文章编号:1671-4598(2020)06-0080-06 DOI:10.16526/j.cnki.11—4762/tp.2020.06.017 中图分类号:TU995;TP273 文献标识码:A

基于Q学习的供热末端自适应PID控制算法

段中兴，赵 莎，马祥双
（西安建筑科技大学 信息与控制工程学院，西安 710055）

摘要：城市建筑集中供热末端“全开”和“全关”控制方式不仅热舒适性差，也造成了较大热损耗；为改善这一问题，在研究了强化学习的基础上提出基于Q学习PID参数的供热末端流量控制算法；首先分析了散热器等的热动态及传热过程，建立了供热房间热平衡数学模型，然后以PID控制算法为基础，温度偏差为控制器输入，调节阀开度控制量为输出，选择温差变化为智能体奖惩的学习策略，通过Q学习算法对PID参数进行在线自适应整定，最后在集中供热末端流量调节的仿真实验中验证了控制器的调控性能，并与传统PID控制结果进行了对比；实验结果表明，基于Q学习的自适应PID流量控制算法能够使室内温度变化和调节阀开度变化更加平缓，且节省约33%的供热量，节能效果较明显。

关键词：供热末端；Q学习；PID控制；流量调节

Heating End Adaptive PID Control Algorithm Based on Q Learning

Duan Zhongxing, Zhao Sha, Ma Xiangshuang
(School of Information & Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China)

Abstract: In the central heating of urban buildings, the "full open" and "fully closed" control modes at the end of heating not only have poor thermal comfort, but also cause large heat loss. For improving this problem, This paper proposes that the heating end flow control algorithm based on Q learning PID parameters is based on the study of reinforcement learning. First, we analyzed the thermal dynamics and heat transfer process of the radiator, and established a mathematical model of heat balance in the heating room. Then, the temperature deviation is the controller input, and the control valve opening degree control quantity is output based on the PID control algorithm. The temperature difference change is selected as the learning strategy of the intelligent body reward and punishment, and the PID parameters are adaptively adjusted online by the Q learning algorithm. Finally, we verified the controller's regulation performance in the simulation experiment of central heating end flow regulation and compared with the traditional PID control results. The experimental results show that the adaptive PID flow control algorithm based on Q learning can make the indoor temperature change and the adjustment valve opening change more gradual, and save about 33% of the heat supply, and the energy saving effect is more obvious.

Keywords: heating end; Q learning; PID control; flow regulation

0 引言

近年来，随着我国城市规模的快速发展和城镇化率的提高，北方城市市政集中供暖建筑面积不断增加，随之而来的是建筑供暖能耗的快速增长。当前，建筑供暖末端的调节阀多为手动调节阀，且大多处于“全开”和“全关”的运行状态，这种“全开”和“全关”的控制方式一方面给用户带来不良的热舒适体验，另一方面也造成建筑供暖能量的大量浪费。因此，建筑供暖节能存在巨大潜力，而如何实现供暖末端的高效调控，既是改善供暖室内环境热舒适性、降低建筑能耗的关键，也是集中供暖系统节能亟待解决的问题。

针对集中供暖系统与供暖末端的调控问题，国内外学者开展了大量研究，如I. H. Yang[1]等人研究了人工神经网络（ANN）在供暖系统中的应用，针对温控系统的时间滞后问题，采用ANN来估算供暖系统的启动时间以加快系统响应，提高用户的热舒适性；L. Z. Li[2]等人采用6种不同的混合控制策略对锅炉系统的燃油燃烧速率、热水流量和热水温度进行控制，取得了近17%的节能效果；徐宝萍[3]等综述及评价了国内外末端控制相关研究情况，提出了突破单一用户室温控制、兼顾供暖系统水力工况及回水温度变化的系统优化控制思路；王娇[4]等采用模糊控制理论，设计了根据各参数隶属度函数及参数调节规则的自校正模糊控制器；李琦[5]等在分析集中供热系统运行机理的基础上，建立热源总热量生产优化问题的数学描述，利用双启发式动态规划（DHP）算法和质量并调的控制策略求解，获得热源供水流量和供水温度的优化设定值；刁成玉琢[6]等采用实验研究方法对比分析了风机盘管、顶板辐射、侧墙辐射、地板辐射4种不同供暖末端时的室内温湿度、空气流速和壁面温度等数据，获得了4种供暖末端的热舒适性结论。上述研究取得了许多积极成果，对本文研究的开展具有较好的借鉴意义。

