专家监督的SAC强化学习重载列车运行优化控制

杨 辉†, 王 禹, 李中奇, 付雅婷, 谭 畅

(华东交通大学 电气与自动化工程学院, 江西 南昌 330013;

江西省先进控制与优化重点实验室, 江西 南昌 330013)

摘要: 重载列车是我国大宗商品运输的重要方式, 因载重大、车身长、线路复杂等因素导致重载列车的控制变得困难. 本文将列车运行过程分为启动牵引、巡航控制、停车制动3个阶段, 基于多质点重载列车纵向动力学模型, 考虑常用空气制动, 利用(SAC)强化学习方法, 结合循环神经网络对专家经验数据进行行为克隆, 并将克隆出的专家策略对强化学习训练进行监督, 训练了一种新的智能驾驶操控策略. 本文的策略可以高效学习驾驶经验数据, 不断从学习中提高目标奖励, 得到最优控制策略. 仿真结果表明: 本文所提的控制策略比未受专家模型监督的强化学习算法更优, 奖励提升的周期更快, 并能获得更高的奖励, 训练出的控制器运行效果更加高效、稳定.

关键词: 重载列车; 强化学习; 行为克隆; 专家策略

引用格式: 杨辉, 王禹, 李中奇, 等. 专家监督的SAC强化学习重载列车运行优化控制. 控制理论与应用, 2022, 39(5): 799 – 808

DOI: 10.7641/CTA.2021.10132

Supervised SAC reinforcement learning method for heavy haul train optimization control

YANG Hui†, WANG Yu, LI Zhong-qi, FU Ya-ting, TAN Chang

(School of Electrical and Automation, East China Jiaotong University, Nanchang Jiangxi 330013, China;
Key Laboratory of Advanced Control and Optimization of Jiangxi Province, Nanchang Jiangxi 330013, China)

Abstract: Heavy haul train is an important transportation way of bulk commodity in our country. The control of heavy haul train becomes difficult due to factors such as heavy load, long body length, and complex line conditions. In this paper, the train operation process is divided into three stages: startup mode, cruise mode, and brake mode. Based on the longitudinal dynamics model of the multi-point mass heavy haul train, the common air brake is considered, using soft actor-critic (SAC) reinforcement learning method, combined with expert contorl strategy that trained by recurrent neural network fitting with expertise data, which called “behavior clone”, to supervise reinforcement learning process. A new intelligent driving control strategy is trained. The strategy in this paper can efficiently learn the driving experience data, continuously improve the total reward from the learning, and obtain the optimal control strategy. The result of simulation shows that the control strategy proposed in this paper is better than the reinforcement learning algorithm that is not supervised by the expert model, the period of reward promotion is faster, higher rewards can be obtained, and the training controller operates more efficiently and stably.

Key words: heavy haul train; reinforcement learning; behavior clone; expertise strategy

Citation: YANG Hui, WANG Yu, LI Zhongqi, et al. Supervised SAC reinforcement learning method for heavy haul train optimization control. *Control Theory & Applications*, 2022, 39(5): 799 – 808

1 引言

目前为止中国铁路电气化里程已超过10万公里. 作为货运专线铁路的顶级种类, 重载铁路是大宗商品的运输的重要渠道, 从大同出发到秦皇岛的大秦铁路更被誉为能源运输的大动脉. 由于重载列车牵引质量大、编组长和线路复杂等特点, 如何安全、稳定的控制

收稿日期: 2021−02−10; 录用日期: 2021−08−05.

†通信作者. E-mail: yhshuo@263.net.

本文责任编委: 刘允刚.

国家自然科学基金项目(U2034211, 62003138, 61803155), 江西省自然科学基金项目(20202BAB202005), 江西省科技专项(20203AEI009), 江西省青年科学基金重点资助项目(20192ACBL21005)资助.

Supported by the National Natural Science Foundation of China (U2034211, 62003138, 61803155), the National Natural Science Foundation of Jiangxi Province (20202BAB202005), the Science and Technology Projects of Jiangxi Province (20203AEI009) and the Youth Sicence Foundation of Jiangxi Province (20192ACBL21005).

列车运行得到了广泛的研究. 目前研究列车智能驾驶的主流方法主要有经典智能决策算法、监督学习决策算法和强化学习决策算法.

在经典智能决策算法的思想中, 多数是按照设计行驶曲线结合控制算法对列车进行控制. ZHANG等[1]以带有电控空气制动系统的重载列车为研究对象, 优化重载列车的空电联合制动为目标, 考虑列车安全、高效和节能因素, 建立并优化了列车多质点模型, 提出了长远距离运输的模型预测控制方法. TANG[2–3]使用鲁棒控制器, 针对高速列车的非线性, 列车质量、参数的不确定性等问题, 进行速度跟踪控制使列车平稳运行. 何之煜[4]利用基于迭代学习控制的自适应控制算法研究了高速列车在时变扰动情况下的追踪控制, 得到跟踪精度高、收敛速度快的控制效果.

在监督学习决策算法的研究中, 王悉在文献[5–6]中对数据不平衡下的重载列车智能操控的研究, 利用EasyEnsemble的思想对数据样本进行随机欠抽样, 再使用(K-nearest neighbor, KNN)算法对数据进行降噪处理, 使重载列车能够在缺少足够类型的样本数据的情况下对列车进行稳定操控, 掌握真实行驶数据情况下, 能够使机器学习的算法更加可靠, 减少因缺少数据产生的偏差.

在强化许学习决策算法的研究中, WANG[7]利用Q学习方法对多质点的重载列车在大下坡的情况下进行了训练, 并在固定容量的Q表下对计算消耗进行了优化, 完成了其最小化纵向车钩力, 准时和安全的目标; TANG[8]利用双Q表切换策略对重载列车进行控制, 根据Q学习离散的控制特性, 可以适用于不同档位的情况; 张淼[9–10]分别提供了一种基于策略梯度和Q学习的强化学习方法, 通过专家经验对高速列车运行环境约束, 达到节能、准点运行的目标.

上述研究使用不同的优化控制方法从多个方面对列车控制进行优化, 强化学习中的策略梯度算法为在线策略, 对历史的运行数据的利用率低, 智能体获得的奖励收敛会相对较慢. 上述文献中的控制对象HXD3机车的控制力给定为调整档位, 是离散的动作空间, Q学习的方法能够训练出较好的控制策略, 而HXD1机车的控制力为给定当前控制力的百分比, 是连续的动作, 使用Q学习的方法无法给出精度较高的控制策略.

针对强化学习在线与离线策略的问题, 最大熵框架提供了一类能够高效稳定训练的算法[11], 本文采用具有稳定的训练过程、能在连续动作空间上训练并有效利用智能体经验数据的(soft actor-critic, SAC)算法[12]架构, 结合钩缓约束和常用空气制动, 对多质点重载列车模型的运行划分成启动牵引、巡航、停车制动3个工况, 分别训练, 并加以专家经验对强化学习过程进行监督, 得到最优的重载列车控制策略. 控制的效果满足安全、稳定、高效的要求.

