控制与决策
Control and Decision

基于强化学习的地铁站空调系统节能控制

焦焕炎, 冯浩东, 魏东, 冉义兵, 胡朝文

引用本文:
焦焕炎,冯浩东,魏东,冉义兵,胡朝文. 基于强化学习的地铁站空调系统节能控制[J]. 控制与决策, 2022, 37(12): 3139−3148.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2021.0778

基于强化学习的地铁站空调系统节能控制

焦焕炎[1], 冯浩东[1], 魏 东[1,2†], 冉义兵[1,2], 胡朝文[1,3]

(1. 北京建筑大学 电气与信息工程学院，北京 100044；2. 建筑大数据智能处理方法研究北京市重点实验室，北京 100044；3. 北京兴创置地房地产开发有限公司，北京 102600)

摘 要: 地铁站空调系统能源消耗较大,传统控制方法无法兼顾舒适性和节能问题,控制效果不佳,且目前地铁站空调控制系统均是对风系统和水系统单独控制,无法保证整个系统的节能效果. 鉴于此,提出基于强化学习的空调系统节能控制策略. 首先,采用神经网络建立空调系统模型,作为离线训练智能体的模拟环境,以解决无模型强化学习方法在线训练收敛时间长的问题;然后,为了提升算法效率,同时针对地铁站空调系统多维连续动作空间的特点,提出基于多步预测的深度确定性策略梯度算法,设计智能体框架,将其用于与环境模型进行交互训练;此外,为了确定最佳的训练次数,设置了智能体训练终止条件,进一步提升了算法效率;最后,基于武汉某地铁站的实测运行数据进行仿真实验,结果表明,所提出控制策略具有较好的温度跟踪性能,能够保证站台舒适性,且与目前实际系统相比能源节省约17.908 %.

关键词: 强化学习；深度确定性策略梯度法；神经网络；多步预测；地铁站空调系统；节能控制

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2021.0778 **开放科学(资源服务)标识码(OSID):**

引用格式: 焦焕炎,冯浩东,魏东,等. 基于强化学习的地铁站空调系统节能控制[J]. 控制与决策, 2022, 37(12): 3139-3148.

Energy saving control for subway station air conditioning systems based on reinforcement learning

JIAO Huan-yan[1], *FENG Hao-dong*[1], *WEI Dong*[1,2†], *RAN Yi-bing*[1,2], *HU Chao-wen*[1,3]

(1. School of Electrical and Information Engineering，Beijing University of Civil Engineering and Architecture，Beijing 100044，China；2. Beijing Key Laboratory of Intelligent Processing for Building Big Data，Beijing 100044，China；3. Beijing Xingchuang Land Real Estate Development Co., Ltd, Beijing 102600，China)

Abstract: The subway station air conditioning system consumes a lot of energy, and traditional control methods cannot take into account the comfort and energy saving issues together, resulting in poor control effect. Moreover, the current subway station air conditioning control systems control the air system and water system separately, which cannot guarantee the energy saving effect of the whole system. Therefore, this paper proposes an energy-saving control strategy for the system based on reinforcement learning. Firstly, this paper uses a neural network to establish an air conditioning system model as a simulation environment for offline training of the agent to solve the problem of long convergence time of model-free reinforcement learning methods for online training. Then, in order to improve the efficiency of the algorithm and also to address the characteristics of the multidimensional continuous action space of the air conditioning systems, this paper proposes a deep deterministic policy gradient algorithm based on multi-step prediction and designs an agent framework that will be used to interact with the environment model for training. In addition, in order to determine the optimal number of training times, the agent training termination condition is also set, which further improves the algorithm efficiency. Finally, simulation experiments are conducted based on the measured operational data of a subway station in Wuhan, and the results show that the proposed control strategy has better temperature tracking performance and can ensure the comfort of the platform, and the energy saving is about 17.908% compared with the current actual system.

Keywords: reinforcement learning；deep deterministic policy gradient；neural networks；multi-step prediction；subway station air conditioning systems；energy saving control

收稿日期: 2021-05-04；**录用日期:** 2021-08-27.

基金项目: 北京市属高校高水平创新团队建设计划项目 (IDHT20190506)；北京市教委科技计划重点项目 (KZ201810016019)；北京建筑大学市属高校基本科研业务费专项资金项目 (X20068).

责任编委: 孙秋野.

†通讯作者. E-mail: weidong@bucea.edu.cn.

0 引 言

地铁站作为实现城市轨道交通功能性的必要环节,对人们的日常生活具有重要意义. 近年来,随着众多地铁站的快速建设、运营,其相应的能耗也迅速增长,能耗问题日益凸显. 其中,暖通空调(heating, ventilation and air conditioning, HVAC)系统是主要的能耗来源,约占车站总能耗的40%以上,仅次于列车牵引系统[1].

地铁站空调系统的设备一般按照远期高峰小时运行情况进行配置,在运行初中期,客流及行车对数远没有达到设计水平,因此设备选型有较大的富余量,导致能源浪费. 此外,目前国内大部分地铁站仍然依赖用于低层设备的PID调节器,以及用于高层监控系统的基于规则的控制方案[2]. PID控制方法存在参数设定和调试困难的问题,在空调系统负荷和工况发生变化时极易产生振荡,控制效果不佳. 基于规则的控制方法是指根据地铁运行时刻表对各设备采取固定模式的变频技术,该方法存在无法根据实际负荷需求实时调整控制参数的问题,不仅会消耗更多能源,还会使得夏季地铁站台温度偏低,造成人员舒适性差. 另一方面,目前地铁站空调通常对风系统和水系统单独进行控制,而风系统与水系统之间存在耦合关系,单独控制难以实现系统精准节能,也很难保证人员的舒适性要求. 要降低地铁站空调系统的运行能耗,就必须在保证车站舒适度的前提下采取合理可行的节能控制方案.

已有研究表明,智能控制方法具有自适应、自学习和自协调能力,能够提升空调系统的性能和节能效果. 其中,强化学习(reinforcement learning, RL)[3]中的智能体通过与环境之间的直接交互最大化奖励信号,能够实现复杂系统的全局优化控制,是充分发挥空调系统节能潜力的有效方法之一.

近年来,多位学者研究了基于强化学习的空调系统节能控制方法. 戴小燕等[4]提出了一种基于蒙特卡罗算法的物联网架构云端数据中心的空调系统节能控制策略,在保证人员舒适度的前提下,经测试可实现15%～20%的节能率. 闫军威等[5]选取离散化的冷冻水出水温度和冷冻泵频率作为控制变量,利用Double-DQN算法实现了中央空调系统的节能优化运行. Yuan等[6]将基于规则的控制算法与基于RL的控制算法相结合,应用于变风量空调系统运行优化. 结果表明,RL控制器在空调系统的舒适性和能耗方面表现更好,与基于规则和PID策略相比,系统的总能耗分别降低了7.7%和4.7%. Dalamagkidis等[7]采用无模型强化学习方法控制通风风扇速度和窗户开启程度,实验结果表明,该策略取得了比开关控制和模糊PD控制器更好的性能.

