基于强化学习的智慧楼宇机房冷却系统优化算法

郭振君[1] 韩明涛[1] 李井鹏[1] 董文吉[1]
GUO Zhenjun HAN Mingtao LI Jingpeng DONG Wenji

摘 要 传统智慧楼宇机房冷却系统控制方法通常依赖于机械制冷、电力和暖通方面的近似模型，这些模型难以设计，且不具备通用性。为此，提出一种基于 Actor-Critic 框架的策略优化算法，用于自主控制机房冷却系统。通过持续与环境互动获得经验，并利用这些经验优化控制策略，以更节能的方式确保机房正常运行。与传统控制算法相比，基于强化学习方法无需明确模型知识，只需设计奖励信号即可自动优化系统性能。在模拟平台上评估所提出的算法，实验结果表明，相较于 PID 控制算法，所提出的方法在冷却效率上提高了 13.9%。

关键词 智慧楼宇机房；冷却系统；强化学习；Actor-Critic；PID 控制算法；策略优化

doi：10.3969/j.issn.1672-9528.2024.10.052

0 引言

随着大数据时代的到来，智慧楼宇机房建设已成为关键环节，其规模快速增长，电力消耗也逐年上升。机房以确保 IT 设备安全为首要任务，冷却设备必须时刻运作，故在保证制冷效果的基础上降低制冷功耗具有十分重要的意义。

陈炯德等人[1]通过非线性回归网络建立室内温度模型，并使用粒子群算法优化控制量，提升了系统节能效率，但实际应用中存在局限性。Yang 等人[2]提出了一种结合监督学习和优化策略的机房能效提升方法，用于预测和管理冷却系统，但依赖大量高质量数据，并在面对环境动态变化时适应性不足。针对以上问题，本文提出了一种基于 Actor-Critic 框架的冷却系统策略优化算法（actor-critic cooling optimization，ACC-CO）。ACC-CO 算法不需要构建显式模型知识，具有通用性和灵活性。除此之外，通过与环境的不断交互，自主学习最优控制策略，减少对训练数据的依赖。

1. 浪潮通信信息系统有限公司 山东济南 250013

[11]HOU Q, ZHOU D, FENG J.Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.Piscataway:IEEE,2021:13713-13722.

[12]WOO S, PARK J, LEE J Y, et al.CBAM: convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV).[S.l.]:[s.n.],2018:3-19.

[13]GEVORGYAN Z.SIoU loss: More powerful learning for bounding box regression[EB/OL].(2022-05-25)[2024-05-11]. https://arxiv.org/pdf/2205.12740.

[14]LIN T, MAIRE M, BELONGIE S, et al.Microsoft coco: common objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13.Berlin:Springer International Publishing,2014:740-755.

[15]REDMON J, FARHADI A.YOLOv3: an incremental improvement[EB/OL].(2018-04-08)[2024-05-11].https://arxiv.org/abs/1804.02767.

[16]THUAN D. Evolution of YOLO algorithm and YOLOv5: the state-of-the-art object detention algorithm[J/OL]. Computer science,2021.[2024-05-16].https://www.theseus.fi/bitstream/handle/10024/452552/Do_Thuan.pdf?isAllowed=y&sequence=2.

[17]LI C, LI L, JIANG H, et al.YOLOv6: a single-stage object detection framework for industrial applications[EB/OL]. (2022-09-07)[2024-05-11].https://arxiv.org/abs/2209.02976.

[18]WANG C, BOCHKOVSKIY A, LIAO H.YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Piscataway:IEEE,2023:7464-7475.