2 重载列车数学模型

通过对重载列车进行机理建模, 搭建出重载列车的运动学模型和车钩力计算方法. 模型搭建完成后将作为强化学习的环境的一部分, 支持后续控制器的训练. 下面对重载列车运动学建模过程进行详细介绍.

2.1 重载列车多质点纵向动力学模型

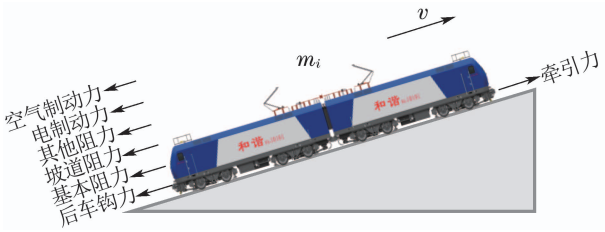
单元万吨重载列车由一辆重载机车在头部牵引, 后方通过车钩连接挂载105辆货运车辆组成, 列车动力单元集中在头部机车. 图1为列车的机车与货车在上坡时的受力情况, 可得重载列车的纵向动力学方程为

$$m_i\ddot{x}_i = F_{\rm T} + F_{i-1} - F_{i+1} - F_{\rm pne} - F_{\rm e} - F_{\rm w} - F_{\rm R}, \tag{1}$$

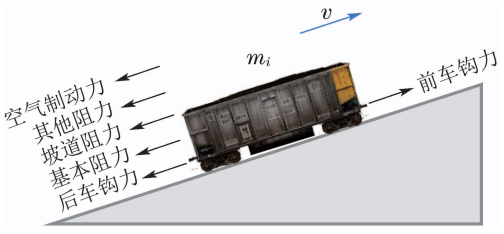
其中: m_i为第i辆车的质量; $\ddot{x}_i$为第i辆车的加速度; $F_{\rm T}$为机车牵引力; $F_{\rm e}$为机车电制动力; F_{i-1}为前车钩力; F_{i+1}为后车钩力; $F_{\rm pne}$为空气制动力; $F_{\rm w}$为车辆运行基本阻力; $F_{\rm R}$为列车的附加阻力, 并且列车的附加阻力由线路弯道附加的弯道阻力、线路坡度造成的坡度阻力和列车通过隧道时的阻力构成, 表达式为

$$F_{\rm R} = F_{\rm Rc} + F_{\rm Ri} + F_{\rm Rt}, \tag{2}$$

其中: $F_{\rm Rc}$为曲线附加阻力, $F_{\rm Ri}$为坡道附加阻力, $F_{\rm Rt}$为隧道阻力.



(a) 机车受力分析



(b) 货车受力分析

图 1 多质点列车受力示意图

Fig. 1 Schematic diagram of multi-mass train force

HXD1型机车牵引制动特性受到机车黏着力和牵引电机功率联合限制, 其牵引制动特性曲线呈现分段式变化, 最大牵引/制动特性式分别为式(3)–(4). 其牵引制动的上限是设计控制器所需的重要约束, 牵引特性与制动特性分别为

$$F_{\mathrm{T}}=\begin{cases}760, & 0\leqslant v\leqslant 5,\\ 779-3.8v, & 5<v\leqslant 65,\\ 34560/v, & v>65,\end{cases}\tag{3}$$

$$F_{\mathrm{b}}=\begin{cases}-153.3v, & 0\leqslant v\leqslant 3,\\ -460, & 3<v\leqslant 75,\\ -34560/v, & v>75,\end{cases}\tag{4}$$

其中v表示列车的速度km/h.

2.2　钩缓装置模型

为了计算多质点受力式(1)中的车钩力大小, 需要对重载列车的钩缓装置进行建模. 根据文献[13]中研究的缓冲器特性对纵向冲动的影响, 精确描述缓冲器的模型有重要意义. 本文参考付雅婷[14]的钩缓装置建模过程, 使用考虑钩缓切换速度的重载列车钩缓装置模型计算每节列车间的车钩力大小. 车钩缓冲器的数学特性通过式(5)–(6)表示.

$$F_{\mathrm{C}}=\begin{cases}f_{\mathrm{u}}, \ (\Delta x*\Delta v\geqslant 0)\cap(|\Delta v|\geqslant v_{\mathrm{e}}),\\ f_3+(f_{\mathrm{u}}-f_3)\dfrac{\Delta v}{v_{\mathrm{e}}}\operatorname{sgn}(\Delta v),\\ \qquad -v_{\mathrm{e}}<\Delta v<v_{\mathrm{e}},\\ f_{\mathrm{l}}, \ (\Delta x*\Delta v<0)\cap(|\Delta v|\geqslant v_{\mathrm{e}}),\end{cases}\tag{5}$$

$$f_3=\frac{f_{\mathrm{u}}+f_{\mathrm{l}}}{2},\tag{6}$$

式中: F_{C}为通过钩缓特性计算出的车钩力, 其值便用作式(1)中F_{i-1}与F_{i+1}; f_{u}为加载时缓冲器的阻抗力; f_{l}为卸载时缓冲器的阻抗力; f_3为缓冲器阻抗力均值; Δv为相邻两车速度差; Δx为相邻两车的运行距离差; v_{e}为缓冲器转换速度.

2.3　空气制动力计算

为了计算多质点受力式(1)中的空气制动力F_{pne}大小, 需要对重载列车空气制动系统进行建模. 重载列车空气制动系统由牵引机车的总风缸、列车管与货车上的副风缸等机构构成. 当重载列车发出空气制动信号时, 总风缸与连通的列车管内的压缩空气向大气排出一部分, 当列车管内气压降低时, 置于货车上的副风缸和列车管形成压力差, 将货车上的闸瓦压挤压向轮毂产生制动力, 此过程称为列车空气制动. 当不需要制动时, 位于机车上的空气压缩机向总风缸内充气, 提升列车管内的气压, 将副风缸与列车管的气压差逐渐降至零, 此过程称为空气制动的缓解.

根据《列车制动》[15]内容, 列车常用制动的计算式为

$$b_{\mathrm{c}}=\beta_{\mathrm{c}}\cdot b=1000\varphi_{\mathrm{h}}\vartheta_{\mathrm{h}}\beta_{\mathrm{c}}\ (\mathrm{N/kN}),\tag{7}$$

本文只考虑常用空气制动, 故$F_{\mathrm{pne}}=b_{\mathrm{c}}$, 其中$\beta_{\mathrm{c}}$为常用制动系数, 表1给出了部分常用制动系数, b_{c}为列车单位制动力, φ_{h}为换算摩擦系数, ϑ_{h}为列车的制动率, b为紧急制动力.