从以上研究成果看,应用强化学习方法控制地铁站空调系统可以有效提升系统的节能效果,不过目前还有两个问题有待解决: 1) 基于无模型的强化学习方法在线训练智能体的收敛时间较长,例如文献[7]的训练时间达到48个月. 为解决这一问题,本文获取武汉某地铁站空调系统一年的运行数据,采用神经网络建立系统模型,将其作为离线训练强化学习智能体的模拟环境,以缩减智能体训练时间. 2) 地铁站空调系统的状态空间和动作空间都是多维连续的,然而目前大多数相关研究只处理参数空间有限的问题,且只针对单个离散控制变量产生控制律,这限制了它们对复杂系统控制的适用性. 针对这一问题,本文提出了基于多步预测的深度确定性策略梯度(deep deterministic policy gradient, DDPG)算法,对地铁站空调风系统和水系统进行全局优化控制. DDPG算法可以适应多维连续动作空间的系统,同时对传统DDPG算法进行改进,基于多步预测,使智能体择优更新参数,提升了算法的学习效率. 此外,利用邻近训练过程中总奖励值的变化趋势设置智能体训练终止条件,进一步减少算法的训练时间. 最后,以武汉某地铁站的空调系统为研究对象,以满足站台温度需求和降低系统整体能耗为目标,将空调水系统和风系统视为一个整体,基于系统实际运行数据设计智能体训练方法,进行仿真实验. 结果表明,所提出控制策略与传统DDPG算法相比,智能体训练次数减少86次,且能够在系统负荷变化的情况下使系统稳定运行,满足车站温度需求,同时与目前实际工程中的运行系统相比,节能约17.908%.

1 系统描述

1.1 研究对象

地铁站空调系统由大系统(末端风系统)、小系统和水系统构成,3部分组成一个有机的整体,共同作用完成车站环境参数的调节. 大系统为车站公共区的通风空调系统,其服务区域为车站站厅和站台,主要用于维持温度、湿度和二氧化碳浓度在所要求的区间范围内,以满足乘客过渡性舒适需求. 小系统的服务区域为车站辅助功能用房,包括车站管理用房及设备用房,主要负责提供舒适的人员工作环境和适宜的设备运行条件. 与大系统相比,小系统承担的服务面积和系统负荷较小,其能耗低,节能空间较小. 水系统为车站组合空调机组完成热交换过程提供冷源,从而

实现地铁站的温度调节.

鉴于小系统节能空间较小,本文以大系统和水系统为被控对象,研究基于强化学习的地铁站空调系统节能控制策略. 整个被控对象包括水系统的冷水机组、冷冻泵、冷却泵和冷却塔,以及大系统末端风机,系统示意图如图1所示.

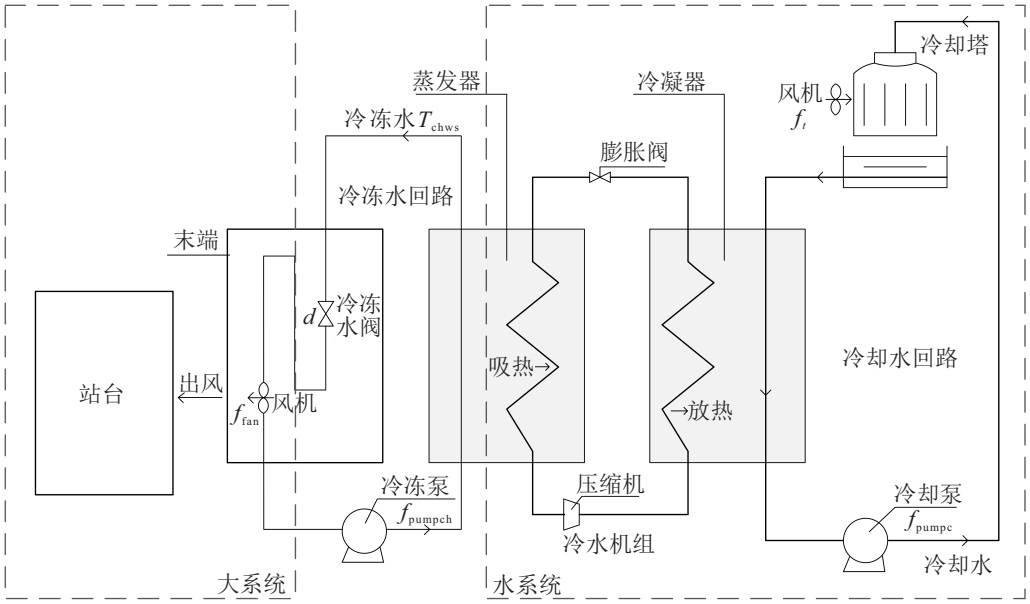


图 1　地铁站空调系统示意图

1.2　系统控制目标

地铁站空调系统的主要能耗设备包括冷水机组、冷冻水泵、冷却水泵、冷却塔风机和末端风机. 其中,末端风机和冷水机组是最主要的耗能设备,二者共占据约80%的能耗,冷冻水泵、冷却水泵和冷却塔占据约20%的能耗.

地铁站空调系统的控制目标是在满足地铁站台温度需求的前提下尽可能节约能源,使空调系统能效比EER (energy efficiency ratio)值达到最大. EER值越高,表示空调系统能够在消耗越少的电能情况下提供越多的冷量. 空调系统能效比EER计算为

$$\mathrm{EER} = Q_{\mathrm{ch}}/P_{\mathrm{total}}\,. \tag{1}$$

其中: Q_{ch} 为冷水机组制备的冷量,单位kW; P_{total} 为空调系统各设备的运行功率总和,单位kW,有

$$P_{\mathrm{total}} = P_{\mathrm{chiller}} + P_{\mathrm{pumpch}} + P_{\mathrm{pumpc}} + P_{\mathrm{tower}} + P_{\mathrm{fan}}\,, \tag{2}$$

P_{chiller} 为冷水机组的运行功率, P_{pumpch}、P_{pumpc} 分别为冷冻水泵运行功率和冷却水泵运行功率, P_{tower} 为冷却塔风机运行功率, P_{fan} 为末端风机的功率,单位均为kW.

