文章编号：1001-4632（2022）05-0146-11

引用格式：王荣笙，张琦，张涛，等．基于蒙特卡罗树搜索-强化学习的列车运行智能调整方法[J]．中国铁道科学，2022，43(5)：146-156.

Citation：WANG Rongsheng，ZHANG Qi，ZHANG Tao，et al. Intelligent Adjustment Approach for Train Operation Based on Monte Carlo Tree Search-Reinforcement Learning [J]. China Railway Science，2022，43 (5)：146-156.

基于蒙特卡罗树搜索-强化学习的列车运行智能调整方法

王荣笙[1,2,3]，张 琦[2,3]，张 涛[2,3]，王 涛[2,3]，丁舒忻[2,3]

（1. 中国铁道科学研究院 研究生部，北京 100081；
2. 中国铁道科学研究院集团有限公司 通信信号研究所，北京 100081；
3. 中国铁道科学研究院集团有限公司 国家铁路智能运输系统工程技术研究中心，北京 100081）

摘 要：为提升突发事件下高速铁路应急处置效率，以列车运行图为研究对象，提出晚点场景下蒙特卡罗树搜索－强化学习（MCTS-RL）的列车运行智能调整方法，包括MCTS-RL的列车运行智能调整离线训练模型、强化学习方法、MCTS的发车次序决策方法和冲突消解启发式规则。基于高速铁路列车运行调整数学模型构建强化学习环境，包括状态集、动作集、状态转移概率和奖励函数。先设计启发式规则，生成可行发车次序，作为蒙特卡罗树搜索博弈树结构的节点，应用MCTS输出列车运行调整的最优发车次序。之后再设计启发式规则，消解列车在车站和区间的运行冲突。以线路上列车总晚点时间最短为目标函数，基于MCTS-RL一次性离线训练生成在线调整模型，用于实时调整各次列车在各个车站的发车次序。以京沪高速铁路北京南—泰安段为例，设置到站晚点和发车晚点场景，分别应用先到先服务、CPLEX求解器和MCTS-RL方法进行求解。结果表明：与CPLEX求解器下的方案相比，MCTS-RL方法能在0.001 s内给出同样最优的列车运行调整方案。

关键词：高速铁路；列车运行调整；人工智能；强化学习；蒙特卡罗树搜索

中图分类号：U292.41 **文献标识码**：A **doi**：10.3969/j.issn.1001-4632.2022.05.16

我国高速铁路已进入大规模网络化运营时期，其路网规模、行车密度、场景工况、旅客发送量以及运输组织复杂性均为世界高铁之最。巨大的客流压力和多变的运营场景下，高铁路网呈现出前所未有的时空复杂度。同时，我国高铁跨越高原、高热、高湿、大风、地震等复杂工况地区，可能导致列车产生大范围延误，此时需要进行列车运行调整工作，恢复正常运行秩序。目前，我国高速铁路列车运行调整仍以列车调度员凭经验处置为主，现场工作强度较大，也难以同时保证调整策略的实时性和近似最优性。

高速铁路列车运行调整问题具有NP难（NP-hard）特性[1-2]，该问题是指受突发事件影响，调整列车运行计划使列车恢复有序运行状态[3]。问题求解过程中列车和车站数量的增加会导致求解时间呈现指数级甚至阶乘式增长。国内外学者通常以晚点较小的扰动场景或晚点严重的干扰场景为出发点[4-5]，或基于运筹学方法[6-8]，或基于进化算法[9-10]，对突发事件下各列车在各车站的进路、接发车时刻、发车次序进行调整或协同优化[11-13]，力求获取近似最优的调整策略。但上述方法均需自行设计模型分支定界或启发式规则，模型构造严重依赖于个体经验，同时得到的模型在加快算法收敛速度和搜索近似最优解等方面的表现仍不理想。

收稿日期：2021-06-29；修订日期：2022-03-31

基金项目：国家自然科学基金高铁联合基金资助项目（U1834211，U1934220）；中国国家铁路集团有限公司科技研究开发计划重大课题（K2019G043）

第一作者：王荣笙（1994—），男，辽宁辽阳人，博士研究生。E-mail：wrs20138437@126.com

通讯作者：张 琦（1968—），男，上海人，研究员，博士。E-mail：gorgeous@139.com

以强化学习为代表的人工智能方法在实时求解列车运行最优调整方案上具有独特优势。强化学习方法通过智能体与环境之间的不断试错学习，以获取奖励函数最大（目标函数最优）的学习策略，生成的离线训练模型可直接用于问题的在线实时求解，无须对研究问题重新建模[14]，即采用离线训练、在线调整的形式就能很好地同时满足调整策略在实时性和近似最优性方面的需求。目前，强化学习在软件项目分配方案、库存管理、车间作业调度等调度优化问题中得到广泛应用[15−17]，部分学者也将其应用到列车运行调整问题中。如文献［18］通过分析铁路设施基础布局构建强化学习环境，离线训练生成的模型能实时优化初始晚点下的时刻表；文献［19−20］基于强化学习方法确定了不同优先级列车占用车站股道的次序；文献［21］利用深度强化学习方法优化列车在车站的发车次序，生成了列车总晚点时间最短的运行图调整方案。目前的研究虽然从宏观、微观不同角度构建出列车运行调整的强化学习环境，但在强化学习策略最优性验证方面的研究较少，仍存在极大的改善空间。

本文面向人工智能方法应用于列车运行调整的迫切需求，基于列车调度员的调图视角提出蒙特卡罗树搜索-强化学习（Monte Carlo Tree Search-Reinforcement Learning，MCTS-RL）的列车运行智能调整方法，包括MCTS-RL的列车运行智能调整离线训练模型、强化学习方法、MCTS的发车次序决策方法和冲突消解启发式规则。通过MCTS-RL方法一次性离线训练生成在线调整模型，用于实时调整晚点场景下的实绩运行图，并通过与CPLEX求解器下的运行图调整方案进行对比，验证MCTS-RL方法下学习策略的最优性。

1　高速铁路列车运行调整数学模型

1.1　问题描述

高速铁路列车运营中，突发事件会造成列车在车站的到达晚点或出发晚点，在列车运行图中表现为列车运行线的偏移，此时需要综合考虑列车的运行情况，通过调整各列车在各车站的接、发车时刻和发车次序，给出总晚点时间最短的列车运行调整策略，以保证列车运行效率。

列车在车站和区间的作业时间示意图如图1所示。图中：L为线路上列车总数，列车$l\in\{1,2,\cdots,L\}$；S为线路上车站数量，车站$s\in\{1,2,\cdots,S\}$；$\Gamma^{\mathrm{a}}_{l,s}$和$\Gamma^{\mathrm{d}}_{l,s}$分别为列车$l$在车站$s$的实际到站时刻和实际发车时刻；$\Gamma^{\mathrm{a}}_{l,s+1}$为列车$l$在车站$s+1$的实际到站时刻；$t_{l,s,s+1}$为列车$l$在区间$(s,s+1)$的实际区间运行时间；$\Gamma^{x}_{l,s,s+1}$和$\Gamma^{x}_{l+1,s,s+1}$分别为相邻2列列车$l$和$l+1$在区间$(s,s+1)$内通过任意位置$x$的通过时刻。由图1可知：以列车调度员调整列车运行图为视角，可将列车运行调整过程拆解为2个阶段：首先选择列车在车站的发车次序，之后消解列车在车站和区间的运行冲突，这样一来，合理调整接、发车时刻$\Gamma^{\mathrm{a}}_{l,s}$和$\Gamma^{\mathrm{d}}_{l,s}$，可使所有列车在各站的总晚点时间最短。由此，列车运行调整可描述为以列车总晚点时间最短为优化目标，按时间顺序给出列车在沿线各车站最优发车次序的动态规划过程。