表 1　部分常用空气制动系数

Table 1　Partial common air brake coefficent

减压量r/kPa	50	60	70	80	90
制动系数	0.17	0.28	0.37	0.46	0.53

大秦线上的重载列车采用高摩合成闸瓦, 换算摩擦系数为

$$\varphi_{\mathrm{h}}=0.322\cdot\frac{v+150}{2v+150}.\tag{8}$$

2.4　翟方法多质点状态更新

为了计算出每个时刻的车辆受力, 需要对多质点的速度与位移进行更新. 重载列车对算法的实时性有较高的要求, 设计控制算法要求算法能够在160 ms内迭代更新一次. 本文采用翟方法[16]显示数值积分算法对列车非线性动力学方程(1)进行迭代运算, 计算每个质点当前的速度与位置: 将160 ms等分成100个小步骤进行迭代积分, 满足毫秒级的计算时间与高精度要求. 翟方法更新速度与位置如式(9)所示:

$$\begin{cases}X_n+1=X_n+V_n\Delta t+(\dfrac{1}{2}+\psi)A_n\Delta t^2-\\ \qquad\qquad \psi A_{n-1}\Delta t^2,\\ V_{n+1}=V_n+(1+\varphi)A_n\Delta t-\varphi A_{n-1}\Delta t,\end{cases}\tag{9}$$

式中: X_{n+1}为车辆下一时刻位移量, X_n为车辆当前时刻位移量, V_{n+1}为车辆下一时刻速度量, V_n为车辆当前时刻速度量, A_n为车辆当前时刻加速度量, A_{n-1}为车辆上一时刻加速度量, Δt为时间积分步长, 下标$n-1, n, n+1$分别代表上一步$t=(n-1)\Delta t$时刻, 当前$t=n\Delta t$时刻, 下一步$t=(n+1)\Delta t$; ψ, φ是控制积分方法特性的独立参数.

3　重载列车的强化学习优化控制方法

为简化控制模型设计, 本文使用无模型的强化学习方法训练列车控制器. 另外, 为了让强化学习训练的更加快速稳定, 本文通过克隆专家行为对强化学习的训练过程进行监督. 下面对专家行为克隆和列车控制器的训练进行详细阐述.

3.1　RNN网络克隆专家行为

行为克隆是模仿学习中最简单的一种方法, 利用神经网络将环境状态到动作的映射学习出来, 把状态作为特征, 动作作为被预测的值. 由于重载列车具有非马尔可夫特性, 本文采用循环神经网络对专家行为进行克隆.

循环神经网络(recurrent neural network, RNN)是一个具有时间动态行为的神经网络. 在重载列车人工控制时, 列车乘务员在驾驶列车时的手柄给定位置具有明显的时间动态过程. 采用RNN网络结构对乘务员操作数据进行学习, 意从经验控制数据中学习到某段区域的牵引制动时机以及控制力给定大小. 通过训练出符合驾驶员操作习惯的RNN网络作为指导专家, 用作强化学习控制网络输出的监督, 加速智能体的训练.

专家网络的模型结构为多输入单输出的RNN模型. 根据专家经验, 专家网络输入为最近50个时间步长(8 s)的列车状态、动作序列, 每个时间步的参数为当前时刻列车的速度、位置, 输出为当前时刻采取的动作.

3.2 基于专家监督的SAC算法

本文采用SAC算法[13]架构, 对重载列车控制器进行训练. 该算法是面向最大熵的强化学习开发的一种离线策略算法, 本文的训练过程如图2所示, 主要由强化学习网络、重载列车运行仿真环境和记忆库组成. 强化学习网络利用列车当前的状态输出控制指令, 重载列车运行仿真环境由第2节介绍的重载列车机理模型构成, 仿真环境接受控制指令输出列车的状态以及奖励作为下一时刻强化学习网络控制器的输入, 并存入记忆库中以供网络学习时调用. 在训练控制器参数时, 通过从记忆库中采样数据分别输入到强化学习网络与专家网络中, 通过专家的输出监督强化学习网络的学习.

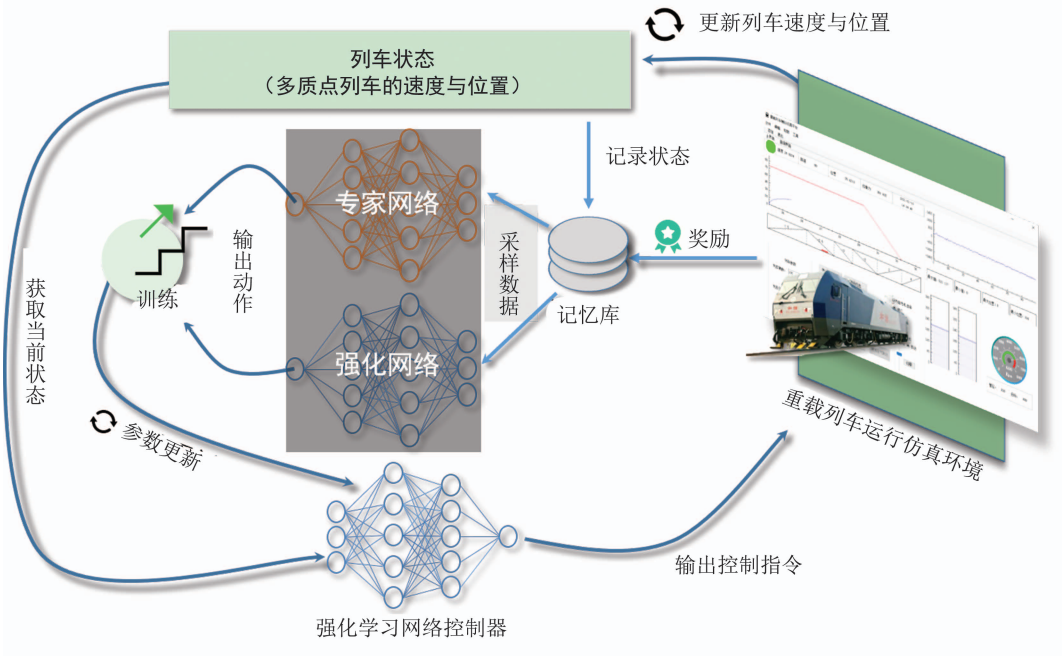


图 2 重载列车强化学习过程

Fig. 2 Reinforcement learning process of heavy haul train

SAC算法的最优策略表达式由式(10)表示, 描述了最优策略跟奖励与熵值之间的关系, 最优策略分为奖励和熵正则化项, 在训练初期的策略随机, 获得的奖励小, 通过增加熵值可以探寻更好的策略; 随着获得的奖励变大, 应当将熵值减小使好的策略能够保持, 直至训练末期, 收获最大的奖励与最小熵, 获得稳定的最优策略.

$$\pi^* = \arg\max_{\pi} \sum_t E_{(s_t,a_t)\sim\rho_\pi}[r(s_t,a_t)+ \alpha\mathcal{H}(\pi(\cdot|s_t))], \tag{10}$$

其中$r(s_t,a_t)$为当前重载列车运行状态下采取控制指令的奖励; $\mathcal{H}(\pi(\cdot|s_t)) = E_{a\sim\pi(\cdot|s)}[-\log(\pi(a|s))]$表示动作的熵值, 熵值前面的系数$\alpha$为温度系数, 控制着控制指令的随机性, 温度系数$\alpha$越大则控制指令越随机, 更有利于重载列车运行状态空间的探索. 每个时刻的状态与动作组合在一起形成了一条轨迹, 对该轨迹求最解大的期望就是期望得到的最优策略.

