基于强化学习的空调系统运行优化

王尉同[1] 潘毅群[1] 黄治钟[2]

（[1]同济大学 机械与能源工程学院，上海 201804

[2] 同济大学 中德工程学院，上海 201804）

摘要：空调系统运行优化是建筑节能的重要组成部分。本文将强化学习算法应用在空调系统运行优化中。首先建立空调系统运行优化任务的马尔科夫决策过程模型，接下来确立规则对探索空间进行进一步缩减，使探索过程尽可能满足系统设备运行和被控区域舒适度的约束，最后应用 Q-Learning 算法探索学习不断改进控制策略。通过仿真案例对算法进行了测试。采用动态仿真软件 TRNSYS 和 MATLAB 进行联合仿真，对被控区域送风量进行了优化，仿真结果显示，与基于固定规则的控制策略和 PID 控制相比，本文提出的基于强化学习的控制方法可以在小幅提升被控区域舒适度的同时可将运行费用分别降低超过 7%和 4.5%。

关键词：空调系统；运行优化；强化学习；Q-Learning

OPTIMIZATION OF HVAC SYSTEM OPERATION BASED ON REINFORCEMENT LEARNING

Abstract: The operation optimization of HVAC system is an essential part of building energy conservation. In this work, reinforcement learning technique is applied to optimize HVAC system operation. Firstly, we formulate the HVAC system operation as a Markov decision process. Secondly, rules are defined to shrinkage the exploration space in order to avoid violating equipment operation constraints and conditioned zones comfort constraints during exploration. Finally, Q-Learning algorithm is deployed to improve the control strategy through interacting with the environment. This study tests the algorithm by a simulation case. Dynamic simulation program TRNSYS and MATLAB are utilized to conduct co-simulation. Results show that, compared with fixed rule-based control strategy and PID controller, reinforcement learning controller can reduce energy cost by over 7% and 4.5% separately by adjusting supply air flow rate, and at the same time improve occupants comfort level slightly.

Keywords: HVAC system; operation optimization; reinforcement learning; Q-Learning

0. 引言

空调系统运行优化是建筑节能的重要组成部分。建筑运行优化的基本目标是在满足被控区域舒适度要求的前提下尽可能的降低能耗或者能耗消费。为了适应室外不断变化的天气情况和室内负荷变化情况，如何动态调整系统中的设定值或者运行规则从而提高建筑能效一直是建筑运行优化的研究重点。

经典控制器包括基于规则的控制（如启停控制），PID 控制等由于结构简单，初投资较低被广泛应用，然而空调系统是典型的高度非线性、不确定性、时变性、耦合性的复杂系统，导致经典控制方式的表现并不理想[1]。近年来模型预测控制（MPC）在空调系统中的

应用受到广泛的关注，MPC 的基本思想是在每个时间步长，通过对未来一段时间窗内进行滚动优化得到最优的控制策略。通过对未来的室内扰动和室外天气状况进行预测，可以显著提高建筑能效[2]。然而 MPC 的实际运行效果严重依赖模型的准确度，尤其是对于建筑热湿环境控制问题，难以建立精确且能应用在实时优化控制中的建筑动态模型，一旦数学模型与实际情况存在较大偏差，MPC 计算得到的控制策略的效果将难以保证。Killian[3][4][5]等人针对将 MPC 实际应用在建筑运行中遇到的挑战进行了阐述与总结。

随着强化学习在算法和理论上的研究深入，强化学习算法在实际的工程优化和控制中得到了广泛的应用[6]。与 MPC 的特点相反，基于强化学习的免模型控制方法是一种数据驱动的控制方法，通过不断尝试，利用反馈的信息来更新控制策略，降低了对先验知识的依赖。基于强化学习的控制方法与人类通过学习获取经验的方式非常相似[7]，吸引了部分研究人员对将其应用在建筑运行优化上进行了初步的尝试。Liu[8]等人将强化学习应用在空调蓄能装置运行优化中，作者指出使用的经典 Q-Learning 算法存在对于高维度空间学习效率低下的问题。Dalamagkidis[9]等人将经典的强化学习控制方法与传统的启停控制，和模糊 PD 控制进行了对比，结果表明，在使用几年的仿真数据训练之后的强化学习控制器的控制表现与其他控制器控制效果大致相等或表现更加出色。Li[10]等人提出了多尺度强化学习，用来加速最优控制策略求解过程。Yang[11]等人将强化学习控制器应用在低烟建筑中，通过仿真进行了验证，但计算过程并未考虑建筑的热惰性。Ruelens[12]等人使用强化学习控制器确定温度设定重置策略，提升热泵运行效率，仿真结果显示与固定温度状态点的方式相比节能显著。Costanzo[13]等人将强化学习控制器应用在建筑需求相应中，并通过实验进行了验证，定量分析表示强化学习控制策略可以达到 90%的全局最优解。Wei[14]等人提出了基于深度强化学习的空调系统控制方法，并验证了深度强化学习控制器的可扩展性，但深度强化学习控制器存在训练所需的时间过长的问题。总体来说，目前相关研究主要集中在验证强化学习算法在不同的空调系统运行优化场景中的表现，其应用方式缺乏系统性的梳理，如何进一步提升强化学习控制器的实用性仍然是目前主要的研究方向。