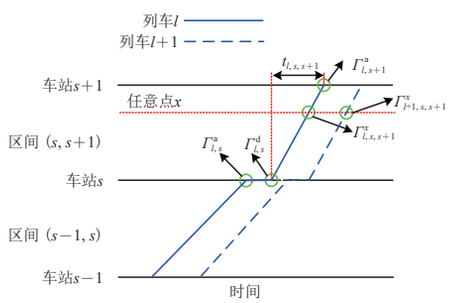


图1　列车在车站和区间的作业时间示意图

同时，为研究方便且不失高速铁路列车运行调整的一般实际性，做出如下基本假设：

（1）初始晚点发生后，线路将不再产生向其他线路传播的晚点；

（2）列车在车站的实际到达和出发时刻不早于图定时刻；

（3）相邻2列列车的到达—发车和发车—到达作业若发生在同一股道，会产生作业时间冲突。因这种情况在实际中极少，可视为以上2种作业全部在不同股道进行，互不影响。

1.2　数学模型

1.2.1　目标函数

突发事件引起列车晚点时，铁路运营方关注更多的是在调整各列车在各车站的接、发车时刻后，使线路上列车总晚点时间最短。故定义高速铁路列车运行调整数学模型的目标函数Z为列车实际到站、发车时刻与图定到站、发车时刻的偏差之和的最小值，即

$$Z=\min\sum_{l=1}^{L}\sum_{s=1}^{S}\left[(\Gamma^{\mathrm{a}}_{l,s}-\bar{\Gamma}^{\mathrm{a}}_{l,s})+(\Gamma^{\mathrm{d}}_{l,s}-\bar{\Gamma}^{\mathrm{d}}_{l,s})\right]\quad(1)$$

式中：$\Gamma^{\mathrm{a}}_{l,s}$和$\bar{\Gamma}^{\mathrm{a}}_{l,s}$分别表示列车$l$在车站$s$的实际和

图定到站时刻；$\Gamma_{l,s}^{\mathrm{d}}$和$\bar{\Gamma}_{l,s}^{\mathrm{d}}$分别表示列车$l$在车站$s$的实际和图定发车时刻。

1.2.2 约束条件

高铁列车在线路上运行时，需要考虑车站作业时间约束和区间作业时间约束。

1）车站作业时间约束

为保证列车在车站到站、发车和接发旅客等基础作业的可行性，根据假设（2），列车l实际的到站时刻和发车时刻不应早于对应的图定时刻，即

$$\Gamma_{l,s}^{\mathrm{a}} \geqslant \bar{\Gamma}_{l,s}^{\mathrm{a}} \tag{2}$$

$$\Gamma_{l,s}^{\mathrm{d}} \geqslant \bar{\Gamma}_{l,s}^{\mathrm{d}} \tag{3}$$

对于经停车站s的列车l，其实际停站时间应符合最小值约束，即列车l在车站s的实际停站时间不小于该车在该站的最小停站时间。值得注意的是，停站列车应保证旅客的正常上下车，故停站列车的作业不能由“停站”改为“通过”，但通过作业的列车若为低等级列车，可将其作业由“通过”改为“停站”，供后行高等级列车越行

$$\Gamma_{l,s}^{\mathrm{d}} - \Gamma_{l,s}^{\mathrm{a}} \geqslant t_{l,s}^{\min} \tag{4}$$

对于列车l经停的车站s，其接发列车数量n_s应符合最大值$n_s^{\max}$约束，即n_s不大于车站s可接发列车的最大数量$n_s^{\max}$

$$n_s \leqslant n_s^{\max} \tag{5}$$

当有相邻2列列车l和$l+1$在车站s相继执行到达、通过和发车作业时，涉及到的车站作业间隔时间共有7种，分别为：通过—通过间隔时间$I_{l,l+1,s}^{\mathrm{pp}}$；通过—发车间隔时间$I_{l,l+1,s}^{\mathrm{pd}}$；通过—到达间隔时间$I_{l,l+1,s}^{\mathrm{pa}}$；到达—到达间隔时间$I_{l,l+1,s}^{\mathrm{aa}}$；到达—通过间隔时间$I_{l,l+1,s}^{\mathrm{ap}}$；发车—发车间隔时间$I_{l,l+1,s}^{\mathrm{dd}}$；发车—通过间隔时间$I_{l,l+1,s}^{\mathrm{dp}}$。7种车站作业间隔时间均存在最小值约束，不同类型车站间隔时间的最小值与车站类型、道岔操作方式等因素有关。为研究方便且不失实际性，令上述7种车站作业间隔时间的最小值均为$I_s^{\min}$（实际可根据车站具体要求进行修改），即

$$\begin{cases} I_{l,l+1,s}^{\mathrm{pp}} = \Gamma_{l,s}^{\mathrm{p}} - \Gamma_{l+1,s}^{\mathrm{p}} & I_{l,l+1,s}^{\mathrm{pp}} \geqslant I_s^{\min} \\ I_{l,l+1,s}^{\mathrm{pd}} = \Gamma_{l,s}^{\mathrm{p}} - \Gamma_{l+1,s}^{\mathrm{d}} & I_{l,l+1,s}^{\mathrm{pd}} \geqslant I_s^{\min} \\ I_{l,l+1,s}^{\mathrm{pa}} = \Gamma_{l,s}^{\mathrm{p}} - \Gamma_{l+1,s}^{\mathrm{a}} & I_{l,l+1,s}^{\mathrm{pa}} \geqslant I_s^{\min} \\ I_{l,l+1,s}^{\mathrm{aa}} = \Gamma_{l,s}^{\mathrm{a}} - \Gamma_{l+1,s}^{\mathrm{a}} & I_{l,l+1,s}^{\mathrm{aa}} \geqslant I_s^{\min} \\ I_{l,l+1,s}^{\mathrm{ap}} = \Gamma_{l,s}^{\mathrm{a}} - \Gamma_{l+1,s}^{\mathrm{p}} & I_{l,l+1,s}^{\mathrm{ap}} \geqslant I_s^{\min} \\ I_{l,l+1,s}^{\mathrm{dd}} = \Gamma_{l,s}^{\mathrm{d}} - \Gamma_{l+1,s}^{\mathrm{d}} & I_{l,l+1,s}^{\mathrm{dd}} \geqslant I_s^{\min} \\ I_{l,l+1,s}^{\mathrm{dp}} = \Gamma_{l,s}^{\mathrm{d}} - \Gamma_{l+1,s}^{\mathrm{p}} & I_{l,l+1,s}^{\mathrm{dp}} \geqslant I_s^{\min} \end{cases} \tag{6}$$

