控制与决策
Control and Decision

基于DDPG的冷源系统节能优化控制策略

闫军威， 黄琪， 周璇

引用本文:
闫军威, 黄琪, 周璇. 基于DDPG的冷源系统节能优化控制策略[J]. 控制与决策, 2021, 36(12): 2955−2963.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2020.0734

基于DDPG的冷源系统节能优化控制策略

闫军威, 黄 琪, 周 璇†

(华南理工大学 机械与汽车工程学院，广州 510641)

摘 要: 针对传统冷源系统节能优化方式机理建模复杂,缺乏自我学习能力,优化速度较慢等问题,提出一种基于数据驱动和自我学习机制的冷源系统节能优化控制策略,设计冷源马尔可夫决策过程模型,并采用深度确定性策略梯度算法(DDPG)解决维数灾难与避免控制动作离散化问题. 以夏热冬暖地区某大型办公建筑中央空调冷源系统为研究对象,对冷源系统控制策略进行节能优化,实现在满足室内热舒适性要求的前提下,减少系统能耗的目标. 在对比实验中,DDPG控制策略下的冷源系统总能耗相比PSO控制策略和规则控制策略减少了6.47 %和14.42 %,平均室内热舒适性提升了5.59 %和18.71 %,非舒适性时间占比减少了5.22 %和76.70 %. 仿真结果表明,所提出的控制策略具备有效性与实用性,相比其他控制策略在节能优化方面具有较明显的优势.

关键词: 冷源系统；强化学习；DDPG算法；节能优化控制策略；马尔可夫决策过程；策略梯度

中图分类号: TU831.3 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.0734 **开放科学(资源服务)标识码(OSID):**

引用格式: 闫军威,黄琪,周璇. 基于DDPG的冷源系统节能优化控制策略[J]. 控制与决策, 2021, 36(12): 2955-2963.

Energy-saving optimization control strategy of cold source system based on DDPG algorithm

YAN Jun-wei, HUANG Qi, ZHOU Xuan†

(School of Mechanical & Automotive Engineering，South China University of Technology，Guangzhou 510641，China)

Abstract: An energy-saving control strategy based on data-driven and self-learning mechanism is proposed to solve the problems of complex mechanism modeling, lack of self-learning ability and slow optimization speed of traditional energy-saving optimization methods for cold source systems. The Markov decision process model of cold source is designed and the deep deterministic policy gradient (DDPG) algorithm from policy gradient is used to solve the problem of dimensionality curse and can avoid discretization of control actions. In this paper, the central air conditioning cold source system of a large office building in the hot summer and warm winter area is selected as the research object, and the control strategy of the cold source system is optimized. The results show that under the premise of meeting the indoor thermal comfort requirement，the energy-saving control strategy of the system is realized with the goal of minimizing the energy consumption. In the comparison experiment, the total energy consumption of the cold source system under the DDPG control strategy is reduced by 6.47 % and 14.42 % compared with the PSO control strategy and the rule based control strategy, the average indoor thermal comfort is increased by 5.59 % and 18.71 %, and the proportion of total uncomfortable time is decreased by 5.22 % and 76.70 %, respectively. The simulation results show that the proposed control strategy has effectiveness and practicality, which has obvious advantage in energy-saving optimization compared with other control strategies.

Keywords: cold source system；reinforcement learning；DDPG algorithm；energy-saving optimization control strategy；Markov decision process；policy gradient

0 引 言

中央空调冷源系统能耗水平直接影响空调系统的高效运行与建筑节能工作的落实[1]. 对既有建筑冷源系统的控制策略进行节能优化控制是降低其能耗水平的有效技术手段.

目前,已有诸多学者开展了冷源系统控制策略节能优化问题的研究,常规方法分为3个步骤:首先建立设备模型,然后采用寻优算法搜寻最优参数,最后制定冷源系统的节能优化控制策略[2-3]. 上述研究方法的难点在于[4]:系统设备数量较多,运行参数相互

收稿日期: 2020-06-11；**修回日期:** 2020-09-25.

基金项目: 广东省自然科学基金项目(2017A030310162, 2018A030313352).

责任编委: 张国山.

†通讯作者. E-mail: zhouxuan@scut.edu.cn.

耦合,机理建模过程较为复杂,且随着运行时间的推移,系统参数发生缓慢时变,原有的模型难以满足节能优化控制需求;其次,部分模型虽然能够在线优化运行参数,但是由于不具备自我学习能力,每一步均需要进行控制参数寻优,对算法要求较高,难以满足控制实时性要求;此外,由于优化算法的随机性,寻优过程可能发生优化速度慢、陷入局部最优解、难以收敛等问题.

随着国家相关政策的推动与物联网技术的高速发展,现有建筑系统已经积累了大量冷源系统运行数据,为探索数据驱动的节能优化控制策略提供了强大支持. 强化学习(reinforcement learning, RL)是一种无模型、自适应的机器学习方法,基于强化学习的控制器通过“动作和奖赏”机制,能够根据系统环境变化,实时给出最优的控制策略. 同时,其无模型的特性在一定条件下能够避免复杂的系统建模过程,是一种面向数据驱动的智能控制方法[5]. 部分学者与研究人员已将其应用在中央空调系统节能优化领域中. 胡龄爻等[6]采用Q-learning算法,通过控制空调启停与窗户开关实现了建筑的节能舒适运行. Chen等[7]以自然通风的办公建筑为研究对象,将室外气象参数与室内温度作为控制器的输入,利用Q-learning算法输出HAVC系统的启停控制策略,与规则控制策略相比,两个仿真实验中的建筑能耗分别减少了13 %与23 %. 传统Q-learning算法应用于实际控制问题中,由于系统的状态空间与动作空间维数较大,可能会导致维数灾难问题. 为此,部分学者改进原有算法,利用神经网络逼近Q值函数,提出了DQN (deep Q network)算法. Wei等[8]利用DQN算法,实现水冷式中央空调系统送风量的优化控制,在保证室内热舒适的前提下降低了建筑能耗. Qiu等[9]以室内湿球温度与冷负荷为状态,以风机与水泵的频率为动作,提出基于DQN的冷却水系统控制策略,通过4种不同控制策略的仿真对比实验,验证了DQN控制策略的优越性.