比例—积分—微分（PID）控制以其结构简单，鲁棒性

收稿日期:2019-11-20； **修回日期**:2019-12-07。
基金项目:“十三五”国家重点研发计划(2017YFC0704207)。
作者简介:段中兴(1969-),男,湖南茶陵人,工学博士,教授,主要从事智能控制与智能信号处理、网络化控制系统、嵌入式系统方向的研究。

好和工作可靠性高的特点而在控制领域得到了广泛应用，但传统 PID 的参数一旦确定就无法在线调整，难以满足时变系统的控制要求，如何高效地调整和优化 PID 的控制参数成了人们竞相研究的问题。近年来兴起的强化学习为 PID 参数自适应调整提供了新的思路和方法，并取得了较好的应用效果[7-10]。本文在分析现有研究成果的基础上，以 PID 控制算法为基础，针对集中供暖末端控制系统存在大滞后、强耦合的特点，引入强化学习算法，提出一种基于 Q 学习的 PID 参数在线优化的供暖末端流量控制算法，旨在利用 Q 学习算法对 PID 参数进行整定与寻优，从而获得更优的控制参数，并在仿真实验中验证该方法的有效性和节能效果。

1　PID 控制器

典型的 PID 控制器原理如图 1 所示。

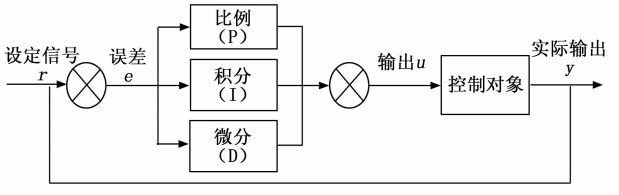

图 1　PID 控制系统原理图

典型的 PID 控制系统由控制器、被控对象和反馈回路组成。PID 控制器根据设定值和实际输出值之间的偏差，对偏差进行同比例放大（或缩小）、积分以及微分后，通过线性组合构成控制量，进而对被控对象进行控制，其控制规律如下：

$$u(t) = K_p\left[e(t) + \frac{1}{K_I}\int_0^t e(t)\mathrm{d}t + \frac{K_D de(t)}{\mathrm{d}t}\right] \tag{1}$$

式中，$e(t) = r(t) - y(t)$ 为控制量；K_p 为比例系数；K_I 为积分时间常数；K_D 为微分时间常数[11]。

2　供热末端的热平衡模型

由传热学理论可知，供热末端—采暖房间的热平衡方程可表示为：

$$Q = Q_{得} - Q_{失} \tag{2}$$

式中，$Q_{得}$ 为采暖房间总得热量，即散热器散热量；$Q_{失}$ 为采暖房间总失热量，主要包括房间维护结构传热耗热量 Q_1 和门窗缝隙渗入的室外空气吸热量 Q_2；Q 为采暖房间的最终热量，且有：

$$Q = C_k\frac{\mathrm{d}t_n}{\mathrm{d}t} \tag{3}$$

式中，C_k 为采暖房间空气的热容，$C_k = c_1 \cdot \rho_1 \cdot V$，$\rho_1$ 为室内温度下的空气密度，其取值一般通过查询《传热学附表》可得。

散热器释放热量为：

$$Q_{得} = Gc_p(t_g - t_h) \tag{4}$$

式中，t_g 为散热器进口热水温度（℃）；t_h 为散热器出水口热水温度（℃）；G 为散热器进水流量（m^3/s）；c_p 为热水比热。

室内外通过围护结构传递的热量为：

$$Q_1 = \frac{S \cdot k_1(t_n - t_w)}{L} \tag{5}$$

式中，t_n 为用户室内当前温度（℃）；t_w 为户外温度，S 为围护结构的传热面积（m^2），k_1 为围护结构（外墙）的平均传热系数（$W/m^2 \cdot ℃$），L 为墙体厚度 m。

室内外空气对流换热量为：

$$Q_2 = \lambda \cdot \nu \cdot \rho_2 \cdot c_2(t_n - t_w) \tag{6}$$

式中，λ 为单位换算系数，1 KJ/h＝0.278 W；υ 为门、窗缝隙渗入室内的总空气量（m^3/h），$\upsilon = M \times H \times \beta$；其中：$M$ 为每米门、窗缝隙渗入室内的总空气量（$m^3/h \cdot m$），H 为门、窗缝隙的计算长度（m），β 为修正系数，根据《供热工程》附录查阅可知西安地区渗透量的修正系数为 0.7。ρ_2 为冷空气的定压密度，c_2 为冷空气的定压比热。将式（3）～（6）代入式（2）可得：

$$C_k\frac{\mathrm{d}t_n}{\mathrm{d}t} = Gc_p(t_g - t_h) - \frac{S \cdot k_1(t_n - t_w)}{L} - \lambda \cdot \nu \cdot \rho_2 \cdot c_2(t_n - t_w) \tag{7}$$