式中：$\Gamma_{l,s}^{\mathrm{p}}$和$\Gamma_{l+1,s}^{\mathrm{p}}$分别为列车$l$和列车$l+1$在车站$s$的通过时刻。

根据假设（3），故式（6）中不再考虑到达—发车间隔作业和发车—到达间隔作业。

2）区间作业时间约束

令$\Gamma_{l,l+1,s}^{x}$为相邻2列列车l和$l+1$在区间（s，$s+1$）的实际追踪列车间隔时间，那么列车在该区间的实际区间运行时间$t_{l,s,s+1}$和这2列列车的追踪列车间隔时间$I_{l,l+1,s}^{x}$应分别满足最小值约束，即

$$t_{l,s,s+1} = \Gamma_{l,s+1}^{\mathrm{a}} - \Gamma_{l,s}^{\mathrm{d}} \qquad t_{l,s,s+1} \geqslant t_{l,s,s+1}^{\min} \tag{7}$$

$$I_{l,l+1,s}^{x} = \Gamma_{l,s,s+1}^{x} - \Gamma_{l+1,s,s+1}^{x} \quad I_{l,l+1,s}^{x} \geqslant I_{l,l+1}^{\min} \tag{8}$$

式中：$t_{l,s,s+1}^{\min}$为列车l在区间（s,$s+1$）的最小区间运行时间；$I_{l,l+1}^{\min}$为列车l和$l+1$在区间（s,$s+1$）的最小追踪列车间隔时间。

2 列车运行智能调整方法

要采用强化学习求解建立的高速铁路列车运行调整数学模型，需要分析强化学习机制与列车运行调整过程之间的对应关系，构建列车运行智能调整离线训练模型中的强化学习环境和智能体。对列车运行调整方案求解时，为了计算模型中列车总晚点时间最短下的列车发车次序，提出蒙特卡罗树搜索的发车次序决策方法；为了消解模型中列车在车站和区间的运行冲突，提出启发式规则。

2.1 蒙特卡罗树搜索-强化学习的列车运行智能调整离线训练模型

列车运行调整过程具有马尔可夫性质，即未来车站状态下的发车次序信息仅与当前车站状态有关，与过去车站状态的历史信息无关。强化学习方法本质上是1种基于动态规划思想且具有马尔可夫性质的半监督机器学习方法[14]，包括智能体和环境。智能体相当于决策者；环境包括状态集、动作集和奖励函数。采用强化学习离线训练—在线调整的机制，学习该过程的列车运行最优调整策略。

对于图1所示的列车运行调整过程来说，前一阶段选择列车在车站的最优发车次序时，采用蒙特卡罗树搜索（Monte Carlo Tree Search，MCTS）方法，该方法基于博弈树结构，整合了广度优先搜索和深度优先搜索的各自优点，被视为求解决策过程最优化的高效快速搜索方法之一[22]，并已在围棋人工智能AlphaGo的策略选择问题中得到充分应用[23-24]；后一阶段消解列车在车站和区间的运行冲突时，设计并运用启发式规则。

基于高速铁路列车运行调整数学模型和列车运

行调整过程，构建强化学习方法的智能体和环境。其中：环境中的MCTS方法和启发式规则先后用于生成列车发车次序和消解列车运行冲突；智能体与环境不断交互学习生成最终离线训练模型。在列车运行调整过程中，当输入列车接车或者发车晚点时，该模型可直接用于列车运行调整问题的实时求解，无须重新离线训练。由此得到基于蒙特卡罗树搜索-强化学习（Monte Carlo Tree Search-Reinforcement Learning，MCTS-RL）方法下的列车运行智能调整离线训练模型，其流程图如图2所示。图中：$\boldsymbol{\mathcal{S}}_s$，$A_s$，$R_s$分别为强化学习训练至车站$s$时的状态集、动作集和奖励函数。

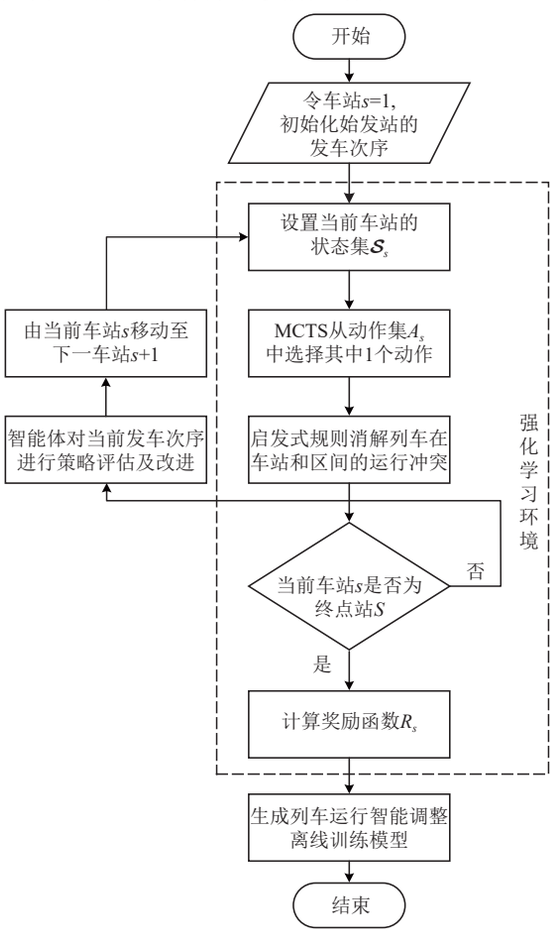


图2　列车运行智能调整离线训练模型流程图

图2描述了智能体与列车运行调整强化学习环境不断交互，搜索列车运行最优调整策略的离线训练过程，步骤如下。

步骤1：智能体观测当前车站s的状态集$\boldsymbol{\mathcal{S}}_s$，并基于MCTS从动作集$A_s$中随机选择1个动作；

步骤2：应用启发式规则检测并消解当前车站s及下一区间（s,s+1）的列车运行冲突，然后判定当前车站s的状态集$\boldsymbol{\mathcal{S}}_s$是否为终止状态（是否调整至终点站）；

步骤3：若当前车站s的状态集$\boldsymbol{\mathcal{S}}_s$不是终止状态，则更新至下一车站$s+1$的状态集$\boldsymbol{\mathcal{S}}_{s+1}$，并继续确定所有列车在该车站的动作集；

步骤4：若当前车站s的状态集$\boldsymbol{\mathcal{S}}_s$处于终止状态（即已调整至终点站$S$），则表明从始发站训练至终点站$S$的1次训练片段结束，此时记录所有列车在之前所有车站（1,…,s,…,S）的动作集集合，组成强化学习策略，计算奖励函数R_s（即目标函数值）并传递给智能体，供其评估和改进学习策略，然后进入下一次训练片段，形成智能体和强化学习环境试错学习的闭环反馈过程；

