(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115167102 A
(43)申请公布日 2022.10.11

(21)申请号 202210547986.9

(22)申请日 2022.05.18

(71)申请人 上海电力大学
地址 201306 上海市浦东新区沪城环路1851号
申请人 甘肃华电福新能源有限公司民乐分公司

(72)发明人 杜海舟 徐野 郭晓晗 田飞 李鑫 李建鹏 贺正良

(74)专利代理机构 南京禹为知识产权代理事务所(特殊普通合伙) 32272
专利代理师 范晓翠

(51)Int.Cl.
G05B 11/42(2006.01)

权利要求书3页 说明书10页 附图5页

(54)发明名称

一种基于并行优势动作评价的强化学习自适应PID控制方法

(57)摘要

本发明公开了基于并行优势动作评价的强化学习自适应PID控制方法,包括设置采样周期,A2C算法的线程个数n,初始化网络参数;根据当前状态误差,利用状态转换器构造第i个智能体的输入状态向量;利用RBF网络同时实现第i个智能体策略函数和值函数的学习,参数值修正后计算系统输出,并观测下一采样时间系统误差$e_i(t+1)$,计算奖励函数$r_i(t)$;判断是否更新参数,数据总量达到batch size,停止采样,输出状态估计值并更新权值,将n个智能体上传的梯度汇总并求平均,更新Global Net参数,Global Net传递给Actor(i)和Critic(i)新的权值;迭代重复,输出算法的最优解。通过本发明提供的方法,可以有效的克服大超调,非线性和滞后性对PID控制器带来的影响。

CN 115167102 A

1.一种基于并行优势动作评价的强化学习自适应PID控制方法,其特征在于:包括以下步骤,

设置采样周期,A2C算法的线程个数n,初始化网络参数;

获取当前状态误差,误差$e_i(t)$=yd(t)-y(t),由调度中心下发的计划出力值yd(t)和发电站实际出力值y(t)的差值确定,误差$e_i(t)$作为输入信号进入控制器;

根据当前状态误差,利用状态转换器构造第i个智能体的输入状态向量$\Theta_i(t)$;

利用RBF网络同时实现第i个智能体策略函数和值函数的学习,参数值修正后计算系统输出,并观测下一采样时间系统误差$e_i(t+1)$,计算奖励函数$r_i(t)$;

判断是否更新参数,数据总量达到batch size,停止采样,输出状态估计值并更新权值,将n个智能体上传的梯度汇总并求平均,更新GlobalNet参数,Global Net传递给Actor(i)和Critic(i)新的权值;

不断迭代重复,直到满足停止条件,输出算法的最优解。

2.根据权利要求1所述的基于并行优势动作评价的强化学习自适应PID控制方法,其特征在于:根据当前状态误差,利用状态转换器构造第i个智能体的输入状态向量包括,PID控制器的控制规律如下:

$$u(t) = k_P e(t) + k_I \int_0^t e(t)\,dt + k_D \frac{de(t)}{dt}$$

将连续函数离散化,离散化后形式:

$$u_i(t) = k_P e_i(t) + k_I \sum_{j=0}^{t} e_i(j) + k_D\big(e_i(t) - e_i(t-1)\big)$$

$e_i(t)$=yd(t)-y(t)

根据递推原理:

$$u_i(t-1) = k_P e_i(t-1) + k_I \sum_{j=0}^{t-1} e_i(j) + k_D\big(e_i(t-1) - e_i(t-2)\big)$$

$\Delta u_i(t) = u_i(t) - u_i(t-1)$

故:

$\Delta u_i(t) = k_p \Delta e_i(t) + k_I e_i(t) + k_D \Delta^2 e_i(t) = K\Theta_i(t)$