【作者简介】

周秋航（1997—），男，湖北宜昌人，硕士研究生，研究方向：计算机视觉。

（收稿日期：2024-06-28）

1 冷却系统介绍

智慧楼宇机房冷却系统通常分为风冷空调、水冷空调与液冷空调。水冷空调系统在冷却效率、环境友好性和成本方面表现良好，是一种适用于大多数机房的有效冷却方式。故本文选用常用的水冷空调作为研究对象，如图 1 所示。

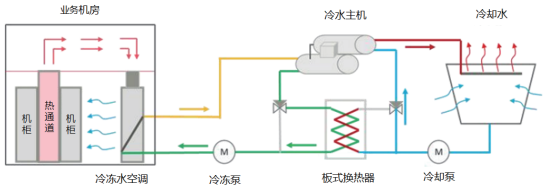

图 1 水冷冷却系统示意图

2 基于 Actor-Critic 框架的策略优化算法

2.1 Actor-Critic 框架介绍

Actor-Critic 框架是一种强化学习算法 [3]，如图 2 所示，通过结合策略梯度和价值函数能够高效处理连续动作空间优化问题。Actor 网络通过策略网络生成动作概率分布，并根据与环境的交互反馈更新策略，以优化动作选择。Critic 网络计算当前策略的期望回报值，评估策略的效果，并将评估结果反馈给 Actor 网络以指导策略改进。环境反馈的奖励信息经过折扣因子处理后传递给 Critic 网络，帮助其提高回报值的评估能力。

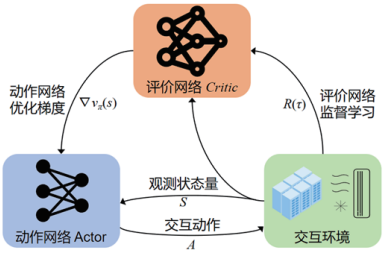

图 2 Actor-Critic 框架示意图

对该智能体进行训练的最终优化目标是寻找一个动作决策网络来搜寻收益最高的动作策略 π^*，该策略可评价网络的评估达到最大值，即目标函数为：

$$\pi^* = \arg\max_{\pi} J(\pi) \approx \arg\max_{\pi} v_\pi(s_0) \tag{1}$$

式中：$J(\pi)$ 表示累计奖励的期望值，s_0 为初始状态。

为了提高网络训练的稳定性，在梯度优化函数中采用 TD（temporal difference）网络优化方法，通过在不同时间点之间计算差异来调整和优化网络权重，有效地提升训练效率。由此得到的 Actor 网络的优化梯度为：

$$\nabla_\theta J(\theta) = \mathrm{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(a_t \mid s_t) A(s_t, a_t)] \tag{2}$$

$$A(s_t, a_t) = Q^{\pi_\theta}(s_t, a_t) - V^{\pi_\theta}(s_t) \tag{3}$$

式中：θ 为策略参数；E 表示在策略 π_θ 下所有可能的状态和动作的期望；a_t 表示在时间步 t 采取的动作；s_t 表示在时间步 t 观察到的状态；$A(s_t, a_t)$ 表示优势函数，衡量在状态 s_t 下选择动作 a_t 相较于原状态价值评估的优越性。

2.2 环境建模

为了保证算法能取得良好的效果，需要对机房冷却系统环境进行建模，包括定义状态空间、动作空间与奖励函数。

状态空间用于描述环境当前状态的所有可能状态的集合。在机房的冷却系统中，状态空间需要能够全面反映系统运行状态。假设冷却系统 i 负责 n 个机架的散热调控，在每个机架中部署温度传感器用于监测，衡量服务器的运行状态。另外，冷却系统的功耗同样是我们的关注指标。所以，在固定窗口时间采样周期 T 内，对于冷却系统 i 的状态空间可以表示为 $S_i=(t_1, t_2,\cdots, t_n, p_i)$，其中 t 表示冷却系统 i 负责 n 个机架安装的传感器在周期 T 内的平均温度；功耗表示冷却系统 i 在周期 T 内的功耗。

状态空间元素与控制空调设定温度与设定风速密切相关。机房中通常有多个冷却系统，每个冷却系统 i 的动作空间为 A_i，空调 i 执行动作 $a_i=(\mathrm{temp}_i, \mathrm{speed}_i)$，$a_i \in A_i$。$\mathrm{temp}_i$ 与 speed_i 分别表示第 i 台空调的温度与风速，两者均在安全范围内进行调节。