强化学习中Critic网络计算损失常用的是时间差分(temporal difference, TD)方法, 该方法计算开销小, 且无需系统模型, 网络可在状态和奖励返回后立即更新. Critic网络的损失由式(11)给出, 其中TD学习表达式中的Q_{target}中下一步的Q值在SAC算法中替换成了价值函数$V_{\bar{\theta}}(s_{t+1})$.

$$J_Q(\theta) = E_{(s_t,a_t)\sim\mathcal{D}}[\frac{1}{2}Q_\theta(s_t,a_t)- (r(s_t,a_t)+\gamma E_{s_{t+1}\sim p}[V_{\bar{\theta}}(s_{t+1})])^2], \tag{11}$$

其中: γ为折扣因子, $V_{\bar{\theta}}(s_{t+1})$为target critic网络输出的$Q$值减去熵值, target critic网络和Critic网络结构相

同, 但是参数更新滞后Critic网络. SAC算法中有两个Critic网络Q_{θ_1}, Q_{θ_2}, 在训练时选取最小的Q值作为$Q_\theta(s_t, a_t)$, 减少过估计.

本文采用两种损失函数同时训练Actor网络: 一种由Critic网络输出的价值对Actor网络进行训练, 期望动作价值最大化; 另一种由专家网络在相同环境状态下输出动作, 与Actor网络输出的动作的均方差损失来训练, 以期训练出与专家相似的决策.

评判网络对策略网络的损失由式(12)给出,

$$J_{\text{ori}}(\phi) = E_{s_t\sim\mathcal{D},e_t\sim\mathcal{N}}[\alpha\log\pi_\phi(f_\phi(e_t;s_t)|s_t)- Q_\theta(s_t, f_\phi(e_t;s_t))], \tag{12}$$

其中: π_ϕ是根据f_ϕ隐式定义的, $f_\phi(e_t;s_t)$为重参数化后的控制指令, 帮助网络将误差反向传递; 是从固定分布(如球形高斯分布)中采样的噪声.

专家网络对策略网络的损失由式(13)给出

$$J_{\text{sup}}(\theta^\mu) = E_{s_t\sim\mathcal{D},a_t\sim\pi}[\frac{1}{2}\|f_\phi(e_t;s_t)-\mu(s_t)\|^2], \tag{13}$$

其中: $\mu(s_t)$为专家网络在当前状态s_t下输出的控制指令; $f_\phi(e_t;s_t)$为控制网络在状态s_t下输出的控制指令, 通过计算二者的均方根误差作为专家的监督损失; $\mathcal{D}$, π为状态、策略的空间.

将评价网络输出对控制网络的损失和专家网络输出对控制网络输出的损失乘上比例系数之和得到控制网络的综合损失, 如式(14):

$$J = J_{\text{ori}} + \lambda\cdot J_{\text{sup}}, \tag{14}$$

其中比例系数在智能体找到一条能够完整完成任务的控制策略之后, 将逐步递减, 以实现优化目的.

温度参数决定了熵的大小, 在训练过程中需要自动调整温度参数使模型可以稳定训练, 所以将温度参数作为约束对象, 当作一个优化问题: 最大化期望收益的同时, 保持策略的熵大于一个阈值. 需要优化的表达式如式(15)所示:

$$\begin{aligned}&\max_{\pi_{0:T}} E_{\rho_\pi}[\sum_{t=0}^{T} r(s_t,a_t)],\\ &\text{s.t. } E_{(s_t,a_t)\sim\rho_\pi}[-\log\pi_t(a_t|s_t)] \geqslant \bar{\mathcal{H}};\ \forall t,\end{aligned} \tag{15}$$

其中: $\sum_{t=0}^{T} r(s_t,a_t)$为0到$T$时刻的累计奖励, 对其求解最大期望$E_{\rho_\pi}(\cdot)$, ρ^π表示状态到动作的映射; $E_{(s_t,a_t)\sim\rho_\pi}[-\log\pi_t(a_t|s_t)]$为熵的期望; $\bar{\mathcal{H}}$为最小期望熵, 作为从0时刻到T时刻整条轨迹奖励$\sum_{t=0}^{T} r(s_t, a_t)$期望最大的约束.

根据式(15), 最终得到需要优化的损失函数(16)

$$J(\alpha) = E_{s_t\sim\mathcal{D}}[E_{a\sim\pi_\phi(\cdot|s_t)}[-\alpha\log\pi_t(a|s_t)-\alpha\overline{\mathcal{H}}]]. \tag{16}$$

3.3　状态空间设计

考虑到万吨重载列车由一组车头牵引着105辆货运拖车组成, 而且每一个训练周期都有多个状态转移的步骤. 为简化计算复杂度, 本文状态空间只包含速度与位置, 表示为式(17)

$$\begin{aligned}&s_t = \{v_{t,1}, p_{t,1}\},\ v_{t,1}\in[V_L^{\lim}, V_H^{\lim}],\\ &p_{k,1}\in[P_{\text{st}}, P_{\text{nd}}],\end{aligned} \tag{17}$$

其中: s_t为t时刻智能体所处的状态, $v_{t,1}$表示t时刻头部机车的速度, $p_{t,1}$表示t时刻头部机车所处的位置. $V_L^{\lim}, V_H^{\lim}$分别表示速度的下限与上限, $P_{\text{st}}, P_{\text{nd}}$分别表示列车的起始位置与结束位置, 限速根据路段和运行时间变化, 列车的起始与终结位置则根据训练功能进行调整. $t = 1, 2, \cdots, N_{\text{done}}$, N_{done}为触发了终止条件的时刻.

3.4　奖励设计

好的奖励能够帮助智能体学习, 稀疏的奖励会使智能体在抵达目标前无法获得任何奖励, 加大训练难度, 奖励分布的方差太大也会使得策略梯度太大导致学习不平稳, 将奖励归一化能够有效提升学习效率.

虽然强化学习的本质就是积累奖励使奖励最大化, 奖励的设置与要实现的目标有很强的关联性, 负奖励有利于有限步数内快速结束该回合, 正奖励鼓励智能体不断累积奖励以维持最高奖励状态.