式（7）即为供暖房间的热平衡数学模型。由式（7）可知，当供暖房间面积、围护结构参数等确定后，散热器入口流量决定室温变化率，由于室温设定值为人为设置，则通过控制流量大小控制房间温度变化。

3　基于 Q 学习的自适应 PID 算法

3.1　强化学习

强化学习算法（RL 算法）是机器学习的一个重要分支，其区别于深度学习中的有监督学习和无监督学习，通过试错与环境交互获得策略的改进，进行自学习和在线学习[12]。其受到大脑学习本质的启发，只通过智能体与环境交互而不知道系统模型的基础，模拟动物学习行为过程中大脑的学习过程，通过智能体（即实际运用中的传感器）与环境条件相互作用获得先前数据，独立自主进行动作选择，生成控制策略，不断循环，使智能体具有自主学习能力。强化学习过程如图 2 所示，智能体（Agent）不断与环境（environment）进行信息交互。智能体 Agent 感知环境当前状态 $S_t \in S$，根据初始策略施加一个动作 $a_t \in a$ 给环境 Environment，环境在该动作的作用后，更新状态为 $S_{t+1} \in S$，同时根据奖惩计划提供一个奖励或惩罚以更新策略，然后智能体 Agent 再次感知环境新状态 $S_{t+1} \in S$ 选择新的动作 $a_{t+1} \in a$，直到到达终端状态 $S_T \in S$。智能体 Agent 的目标就是获得最大化奖励的概率下得到一个最优控制策略。

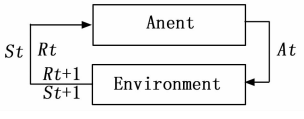

图 2　RL 中智能体－环境交互的图示

强化学习是一种基于马尔可夫决策过程的无模型增量式动态规划，其属性为：t 时刻状态信息足够以供智能体 A-

gent 进行决策生成 $t+1$ 时刻动作，从而决定进行决策 $t+1$ 时刻状态[13]。假定环境的所有可能状态是一个有限状态的离散马尔可夫过程，强化学习系统对每一步动作的选取为单步进行，环境在接受动作后便发生状态转移，并得到评价函数，其中状态转移的概率为：

$$P[s_{t+1}\ \overline{a_t}s_t] = P[s_t, a_t, s_{t+1}] \tag{8}$$

策略 π 下给定状态下的状态值函数定义为：

$$V_\pi(s) \doteq E_\pi[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1} \mid S_t = s] \tag{9}$$

其中：$\gamma\in$（0，1］是权衡下一步回报率的折扣因子，E_π 表示策略 π 下的期望值。因为在动态规划中至少得保证有一个策略 π^*，并有：

$$V^{\pi*}(s_t) = \max\{r(\pi(s_t)) + \gamma\sum P[s_t, a_t, s_{t+1}]V^{\pi*}(s_t)\} \tag{10}$$

类似的，在策略 p 下的状态 s 中采取动作 a 的动作值函数 Q_π 可以定义为：

$$Q_\pi(s,a) \doteq E_\pi[G_t \mid S_t = s, A_t = a] = E_\pi[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1} \mid S_t = s, A_t = a] \tag{11}$$

在所有动作值函数中，最佳动作值函数定义为：

$$Q_{\pi*}(s,a) \doteq \max_\pi Q_\pi(s,a) \tag{12}$$

式中，π^* 为最优策略，当策略为 π^* 时，动作函数值 Q_π（s，a）最大。在最佳动作值函数最大时的 π^* 为最优策略，根据生成的最优策略 π^*，确定最优 PID 增益（K_p（t），K_i（t），K_d（t））进行室温控制。

3.2 Q 学习算法

Q 学习算法是一种基于时间差分方法的无模型控制算法，是 RL 领域最重要的进步之一[14]。Q 学习使用状态－动作值函数 Q（S_t，A_t）来查找最优策略 π^*，动作值函数 Q（S_t，A_t）的定义如下：

$$Q(S_t,A_t) = Q(S_t,A_t) + \alpha[R_{t+1} + \gamma\max_a Q(S_{t+1},a) - Q(S_t,A_t)] \tag{13}$$

式中，$\alpha\in$（0，1］是学习率。Q 学习算法的伪代码如算法 1 所示。

算法 1：Q 学习算法

Step1：初始化任意 $\boldsymbol{Q(s,a)}$，$\forall \boldsymbol{a}\in\boldsymbol{A}$，$\forall \boldsymbol{s}\in\boldsymbol{S}$；

