文章编号：1672-4747（2020）04-0053-08

基于强化学习的多时隙铁路空车实时调配研究

谭　雪[1]，张小强[1,2]，石红国[1,2]，成嘉琪[3]

（1. 西南交通大学，交通运输与物流学院，成都 611756；
2. 综合交通运输智能化国家地方联合工程实验室，成都 611756；
3. 上海市政工程设计研究总院（集团）有限公司，上海 200000）

摘　要：铁路空车调配计划是进行运输组织的基础和重要条件，空车供求关系的时空变化特性和运输生产的动态性，使求解多时隙空车实时调配最优策略变得困难。强化学习中的 Q-learning 时序差分算法能较好地解决不完全信息下的大规模序列决策问题，故本文将决策周期划分为若干个时隙，提出多时隙空车实时调配模型：首先利用空车实际调配的局部马尔科夫特性改进 Q-learning 算法，进行“单一空车调配策略评估”以量化单一空车在决策周期内所有时空状态下采取不同行动的长期回报；然后提出空车实时优先调配算法，求解决策周期全局最优的调配策略。算例表明模型可以兼顾实时调配长期回报最大、空走距离小、即时需求响应程度高，求解出每时隙下最优且决策周期全局最优的实时调配策略，以使运输部门快速适应变化的货运市场需求、提供科学合理的空车实时调配策略是可行的。

关键词：铁路运输；空车实时调配；强化学习；空车；多时隙

中图分类号：U292.8　　**文献标志码：**A　　**DOI：**10.3969/j.issn.1672-4747.2020.04.007

Reinforcement-learning-based Multi-slot Rail Empty Wagon Real-time Distribution

TAN Xue[1], ZHANG Xiao-qiang[1,2], SHI Hong-guo[1,2], CHENG Jia-qi[3]

(1. School of Transportation and Logistics, Southwest Jiaotong University, Chengdu 611756, China;
2. National United Engineering Laboratory of Integrated and Intelligent Transportation, Chengdu 611756, China;
3. Shanghai Municipal Engineering Design Institute Co., Ltd., Shanghai 200000, China)

Abstract: Rail empty wagon distribution is critical to a transportation enterprise. The spatio-temporal characteristics of the supply and demand of empty wagons and the dynamics of transportation generate difficulties in developing an optimal strategy for multi-slot empty wagon real-time distribution. A Q-reinforcement-learning algorithm can solve large-scale sequence decision problems using incomplete information. In this study, the decision period is divided into multi-slots, and a multi-slot empty wagon distribution model is proposed. First, based on local Markov characteristics of empty wagon distribution, an improved Q-learning algorithm is designed, and a single empty wagon strategy evaluation is performed to

收稿日期：2020-06-07
基金项目：国家铁路局科技开发项目（KF2019-101-B）
作者简介：谭　雪（1997—），女，汉族，安徽亳州人，硕士，研究方向：机器学习、数据挖掘，E-mail：779495316@qq.com
通信作者：张小强（1975—），男，汉族，江西石城人，副教授，博士后，研究方向：铁路运营管理，人工智能与智慧物流，E-mail：xqzhang@swjtu.edu.cn
引文格式：谭雪，张小强，石红国，等. 基于强化学习的多时隙铁路空车实时调配研究[J]. 交通运输工程与信息学报，2020, 18(4): 53-60

evaluate a single wagon's long-term gains under all spatio-temporal states during the decision period. Second, an empty wagon real-time priority distribution algorithm is proposed to solve the strategy for each slot. A case study of multi-slot empty wagon real-time distribution shows that our proposed model can maximize long-term gains as well as minimize unloaded distances of a real-time distribution. Thus, providing rail transportation enterprises with scientific real-time empty wagon distribution strategies is feasible.