为了评估冷却系统选择策略的好坏，需要设置奖励函数作为评价标准。除了保证机柜服务器在正常工作温度范围内工作，还需要把冷却系统的功耗降到最低。所以，将制冷系统的奖励函数设计为两部分，分别是温度惩罚代价与冷却系统功耗。为了方便模型计算，将奖励函数设置为负。

空调 i 在执行动作 a_i 后，能耗为滑动窗口周期内开始节点与终止节点的电表数据之差。温度违约惩罚与设备温度超出设定范围相关，温度违约代价的计算公式为：

$$t_{\mathrm{cost}} = \mathrm{cost}_1 + \mathrm{cost}_2 + \ldots + \mathrm{cost}_n \tag{4}$$

$$\mathrm{cost}_j = \begin{cases} t_j - t_{\max}, t_j > t_{\max} \\ t_{\min} - t_j, t_j < t_{\min} \\ 0, t_{\min} \le t_j \le t_{\max} \end{cases}, 1 \le j \le n \tag{5}$$

式中：$t_{\min}$ 与 $t_{\max}$ 分别表示设备安全阈值的最小温度与最大温度。

所以，奖励函数的计算公式表示为：

$$R_i = -(\alpha t_{\mathrm{cost}} + (1-\alpha) p_i \times w) \tag{6}$$

$$w = \frac{p_{\max} - p_{\min}}{(t_{\mathrm{cost}})_{\max} - (t_{\mathrm{cost}})_{\min}} \tag{7}$$

式中：α 表示奖励函数中空调功耗在奖励中所占的权重，若 α 变大，则表示温度违约惩罚更大；若 α 变小，则表示冷却系统功耗过大的惩罚更大。w 表示标准化系数，保证温度惩罚与冷却系统功耗在同一个量级。

2.3 冷却系统策略优化算法

强化学习算法在训练过程中需要大量探索和尝试，这可能会导致机房的环境温度出现异常波动，进而影响设备的正

常运行和机房的整体性能，甚至对设备产生不可逆的伤害。所以，在设计冷却系统优化策略时，需要充分考虑算法的稳定性和可靠性，故设计 ACC-CO 的算法流程如下。

（1）初始化环境，经验回放池 D，Actor 网络参数和 Critic 网络参数

（2）for step=1 to Last do

（3）获取当前状态 s 作为输入

（4）根据 Actor 网络，按 ε -greedy 策略选择动作 a

（5）执行动作 a，观察环境反馈下一状态 s_{next} 和奖励 r

（6）将（s, a, r, s_{next}）存入经验回放池 D

（7）从经验回放池 D 中随机采样出 N 个样本（s, a, r, s_{next}）

（8）for i=1 to N do

（9）使用 Critic 网络计算当前 Q 值：$Q_{current}$=Critic(s, a)

（10）使用 Critic 网络计算下一个状态 Q 值，Q_{next}=Critic(s_{next}, Actor(s_{next}))

（11）计算目标 Q 值：$Q_{target} = r + \gamma \times Q_{next}$

（12）计算 Critic 网络的损失：$L_{critic}=(Q_{target} - Q_{current})^2$

（13）使用梯度下降法更新 Critic 网络参数，以最小化 L_{critic}

（14）计算 Actor 网络的损失：L_{actor} = -Critic(s, Actor(s))

（15）使用梯度上升法更新 Actor 网络参数，以最小化 L_{actor}

（16）end for

（17）状态转移：$s=s_{next}$

（18）end for

ACC-CO 算法不断与环境的交互过程中，通过反复采样和更新网络参数，使 Actor 网络生成的策略和 Critic 网络的价值评估不断改进，最终实现机房冷却系统在能耗最优的情况下保证设备正常工作。

3 实验分析

3.1 实验环境

实验环境在仿真软件 Reality DC Design Pro 中进行。其是一款专业的设计软件，能够帮助用户创造高度真实的虚拟现实场景，如图 3 所示。共设置四排机架，每排包含 7 个机柜，具体参数请见表 1。