上述文献中采用的算法核心是强化学习中的状态-动作值函数,智能体在决策时首先对比不同环境-动作对的回报大小,然后选取回报最高的动作作为策略,因此,需要对控制动作进行离散化. 值得注意的问题是:离散化的程度与方法对最终结果可能产生较大的影响;动作空间的复杂度将随着动作的增多而呈几何倍数增加,限制了其在复杂控制问题中的使用. 基于状态值函数的策略梯度法(policy gradient, PG)是另一类强化学习算法,智能体学习的对象是环境-动作的映射函数,无需对控制动作进行离散化,在解决复杂问题上具有较大的应用潜力[10]. 目前策略梯度法已应用于自动驾驶[11-12]、工业机器人[13]、电网调度[14]等领域,但在中央空调领域的研究相对较少.

为此,针对冷源系统这类复杂系统的节能优化控制,本文拟分析冷源系统特点,建立其马尔可夫决策模型,并采用深度确定性策略梯度算法(deep deterministic policy gradient, DDPG),以避免控制动作离散化与解决维数灾难问题;以某办公建筑冷源系统为研究对象,建立仿真实验平台,验证算法的有效性,并分析算法的节能优化效果.

1 冷源马尔可夫决策过程模型

强化学习是一种由多学科交叉发展而来的机器学习方法,是目前人工智能领域最为活跃的分支之一. 强化学习使用的策略是智能体与环境不断交互,采用不同的行为序列,从环境对行为序列的回馈信号中产生动作评价,依据评价去指导智能体以后的行动[5],其基本原理是:如果智能体的某个行为收获环境对智能体的正向回报,则智能体在该环境条件下采用这个行为的趋势会得到加强,反之则会减弱,在尝试中最终获得最优策略,其本质是动作到环境的复杂非线性高维映射.

本文选取的研究对象为夏热冬暖地区某大型办公建筑,由主楼、东楼和西楼3栋建筑组成,总建筑面积为14.75万m^2,中央空调使用面积约为10.26万m^2. 冷源系统设备参数如表1所示.

表1 冷源系统设备参数

序号	设备名称	详细参数	台数
1	冷水主机	额定制冷量3 517 kW,功率647 kW	2
2	冷水主机	额定制冷量1 758 kW,功率322 kW	1
3	冷冻水泵	功率90 kW,流量720 m^3/h,扬程40 m	4
4	冷冻水泵	功率75 kW,流量600 m^3/h,扬程32 m	2
5	冷却水泵	功率110 kW,流量720 m^3/h,扬程40 m	4
6	冷却水泵	功率80 kW,流量600 m^3/h,扬程32 m	2
7	方形冷却塔	功率11 kW,流量400 m^3/h	8

该建筑的中央空调系统已于2014年完成节能改造,累积了大量运行数据,本文的仿真实验对象为该建筑的水冷式中央空调冷源系统.

冷源系统节能优化控制与强化学习结合的前提是建立冷源马尔可夫决策过程(Markov decision process, MDP)模型,通常MDP模型被描述为一个四元组$\{S, A, P, R\}$[15],定义冷源MDP模型为:S是由冷源供冷区域及外部环境参数表述的状态空间;A是冷源控制器的可用控制指令组成的动作空间;P是冷源

系统不同环境状态之间的转移概率; R是在状态S时采取不同控制动作a所能得到的即时回报,通常是带有惩罚项的函数形式. 冷源MDP模型如图1所示.

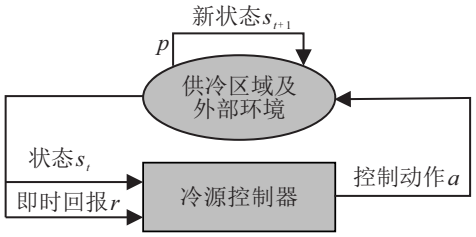


图 1　冷源马尔可夫决策过程

针对研究对象的实际情况,MDP模型的具体参数规定如下:

1) 状态空间S.

选择室内温度、室内相对湿度、室外温度、室外相对湿度和建筑运行时间表作为状态空间参数. 为作出周全决策,冷源控制器需要考虑过去、当前和未来3个方面的影响因素,类似于人的思考方式,因此选取当前时刻室内温度、室内相对湿度、室外温度、室外相对湿度和当前建筑运行时间表作为对当前系统状态的描述;选取前两个时刻的室内温度、室内相对湿度、室外相对湿度和前两个时刻建筑运行时间表作为对过去系统状态的描述;选取下两个时刻室外温度、室外相对湿度和下两个时刻建筑运行时间表作为对未来系统状态的描述. 因此,系统状态是一个由21维特征描述的状态空间.

2) 动作空间A.

根据动作选择与实际被控变量相符的原则,结合研究对象实际情况,选择冷冻水供水温度,冷冻水供回水压差和冷却水回水温度作为控制动作.

3) 转移概率P.

转移概率取决于执行控制动作后系统环境的真实状态, 智能体需要依靠多次蒙特卡洛采样对其作出无偏估计.

4) 即时回报R.

建筑物室内人体热舒适性与系统运行能耗是评价运行情况的主要指标,即时回报计算公式定义为

$$R = -\mathrm{Energy}(s_t,a_t)-\lambda[|\bar{P}-\mathrm{PMV}|+ |\underline{P}-\mathrm{PWV}|-(\bar{P}-\underline{P})]. \tag{1}$$

其中: $\mathrm{Energy}(s_t,a_t)$为该时刻冷源系统能耗,由于需要尽可能减少能耗,一般取为负值; PMV是室内热舒适性评价指标,越接近0表明室内越舒适. 为了保证建筑热工环境在舒适性范围内,需要附加关于PMV的罚函数,其中λ为热舒适性惩罚系数, $\bar{P}$和$\underline{P}$为PMV指标的上限与下限.