3.4.1　速度奖励设计

智能体学习操控列车有两种目标: 在启动至巡航阶段, 应当获得正奖励以累积奖励值, 鼓励智能体将奖励最大化; 而在制动过程中操控的目标是在安全操纵下停在指定地点, 这时应该将奖励设计为负值, 以期智能体快速达到目标状态. 本文将速度奖励设计为式(18)

$$r_v = \begin{cases}(1-\bar{d}^\eta)(1-\bar{v})^{\max(\bar{d},0.2)}, & d<2,\\ k_1\bar{v}, & 0<\bar{v}<0.875,\\ k_2(1-\bar{v}), & \bar{v}>0.875,\end{cases} \tag{18}$$

其中: $\bar{v} = \dfrac{V_H^{\lim}(p_{t+1,1}) - v_{t+1,1}}{V_H^{\lim}(p_{t+1,1})}$, $\bar{d} = \dfrac{p_{t+1,1}-p_{\text{st}}}{p_{\text{nd}}-p_{\text{st}}}$. d为当前位置到停车点的距离, 单位km; $\bar{v}$和$\bar{d}$为归一化后的速度与距离; η为距离$\bar{d}$的指数参数, 调整η大小可以改变制动阶段到停车点距离的奖励变化斜率; k_1, k_2为有关速度的缩放系数. 距终点2 km以内为停车制动工况下的速度奖励函数, 在启动牵引和巡航工况下的速度奖励为式(20)的后两项. 制动条件下的奖励函数表达式包含了速度和距离两个维度的参数, 距离越近, 速度越低则奖励越高, 如图3所示. 为使列车能够安全行驶, 且速度不超过速度上限, 在给出的每个时

刻速度最大的奖励设定在最高限速的87.5%, 给出一定的冗余空间作为列车安全运行的保障.

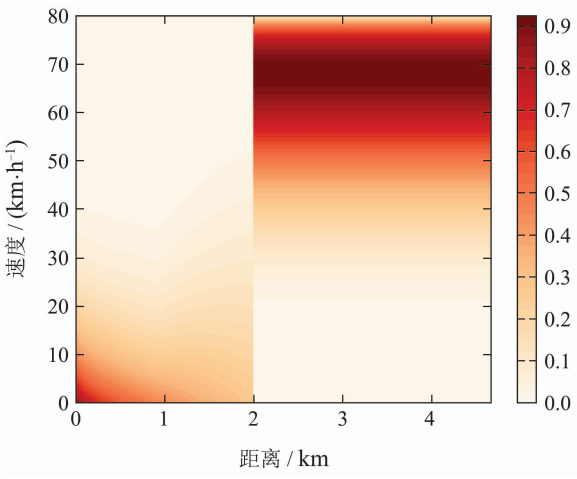


图 3 速度距离相关奖励

Fig. 3 Speed-distance reward

图3为机车速度、机车与下一个站点的距离有关奖励的等高线图, 预设为距离终点2 km处速度上限开始下降, 分界线右侧为当最高限速为80 km/h时, 速度的奖励距离终点2 km外启动牵引和巡航工况的速度奖励, 红色越深则奖励值越大.

为加快算法的训练速度, 对目标探索空间进行约束. 参考TANG[8]的奖励设置, 给定了智能体探索的上限与下限. 上限为机车满级位运行的速度–位移曲线, 下线为以初始40%牵引力, 每隔200步递减1‰的牵引力运行的速度–时间曲线, 对超出期望探索范围的状态–动作进行惩罚.

如果当前速度$v_{t+1,1}$大于当前的速度上限$V_H^{\lim}(p_{t+1,1})$时, 有关最高限速曲线的惩罚为式(19)

$$r_{t+1}^h = -c_1(v_{t+1,1} - V_H^{\lim}(p_{t+1,1}) + c_2)^{c_3} - c_4, \tag{19}$$

否则$r_{t+1}^h = 0$.

如果当前速度$v_{t+1,1}$小于当前的速度下限$V_L^{\lim}(t+1)$时, 有关最低限速曲线的惩罚为式(20)

$$r_{t+1}^l = -c_5 \cdot (V_L^{\lim}(t+1) - v_{t+1,1} + c_6)^{c_7} - c_8, \tag{20}$$

否则$r_{t+1}^l = 0$. 其中r_{t+1}^h为$t+1$时刻最高限速的奖励函数, r_{t+1}^l为$t+1$时刻最低限速的奖励函数, $c_1 \sim c_8$为常数, $v_{t+1,1}$为$t+1$时刻机车的速度, $V_H^{\lim}(p_{t+1,1})$为$t+1$时刻所在位置的速度上限, $V_L^{\lim}(t+1)$为$t+1$时刻速度下限值.

3.4.2 车钩力奖励设计

列车运行过程中, 车钩力应当在最大应力极限范围内变化, 以免造成脱钩事故. 为避免智能体一直施加最小的控制力, 本文将车钩力的奖励函数进行分段处理, 在正常车钩力范围内奖励函数为定值, 当车钩力大于1000时, 车钩力奖励逐渐下降. 最大车钩力的奖励$r_{\rm c}$构造为式(21)

$$r_{\rm c} = \begin{cases} 0.5, & \max F_{\rm c} < 1000, \\ \dfrac{1}{2}{\rm e}^{\frac{\max F_{\rm c} - 1000}{200}}, & 1000 \leqslant \max F_{\rm c} < 2000, \end{cases} \tag{21}$$

其中$\max F_{\rm c}$为整列车的最大车钩力, 车钩力的奖励是将最大车钩力约束到$[-1, 1]$的区间内.

3.4.3 控制力变化奖励设计

车钩力的约束仅在重载列车产生了较高的车钩力时做出惩罚, 为了能够获得一个更为平稳的控制策略, 本文对控制策略进行了约束, 如式(22):

$$r_{\rm F} = \begin{cases} 1, & |F - F'| \leqslant 50, \\ -2, & |F - F'| > 50, \end{cases} \tag{22}$$

其中: F为当前控制力, F'为上一时刻的控制力, 如果当前控制力与上一时刻的控制力差值在50 kN以内, 对智能体给出奖励, 否则惩罚智能体的该动作.

为了将上述奖励函数都发挥作用, 达成多目标优化的效果, 将上述多个目标的奖励线性组合为单个目标奖励以方便目标优化. 奖励的形式为式(23)

$$r = \theta[r_{\rm v} \ \ r_{t+1}^h \ \ r_{t+1}^l \ \ r_{\rm c} \ \ r_{\rm F}]^{\rm T}, \tag{23}$$

参考DSQ方法控制重载列车[8]的奖励参数分配, 并且结合本文定义的奖励对象, 本文各奖励参数分配分别为$\theta = [3 \ \ 1 \ \ 1 \ \ 1 \ \ 1.5]$.

3.5 记忆库采样

在SAC算法中需要引入记忆库存储智能体经验数据来训练智能体. 记忆库的保存方式为$(s, a, r, s'$, done, period), 保存与环境互动产生的当前状态、采取的动作、获得的奖励, 下一时刻的状态、回合终止标志和每个动作状态对应的周期.

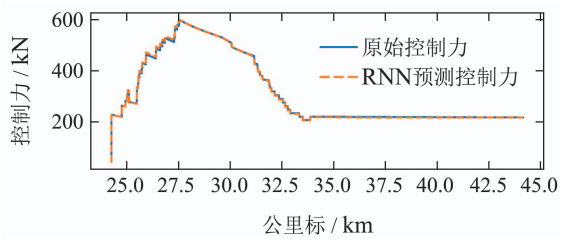
由于专家网络输入为时序结构, 数据采样需要做预处理, 满足采样表示的准确度. 获取记忆库索引, 并且初始化采样计数器. 在当前索引基础上递归向前收集数据, 若采样计数器到达50时, 则采样结束; 若采样计数器为$i(i < 50)$且到达该数据周期头部, 则在该采样数据头部在添加$(50 - i)$个零, 以补全50个数据, 以提供专家策略网络输入.

