基于 DDPG 模型的建筑能耗控制方法

周 鑫 陈建平 傅启明*
(苏州科技大学电子与信息工程学院 江苏 苏州 215009)
(江苏省建筑智慧节能重点实验室 江苏 苏州 215009)

摘 要 针对居民建筑能耗逐渐增加、传统控制方法效率低下的问题,提出一种基于深度确定性策略梯度的建筑能耗控制方法。该方法利用深度强化学习模型,将建筑电力使用问题建模为强化学习的控制问题,解决负荷降低和成本最小化的问题。根据某开源数据库中居民的能耗使用数据,结合深度 Q 网络、确定性策略梯度和深度确定性策略梯度算法进行实验验证。实验结果表明,该方法能够有效降低负荷峰值与电力能源使用成本,实现建筑节能的目的。

关键词 深度强化学习 深度确定性策略梯度 策略优化 建筑节能

中图分类号 TP391 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2023.02.007

A BUILDING ENERGY CONSUMPTION CONTROL METHOD BASED ON DDPG MODEL

Zhou Xin Chen Jianping Fu Qiming*
(*School of Electronics and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, Jiangsu, China*)
(*Jiangsu Province Key Laboratory of Intelligent Building Energy Efficiency, Suzhou 215009, Jiangsu, China*)

Abstract In order to solve the problem of the increasing energy consumption of residential buildings and low efficiency of traditional control methods, this paper proposes a building energy consumption control method based on deep deterministic policy gradient. This method used the deep reinforcement learning model to model the problem of building power use as the control problem of reinforcement learning to solve the problem of load reduction and cost minimization. Based on the energy consumption data of residents in an open source database, DQN, DPG and DDPG algorithms were combined for experimental verification. The experimental results show that the method can effectively reduce the peak load and power energy use cost and achieve the purpose of building energy conservation.

Keywords Deep reinforcement learning Deep deterministic policy gradient Policy optimization Building energy efficiency

0 引 言

为了应对日益增加的建筑能耗问题,单纯依靠政策宣传节能等粗放的手段难以有效地解决节能问题。新一代的人工智能技术,已成为智能电网背景下建筑节能的又一大研究趋势。但目前的人工智能方法在建筑节能领域尚处在初级阶段,探索如何使用新技术实现建筑节能,是一个亟待解决的问题。

深度强化学习(Deep Reinforcement Learning,DRL)[1]是人工智能方法中的一种,因其在多个领域的应用价值,使其成为主要的研究方向。深度强化学习是由具有决策能力强化学习(Reinforcement Learning,RL)[2]与具有特征提取能力的深度学习(Deep Learning,DL)[3]结合而成,具有很强的通用性[4]。在之后的研究中,深度强化学习在各个领域被广泛运用,如游戏[5]、机器人控制[6-7]等。

Mnih 等[8-9]将神经网络与 RL 算法结合,提出了

收稿日期:2020-07-01。国家自然科学基金项目(61772357,61672371);江苏省研究生科研创新计划项目(SJCX18-0881)。
周鑫,硕士生,主研领域:深度强化学习。**陈建平**,教授。**傅启明**,讲师。

深度 Q 网络模型(Deep Q-Network,DQN),用于处理视觉感知的控制任务。之后,DQN 算法出现了多种改进版本[10],包括对算法的改进[11]、神经网络模型的改进[12]、学习机制的改进[13-14],以及新的 RL 算法的改进[15]。然而,这些算法适用于离散动作空间的 RL 任务,在连续动作空间中,基于确定性策略梯度(Deterministic Policy Gradient,DPG)[16]的算法可以获得更好的效果。因此,Deep Mind 团队提出了深度确定性策略梯度算法(Deep Deterministic Policy Gradient,DDPG)[17],结合深度神经网络来处理大规模状态空间的问题,并在该算法的基础上提出了多智能体的 DDPG 算法[18],取得了显著效果。陈建平等[19]提出一种增强型深度确定性策略梯度算法,加快了算法的收敛速度。何明等[20]提出了基于多智能体 DDPG 算法的经验优先抽取机制,提高了算法的训练速度。邹长杰等[21]提出了基于多智能体 DDPG 模型的分组学习策略,提高了多智能体的学习效率。

综上,针对现有建筑节能方法比较粗放的问题,基于深度强化学习的理论,提出更加智能化的控制策略,用于解决建筑节能问题。本文提出一种基于 DDPG 算法的建筑能耗策略优化方法,利用强化学习构建成本最小化与电力负荷峰值降低的关系模型,解决连续动作空间下的策略优化问题。通过对开源的建筑能耗使用数据进行实验验证,该方法能够有效降低电力负荷与使用成本,最终实现建筑节能。

1　相关理论

1.1　马尔可夫决策过程

满足马尔可夫性质的强化学习任务被称为马尔可夫决策过程(Markov Decision Process,MDP)或 MDP,因此,利用马尔可夫决策过程对强化学习进行建模,可以有效完成序贯决策任务。通常,MDP 可以用一个四元组$\{S,A,T,R\}$表示,其中:S 是所有环境状态的集合;A 是 agent 可执行动作的集合;T 是状态转移函数;R 是奖赏函数。对一个 MDP 问题,在任意时刻 t,其状态为 $S_t \in S$,选择并执行动作 $a_t \in A$,获得立即奖赏 $r(s_t,a_t) \in \mathbf{R}$,通常可以简写为 r_t,且转移到下一状态 $s_{t+1} \in S$,状态转移 $T(s_t,a_t,s_{t+1})$ 的概率为 $Pr(s_t,a_t,s_{t+1})$。