2　系统建模

为减少智能体的训练时间,首先需要对系统进行建模,构建与智能体交互的模拟环境. 空调系统设备众多,系统的状态参数与设备控制参数之间呈非线性关系,使用传统机理建模方法较为困难[8]. 为此,研究人员提出了多种基于数据驱动的建模方法,如数据挖掘算法(人工神经网络-ANN[9]、支持向量机-SVM[10])、统计模型(回归[11]等)、几何模型[12]以及随机模型(概率密度函数逼近[13])等. 在这些建模方法中,神经网络算法无需繁冗的建模过程,且模型精度较高[14],相比于其他方法,在非线性系统建模方面更具优势. 因此,本文采用神经网络获取系统模型,建立神经网络模型需要合理选择建模参数,以提高模型的可理解性、可扩展性和准确性[15].

2.1　系统模型参数

由于控制目标是使系统在满足舒适性要求的前提下使EER尽可能大,神经网络模型输出应为系统下一时刻站台温度和EER,输入量包括状态变量和控制变量.

地铁站一般位于地下,影响空调系统能效的因素包括室外环境、客流量、列车产热、隧道换热和设备散热等[16]. 客流量对地铁站空调系统负荷具有显著影响,而本文无法获取武汉该地铁站的客流量数据信息. 通过文献[17]发现,地铁站内客流变化与站内系统负荷变化存在近似正比关系,且呈现相对规律的周期性变化,如图2所示. 因此,本文以系统负荷代替客流量,作为模型的一个输入参数. 列车产热、隧道换热和设备等的发热量对系统总负荷不具有显著影响. 最终,选取影响地铁站空调系统温度和EER的主

要状态变量为室外温度、室外相对湿度和系统负荷.

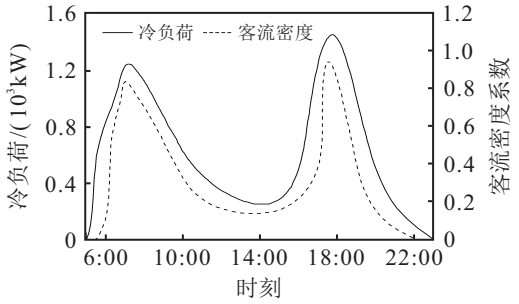


图 2 车站冷负荷与客流密度关系

另一方面,水泵的控制变量是水泵流量,末端风机的控制变量是输出给相应变频器的频率信号,冷却塔风机的控制变量是控制电压信号. 而冷水机组结构复杂,其内部有厂家设置好的控制器,不允许外部控制器对其进行控制,因此选择冷冻水供水温度设定值作为冷水机组的控制变量,利用控制算法计算使系统优化目标达到最佳时的冷冻水供水温度设定值. 当优化后的设定值传给冷水机组后,冷水机组会自动通过内部控制器,使其出水温度跟上冷冻水的设定值. 另外,为了实现风水联动控制,通过末端组合式空调上的冷冻水阀将大系统与水系统有机地结合起来,以实现整个系统的协调工作和动态水力平衡控制[18]. 综上所述,系统控制变量选择为冷冻水供水温度、冷冻水泵流量、冷却水泵流量、冷却塔风机电压信号、末端风机频率和冷冻水阀开度. 由于当前时刻的温度和能效比对系统下一时刻状态也有影响,最终空调系统预测模型的输入参数有: 1) 室外温度 $T_{\text{out}}[k]$; 2) 室外相对湿度 $\text{RH}_{\text{out}}[k]$; 3) 系统负荷 $L[k]$; 4) 系统能效比 $\text{EER}[k]$; 5) 站台温度 $T_{\text{in}}[k]$; 6) 冷冻水供水温度 $T_{\text{chws}}[k]$; 7) 冷冻水泵流量 $f_{\text{pumpch}}[k]$; 8) 冷却水泵流量 $f_{\text{pumpc}}[k]$; 9) 冷却塔风机电压 $f_{\text{t}}[k]$; 10) 末端风机频率 $f_{\text{fan}}[k]$; 11) 冷冻水阀开度 $d[k]$. 其中 k 为当前时刻. 模型的输出为下一时刻系统能效比 $\text{EER}[k+1]$ 和站台温度 $T_{\text{in}}[k+1]$.

2.2 数据采集和预处理

选取夏季6月～9月系统实测样本数据进行系统建模,采样周期为5 min,这些数据均由地铁站监控管理系统提供,部分数据如表1所示. 所有样本数据被随机分成两个不同的数据集,用于对神经网络模型的训练和测试,数据占比分别为70 %和30 %.

表 1 部分数据信息

参数	个数									
	1	2	3	4	5	6	7	8	9	10
能效比EER	4.58	4.54	4.49	4.63	4.23	4.14	4.19	4.17	4.2	4.06
站台温度/°C	23.9	23.89	23.85	23.95	23.78	23.7	23.74	23.74	23.74	23.6
室外温度/°C	32.6	31.9	31.6	31.6	31	30.9	31	30.9	30.9	30.9
室外相对湿度/ %	68.4	68.1	70.7	73.4	67.2	66.2	66.9	66.2	69.1	67.8
系统负荷/kW	556	551	546	558	538	528	533	533	534	515
冷冻水供水温度/°C	8.2	8.2	8.2	8.1	7.8	7.9	7.9	8	8	8.1
冷却塔风机电压/V	5.7	5.7	5.5	5.4	5.3	5.3	5.3	5.4	5.4	5.4
冷冻水泵流量/(m^3/h)	111	110	109	109	110	108	109	109	109	108
冷却水泵流量/(m^3/h)	162	161	159	163	163	158	159	163	163	159

在神经网络的训练过程中,由于模型各输入变量的量纲不同,且数据值的大小及范围差异较大,会使网络训练速度变慢,甚至出现最终无法收敛的情况,需要对实测样本数据进行归一化和反归一化处理. 在进行处理时,采用线性函数转换方法[19],将数据转换成0～1范围内的数值.

2.3 神经网络结构

选用3层前馈神经网络(1个输入层、1个隐含层和1个输出层)建立系统模型,3层前馈神经网络已被证明能以高精度有效逼近任何一个非线性过程[20]. 网络隐层神经元数目对预测模型的性能有显著影响,然而,现有文献并没有提出明确的解析函数来预先确定隐层神经元的数量,一般可通过基于下式的试错法[21]计算得到隐层神经元数量:

$$m = \sqrt{n+l} + \alpha. \tag{3}$$

其中: n 和 l 分别为输入层和输出层节点数; m 为隐含层节点数; α 为试凑常数, 取值范围通常是 $[1, 10]$.