步骤5：若当前训练次数未达到最大值时，则令当前的调整车站为始发站，转至步骤1继续训练；否则，输出此时的列车运行智能调整离线训练模型，模型中的学习策略可直接用于列车运行图的实时调整。

2.2　强化学习方法

根据建立的数学模型和图2所示的离线训练模型流程图，设计列车运行智能调整的强化学习环境。

1）状态集$\boldsymbol{\mathcal{S}}_s$

列车l在车站s通过时，列车l的发车时刻$\Gamma^{\mathrm{d}}_{l,s}$等于对应的到站时刻$\Gamma^{\mathrm{a}}_{l,s}$；否则，当列车$l$在车站$s$停站时，其发车时刻$\Gamma^{\mathrm{d}}_{l,s}$与到站时刻$\Gamma^{\mathrm{a}}_{l,s}$应满足式（4）最小停站时间的约束。因此按列车调度员调整列车实际接发车时刻的视角，将$\boldsymbol{\mathcal{S}}_s$设置为智能体训练学习至车站$i$时列车的实际接发车时刻$\Gamma^{\mathrm{a}}_{1,s}$，$\Gamma^{\mathrm{d}}_{1,s}$，…,$\Gamma^{\mathrm{a}}_{l,s}$，$\Gamma^{\mathrm{d}}_{l,s}$，…,$\Gamma^{\mathrm{a}}_{L,s}$，$\Gamma^{\mathrm{d}}_{L,s}$组成的矩阵，即

$$\boldsymbol{\mathcal{S}}_s=\begin{pmatrix}\Gamma^{\mathrm{a}}_{1,s} & \Gamma^{\mathrm{d}}_{1,s}\\ \cdots & \cdots\\ \Gamma^{\mathrm{a}}_{l,s} & \Gamma^{\mathrm{d}}_{l,s}\\ \cdots & \cdots\\ \Gamma^{\mathrm{a}}_{L,s} & \Gamma^{\mathrm{d}}_{L,s}\end{pmatrix} \tag{9}$$

式中：$\boldsymbol{\mathcal{S}}_s$中第1列表示所有列车在当前车站$s$下的到站时刻，由上一车站状态集$\boldsymbol{\mathcal{S}}_{s-1}$下的发车时刻（$\Gamma^{\mathrm{d}}_{1,s-1}$,…,$\Gamma^{\mathrm{d}}_{l,s-1}$,…,$\Gamma^{\mathrm{d}}_{L,s-1}$）和上述所有列车在区间（$s$−1,$s$）的实际区间运行时间决定；$\boldsymbol{\mathcal{S}}_s$中第2列表示所有列车在当前车站$s$下的发车时刻，由动作集$A_s$决定。

2）动作集A_s

将A_s设置为列车在车站s所有发车次序情形的集合，即所有列车在车站s的第1种发车次序为e_1，第2种发车次序为e_2，一直到第$L!$种发车次序为$e_{L!}$，有

$$A_s=\{e_1,e_2,\cdots,e_{L!}\} \tag{10}$$

结合式（9），调整$\boldsymbol{\mathcal{S}}_s$中第2列（实际发车时刻）的向量顺序，形成不同发车次序下的动作集。

3）状态转移概率$P(\boldsymbol{\mathcal{S}}_{s+1}|\boldsymbol{\mathcal{S}}_s,A_s)$

表示当列车处于当前车站s的状态集$\boldsymbol{\mathcal{S}}_s$和动作集$A_s$时，转移到下一车站$s+1$的状态集$\boldsymbol{\mathcal{S}}_{s+1}$的概率。若当前车站不是终点站，则一定会发生状态转移，由$\boldsymbol{\mathcal{S}}_s$转移至$\boldsymbol{\mathcal{S}}_{s+1}$，即$P(\boldsymbol{\mathcal{S}}_{s+1}|\boldsymbol{\mathcal{S}}_s,A_s)=1$；若当前车站是终点站，则一次训练片段结束，不再进行状态转移，即$P(\boldsymbol{\mathcal{S}}_{s+1}|\boldsymbol{\mathcal{S}}_s,A_s)=0$，此时输出奖励函数。

4）奖励函数R

将R视为高速铁路列车运行调整数学模型的目标函数，对应于式（1）列车总晚点时间$\sum_{l=1}^{L}\sum_{i=s}^{S}\left[(\Gamma_{l,s}^{\mathrm{a}}-\bar{\Gamma}_{l,s}^{\mathrm{a}})+(\Gamma_{l,s}^{\mathrm{d}}-\bar{\Gamma}_{l,s}^{\mathrm{d}})\right]$，奖励函数$R$设置为列车总晚点时间的负值，即

$$R=-\sum_{l=1}^{L}\sum_{i=s}^{S}\left[(\Gamma_{l,s}^{\mathrm{a}}-\bar{\Gamma}_{l,s}^{\mathrm{a}})+(\Gamma_{l,s}^{\mathrm{d}}-\bar{\Gamma}_{l,s}^{\mathrm{d}})\right] \quad (11)$$

列车总晚点时间越短，奖励函数R值越大，说明列车运行调整策略越优。

5）智能体

强化学习智能体针对突发事件下列车晚点情况，在环境对约束条件（列车的车站作业时间和区间作业时间）的有效表征下，调整各列车在各车站的接发车时刻，故智能体相当于实际中给出列车运行调整计划的列车调度员。基于高速铁路列车运行调整数学模型设计强化学习方法的智能体和环境，智能体与环境的不断交互，最终生成总晚点时间最短的列车运行智能调整离线训练模型，模型策略可直接用于问题实时求解，无须重新离线训练。智能体中的学习策略π是所有状态下沿线各车站动作集的集合，表示从始发站调整至终点站1个完整的发车次序集合，故某个车站选择的发车次序不同导致每次训练的学习策略也不同。

2.3 发车次序决策方法

2.3.1 可行发车次序的启发式规则

从运行图来看，当晚点列车的运行线发生偏移后，智能体会综合考虑不同的发车次序构成的不同强化学习策略，并从中选择使列车总晚点最短的列车运行调整策略。车站发车次序总数等于列车总数L的阶乘，但并非所有$L!$种发车次序结果都是可行的，原因有二：其一，通过作业的2列列车在车站不可能改变列车运行顺序；其二，某些发车次序并不满足车站作业间隔时间的约束。以图3为例说明这种不可行的发车次序。由图3可知：对于接连经过车站$s+1$的停站列车和通过不停站列车，因存在车站作业间隔时间的约束关系，后行通过的列车$l+1$无法越行当前停站时间只有2 min的停站列车l，因此车站$s+1$可行的发车次序有且只有{列车l，列车$l+1$}。故设计启发式规则对各车站的发车次序集合进行“剪枝”，剔除其中不可行的发车次序，以便最终输入到强化学习环境动作集中的沿线车站所有发车次序均是可行的。

