计算机科学
COMPUTER SCIENCE

一种基于深度学习的供热策略优化方法

李鹏，易修文，齐德康，段哲文，李天瑞

引用本文

李鹏，易修文，齐德康，段哲文，李天瑞. 一种基于深度学习的供热策略优化方法[J]. 计算机科学, 2022, 49(4): 263-268.

LI Peng, YI Xiu-wen, QI De-kang, DUAN Zhe-wen, LI Tian-rui. Heating Strategy Optimization Method Based on Deep Learning[J]. Computer Science, 2022, 49(4): 263-268.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210300155

一种基于深度学习的供热策略优化方法

李　鹏[1,2]　易修文[2]　齐德康[1,2]　段哲文[2,3]　李天瑞[1]
1 西南交通大学计算机与人工智能学院　成都 611756
2 北京京东智能城市大数据研究院　北京 100176
3 西安电子科技大学计算机科学与技术学院　西安 710071
(lipengsx@my.swjtu.edu.cn)

摘　要　在中国北方,冬季楼宇集中供暖采用的策略通常为气候补偿器,但是该策略严重依赖人工经验,调节相对粗放,如何优化供热控制策略对于保持楼宇室温的稳定舒适十分重要。对此,提出了一种基于深度学习的供热策略优化方法,通过学习历史真实数据信息从而对原始控制策略进行优化。首先以学习室内温度变化的热力学规律为目标,提出了一种深度多时差分网络MTDN(Multiple Time Difference Network)来对下一时刻的室温进行预测,该网络不仅准确率高,而且符合物理规律;然后将MTDN当成模拟器,以表征人体热反应的评价指标作为相关奖励项,使用基于最大熵强化学习思想的SAC(Soft Actor Critic)算法作为策略优化器与之交互训练,从而学习到一个稳定优秀的供热控制策略;最后基于天津某个换热站的真实数据,设计相关实验分别对模拟器预测能力和策略优化器策略控制能力进行评估。验证得出:相比其他类型的预测模拟器,该模拟器不仅预测精度高,并且符合物理规律;同时,相比原始策略,该策略优化器所学的策略在随机采样的多个时段内均可以保证室内温度更加稳定舒适。

关键词:集中供暖;供热优化;深度学习;深度强化学习;城市计算
中图法分类号　TP399

Heating Strategy Optimization Method Based on Deep Learning

LI Peng[1,2],YI Xiu-wen[2],QI De-kang[1,2],DUAN Zhe-wen[2,3] and LI Tian-rui[1]
1 School of Computing and Artificial Intelligence,Southwest Jiaotong University,Chengdu 611756,China
2 JD Intelligent Cities Research,Beijing 100176,China
3 School of Computer Science and Technology,Xidian University,Xi'an 710071,China

Abstract　Typically,the strategy of central heating for buildings in winter is climate compensator. However,this strategy heavily relies on manual experience with a relatively simple regulation. Therefore,how to optimize the heating control strategy is very important to keep the indoor temperature stable and comfortable. For this task,this paper proposes a heating strategy optimization method based on deep learning and deep reinforcement learning,which can optimize the original control strategy based on real historical data. The paper first develops a deep MTDN (Multiple Time Difference Network) as the simulator to predict the next time slot's room temperature. By learning the thermodynamic law of indoor temperature change,the network has high accuracy and confirms the physical laws. After that,the SAC (Soft Actor-Critic) algorithm based on maximum entropy reinforcement learning is employed as the strategy optimizer to interact with the simulator. Here,we use the evaluation index of the human body's thermal response as the reward to train and optimize the heating control strategy. Based on the real data of a heat exchange station in Tianjin,we evaluate the predictive ability of the simulator and the control ability of the strategy optimizer,respectively. The results verify that,compared with other types of prediction simulators,this simulator not only has high prediction accuracy but also conforms to physical laws. At the same time,compared with the original strategy,the strategy learned by the strategy optimizer can ensure that the indoor temperature is more stable and comfortable in multiple time periods of random sampling.