强化学习中,策略 π 是指在状态 s 下采取动作 a 的概率,表示为 $\pi(s,a)$。判断某一策略 π 的优劣程度,基本上是通过计算估计动作值函数的值进行判断。其中,估计动作值函数根据未来累积奖赏进行计算评估,定义如下:

$$Q_\pi(s,a)=E_\pi\{R_t \mid s_t=s,a_t=a\}=\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1} \mid s_t=s,a_t=a\right\} \tag{1}$$

式中:γ 是折扣率,决定着未来奖赏的当前价值。如果选择的策略是最优策略,则用最优动作值函数 Q^* 进行表示,定义如式(2)所示。

$$Q^*(s_t,a_t)=E\{r_{t+1}+\gamma\max_{a_{t+1}}Q^*(s_{t+1},a_{t+1}) \mid s_t=s,a_t=a\}=\sum_{s_{t+1}}T^a_{s_ts_{t+1}}\left[R^a_{s_ts_{t+1}}+\gamma\max_{a_{t+1}}Q^*(s_{t+1},a_{t+1})\right] \tag{2}$$

1.2　DDPG 方法

DDPG 算法融合了 DPG 算法与 DQN 算法的优点,利用神经网络来模拟策略函数和 Q 值函数,通过训练,能够提高非线性模拟函数的准确性和高效性。此外,利用 DPG 算法中行动者评论家方法(Action-Critic,AC)的优势,结合 DQN 算法中的经验池和双网络结构,以及目标网络参数的"软更新"方式,提高神经网络的学习效率,在连续状态空间问题中取得了较好的实验效果。其中,DPG 算法利用近似函数 $\mu(s \mid \theta^\mu)$ 表示动作选择,其梯度定义如下:

$$\nabla_\theta J(\mu_\theta)=\int_S \rho^\mu(s)\nabla_\theta\mu_\theta(s)\nabla_a Q^\mu(s,a)\mid_{a=\mu_\theta(s)}\mathrm{d}s \tag{3}$$

在随机策略中,状态和动作的值会影响策略梯度的计算,而在确定策略中,只有状态值才会影响策略梯度。相较而言,DPG 算法在达到收敛条件时所需要的样本较少。DDPG 算法利用式(3)更新策略网络参数,并通过式(4)对网络参数进行更新。但是,如果直接使用式(4)进行更新会导致收敛不稳定,因为在更新 $Q(s,a \mid \theta^Q)$ 的过程中,其目标值也在同步计算,即式(5)中的 y_t。

$$L(\theta^Q)=E_{s_t\sim\rho^\pi,a_t\sim\pi,r_t\sim E}[(Q(s_t,a_t \mid \theta^Q)-y_t)^2] \tag{4}$$

$$y_t=r(s_t,a_t)+rQ(s_{t+1},\mu(s_{t+1}) \mid \theta^Q) \tag{5}$$

针对这个问题,DDPG 采用"软更新"的方式,即创建新的 AC 网络($Q'(s,a \mid \theta^{Q'})$,$\mu'(s \mid \theta^{\mu'})$)用于目标参数的更新。目标策略网络和目标值网络中参数的更新规则为 $\theta^{\mu'}\leftarrow\alpha\theta^\mu+(1-\alpha)\theta^{\mu'}$,$\theta^{Q'}\leftarrow\alpha\theta^Q+(1-\alpha)\theta^{Q'}$,$\alpha \ll 1$,该方法可以降低目标值的更新速度,从而提高算法的收敛稳定性。此外,DDPG 算法引入经验回放机制打破样本之间的相关性,以提高算法的学习效率。不仅如此,DDPG 算法还通过引入随机噪声 N 来完成策略探索,使动作的选择具有一定的随机性,从而在一定程度上提高探索环境的效率,具体如式(6)所示。

$$\mu'(s_t) = \mu(s_t \mid \theta_t^{\mu} + N) \tag{6}$$

1.3 Softmax 策略选择方法

判断 agent 是否选择最优动作之后,需要对策略选择方法加以改进,或者选择合适的动作选择策略。UCB 策略、ε-greedy 方法、与 Softmax 都是强化学习中比较重要的动作选择策略。前两种策略缺陷都在于不能够有效地优化动作的选择概率。因此,一个比较有效的方法就是将选择动作的概率更改为估计值的一个分级函数,即将最高的选择概率分配给贪心动作,但是,除此之外的其他动作则根据其值的估计进行比较并分配权重,这称为软最大化动作选择规则。

动作选择概率的一般表达式可以写为它在某次操作选择动作 a_j 的概率:

$$\mathbf{P}(a_j) = \frac{e^{Q(a_j)/\tau}}{\sum_{i=1}^{N} e^{Q(a_i)/\tau}} \tag{7}$$

式中:τ 指代的是温度系数,当 $\tau \to 0$ 时,软最大化动作选择方法就变得与贪心方法一样。

2 基于 DDPG 模型的能耗控制算法

2.1 问题建模

本文旨在降低电网负载峰值,最大程度地降低能源成本。令 B 表示建筑物集合,例如 $B_i \in B, \forall i \in \mathbf{N}$ 代表建筑物的索引。建筑物总能耗 E_i 是所有发电 P^+ 和特定时间间隔 Δt 内的能耗之和。其中,基于建筑物中存在的电气设备的可移动性,区分了弹性能源消耗 P_d^-,例如电气设备 $d \in \{1,2,\cdots,m_i\}$ 的消耗功率及固定能源消耗 P^-。