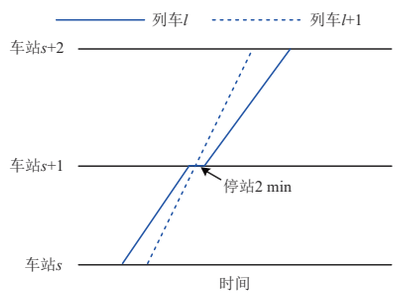


图3 不可行发车次序示意图

2.3.2 可行发车次序树结构

通过启发式规则，输出各车站可行的发车次序。以相邻3列列车l，$l+1$和$l+2$为例，设计得到蒙特卡罗树搜索算法下博弈树的数据结构如图4所示。由图4可知：始发站（车站1）的发车次序为树结构的根节点，车站2的3种发车次序为连接始发站根节点的3个子节点，以此类推，最终可遍历至终点站S发车次序的子节点。

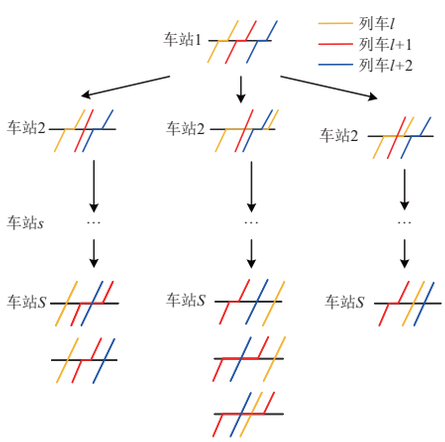


图4 发车次序树结构示意图

2.3.3 蒙特卡罗树搜索的最优发车次序算法

结合上述发车次序的博弈树结构，提出MCTS的列车最优发车次序算法，步骤如下。

步骤1：输入始发站（车站序号$s=1$）根节点状态$\boldsymbol{\mathcal{S}}_1$。

步骤2：判断其后的车站子节点（发车次序）

是否被访问过，若已被访问，转步骤3；否则转步骤4。

步骤3：利用上限置信区间（UCT）算法求出各子节点函数值（算法和求解方法可参考文献[22]），选取函数值最大的子节点（发车次序）作为当前节点动作并转步骤4；若函数值相等，则随机选择1个子节点，转步骤5。

步骤4：随机选择1个未被访问的子节点，转步骤5。

步骤5：判定当前节点是否为终点站子节点，若是，转步骤6；否则，转移至下一车站，转步骤2。

步骤6：扩展生成终点站子节点的动作（可行发车次序），随机选择1个动作并在树结构中加入该动作的新状态，转步骤7。

步骤7：从根节点到当前节点，完成1次完整回合的模拟训练，转步骤8。

步骤8：将模拟训练的胜负结果回溯至树中，更新UCT算法参数，若当前回合数未达到所设定的最大值，转步骤1；若已达到设定的最大值，终止模拟，输出列车运行调整的最优发车次序。

2.4　冲突消解启发式规则

在MCTS给出最优发车次序后，晚点列车和运行线发生偏移可能与后行列车在区间或者车站产生冲突，在列车运行图中表现为冲突列车的运行线在区间产生交点，或冲突列车在车站不满足车站作业间隔时间的最小值约束，严重影响行车安全。因此在蒙特卡罗树搜索生成列车在车站的发车次序后，基于消解列车运行冲突的传统方法[25]设计启发式规则，将其运用于列车在车站和区间运行冲突的消解，步骤如下。

步骤1：在强化学习环境中，输入晚点场景下的实际接发车时刻矩阵$\boldsymbol{S}_s$。

步骤2：检测当前相邻2列列车l和$l+1$在车站s的实际发车间隔时间（即$I^{\mathrm{dd}}_{l,l+1,s}$，$I^{\mathrm{dp}}_{l,l+1,s}$，$I^{\mathrm{pd}}_{l,l+1,s}$，$I^{\mathrm{pp}}_{l,l+1,s}$）是否满足最小车站作业间隔时间$I^{\min}_s$的约束，若满足，转步骤3；否则，转步骤4。

步骤3：转移至下一组的相邻2列列车，继续检测发车间隔时间是否满足$I^{\min}_s$的约束，若不满足，转步骤2；否则将继续检测该站所有其他列车，直到所有列车完成检测后，转步骤5。

步骤4：令实际发车间隔时间（即$I^{\mathrm{dd}}_{l,l+1,s}$，$I^{\mathrm{dp}}_{l,l+1,s}$，$I^{\mathrm{pd}}_{l,l+1,s}$和$I^{\mathrm{pp}}_{l,l+1,s}$）满足$I^{\min}_s$的约束，并调整当前相邻2列列车的发车时刻，转步骤3。

步骤5：检测列车l与后行受晚点影响列车在区间（s,$s+1$）是否存在冲突，若存在冲突，则运用启发式规则消解冲突；否则，转步骤6。

步骤6：检测当前相邻2列列车l和$l+1$在车站s的到站间隔时间（即$I^{\mathrm{aa}}_{l,l+1,s}$，$I^{\mathrm{ap}}_{l,l+1,s}$和$I^{\mathrm{pa}}_{l,l+1,s}$）是否满足最小车站作业间隔时间$I^{\min}_s$的约束，若满足，转步骤7；否则，转步骤8。

步骤7：转移至下一组的相邻2列列车继续检测到站间隔时间是否满足$I^{\min}_s$的约束，若不满足，转步骤6；否则将继续检测该站所有其他列车，直到所有列车完成检测后，转步骤9。

步骤8：令实际到站间隔时间（即$I^{\mathrm{aa}}_{l,l+1,s}$，$I^{\mathrm{ap}}_{l,l+1,s}$和$I^{\mathrm{pa}}_{l,l+1,s}$）满足$I^{\min}_s$的约束，并调整当前相邻2列列车的到站时刻，转步骤6。

步骤9：基于MCTS-RL方法调整所有列车在车站s的发车次序，并选择其中1种，当$s=S$时，转步骤10；否则，转步骤2。

步骤10：计算列车总晚点时间下的奖励函数值，输出列车运行调整策略。

3　算例仿真

以京沪高铁北京南—泰安段的某日计划运行图作为初始数据输入，设置大量晚点场景并选择其中2个作为典型场景，基于前述数学模型和列车运行调整过程，构建得到强化学习环境与智能体，并令其不断交互学习；基于MCTS-RL一次性生成离线训练模型得到列车运行智能调整方法（简称MCTS-RL法）。在列车运行调整过程中，当输入列车接车或者发车晚点时，该离线训练模型可直接用于列车运行调整问题的实时求解，无须重新建模求解。将MCTS-RL方法下的方案与同样应用本文数学模型、但求解时分别采用先到先服务（First-Come-First-Served，FCFS）法[6]和CPLEX求解器得到的调整方案进行对比，验证本文提出方法的有效性和最优性。

3.1　参数设置和晚点场景

以京沪高铁北京南—泰安段沿线的北京南、廊坊、天津南、沧州西、德州东、济南西和泰安7个车站为背景，某日线上共开行列车79列。列车在6个站间的最小区间运行时间分别为15，14，14，21，17和15 min；最小停站时间$t^{\min}_{l,s}$和最小车站作业间隔时间$I^{\min}_s$均设置为2 min。