式中:$i \in [1,n]$,n表示智能体的总数,i代表第i个智能体;$K=[k_I\ k_P\ k_D]$为PID控制器的三个参数值;yd(t)为设定目标值;y(t)为实测系统反馈值;$e_i(t)$为本次采样误差;$e_i(t-1)$为上一次采样误差;$e_i(t-2)$为上上次采样误差;$\Delta u_i(t)$为本次控制量增量;$\Theta_i(t)=[e_i(t)\ \Delta e_i(t)\ \Delta^2 e_i(t)]$为RBF网络的输入向量;$\Delta e_i(t)=e_i(t)-e_i(t-1)$为$e_i(t)$的一次差分;$\Delta^2 e_i(t)=e_i(t)-2e_i(t-1)+e_i(t-2)$为$e_i(t)$的二次差分。

3.根据权利要求1或2所述的基于并行优势动作评价的强化学习自适应PID控制方法,其特征在于:利用RBF网络同时实现第i个智能体策略函数和值函数的学习,参数值修正后计算系统输出,并观测下一采样时间系统误差$e_i(t+1)$,计算奖励函数$r_i(t)$包括,所述RBF网络由输入层、隐含层和输出层构成,选择RBF网络作为参数化手段,设置神经网络中心,随机配置初始Actor和Critic网络权值参数,对应网络输入$\Theta_i(t)=[e_i(t)\ \Delta e_i(t)\ \Delta^2 e_i(t)]^T$,输出为$Y=[k_I'k_P'k_D'V(t)]^T$。

4.根据权利要求3所述的基于并行优势动作评价的强化学习自适应PID控制方法，其特征在于：所述输入层包括三个输入节点，三个输入节点的输入分别为状态向量$\Theta_i(t)$的一个分量，分别代表系统输出误差、误差的一次差分、误差的二次差分。

5.根据权利要求3或4所述的基于并行优势动作评价的强化学习自适应PID控制方法，其特征在于：所述隐含层取5个节点，激活函数选用高斯型核函数，则第j个隐节点的输出为：

$$\Phi_j(t) = exp\left(-\frac{\| x(t) - \mu_j(t) \|^2}{2\sigma_j^2}\right), j = 1, \ldots, 5$$

其中：$\mu_j=[\mu_{1j}\ \mu_{2j}\ \mu_{3j}]^T$为第j个隐节点的中心向量，$\sigma_j$为第j个节点的宽度向量。

6.根据权利要求5所述的基于并行优势动作评价的强化学习自适应PID控制方法，其特征在于：所述输出层由Actor和Critic组成，Actor和Critic共享RBF网络的输入层和隐层的资源，共四个输出节点，其中前三个输出为Actor部分的输出K′(t)的三个分量，第四个节点的输出为Critic部分的值函数V(t)：

$$K_m^{'(t)} = \sum_{j=1}^{5} w_{jm}(t)\,\Phi_j(t),\ m = 1,2,3$$

$$V(t) = \sum_{j=1}^{5} w_{j4}(t)\,\Phi_j(t)$$

其中，j＝1,2,…,5为隐含层节点编号；m＝1,2,3，为输出层节点编号；wj1,2,3为隐含层第j个节点分别与输出层Actor第1,2,3个节点之间的权值；wj4为隐含层第j个节点与输出层Critic(第四个节点)之间的权值。

7.根据权利要求6所述的基于并行优势动作评价的强化学习自适应PID控制方法，其特征在于：参数值修正后计算系统输出，并观测下一采样时间系统误差$e_i(t+1)$，计算奖励函数$r_i(t)$包括，控制作用为：

$\Delta u(k)=k_p\Delta e(t)+k_I e(t)+k_D\Delta^2 e(t)=K\Theta(t)$

Actor通过高斯干扰K_ξ对K′(t)进行修正，得到最终K(t)，修正公式为：

$K(t)=K'(t)+K_\xi$

Critic用于评估值函数，AC学习中的误差δ_{TD}与状态转移中相邻状态的值函数和回报函数有关，回报函数$r_i(t)$，用来反映选择动作的好坏，定义为：

$$r_i(t) = \begin{cases} 1, |e_i(t)/e_i(t-1)| < 1 \\ 0, |e_i(t)/e_i(t-1)| = 1 \\ -1, |e_i(t)/e_i(t-1)| > 1 \end{cases}。$$