(1) 降低成本问题。本节假设 B 空间上有两个价格分量,λ_t^- 是在时间间隔 t 内电力公司设置的价格,而 λ_t^+ 代表电力公司在时间间隔 t 从终端用户购买能源的价格。因此,在优化的时间范围 T 内,时间 t 与用户 i 相关联的最优成本可以计算为:

$$\min \sum_{t=1}^{T}\left(\lambda_t^+ \sum_{i=1}^{n} P_{i,t}^+ - \lambda_t^- \sum_{i=1}^{n}\left(P_{i,t}^- + \sum_{d=1}^{m_i} a_{i,d,t} P_{i,d,t}^-\right)\right) \tag{8}$$

$$\text{s.t.} \quad \sum_{t=1}^{T} P_i^- \Delta t = E_i \quad \forall i \in \mathbf{N}, \forall t \in \mathbf{N}$$

$$\sum_{t=1}^{T} P_d^- \Delta t = E_d \quad \forall d \in \mathbf{N}, \forall t \in \mathbf{N}$$

$$a_{i,d,t} = \{1,0\} \quad \forall a \in \mathrm{A}, \forall i \in \mathbf{N}, \forall d \in \mathbf{N}, \forall t \in \mathbf{N}$$

$$P_{i,t}^+, P_{i,t}^-, P_{i,d,t}^- \geqslant 0 \quad \forall t = [1:T] \in \mathbf{N}$$

$$\lambda_t^+, \lambda_t^- \geqslant 0 \quad \forall t = [1:T] \in \mathbf{N}$$

式中:如果电气设备在特定时间打开,则 $a_{i,d,t} = 1$,否则为 0。此外,本节提出的方法中,$a_{i,d,t}$ 等同于对动作的估计。

(2) 降低负荷峰值问题。对于太阳能发电和能源消耗,在价格不变的特殊情况下,即当 $\lambda_t^+ = \lambda_t^-$ 时,将成本最小化问题转为峰值降低的问题,定义如下:

$$\min \sum_{t=1}^{T}\left(\sum_{i=1}^{n} P_{i,t}^+ - \sum_{i=1}^{n}\left(P_{i,t}^- + \sum_{d=1}^{m_i} a_{i,d,t} P_{i,d,t}^-\right)\right) \tag{9}$$

因此,式(8)的约束条件将同时对两个问题都有效。但是,基于不同类型的电气设备之间的差异,约束条件的整个范围会变大,如下所述。

电气设备的约束条件:假设三种类型的消耗曲线。首先,考虑时间缩放负载。对此,本节的分析仅限于空调负荷(d_{AC}),作为每栋建筑物中较大的一组电气设备的代表,可以在优化范围内将开关次数约束为有限时间,例如灯、电视等电气设备。先前的研究表明,短期时间内减少空调的使用对最终用户舒适度的影响可忽略不计。其次,本章包括时移负载,也称为可延迟负载,即能够实现用电时间的转移,它必须在给定的时间间隔内消耗最少的电量。其中,本节将洗碗机(d_{DW})建模为不间断负载,该模型需要多个连续的时间步长。最后,电动汽车(d_{EV})被建模为可移动负载。就本节而言,根据随时间变化的设备约束 $a_{d,t}$ 的定义,提出以下假设:

假设 1:对于所有 d,在 P_d^- 时间缩放负载下,在优化范围内存在 $\delta_d \in \mathbf{R}_+$,使得:

$$\begin{cases} \sum_t P_d^- \leqslant \delta_d & p(P_d^- = 0 \mid t) \in (0,1] \\ \sum_t P_d^- = \delta_d & p(P_d^- = 0 \mid t) = 0 \end{cases} \tag{10}$$

式中:$p(P_d^- = 0 \mid t)$ 是电子设备 d 在任意时间 $t = [1:T] \in \mathbf{N}$ 内,在 t 时刻处于活动状态的概率。

假设 2:对于所有电子设备 d,在 P_d^- 时移负载下,存在 δ_d 常数,使得对于所有时间 $t = [1:T] \in \mathbf{N}$,有 $\sum_t P_d = \delta_d$。

条件 1:在本节中,P^+ 被认为是不可忽视的部分。

条件 2:所有电动汽车 d_{EV} 及其相关的消耗 P_d^- 均被视为在假设 1 和假设 2 的前提下,其工作随时间扩展和移动的负载。

在本节中,使用 DRL 方法作为建筑能耗控制的优化方法,以便在不同复杂程度上执行最佳建筑能耗控制策略。DRL 可以通过自动提取模式,例如能源消耗的数据,来学习比标准 RL 更好的行为策略。简而言之,可以从总体框架的角度将 DNN 方法表示为在给定输入分布上具有良好泛化能力的黑匣子模型,如式

(10)所示。

$$Input \xrightarrow{\text{data}} DNN_{(k)} \xrightarrow{\text{data estimation}} Output \tag{11}$$

2.2　奖赏函数构造

针对本节所解决的多目标优化问题,在一天结束时计算一个准确的奖赏函数,而不是在一天的每个时间步长都计算奖赏函数。因此,推导出了一个简单的包含三个奖赏组成的多任务联合奖赏:

1) 对于所有 $\psi=(s_t,a_t,r_t)$,奖赏向量将能够控制三种类型的能源消费行为,从而控制住户总的可移动和扩展负载 $\sum_{d=1}^{m} a_t P_{d,t}^{-}$,因此,应使用差异化的奖赏:

$$\begin{aligned}
r_{a_1} &= \begin{cases} -n_{a_1^+} & n_{a_1^+}>10 \\ \zeta_1 & n_{a_1^+}\in[1,10] \\ \zeta_2 & n_{a_1^+}<1 \end{cases} \\
r_{a_2} &= \begin{cases} -4\,|\,n_{a_2^t}-n_{a_2^+}\,| & n_{a_2^+}\neq n_{a_2^t},\ \forall\, n_{a_2^t}\in\mathbf{N} \\ n_{a_2^+} & n_{a_2^+}=n_{a_2^t},\ \forall\, n_{a_2^t}\in\mathbf{N} \end{cases} \\
r_{a_3} &= \begin{cases} -n_{a_3^+} & n_{a_3^+}>2 \\ \zeta_1 & n_{a_3^+}\in[1,2] \\ \zeta_2 & n_{a_3^+}<1 \end{cases}
\end{aligned} \tag{12}$$