针对该日计划运行图中的全部79列列车，随机设置10～30 min的大量发车晚点和到站晚点场

景，并从中选择2个较具代表性的场景见表1。

表1 典型晚点场景

场景序号	影响列车	始发站发车时刻	晚点情况
1	第20列	9:20	在北京南发车晚点12 min
	第23列	9:50	到达廊坊站晚点18 min
2	第45列	13:30	到达天津南晚点12 min
	第46列	13:40	在北京南站发车晚点8 min
	第49列	14:00	在北京南站发车晚点15 min

3.2 计算结果

针对设置的大量晚点场景，基于Python 3.6.5编写强化学习环境，在Intel Core i7-4710MQ@2.5 GHz，12 GB的电脑上一次性离线训练，生成最终的MCTS-RL在线调整模型。强化学习训练时，列车运行图采用深度卷积神经网络进行状态集输入特征的学习，深度学习框架TensorFlow版本为tensorflow-gpu 1.8.0。

经过多次强化学习训练交叉验证后，确定其训练参数见表2。表中：探索开发比表示训练阶段随机搜索策略占所有策略的比值；折算因子表示某个训练片段中随着车站状态集不断向前推移，奖励函数值所呈现的指数衰减趋势（即距离当前状态越远的车站状态集，对智能体影响越小）。

表2 强化学习的训练参数

参数	步长	折算因子	训练片段个数	初始探索开发比
数值	0.001	0.15	1 500	0.75

以表1中的2个典型场景为例，分别采用FCFS法、CPLEX求解器方法（简称CPLEX法）以及MCTS-RL法求解列车运行调整方案。FCFS法用于验证MCTS-RL法在减小列车总晚点时间上的有效性。考虑到CPLEX法的求解结果一定最优，故以CPLEX下的调整方案（即最优方案）验证MCTS-RL法下调整方案的最优性。

为表达FCFS法（或MCTS-RL法）下调整方案与CPLEX下最优方案之间的总晚点差值（Gap），引入η

$$\eta=\frac{\tau-\tau_{\text{opt}}}{\tau} \tag{13}$$

式中：τ为FCFS法/MCTS-RL法下调整方案的总晚点时间，min；τ_{opt}为CPLEX最优方案的总晚点时间，min。

1）3种方法下求解指标对比

FCFS、CPLEX和MCTS-RL这3种方法下，求解得到调整方案的总晚点时间及求解时间对比见表3。由表3可得到如下结论。

表3 FCFS，CPLEX和MCTS-RL的求解指标对比

场景序号	求解方法	总晚点时间/min	η/%	求解时间/s
1	FCFS	254	5.51	0.005
	CPLEX	240		24.044
	MCTS-RL	240	0	<0.001
2	FCFS	214	22.43	0.013
	CPLEX	166		24.605
	MCTS-RL	166	0	<0.001

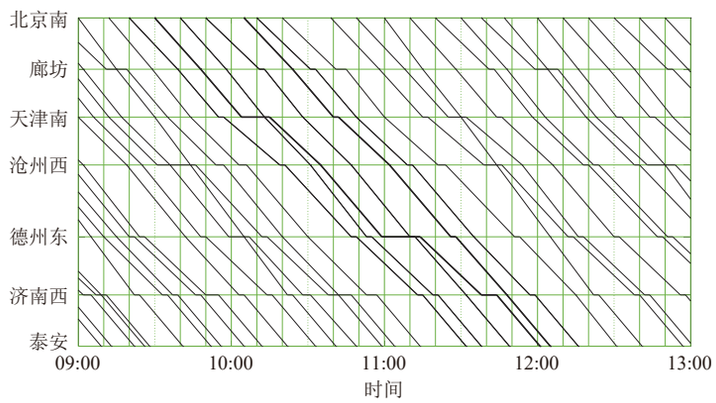
（1）在列车总晚点时间方面，CPLEX和MCTS-RL方法下的更短，分别比FCFS法缩短14 min和48 min；这意味着在2个典型晚点场景下，CPLEX和MCTS-RL方法下最优方案能够分别缩短5.51%和22.43%的晚点时间。

（2）在列车运行调整求解实时性方面，FCFS法能分别在0.005 s和0.013 s内给出与图定发车次序相同的调整策略，具有较好的实时性；CPLEX求解器虽然得到总晚点时间最短的最优调整策略，但求解时间分别达24.044 s和24.605 s，考虑到案例涉及参数、变量较少，若将其运用于真实场景下，求解时间可能会随着车站、列车数量的增加而呈现指数级增长；MCTS-RL虽消耗大量时间用于试错学习的离线训练，但训练结束后可产生总晚点时间最短的列车运行调整学习策略，智能体凭借该策略能够在短于0.001 s时间内给出同样最优的列车运行调整策略。相较于CPLEX法，MCTS-RL法的求解效率高很多。

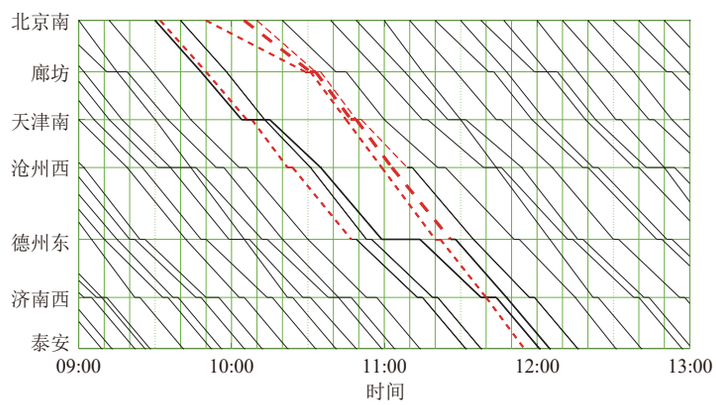
2）列车运行图调整结果

针对2个典型晚点场景，FCFS、CPLEX和MCTS-RL这3种方法下的运行图调整结果对比，分别如图5和图6所示。图中：实线和虚线分别表示该方法下不需要调整、应进行调整的列车运行线；线型粗细用于区分运行线归属于不同列车。由图5和图6可得到如下结论。

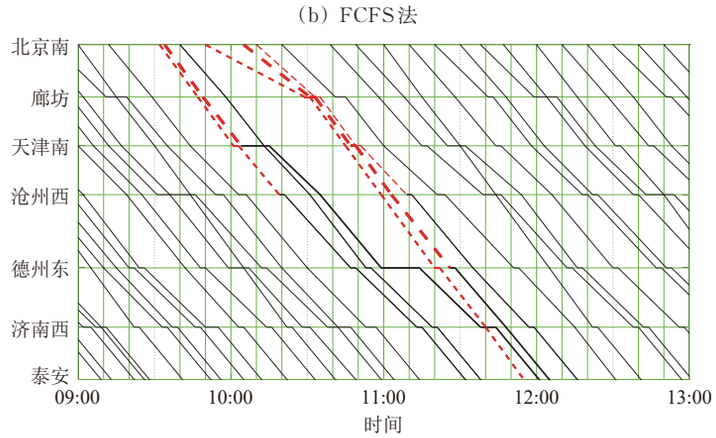
（1）CPLEX求解器和MCTS-RL方法下各列车在各车站的发车次序相同，这说明2种方法下运行图调整结果是相同的，进一步说明本文所提出MCTS-RL方法能给出同样最优的调整策略；相比于CPLEX，MCTS-RL的优势在于无须每次重新求解新问题，而是可直接根据离线训练模型下的学习策略，在线实时生成列车运行调整方案。