2　深度确定性策略梯度算法

2.1　确定性策略梯度

策略梯度法中,智能体将策略π直接进行参数化表示,学习环境-动作的映射关系的状态值函数[16]为

$$V_\pi(s,\theta)=E\Big[\sum_{t=0} r(s_t,a_t);\pi_\theta\Big]=\sum_\tau P(\tau,\theta)R(\tau). \tag{2}$$

其中: θ表示函数参数; π_θ表示参数化策略; $\gamma(s_t,a_t)$表示当前时刻即时回报; τ表示一组状态-动作轨迹序列; $P(\tau,\theta)$表示出现轨迹序列τ的概率; $R(\tau)$表示轨迹序列τ的累计回报.

对式(2)求导可得

$$\nabla V_\pi(s;\theta)=\sum_\tau P(\tau;\theta)\nabla\log P(\tau;\theta)R(\tau). \tag{3}$$

其中: $\nabla\log p(\tau,\theta)$是梯度方向,策略梯度最终成为求$\nabla\log P(\tau;\theta)R(\tau)$的期望,当使用策略$\pi_\theta$采样$n$条轨迹后,可以利用这$n$条轨迹的平均经验逼近策略梯度,如下式所示:

$$\nabla V_\pi(s;\theta)=\frac{1}{n}\sum_{i=1}^{n}\nabla\log p_i(\tau,\theta)R_i(\tau). \tag{4}$$

$\nabla\log p_i(\tau,\theta)$是第$i$条轨迹的梯度方向,它代表变化最快的方向,假如参数沿着该方向更新,则轨迹τ出现的几率将会增加,反之出现的概率将会减小. $R_i(\tau)$决定参数更新的步长,当$R_i(\tau)$为正回报时,智能体在今后将会增加采用该轨迹的频率,正向回报越大越容易出现该轨迹,反之则会降低该轨迹的频率[10]. 智能体会提高回报高的区域路径出现的概率,低回报区域路径则会在迭代更新中逐渐减少直到消失.

在Q-learning算法中,状态-动作值函数表示累计回报在状态-动作对处的期望值,被称为Q值函数,如下式[15]所示:

$$Q^\pi(s,a)=E_\pi\Big[\sum_{t=0}^{\infty}\gamma^t r_t|s_0=s,a_0=a\Big]. \tag{5}$$

其中: E_π表示累计回报期望; $\gamma\in(0,1)$为折扣因子,表示即时回报随着系统状态转移衰减的程度, r_t为即时回报; s_0表示系统初始状态; a_0表示初始动作; t表示系统时刻.

用确定性策略$\mu_\theta(a|s)$代替式(4)中的概率,用Q值函数替代轨迹回报,可以得到确定性策略梯度(deterministic policy gradient, DPG)计算公式[17]如下:

$$\nabla V_\pi(s;\theta)=\nabla_{\mu_\theta}(a|s)\nabla Q^\mu(s,a). \tag{6}$$

2.2　DDPG算法

与随机策略梯度相比,确定性策略梯度的计算效率得到显著提升,但却存在无法学习的问题,具体表

现为: 当系统初始状态确定时, 被智能体所采样的轨迹是固定的, 无法访问其他系统状态, 这样显然违背了强化学习在探索中学习的初衷. 为了解决这一问题, 文献[17]提出了异策略学习方法, 即智能体在学习阶段时, 行动策略采用随机策略从而保证探索的充分性, 评估策略则采用确定性策略保证训练效率, 其基本思想与异策略Q-leaning算法一致, 因此, DPG可以看成是Q-leaning在连续动作问题中的推广.

基于以上理论, 确定性策略梯度算法采用演员-评论家(actor-critic, AC)框架[16]实现, 其中actor负责给出不同环境状态下的动作, critic则负责评判此次动作的好坏, 两者是互相监督互相学习的关系, 如图2所示. 从本质上看, 前者从环境中学习的是状态值函数, 后者学习的是状态-动作值函数. 智能体在完成学习任务后, 只需要actor就可以完成特定的强化学习任务.

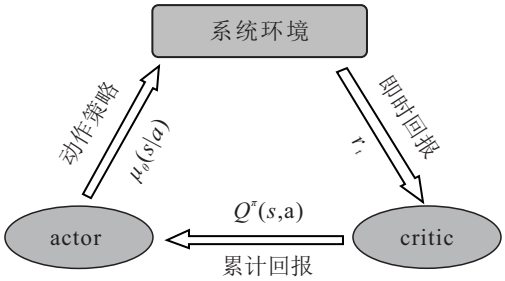


图2 AC框架示意

为了解决critic面临的维数灾难与actor策略参数化问题, 有关学者将深度神经网络引入到DPG中, 进一步提出了深度确定性策略梯度算法[18], 利用深度神经网络参数化逼近actor的状态值函数与critic的状态-动作值函数, 其神经网络权重参数主要更新公式如下所示:

$$\omega_{t+1}=\omega_t+\alpha_\omega[r_t+\gamma Q(s_{t+1},a_{t+1})-Q(s_t,a_t)],$$
$$\theta_{t+1}=\theta_t+\alpha_\theta\nabla\mu_\theta(s_t)\nabla Q(s_t,a_t). \tag{7}$$

其中: ω为critic权重参数; θ为actor权重参数; α_ω为critic学习率; α_θ为actor学习率; $r_t+\gamma Q(s_{t+1},a_{t+1})-Q(s_t,a_t)$被称为TD偏差.

DDPG算法的更新步骤如表2所示.