Step2：循环所有 epsode；

重复

Step3：更新状态 S_t；

重复

Step4：执行动作 A_t，观察 S_{t+1} 和 R_{t+1}；

Step5：根据式(13)更新 Q 值；

Step6：$S_t \leftarrow S_{t+1}$；

Step7：直到 S_t 达到最终状态 S_T；

Step8：直到 episode 结束。

3.3 供热末端自适应 PID 控制器设计

基于 Q 学习的供热末端自适应 PID 控制系统结构如图 3 所示，包含 PID 控制器和学习 Q 表两个部分。PID 控制器实现供热流量的调节，控制器参数 K_p，K_i，K_d 通过在线学习的 Q 表进行自适应调整。

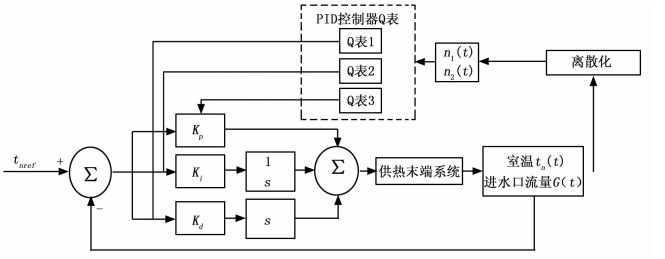

图 3 基于 Q 学习 PID 的供热末端系统控制器结构

室温设定值 $t_{n_{ref}}$ 作为输入，将室温 t_n（t）与设定值的偏差作为控制量，进行 PID 控制。便于数据采样，将室温 t_n（t）和进水口流量 G（t）离散化得到 n_1（t）和 n_2（t）作为状态，进行 Q 学习，生成 3 个 Q 表，每个 Q 表分别与 PID 控制器的比例增益 K_p、微分增益 K_i 和积分增益 K_d 相对应，当给定当前状态时，每个学习的 Q 表生成 PID 控制器增益的最佳值。

3.4 结合 Q 学习的 PID 控制算法

本文中基于 Q 学习 PID 的关键是对 PID 增益参数 Q 表的训练，通过 Q 表将不同环境状态映射到不同的 PID 的增益上。为加快 Q 表学习过程，采用了适应模型参数的启发方式——Delta－Bar－Delta [15] 自适应学习率方法。训练出当前状态下最优的 PID 增益后，根据式（2）～（7）计算出控制量 u（t），在控制量作用后再观察新状态下的流量和室温，比较前后时刻状态获得奖励 R_p，并继续进行训练学习，不断通过观察状态训练 Q 表，得出每个状态下的 PID 增益以控制阀门开度改变环境状态。故结合 Q 学习 PID 控制算法的伪代码如算法 2 所示。

算法 2：结合 Q 学习的 PID 控制算法

Step1：初始化任意 $Q_i(s,a)=0$，$\forall a\in A$，$\forall s\in S$，$i=1,2,3\cdots6$；

Step2：更新学习率 a_1 和 a_2；

Step3：更新 ε－greedy 策略的 ε；

Step4：当 episode＜maxepisode 执行：

Step5：$t=0$；

Step6：更新 $S_t(\theta(t),\dot{\theta}(t),x(t),\dot{x}(t))$；

Step7：ε 衰变，(当 episode＞0.6×maxepisode，ε＝0)；

Step8：for $t=1$；$t\leqslant$maxtime，$t++$；

Step9：将 S_t 离散化，获得：$n_1(t)$，$n_2(t)$；

Step10：for $i=1$；$i\leqslant3$，$i++$

Step11：根据 $n_1(t)$，$n_2(t)$选择动作 A_i，遵循 ε－greedy 政策；

end

Step12：根据式(2)～式(7)，获得完整的输出 $u(t)$；

Step13：观察新状态 $S_{t+1}(t_n(t+1),G(t+1))$；

Step14：获得 $Q_1(s,a)$，$Q_2(s,a)$和 $Q_3(s,a)$的奖励 R_p；

Step15：将 S_{t+1} 离散化，

获得：$n_1(t+1)$，$n_2(t+1)$；

Step16：更新 $Q_1(s,a)$，$Q_2(s,a)$和 $Q_3(s,a)$的学习率 α_1；

Step17：用 R_p 和 α_1 更新 $Q_1(s,a)$，$Q_2(s,a)$和 $Q_3(s,a)$；

Step18：$S_t \leftarrow S_{t+1}$；

End

End

3.4.1　离散化

为加快训练速度，对于效果相同的情况可为同一控制参数进行调节，故将每个连续变量被分成几个区间，同一区间内的值被视为一个相同的状态。使用相同的规则设置存储区间定义为：

$$n=\begin{cases}1 & \text{if} \quad x_{con}<X_{min}\\ 10 & \text{if} \quad x_{con}>X_{max}\\ \left[\dfrac{x_{con}}{X_{max}-X_{min}}\times N\right]+1 & \text{if} \quad X_{max}\leqslant x_{con}\leqslant X_{min}\end{cases} \tag{14}$$