本文将强化学习算法应用在空调系统运行优化中，首先对目前的空调系统运行优化方法进行了梳理，重点介绍基于强化学习的系统运行优化目前的研究情况，其次介绍强化学习相关理论和算法，提出强化学习控制器在空调系统中完整的应用流程，最后通过仿真案例进行验证，以一小型办公建筑作为研究对象，使用强化学习控制器动态调节房间送风量，对其应用效果进行分析。

1. 强化学习与 Q-Learning 算法

1.1 强化学习

强化学习是一种以环境反馈作为输入的、特殊的、适应环境的机器学习方法，它的主要思想是与环境交互和试错，利用评价性的反馈信号实现决策的优化[15]。图 1 为强化学习的一个简单图示。强化学习任务通常用马尔科夫决策过程来描述：包括 S 状态集，A 动作集，P 转移函数和 R 奖赏函数。在每个决策时刻 t，智能体选择执行动作 $a_t(a_t \in A)$，经过一个时间步长 Δt，环境 E 从状态 s_t 转变为 s_{t+1}，在 t+1 时刻，智能体根据奖赏函数计算此时间步长内的奖赏 r_t。转移函数描述在执行动作 $a \in A$ 后，状态从 s 转移到 s' 的概率，在马尔科夫决策过程中，环境的下个状态仅与当前的状态和当前采取的动作有关。强化学习的目标是获取一个策略函数 $a = \pi(s)$，通过执行策略函数即可最大化累计奖赏[16]，此策略函数即为学习到的最优控制策略。强化学习方法具有目标导向特点，善于求解未知环境下的决策链路优化问题。

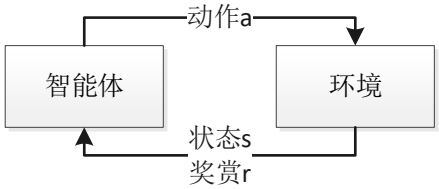

图 1 强化学习图示

Fig 1 Reinforcement Learning Problem

1.2 表格型 Q-Learning 算法

Q-Learning 算法是一种可以求解具有不完全信息的马尔科夫决策问题的强化学习方法。算法考察的对象是状态—动作对所对应的值函数，即 Q 值函数，用 $Q(s,a)$ 表示，代表状态 s 下执行动作 a 系统将获得的累计奖赏值。表格型 Q-Learning 是指算法将有限个状态—动作对的 Q 值存储在一张表中，初始阶段 Q 值为随机初始值，在智能体与环境不断交互的过程中通过式（1）对表格中的 Q 值进行更新：

$$\begin{gathered}\Delta Q(s,a)=\alpha\left(R(s)+\gamma\max_{a^{'}\in A(s^{'})}Q(s^{'},a^{'})-Q(s,a)\right)\\ Q(s,a)\leftarrow Q(s,a)+\Delta Q(s,a)\end{gathered}\tag{1}$$

其中，$A(s)$ 为在状态 s 可以采取的动作集合，$s^{'}$ 为在状态 s 下执行动作 a 后到达的下一状态，奖赏折扣因子 $\gamma\in[0,1]$，表示当前动作对未来奖赏的影响程度，当 $\gamma=0$ 时，智能体只考虑立即奖赏，而 $\gamma=1$ 则表示智能体赋予长期奖赏和立即奖赏同样的重要性。学习率 $\alpha\in(0,1)$为 Q 值的更新速度，较大的学习率可以提高算法的收敛速度，较小的学习率可以提升算法的稳定性[17]。