Key words: railway transportation; empty wagon real-time distribution; reinforcement learning; empty wagon; multi-slot

0 引 言

空车调配计划是铁路技术计划的重要组成部分，合理确定空车调配数量和调配方向，减少空车走行公里对铁路降本增效至关重要。铁路空车调配受运输生产动态性、路网结构复杂性和空车供需不确定性等复杂因素的影响，属于不完全信息下的时变决策问题，因此优化决策周期内的空车实时调配策略较为困难。

空车调配算法分为静态调配模型和动态调配模型，模型目标一般是决定调配起讫点、空车数量和输送路径。静态调配模型是依据已知的空车供需确定性信息优化当前调配策略[1-4]，直观性强且容易实施，但不适合处理实际中空车供求状况随时空动态变化的实时调配过程。动态调配以基于时空网络的实时调配模型为主，指在一个决策周期内，依据当前和未来时隙的空车供求信息来优化调配策略。比如文献[5]同时考虑了决策周期内的固定需求及各时隙新产生的空车需求，分两阶段求解实时调配策略；文献[6]从动态优化的角度构建多时点调配模型。上述两种实时调配模型降低了空车调配时变系统研究复杂性，可为决策周期内每一时隙调整调配策略提供依据。但是由于铁路空车供求关系的时空不匹配性和不确定性，按上述方法求解出的实时调配策略从调配决策周期全局看不一定是最优解。

综上所述，对铁路空车调配决策周期内建立全局最优的实时调配模型研究很少。Q-learning是强化学习[7-11]中应用最为广泛的一种时序差分算法：智能体通过状态观测值、行动和即时回报序列与环境持续交互学习，构建对环境的认知，完成策略评估—策略改进—迭代收敛，进而求解马尔科夫决策过程（Markov Decision Process, MDP）的最优决策序列。空车实时调配本质属于不完全信息下的 MDP 问题，所以 Q-learning 算法可以量化单一空车在决策周期内所有时空状态下的调配动作价值函数，并用之优化实时调配策略。因此，本文将铁路空车实时调配转化为多时隙大规模序列决策问题，应用强化学习构建多时隙空车实时调配模型，求解时空动态变化和不完全空车供需信息下，兼顾决策周期全局最优和各时隙最优的多时隙铁路空车实时调配策略，最后通过仿真算例验证模型的有效性。

1 多时隙空车实时调配模型

针对铁路空车需求时空变化特征和实际调配过程的马尔科夫特性，将决策周期拆解为多时隙，提出多时隙空车实时调配模型：（1）以实际空车调配的局部马尔科夫特性，改进 Q-learning 算法，进行“单一空车调配策略评估”以量化单一空车在决策周期内所有时空状态下采取不同行动（站内停留或站间调配）的长期回报；（2）在每个时隙下的实时调配阶段，将所有空车视为多智能体系统，在综合考虑货主即时需求响应程度高、空车走行距离小、铁路运输企业长期回报最大的基础上，使用优先调配算法求解该时隙下最优且决策周期同样最优的站间空车调配数量和

调配方向。

1.1　基于局部 MDP 的单一空车调配模型

马尔科夫决策过程常用于完全信息下序列决策问题建模，其基本思想是，智能体在离散时间序列中与环境持续交互使得智能体不断更新。即智能体在时刻 t 所处状态 s_t，按照已知的转移概率 $p_\pi(a_t|s_t)$ 选择 a_t，更新至 s_{t+1} 并获得即时奖励 r_{t+1}，持续交互构成状态、动作、奖励序列。

当智能体不能提前获知状态转移概率时，该过程是不完全信息下的 MDP（又称局部 MDP）。显然，单一空车调配为局部 MDP 模型，针对空车需求时空变化特征和实际调配过程，合理构建该局部 MDP 是基于 Q-learning 的单一空车调配策略评估和求解实时调配策略的基础。

（1）状态　这里将单一空车视为智能体，决策周期由若干个离散时隙 $[0,\mathrm{L}\ ,t,\mathrm{L}\ ,T-1]$ 构成。*stat* 包括空车剩余/卸空站 *stat0* 和空车需求站 *stat1*。智能体的状态 s 由时空维度共同决定：$s_t=(i,t)\in S$，其中 $i\in\{stat0\}$ 和 t 分别代表智能体所处空车剩余站点索引和决策周期内的时间索引，可见单一空车调配的 MDP 模型共有 $|S|=|stat|\times|T|$ 个状态。

（2）动作及状态转移　智能体在每一个调配初始状态 $s_t=(i,t)$ 下，有两种动作可供选择：

① 在当前状态下，智能体不参与站间调配，而继续在站点等待，此时自动转到下一个时隙状态 $s_{t+1}=(i,t+1)$。

② 假设单一空车完整调配策略为：空走—货物作业—重走—卸空，则空车在当前状态下按调配策略先由空车剩余站 i 空走至空车需求站 $j\in stat1$，货物作业后重走至卸车站 $k\in stat0$。货物作业时间、走行时间、技术作业时间等之和 Δt 根据历史车流数据查定。此时空车执行此完整调配动作，状态便由 $s_t=(i,t)$ 转移至 $s_{t+\Delta t}=(k,t+\Delta t)$。