其中：$[x]=max\{n\in Z\mid n\leqslant x\}$；$n$ 表示离散变量；x_{con} 表示连续变量；X_{min} 和 X_{max} 分别是 x_{con} 的下限和上限；N 表示每个变量被分成的区间数，在这种情况下 $N=10$。区间的数量取决于模拟性能。

本文需将室内温度 t_n 和阀门开度 K 通过式（14）区间划分进行离散化处理，离散化设定的值如表 1 所示。

表 1　系统离散化约束

变量	下限	上限
室内温度 t_n	5℃	30℃
阀门开度 K	0.05	1.0

3.4.2　ε－greedy 方法

为保证奖励最大化，采用当前 Q 值最大的动作，因为在 ε－greedy 策略中，ε 的值越大，表示采用随机的一个动作的概率越大。故当给定当前状态时，三个 Q 表都根据 ε－greedy 方法生成动作，此方法被定义为：

$$A=\begin{cases}\text{随机动作} & \text{如果 } \xi<\epsilon\\ \underset{a}{\arg\max}Q(s,a) & \text{其他}\end{cases} \tag{15}$$

式中，$\xi\in[0,1]$ 是一个正态分布的随机数。

为了加速收敛，ϵ 值随着训练事件的增加而衰减，并且在某些事件之后被设置为零，而事件的数量根据训练表现来决定。细节定义为：

$$\epsilon(eps)=\begin{cases}\dfrac{1}{1+e^{eps}+0.001} & eps<0.6\times maxepisode\\ 0 & \text{其他}\end{cases} \tag{16}$$

其中：eps 是当前 episode，而 $maxepisode$ 是 episode 的最大值。

3.4.3　奖励策略

奖励策略根据应用实际情况而定。本文根据室内供热末端系统将奖励函数分为 3 种情况：调控后室温趋于设定温度，室温远离设定温度，室温无变化。

1）调控后室温趋于设定温度。根据 a_t 得到的增益调控所得室温 t_n（t）与设定值 $T_{设}$ 的差值小于 $t-1$ 时刻室温 t_n（$t-1$）与 $T_{设}$ 的差值，即说明此次调控有效，给予其调控所达效果的奖励值，即为前后时刻室温变化值。

2）调控后室温远离设定温度。根据 a_t 得到的增益调控所得室温 t_n（t）与设定值 $T_{设}$ 的差值大于 $t-1$ 时刻室温 t_n（$t-1$）与 $T_{设}$ 的差值，即说明此次调控为干扰调控，奖励负值。

3）调控后室温无变化。根据 a_t 得到的增益调控所得室温 t_n（t）与设定值 $T_{设}$ 的差值等于 $t-1$ 时刻室温 t_n（$t-1$）与 $T_{设}$ 的差值，即说明此次调控无效，即不奖励不惩罚。

所以奖励计划如下：

$$r(t)=\begin{cases}|t_n(t)-t_n(t-1)| & \text{当}\ |T_{设}-t_n(t)|<|T_{设}-t_n(t-1)|\\ 0 & \text{当}\ |T_{设}-t_n(t)|=|T_{设}-t_n(t-1)|\\ -|t_n(t)-t_n(t-1)| & \text{当}\ |T_{设}-t_n(t)|>|T_{设}-t_n(t-1)|\end{cases} \tag{17}$$

3.4.4　自适应学习率

为了提高收敛效率，采用 Delta－Bar－Delta[15] 自适应学习率算法。算法定义为：

$$\Delta\alpha_t=\begin{cases}\kappa & \text{if} \quad \bar{\delta}_{t-1}\delta_t>0\\ -\phi\alpha_t & \text{if} \quad \bar{\delta}_{t-1}\delta_t<0\\ 0 & \text{if} \quad \bar{\delta}_{t-1}\delta_t=0\end{cases} \tag{18}$$

式中，$\Delta\alpha_t$ 是时间步 t 中学习速率的增量；k 是提高学习率的正常数值；ϕ 是表示折扣因子的正常数值；δ_t 是时间步长 t 中的时间差（TD）误差，$\delta_t=R_{t+1}+\gamma maxQ(S_{t+1},a)-Q(S_t,A_t)$；$\bar{\delta}_t=(1-\varphi)\delta_t+\varphi\bar{\delta}_{t-1}$。