在学习过程中，智能体通过不断尝试获取经验，经历一段时间的学习之后，智能体得到具有一定性能的策略，为了测试是否还有改进策略的可能，需要对已有策略建议之外的部分进行探索，然而探索过程具有牺牲性能的风险，若仅利用已有知识则无法进一步改进策略，ϵ-贪心探索是一种用于平衡探索和利用的方法[11]，即智能体在每个决策时刻以 ϵ 的概率随机选择一个动作，以 $1-\epsilon$ 的概率选择 Q 值最高的动作。在智能体进行了足够多的探索之后，则不再探索，按照贪婪控制准则执行最优控制策略。

1.3 Q-Learning 算法的改进

表格型 Q-Learning 算法是针对离散状态和动作空间的马尔科夫决策过程，但实际工程中的许多优化决策问题都具有大规模或连续的状态和行为空间，因此表格型 Q-Learning 会面临维数灾难[6]。为了解决这个问题，通常将泛化方法例如神经网络用于强化学习的 Q 值函数逼近。常见的用于 Q 值估计的神经网络结构如图 2 所示。

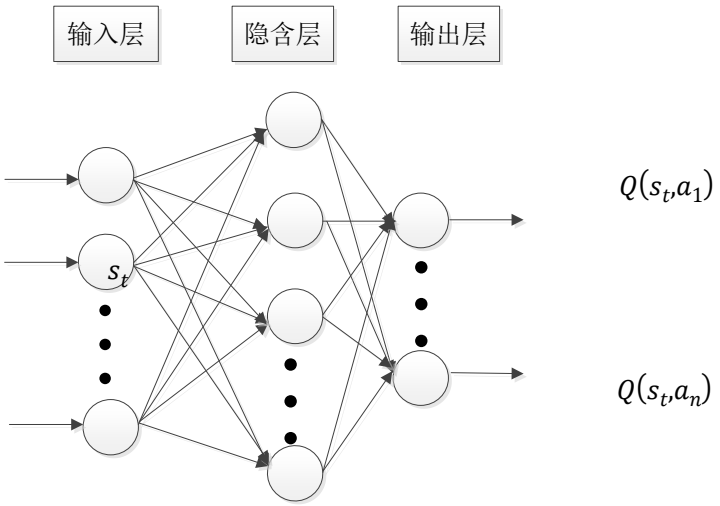

图 2 一种用于 Q 值估计的神经网络

Fig 2 A Neural Network for Q-value Approximation

训练过程通常采用经验回放的方式[18]，将最近的 N 条数据$(s_t,a_t,r_{t+1},s_{t+1})$存储在回放存储器中，在训练过程中，每次随机挑选 M（M<N）条数据，被选取的数据被称作微批数据。通过式（2）计算微批数据中每个样本的期望输出值，最后使用微批数据中的 s_t 和计算得到的期望输出值更新神经网络的权重。

$$label(s_t)=\begin{cases}Q(s_t,a)+\Delta Q(s_t,a),\ if\ a=a_t\\ Q(s_t,a),\ if\ a\neq a_t\end{cases} \tag{2}$$

2. 基于强化学习的空调系统运行

基于强化学习的空调系统运行将环境视为“黑箱”，通过收集到的信号反馈不断改进控制策略，降低了对先验知识（如建筑热惰性，设备特性等）的依赖。然而将强化学习应用在空调系统运行优化时需要注意两个问题，其一如何实现有效的状态特征表达，当状态所包含的信息越完整，则智能体所学习到的策略越接近最优策略。针对建筑运行问题，通常需要完整考虑室外天气状况，室内环境状态和空调系统运行情况。其二强化学习通过不断尝试来改进策略，本文提出通过确立规则对探索空间进行缩减，从而确保智能体在探索过程中避免损害设备，同时尽可能的满足被控区域舒适度的要求。在考虑以上两个问题的基础上并结合改进的 Q-Learning 算法，本文提出的基于强化学习的空调系统运行流程如表 1 所示，接下来通过仿真案例对强化学习控制算法进行验证和分析。