利用式(3)计算得出系统模型的隐含层节点数取值范围为[5,13]. 综合考虑网络误差和网络泛化性能,通过实验测定发现,当隐含层节点数为10时网络训练效果最佳,均方根误差最小,实验比较结果如表2所示.

表 2　不同情况下神经网络模型的均方根误差

隐含层神经元个数	RMSE
8	0.013 1
9	0.014 7
10	0.012 7
11	0.015 0
12	0.013 2
13	0.014 5

最终,本文神经网络模型的拓扑结构与各参数设置如表3所示.

表 3　系统模型拓扑结构与神经网络参数设置

参数名	参数设置
神经网络类型	三层全连接层网络
输入层节点数	11
隐含层节点数	10
输出层节点数	2
激活函数	隐含层RELU, 输出层SIGMOID函数
损失函数	MSE(均方差)
优化器	ADAM[22]
网络评价指标	MAE(平均绝对误差)

2.4　模型性能分析

利用所构建的系统模型进行测试,测试集数据共1 000组,测试结果如表4所示. 由表4可见,模型测试输出与所对应的目标输出之间误差较小,表明所构建的空调系统模型可以以较高的精度反映系统输入输出样本数据对中固有的非线性映射关系,且具有结构简单的特点,同时能够避免进行繁琐的计算过程,具有工程实用价值.

表 4　系统模型拓扑结构与神经网络参数设置

模型输出	站台温度	能效比EER
误差范围	−0.5～0.5°C	−0.3～0.4
MAE(平均绝对误差)	0.161°C	0.136
MRE(平均相对误差)	0.665 %	2.861 %
MSE(均方差)	0.086	0.04

3　基于多步预测的强化学习控制

地铁站空调系统状态空间和动作空间均是多维连续的,强化学习深度确定性策略梯度(DDPG)法对于解决这一类问题非常有效. DDPG算法属于强化学习Actor-Critic(AC)方法[23]. AC方法智能体共包含两部分,即Actor部分和Critic部分. 智能体的唯一目标是找到一个最优策略最大化其长期回报,即价值. 为了实现该目标,智能体与环境不断进行交互训练,智能体通过决策选择动作,环境根据这些动作做出相应的响应,并反馈给智能体.

3.1　强化学习模型要素

利用强化学习方法解决地铁站空调系统节能控制问题,首先需要将该问题转化为强化学习模型. 在强化学习模型中,有几个核心要素:状态、动作、奖励信号、策略和价值函数.

地铁站空调系统需要维持地铁站台温度在一定范围内,而每个时间步长的站台温度仅由当前环境状态和控制设备输入决定,独立于系统以前的状态. 确定强化学习智能体的状态就是系统模型的状态变量,动作就是系统的控制变量,即此处状态和动作是第2节中系统模型的输入变量,具体为:状态 $S=[T_{\text{out}}, \text{RH}_{\text{out}}, L, \text{EER}, T_{\text{in}}]$,动作 $A=[T_{\text{chws}}, f_{\text{pumpch}}, f_{\text{pumpc}}, f_{\text{t}}, f_{\text{fan}}, d]$. 地铁站空调系统控制的目标是使站台温度实时跟踪设定值,同时使系统能效比最大,设置奖励信号 R 为

$$R=-|T_{\text{in}}-T_{\text{in\_set}}|+\text{e}^{\text{EER}}/100, \tag{4}$$

其中 $T_{\text{in\_set}}$ 为站台温度设定值,根据地铁站环境控制要求,将地铁站台夏季的设计温度值定为24°C,即 $T_{\text{in\_set}}=24°\text{C}$. 式(4)中前1项 $-|T_{\text{in}}-T_{\text{in\_set}}|$ 表示当站台实际温度越接近设定值时奖励值越大,后一项 e^{EER} 表示系统能效比的指数函数,能效比越大奖励值越大,同时随着EER越来越大,奖励值的变化也越来越大,最后用该值除以100是为了防止奖励值过大,不利于计算.

在DDPG算法中,智能体Actor部分和Critic部分一般均用神经网络表示,Actor网络映射策略函数,Critic网络映射价值函数,算法训练的过程即为更新智能体网络的过程,最终目的是寻找出最优策略网络.

3.2　基于多步预测的DDPG算法

传统DDPG算法中智能体所有的数据均来自环境模型的反馈,智能体只利用过去的数据对当前行为进行优化和提升,限制了智能体的学习速度. 模型预测控制[24]作为一种智能优化控制算法,采用多步预测、滚动优化和反馈校正等策略. 其中,滚动优化与传统的全局优化不同,其在每一时刻的优化性能指标只涉及从该时刻起到未来有限的时间内,而到下一时刻,这一优化时间同时向前推移,不断地进行在线优化,具有鲁棒性强、对模型精确性要求不高等优点. 滚动优化以多步预测为基础,在每一优化时刻,算法利用系统模型预测未来有限时间内的系统状态和动作,再求解优化性能指标. 利用基于多步预测的思想能够使系统提前采取行动,选择最优结果,提高算法的学习效率.

对于强化学习智能体而言,为了使它不必局限于只从与环境模型的交互中获取数据,让智能体能够利用所预测的数据择优更新参数,本文基于多步预测滚动优化的思想提出了基于多步预测的DDPG算法.

算法1 基于多步预测的DDPG算法.

输入: Actor当前网络$\pi(s,\theta)$,Actor目标网络$\pi'(s,\boldsymbol{\theta}')$,Critic当前网络$q(s,a,\boldsymbol{w})$,Critic目标网络$q'(s,a,\boldsymbol{w}')$, 参数分别为$\boldsymbol{\theta}$、$\boldsymbol{\theta}'$、$\boldsymbol{w}$ 和$\boldsymbol{w}'$;随机噪声ξ,经验回放池集合D.

算法参数: 批量梯度下降的样本数m,目标网络参数更新频率C,最大迭代次数N,步长$\alpha^{\boldsymbol{\theta}}>0$,$\alpha^{\boldsymbol{w}}>0$,折扣因子$\gamma$,软更新系数$\tau$,预测组数$n$,预测步数$p$.