表 1 机房组件参数

名称	设定值
机柜电力限额	5 kW
机柜服务器个数	8
服务器标称功率	800 W
设备最高温度阈值	32 ℃
通风地板尺寸	0.62 m×0.62 m
空调最大感热冷却能力	60 kW
空调最大流量	3.6 m^3/s

为了使仿真实验更接近真实环境，实验采用了阿里巴巴 2018 年的集群追踪数据，模拟真实机房的负载变化。

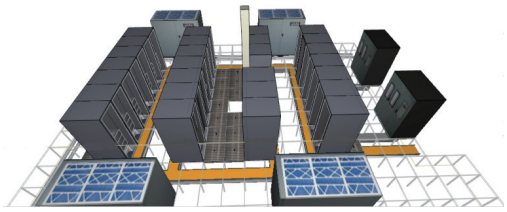

图 3 仿真环境

四个空调以送风温度 18 ℃，送风速度 100% 作为初始化条件，然后交由仿真软件进行仿真模拟。另外，由于服务器送风口冷热空气质量差异，机柜上下温度存在差差异，如图 4 所示，故传感器选择安装在送风口中间点位置。

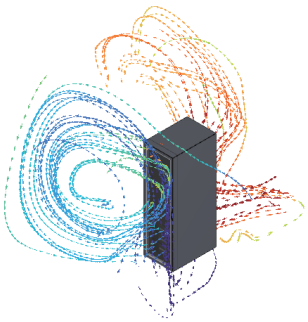

图 4 机柜温度流线图

3.2 算法参数

算法 Actor 网络共分为四层，输入层神经元个数为状态空间的维度，包括各个温度传感器采集温度与冷却系统能耗；第一隐藏层包含 64 个神经元，激活函数选择 ReLU 函数，引入非线性并提高网络的表示能力；第二隐藏层包含 32 个神经元，进一步提取特征，减少参数数量和计算复杂度，同样使用 ReLU 函数；输出层包含两个神经元，即分别对应空调的送风温度与送风速度。

Critic 网络同样分为四层，输入神经元个数为状态空间和动作的组合，用于接受当前环境状态和 Actor 网络输出的动作；第一隐藏层包含 128 个神经元，使 Critic 网络能够充分学习状态和动作的复杂关系；第二隐藏层包含 64 个神经元，进一步提取特征；输出层为一个神经元，输出当前状态和动作的价值。训练过程的超参数设置请见表 2。

表 2 训练参数

参数	初始值	含义
lr_a	0.001	Actor 网络学习率
lr_c	0.001	Critic 网络学习率
α	0.5	惩罚权重系数
ε	0.9	初始探索率
γ	0.95	奖励折扣因子
n_r	0.1	动作添加噪声

3.3 实验结果分析

本文提出的 ACC-CO 算法与 PID 控制算法进行了对比实验，如图 5，每组实验的实验长度为 800 步。分别从瞬时奖励、温度违约惩罚与空调功耗进行对比分析。

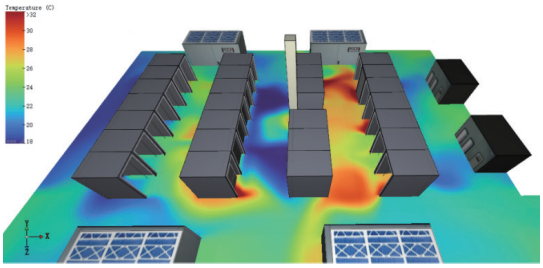

图 5 机房温度场仿真示意图

图 6 表示 ACC-CO 算法与 PID 控制算法的瞬时奖励折线图。PID 控制算法是根据控制器控制策略决定的，不涉及探索过程，相对比较稳定；ACC-CO 算法初始以 90% 的概率进行动作探索，以 10% 的概率采用 Actor 网络评估的动作策略，保证能够探索动作空间的张度。算法开始时，ACC-CO 为探索阶段，需要经过一段时间的训练才能输出最优的动作。训练过程中，探索概率以 0.5% 逐步递减，经过 180 步左右的迭代，探索概率达到下限值 1%。随着时间的变化，ACC-CO 算法逐渐趋于稳定，其瞬时奖励也高于控制器算法，达到了更好的控制效果。