Keywords　Central heating,Heating optimization,Deep learning,Deep reinforcement learning,Urban computing

到稿日期:2021-03-15　返修日期:2021-07-25
基金项目:国家重点研发计划(2019YFB2101801);国家自然科学基金面上项目(61773324)
This work was supported by the National Key R & D Program of China(2019YFB2101801) and National Natural Science Foundation of China (61773324).
通信作者:易修文(yixiuwen@jd.com)

1 引言

在中国北方，冬季会使用集中供热系统对楼宇建筑进行供暖。如图 1 所示，在该系统中，热源厂生产的热水通过一次管网送入换热站，由于一、二次管网不互通，所以两者通过换热站进行热量传递，二次管网中的水升温并输送到热用户室内的散热器，对室内空气进行加热。供热系统优秀与否，核心标准在于其控制策略是否可以在气候条件变化的情况下，通过调整换热站二次供水温度，使室内温度稳定在合理舒适的范围，满足用户的生产生活需求。

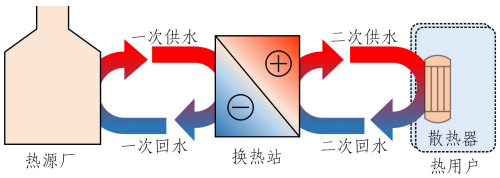

图 1 集中供热系统

Fig. 1 Central heating system

目前多数供热系统的供热策略采用传统气候补偿器[1]。它可以根据室外温度变化情况对二次供水温度进行调控，从而补偿室外温度对室温的影响。但这种做法存在严重依赖人工经验、调节粗放、反应慢、缺少室温反馈等诸多缺点。针对气候补偿器的优化主要有 3 种类型：1）使用物理模拟软件 EnergyPlus[2] 等搭建供热系统，借助物理引擎模拟器，间接优化原始策略。这种方式不仅要对真实环境参数进行详细调研，还需要具备大量的建筑暖通领域专业知识。2）直接对供热策略进行优化，例如基于 PID（Proportional，Integral，Differential）[3] 等方法直接和真实环境做长期交互，对原始控制策略进行优化，这种方法调节粗放，参数整定难，不能很好地处理非线性问题，策略异常时会直接对真实环境产生不良影响。3）使用回归类预测模型预测相关工况，借助预测模型对供热策略做优化。这种方式得不到预期的系数组合，会造成其泛化能力差，不符合物理规律。

深度学习[4] 和强化学习[5] 作为前沿的人工智能方法，也可以用于供热策略优化。深度学习是一种非线性的、能够准确地拟合数据内部逻辑的机器学习技术，近年来，由于大数据和硬件算力的加持，其在模拟预测领域展现出强劲的能力。而强化学习是一种从环境状态映射到动作的学习方式，其目标是使智能体在和环境交互的过程中获取最大的奖励。深度强化学习使用深度学习强大的表征能力赋能强化学习，以处理复杂决策类任务。近年来，深度强化学习在游戏、机器人、推荐系统等方面都取得了很好的效果。由于 PPO[6]，DDPG[7]（Deep Deterministic Policy Gradient），SAC（Soft Actor-Critic）[8] 等连续控制类算法的提出，人们已将深度强化学习扩展到工业控制等多个实际场景中。

针对供热优化的难点，本文提出了一种基于深度学习的供热策略优化方案。首先以学习室内温度变化的热力学规律为目标，提出了一种深度多时差分网络 MTDN。通过使用共享权重的神经网络对多时差分数据集进行学习，从而预测各个因素对室内温度变化的影响。该方法不仅精确度高，还具备良好的泛化能力，符合物理规律。然后将上述模拟器作为实际供热场景的模拟环境，使用基于最大熵强化学习思想的深度强化学习算法 SAC 作为策略优化器与之交互，以建筑学领域人体热感舒适度指标 PMV（Predicted Mean Vote）[9] 作为相关奖励项，训练过程中加入动作熵值，使其学到了更加稳定的维持室温舒适的策略。最后基于天津某小区供热真实数据设计实验，分别对模拟器的精准度、泛化能力、策略优化器学习到的供热策略调控能力进行验证，证明本文方案有着良好的效果。

2 相关工作

针对供热系统建模及控制策略优化的问题，国内外学者展开过大量的研究。Fazlollahi 等[10] 提出了一种多目标、多周期的区域能源系统设计与运行策略优化模型。Li 等[11] 使用回归分析的方法建立了供热负荷的多元回归模型。Bai 等[12] 基于 PID 控制方法对热网回水温度控制系统进行优化调节，以解决传统供热策略反应慢、调整粗等问题。Wu 等[13] 使用热力学流体系统分析软件 Flowmaster 对供热系统进行建模，并基于该模型提出了优化调节方案。