式中:$n_{a_1^+}$、$n_{a_2^+}$和 $n_{a_3^+}$代表执行与该设备相对应的动作的次数,而 $n_{a_2^t}$是电动汽车每天的目标负荷;ζ_1和 ζ_2系数的选择是基于反复实验的过程,获得的值为 $\zeta_1=40$,$\zeta_2=-50$。

2) 对式(9)中定义的总能耗进行如下设置:

$$r=\begin{cases} -3\zeta_2+4[\max(P^-)-\max(\tilde{P}^-)] & \\ \qquad\qquad \max(\tilde{P}^-)<\max(P^-) & \\ -3\zeta_1-1 \qquad 其他 & \end{cases} \tag{13}$$

此外,根据式(8),当有更多的能源产生时,通过时间转移能源消耗量:

$$r=\begin{cases} \dfrac{\zeta_1}{2}-|\min(\tilde{P}^-)| & \tilde{P}^->0 \\ -\dfrac{\zeta_1}{2} & 其他 \end{cases} \tag{14}$$

空调的控制由假设 2 以及式(15)给出:

$$r=\begin{cases} \dfrac{\zeta_1}{8}+2[\max(\tilde{P}_{AC}^-)-\max(P_{AC}^-)] & \tilde{P}^-<0 \\ -\dfrac{\zeta_1}{10} & 其他 \end{cases} \tag{15}$$

3) 总成本 C 的计算如下:

$$r=\begin{cases} 5\,|\,\tilde{C}-C\,| & \tilde{C}<C \\ -3\zeta_1-1 & 其他 \end{cases} \tag{16}$$

故本节用奖赏函数的 1)和 2)求解式(9),用奖赏函数的 1)和 3)求解式(8)。

联合奖赏函数可以被简易地概括为执行任意数量的任务。然而,在式(12)中考虑的 $n_{a_1^+}$和 $n_{a_3^+}$的范围间隔,以及式(12)至式(16)中考虑的正负系数(即 ζ_1和 ζ_2)都取决于实际的情况。同样,在式(12)中,如果放松舒适度的限制,$n_{a_1^+}$和 $n_{a_3^+}$的范围就会扩大。

2.3　基于 DDPG 模型的策略选择算法

之前有将基于值函数差异的探索与 Softmax 动作选择结合在一起,利用学习过程中产生的值差异来衡量 agent 对环境的不确定性,以适应在线探索参数。事实证明,这种方法可以极大地优化多臂赌博机问题的求解。但是,这种探索策略的缺点是必须记录每种状态的探索参数,在遇到大规模连续状态或动作空间时效率低下。因此,本节提出一种基于 Softmax 方法的策略选择方法,即 S-DDPG,该方法根据 agent 与环境之间的交互过程中的动作值和平均动作值动态地调整探索参数。

策略选择方法的核心思想是根据 agent 达到目标状态的成功数量和成功率来鼓励探索。一方面,当 agent 获得越来越高的奖赏时,策略应该更多地被利用。另一方面,当 agent 由于环境变化而停止获得奖赏时,应该再次鼓励探索。因此,策略搜索算法如式(17)所示。

$$\pi(s)=\begin{cases} \text{Softmax } action\ a & \xi<\varepsilon \\ \arg\max\limits_{a\in A} Q(s,a) & 其他 \end{cases} \tag{17}$$

式中:ε 的取值来源于 ε-greedy 方法。

该方法的总体框架如图 1 所示,基于 DDPG 模型的策略选择算法如算法 1 所示。

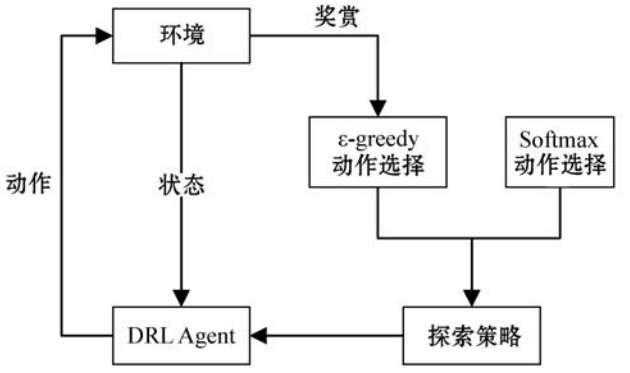

图 1　策略搜索方法框架

算法 1　基于 DDPG 模型的策略选择方法

输入:状态信息数据。

输出:动作的概率。

1) 初始化 Actor、Critic 网络模型的超参数(α,γ,ζ);网络权重 θ

2) for *episode* = 1 to M do:初随机状态 s

3) for $t=1$,T:actor 根据策略方法选择动作;执行动作,返回奖赏随机状态 r 以及下一状态 s';并将状态转移信息存入经

验池;从经验池中随机选择数据进行训练

4) 通过损失函数更新 Critic 网络;使用样本的策略梯度更新 Actor 网络

5) end for

6) end for

3 实验及结果分析

3.1 网络模型

为了在离散和连续动作空间下令 DQN、DPG 和 DDPG,以及 S-DDPG 进行公平的比较,模型所使用的深度神经网络的架构相似,并且具有以下特征:每个强化学习状态由一个时间窗口的两个连续时间步长给出。因此,在峰值降低问题的情况下,输入层具有 11 个神经元,即时间步长 t,以及在 $t-1$ 到 t 时刻的基本负荷、光伏发电、空调状态、电动汽车和洗碗机的状态。需要注意的是,除了固定的基本负荷和发电量外,其他状态分量不是由智能电表测量的初始值直接给出,而是通过学习过程中获得的值动态调整。对于成本最小化的问题,输入层有一个额外的神经元,用于对分时电价进行编码。此外,该网络具有三层隐藏的神经元层,各层都包含 100 个神经元,其中以整流线性单元(ReLU)作为神经网络的激活函数。