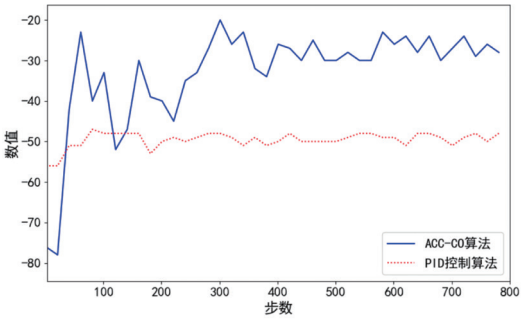

图 6 瞬时奖励折线图

在机房能效管理中，PUE 是关键的评价指标。PUE 定义为机房总能耗与 IT 设备能耗的比值，理想情况下，PUE 值越接近 1，表示机房的能效比越高。将达到稳态的 AC 模型与 PID 控制算法进行比较，如图 7 所示。

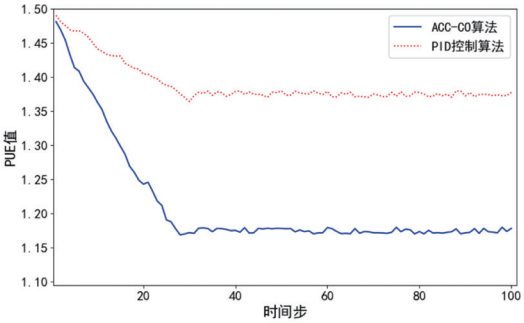

图 7 PUE 折线图

从实验结果可以看出，PID 控制算法虽然能够快速响应温度变化，控制策略相对固定，保证了设备的稳定运行，但在能耗优化方面同样存在局限性。AC 模型能够通过持续学习与优化，动态调整机房的能耗管理策略。尽管在初始阶段由于较高的探索概率，瞬时奖励的波动较大，表现出一定的不稳定性，但这种策略为动作空间提供广度和深度搜索。通过这样的探索过程，AC 模型能够找到更优的动作组合，实现设备正常工作的同时，显著降低能耗。经过充分的训练后，PID 控制算法的 PUE 值稳定到 1.362，AC 模型在稳态时的 PUE 值稳定到 1.173。因此，应用 ACC-CO 算法的机房冷却系统的节能效率较 PID 控制系统提高 13.88%，为机房的绿色发展提供有力的支持。

4 总结

为了解决智慧楼宇机房冷却系统能效优化问题，本文提出了一种基于 Actor-Critic 框架的机房冷却系统策略优化算法。Actor 网络生成冷却系统的动作组合，Critic 网络评估 Actor 网络选择动作对性能的影响，算法可以同时学习如何选择最优动作以及如何评估这些动作的长期影响，从而在训练过程中逐步优化冷却系统的能效。在与环境的交互中逐步优化冷却系统的能效，同时确保系统稳定运行和性能优化的双重目标。

参考文献：

[1] 陈炯德，王子轩，姚晔，等．变风量空调系统用非线性模型预测控制方法研究 [J]. 制冷学报 ,2019,40(6):62-69.

[2]YANG Z, DU J, LIN Y, et al. Increasing the energy efficiency of a data center based on machine learning[J]. Journal of industrial ecology, 2022, 26(1): 323-335.

[3]JIA Y, ZHOU X Y. Policy gradient and actor-critic learning in continuous time and space: theory and algorithms[J]. Journal of machine learning research, 2022, 23(275): 1-50.

【作者简介】

郭振君（1983—），男，山东烟台人，本科，研究方向：数据中心能耗优化、强化学习。

韩明涛（1986—），男，山东日照人，本科，研究方向：数据中心能耗优化、强化学习。

李井鹏（1979—），男，山东泰安人，本科，研究方向：数据中心能耗优化、强化学习。

董文吉（1997—），男，山东潍坊人，硕士，研究方向：数据中心能耗优化算法、强化学习。

（收稿日期：2024-07-16）