除了上述传统控制领域优化方式，国内外诸多学者也将深度学习和强化学习应用于供热策略优化领域。Li 等[14] 基于深度学习，使用 LSTM 神经网络对热力站进行建模，使用 DDPG 实现对热力站热量的分配。Zhang[15] 提出了一种基于模型的强化学习方法，使用神经网络学习系统动力学，采用 MPC 模型预测控制方法对 HVAC 空调系统策略进行优化。Zhang 等[16] 利用深度强化学习算法 A3C 和建筑物 EnergyPlus 仿真模型交互，实现了对 HVAC 供热系统策略的优化。Wei 等[17] 开发了一种数据驱动的方法，以深度强化学习算法 DQN 为基础，有效学得 HVAC 系统控制策略。

在这些研究中，传统模拟器 EnergyPlus 等需要个人对建筑结构材料等有细致了解；回归类模拟器存在泛化性不够的局限；相比 SAC，采用的 A3C，DDPG 等策略优化方法并不能探索出并获取到所有的最优控制策略，难以保证策略的稳定性。本文基于热力学定律构建多时差分网络对室温变化过程进行模拟，在预测精度较高的同时保证模型符合物理规律，同时使用基于最大熵强化学习思想的 SAC 算法作为策略优化器，保证所学策略的稳定性，满足供热场景。

3 问题定义

我们使用 i_t 表示 t 时刻室内温度，h_t 表示 t 时刻换热站二次供水温度，o_t 表示 t 时刻室外温度，w_t 表示 t 时刻风速，l_t 表示 t 时刻光照强度。供热策略定义为：为使室内温度 i 可以长期处于舒适状态，在不同时刻环境下，调控供热温度 h 的策略。

本文目标定义如下。

目标 1：给定时间长度 T，使用数据序列 $\{i_t, h_t, o_t, w_t, l_t \mid t \in \{1, \cdots, T\}\}$ 对下一时刻室内温度 i_{t+1} 进行预测，使该预测模型可以作为目标 2 训练交互的环境 E，E 可以根据供热温度等因素的变化预测室内温度的变化情况。

目标 2：构建策略 π 与环境 E 交互，π 通过调整供热温度

h 获得 E 预测的室内温度 i。给定奖励 $r_t=f(\{i_m \mid m\in\{m,\cdots,t\}\})$，表示时刻 t 奖励 r_t 和前 m 个时刻室内温度有关，使得从当前时刻 t_0 开始，未来 k 个时刻总奖励 $R=\sum_{t=t_0}^{t_0+k} r_t$ 最大。

4　模拟器和策略优化器

4.1　深度多时差分网络 MTDN 模拟器

为优化供热策略，需要构建一个真实环境的模拟器，从而基于历史时刻数据准确预测下一时刻的室内温度。单纯使用真值数据构建的回归类模型虽然预测精度高，但泛化能力差，纯粹的热力学模型虽具备泛化能力，但预测精度太差。基于热力学第一定律，即物体内能的增加等于物体吸收的热量和对其做的功的总和，同时，考虑到空气比热容非常稳定，因此室内温度的变化仅取决于其所含热量的变化。使用 $x=\{h,o,w,l\}$ 表示除室温度外的其他数据，可得在任意两个时刻 t_1，t_2 下，其他因素的变化引起的室温变化，即 $\Delta x_{(t_1-1,t_2-1)}\rightarrow\Delta i_{(t_1,t_2)}$，该过程符合热力学规律。

为了刻画室内温度变化的热力学规律，本文提出了一种多时差分网络 MTDN，结构如图 2 所示。该网络可以分为 3 部分。首先我们使用历史数据构建所需特征标签，特征包括当前时刻特征 x_t、相邻时刻特征 x_{t-1}，以及随机 s 时刻之前的特征 x_{t-s}；标签分别为相邻时刻室内温度差分 $\Delta i_{(t,t+1)}$ 和随机时刻室内温度差分 $\Delta i_{(t-s+1,t+1)}$。