（a）计划运行图

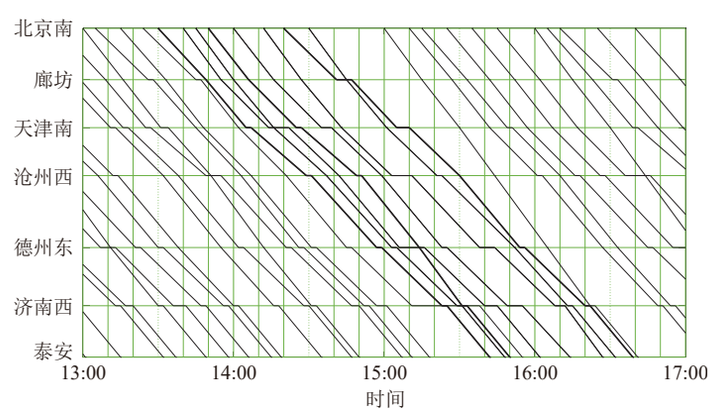


（b）FCFS法

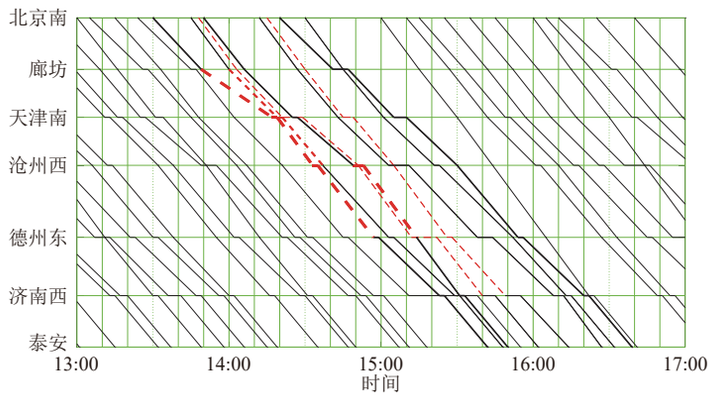


（c）CPLEX法/MCTS-RL法

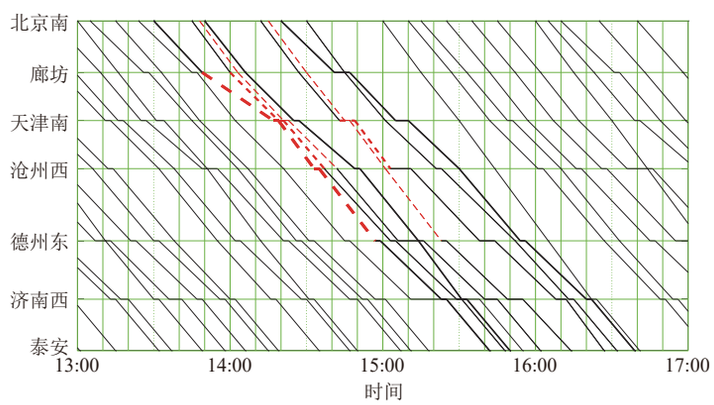
图5　典型晚点场景1下计划运行图和FCFS法、CPLEX法/MCTS-RL法得到的运行图调整结果



（a）计划运行图



（b）FCFS法



（c）CPLEX法/MCTS-RL法

图6 典型晚点场景2下计划运行图和FCFS、CPLEX法/MCTS-RL法得到的运行图调整结果

（2）与FCFS法相比，CPLEX法和MCTS-RL法均能够通过调整列车在车站的接发车时刻，生成总晚点最短的列车运行调整策略。例如图5中，最优方案调整了第20列和21列列车（图定9:30始发）在北京南的发车次序和时刻，这样第20列列车能够在沧州西站更早地恢复正点，但各列车在其余车站的发车次序与图定相同；图6中，最优方案调整了第47列列车（晚点后13:45始发）与第48列列车（13:50始发）在天津南站的发车次序和发车时刻，增加了第50列列车（14:12始发）在天津南站的停站时间，令第49列列车（晚点后14:15始发）在该站越行，使列车总晚点时间最短。

4 结 语

针对路网中列车的到站和发车晚点，根据高速铁路列车运行调整数学模型，提出MCTS-RL的列车运行智能调整方法，设计由状态集、动作集、状态转移概率和奖励函数组成的强化学习环境。MCTS可给出总晚点时间最短下各列车在各车站的发车次序，然后设计启发式规则消解列车运行冲突。MCTS-RL通过离线训练—在线调整的学习机制，实时辅助列车调度员调整列车运行图，提升晚点场景下应急处置效率。仿真结果表明，典型晚点场景下，MCTS-RL方法下的在线调整模型能够在0.001 s内给出与CPLEX求解器同样最优的列车运行调整策略；与FCFS方案相比，MCTS-RL下最优调整策略的总晚点时间又分别缩短14 min和48 min。

与既有研究不同的是，本文研究基于列车调度员的宏观调图视角，后续工作可考虑车站进路、线路信号设备布置和列车运行状态等实际微观约束，同时还可进一步研究严重晚点场景下动车组运用计划和列车运行图的协同调整。

参 考 文 献

[1] MASCIS A, PACCIARELLI D. Job-Shop Scheduling with Blocking and No-Wait Constraints [J]. European Journal of