8.根据权利要求7所述的基于并行优势动作评价的强化学习自适应PID控制方法，其特征在于：判断是否更新参数，数据总量达到batch size，停止采样，输出状态估计值并更新权值，将n个智能体上传的梯度汇总并求平均，更新GlobalNet参数，GlobalNet传递给Actor(i)和Critic(i)新的权值包括，

TD误差δ_{TD}为：

$\delta_{TD}=q_t-V(S_t,W'_v)$

$q_t=r_{t+1}+\gamma r_{t+2}+\ldots+\gamma^{n-1}r_{t+n}+\gamma^n y(S_{t+n},W'_v)$

其中，γ为折扣因子，0＜γ＜1，用来确定延迟回报与立即回报的比例，此处取γ＝0.99。W'_v为Critic网络权值，δ_{TD}反映Actor网络所选动作的优劣程度；

系统的学习性能指标为：

$$E(t)=\frac{1}{2}\delta_{TD}^2(t);$$

梯度下降法进行权值迭代更新，具体公式为：

$$w_{j(t+1)}=w_j(t)+\alpha\frac{\partial E}{\partial w}$$

其中，α为学习率，是可调节的常数；$\frac{\partial E}{\partial w}$为策略梯度，又有：

$$\frac{\partial E}{\partial \delta_{TD}}=\frac{\partial\left(\frac{1}{2}\delta_{TD}^2\right)}{\partial \delta_{TD}}=\delta_{TD}$$

故RBF网络中Actor与Critic权重更新公式为：

$w_{j(t+1)}=w_j(t)+\alpha_A\delta_{TD}(t)Y_j(t)$，j＝1,2,3

$w_{4(t+1)}=w_{4(t)}+\alpha_c\delta_{TD}(t)Y_j(t)$

其中，α_A为Actor的学习率，α_c为Critic的学习率。

9.根据权利要求7或8所述的基于并行优势动作评价的强化学习自适应PID控制方法，其特征在于：判断是否更新参数，数据总量达到batch size，停止采样，输出状态估计值并更新权值，将n个智能体上传的梯度汇总并求平均，更新GlobalNet参数，GlobalNet传递给Actor(i)和Critic(i)新的权值包括，

隐含层节点的中心和宽度的更新的具体公式为：

$$\mu_{jm}(t+1)=\mu_{jm}(t)+\alpha_\mu\delta_{TD}(t)w_{j4}(t)\Phi_j(t)\frac{x_i(t)-\mu_{ij}(t)}{\sigma_j^2(t)}$$

$$\sigma_{j4}(t+1)=\sigma_{j4}(t)+\alpha_\sigma\delta_{TD}(t)w_{j4}(t)\Phi_j(t)\frac{\| x_i(t)-\mu_{ij}(t)\|^2}{\sigma_j^3(t)}$$

其中，α_μ，α_σ分别为中心和宽度的学习率；

Actor网络结构与Critic网络结构利用自身梯度更新中央大脑网络Global Net中存储的的网络参数，更新公式为：

$$dW_a=dW_a+\alpha_a\left(\nabla W'_a ln\pi\left(a|s;W'_a\right)\delta_{TD}\right)$$

$$dW_v=dW_v+\alpha_c\left(\frac{\partial {\delta_{TD}}^2}{W'_v}\right)$$

其中，W_a为中央网络存储的的Actor网络权值，W'_a为每个Actor-Critic结构的Actor网络权值，W_v为中央网络存储的的Critic网络权值，W'_v为每个Actor-Critic结构的Critic网络权值，α_A为Actor的学习率，α_c为Critic的学习率。

10.根据权利要求9所述的基于并行优势动作评价的强化学习自适应PID控制方法，其特征在于：通过调整Actor和Critic网络结构的学习率，提高PID控制器参数整定系统结构的收敛性。