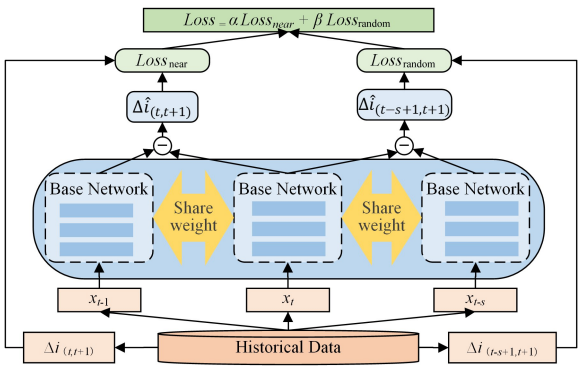

图 2　深度多时差分网络 MTDN

Fig. 2　Multiple time difference network

然后使用 3 个全连接神经网络作为基础网络，输入分别为 3 个时刻的真值特征 x_t，x_{t-1}，x_{t-s}，输出分别为 3 个时刻的室内温度预测值 $\hat{i}_{t+1}$，$\hat{i}_t$，$\hat{i}_{t-s+1}$，这 3 个网络不仅网络结构及初始网络参数相同，并且在训练过程中共享神经网络权重。

最后是 $Loss$ 输出部分。正向传播过程中，对当前时刻基础网络输出和其他两个基础网络输出分别求差，表示为 $\Delta\hat{i}_{(t,t+1)}$ 和 $\Delta\hat{i}_{(t-s+1,t+1)}$。式(1)$Loss_{\text{near}}$ 表示相邻时刻差分真值和预测值之间的均方误差，式(2)$Loss_{\text{random}}$ 表示随机时刻差分真值和预测值之间的均方误差，然后加权求和生成总 $Loss$，如式(3)所示。实际训练过程中，这两部分权重都设为 1。

$$Loss_{\text{near}}=\frac{1}{m}\sum_{t=1}^{m}(\Delta i_{(t,t+1)}-\Delta\hat{i}_{(t,t+1)})^2 \tag{1}$$

$$Loss_{\text{random}}=\frac{1}{m}\sum_{t=1}^{m}(\Delta i_{(t-s+1,t+1)}-\Delta\hat{i}_{(t-s+1,t+1)})^2 \tag{2}$$

$$Loss=\alpha Loss_{\text{near}}+\beta Loss_{\text{random}} \tag{3}$$

本文模型将多个时刻的真值数据集送入 3 个共享权重的基础网络中，各自独立进行正向传播，使该模型不会损失真值数据的内在信息，从而进行高精度预测；同时因为多时差分数据集参与反向传播，使得模型整体目标是对室内温度的热力学规律进行学习，具有优秀的泛化能力，符合物理规律。

实际模拟预测时，我们仅使用 MTDN 内 3 个基础网络中的 2 个，即以当前时刻特征 x_t 和上一时刻特征 x_{t-1} 作为输入预测室内温度差值 $\Delta i_{(t,t+1)}$，并加上当前时刻室内温度 i_t 生成下一时刻室内温度 i_{t+1}。使用 OpenAI-Gym[18] 框架将 MTDN 预测模型包装成室内温度模拟器，该模拟器可以稳定准确地模拟出供热温度等因素变化时室内温度的变化情况，最终将其用作强化学习策略优化器的训练交互环境。

4.2　SAC 算法策略优化器

深度强化学习集成了深度学习和强化学习的优点，处理连续动作空间的控制任务效果非常优秀。SAC 算法是该领域基于最大熵思想的一种算法，相比于其他连续控制优化算法，SAC 可以对所有的最优路径进行覆盖，稳定性更强，可直接应用于真实环境，更加符合供热任务。因此本文以 SAC 算法作为策略优化器。

如图 3 所示，交互过程中，环境给出的 state $s=\{i,o,w,l\}$ 分别表示室内温度、外界温度、风速、光照强度。Done 表示该回合是否结束。SAC 策略优化器输出 action $a=\{\Delta h\}$ 表示策略给出的室温变化差值。Reward 引入了预测平均评价指标 PMV，PMV 是一个基于人体热平衡公式以及心理生理学主观热感觉的全面评价指标，一般情况下室温 PMV 指数处于[−0.5，0.5]表示舒适[9]。Reward 分为两部分，一部分专注于当前时刻室内舒适度，如式(4)所示，其中 p 表示 PMV 计算函数，我们这里令单个时间 PMV 指数在[−0.1，0.1]之间，其 reward 为 10，其他时刻 PMV 指数越偏离 0，reward 越低；另一部分专注于室温多个时刻内变化幅度，如式(5)所示，取最近 n 个时刻室温序列，求其方差，方差越小，则 reward 越大，具体实验中 $n=5$。这两部分奖励加权求和构成总体奖励，如式(6)所示，具体实验中权重都设置为 1。