表 1 基于强化学习的空调系统运行流程

Table 1 The Process of HVAC System Operation Based on Reinforcement Learning

准备阶段：

1) 将空调系统运行优化问题建模为马尔科夫决策过程。
2) 定义规则缩减探索空间。
3) 初始化用于 Q 值估计的神经网络。

探索阶段：

1) 获取当前状态量 s。
2) 重复以下过程，并且每隔固定时间步长从回放储存器中随机抽取 M 条数据更新用于 Q 值估计的神经网络的权值：
   a) 计算状态为 s 下不同动作的 Q 值。
   b) 按照ϵ-贪心探索选择动作 a，执行动作 a 直至下一决策时刻。
   c) 测量和降维获得当前时刻状态量 $s^{'}$，并计算此控制时间步长内奖赏 r，将

(s,a,r,s')存入回放存储器。

d) 将 s' 赋值给 s 。

运行阶段：

1) 获取当前状态量 s。

2) 重复以下过程，并且每隔固定时间步长从回放储存器中随机抽取 M 条数据更新用于 Q 值估计的神经网络的权值：

a) 计算状态为 s 下不同动作的 Q 值。

b) 选取 Q 值最高的动作 a，执行动作 a 直至下一决策时刻。

c) 测量和降维获得当前时刻状态量 s'，并计算此控制时间步长内奖赏 r，将 (s,a,r,s')存入回放存储器。

d) 将 s' 赋值给 s 。

3. 仿真实验

3.1 仿真案例简介

本文选取的研究对象为一单层办公建筑，总空调面积为 475 m^2，空调系统为变风量空调系统，强化学习控制器负责调节被控区域送风量，满足室内舒适度需求的同时尽可能的降低能耗消费。

通过 TRNSYS 与 MATLAB 进行联合仿真，TRNSYS 负责建筑及空调系统仿真，强化学习控制策略通过 MATLAB 实现，二者通过 TRNSYS 组件 Type155 进行通信。围护结构热工参数的设置参照 GB50189－2005《公共建筑节能设计标准》，具体的热工参数包括：外墙为 0.888 $W/m^2·K$，屋顶为 0.638 $W/m^2·K$，外窗为 2.73 $W/m^2·K$。内热源部分设置人员密度为 0.25 p/m^2，设备密度为 20 W/m^2，照明密度为 11 W/m^2。建筑运行时间为制冷季 6.1 - 9.20 间每天 8:00 - 18:00。设定室内送风温度为 15℃，人均新风量为 30 $m^3/(h·p)$。建筑依靠额定 COP 为 5 的离心式制冷机组提供冷量，冷冻水泵为变频水泵，冷却水泵为定频水泵。在 TRNSYS 中建立好建筑空调系统仿真模型，如图 3 所示。