表2 DDPG算法主要更新步骤

步骤	过程
1	从数据库中随机抽取N条状态转移数据(s_t,a_t,r_t,s_{t+1})
2	actor根据s_{t+1}向critic提供a_{t+1}
3	critic根据(s_t,a_t)与(s_{t+1},a_{t+1})得到$Q(s_t,a_t)$与$Q(s_{t+1},a_{t+1})$, 计算TD偏差,
4	critic根据TD偏差, 按照式(7)进行权重参数更新
5	actor根据critic提供的$Q(s_t,a_t)$, 按照式(7)进行权重参数更新

3 冷源系统节能优化控制流程

针对本文的水冷式中央空调冷源系统, 节能优化运行流程可以分为5个部分, 流程如图3所示.

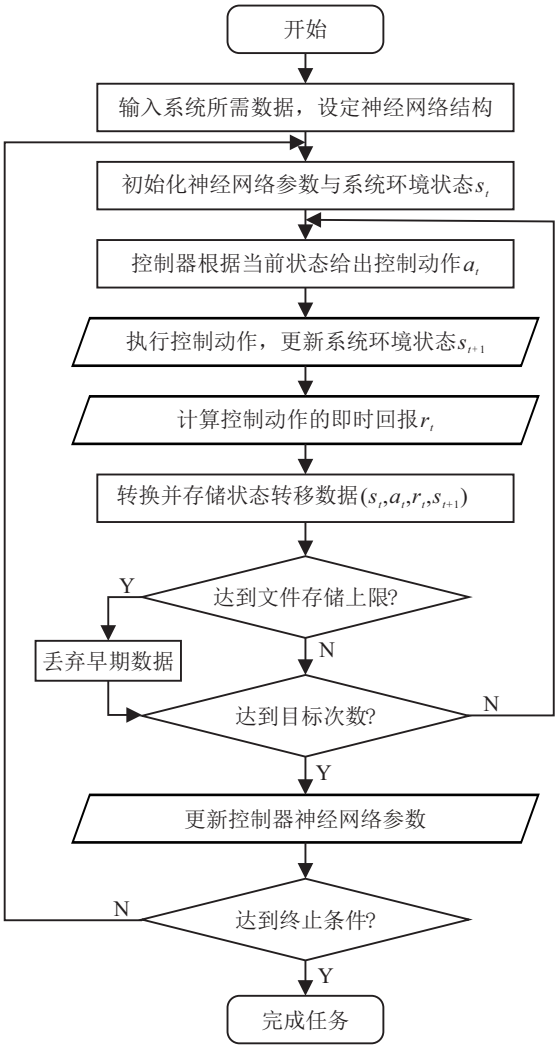


图3 节能优化流程

1) 系统初始化.

输入系统所需的气象参数、建筑时间表和冷源运行参数等数据, 设定冷源控制器的神经网络结构; 初始化控制器的神经网络参数与系统环境状态.

2) 经验产生.

冷源控制器根据当前系统环境状态, 依靠随机策略给出下一时刻控制策略, 冷源系统执行策略后系统环境状态发生改变, 控制器得到一个即时回报.

3) 数据转化与存储.

数据库接收系统状态数据, 将其转换为(s_t,a_t,s_{t+1})形式的数据格式, 其中s_t为当前时刻系统环境状态, a_t为冷源控制器给出的控制动作, r_t为当前动作的即时回报, s_{t+1}为执行控制动作后下一时刻系统环境状态; 转换完成后存入数据文件中, 当数据文件的容量达到上限时, 将会丢弃早期存储的数据.

4) 策略更新.

每隔固定的迭代次数,冷源控制器将从数据库中随机抽取 N 条数据,利用梯度下降法更新自身神经网络参数,达到改进控制策略的目的.

5) 完成任务.

随着迭代次数的逐渐增加,神经网络的参数会逐步收敛稳定,控制器给出的控制动作会越来越稳定,当系统累计回报不再发生明显变化时即认为学习过程结束,此时已经得到了最优策略,可以投入到正式使用中;此外,当系统达到最大迭代次数时,不管此时是否已经学习到最优策略,任务都将强行结束.

4　仿真验证与分析

4.1　仿真平台搭建

本研究团队搭建了该建筑的仿真平台系统,平台基于python语言编写,由数据交互模块、室内温度仿真模块、室内相对湿度仿真模块、冷源系统能耗仿真模块、气象参数模块和即时回报计算模块组成,如图4所示.

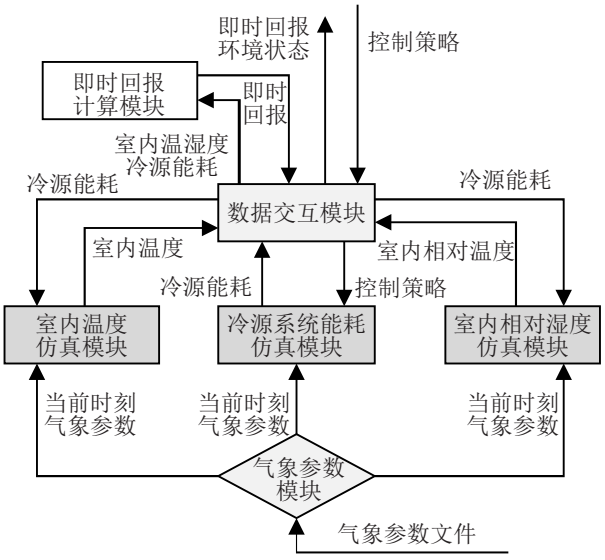


图 4　仿真平台设计及数据流向

平台中的仿真模块基于2018年4月至2019年9月的建筑实际运行数据进行训练,其中室内温度的仿真误差为0.98 %,室内相对湿度的仿真误差为5.03 %,冷源系统能耗的仿真误差为3.59 %;因此,可以认为由平台得到的仿真数据能够代表建筑的实际运行情况,满足使用要求.

4.2　仿真平台参数设置

仿真参数的设置可以分为以下两个部分:

1) 仿真平台参数设置.