$$r_{pmv}=\begin{cases}10, & -0.1<p(temp)<0.1\\ 10-50^{|p(i_t)|}, & p(i_t)<-0.1 \text{ or } p(i_t)>0.1\end{cases} \tag{4}$$

$$r_{\text{var}}=-\frac{\sum_{t=1}^{n}(i_t-\bar{i})^2}{n-1}\times 10 \tag{5}$$

$$r=\alpha r_{pmv}+\beta r_{\text{var}} \tag{6}$$

SAC 策略优化器基于 Actor-Critic 结构，如图 3 所示，其中有 1 个 Actor 网络和 4 个 Critic 网络，还有 1 个经验缓冲池(experience replay buffer)用来存储和模拟器交互过程中的历史数据，以从中抽样训练，打破样本之间的相关性。Actor 为 Policy 网络，负责和模拟器交互，输入为 state，输出为 action，表示为 $\pi_\varphi(a_t\mid s_t)$，π 为策略 Policy，ϕ 为网络参数。Critic 网络为 Q 值网络，输入为(state，action)，输出为 Q 值，表示为 $Q_\theta(a_t,s_t)$，θ 为网络参数，用于对 Actor 网络输出的 action 做评价。4 个 Q 网络分为 Q_1，target Q_1，Q_2 以及 target Q_2，使用两个估值网络求 $\min(Q_{\theta_1}(s,a),Q_{\theta_2}(s,a))$ 可解决值函数高估

问题。target 网络可用于 TD-error 更新。H 为策略 π 的熵值，计算式为 $-E\log\pi(a_{t+1}|s_{t+1})$。

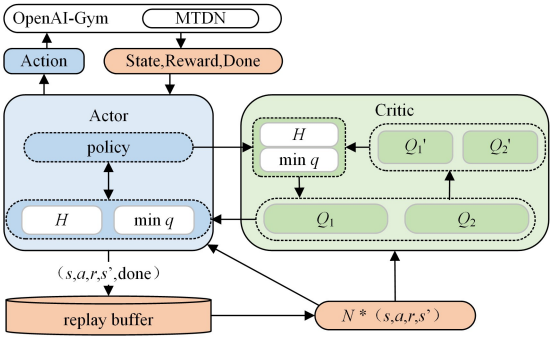

图 3 Soft Actor-Critic 策略优化器

Fig. 3 Soft Actor-Critic strategy optimizer

SAC 策略优化器训练过程中，Q 网络训练目标函数表示为式(7)，其中 $\bar{\theta}$ 表示 target 网络的参数，该数据通过 θ 软更新得到；V 为状态价值网络，可以使用 Q 网络表示为式(8)，$(s_t,a_t,s_{t+1})\sim D$，$a_{t+1}\sim\pi_\phi$，D 表示经验缓冲池。Policy 网络训练时的目标函数为式(9)，其中 $\boldsymbol{\varepsilon}_t$ 是一个高斯噪音向量。

$$J_Q(\theta)=E_{s_t,a_t,s_{t+1}}\left[\frac{1}{2}\left(Q_\theta(s_t,a_t)-\left(r(s_t,a_t)+\gamma V_{\bar{\theta}}(s_{t+1})\right)\right)^2\right] \tag{7}$$

$$V_{\bar{\theta}}(s_{t+1})=E_{s_{t+1},a_{t+1}}\left[Q_{\bar{\theta}}(s_{t+1},a_{t+1})-\alpha\log\pi_\phi(a_{t+1}|s_{t+1})\right] \tag{8}$$

$$J_\pi(\phi)=E_{s_t,\boldsymbol{\varepsilon}_t}\left[\alpha\log\pi_\phi(f_\phi(\boldsymbol{\varepsilon}_t;s_t)|s_t)-Q_\theta(s_t,f_\phi(\boldsymbol{\varepsilon}_t;s_t))\right] \tag{9}$$

由于 SAC 算法学得的策略 π 不仅希望累计 reward 期望最大，还需要 policy 每次输出的 action 熵最大，因此在寻找最优策略的同时还需保持探索能力，能够覆盖到所有有用的解决方法，不会过早陷入局部最优解，使学到的策略更具稳定性。同时，SAC 更强的探索能力也使得其在追求多模态奖励时，可以更好地找到最优策略。