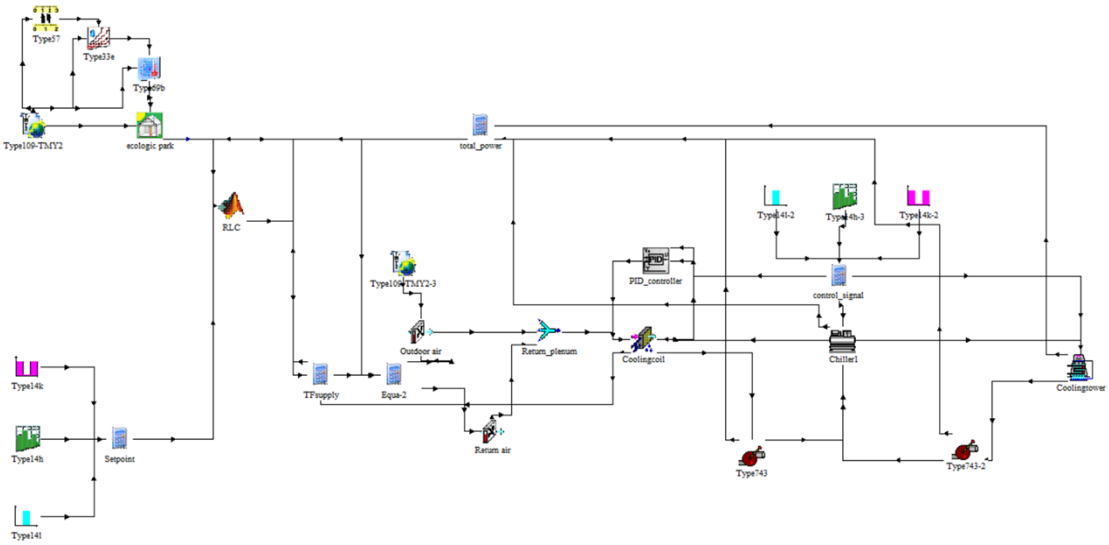

图 3 TRNSYS 中建筑空调系统仿真模型

Fig 3 The Simulation Model of Building and HVAC System in TRNSYS

在准备阶段，第一步先建立此运行优化问题的马尔科夫决策过程模型，首先定义动作，即被控区域送风量。确定最小送风量为室内人员所需新风量，最大送风量确定为满足室内 6 次换气次数，将动作区间离散为四个水平$\{a_1,a_2,a_3,a_4\}$，设置控制时间步长为 0.2h（12 分钟）。接下来定义状态，本案例选取时间，室外天气情况，被控区域温度作为状态。其中时间用于反映如室内热扰等与时间相关信息。室外天气状况选取当前时刻室外温

度，当前室外太阳辐射和预测的未来一小时的温度变化值。被控区域温度应包含室内空气温度和壁面温度两部分，其中壁面温度用于反映建筑蓄热情况，然而在实际工程中通常无法测量，本文采用过去五个时间步长的室内空气温度的均值来近似反映建筑蓄热情况。最后定义奖赏函数，奖赏函数如式（3）所示分为两个部分，一部分为此控制时间步长内的能耗消费，其中能耗为冷机，冷冻水泵，冷却水泵，冷却塔能耗之和，能耗费用按照上海市分时电价计算，另一部分为室内温度超出舒适度范围时的惩罚项，按照 GB50736-2012《民用建筑供暖通风与空气调节设计规范》，夏季室内温度应保持在 24-28℃之间，故取温度上限为 28℃，下限为 24℃。第二步定义规则对探索空间进行缩减，本案例中我们定义两条规则用于确保系统运行满足舒适度约束，当室内温度大于 27.5℃时，按照最大送风量送风，避免温度进一步上升，当室内温度小于 24.5℃时则按照最小送风量送风，降低能耗开销。第三步初始化用于 Q 值估计的神经网络，本文采用的神经网络包括三个隐藏层，每个隐藏层的神经元数分别为 32,64,32。前两个隐藏层采用修正线性单元作为激活函数，最后一个隐藏层采用线性单元作为激活函数。

$$r_t = -\,cost\left(s_{t-1}, a_{t-1}\right) + penalty(T_t)$$

$$penalty(T_t) = \begin{cases} -\beta\left(T_t - T_{lower\ bound}\right)^2, if\ T_t < T_{lower\ bound} \\ -\beta\left(T_t - T_{upper\ bound}\right)^2, if\ T_t > T_{upper\ bound} \\ 0, if\ T_{lower\ bound} \le T_t \le T_{upper\ bound} \end{cases}$$

在应用阶段，设置第一年运行为探索阶段，即采用 ϵ-贪心探索选择动作。从第二年起即转为运行阶段。本案例用上海市典型气象年数据模拟气象参数，温度预测值采用真实温度加入 $N(0, 0.2)$ 正态分布随机误差计算得出。强化学习控制器相关参数见表 2。