（3）奖励　r 与 s_t 和 a_t 密切相关，在单一空车调配问题中，每个状态下的动作奖励应分为两种情况计算：

① 假设不考虑受车站作业能力约束或线路能力约束等外在因素导致空车无法调配以及单一空车要满足下一时刻本站装车产生等待的情况，在此 MDP 模型构建中，统一规定空车采取站内等待动作(*w*)时，即时奖励设为 0，状态转移方程用 $(s_t,w,r_{t+1},s_{t+1},\mathrm{L}\ )$ 表示，其中 $r_{t+1}=0$。

② 当空车执行一次完整调配时，奖励计算方法如式（1）-（3）所示：

$$\Delta t=\Delta t_1+\Delta t_2 \tag{1}$$

$$r_{\Delta t}=0+r_{\Delta t_2} \tag{2}$$

$$r_{\Delta t}^{\gamma}=\sum_{t=0}^{\Delta t-1}\gamma^t\frac{r_{\Delta t}}{\Delta t} \tag{3}$$

式（1）中：Δt 指单一空车执行一次完整调配总时间；Δt_1 指空车从 i 站空走至 j 站，完成货物作业、技术作业后，再从 j 站发出经历的总时间；Δt_2 指空车由 j 站重走至 k 站及在 k 站卸空经历的总时间。式（2）中：$r_{\Delta t}$ 指不考虑折扣因子下完成此次调配动作获得的奖励，其中空车从 i 站走行至 j 站不产生货运效益，$r_{\Delta t_2}$ 指空车装货后重走产生的货运收入。式（3）中，$r_{\Delta t}^{\gamma}$ 指折扣动作奖励，由于在实际应用中，长周期序列决策通常在计算价值函数时会产生较大方差，因此在策略迭代时常对原始动作奖励做此折扣处理[10]。

以下提供单一空车调配局部 MDP 模型构建的算例。

假设 $t=0$ 时刻空车在站点 A $\left(s_t=\left(A,0\right)\right)$ 采取调配方案是先空走至 B 站再重走至 C 站，预计完成一次调配的 $\Delta t=\Delta t_1+\Delta t_2=1+3=4$ 天，空车的时空状态转移至 $s_{t+\Delta t}=\left(C,4\right)$ 并获得即时奖励 $r_{\Delta t}=0+r_{\Delta t_2}=2500$ 元，设折扣因子 $\gamma=0.9$，则执行此完整调配策略的折扣动作奖励为：$r_{\Delta t}^{\gamma}=\frac{2\,500}{4}+\frac{2\,500}{4}\times$

$$0.9+\frac{2\,500}{4}\times 0.9^2+\frac{2\,500}{4}\times 0.9^3=2\,149.375\text{ 元}$$

1.2 基于 Q-learning 的单一空车调配策略评估

在 Q-learning 中，将求解动作价值函数的过程视为策略评估。针对单一空车，希望找到每一个时空状态下价值最大化的调配行动，即 $a^*=\max_{a\in A}Q_\pi(s,a),\forall s$。注意，在单一空车调配策略评估阶段不考虑决策周期内智能体需满足站点的实际空车需求导致 Q-learning 迭代提前终止的情况，而是假定智能体可以。在一个周期内持续探索、迭代获得各时隙下所有站点的状态值函数，并将之用于空车实时优先调配才有意义。

为简化单一空车探索动作集合，将其完整调配动作拆分为 Δt_1 空走、Δt_2 重走两阶段交替序列分别计算动作价值函数。受调配决策周期总时长 T 的限制，在原始 Q-learning 模型上对决策序列结束状态加以改进（见表 1 步骤 11~23），延伸出局部 MDP 下的铁路单一空车调配策略评估伪代码。