5 实验设计

5.1 参数及实验说明

为验证整体方案的有效性，我们分别对 MTDN 模拟器和 SAC 策略优化器进行了评估，实验数据来自天津某换热站及其对应小区。室温数据来自一部分房屋内部署的室温传感器，我们对其做了平均，代表换热站对应小区的平均室温；其他数据分别来自于换热站传感器、室外温度传感器和气象传感器，收集到的数据包括供热温度、室外温度、风速、光照强度。数据每小时采集一条，范围涵盖 2018-11-05—2019-03-15 整个供暖季。

实验的服务器配置为：NVIDIA Tesla V100 GPU，Intel Xeon CPU E5-2650 v4 CPU，16 GB 运行内存，Python 3.7，PyTorch 1.7.1。PMV 使用专用库 pythermalcomfort[19] 计算。

模拟器内的基础神经网络共有 3 个隐藏层，每层含 64 个神经元，采用 ReLU 作为激活函数。训练过程中学习率为 0.01，网络优化器为 Adam，数据采用 standard 标准化。将每个月前 3 周数据作为训练集，最后 1 周数据作为测试集，训练集中取 1/10 的数据用于训练过程早停。

策略优化器内 actor 网络、Critic 网络中间均有两个隐藏层网络，各包含 128 个神经元，采用 ReLU 作为激活函数。在训练过程中，actor 网络学习率为 0.0005，critic 网络学习率为 0.001，温度系数 α 初始值为 0.01，其学习率为 0.001，γ 为 0.98，经验池大小为 2000，用于 target Q 网络软更新的平滑系数为 0.01。Actor 模型在具体实验中的输出是一个高斯分布，从该分布采样后，使用 tanh 函数将采样值裁剪至[−1,1]范围，对其乘以 2 作为输出动作，表示供温调整值 Δh。

5.2 MTDN 模拟器评估

本文通过和历史数据做对比来计算不同模型的预测精度。在保持其他特征不变的情况下，仅改变供热温度，通过观察室内温度变化的幅度是否符合物理规律，判断模型泛化能力。具体实验中，供热温度变化的最小单位为±0.1℃，最高变化范围为±2℃。参与对照的模型如下。

(1)可解释性模型：以当前时刻 h_t，o_t，w_t，l_t 作为输入特征，以下一时刻室温数据 i_{t+1} 为标签进行建模。之所以称之为可解释性模型，是因为其输入特征没有引入上一时刻室内温度，可以较好地刻画各因素对室温的影响关系。

(2)True Feature Model：使用当前时刻 h_t，o_t，w_t，l_t，i_t 作为输入，以下一时刻室温数据 i_{t+1} 为标签进行训练测试，相比可解释性模型，该模型引入了上一时刻的室内温度。

(3)Diff feature Model：使用差分特征数据 $\Delta x_{(t-s,t)}$ 作为输入特征，以下一时刻室内温度差值 $\Delta i_{(t-s+1,t+1)}$ 为标签进行建模测试。

表 1 为各个模型预测精度对比，MAE 和 MAPE 分别表示平均绝对误差和平均百分比误差。可以看出：可解释性模型预测精度不佳，因为其没有引入上一时刻室温数据，原始数据量较小使其无法完全刻画各因素对室温的影响关系；True Feature Model 引入了上一时刻室温数据，预测精度最高；Diff feature Model 仅使用多时差分特征作为训练数据，预测精度仍旧不高；MTDN model 同时使用真值和差分数据，相比 True Feature Model，其预测精度稍低，但相比其他模型，其精度更高。

表 1 各个模型的预测结果对比

Table 1 Comparison of prediction results of each model

Model	MAE	MAPE
True Feature Model	**0.1429**	**0.7122**
可解释性模型	0.8536	4.2486
Diff feature Model	0.4983	2.5227
MTDN	**0.1818**	**0.8905**

图 4 给出了各模型泛化能力的对比结果，横轴为在其他条件保持不变的情况下供热温度的变化量，纵轴为相对应室内温度的变化量。可以看出：可解释性模型可以较好地刻画建筑物热力学特点，具备良好的泛化性，符合物理规律；True Feature Model 泛化能力极差，不符合物理规律；Diff feature Model 有一定泛化能力，但是室温变化幅度有限；MTDN

model 具备优秀的泛化能力,符合物理规律,这是因为它以多时差分特征进行建模,学到了室内热力学规律。

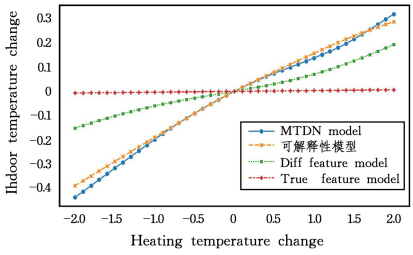

图 4　各模型泛化能力对比

Fig. 4　Comparison of extensive ability of each model

综上,MTDN 同时用到了真值特征和长短期差分特征,既没有损失原始数据内在信息,又刻画了建筑物热力学特性,完全可以作为该小区实际供热场景下的模拟器,用于策略优化器交互训练。