当学习速率变得太大时，学习速率的增加改变符号并降低学习速率。另一方面，如果学习速率太小，则学习速率在先前趋势中保持变化并加速收敛。所以本文通过将当前 TD 误差与先前步骤中的累积 TD 误差进行比较来更新学习速率，即时间步骤 $t+1$ 中的学习速率为：

$$\alpha_{t+1}=\alpha_t+\Delta\alpha_t \tag{19}$$

4　仿真实验

4.1　仿真环境

实验环境为西安地区高 3 m，宽 7 m，长 10 m 的供暖房间，故采暖房间体积为 $V=210\ \text{m}^3$，窗户为 1 800 mm×1 500 mm 单层金属窗，其墙体主要为钢筋混凝土制造，墙体厚度为 $L=0.2$ m，查阅《供热工程》附录可知，钢筋混凝土围护结构（外墙）的平均传热系数为 $k_1=1.74\ \text{W/m}^2$，西安地区空气渗透量修正系数 $\beta=0.7$。根据我国《采暖通风与空气调节设计规范》查阅，设定温度设置为 18℃，西安城区冬季未供暖下平均室温为 5℃，即实验中初始室温为 5℃。仿真实验中各参数变量的取值如表 2 所示。

将表 2 实验环境数据代入式（7），可得到：

$$253.4112\frac{\mathrm{d}t_n}{\mathrm{d}t}=125.478G-887.4(t_n-2)-1.6607(t_n-2) \tag{20}$$

整理得到：

$$253.4112\frac{\mathrm{d}t_n}{\mathrm{d}t}+889.0607t_n=125.478G+1778.1214 \tag{21}$$

将式（21）拉氏变化可得：

$$(253.4112s+889.0607)T_n(s)=125.478G(s)+1778.1214 \tag{22}$$

表 2 实验环境参数取值

变量	含义	取值
c_1	室温空气比热	1.0056kJ/kg·℃
ρ_1	室温的空气密度	$\rho_1=1.2\text{kg/m}^3$
V	采暖房间的体积	210m^3
t_g	进口热水温度	70℃
t_h	出水口热水温度	40℃
c_p	热水比热	4.1826kJ/kg·℃
t_w	室外环境温度	2℃
S	传热面积	102m^2
k_1	围护结构平均传热系数	1.74W/m^2
L	墙体厚度	0.2m
λ	单位换算系数	0.278
M	缝隙每米渗入室内的总空气量	$1.0\text{m}^3/\text{h}\cdot\text{m}$
H	缝隙的计算长度	6.6m
β	修正系数	0.7
ρ_2	冷空气的定压密度	1.293
c_2	冷空气的定压比热	1KJ/kg·℃

由于本文仅考虑热水流量控制对室温调节的影响，即当实验环境确定时，即房间结构参数、室外温度和室内初始温度确定时，供暖房间的热平衡数学模型如式（22）所示。

4.2 实验结果分析

本文在 Simulink 中搭建室内热平衡模型，在 Matlab 中利用传统 PID 和基于 Q 学习的改进 PID 算法对模拟实验环境下的供热末端控制系统式（22）进行仿真。分别比较了其输出量室温和控制量阀门开度的变化，也比较了控制过程中热水总流量，并且从系统的性能指标上进行了对比。

对比图 4 可以看出，调节过程中基于 Q 学习 PID 控制的室温变化明显比传统 PID 控制策略超调量更小，所以其在热量利用率会相对更高；其振荡次数更少，人体对室温的舒适度更好。不过基于 Q 学习改进 PID 控制策略使室温达到稳态的时间较长，其主要原因是基于 Q 学习实现 PID 参数在线调节的过程中数据计算量大。

在开度调节方面，对比图 5 可以看出基于 Q 学习改进 PID 控制策略下的阀门开度变化更加平缓，其调节过程中所需要的供热流量为 $G_{总}=626.1836\ \text{m}^3$，而传统 PID 控制策下阀门调节后，整个控制过程所需的供热流量为 $G_{总}=934.421\ \text{m}^3$，基于 Q 学习的自适应 PID 控制系统节约了 32.99%的供热量。从阀门损耗角度而言，对阀门的损耗会更小，阀门使命寿命也会得到增长。