输出: 最优Actor当前网络参数$\boldsymbol{\theta}$和Critic当前网络参数$\boldsymbol{w}$.

step 1: 随机初始化参数$\boldsymbol{\theta}$、$\boldsymbol{w}$、$\boldsymbol{\theta}'=\boldsymbol{\theta}$、$\boldsymbol{w}'=\boldsymbol{w}$,清空经验回放池$D$.

step 2: 初始化状态S,令$i=1,k=1$.

step 3: 基于Actor当前网络生成一组动作$A_{i,k}$,并添加随机噪声ξ.

step 4: 利用系统模型,输入当前状态和$A_{i,k}$,得到下一时刻系统T_{in}和EER,利用环境计算奖励值,并将下一时刻系统状态作为当前状态.

step 5: 若$k=p$,则转至step 6,否则令$k=k+1$,转至step 3.

step 6: 若$i=n$,则转至step 7,否则,令$i=i+1,k=1$,转至step 3;

step 7: 计算每一组的总奖励值$R_{\text{total}}(A_{i,k})$,共$n$个,令$A=\arg\max(R_{\text{total}}(A_{i,k}))$.

step 8: 利用模拟环境执行动作A,得到下一时刻状态S'和奖励R.

step 9: 将$\{S,A,R,S'\}$四元组存入经验回放池D.

step 10: $S\leftarrow S'$.

step 11: 从集合D中随机采样m个样本$\{S_j,A_j,R_j,S'_j\},j=1,2,\cdots,m$,由式(6)计算当前目标$q$值$y_j$.

step 12: 由式(7)计算均方差损失函数$J(\boldsymbol{w})$,更新$\boldsymbol{w}$:$\boldsymbol{w}\leftarrow\boldsymbol{w}-\alpha^{\text{w}}\nabla J(\boldsymbol{w})$.

step 13: 由式(8)计算损失函数$J(\boldsymbol{\theta})$,更新$\boldsymbol{\theta}$:$\boldsymbol{\theta}\leftarrow\boldsymbol{\theta}-\alpha^{\theta}\nabla J(\boldsymbol{\theta})$.

step 14: 如果$N\%C=1$,则由式(9)和(10)更新目标网络参数.

step 15: 若S为非终止状态,则令$i=1,k=1$,转step 3,否则,转至下一步.

step 16: 若迭代次数小于N,则转至step 2,否则结束.

算法1详细描述了基于多步预测的DDPG算法的流程,其中Actor和Critic为两个神经网络,分别用$\pi(s,\boldsymbol{\theta})$和$q(s,a,\boldsymbol{w})$表示,即策略和价值. Actor网络将状态$s$映射到动作$a$,而Critic网络通过遵循当前状态对应的策略计算预期价值q. 在DDPG算法中,Actor网络的输出策略便是系统的控制动作,即$\pi(s,\boldsymbol{\theta})=a$. 算法随机初始化每个网络的权重$\boldsymbol{\theta}$和$\boldsymbol{w}$. 为提高算法训练的稳定性,DDPG算法中共包含4个网络,除Actor和Critic当前网络外,另两个网络分别为Actor和Critic的目标网络$\pi'(s,\boldsymbol{\theta}')$和$q'(s,a,\boldsymbol{w}')$,这两个网络用于在更新当前网络权值时计算目标值. 目标网络的权值$\boldsymbol{\theta}'$和$\boldsymbol{w}'$初始化为与当前网络相同,然后每隔一段时间算法将当前网络参数复制到目标网络进行更新.

为了在智能体训练时去除相邻时刻训练数据之间的相关性,同时提高数据利用率,算法中人为定义了一个有限的缓存区D作为经验回放池,用于将每次与环境交互得到的奖励与状态更新情况都保存起来. 在每次更新网络、计算目标值时,算法不是使用在每个决策时刻立即收集的转换样本,而是从经验回放池D中随机抽取少量的转换样本,对网络进行训练.

为了使算法具有一定的在线探索性,避免错过其他较好的动作,在step 3使用的策略表示为

$$A=\pi(S,\boldsymbol{\theta})+\xi. \tag{5}$$

其中:A为智能体施加给环境的动作;ξ为随机高斯噪声,将其添加到动作中的目的是确保探索性并防止算法收敛到局部最优解.

传统DDPG算法在step 3后直接执行step 8. 本文在step 3后添加了step 4$\sim$step 7,引入多步预测的思想,即先基于Actor网络生成多组(n组)动作,并为每一组动作添加随机噪声,然后每组基于第1个动作利用环境模型和Actor当前网络向前预测p步,最终产生p个未来时刻的状态、动作和奖励值,共有n组;然后计算每组的总奖励值,即预测的p步的奖励之和,这样便有n个总奖励值,算法从中选择最大的一个值,该组的第1个动作被作为传统算法step 3的返回值,接着执行step 8. 经过上述改进之后,算法每次都能够选择具有更高奖励值的数据进行训练,可以提升其学习效率. 环境执行动作A后,系统得到下一时刻系统状态S'和奖励值R,并将$\{S,A,R,S'\}$四元组存入经验回放池.

更新网络参数时,智能体从经验回放池中随机采样m个样本$\{S_j,A_j,R_j,S'_j\}(j=1,2,\cdots,m)$,计算当前目标$q$值$y_j$,有

$$y_j=R_j+\gamma q'(S',\pi'(S',\boldsymbol{\theta}'),\boldsymbol{w}'), \tag{6}$$

其中: γ 为折扣因子, 决定了未来奖励的比重, $0 \leqslant \gamma \leqslant 1$.

Critic 网络的损失函数为

$$J(\boldsymbol{w}) = \frac{1}{m}\sum_{j=1}^{m}(y_j - q(S_j, A_j, \boldsymbol{w}))^2, \tag{7}$$

表示使式 (6) 中的目标 q 值与 Critic 网络输出的期望价值之间的误差最小.

根据策略梯度法, Actor 网络的损失函数为

$$J(\boldsymbol{\theta}) = -\frac{1}{m}\sum_{j=1}^{m} q(S_j, A_j, \boldsymbol{w}). \tag{8}$$

算法采用梯度下降法对 Actor 和 Critic 网络的权值进行更新. 当满足目标网络更新频率 C 时, 算法使用如下软更新公式更新目标网络参数:

$$\boldsymbol{w}' \leftarrow \tau\boldsymbol{w} + (1-\tau)\boldsymbol{w}', \tag{9}$$

$$\boldsymbol{\theta}' \leftarrow \tau\boldsymbol{\theta} + (1-\tau)\boldsymbol{\theta}', \tag{10}$$

其中 τ 为软更新系数, 这样可以避免损失值出现较大波动.

3.3　控制系统设计

图 3 所示为所提出的基于强化学习的控制策略智能体训练系统结构, 图中模拟环境利用第 2 节构建的神经网络空调系统模型预测下一时刻站台温度和 EER. 同时, 由于是离线训练智能体, 无法获取实际室外温度、湿度等不可控变量, 且难以预测, 在环境中这些变量是根据时间变化从往年提供的真实数据表格中进行读取.