5.3　SAC 策略优化器评估

我们分两部分对 SAC 策略优化器进行评估。一是模型收敛效果,对比其和经典深度确定性策略 DDPG 算法的奖励值回报,其中 DDPG 除了 policy 网络输出为确定性 action 之外,其他参数以及网络结构均和 SAC 一致;二是模型策略优化效果,通过与历史室内温度作对比,评价其控制下模拟器输出的室内温度稳定性和舒适度,其中舒适度采用 $|p(i)|$ 作为评价标准,该值越接近 0 表明室内温度越舒适。

图 5 为 SAC 和 DDPG 策略返回奖励的对比,横轴表示训练回合数,纵轴表示训练过程测试模式下的回合总返回奖励。两个模型均训练 900 回合(episode),每回合最多包含 48 个 step,表示交互 48 h。网络在每回合训练之后进行更新,并以测试模式和环境交互 50 个回合,收集其奖励回报数据。可以看出,SAC 相比于 DDPG 震荡更小,收敛速度更快,整体更稳定,单个回合奖励更高,说明 SAC 策略优化器训练出来的策略更加稳定高效,符合供热控制场景。这两个策略最终都可以收敛,也侧面证明了模拟器非常稳定。

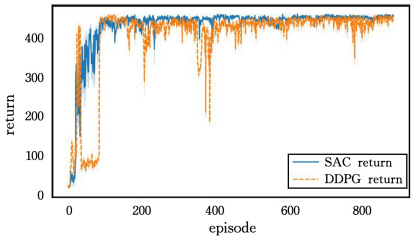

图 5　各算法奖励回报对比

Fig. 5　Comparison of return of each algorithm

使用训练好的策略优化器和模拟器交互,随机采样交互数据,然后对比其室温稳定性和室温舒适度。图 6 横轴表示回合交互步数,纵轴表示室内温度,model 标签数据表示在 SAC 训练好的策略控制下模拟器给出的室内温度,true 标签数据为原始策略下的室内温度,9 个子图表示随机采样的不同时刻的室温对比。可以看出,相比原始策略,优化后策略调控下的室内温度变化更稳定,室温不会出现大幅升高或降低的情况。

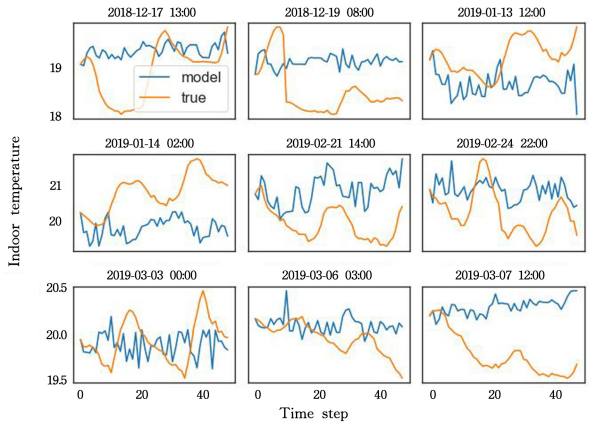

图 6　优化后策略和原始策略的室内温度对比

Fig. 6　Comparison of indoor temperature between optimized strategy and original strategy

图 7 中的 model_PMV 和 true_PMV 标签与图 6 中的 model 和 true 标签相对应,横轴表示回合交互步数,纵轴表示室内温度舒适度,9 个子图分别表示图 6 各个时刻下的室温舒适度。可以看出,优化后策略下的室温整体更加舒适。这表明 SAC 优化器学到的策略相比原始策略有良好的提升效果,可以更好地保持室内温度变化平稳,使其长期处于舒适状态。