表 1 局部 MDP 下单一空车调配 Q-learning 策略评估伪代码

Fig.1 Pseudocode for pail empty wagon distribution evaluation in local MDP

步骤	伪代码
Input	state transition equation (s,a,r,s')
Output	action-value function $Q(s,a)$
1	Initialize $Q(s,a)$ and $Q(terminal,\cdot)=0$, $\forall s\in S,a\in \mathrm{A}(s)$
2	def judge_action (s,a):
	If $a=s[0]$: $r_{\Delta t}^{\gamma}=0$ # wait in the same station
	Else $a\neq s[0]$:
	if unloaded: $r_{\Delta t}^{\gamma}=0$ # unloaded transport
	Elif loaded: $r_{\Delta t}^{\gamma}$ # loaded transport
	return $r_{\Delta t}^{\gamma}$
3	for $total\_time=T$ to 0 do:
4	Initialize s
5	Repeat for each step of episode:
6	Choose a from s using $\varepsilon-greedy$ policy derived from Q
7	Take action a, judge_action (s,a), observe $r_{\Delta t}^{\gamma},s',\Delta t$
8	$Q(s,a)\leftarrow Q(s,a)+\alpha[r_{\Delta t}^{\gamma}+\gamma^{\Delta t}\max_{a'\in \mathrm{A}}Q(s',a')-Q(s,a)]$
9	time += Δt
10	$s\leftarrow s'$
11	If $time==total\_time$ then:
12	s is terminal
13	return $Q(s,a)$
14	break
15	Elif $time>total\_time$ then:
16	$last\_time=total\_time-time+\Delta t$
17	for each $a\in A$:
18	If $\Delta t(s',a,s'')<=last\_time$:
19	$Q_{last}\leftarrow Q_{last}(s',a)$
20	Choose a_{last} from s' using $\varepsilon-greedy$ policy derived from Q_{last}
21	Take action a_{last}, judge_action (s',a_{last}), observe $r_{\Delta t}^{\gamma},s'',\Delta t$
22	$Q(s',a_{last})\leftarrow Q(s',a_{last})+\alpha[r_{\Delta t}^{\gamma}+\gamma^{\Delta t}\max_{a''\in \mathrm{A}}Q(s'',a'')-Q(s',a_{last})]$
23	s'' is terminal
24	return $Q(s',a_{last})$
25	break

以上 $Q(s,a)$ 表示单一空车在 s 执行动作 a（等待或空/重走行）后可获得的长期回报，包含了两部分：动作的即时收益 $r_{\Delta t}^{\gamma}$ 和该空车执行动作后转移至新状态 s' 下的折扣状态价值，即 $\gamma^{\Delta t}\cdot V_{\pi^*}(s')=\gamma^{\Delta t}\cdot\max_{a'\in A}Q(s',a')$。

把每个时隙所有的 $Q(s,a)$ 记录在 Q 表中，可获得决策周期内所有时隙下单一空车初始状态的价值函数及所有的动作价值函数。

1.3　空车实时优先调配算法

从强化学习的角度分析，每一辆空车是相互独立的，每一时隙也是相互独立的，分而治之，将决策周期内每一个时隙的所有空车（下称空车）调配拆解为单一空车的实时调配合集，调配系统的目标函数是最大化多时隙初始状态下所有单一空车调配动作价值：

$$\max\ \sum_{t=0}^{T-1}\sum_{i=0}^{m}\sum_{j=0}^{n}Q_t(s_0,a_{ij})x_{ij} \tag{4}$$

为降低求解复杂度，确保空车调配系统全局最优，对传统运输问题的目标函数加以改进。建立空车实时优先调配算法，为防止对流，假定在每个时隙满足本站空车需求基础上，再确定剩余空车站间优先调配量和调配方向，具体模型如下：

$$\max\ \sum_{i=0}^{m}\sum_{j=0}^{n}A_\pi(i,j)x_{ij}\qquad t=0,\mathrm{L}\ ,T-1 \tag{5}$$

$$\text{s.t.}\ \sum_{i=0}^{m}x_{ij}\leqslant b_i\qquad j=0,1,\mathrm{L}\ ,n \tag{6}$$

$$\sum_{j=0}^{n}x_{ij}\leqslant c_j\qquad i=0,1,\mathrm{L}\ ,m \tag{7}$$

$$\begin{aligned}A_\pi(i,j)&=r_{\Delta t_{i\to j}}^{\gamma}+\gamma^{\Delta t_{i\to j}}V(s_j')-V(s_i)\\&=0+\gamma^{\Delta t_{i\to j}}\max_{k'\in k}Q_{\pi^*}(j,k')-\max_{j'\in j}Q_{\pi^*}(i,j')\end{aligned} \tag{8}$$