Operational Research, 2002, 143 (3): 498-517.
[2] MIN Y H, PARK M J, HONG S P, et al. An Appraisal of a Column-Generation-Based Algorithm for Centralized Train-Conflict Resolution on a Metropolitan Railway Network [J]. Transportation Research Part B: Methodological, 2011, 45 (2): 409-429.
[3] 章优仕,金炜东.单线列车运行调整目标体系与算法模型的研究[J].东南大学学报:自然科学版,2009,39(增1):248-254.
(ZHANG Youshi, JIN Weidong. Study on Target System of the Railway Rescheduling Based on the Theory of Satisfactory Optimization [J]. Journal of Southeast University: Natural Science Edition, 2009, 39 (Supplement 1): 248-254. in Chinese)
[4] ZHANG H M, LI S K, WANG Y H, et al. Real-Time Optimization Strategy for Single-Track High-Speed Train Rescheduling with Disturbance Uncertainties: a Scenario-Based Chance-Constrained Model Predictive Control Approach [J]. Computers & Operations Research, 2021, 127: 105135.
[5] ZHU Y Q, GOVERDE R M P. Dynamic and Robust Timetable Rescheduling for Uncertain Railway Disruptions [J]. Journal of Rail Transport Planning & Management, 2020, 15: 100196.
[6] XU P J, CORMAN F, PENG Q Y, et al. A Train Rescheduling Model Integrating Speed Management during Disruptions of High-Speed Traffic under a Quasi-Moving Block System [J]. Transportation Research Part B: Methodological, 2017, 104: 638-666.
[7] 王涛,张琦,赵宏涛,等.基于替代图的列车运行调整计划编制及优化方法[J].中国铁道科学,2013,34(5):126-133.
(WANG Tao, ZHANG Qi, ZHAO Hongtao, et al. A Method for Generation and Optimization of Train Operation Adjustment Plan Based on Alternative Graph [J]. China Railway Science, 2013, 34 (5): 126-133. in Chinese)
[8] 李智,端嘉盈,曾壹,等.基于智能化应用的列车运行调整模型[J].中国铁道科学,2021,42(2):173-182.
(LI Zhi, DUAN Jiaying, ZENG Yi, et al. Train Operation Adjustment Model Based on Intelligent Application [J]. China Railway Science, 2021, 42 (2): 173-182. in Chinese)
[9] EATON J, YANG S X, GONGORA M. Ant Colony Optimization for Simulated Dynamic Multi-Objective Railway Junction Rescheduling [J]. IEEE Transactions on Intelligent Transportation Systems, 2017, 18 (11): 2980-2992.
[10] 林博,俞胜平,刘子源,等.基于改进粒子群算法的高铁列车动态调度[J].控制工程,2021,28(7):1334-1341.
(LIN Bo, YU Shengping, LIU Ziyuan, et al. High-Speed Train Dynamic Scheduling Method Based on Improved Particle Swarm Optimization Algorithm [J]. Control Engineering of China, 2021, 28 (7): 1334-1341. in Chinese)
[11] 马驷,孙建康,鲁工圆.高速铁路车站列车进路分配方案的优化与调整[J].中国铁道科学,2018,39(1):122-130.
(MA Si, SUN Jiankang, LU Gongyuan. Optimization and Adjustment of Train Route Allocation Scheme for High Speed Railway Station [J]. China Railway Science, 2018, 39 (1): 122-130. in Chinese)
[12] CACCHIANI V, HUISMAN D, KIDD M, et al. An Overview of Recovery Models and Algorithms for Real-Time Railway Rescheduling [J]. Transportation Research Part B: Methodological, 2014, 63: 15-37.
[13] 季学胜,孟令云.列车到发时刻与进路同步优化的高速铁路列车运行调整模型[J].中国铁道科学,2014,35(4):117-123.
(JI Xuesheng, MENG Lingyun. Train Operation Adjustment Model for Synchronously Optimizing Train Arrival/Departure Time and Route on High Speed Railway Network [J]. China Railway Science, 2014, 35 (4): 117-123. in Chinese)
[14] SUTTON R S, BARTO A G. Reinforcement Learning: an Introduction [M]. London: Massachusetts Institute of Technology Press, 2018.
[15] SHEN X N, MINKU L L, MARTURI N, et al. A Q-Learning-Based Memetic Algorithm for Multi-Objective Dynamic Software Project Scheduling [J]. Information Sciences, 2018, 428: 1-29.
[16] KARA A, DOGAN I. Reinforcement Learning Approaches for Specifying Ordering Policies of Perishable Inventory Systems [J]. Expert Systems with Applications, 2018, 91: 150-158.
[17] SHAHRABI J, ADIBI M A, MAHOOTCHI M. A Reinforcement Learning Approach to Parameter Estimation in Dynamic Job Shop Scheduling [J]. Computers & Industrial Engineering, 2017, 110: 75-82.
[18] ŠEMROV D, MARSETIČ R, ŽURA M, et al. Reinforcement Learning Approach for Train Rescheduling on a Single-Track Railway [J]. Transportation Research Part B: Methodological, 2016, 86: 250-267.
[19] KHADILKAR H. A Scalable Reinforcement Learning Algorithm for Scheduling Railway Lines [J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20 (2): 727-736.
[20] ZHU Y Q, WANG H R, GOVERDE R M P. Reinforcement Learning in Railway Timetable Rescheduling [C]// 2020 IEEE 23rd International Conference on Intelligent Transportation Systems. Rhodes, Greece. New York: IEEE Press, 2020: 1-6.
[21] NING L B, LI Y D, ZHOU M, et al. A Deep Reinforcement Learning Approach to High-Speed Train Timetable

Rescheduling under Disturbances [C]// 2019 IEEE Intelligent Transportation Systems Conference. Auckland, New Zealand. New York: IEEE Press, 2019: 3469-3474.
[22] BROWNE C B, POWLEY E, WHITEHOUSE D, et al. A Survey of Monte Carlo Tree Search Methods [J]. IEEE Transactions on Computational Intelligence and AI in Games, 2012, 4 (1): 1-43.
[23] SILVER D, HUANG A, MADDISON C J, et al. Mastering the Game of Go with Deep Neural Networks and Tree Search [J]. Nature, 2016, 529 (7587): 484-489.
[24] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the Game of Go without Human Knowledge [J]. Nature, 2017, 550 (7676): 354-359.
[25] HANSEN I, PACHL J. Railway Timetabling and Operations: Analysis, Modelling, Optimisation, Simulation, Performance, Evaluation [M]. Hamburg: DVV Media Eurail Press, 2014.

Intelligent Adjustment Approach for Train Operation Based on Monte Carlo Tree Search-Reinforcement Learning

WANG Rongsheng[1,2,3], ZHANG Qi[2,3], ZHANG Tao[2,3], WANG Tao[2,3], DING Shuxin[2,3]

(1. Postgraduate Department, China Academy of Railway Sciences, Beijing 100081, China;
2. Signal and Communication Research Institute, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China;
3. The Center of National Railway Intelligent Transportation System Engineering and Technology, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China)

Abstract: In order to improve the emergency response efficiency of high-speed railways in emergencies, taking train timetable as the research object, an intelligent adjustment approach for train operation based on Monte Carlo Tree Search-Reinforcement Learning (MCTS-RL) was proposed under the scenarios of delay, including the off-line training model of train operation intelligent adjustment, reinforcement learning method, the departure sequence decision method based on MCTS and heuristic rules for conflict resolution. This paper established a reinforcement learning environment based on the mathematical model for train operation adjustment of high-speed railway, involving state set, action set, state transition probability and reward function. Firstly, heuristic rules were designed to generate the feasible departure sequence used as the game tree node of MCTS. The optimal departure sequence of train operation adjustment was output by MCTS. Then, heuristic rules were designed to resolve the train operation conflicts in stations and block sections. With the objective function of minimizing the total delay time of trains along the line, an online adjustment model was generated by MCTS-RL during the one-time offline training and used to adjust the train departure sequence at each station in real-time. Taking the Beijingnan-Tai′an section of Beijing-Shanghai high-speed railway as an example, the scenarios of the arrival and departure delay were set and solved by first-come-first-served, the CPLEX solver, and the MCTS-RL approach respectively. The results indicated that, compared with the solution obtained by the CPLEX solver, the MCTS-RL approach could provide the same optimal adjustment solution of train operation within 0.001 s.

Key words: High-speed railway; Train operation adjustment; Artificial intelligence; Reinforcement learning; Monte Carlo Tree Search

（责任编辑 耿枢馨）