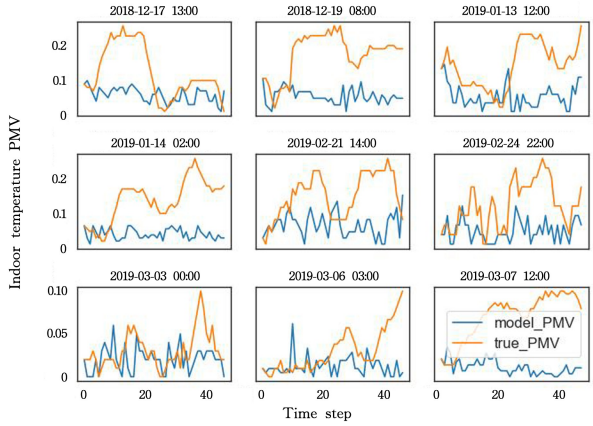

图 7　优化后策略和原始策略的室内温度舒适度对比

Fig. 7　Comparison of indoor temperature comfort between optimized strategy and original strategy

结束语　针对现有供热场景中换热站供热策略优化方式的诸多问题,本文提出了一种基于深度学习和深度强化学习的供热策略优化方案,准确地对室内温度变化的热力学规律进行了模拟仿真,生成了精准稳定的模拟器,并使用基于最大熵强化学习思想的策略器与模拟器进行交互训练,得到了更优的供热策略,可以更好地保持室内温度稳定舒适。在后续工作中,我们还将实现该方案的落地,在实测中发现问题并改进,并考虑多目标优化,在维持室温舒适的情况下同时降低能源消耗。

参 考 文 献

[1]　CHENG L. Application of climate compensator in heating system [J]. Building Science, 2010, 26(10): 42-46.

[2] CRAWLEY D B, LAWRIE L K, WINKELMANN F C, et al. EnergyPlus: creating a new-generation building energy simulation program[J]. Energy and buildings, 2001, 33(4): 319-331.

[3] LI Y, ANG K H, CHONG G C Y. PID control system analysis and design[J]. IEEE Control Systems Magazine, 2006, 26(1): 32-41.

[4] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.

[5] 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.

[6] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv: 1707.06347, 2017.

[7] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J]. arXiv: 1509.02971, 2015.

[8] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]// International Conference on Machine Learning. PMLR, 2018: 1861-1870.

[9] DEAR R D, BRAGER G. Developing an adaptive model of thermal comfort and preference[J]. Ashrae Trans, 1998, 104(1): 73-81.

[10] FAZLOLLAHI S, BECKER G, MARECHAL F. Multi-objectives, multi-period optimization of district energy systems: III. Distribution networks[J]. Computers & Chemical Engineering, 2014, 66(4): 82-97.

[11] LI S Q, JIANG Z J. Heating load forecasting model based on Neural Network[J]. District Heating, 2018, (4): 42-46.

[12] BAI H, WANG Y, FAN W Q, et al. Backwater Temperature Control System of Heat Network Based on PID[J]. District Heating, 2019, (3): 132-136.

[13] WU J X, ZHAO T, LIU L S, et al. Research on Heat-exchange Station Operation Based on Flowmaster Simulation[J]. District Heating, 2019, (4): 144-150.

[14] LI Q, HAN B C. Optimal Control of Primary Side of Thermal Power Station Based on Deep Deterministic Policy Gradient[J]. Science Technology and Engineering, 2019, 19(29): 193-200.

[15] ZHANG C, KUPPANNAGARI S R, KANNAN R, et al. Building HVAC scheduling using reinforcement learning via neural network based model approximation[C]// Proceedings of the 6th ACM International Conference on Systems for Energy-efficient Buildings, Cities, and Transportation. 2019: 287-296.

[16] ZHANG Z, CHONG A, PAN Y, et al. Whole building energy model for HVAC optimal control: A practical framework based on deep reinforcement learning[J]. Energy and Buildings, 2019, 199: 472-490.

[17] WEI T, WANG Y, ZHU Q. Deep reinforcement learning for building HVAC control[C]// Proceedings of the 54th Annual Design Automation Conference 2017. 2017: 1-6.

[18] BROCKMAN G, CHEUNG V, PETTERSSON L, et al. Openai gym[J]. arXiv: 1606.01540, 2016.

[19] TARTARINI F, SCHIAVON S. pythermalcomfort: A Python package for thermal comfort research[J]. SoftwareX, 2020, 12: 100578.

LI Peng, born in 1996, postgraduate. His main research interests include deep learning and deep reinforcement learning.

YI Xiu-wen, born in 1991, Ph. D, data scientist, researcher, is a member of China Computer Federation. His main research interests include spatio-temporal data mining and deep learning.

(责任编辑:柯颖)