表 2 强化学习控制器相关参数

Table 2 Parameters Setting for Reinforcement Learning Controller

参数	值
学习率 α	0.05
奖赏折扣因子 γ	0.1
回放存储器容量 M	1200
微批数据容量 N	50
探索比例 ϵ	0.1

3.2 仿真结果讨论与分析

图 4 为在运行阶段室内温度变化情况，可以看出在室外温度不断变化的过程中，强化学习控制器可以保证在绝大多数时间内满足舒适度的需求。

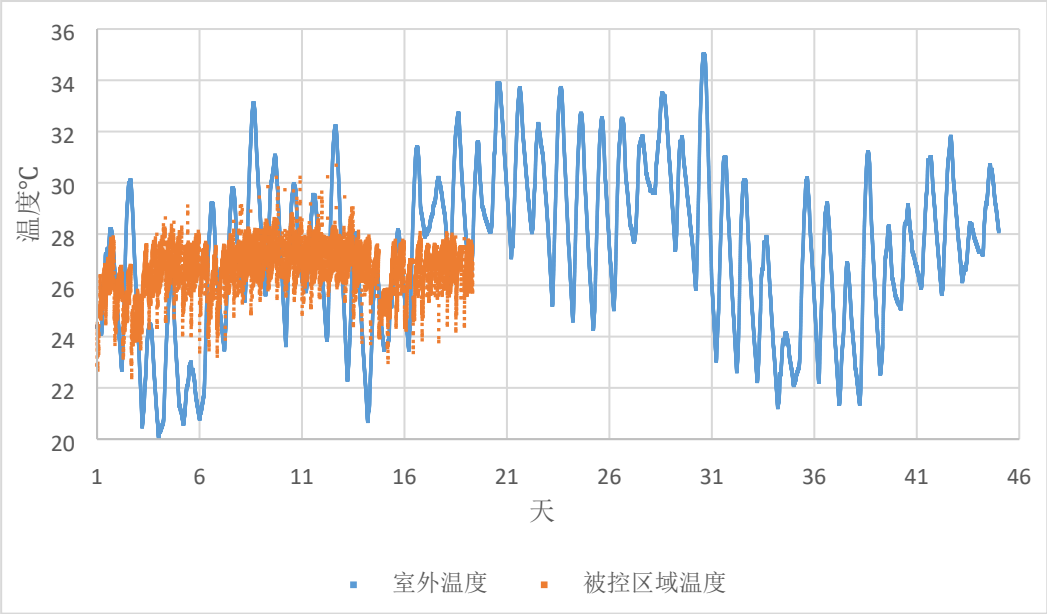

图 4 被控区域温度变化情况

Fig 4 Temperature of the Conditioned Zone

为了量化强化学习控制器的控制效果，基准控制策略选取类似于启停控制的基于规则的控制策略（RBC）和固定温度设定值的 PID 控制器作为对比对象，RBC 具体设置为当被控区域温度超过 27.5℃时，保持最大送风量送风，当被控区域温度低于 24.5℃时，则保持最小送风量送风，PID 控制器的温度设定值为 26℃，通过对送风量的调整维持室内温度的稳定。从能耗消费和被控区域舒适度情况对三种控制策略进行对比，结果如图 5 所示，

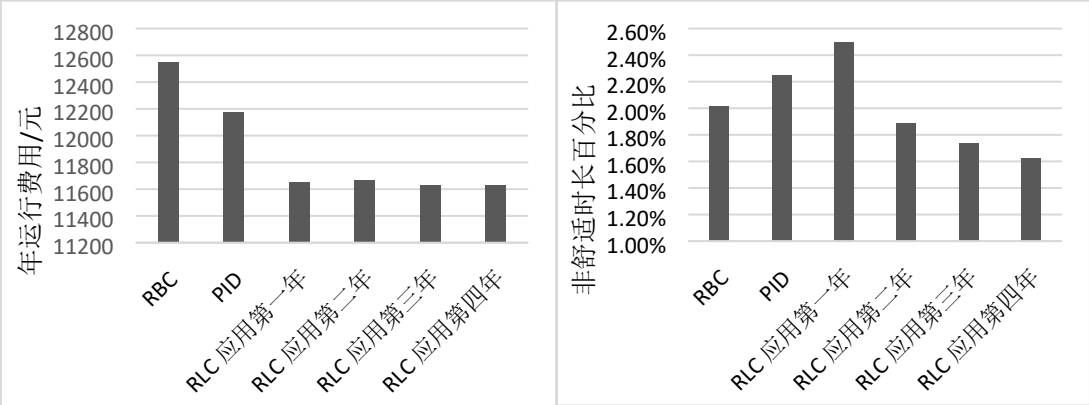

图 5 控制效果对比

Fig 5 Controller Performance Comparision

从结果对比可以看出，在探索阶段，由于强化学习控制器（RLC）通过不断尝试更新策略，导致非舒适时长百分比有所增加。但在运行阶段，利用已经学习到的经验，强化学习控制器的表现均优于基准策略。与 RBC 和 PID 控制相对比，在小幅提升室内舒适度水平的同时，分别将运行费用降低超过 7%和 4.6%。并且随着运行时间的推进，强化学习控制器仍可以不断改进控制策略。

在应用过程中，强化学习控制器利用反馈回来的信号不断改进控制策略，自适应性强，降低了对先验知识的依赖，与基于模型的控制方法相比，无须进行滚动优化，故计算速度方面具有优势。但是强化学习也存在学习效率不高的问题，然而建筑运行具有长生命周期的特点，通过大量的监测数据反馈，强化学习控制器可以提炼运行经验，实现空调系统精细化运行，提高建筑能效。