设定建筑办公时间为9点至18点,没有设置周末;室外气象参数数据采用2019年7月的实际采集数据;设定即时回报中PMV指标的计算参数为:服装热阻0.52 clo,人体新陈代谢率1met,室内风速0.1 m/s;PMV上限为0.5,下限为−0.5;设定每次仿真最大幕次数为7幕,幕指系统从仿真起点到仿真终点的一次过程[19].

2) 算法参数设置.

与算法相关的参数根据仿真需求与人工经验进行设置,分为神经网络参数设置与算法超参数设置,相关参数均已经过寻优处理.

① 神经网络参数设置.

由于采用AC框架,算法拥有两个神经网络,参数设置如表3所示.

表 3　神经网络参数设置

参数名称	actor	critic
输入层神经元个数	200	400
隐藏层神经元个数	300, 150	300
输出层神经元个数	3	1
学习率	0.000 1	0.01
激活函数	Relu激活函数	

②算法超参数设置.

超参数是指算法在开始训练过程之前设置的参数,参数设置如表4所示.

表 4　DDPG算法超参数设置

超参数	数值	超参数	数值
折扣因子	0.3	初始噪声强度	0.8
热舒适性惩罚系数	10 000	批处理量	80
初始探索率	0.9	批迭代次数	10

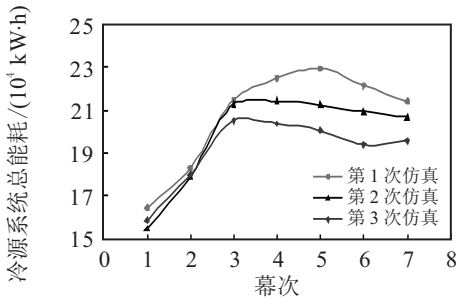
为了便于评价控制策略的节能优化效果,本文规定了3个指标对控制策略进行评价,分别为冷源系统总能耗,平均室内热舒适性和非舒适性时间占比. 每个指标的定义为:冷源系统总能耗是指在一幕内冷源系统消耗的总能耗;平均室内热舒适性是指在一幕内,建筑处于办公时间时室内PMV指标绝对值的平均值;非舒适性时间占比是指在一幕内,建筑处于办公时间时室内PMV指标超过即时回报规定的上下限时间占累计办公时间的百分比.

4.3　仿真结果分析

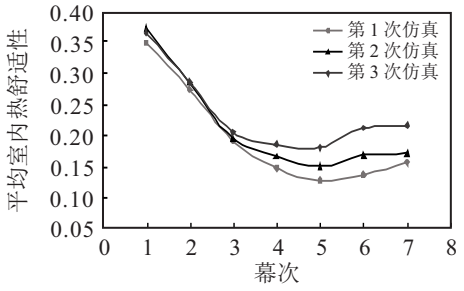
4.3.1　训练过程分析

利用仿真平台对冷源控制器进行3次仿真训练,每次仿真过程中冷源系统总能耗、平均室内热舒适性和非舒适性时间占比如图5所示.

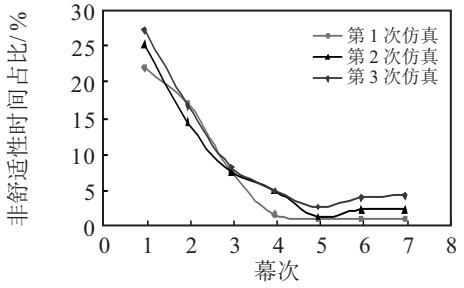
由图5可以看出,3次仿真训练过程中评价指标变化趋势基本相同,反映出控制策略的节能优化具有一定规律性. 第1幕结束时,由于控制器处于初始学习阶段,此时的控制效果并不理想,虽然冷源系统总能耗较低,但却牺牲了室内热舒适性;第2幕结束时,



(a) 冷源系统总能耗



(b) 平均室内热舒适性



(c) 非舒适性时间占比

图5 3次训练过程中评价指标变化

控制器已经学习到之前与环境交互所获得的经验,为了满足室内热舒适性要求而增加了冷源系统总能耗,让平均室内热舒适性有明显提升,非舒适性时间占比出现大幅下降;第3幕结束时,冷源控制器仍在继续学习,冷源系统总能耗明显上升,但室内热舒适性持续提高;第4幕结束时,冷源系统总能耗与上一幕相差较小,但室内热舒适性仍然有较大幅度的提升,说明此时控制器已经基本掌握较优的控制策略,它依据不同的系统环境,尽可能合理分配冷源系统能耗,用基本相同的能耗进一步提高了室内热舒适性;第5幕结束时,控制器继续尝试在满足室内热舒适性要求的前提下,尽可能减少冷源系统能耗,除第一次仿真外,其余两次仿真冷源系统总能耗相比第4幕均有小幅下降,控制器给出的控制策略已经接近于最优策略;第6幕和第7幕结束时,冷源系统总能耗持续下降,但室内热舒适性指标却开始上升,出现偏热现象,说明在3次仿真训练中策略梯度在第5幕时已经来到“顶峰”附近,在第6幕与第7幕时已经错过策略梯度“顶峰”而呈现下降趋势.

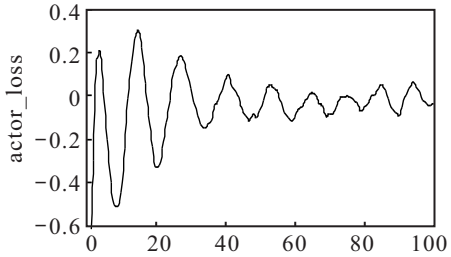
由图5还可以发现,控制策略具有一定的波动性,每一次训练过程不会完全相同,得到的各类评价指标会有不同程度的差别,取第5幕为最终训练结果,3次仿真中,冷源系统总能耗最低的是第3次仿真的200 532 kWh,平均室内热舒适性最佳的是第1次仿真的0.129,非舒适性时间占比最少的是第1次仿真的1.29 %,但是相对应的第1次仿真的冷源系统总能耗是最高的,第3次仿真的室内热舒适性是效果最差的,而第2次仿真可以视为第1次与第3次仿真的结合,较好地兼顾了冷源系统能耗与室内热舒适性两方面的即时回报因素,得到的评价指标为:冷源系统总能耗214 434 kWh,平均室内热舒适性0.151,非舒适性时间占比1.61 %.综上所述,取第2次训练第5幕结果作为最终的冷源控制器.