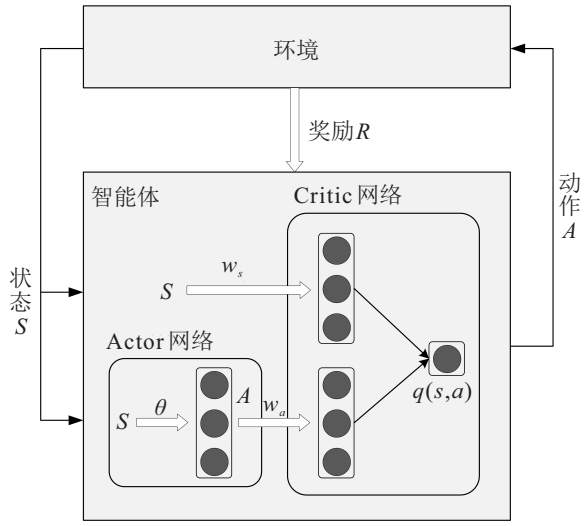
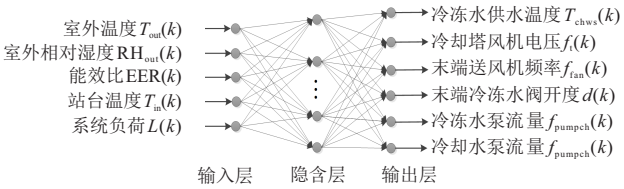
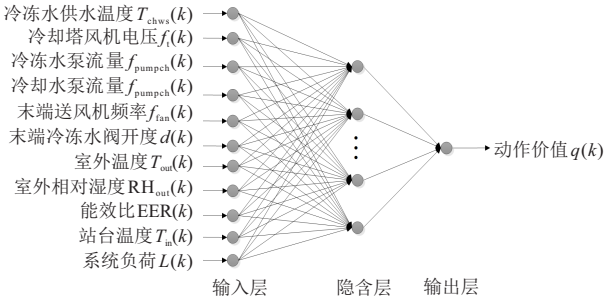


图 3　强化学习控制系统结构

在智能体训练之前还需确定智能体的网络结构, 包括 Actor 网络和 Critic 网络. 如图 3 所示, Actor 网络以系统状态为输入, 控制动作为输出, Critic 网络以系统状态和动作作为输入, 动作价值函数 q 为输出. 设置智能体每一个 Actor 网络和 Critic 网络均由 3 层全连接层网络构成, 具体的网络结构分别如图 4(a) 和 (b) 所示.



(a) Actor 网络示意图



(b) Critic 网络示意图

图 4　智能体的网络示意图

4　仿真实验

4.1　智能体训练

为实现所提出的改进 DDPG 算法, 使用 Pycharm 软件, 基于 Tensorflow 框架, 根据算法 1 编写算法程序进行仿真实验, 具体流程如图 5 所示.

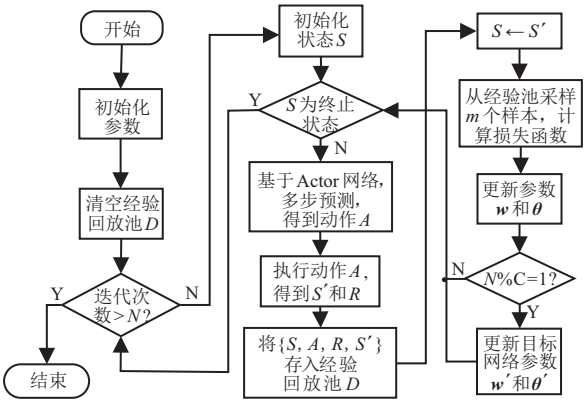
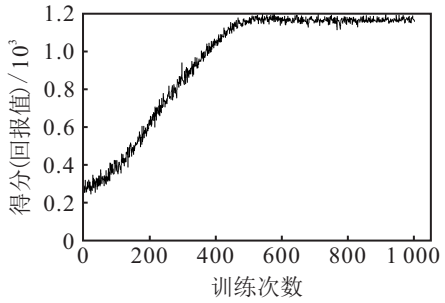


图 5　智能体训练流程

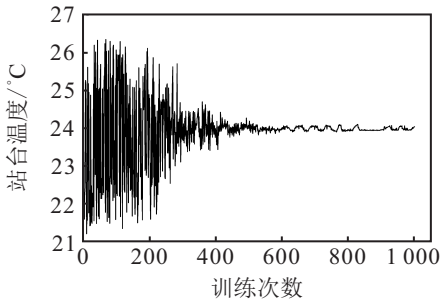
4.2　训练结果

图 6(a) 给出基于多步预测的 DDPG 算法训练 1 000 次过程中的得分 (总奖励值). 可以看出, 在训练过程中, 每次的奖励值是有波动的, 造成这种现象的原因主要有两个, 一是每次训练的初始环境不同, 二是算法为每次策略探索添加了随机噪声. 但是, 从整体奖励值的变化趋势看, 在训练过程中, 总奖励值呈稳步上升的趋势, 并在大约第 500 次训练后达到饱和值, 总奖励值接近 1 200, 表明智能体已经训练完成. 训练过程中的站台温度变化情况如图 6(b) 所示, 可以看出, 在第 500 次训练之前, 温度波动较大, 智能

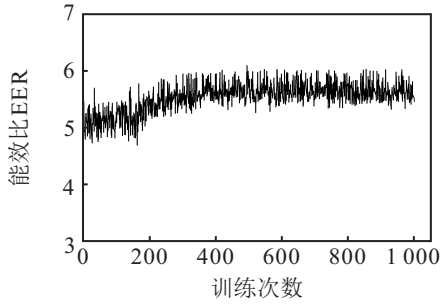
体在不断探索,寻求更大的奖励,之后温度趋于稳定,稳定在设定值24°C左右. 训练过程中系统能效比的变化情况如图6(c)所示,可以看出,在整个训练过程中,EER一直在不断探索以获得较大值,并最终能达到接近6.



(a) 训练得分



(b) 温度变化曲线



(c) EER变化曲线

图6 训练过程中变化曲线

在智能体训练过程中,系统各控制变量均存在探索过程并最终达到稳定值. 在系统实际运行时,各控制参数存在阈值约束,其限定范围如下:

冷冻水供水温度: $7.2\,^{\circ}\mathrm{C} \leqslant T_{\mathrm{chws}} \leqslant 16\,^{\circ}\mathrm{C}$;

冷却塔风机电压: $5\ \mathrm{V} \leqslant f_{\mathrm{t}} \leqslant 8.9\ \mathrm{V}$;

冷冻水泵流量: $70\ \mathrm{m^3/h} \leqslant f_{\mathrm{pumpch}} \leqslant 192\ \mathrm{m^3/h}$;

冷却水泵流量: $90\ \mathrm{m^3/h} \leqslant f_{\mathrm{pumpc}} \leqslant 250\ \mathrm{m^3/h}$;

末端风机控制信号: $9\,\% \leqslant f_{\mathrm{fan}} \leqslant 80\,\%$;

冷冻水阀开度: $23\,\% \leqslant d \leqslant 74\,\%$.