为了更精确分析两种控制策略的控制效果，结合室温变化仿真结果做了控制性能指标分析。

表 3 控制性能指标分析

控制策略	超调量/%	调节时间/s	稳态时间/s
传统 PID	34.11	284	407
基于 Q 学习的改进 PID	29.17	259	585

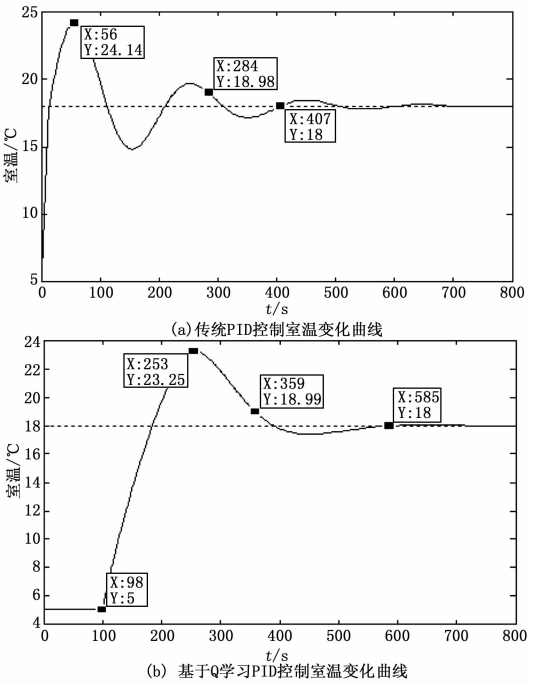

图 4 室温变化仿真结果

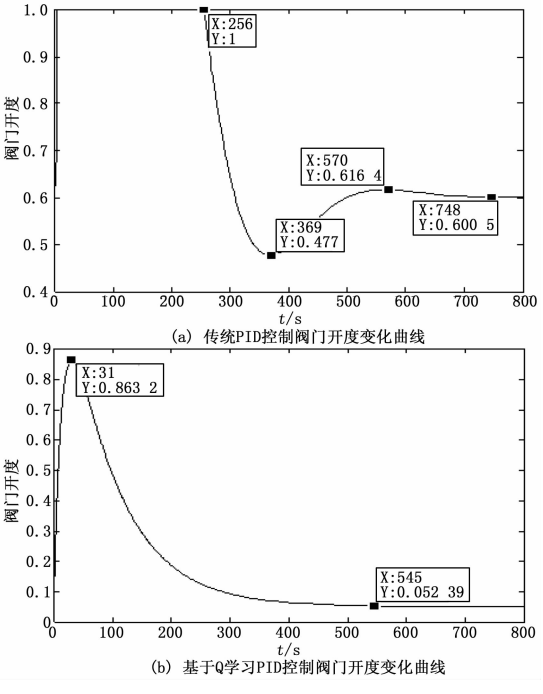

图 5 开度变化仿真结果

根据控制性能指标分析可知基于 Q 学习改进 PID 的控制策略稳态时间在 9.75 min，传统 PID 控制策略稳态时间在 6.78 min，考虑实际情况下，15 min 内达到设定温度可以满足供热用户的需求。

由于室内供暖过程中突变环境较为复杂频繁，如当室温达到设定值后，由于外来人员的突然闯入或开窗使得外来冷空气渗入导致室内温度骤降等。为得知基于 Q 学习

PID 控制策略在环境突变下的控制效果，本文在 $t=800$ s 时，室内温度发生突变骤降为 14℃后，比较基于 Q 学习 PID 控制策略和传统 PID 控制策略的控制效果，仿真结果如图 6 所示。