由于离散动作空间和连续动作空间的任务模型不同,即 DQN 模型和 S-DDPG 等模型的输出层不同。对于 DQN 模型,设置输出层为 8 个神经元,每个神经元代表一个组合动作的 Q 值。每个组合动作都是多个设备的可能组合,即空调(a1)、电动汽车(a2)、洗碗机(a3)的启动或者关闭。相比之下,S-DDPG 输出层只有三个神经元,每个神经元代表一个设备动作。更准确地说,它输出的是在特定输入状态执行与设备相关联的动作的概率。这是 S-DDPG 方法相对于 DQN 方法的一个明显优势,因为 S-DDPG 与设备的数量成线性比例。

超参数设置:在所有执行的实验中,学习率均设置为 $\alpha=0.01$,折扣因子设置为 $\gamma=0.99$,$\eta=0.01$。本节训练了 5 000 个情节的模型,其中每一个情节由随机 20 天内的数据组成。网络结构的权重参数每两个情节更新一次。

3.2 数据描述

本节结合改进算法验证了所提出的模型,并在大型真实数据库中分析了该模型性能。首先,描述数据库。然后,针对各种建筑物的降低负荷峰值问题和最小化成本问题,给出两个问题的实验对比结果。

(1) 建筑能耗模型。数据集中包含用户每天使用能源的数据,将用电记录进行清洗分割,得到两千多万条数据,并将这些能耗数据用于构建特定的设备模型。图 2 和图 3 列出了每 15 分钟一次记录的两种不同类型建筑(B1 和 B2)的能源数据模型。在不同的建筑能源数据模型中,光伏发电的不确定性以及用户消耗能源的行为特征非常明显。在本文的实验中,使用了 2010 年 1 月至 2016 年 12 月之间的数据。

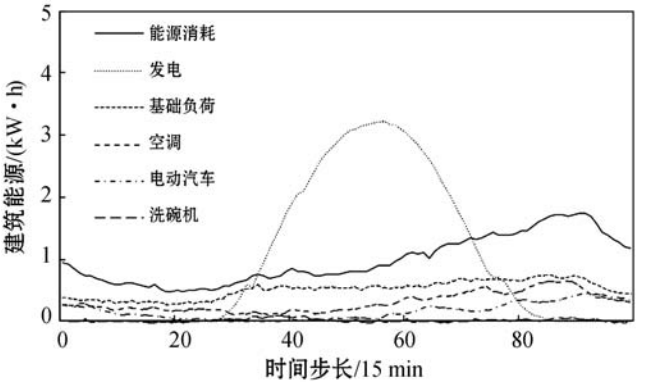

图 2 B1 型建筑的能源数据模型

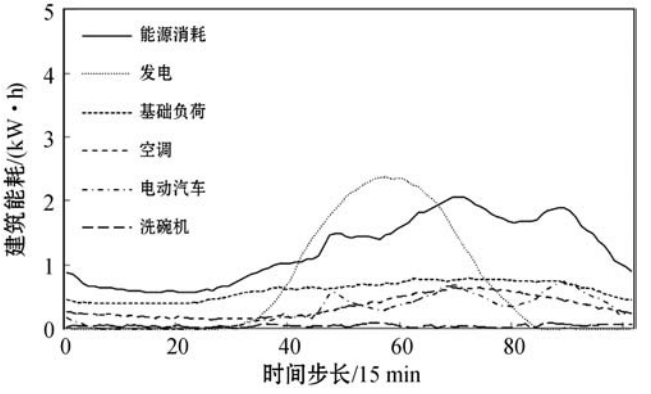

图 3 B2 型建筑的能源数据模型

(2) 价格数据。本文使用建筑能耗数据当地的电网公司为用户提供的分时电价。夏天的电价由高峰、中峰、低谷时段的电价组成,冬季的电价由高峰低谷电价组成。此外,在建筑上进行自发电的客户将收到由电网公司支付的光伏发电费用。

3.3 实验分析

表 1 和图 4、图 5 显示了两种类型建筑物(Bi)在一年内以 15 min 的频率采样,显示有关单个建筑物级别的降低负荷峰值的对比结果。表 1 中,第一列表示峰值,第二列是优化方法,第三列是某一类型的建筑,第三、第四列中 Mean 与 S. d 分别代表平均值和标准差。对于原始数据,计算日均负荷峰值的平均值和标准差。在将四种深度强化学习算法应用于建筑能耗优化控制之后,负荷峰值均有一定程度的降低。其中:DPG 方法比 DQN 方法的效果好;S-DDPG 算法的优化效果比 DQN、DPG、DDPG 等方法更好,优化之后负荷峰

值明显降低。这是因为,DQN 方法主要解决的是离散动作空间下的动作选择问题,在应对连续大规模状态空间的问题时,无法及时采取最优策略,只能进行离散化的动作选择,最终导致优化效果较差,而 DDPG 方法与 S-DDPG 方法能够在该状况下取得较好的实验结果。

