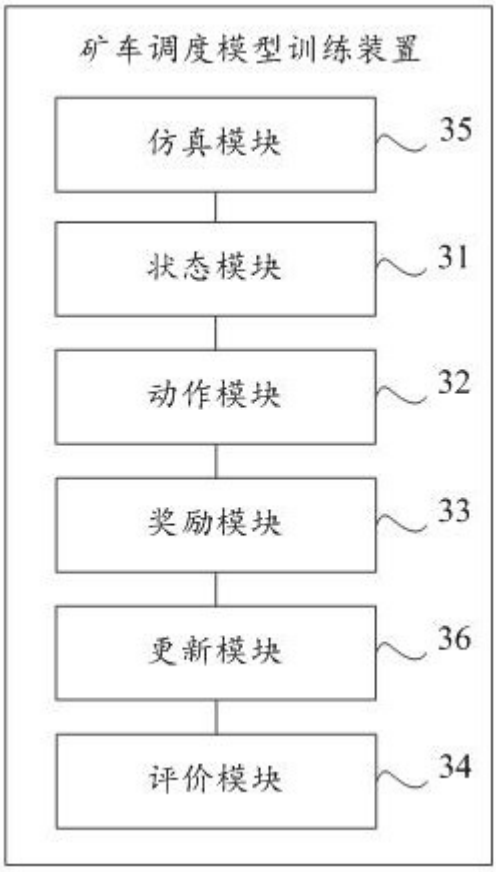

矿车调度模型训练装置
仿真模块
35
状态模块
31
动作模块
32
奖励模块
33
更新模块
36
评价模块
34

图6

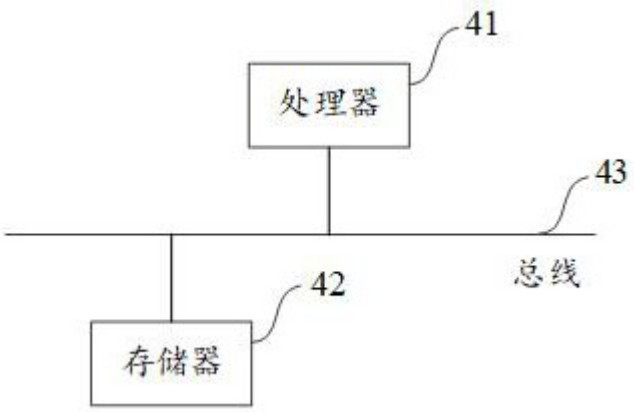

41
处理器
43
总线
42
存储器

图7

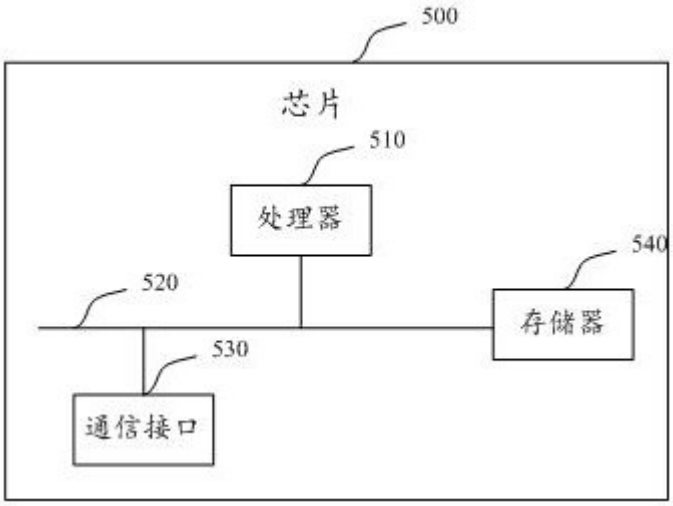

500
芯片
510
处理器
540
520
存储器
530
通信接口

图8

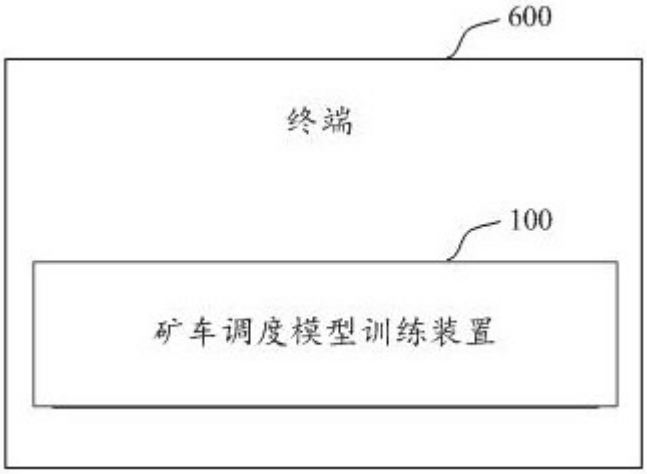

600
终端
100
矿车调度模型训练装置

图9

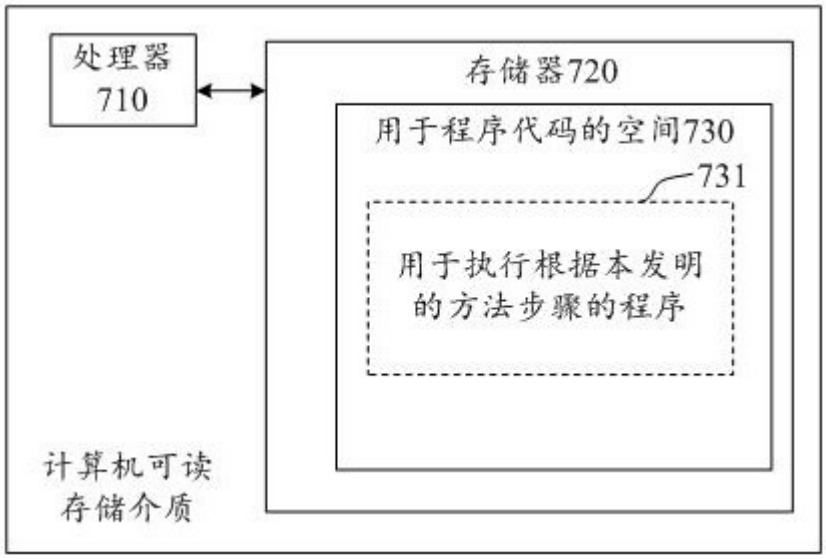

处理器
710
存储器720
用于程序代码的空间730
731
用于执行根据本发明
的方法步骤的程序
计算机可读
存储介质

图10