在智能体训练结束后,冷却塔风机控制电压、冷冻水泵流量和冷却水泵流量均稳定在最大值处,而冷冻水供水温度稳定在10 °C左右,末端风机控制信号稳定在约40 %,表明冷水机组和末端风机是耗能的主要设备,整个系统的节能主要就是控制这两个设备进行优化的. 冷冻水阀开度稳定在约74 %处,通过其可以实现系统风系统和水系统的协调控制.

由图6(a)可见,在约500次训练后,智能体得分已接近饱和值,因此后几百次的训练是多余的. 为了确定智能体训练完成的确切次数,避免人为设置训练次数过多,影响算法效率,每次训练结束后增加一个判断条件. 通过观察图6(a) 1 000次训练结果发现,在回报值达到饱和之前,相隔100次的两次训练回报值之差都超过100,达到饱和之后,该差值低于50,故将第i次的判断条件设置为计算第i次与第$i-100$次的回报值之差,若连续3次该差值均小于50,则判定智能体训练结束. 图7为应用该终止条件后的训练结果,可以看出,在第530次时训练结束,智能体得分达到饱和值1 169.2,这大大缩减了智能体的训练时间,并取得了与之前同样的控制效果.

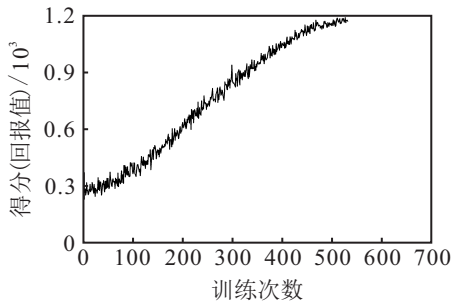


图7 设置终止条件后的训练过程

4.3 改进算法与传统DDPG算法性能比较

传统DDPG算法与改进算法的训练得分过程比较如图8所示,当智能体训练次数达到616次时,传统算法才训练结束,相比于改进的基于多步预测的DDPG算法增加了86次,可见所提出的改进算法提升了系统的学习效率. 除减少了智能体的训练次数外,二者在训练完成时对系统状态温度和EER取得了同样的控制效果.

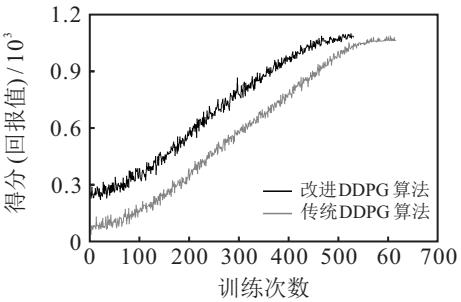


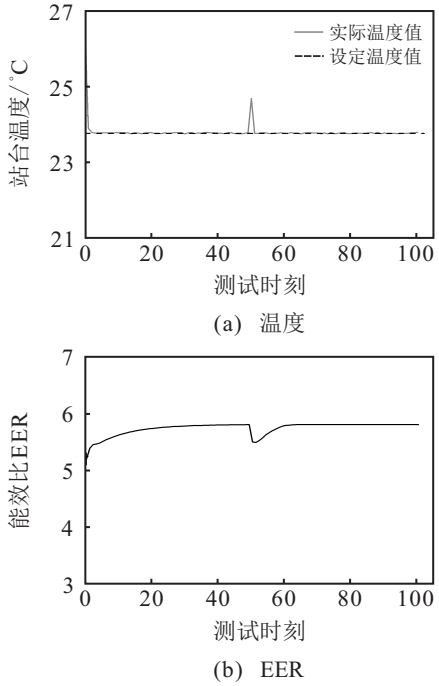
图8 传统算法与改进算法训练过程比较

4.4 测试结果

为了评价所提出基于多步预测的DDPG算法的性能,将训练完成的智能体用于空调系统控制仿真实验,并在第50个测试时刻人为加入干扰,以观察地铁

站台温度和能效比的输出结果.

图9(a)为系统经过100个测试时刻后站台温度的变化曲线. 可以看出,在前50个时刻,温度能够从初始值约27 °C很快到达设定值24 °C,平均绝对误差为0.013 7 °C,且过程较为平稳,控制效果较佳. 在第50个测试时刻,假设地铁站内客流量增大,温度从24 °C升到25 °C,可以看出,控制系统同样能够使温度快速跟踪到设定值,表明该控制策略能够使系统根据站内实时负荷需求调整设备参数,满足系统实时控制的需求. 图9(b)为系统能效比EER的测试结果,EER最大值能达到接近6左右,且较为稳定,在第50个测试时刻后,EER发生变化是因为此时站台温度升高,系统在调整各能耗设备控制参数. 通过计算,在应用所提出的策略后,EER平均值约为5.756 6,而原地铁站空调系统采用PID控制,且风系统和水系统分开控制,其实际运行EER平均值为4.882 3,从这个角度而言,该强化学习控制方案可以节能约17.908 %.



(a) 温度

(b) EER

图9　测试曲线

5 结　论

为解决地铁站空调系统能耗大、能源利用效率不高的问题,本文提出了一种多步预测改进DDPG算法,实现了基于强化学习的空调系统节能控制. 通过将多步预测与DDPG算法相结合,提高了算法的效率. 首先,采用神经网络建立系统模型,作为离线训练智能体的模拟环境,减少了强化学习智能体的训练时间;然后,将地铁站空调水系统与大系统视为一个整体设计智能体框架,在此基础上基于Pycharm软件和Tensorflow框架编写算法程序,训练智能体,在保证地铁站台温度要求的前提下,使整个系统的能效比最大化,为了避免不必要的训练,设置智能体训练终止条件,有效减少了训练时间;最后,基于武汉某地铁站的实测运行数据进行仿真实验,结果表明,所提出的基于多步预测的DDPG算法与传统DDPG算法相比,智能体训练次数减少了86次,提升了系统的计算效率,在将训练好的智能体应用于地铁站空调系统控制时,该控制策略具有较好的温度跟踪性能,且在保证温度效果的前提下,与目前实际系统相比,能源节省约17.908 %. 今后的研究工作将进一步对算法进行改进,以提升智能体的收敛速度.

参考文献(References)

[1] 曾逸婷, 赵蕾. 地铁车站环境热舒适与通风空调系统节能策略研究进展[J]. 铁道标准设计, 2019, 63(3): 178-183.
(Zeng Y T, Zhao L. Research on thermal comfort and energy saving strategies of ventilation and air-conditioning system in underground subway stations[J]. Railway Standard Design, 2019, 63(3): 178-183.)