式中：i,j,k 分别表示空车剩余站索引、空车剩余站（$i=j$，等待）、空车需要站（$i\neq j$，空走结束站）索引和卸车站索引；x_{ij} 和 $\Delta t_{i\to j}$ 分别指从 i 站到 j 站调配/站内等待空车数和动作时间；b_i 指 i 站满足本站装车需要后可供调配的空车剩余量，c_j 指 j 站的空车需求量；$A_\pi(i,j)$ 指单一空车执行从 i 站至 j 站调配动作的优先函数，由动作预期收益与空车在当前时隙下 i 站的状态价值差值计算得到（见式（8））。若 $i=j$，代表仍有一部分剩余空车在 i 站等待至下一时隙 $t+1$，而对 t 时刻的空车流调整不产生任何效益。采用这种优先函数主要考虑因素包括：

① 车辆初始状态价值和调配后的状态价值。空车所处状态值 $V(s_i)$ 对优先函数是负影响，这样保证了剩余空车在低状态价值下可优先参与调配。并且，空走结束站点 $V(s_j')$ 越高，意味着当前状态下剩余空车执行从 i 站至 j 站动作所获得的长期回报越大，在实时调配中该策略具有较高的优先程度。

② 空走距离。假定空走时间 $\Delta t_{i\to j}$ 与空走距离成严格正相关，降低决策周期内空车走行距离（时间）是铁路运输企业追求目标之一，那么对 j 站的状态价值 $V(s_j')$ 引入与空走时间成指数关系的折扣 $\gamma^{\Delta t_{i\to j}}$，即可保证空走时间短的调配动作优先程度更高。

2　模拟计算分析

2.1　算例设计

为了验证算法的有效性，使用广铁集团部分货运数据编写仿真算例进行分析。该案例中有 6 个既是空车剩余站/卸车站又是装车站双重性质的站点，决策周期 $T=5$，确定这一决策周期内所有时隙下 6 个车站最优的空车实时调配方案。

站间运行时间、重走货运收益以及折扣货运收益见表 2，站内等待和空车站间走行不产生货运收益。在每个时隙，6 个站点中既有已满足本站装车的可参与站间调配的剩余空车站点，又有

空车不足需要其余站调拨的站点。各站点剩余空车数、空车需求数见表3。

在此基础上，按照局部 MDP 下铁路单一空车调配策略评估算法，获得所有 $Q(s,a)$ 记录在 Q 表中。根据公式 $V(s)=\max_{a\in A}Q(s,a)$，获得决策周期内不同时空状态下单一空车的状态值函数。由式（8）可计算出 $t=0$ 时刻空车在不同站间走行的优先函数值。

表2 站间运行时间 Δt（天）/货运（重走）收益 $r_{\Delta t}$（元·辆/天）/折扣货运收益 $r_{\Delta t}^{\gamma}$（元/辆）

Tab. 2 Interval running times Δt /freight revenue $r_{\Delta t}$ /discounted freight revenue $r_{\Delta t}^{\gamma}$

站 点	小塘西	阳 春	茂名东	东莞东	汕头北	惠 州
小塘西	0/0/0	1/10/10	2/20/19	4/40/34.39	3/30/27.1	1/10/10
阳 春	1/10/10	0/0/0	1/10/10	3/30/27.1	2/20/19	2/20/19
茂名东	2/20/19	1/10/10	0/0/0	2/20/19	1/10/10	2/20/19
东莞东	3/30/27.1	3/30/27.1	2/20/19	0/0/0	1/10/10	2/20/19
汕头北	2/20/19	2/20/19	1/10/10	1/10/10	0/0/0	1/10/10
惠 州	1/10/10	2/20/19	2/20/19	3/30/27.1	1/10/10	0/0/0