4.3.2 仿真结果分析

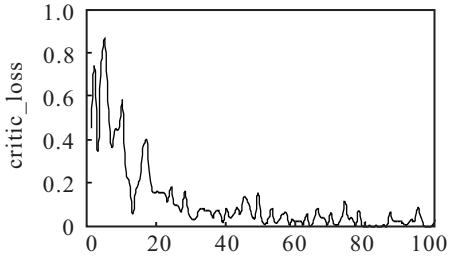
冷源控制器训练过程中actor与critic的反向传播误差变化如图6所示,室外温湿度与室内温湿度在建筑办公时间内变化情况如图7所示,冷源系统能耗与室内PMV值在建筑办公时间内变化情况如图8所示,系统控制策略在建筑办公时间内变化情况如图9所示.

由图6可以看出:actor的误差在0.3～−0.6之间摆动,随着训练次数的增加震荡收敛至0值附近;critic的误差从整体上看呈下降趋势,开始训练时数据震荡幅度较大,说明控制器初始学习时策略的不稳定性较高,随着训练次数的增加,控制器的控制策略逐渐稳定,误差的震荡幅度逐步减小,最后收敛至0值附近.

由图7可以发现,随着室外温度与相对湿度的动态变化,室内温度基本稳定在25 ℃～27 ℃之间,室内相对湿度基本稳定在63 %～78 %之间,满足夏热冬

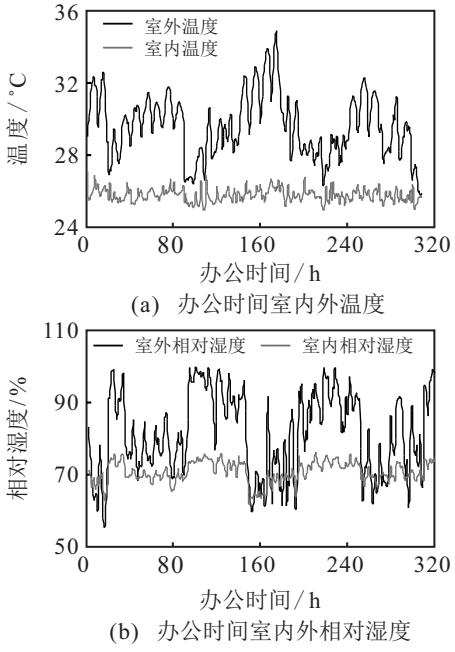


(a) actor的反向传播误差



(b) critic的反向传播误差

图6 控制器训练过程反向传播误差变化



(a) 办公时间室内外温度

(b) 办公时间室内外相对湿度

图7　办公时间室内外温度与相对湿度变化

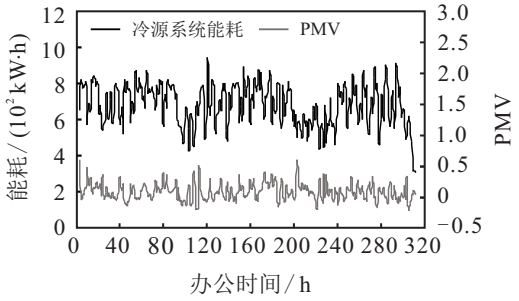
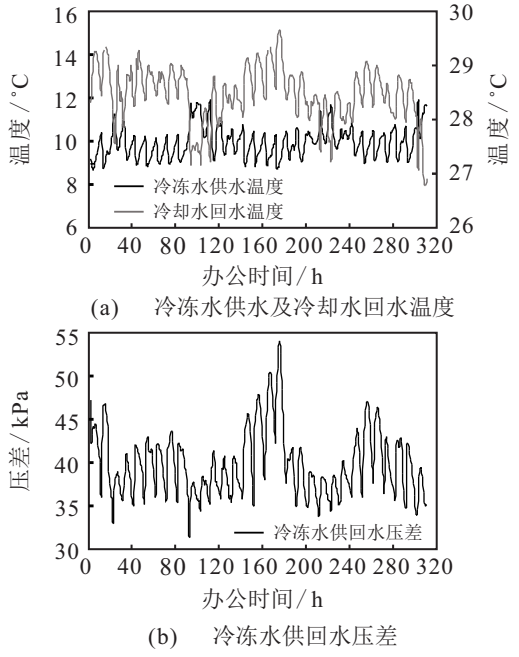


图8　办公时间冷源系统能耗与室内PMV值变化



(a)　冷冻水供水及冷却水回水温度

(b)　冷冻水供回水压差

图9　办公时间系统控制策略变化

冷地区室内人体热舒适性基本要求. 还可以发现,个别时间点出现室内温度或相对湿度偏高或者偏低的情况,这有两方面原因: 一是此时室外温湿度较高或者较低,系统仿真平台对于这种偏离较大的情况可能存在无法较为准确仿真的情况,导致冷源控制器的控制策略出现问题; 二是此类特殊点较少,对于冷源控制器来说属于离群点,内部算法会自动压缩离群点对整体控制策略的影响,因此未能使这些点达到较为理想的室内热舒适性状态. 但是从整个仿真时间来看,控制器给出的控制策略在绝大多数时间内均能满足使用要求,说明了节能优化控制策略的有效性.