表 1　建筑物日均负荷峰值

峰值	优化方法	B1		B2	
		Mean	S. d	Mean	S. d
峰值/(kW · h)		4.06	1.81	4.70	1.54
优化后的峰值/(kW · h)	DQN	2.72	1.45	3.61	1.48
	DPG	2.63	1.41	3.42	1.37
	DDPG	2.54	1.39	3.22	1.36
	S-DDPG	**2.50**	**1.35**	**3.01**	**1.25**

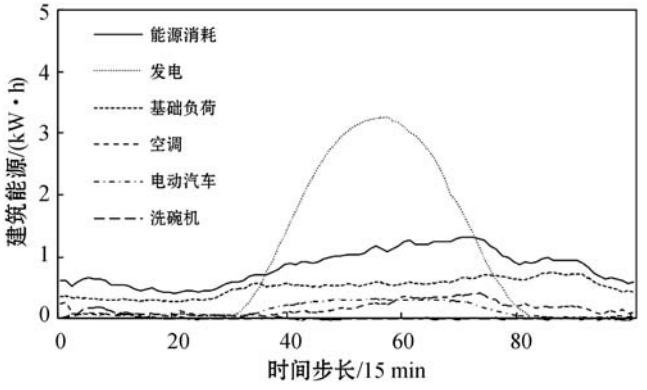

图 4　B1 型建筑降低负荷峰值后的能耗

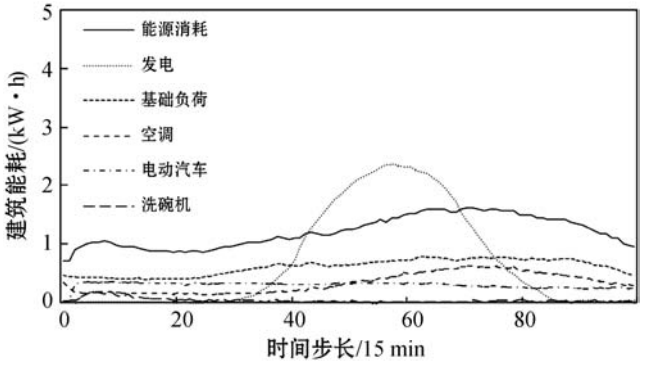

图 5　B2 型建筑降低负荷峰值后的能耗

在图 4 和图 5 中,横坐标表示时间步长(每 15 min 一次),纵坐标表示建筑能源的负荷。与图 2、图 3 的建筑能源模式相比,使用 S-DDPG 算法优化后的曲线值在 60 至 100 个时间步长内已经能够表明下降的趋势,实现了负荷峰值降低的效果。此外,该建筑物中其他的电气设备也都在一定程度上降低了负荷峰值。

表 2 总结了两种不同类型的建筑物日均最小化成本问题的实验对比结果。相较于 DQN、DPG、DDPG 方法,S-DDPG 对建筑能耗使用方法的优化控制,在降低负荷峰值和最小化能源使用成本方面取得了更好的效果。在前面的研究中,假设用户自发电卖出的电价和买入电网公司的电价相等,在首先考虑降低建筑电力能源的使用成本时,则可以将成本最小化问题转化为降低负荷峰值的问题,从而间接地降低负荷峰值。

表 2　建筑物日均最小化成本

峰值和成本	优化方法	B1		B2	
		Mean	S. d	Mean	S. d
峰值/(kW · h)		4.06	1.81	4.70	1.54
优化后的峰值/(kW · h)	DQN	3.23	1.62	3.82	1.39
	DPG	3.19	1.57	3.02	1.35
	DDPG	3.08	1.53	3.14	1.21
	S-DDPG	**2.95**	**1.47**	**2.77**	**1.16**
成本/($ · 天$^{-1}$)		2.37	3.25	3.32	3.94
成本最小化/($ · 天$^{-1}$)	DQN	2.25	3.17	3.04	3.71
	DPG	2.21	3.11	2.98	3.59
	DDPG	2.16	2.93	2.84	3.43
	S-DDPG	**2.02**	**2.81**	**2.69**	**3.21**

因此,对比图 2、图 4 和图 6,以及图 3、图 5 和图 7 可以看出,不同类型的建筑物最小化能源使用成本的解决方案与其负荷峰值降低问题,以及原有的建筑能耗模型相关。此外,对 B1 和 B2 两种类型的建筑进行能耗优化控制,B2 类型的建筑能耗在 S-DDPG 算法的优化控制下,具有更好的表现效果。