4 实验与分析

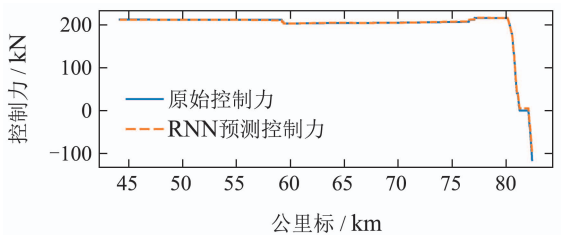
由于线路区段较长, 训练全程的操纵计算的开销非常大. 本文将整个驾驶过程分为3种情况: 启动阶段、巡航阶段、制动阶段, 3段控制的位置为同一区间的连续3段分别为: 24~44 km, 44~83 km, 83~88 km. 通过在3个工况分别训练相应的控制器, 减少训练需要的计算消耗.

4.1　RNN专家网络拟合

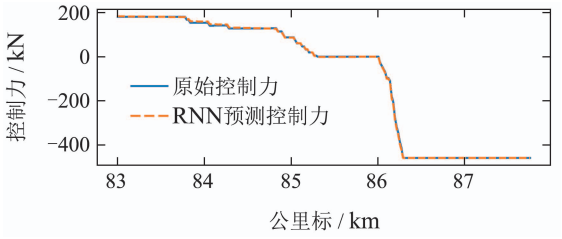
为得到一个具有鲁棒性的预测模型, 在专家经验数据中添加高斯噪声, 训练后的专家模型克隆效果如图4所示. 从图中结果看, 训练后的模型能够很好的拟合专家操控策略.



(a) 启动牵引阶段控制力拟合



(b) 巡航阶段控制力拟合



(c) 停车制动阶段控制力拟合

图 4　RNN拟合专家控制效果

Fig. 4　RNN fitting expert control performance

表2中给出了对重载列车行为克隆之后的专家网络与原始的专家经验之间的误差. 3个运行阶段的控制力的误差控制在3%以内, 误差最大仅有21.63 kN. 对列车的控制来说, 无特别影响.

表 2　专家控制力拟合误差

Table 2　RNN fitting error of expert control force

控制阶段	控制级位最大误差/kN
启动阶段	21.63
巡航阶段	14.63
制动阶段	6.57

4.2　专家监督的SAC强化学习训练

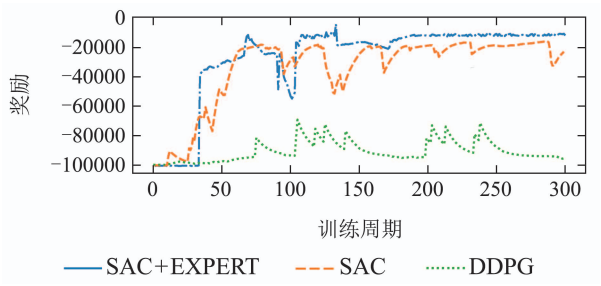
为了使训练后的3种控制器能够联合使用, 让重载列车能够完整的运行, 本文在训练巡航和制动时将对列车初始速度在60~80 km/h的范围内进行一个随机初始化. 本文SAC+EXPERT模型相关参数列于表3中.

表 3　模型相关参数

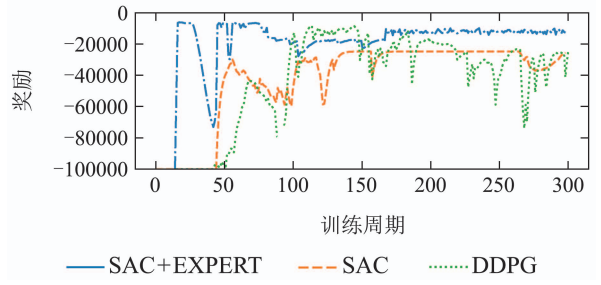
Table 3　Module reference parameters

参数	值
机车质量、货车质量	200 t, 100 t
机车长度、货车长度	35.2 m, 12 m
切换速度v_e	0.1 km/h
翟方法参数ψ, ϕ	0.5, 0.5
初始λ	0.7
温度参数α	0.2
折扣因子γ	0.99
Actor网络每层节点数	2, 256, 128, 3
Critic网络每层节点数	4, 256, 128, 1
超限惩罚参数	$c_1=0.05, c_2=3, c_3=3, c_4=10$ $c_5=0.05, c_6=6, c_7=3, c_8=10$

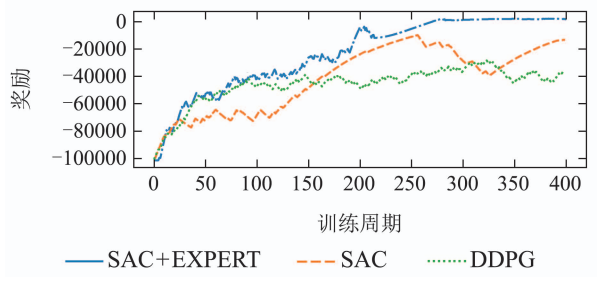
根据图5中3个阶段训练奖励变化的结果, 可以看到在本文提出的SAC+EXPERT算法训练获得的奖励都高于SAC和DDPG两种算法, 在训练的稳定性方面, 通过多次在环境种迭代训练之后, 本文的SAC+EXPERT算法训练的奖励收敛的也都更加稳定, 波动较小.



(a) 启动阶段训练过程



(b) 巡航阶段训练过程



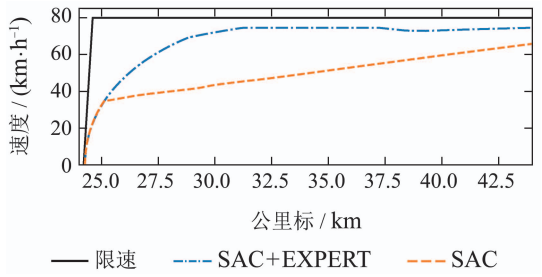
(c) 制动阶段训练过程

图 5　各阶段训练奖励变化

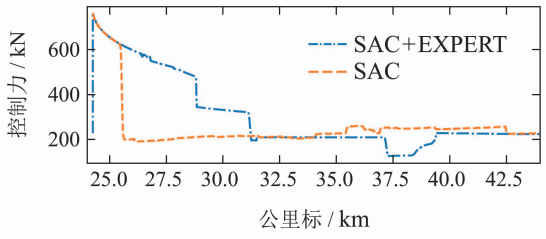
Fig. 5　Reward evalution in three operation training

为了训练一个能够进行巡航制动切换的控制器, 同时参考实际工程环境, 本文选择距离停车点5 km处作为制动阶段的起始点来训练制动控制器, 并设定重载列车停在终点300米以内算作精准停车.