由图8可以看出: 冷源系统能耗变化趋势与室外温度变化趋势基本一致,符合理论与实际使用情况. 室内PMV值在大部分时间内维持在 $-0.3\sim0.5$ 之间,室内整体热感觉略微偏暖,说明冷源智能控制为了减少冷源系统能耗牺牲了部分热舒适性,但优化后的室内舒适性仍然是可以接受的.

由图9可以发现,冷冻水供水温度与室外温度成反比关系,冷却水回水温度与室外温度基本成正比关系,符合理论与工程实际. 还可发现,冷冻水供回水压差受室外温度影响较大,变化趋势与其基本一致,满足系统实际使用情况. 以上仿真结果都充分说明了基于DDPG算法优化的冷源控制策略的实用性.

综上所述,可得到以下两个结论: 一是基于DDPG算法训练的冷源控制器能够在满足室内热舒适性要求的前提下,减少冷源系统能耗,实现节能优化控制策略; 二是DDPG算法在冷源系统节能优化控制领域具有实用性与有效性.

4.4　不同控制策略仿真对比

为了进一步验证DDPG控制策略的节能优化效果,选择粒子群优化(particle swarm optimization, PSO)控制和规则控制(rule based control, RBC)两种控制策略进行仿真对比,PSO控制策略是一种利用PSO算法[20]进行参数优化的控制策略,已经有较多学者对其进行研究. 规则控制策略是一种基于专家经验的简易控制策略,在实际工程项目中应用广泛,本文的研究对象即采用该控制策略.

仿真参数的设定可以分为以下两个部分:

1) 仿真平台参数设置.

室外气象参数数据采用2019-05-10～2009-09-20的实际采集数据,每次仿真最大幕次数为9幕,其余参数与4.2小节相同.

2) 控制器参数设置.

DDPG控制器,PSO控制器与RB控制器的参数设置如下:

①DDPG控制器.

批处理量设为300,批迭代次数设为16,其余参

数与4.2小节相同.

② PSO控制器.

利用PSO算法对不同外部环境下的冷源系统控制变量进行多次寻优,其中冷冻水供回水压差变量的范围是[30, 55],粒子飞行速度的范围是[0.5, 1],冷冻水供水温度变量的范围是[7, 13],粒子飞行速度的范围是[0.1, 0.5],冷却水回水温度变量的范围是[24, 33],粒子飞行速度的范围是[0.1, 0.5],学习因子取2,最大权重因子取0.9,最小权重因子取0.4,采用与DDPG算法相同的即时回报计算公式.

③ RB控制器.

规则控制器的控制策略按照建筑实际冷源系统控制策略进行设置.

利用仿真平台对3种控制器进行训练与仿真,3种控制策略最终得到的冷源系统总能耗,平均室内热舒适性和非舒适性时间占比如表5所示.

表5 不同控制策略仿真结果

控制策略	冷源系统总能耗/(kW·h)	平均室内热舒适性	非舒适性时间占比%
DDPG	102 034 5	0.152	1.26
PSO	109 096 7	0.161	2.64
RB	119 237 3	0.187	5.41

由表5可以看出,DDPG控制策略的3个评价指标均优于其他两种控制策略,与PSO控制策略和规则控制策略相比,冷源系统总能耗分别减少了6.47 %和14.42 %,充分体现出其显著的节能效果. 由表5还可以发现,DDPG控制策略并没有因为节能而牺牲了室内热舒适性,与上述两种控制策略相比,平均室内热舒适性分别提升了5.59 %和18.71 %,非舒适性时间占比减少了5.22 %和76.70 %,达到了在室内热舒适的前提下,减少冷源系统能耗的目标,充分体现出其良好的优化效果. 以上反映了DDPG控制策略的优越性,即更加节能,也更加舒适.

由表5还可以看出,DDPG控制策略的节能优化效果相比规则控制策略有较为明显的提升,考虑到平台的仿真误差较小,因此一定程度上可以认为在实际建筑运行环境中DDPG控制器也将优于RB控制器. 此外,虽然PSO控制策略与DDPG控制策略均能取得不同程度的节能优化效果,但PSO算法计算开销较大,训练效率较为缓慢,并且PSO控制器不具备学习能力,每个控制周期内均需要根据不同的环境条件重新优化控制策略,泛化能力较弱. DDPG控制策略采取先积累经验再集中学习的方式,训练效率显著提升,同时DDPG控制器具有学习能力,经过初始训练后,在实际运行时面对复杂多变的工况均可直接给出控制策略,具有实时性强的优点,更加适用于对系统响应速度要求高的工程项目.

5 结 论

本文首先结合冷源系统运行特点提出了冷源马尔可夫决策过程模型,采用DDPG算法对冷源控制策略进行节能优化,该算法源于强化学习中的策略梯度法,具有无需建模与自学习的特点,能够避免控制动作的离散化与维数灾难,解决较为复杂的控制问题; 然后,以夏热地区某大型办公建筑中央空调冷源系统为研究对象,搭建了仿真实验平台,对算法的节能优化效果进行了仿真研究. 主要得到以下结论:

1) 基于DDPG算法训练的冷源控制器能够在满足室内热舒适性要求的前提下,减少冷源系统能耗,达到节能优化控制要求; 同时,仿真数据表明,该算法在冷源系统控制领域具有实用性与有效性.

2) 与PSO控制策略和规则控制策略相比,基于DDPG控制策略的冷源系统总能耗减少了6.47 %和14.42 %,平均室内热舒适性提升了5.59 %和18.71 %,非舒适性时间占比减少了5.22 %和76.70 %,具有较为明显的优势; 同时,DDPG控制策略具有训练效率高与实时性强等优点,能够适用于对系统响应速度要求高的工程项目.

参考文献(References)

[1] 清华大学建筑节能研究中心. 中国建筑节能年度发展研究报告2019[M]. 北京: 中国建筑工业出版社, 2019.
(Building Energy Conservation Research Center of Tsinghua University. Annual development research report on building energy efficiency of China in 2019[M]. Beijing: China Construction Industry Press, 2019.)