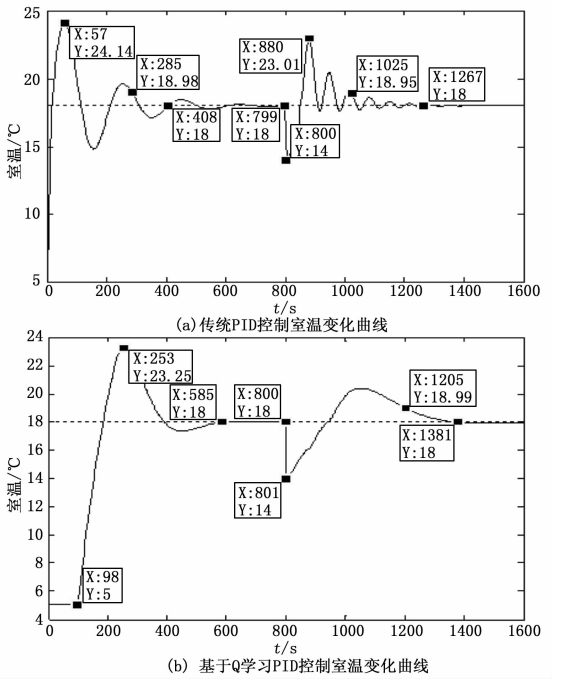

图 6　环境突变下室温变化仿真结果

根据图 6 可以得知，突变后传统 PID 控制策略的调节时间为 3.75 min，稳态时间为 7.78 min，然而基于 Q 学习 PID 控制策略的调节时间为 6.75 min，稳态时间为 9.68 min。显然，基于 Q 学习 PID 控制策略调节时间和稳态时间在 15 min 中内，依然可以保证用户舒适性。另外，基于 Q 学习 PID 控制策略的振荡频率与振荡幅度明显优于传统 PID 控制策略，在达到 18±1℃后后更加平缓的达到稳态值，即更加稳定精确。通过对初始温度－设定温度－发生突变－设定温度整个过程能耗计算可知，基于 Q 学习改进 PID 控制策略下所消耗的供热流量为 $G'_{总}=1\ 192.354\ m^3$，而传统 PID 控制策下整个控制过程所需的供热流量为 $G'_{总}=1\ 703.841\ m^3$，基于 Q 学习的自适应 PID 控制系统节约了 30.02%的供热量。

5　结束语

针对集中供热末端流量调节的 PID 控制参数优化与节能问题，首先依据传热学理论分析和推导了散热器、围护结构和室内外空气对流换热的热动态过程和传热过程，建立了供热房间的热平衡数学模型，在此基础上，以优化 PID 参数和供热末端节能为目标，提出了基于 Q 学习在线优化 PID 参数的供热末端流量控制算法，设计了自适应 PID 控制器，实现了 PID 参数的在线整定。最后通过仿真实验验证了所设计 PID 控制器的调控性能并与传统 PID 控制结果进行了对比，仿真实验结果表明，所提方法能够实现室内温度和调节阀开度的平缓调控，且能节省约 33%的供热量。当发生突变后，基于 Q 学习 PID 控制策略的振荡也优于传统 PID，初始温度－设定温度－发生突变－设定温度整个过程，基于 Q 学习的自适应 PID 控制系统能耗减少了 30.02%。在保证室内环境的热舒适性的基础上对降低建筑供热能耗具有重要的意义。

参考文献：

[1] Yang I H, Yeo M S, Kim K W. Application of artificial neural network to predict the optimal start time for heating system in building [J]. Energy Conversion and Management, 2003, 44 (17): 2791－2809.

[2] Li L Z, Zaheeruddin M. Hybrid fuzzy logic control strategies for hot water district heating systems [J]. Building Services Engineering Research and Technology, 2007, 28 (1): 35－53.

[3] 徐宝萍，张　超. 供暖系统末端控制技术发展及研究综述 [J]. 暖通空调，2019，49 (3)：1－8.

[4] 王　娇，苏　刚，苏　阳，等. 基于模糊 PID 控制的集中供热系统温度控制器的设计 [J]. 天津城建大学学报，2016，22 (5)：356－360.

[5] 李　琦，唐　巍. 基于 DHP 算法的集中供热系统热源优化控制 [J]. 控制工程，2017 (10)：50－55.

[6] 刁成玉琢，李百战，刘　红，等. 不同供暖末端室内热舒适实验研究 [J]. 暖通空调，2018 (7)：98－104.

[7] Shi Q, Lam H K, Xiao B, et al. Adaptive PID controller based on Q－learning algorithm [J]. CAAI Transactions on Intelligence Technology, 2019, 3 (4): 235－244.

[8] 邵俊恺，赵　翾，杨　珏，等无人驾驶铰接式车辆强化学习路径跟踪控制算法 [J]. 农业机械学报，2017 (03)：381－387.

[9] 孙歧峰，任　辉，段友祥. 基于异步优势执行器评价器学习的自适应 PID 控制设计 [J]. 信息与控制，2019，48 (3)：1187－1192.

[10] Carlucho I, De Paula M, Villar S A, et al. Incremental, Q－learning strategy for adaptive PID control of mobile robots [J]. Expert Systems with Applications, 2017, 80: 183－199.

[11] Pandit A, Hingu B. Online tuning of PID controller using black box multi－objective optimization and reinforcement learning [J]. International Federation of Automatic Control. 2018, 51 (32): 844－849.

[12] 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.

[13] Sutton R, Barto A. Reinforcement Learning: an introduction [M]. MIT Press, 1998.

[14] Carlucho I, et al. Incremental Q－learning strategy for adaptive PID control of mobile robots [J]. Expert Systems With Applications, 2017, 80: 183－199.

[15] Jacobs R A. Increased rates of convergence through learning rate adaptation [J]. Neural Networks, 1988, 1 (4): 295－307.