一种基于并行优势动作评价的强化学习自适应PID控制方法

技术领域

[0001] 本发明强化学习,PID控制器参数调优的技术领域,尤其涉及一种基于并行优势动作评价的强化学习自适应PID控制方法。

背景技术

[0002] 在现代工业控制系统研究中,对控制性能指标进行优化是研究控制算法的首要任务之一。常见的工业控制系统一般具有非线性、含时滞、多变量等复杂特性,研究人员提出了模糊PID控制、分数阶PID控制、自抗扰控制等算法,提升控制算法的性能。工程实践中,此类控制算法和控制器的参数整定及优化过程需要工程师大量的实践经验,或通过观察被控对象的响应逐步调整,或通过辨识模型推理计算。参数优化过程繁琐耗时、常有重复性工作。

[0003] 随着人工智能技术的发展,深度学习、强化学习等人工智能理论及技术被广泛应用于语音识别、智能推荐、机器人控制等领域。由于控制理论的“反馈”概念与强化学习的“奖励”概念的相似性,为了增强控制算法性能、减少人工成本,许多学者也尝试在控制理论与控制工程领域引入强化学习。但目前这类研究大多处于理论证明和仿真实验阶段,少有工程实践的验证。

[0004] 而传统PID控制器在控制过程中,控制参数不变。但实际生产过程中,因为非线性、滞后性、扰动信号等因素,系统的过程参数和结构甚至都会发生变化,导致PID控制效果不理想。

[0005] 目前,实现电力系统自动发电控制(Automatic Generation Control,AGC)功能的控制系统依旧采用传统的定参数PID控制器。控制器的目标是使输出量能够跟随输入量的变化而变化,即输出值等于输入的期望值,而存在于期望值与实际输出值中的差值则称为误差。实际上,实际输出值并不完全等于期望值,并且易受扰动影响而只能近似地实现,即误差必然存在。导致这种情况的共有两个原因:一是互联电网的AGC系统具有时变性、非线性以及参数不确定的特点;二是电网的实际用电负荷时刻都在发生变化,机组的各类参数也在随之改变。因此,定参数PID控制策略已经不能满足电网的调频需求。

发明内容

[0006] 本部分的目的在于概述本申请的实施例的一些方面以及简要介绍一些较佳实施例,在本部分以及本申请的说明书摘要和申请名称中可能会做些简化或省略以避免使本部分、说明书摘要和申请名称的目的模糊,而这种简化或省略不能用于限制本申请的范围。

[0007] 鉴于上述和/或现有技术中所存在的问题,提出了本申请。

[0008] 因此,本申请所要解决的技术问题是:定参数PID控制策略已经不能满足电网的调频需求。

[0009] 为解决上述技术问题,本申请提供如下技术方案:一种基于并行优势动作评价的强化学习自适应PID控制方法,包括,

[0010] 设置采样周期,A2C算法的线程个数n,初始化网络参数;

[0011] 获取当前状态误差,误差$e_i(t)=yd(t)-y(t)$,由调度中心下发的计划出力值yd(t)和发电站实际出力值y(t)的差值确定,误差$e_i(t)$作为输入信号进入控制器;

[0012] 根据当前状态误差,利用状态转换器构造第i个智能体的输入状态向量$\Theta_i(t)$;

[0013] 利用RBF网络同时实现第i个智能体策略函数和值函数的学习,参数值修正后计算系统输出,并观测下一采样时间系统误差$e_i(t+1)$,计算奖励函数$r_i(t)$;

[0014] 判断是否更新参数,数据总量达到batch size,停止采样,输出状态估计值并更新权值,将n个智能体上传的梯度汇总并求平均,更新Global Net参数,Global Net传递给Actor(i)和Critic(i)新的权值;

[0015] 不断迭代重复,直到满足停止条件,输出算法的最优解。

[0016] 作为本发明所述的基于强化学习算法的PID控制器参数自整定方法的一种优选方案,其中:根据当前状态误差,利用状态转换器构造第i个智能体的输入状态向量包括,PID控制器的控制规律如下:

[0017] $$u(t)=k_P e(t)+k_I\int_0^t e(t)\,dt+k_D\frac{de(t)}{dt}$$

[0018] 将连续函数离散化,离散化后形式:

[0019] $$u_i(t)=k_P e_i(t)+k_I\sum_{j=0}^{t}e_i(j)+k_D\big(e_i(t)-e_i(t-1)\big)$$

[0020] $e_i(t)=yd(t)-y(t)$

[0021] 根据递推原理:

[0022] $$u_i(t-1)=k_P e_i(t-1)+k_I\sum_{j=0}^{t-1}e_i(j)+k_D\big(e_i(t-1)-e_i(t-2)\big)$$

[0023] $\Delta u_i(t)=u_i(t)-u_i(t-1)$

[0024] 故:

[0025] $\Delta u_i(t)=k_P\Delta e_i(t)+k_I e_i(t)+k_D\Delta^2 e_i(t)=K\Theta_i(t)$

[0026] 式中:$i\in[1,n]$,n表示智能体的总数,i代表第i个智能体;$K=[k_I\ k_P\ k_D]$为PID控制器的三个参数值;yd(t)为设定目标值;y(t)为实测系统反馈值;$e_i(t)$为本次采样误差;$e_i(t-1)$为上一次采样误差;$e_i(t-2)$为上上次采样误差;$\Delta u_i(t)$为本次控制量增量;$\Theta_i(t)=[e_i(t)\ \ \Delta e_i(t)\ \ \Delta^2 e_i(t)]$为RBF网络的输入向量;$\Delta e_i(t)=e_i(t)-e_i(t-1)$为$e_i(t)$的一次差分;$\Delta^2 e_i(t)=e_i(t)-2e_i(t-1)+e_i(t-2)$为$e_i(t)$的二次差分。

[0027] 作为本发明所述的基于强化学习算法的PID控制器参数自整定方法的一种优选方案,其中:利用RBF网络同时实现第i个智能体策略函数和值函数的学习,参数值修正后计算系统输出,并观测下一采样时间系统误差$e_i(t+1)$,计算奖励函数$r_i(t)$包括,所述RBF网络由输入层、隐含层和输出层构成,选择RBF网络作为参数化手段,设置神经网络中心,随机配置初始Actor和Critic网络权值参数,对应网络输入$\Theta_i(t)=[e_i(t)\ \ \Delta e_i(t)\ \ \Delta^2 e_i(t)]^T$,输出为$Y=k_I'\ k_P'\ k_D'\ V(t)]^T$。

[0028] 作为本发明所述的基于强化学习算法的PID控制器参数自整定方法的一种优选方案,其中:所述输入层包括三个输入节点,三个输入节点的输入分别为状态向量$\Theta_i(t)$的一

个分量,分别代表系统输出误差、误差的一次差分、误差的二次差分。

[0029] 作为本发明所述的基于强化学习算法的PID控制器参数自整定方法的一种优选方案,其中:所述隐含层取5个节点,激活函数选用高斯型核函数,则第j个隐节点的输出为:

[0030] $$\Phi_j(t) = exp\left(-\frac{\| x(t) - \mu_j(t) \|^2}{2\sigma_j^2}\right), j = 1, \ldots, 5$$

[0031] 其中:$\mu_j=[\mu_{1j}\ \mu_{2j}\ \mu_{3j}]^T$为第j个隐节点的中心向量,$\sigma_j$为第j个节点的宽度向量。

[0032] 作为本发明所述的基于强化学习算法的PID控制器参数自整定方法的一种优选方案,其中:所述输出层由Actor和Critic组成,Actor和Critic共享RBF网络的输入层和隐层的资源,共四个输出节点,其中前三个输出为Actor部分的输出K'(t)的三个分量,第四个节点的输出为Critic部分的值函数V(t):

[0033] $$K_m^{'(t)} = \sum_{j=1}^{5} w_{jm}(t)\,\Phi_j(t),\ m = 1,2,3$$

[0034] $$V(t) = \sum_{j=1}^{5} w_{j4}(t)\,\Phi_j(t)$$

[0035] 其中,j=1,2,…,5为隐含层节点编号;m=1,2,3,为输出层节点编号;wj1,2,3为隐含层第j个节点分别与输出层Actor第1,2,3个节点之间的权值;wj4为隐含层第j个节点与输出层Critic(第四个节点)之间的权值。