4. 结论

近年来楼宇自动化系统的普及和计算机成本的下降，探索数据驱动的建筑运行优化具有十分重要的研究意义。强化学习控制器利用监测数据的反馈不断改进控制策略，具有免模型，自适应性强，目标导向等特点。本文将强化学习应用在空调系统运行优化中，将应用过程分为准备阶段，探索阶段和运行阶段，描述了各阶段完整的算法流程，算法具备一定的通用性。通过仿真实验验证强化学习控制器的控制效果，在满足被控区域舒适度要求的前提下，可有效地降低运行费用，提高建筑能效。

本文以单区域送风量调节作为验证案例进行分析，如何将强化学习控制器有效的应用在多区域控制，增强其可扩展性是下一步的研究方向，同时考虑到建筑内扰等不确定因素，未来的工作也包括通过实验对强化学习控制策略的效果进行验证。

参考文献：

[1] 刘羽岱, 潘毅群, 黄治钟. 基于模型预测控制的空调系统运行优化仿真研究[C]// 上海市制冷学会 2013 年学术年会. 2013.
[2] Dong B, Lam K P. A real-time model predictive control for building heating and cooling systems based on the occupancy behavior pattern detection and local weather forecasting[J]. Building Simulation, 2014, 7(1):89-106.
[3] Killian M, Kozek M. Ten questions concerning model predictive control for energy efficient buildings[J]. Building & Environment, 2016, 105:403-412.
[4] Široký J, Oldewurtel F, Cigler J, et al. Experimental analysis of model predictive control for an energy efficient building heating system[J]. Applied Energy, 2011, 88(9):3079-3087.
[5] Siddharth Goyal, Herbert A. Ingley, Prabir Barooah. Occupancy-based zone-climate control for energy-efficient buildings: Complexity vs. performance ☆[J]. Applied Energy, 2013, 106(11):209-221.
[6] 陈学松, 杨宜民. 强化学习研究综述[J]. 计算机应用研究, 2010, 27(8):2834-2838.
[7] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning.[J]. Nature, 2015, 518(7540):529.
[8] Liu S, Henze G P. Experimental analysis of simulated reinforcement learning control for active and passive building thermal storage inventory : Part 1. Theoretical foundation[J]. Energy & Buildings, 2006, 38(2):142-147.
[9] Dalamagkidis K, Kolokotsa D, Kalaitzakis K, et al. Reinforcement learning for energy conservation and comfort in buildings[J]. Building & Environment, 2007, 42(7):2686-2698.
[10] Li B, Xia L. A multi-grid reinforcement learning method for energy conservation and comfort of HVAC in buildings[C]// IEEE International Conference on Automation Science and Engineering. IEEE, 2015:444-449.
[11] Yang L, Nagy Z, Goffin P, et al. Reinforcement learning for optimal control of low exergy buildings[J]. Applied Energy, 2015, 156:577-586.
[12] Ruelens F, Iacovella S, Claessens B J, et al. Learning Agent for a Heat-Pump Thermostat With a Set-Back Strategy Using Model-Free Reinforcement Learning[J]. Energies, 2015, 8(8):8300-8318.
[13] Costanzo G T, Iacovella S, Ruelens F, et al. Experimental analysis of data-driven control for a building heating system[J]. Sustainable Energy Grids & Networks, 2016, 6:81-90.
[14] Wei T, Wang Y, Zhu Q. Deep Reinforcement Learning for Building HVAC Control[C]// Design Automation Conference. ACM, 2017:22.
[15] 高阳, 陈世福, 陆鑫. 强化学习研究综述[J]. 自动化学报, 2004, 30(1):86-100.
[16] 周志华. 机器学习[M]. 清华大学出版社, 2016 年 1 月.
[17] Sutton R S, Barto A G. Reinforcement Learning: An Introduction[M]. MIT Press, 1998.
[18] Wawrzyński P, Tanwani A K. Autonomous reinforcement learning with experience replay[J]. Neural Networks the Official Journal of the International Neural Network Society, 2013, 41(5):156-167.

作者基本情况表

通讯地址	上海市嘉定区曹安公路 4800 号同济大学嘉定校区
邮编	201801
联系电话	15201908852
电子邮箱	weitong_wang@126.com