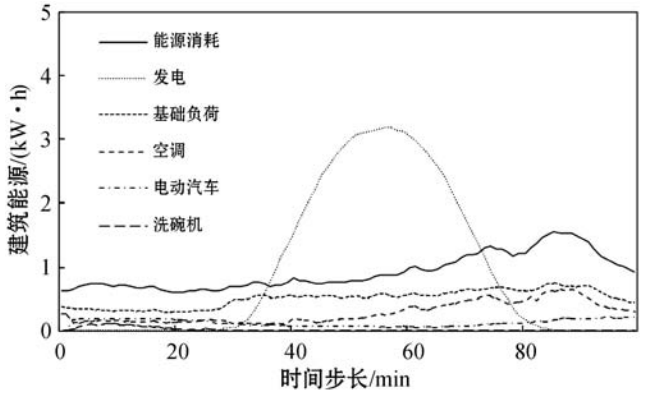

图 6　B1 型建筑最小化成本后的能耗

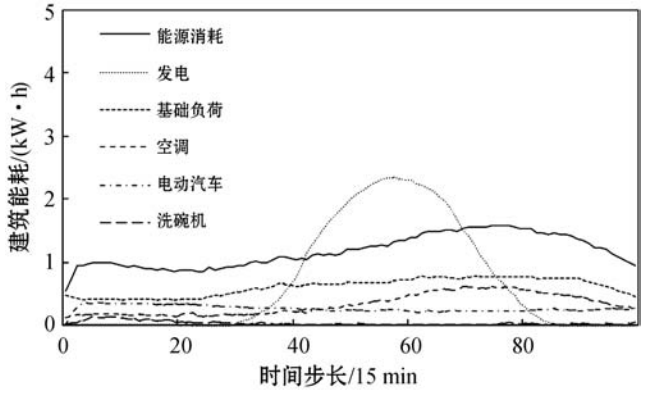

图 7　B2 型建筑最小化成本后的能耗

为了测试在大规模数据下的运行情况,本文使用 S-DDPG 和 DDPG 算法分别分析 10 座和 20 座建筑物的相应结果。表 3 表明,本文方法可以分别用于峰值降低和成本最小化问题。不仅如此,当居民在考虑降低电力使用成本时,也能够隐含地解决负荷峰值问题。在建筑物数量级别相同的前提下,S-DDPG 比 DDPG 算法具有更好的性能。总体而言,在 20 座建筑物的降低成本问题中,S-DDPG 算法将负荷峰值降低了 25.1%,成本降低了 26.9%,而 DDPG 算法将负荷峰值降低了 10.1%,成本降低了 15.6%。为可视化 S-DDPG 算法的性能,图 8 展示了 20 座建筑物中每座建筑物的未优化和优化的年度电力能源成本。可以观察到每个建筑物中居民的电力能源消费行为彼此并不相同,在某些优化效果较好的情况下,将 S-DDPG 算法应用于建筑能耗优化控制,可以将居民的年度电力能源成本降低一半。然而,在一些优化效果较差的情况下,该算法仅仅能够降低几百分点的建筑电力能耗的成本。

表 3 多个建筑物年均成本的优化结果

峰值和成本	优化方法	建筑物数量			
		10		20	
		Mean	S. d	Mean	S. d
峰值/(kW·h)		62.34	6.37	127.27	10.53
优化后的峰值/(kW·h)	DDPG	55.15	5.73	116.92	8.74
	S-DDPG	**45.31**	**4.80**	**92.61**	**7.54**
成本/($·天$^{-1}$)		60.34	21.15	120.58	30.26
成本最小化/($·天$^{-1}$)	DDPG	48.11	17.73	92.88	22.86
	S-DDPG	**44.25**	**15.81**	**82.91**	**19.18**

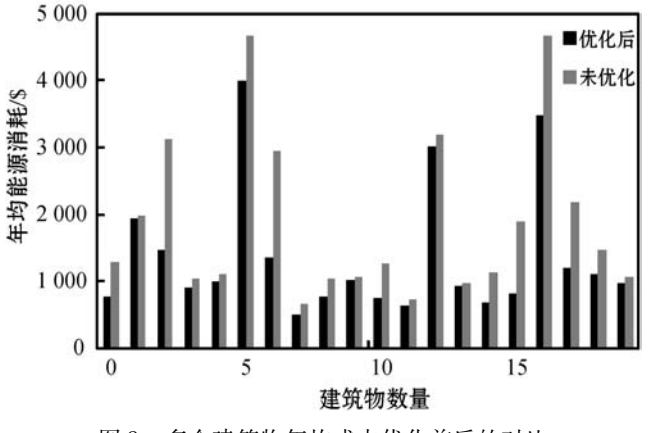

图 8 多个建筑物年均成本优化前后的对比

本节通过对实验情节的多次迭代来评估 S-DDPG 算法的收敛性能。图 9 显示了 S-DDPG 方法在降低负荷峰值方面的学习能力以及降低建筑物的负荷所对应的奖赏值。其中,实验的每个情节表示随机选择的 20 天的平均值。在实验刚开始时,可以观察到奖赏数值增加很快,而在大约 100 个情节之后,奖赏值增加变得缓慢。在大约 100 个情节之后,使用 S-DDPG 方法的平均峰值和优化的平均峰值会趋于收敛。

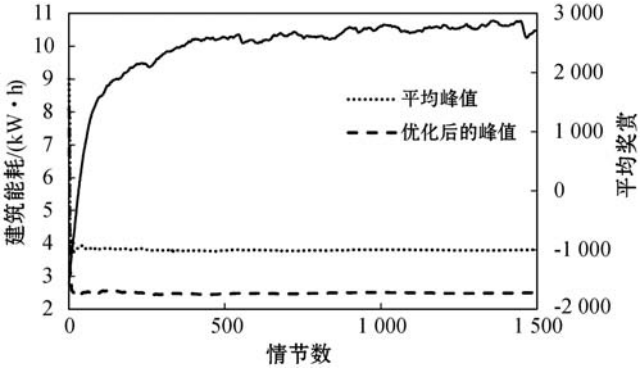

图 9 S-DDPG 方法降低的负荷峰值与奖赏值

4 结 语

本文提出一种基于深度强化学习算法的建筑能耗控制优化方法,该方法通过对建筑能耗负荷建模,在假定发电和消耗的电价相等的基础上,将峰值降低问题和成本最小化问题结合分析,构建三个奖赏函数组合而成的联合奖赏模型,用于建筑能耗控制方法模型。通过对某数据库记载的建筑能耗数据进行处理,并将 DDPG 和 S-DDPG 方法,以及基础的 DQN 算法与 DPG 算法应用于建筑能耗控制方法实验中,实验结果表明,在四种不同的方法进行对比之后,S-DDPG 方法具有更好的建筑能耗优化效果。此外,在下一步的研究计划中,将考虑更加复杂与实际的情况下建筑能耗的优化方法,并希望有更多的学者参与建筑节能的研究。

参 考 文 献

[1] 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.