表3 每个时隙下站点空车剩余数和空车需求数

Tab. 3 Remaining and requisite empty wagons of stations at each slot

站 点	$t=0$		$t=1$		$t=2$		$t=3$		$t=4$	
	空车剩余数	空车需求数	空车剩余数	空车需求数	空车剩余数	空车需求数	空车剩余数	空车需求数	空车剩余数	空车需求数
小塘西	8	20	30	80	1	25	80	24	14	30
阳 春	31	14	85	20	51	57	40	30	20	9
茂名东	11	9	76	55	37	26	65	85	47	55
东莞东	0	33	80	60	78	53	77	100	31	46
汕头北	14	8	80	100	85	75	14	9	59	8
惠 州	45	25	36	72	51	67	50	78	22	45
合 计	109	109	387	387	303	303	326	326	193	193

2.2 实验结果及对比试验

采用空车实时优先调配算法对模型求解，部分时刻的空车调配量、调配方向结果节选见表4。求解结果显示所有时刻的站点空车需求均可满足，站内空车利用总数分别为 64/281/257/255/131 辆，站间调配剩余空车总数分别为 45/106/48/71/62 辆，且均在 2 天内完成站间调配，空车需求响应效率高。

表4 $t=0/t=1/t=2$ 时刻空车调配量、调配方向节选结果表

Tab.4 Excerpts from the results of empty wagons and distribution when $t=0/t=1/t=2$

站 点	小塘西	阳 春	茂名东	东莞东	汕头北	惠 州
小塘西	8/30/1	0/0/0	0/0/0	0/0/0	0/0/0	0/0/0
阳 春	2/29/0	14/20/51	0/0/0	15/0/0	0/0/0	0/36/0
茂名东	2/21/5	0/0/6	9/55/26	0/0/0	0/0/0	0/0/0
东莞东	0/0/19	0/0/0	0/0/0	0/60/53	0/20/0	0/0/6
汕头北	6/0/0	0/0/0	0/0/0	0/0/0	8/80/75	0/0/10
惠 州	2/0/0	0/0/0	0/0/0	18/0/0	0/0/0	25/36/49

为了比较所提实时优先调配算法（M）在空车调配中的性能，分别利用空走距离最小化调配模型（M1）、调配结束状态价值最大化模型（M2）对相同空车供求数据进行反复大量实验，从空车调配预期回报（$G_{预期}$，元）、空走总时间（$t_{空}$，天）、折扣预期回报（$G_{折}$，元）和空车调配后的状态价值减少量（$V_{减}$，元）4 个指标进行对比，计算公式如下：

$$G_{预期}=\sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}\cdot V(s_j) \tag{9}$$

$$t_{空}=\sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}\cdot \Delta t_{i\to j} \tag{10}$$

$$G_{折}=\sum_{i=1}^{m}\sum_{j=1}^{n}\gamma^{\Delta t_{i\to j}}\cdot x_{ij}\cdot V(s_j^{'}) \tag{11}$$

$$V_{减}=\sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}\cdot (V(s_j^{'})-V(s_i)) \tag{12}$$

上式中各变量含义同前。

三种模型在所有时隙下的指标结果如表 5 所示。

表 5　指标对比表

Tab. 5　Results comparison

指标对比	$t=0$			$t=1$			$t=2$			$t=3$			$t=4$		
	M	M1	M2	M	M1	M2	M	M1	M2	M	M1	M2	M	M1	M2
$G_{预期}$	1549	1549	1536	3724	3632	3724	1340	1336	1332	1807	1764	1802	1424	1424	1424
$t_{空}$	96	97	133	163	142	163	95	95	95	135	135	145	67	67	67
$G_{折}$	1239	1239	1230	3171	3155	3171	1097	1087	1079	1520	1487	1525	1274	1274	1274
$V_{减}$	316	317	325	381	473	381	414	418	422	677	720	681	487	487	487

由表 5 可知，在多时隙铁路空车实时调配问题上，所提实时优先调配算法（M）总体比空走距离最小化（M1）和调配结束状态价值最大化（M2）模型性能要优。

从 $t=1$ 指标上看，实时优先调配算法仅以空走总时间比 M1 高出 21 天的代价，获得低于其 19.5%的状态价值减少量和高出其 92 元的调配预期回报；虽然在 $t=4$ 时刻，M1、M2 可以和所提模型相媲美，但在 $t=2$ 时隙，M 求解出的空车站间调配策略在与 M1、M2 方案有相同空走总时间 95 天基础上，调配之后状态价值减少量更低，预期回报更大。

结果直接说明了实时优先调配算法中优先函数（式（8））的合理性。即实时调配时，剩余空车优先从状态价值低的起始站点向调配结束站状态价值高且空走距离短的方向调配，以期获得最大调配长期回报、低空走距离和高响应效率。