[2] Maddalena E T, Lian Y Z, Jones C N. Data-driven methods for building control — A review and promising future directions[J]. Control Engineering Practice, 2020, 95: 104211.

[3] Sutton R S, Barto A G. Reinforcement learning[J]. A Bradford Book, 1998, 15(7): 665-685.

[4] 戴小燕, 张映波, 杲靖, 等. 基于人工智能的节能控制物联网云平台的设计与实现[J]. 电气应用, 2019, 38(11): 97-104.
(Dai X Y, Zhang Y B, Gao J, et al. Design and implementation of artificial intelligence-based energy-saving control system on IoT cloud platform[J]. Electrotechnical Application, 2019, 38(11): 97-104.)

[5] 闫军威, 黄琪, 周璇. 基于Double-DQN的中央空调系统节能优化运行[J]. 华南理工大学学报: 自然科学版, 2019, 47(1): 135-144.
(Yan J W, Huang Q, Zhou X. Energy-saving optimization operation of central air-conditioning system based on double-DQN algorithm[J]. Journal of South China University of Technology: Natural Science Edition, 2019, 47(1): 135-144.)

[6] Yuan X L, Pan Y Q, Yang J R, et al. Study on the application of reinforcement learning in the operation optimization of HVAC system[J]. Building Simulation, 2021, 14(1): 75-87.

[7] Dalamagkidis K, Kolokotsa D, Kalaitzakis K, et al. Reinforcement learning for energy conservation and comfort in buildings[J]. Building and Environment, 2007, 42(7): 2686-2698.

[8] Afram A, Janabi-Sharifi F. Review of modeling methods for HVAC systems[J]. Applied Thermal Engineering,

2014, 67(1/2): 507-519.

[9] 赵静, 王弦, 王奔, 等. 基于神经网络的多类别目标识别[J]. 控制与决策, 2020, 35(8): 2037-2041.
(Zhao J, Wang X, Wang B, et al. Multi-category target recognition based on neural network[J]. Control and Decision, 2020, 35(8): 2037-2041.)

[10] 刘三阳, 吴德. 模糊聚类光滑支持向量机[J]. 控制与决策, 2017, 32(3): 547-551.
(Liu S Y, Wu D. Fuzzy clustering smooth support vector machine[J]. Control and Decision, 2017, 32(3): 547-551.)

[11] Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: Application to the Cox regression model[J]. Statistics in Medicine, 1992, 11(16): 2093-2109.

[12] Anwer N, Ballu A, Mathieu L. The skin model, a comprehensive geometric model for engineering design[J]. CIRP Annals, 2013, 62(1): 143-146.

[13] Zlatanovi I, Gligorevi K, Ivanovi S, et al. Energy-saving estimation model for hypermarket HVAC systems applications[J]. Energy and Buildings, 2011, 43(12): 3353-3359.

[14] Afram A, Janabi-Sharifi F, Fung A S, et al. Artificial neural network (ANN) based model predictive control (MPC) and optimization of HVAC systems: A state of the art review and case study of a residential HVAC system[J]. Energy and Buildings, 2017, 141: 96-113.

[15] Kusiak A, Xu G L, Tang F. Optimization of an HVAC system with a strength multi-objective particle-swarm algorithm[J]. Energy, 2011, 36(10): 5935-5943.

[16] Guan B W, Liu X H, Zhang T, et al. Energy consumption of subway stations in China: Data and influencing factors[J]. Sustainable Cities and Society, 2018, 43: 451-461.

[17] 林晓伟, 王侠. 地铁通风空调系统的优化控制[J]. 城市轨道交通研究, 2012, 15(11): 100-104.
(Lin X W, Wang X. Optimum control of metro air conditioning[J]. Urban Mass Transit, 2012, 15(11): 100-104.)

[18] 王晓保, 杨欣, 袁立新. 地铁车站空调实施风水联动控制技术节能效果分析[J]. 上海节能, 2013(7): 10-14.
(Wang X B, Yang X, Yuan L X. Analysis of energy-saving effect in subway station's air-conditioning implementingair system & water system combined-control technology[J]. Shanghai Energy Conservation, 2013(7): 10-14.)

[19] Nishida T. Data transformation and normalization[J]. Rinsho Byori the Japanese Journal of Clinical Pathology, 2010, 58(10): 990-997.

[20] Jing H U, Amp F R. Predictive modeling of surface skewness and kurtosis based on BP neural network[J]. Surface Technology, 2017, 46(2): 235-239.

[21] Jahirul M I, Senadeera W, Brooks P, et al. An artificial neutral network (ANN) model for predicting biodiesel kinetic viscosity as a function of temperature and chemical compositions[J]. Modsim International Congress on Modelling & Simulation, 2013: 1561-1567.

[22] 李益兵, 宋东林, 王磊. 基于混合PSO-Adam神经网络的外协供应商评价决策模型[J]. 控制与决策, 2018, 33(12): 2142-2152.
(Li Y B, Song D L, Wang L. Based on hybrid PSO-Adam neural networks decision making model for outsourcing supplier evaluation[J]. Control and Decision, 2018, 33(12): 2142-2152.)

[23] Vázquez-Canteli J R, Nagy Z. Reinforcement learning for demand response: A review of algorithms and modeling techniques[J]. Applied Energy, 2019, 235: 1072-1089.

[24] 康铭鑫, 李长平, 刘腾飞. 基于观测器的发动机转矩跟踪模型预测控制[J]. 控制与决策, 2020, 35(4): 791-798.
(Kang M X, Li C P, Liu T F. Observer based model predictive torque tracking control for gasoline engines[J]. Control and Decision, 2020, 35(4): 791-798.)

作者简介

焦焕炎(1996−), 男, 硕士生, 从事建筑节能与安全监控理论及工程的研究, E-mail: jiaohuanyan@foxmail.com;

冯浩东(1995−), 男, 硕士生, 从事建筑节能与安全监控理论及工程的研究, E-mail: fenghaodong@foxmail.com;

魏东(1968−), 女, 教授, 博士, 从事建筑节能与安全监控理论及工程等研究, E-mail: weidong@bucea.edu.cn;

冉义兵(1988−), 男, 硕士生, 从事大数据与人工智能、建筑节能与安全监控理论及工程的研究, E-mail: ranyibing@bucea.edu.cn;

胡朝文(1988−), 男, 硕士生, 从事建筑节能与安全监控理论及工程的研究, E-mail: 1256788751@qq.com.

(责任编辑：郑晓蕾)