[2] Sutton R S, Barto A G. Reinforcement learning: An introduction[M]. MIT Press,2018.

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

[4] 赵星宇,丁世飞. 深度强化学习研究综述[J]. 计算机科学,2018,45(7):1-6.

[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] Levine S, Finn C, Darrell T, et al. End-to-End training of

deep visuomotor policies[J]. Journal of Machine Learning Research,2015,17(39):1 – 40.
[7] Levine S, Pastor P, Krizhevsky A, et al. Learning hand-eye coordination for robotic grasping with large-scale data collection[C]//International Symposium on Experimental Robotics,2016:173 – 184.
[8] Mnih V, Kavukcuoglu K, Silver D, et al. Playing Atari with deep reinforcement learning[EB]. arXiv:1312.5602,2013.
[9] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature,2015,518(7540):529 – 533.
[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.
[11] Hasselt H V, Guez A, Silver D. Deep reinforcement learning with double Q-learning[C]//13th AAAI Conference of Artificial Intelligence,2016:2094 – 2100.
[12] Wang Z, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning[C]//33rd International Conference on Machine Learning, 2016:1995 – 2003.
[13] Schaul T, Quan J, Antonoglou I, et al. Prioritized replay buffer[C]//4th International Conference on Learning Representations,2016:322 – 355.
[14] 陈建平,周鑫,傅启明,等. 基于二阶时序差分误差的双网络 DQN 算法[J]. 计算机工程,2020,46(5):78 – 85,93.
[15] Gu S, Lillicrap T, Sutskever I, et al. Continuous deep Q-learning with model-based acceleration[C]//33rd International Conference on Machine Learning, 2016:2829 – 2838.
[16] Silver D, Lever G, Heess N, et al. Deterministic policy gradient algorithms[C]//31st International Conference on Machine Learning,2014:387 – 395.
[17] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[EB]. arXiv:1509.02971, 2015.
[18] Lowe R, Wu Y, Tamar A, et al. Multi-agent actor-critic for mixed cooperative competitive environments[C]//31th Advances in Neural Information Processing Systems, 2017:6379 – 6390.
[19] 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.
[20] 何明,张斌,柳强,等. MADDPG 算法经验优先抽取机制研究[J]. 控制与决策,2021,36(1):68 – 74.
[21] 邹长杰,郑皎凌,张中雷. 基于 GAED-MADDPG 多智能体强化学习的协作策略研究[J]. 计算机应用研究,2020,37(12):3656 – 3661.

(上接第 6 页)

研究将继续基于分布式用户的多情景业务需求进行系统的改进研究,提高系统及数据服务应用的广泛性。

参 考 文 献

[1] 张开广,孟红玲. 洛阳市城市地下管网系统的设计与实现[J]. 测绘科学技术学报,2003,20(3):201 – 212.
[2] 陆赛群,卢克,岳国英,等. 地下管网可视化 GIS 系统的设计与开发[J]. 中国给水排水,2013,29(9):100 – 104.
[3] 杨怀磊,曾思美,王万东,等. 三维煤气管网仿真系统设计与实践[J]. 中国安全科学学报,2011,21(1):43 – 49.
[4] Deng S, Ma S, Zhang X, et al. Integrated detection of a complex underground water supply pipeline system in an old urban community in China[J]. Sustainability,2020,12(4), 1670.
[5] 杜国明,龚健雅,熊汉江,等. 城市三维管网的可视化及其系统功能实现的关键技术[J]. 武汉大学学报(信息科学版),2002(5):534 – 537.
[6] Wang S,Guo Q,Xu X,et al. A study on a matching algorithm for urban underground pipelines[J]. International Journal of Geo-Information, 2018, 7(1):32.
[7] 杨春宇,纪银晓,胡启亚,等. SketchUp 软件支持下的地下管网三维建模与设计[J]. 测绘通报,2018(5):126 – 130.
[8] Wei L, Yong H, Yu L, et al. Real-time location-based rendering of urban underground pipelines[J]. International Journal of Geo-Information, 2019, 8(8):352.
[9] 卢丹丹,谭仁春,郭明武,等. 城市地下管线三维建模关键技术研究[J]. 测绘通报,2017(5):117 – 119,124.
[10] 张芳,吴思,陈勇,等. 地下管网三维可视化平台设计与实现[J]. 测绘通报,2018(7):101 – 105.
[11] 李运健,李冲,佘东静,等. 城市地下综合管线质检系统设计与实现[J]. 测绘通报,2019(2):121 – 124.
[12] 李勇,王子启,刘甲军,等. 地下管线信息系统建设中共享与保密的实践[J]. 地下空间与工程学报,2018,14(S2):470 – 473.
[13] 井雅,国明,张博尧,等. 云结构智能地下管网管理信息系统[J]. 计算机工程与设计,2018,39(1):288 – 295.
[14] 冯杭建,谢炯,潘雅辉,等. 面向规则的智能空间数据质检模型及实现[J]. 浙江大学学报(理学版),2008,35(1):100 – 104.
[15] 刘增良,陈思,陈品祥. 城市倾斜摄影实景三维模型数据质量检查方法研究与实践[J]. 测绘通报,2019(2):117 – 121.
[16] 刘耀林,赵翔,唐旭. 基于插件技术的多用途土地评价信息系统研究[J]. 中国土地科学,2010(11):29 – 36.
[17] 吕憧憬,蒋冰,曲辉,等. 基于 Web 服务的海洋空间数据共享技术研究与实现[J]. 计算机应用与软件,2016(12):49 – 51,70.