[0036] 作为本发明所述的基于强化学习算法的PID控制器参数自整定方法的一种优选方案,其中:参数值修正后计算系统输出,并观测下一采样时间系统误差$e_i(t+1)$,计算奖励函数$r_i(t)$包括,控制作用为:

[0037] $\Delta u(k)=k_p\Delta e(t)+k_I e(t)+k_D\Delta^2 e(t)=K\Theta(t)$

[0038] Actor通过高斯干扰K_ξ对K'(t)进行修正,得到最终K(t),修正公式为:

[0039] $K(t)=K'(t)+K_\xi$

[0040] Critic用于评估值函数,AC学习中的误差δ_{TD}与状态转移中相邻状态的值函数和回报函数有关,回报函数$r_i(t)$,用来反映选择动作的好坏,定义为:

[0041] $$r_i(t) = \begin{cases} 1, |e_i(t)/e_i(t-1)| < 1 \\ 0, |e_i(t)/e_i(t-1)| = 1 \\ -1, |e_i(t)/e_i(t-1)| > 1。 \end{cases}$$

[0042] 作为本发明所述的基于强化学习算法的PID控制器参数自整定方法的一种优选方案,其中:判断是否更新参数,数据总量达到batch size,停止采样,输出状态估计值并更新权值,将n个智能体上传的梯度汇总并求平均,更新Global Net参数,Global Net传递给Actor(i)和Critic(i)新的权值包括,

[0043] TD误差δ_{TD}为:

[0044] $\delta_{TD}=q_t-V(S_t,W'_v)$

[0045] $q_t=r_{t+1}+\gamma r_{t+2}+\cdots+\gamma^{n-1}r_{t+n}+\gamma^n V(S_{t+n},W'_v)$

[0046] 其中,γ为折扣因子,0<γ<1,用来确定延迟回报与立即回报的比例,此处取γ=0.99。W'_v为Critic网络权值,δ_{TD}反映Actor网络所选动作的优劣程度;

[0047] 系统的学习性能指标为:

[0048] $$E(t)=\frac{1}{2}\delta_{TD}^2(t);$$

[0049] 梯度下降法进行权值迭代更新,具体公式为:

[0050] $$w_{j(t+1)}=w_j(t)+\alpha\frac{\partial E}{\partial w}$$

[0051] 其中,α为学习率,是可调节的常数;$\frac{\partial E}{\partial w}$为策略梯度,又有:

[0052] $$\frac{\partial E}{\partial \delta_{TD}}=\frac{\partial\left(\frac{1}{2}\delta_{TD}^2\right)}{\partial \delta_{TD}}=\delta_{TD}$$

[0053] 故RBF网络中Actor与Critic权重更新公式为:

[0054] $w_{j(t+1)}=w_j(t)+\alpha_A\delta_{TD}(t)Y_j(t)$,j=1,2,3

[0055] $w_{4(t+1)}=w_{4(t)}+\alpha_c\delta_{TD}(t)Y_j(t)$

[0056] 其中,α_A为Actor的学习率,α_c为Critic的学习率。

[0057] 作为本发明所述的基于强化学习算法的PID控制器参数自整定方法的一种优选方案,其中:判断是否更新参数,数据总量达到batch size,停止采样,输出状态估计值并更新权值,将n个智能体上传的梯度汇总并求平均,更新Global Net参数,Global Net传递给Actor(i)和Critic(i)新的权值包括,

[0058] 隐含层节点的中心和宽度的更新的具体公式为:

[0059] $$\mu_{jm}(t+1)=\mu_{jm}(t)+\alpha_\mu\delta_{TD}(t)w_{j4}(t)\Phi_j(t)\frac{x_i(t)-\mu_{ij}(t)}{\sigma_j^2(t)}$$