[2] 吴伟伟, 范东叶, 朱文平, 等. 中央空调系统优化运行研究综述[J]. 建筑热能通风空调, 2019, 38(7): 37-41.
(Wu W W, Fan D Y, Zhu W P, et al. Operation optimal analysis of central air-conditioning system[J]. Building Energy & Environment, 2019, 38(7): 37-41.)

[3] Chen J Y, Sun Y J. A new multiplexed optimization with enhanced performance for complex air conditioning systems[J]. Energy and Buildings, 2017, 156(11): 85-95.

[4] Tang R, Wang S W, Shan K, et al. Optimal control strategy of central air-conditioning systems of buildings at morning start period for enhanced energy efficiency and peak demand limiting[J]. Energy, 2018, 151(5): 771-781.

[5] 刘全, 翟建伟, 章宗长, 等. 深度强化学习综述[J]. 计算机学报, 2018, 41(1): 1-27.
(Liu Q, Zhai J W, Zhang Z C, et al. A survey on deep reinforcement learning[J]. Chinese Journal of Computers, 2018, 41(1): 1-27.)

[6] 胡龄爻, 陈建平, 傅启明, 等. 一种面向建筑节能的强化学习自适应控制方法[J]. 计算机工程与应用, 2017, 53(21): 239-246.

(Hu L Y, Chen J P, Fu Q M, et al. Building energy efficiency oriented reinforcement learning adaptive control method[J]. Computer Engineering and Applications, 2017, 53(21): 239-246.)

[7] Chen Y J, Norford L K, Samuelson H W, et al. Optimal control of HVAC and window systems for natural ventilation through reinforcement learning[J]. Energy and Buildings, 2018, 169(6): 195-205.

[8] Wei T S, Wang Y Z, Zhu Q. Deep reinforcement learning for building HVAC control[C]. The 54th ACM/EDAC/IEEE Design Automation Conference (DAC). Austin: IEEE, 2017: 1-6.

[9] Qiu S N, Li Z H, Li Z W, et al. Model-free control method based on reinforcement learning for building cooling water systems: Validation by measured data-based simulation[J]. Energy and Buildings, 2019, 194(7): 203-221.

[10] 刘建伟, 高峰, 罗雄麟. 基于值函数和策略梯度的深度强化学习综述[J]. 计算机学报, 2019, 42(6): 1406-1438.
(Liu J W, Gao F, Luo X L. Survey of deep reinforcement learning based on value function and policy gradient[J]. Chinese Journal of Computers, 2019, 42(6): 1406-1438.)

[11] Wu Y K, Tan H C, Peng J K, et al. Deep reinforcement learning of energy management with continuous control strategy and traffic information for a series-parallel plug-in hybrid electric bus[J]. Applied Energy, 2019, 247(8): 454-466.

[12] 余伶俐, 魏亚东, 霍淑欣. 基于MCPDDPG的智能车辆路径规划方法及应用[J]. 控制与决策, 2021, 36(4): 835-846.
(Yu L L, Wei Y D, Huo S X. The method and application of intelligent vehicle path planning based on MCPDDPG[J]. Control and Decision, 2021, 36(4): 835-846.)

[13] 张福海, 李宁, 袁儒鹏, 等. 基于强化学习的机器人路径规划算法[J]. 华中科技大学学报: 自然科学版, 2018, 46(12): 65-70.
(Zhang F H, Li N, Yuan R P, et al. Robot path planning algorithm based on reinforcement learning[J]. Journal of Huazhong University of Science and Technology: Natural Science Edition, 2018, 46(12): 65-70.)

[14] Kou P, Liang D L, Wang C, et al. Safe deep reinforcement learning-based constrained optimal control scheme for active distribution networks[J]. Applied Energy, 2019, 239(4): 324-342.

[15] 张峰, 刘凌云, 郭欣欣. 基于改进Q-学习算法的多阶段群体决策模型[J]. 控制与决策, 2019, 34(9): 1917-1922.
(Zhang F, Liu L Y, Guo X X. A multi-stage group decision model based on improved Q-learning[J]. Control and Decision, 2019, 34(9): 1917-1922.)

[16] 陈亮, 梁宸, 张景异, 等. Actor-Critic框架下一种基于改进DDPG的多智能体强化学习算法[J]. 控制与决策, 2021, 36(1): 75-82.
(Chen L, Liang C, Zhang J Y, et al. A multi-agent reinforcement learning algorithm based on improved DDPG in Actor-Critic framework[J]. Control and Decision, 2021, 36(1): 75-82.)

[17] David S, Guy L, Nicolas H, et al. Deterministic policy gradient algorithms[C]. The 31st International Conference on Machine Learning. Beijing: JMLR, 2014: 387-395.

[18] 陈建平, 何超, 刘全, 等. 增强型深度确定策略梯度算法[J]. 通信学报, 2018, 39(11): 106-115.
(Chen J P, He C, Liu Q, et al. Enhanced deep deterministic policy gradient algorithm[J]. Journal on Communications, 2018, 39(11): 106-115.)

[19] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[20] 王玉昆, 陈雪波. 独立局部搜索与多区域渐近收敛的新型PSO算法[J]. 控制与决策, 2018, 33(8): 1382-1390.
(Wang Y K, Chen X B. Improved multi-area search and asymptotic convergence PSO algorithm with independent local search mechanism[J]. Control and Decision, 2018, 33(8): 1382-1390.)

作者简介

闫军威(1968−), 男, 教授, 博士生导师, 从事建筑节能、机器学习等研究, E-mail: mmjwyan@scut.edu.cn;

黄琪(1995−), 男, 硕士生, 从事空调节能控制、强化学习的研究, E-mail: 1196982166@qq.com;

周璇(1976−), 女, 副研究员, 从事建筑节能、数据挖掘等研究, E-mail: zhouxuan@scut.edu.cn.

(责任编辑: 孙艺红)