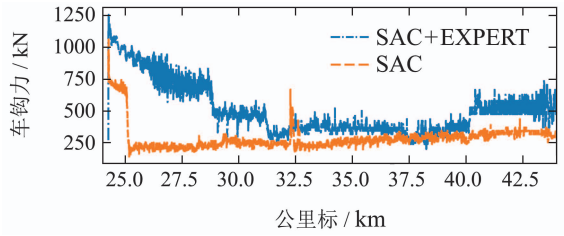
图6–8中的(a)图为SAC+EXPERT和SAC两种算法在启动牵引、巡航控制、停车制动3种工况下控制后得到的速度–位移曲线与对应的控制力曲线. 通过上述3张图可以看出本文提出算法训练出的控制器在3个阶段的平均速度都比较高, 3个阶段的运行效率分别提高了16.75%, 1.88%, 11.18%, 明显的提高了重载列车的运行速度. 在巡航工况下, 本文方法训练出的智能体级位的变化更加平缓. 在制动阶段, 停车位置距离目标停车点300米以内, 达到了预设的停车目标区间.



(a) 启动阶段速度–位移曲线



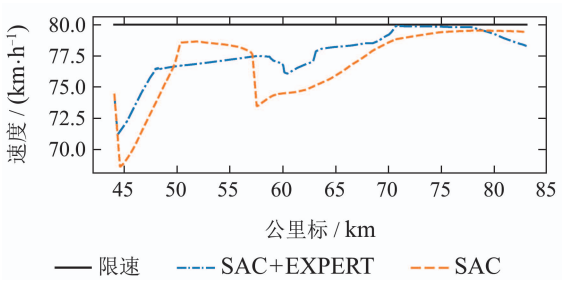
(b) 启动阶段控制力曲线



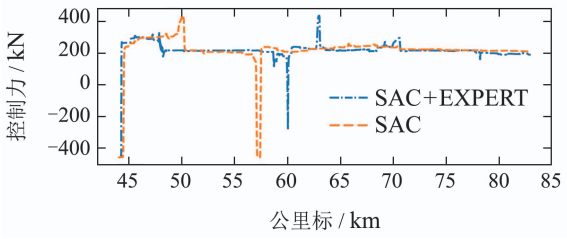
(c) 启动阶段最大车钩力曲线

图 6 启动阶段控制效果

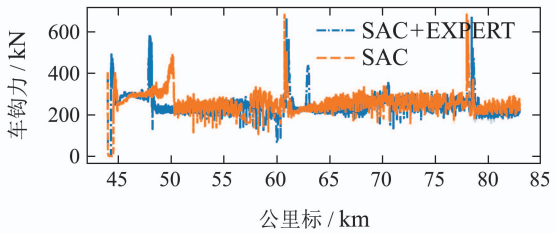
Fig. 6 Performance of control in startup step



(a) 巡航阶段速度–位移曲线



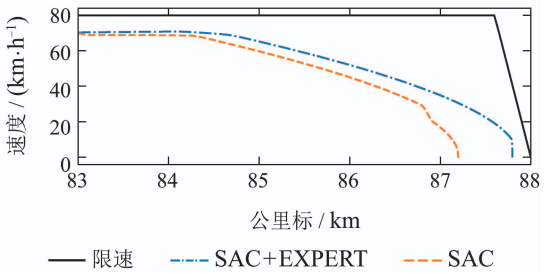
(b) 巡航阶段控制力曲线



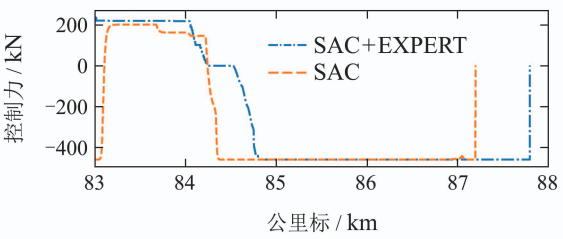
(c) 巡航阶段最大车钩力曲线

图 7 巡航阶段控制效果

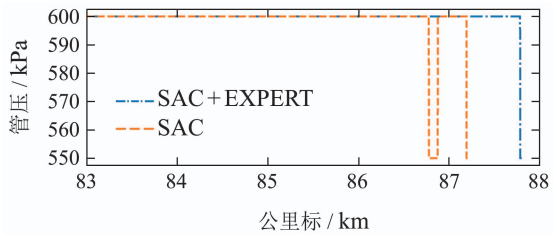
Fig. 7 Performance of control in cruise step



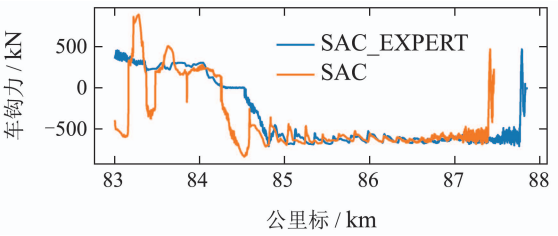
(a) 制动阶段速度–位移曲线



(b) 制动阶段控制力曲线



(c) 制动阶段管压变化



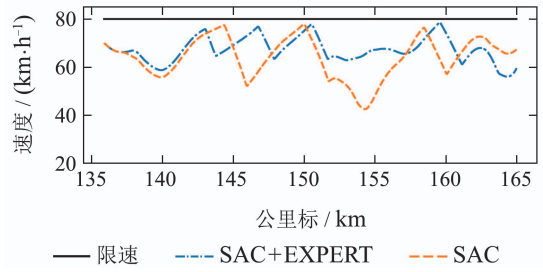
(d) 制动阶段最大车钩力曲线

图 8 制动阶段控制效果

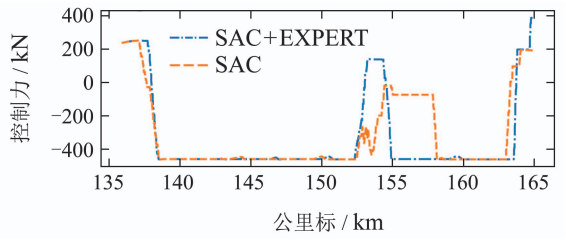
Fig. 8 Performance of control in brake step

图6–8中的(b)图为启动牵引、巡航控制、停车制动3种工况下的控制力变化曲线. 用本文方法训练出的控制器在启动和制动阶段没有出现明显的抖动, 训练出的控制器具有专家控制器的部分特征. 巡航区间中受到专家监督的SAC算法的控制器波动更小, 控制更加稳定. 在制动阶段, 训练出的控制器能够实现通过电制动减速, 并使用空气制动辅助停车的操作方法.

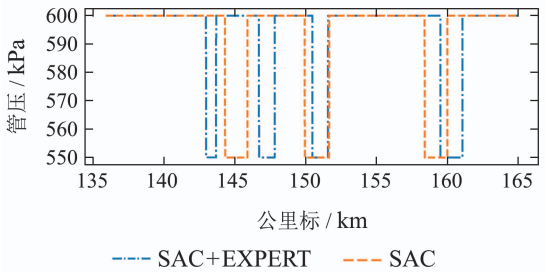
最后, 本文试验了长大下坡路段的循环制动控制, 效果如图9所示. 根据图9(b)(c)(d)中的控制力、目标管压变化以及最大车钩力曲线中, SAC与SAC+EXPERT的运行效果均在车钩最大承受范围内, 较为平稳运行. 在运行效率上结合图9(a)中的速度变化曲线, 带专家监督的SAC策略比SAC的策略高出4.33%.