[0060] $$\sigma_{j4}(t+1)=\sigma_{j4}(t)+\alpha_\sigma\delta_{TD}(t)w_{j4}(t)\Phi_j(t)\frac{\| x_i(t)-\mu_{ij}(t)\|^2}{\sigma_j^3(t)}$$

[0061] 其中,α_μ,α_σ分别为中心和宽度的学习率;

[0062] Actor网络结构与Critic网络结构利用自身梯度更新中央大脑网络Global Net中存储的的网络参数,更新公式为:

[0063] $$dW_a=dW_a+\alpha_a\left(\nabla W_a' ln\pi\left(a|s;W_a'\right)\delta_{TD}\right)$$

[0064] $$dW_v=dW_v+\alpha_c\left(\frac{\partial {\delta_{TD}}^2}{W_v'}\right)$$

[0065] 其中,W_a为中央网络存储的的Actor网络权值,W'_a为每个Actor-Critic结构的Actor网络权值,W_v为中央网络存储的的Critic网络权值,W'_v为每个Actor-Critic结构的Critic网络权值,α_A为Actor的学习率,α_c为Critic的学习率。

[0066] 作为本发明所述的基于强化学习算法的PID控制器参数自整定方法的一种优选方案,其中:通过调整Actor和Critic网络结构的学习率,提高所述PID控制器参数整定系统结构的收敛性。

[0067] 本申请的有益效果:通过本发明提供的方法,可以有效的克服大超调,非线性和滞后性对PID控制器带来的影响,以优化PID控制器在电网领域的适用性。

附图说明

[0068] 为了更清楚地说明本申请实施例的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其它的附图。其中：

[0069] 图1为本发明实施例提供的基于并行优势动作评价的强化学习自适应PID控制方法的基本流程示意图；

[0070] 图2为本发明实施例提供的基于并行优势动作评价的强化学习自适应PID控制方法的PID控制原理框图；

[0071] 图3为本发明实施例提供的基于并行优势动作评价的强化学习自适应PID控制方法的强化学习基本框架；

[0072] 图4为本发明实施例提供的基于并行优势动作评价的强化学习自适应PID控制方法的Advantage-Actor-Critic学习框架；

[0073] 图5为本发明实施例提供的基于并行优势动作评价的强化学习自适应PID控制方法的RBF神经网络结构；

[0074] 图6为本发明实施例提供的基于并行优势动作评价的强化学习自适应PID控制方法的整体网络结构示意图；

[0075] 图7为本发明实施例提供的基于并行优势动作评价的强化学习自适应PID控制方法的控制器Simulink模型；

[0076] 图8为本发明实施例提供的基于并行优势动作评价的强化学习自适应PID控制方法的奖励函数收敛曲线；

[0077] 图9为本发明实施例提供的基于并行优势动作评价的强化学习自适应PID控制方法的仿真实验结果对比图。

具体实施方式

[0078] 为使本申请的上述目的、特征和优点能够更加明显易懂，下面结合说明书附图对本申请的具体实施方式做详细的说明。

[0079] 在下面的描述中阐述了很多具体细节以便于充分理解本申请，但是本申请还可以采用其他不同于在此描述的其它方式来实施，本领域技术人员可以在不违背本申请内涵的情况下做类似推广，因此本申请不受下面公开的具体实施例的限制。

[0080] 其次，本申请结合示意图进行详细描述，在详述本申请实施例时，为便于说明，表示器件结构的剖面图会不依一般比例作局部放大，而且所述示意图只是示例，其在此不应限制本申请保护的范围。此外，在实际制作中应包含长度、宽度及深度的三维空间尺寸。

[0081] 再其次，此处所称的“一个实施例”或“实施例”是指可包含于本申请至少一个实现方式中的特定特征、结构或特性。在本说明书中不同地方出现的“在一个实施例中”并非均指同一个实施例，也不是单独的或选择性的与其他实施例互相排斥的实施例。