3　结　论

本文研究了不完全信息下的铁路空车调配问题，建立了基于强化学习的多时隙空车实时调配全局最优模型，首先，将决策周期划分为若干时隙，再通过“基于 Q-learning 的单一空车调配策略评估”和“空车实时优先调配”两阶段求解每一时隙的实时调配策略，最后通过算例与空走距离最小化和调配结束状态价值最大化模型对比。实验结果表明：所提模型可兼顾实时调配预期回报、调配后状态价值和空走距离求解出每个时隙下最优且决策周期全局最优的调配策略，从而方便铁路运输部门快速适应变化的货运市场需求、进行科学合理的运输组织。后续研究中，可以进一步引入车种代用，分析其对空车调配的影响。

参考文献

[1] HOLMBERG K, JOBORN M, LUNDGREN J T. Improved empty freight car distribution [J]. Transportation Science, 1998, 32 (2): 163-73.

[2] 程学庆. 铁路空车调配综合优化模型及求解[J]. 中国铁道科学, 2012, 33 (6): 115-119.

[3] 薛锋, 孙宗胜. 铁路空车调整模型的 D-W 分解算法[J]. 交通运输工程与信息学报, 2019, 17 (4): 43-48.

[4] 朱健梅, 谭云江, 闫海峰. 铁路空车调整优化模型及其蚁群算法[J]. 交通运输工程与信息学报, 2006 (3): 8-15.

[5] 陈胜波, 何世伟, 刘星材, 等. “实货制”下铁路空车动态调配两阶段优化模型与算法研究 [J]. 铁道学报, 2015, 37 (5): 1-8.

[6] 王波, 荣朝和, 黎浩东, 等. 铁路空车调配的多时点优化模型研究 [J]. 交通运输系统工程与信息, 2015, 15 (5): 157-163, 171.

[7] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518 (7540): 529-533.

[8] ZHU M, WANG X, WANG Y. Human-like autonomous car-following model with deep reinforcement learning [J]. Transportation Research Part C: Emerging Technologies, 2018, 97: 348-368.

[9] MAO C, SHEN Z. A reinforcement learning framework for the adaptive routing problem in stochastic time-dependent network [J]. Transportation Research C: Emerging Technologies Partc: 2018, 93: 179-197.

[10] XU Z, LI Z, GUAN Q, et al. Large-Scale Order Dispatch in On-Demand Ride-Hailing Platforms: A Learning and Planning Approach [C]// 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) . London: Assoc Computing Machinery, 2018: 905-913.

[11] WANG Z, QIN Z, TANG X, et al. Deep Reinforcement Learning with Knowledge Transfer for Online Rides Order Dispatching [C]// 2018 Ieee International Conference on Data Mining. New York: IEEE Press, 2018: 617-626.

（责任编辑：刘娉婷）

上接第 30 页

[18] QIN Z, GAO Y. Uncapacitated p-hub location problem with fixed costs and uncertain flows [J]. Journal of Intelligent Manufacturing, 2017, 28 (1): 1-12.

[19] CAMPBELL J F, De MIRANDA G, De CAMARGO R S, et al. hub location and network design with fixed and variable costs[C]// Hawaii International Conference on System Sciences. 2015: 1059-1067.

[20] O'KELLY M E, CAMPBELL J F, CAMARGO R S D, et al. multiple allocation hub location model with fixed arc costs [J]. Geographical Analysis, 2015, 47 (1): 73-96.

[21] GHODRATNAMA A, ARBABI H R, AZARON A. A bi-objective hub location-allocation model considering congestion [J]. Operational Research, 2018: 1-40.

[22] 朱金福. 航空运输规划[M]. 西安: 西北工业大学出版社, 2008: 238-240.

[23] 中国民用航空局, 国家发展和改革委, 交通运输部. 中国民用航空发展第十三个五年规划[R]. 北京: 中国民用航空局, 2016.

[24] 中国民用航空局. 2016 年民航行业发展统计公报[R]. 北京: 中国民用航空局, 2017.

[25] 孙宏, 张培文, 汪瑜. 基于航线网络运力优化分配的机队规划方法[J]. 西南交通大学学报, 2010, 45 (1): 111-115.

（责任编辑：刘娉婷）