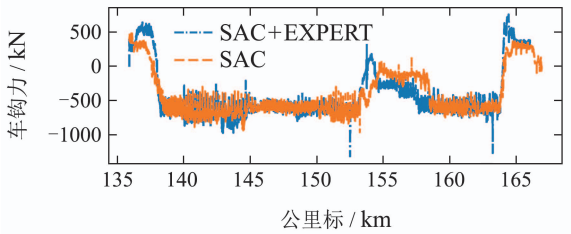
(a) 长达下坡速度曲线



(b) 长达下坡控制力曲线



(c) 长达下坡目标管压变化



(d) 最大车钩力变化情况

图 9　长达下坡控制效果

Fig. 9　Performance of control in long downslope line

5　结语

本文利用强化学习方法对重载列车自动驾驶进行了研究, 本文难点和创新主要围绕以下问题:

1) 针对列车运行距离较长问题, 本文通过将运行过程划分为启动、巡航、制动3个阶段分别进行训练来减少强化学习的学习时间.

2) 针对重载列车运行任务的训练时间长、参数多的问题, 本文通过RNN网络克隆专家经验, 通过比对模型输出与专家(RNN网络)输出对SAC算法训练进行约束, 让网络能够学习到一个可以完成任务的控制器, 并在随着训练进行逐步减小专家监督的比重来优化控制器.

通过以上方法, 本文训练出一个安全、平稳、高效的重载列车运行控制器, 并在最后, 本文对长大下坡的情景进行了实验, 得到一个能够在长达下坡正常运行的循环制动策略.

参考文献:

[1] ZHANG L, ZHUAN X. Optimal operation of heavy haul trains equipped with electronically controlled pneumatic brake systems using model predictive control methodology. *IEEE Transactions on Control Systems Technology*, 2014, 22(1): 13 – 22.

[2] TANG H, WANG Q, FENG X. Robust stochastic control for high-speed trains with nonlinearity, parametric uncertainty, and multiple time-varying delays. *IEEE Transactions on Intelligent Transportation Systems*, 2018, 19(4): 1027 – 1037.

[3] TANG H, GE X, LIU Q, et al. Robust H_∞ control of high-speed trains with parameter uncertainties and unpredictable time-varying delays. *The 35th Chinese Control Conference (CCC)*. Chengdu: IEEE, 2016: 10173 – 10178.

[4] HE Zhiyu, XU Ning. Automatic train operation algorithm based on adaptive iterative learning control theory. *Journal of Transportation Systems Engineering and Information Technology*, 2020, 20(2): 69 – 75.
(何之煜, 徐宁. 基于自适应迭代学习控制的列车自动驾驶算法. 交通运输系统工程与信息, 2020, 20(2): 69 – 75.)

[5] WANG X, LI S, TANG T, et al. Intelligent operation of heavy haul train with data imbalance: A machine learning method. *Knowledge-based Systems*, 2018, 163: 36 – 50.

[6] WANG Xi. *Machine learning based intelligent operation methods for heavy haul train*. Beijing: Beijing Jiaotong University, 2017.
(王悉. 基于机器学习的重载列车智能驾驶方法研究. 北京: 北京交通大学, 2017.)

[7] WANG X, TANG T. Optimal control of heavy haul train on steep downward slope. *IEEE the 19th International Conference on Intelligent Transportation Systems (ITSC)*. Rio de Janeiro: IEEE, 2016: 778 – 783.

[8] TANG H, WANG Y, LIU X, et al. Reinforcement learning approach for optimal control of multiple electric locomotives in a heavy haul freight train: A double-switch-Q-network architecture. *Knowledge-based Systems*, 2020, 190: 105173.

[9] ZHANG Miao, ZHANG Qi, LIU Wentao, et al. A policy-based reinforcemenet learning algorithm for intelligent train control. *Journal of the China Railway Society*, 2020, 42(1): 69 – 75.
(张淼, 张琦, 刘文涛, 等. 一种基于策略梯度强化学习的列车智能控制方法. 铁道学报, 2020, 42(1): 69 – 75.)

[10] ZHANG Miao, ZHANG Qi, ZHANG Zixuan. A study on energy-saving optimization for high-speed railways train based on Q-learning algorithm. *Railway Transport and Economy*, 2019, 1421: 111 – 117.
(张淼, 张琦, 张梓轩. 基于Q学习算法的高速铁路列车节能优化研究. 铁道运输与经济, 2019, 1421: 111 – 117.)

[11] HAARNOJA T, TANG H, ABBEEL P, et al. Reinforcement learning with deep energy-based policies. *The 34th International Conference on Machine Learning (ICML)*. Sydney: PMLR, 2017: 1352 – 1361.

[12] HAARNOJA T, ZHOU A, HARTIKAINEN K, et al. *Soft Actor-Critic Algorithms and Applications*, arXiv preprint, 2018.

[13] ZHAO Xubao, WEI Wei, ZHANG Jun, et al. Influence of segment impedance characteristics of draft gear on longitudinal impulse of heavy haul train. *Journal of the China Railway Society*, 2017, 39(10): 33 – 42.
(赵旭宝, 魏伟, 张军, 等. 缓冲器分段阻抗特性对重载列车纵向冲动的影响. 铁道学报, 2017, 39(10): 33 – 42.)

[14] FU Yating, YUAN Junrong, LI Zhongqi, et al. Optimization of heavy haul train operation process based on coupler constraints. *Acta Automatica Sinica*, 2019, 45(12): 2355 – 2365.
(付雅婷, 原俊荣, 李中奇, 等. 基于钩缓约束的重载列车驾驶过程优化. 自动化学报, 2019, 45(12): 2355 – 2365.)

[15] RAO Zhong. *Traction Computing (2)*. Beijing: China Railway Publishing House, 2010: 63 – 72.
(饶忠. 列车牵引计算(2). 北京: 中国铁道出版社, 2010: 63 – 72.)

[16] ZHAI W M. Two simple fast integration methods for large scale dynamic problems in engineering. *International Journal for Numerical Methods in Engineering*, 1996, 39(24): 4199 – 4214.

作者简介:

杨 辉 教授, 工学博士, 从事复杂工业过程建模、控制与优化、轨道交通自动化与运行优化的研究, E-mail: yhshuo@263.net;

王 禹 硕士研究生, 从事轨道交通运行优化控制方向研究, E-mail: viaytaw@gmail.com;

李中奇 教授, 工学博士, 从事轨道交通自动化与运行优化、控制与运行优化研究, E-mail: 13663699192@163.com;

付雅婷 副教授, 工学博士, 从事轨道交通运行优化控制方向研究, E-mail: fuyating0103@163.com;

谭 畅 副教授, 工学博士, 从事自适应控制、多模型控制、故障诊断与容错控制研究, E-mail: lovetanchang@163.com.