[0082] 实施例1

[0083] 本实施例提供了一种基于强化学习的控制参数优化算法，通过结合多线程优势动作评价算法的奖励机制，RBF的快速学习和控制系统的动态性能指标评价模块，对控制器参

数进行在线自整定及优化。优势动作评价算法(Advantage Actor-critic Algorithm),以下简称为A2C,是深度强化学习领域新一代算法,目的在于解决贯序决策问题。算法的基础是动作-评价强化学习算法(Actor-Critic Algorithm),包括Actor和Critic两个网络,Actor是一个以策略为基础的网络,通过奖惩信息来进行调节不同状态下采取各种动作的概率;Critic是一个以值为基础的学习网络,可以计算每一步的奖惩值。二者相结合,Actor来选择动作,Critic告诉Actor选择的动作是否合适。

[0084] 在这一过程中,Actor不断迭代,得到每一个状态下选择每一动作的合理概率,Critic也不断迭代,不断完善每个状态下选择每一个动作的奖励值。A2C算法创建多个并行的环境,每个并行环境同时运行Actor-Critic,让多个拥有副结构的智能体同时在这些并行环境上更新主结构中的参数。并行中的智能体互不干扰,而主结构的参数更新通过副结构上传汇总各并行智能体的更新梯度实现,所以相对于Actor-Critic算法,A2C算法中数据收敛性显著提高。A2C算法在策略寻优方面体现出了非凡的性能,现阶段被广泛应用于金融、自动控制、游戏等行业,并取得了良好的效果。作为一种动态变参数PID控制算法,具有参数自整定及优化算法的可行性、有效性和普适性。本发明通过利用多线程并行的特性,异步训练多个智能体,经历不同的学习过程,打破样本间的关联性,作为一种动态变参数PID控制算法,解决传统定参数PID控制器的不足。

[0085] 具体的,参照图1～7,一种基于并行优势动作评价的强化学习自适应PID控制方法,包括以下步骤:

[0086] 步骤一:设置采样周期,A2C算法的线程个数n,初始化网络参数。

[0087] 步骤二:在输入端,定义期望输出由调度中心下发的计划出力值yd(t),与被控对象实际输出发电站实际出力值y(t)的差值为状态误差$e_i(t)=yd(t)-y(t)$,误差$e_i(t)$作为输入信号进入控制器。

[0088] 步骤三:根据误差e(t)构建状态向量$\Theta_i(t)=[e_i(t)\ \ \Delta e_i(t)\ \ \Delta^2 e_i(t)]$;

[0089] 参照图2,PID的控制规律为:

[0090] $$u(t)=k_P e(t)+k_I\int_0^t e(t)\,dt+k_D\frac{de(t)}{dt}$$

[0091] k_P-比例环节系数,k_I-积分环节系数,k_D-微分环节系数。

[0092] 计算机控制是采样控制,需将连续函数离散化,离散化后形式:

[0093] $$u_i(t)=k_P e_i(t)+k_I\sum_{j=0}^{t}e_i(j)+k_D\big(e_i(t)-e_i(t-1)\big)$$

[0094] $e_i(t)=yd(t)-y(t)$

[0095] 根据递推原理:

[0096] $$u_i(t-1)=k_P e_i(t-1)+k_I\sum_{j=0}^{t-1}e_i(j)+k_D\big(e_i(t-1)-e_i(t-2)\big)$$

[0097] $\Delta u_i(t)=u_i(t)-u_i(t-1)$

[0098] 故:

[0099] $\Delta u_i(t)=k_P\Delta e_i(t)+k_I e_i(t)+k_D\Delta^2 e_i(t)=K\Theta_i(t)$

[0100] 式中:

[0101] $i\in[1,n]$,n表示智能体的总数,i代表第i个智能体;

[0102] $K=[k_I\ k_P\ k_D]$,PID控制器的三个参数值;

[0103] yd(t),设定目标值;y(t),实测系统反馈值;

[0104] $e_i(t)$,本次采样误差;$e_i(t-1)$,上一次采样误差;

[0105] $e_i(t-2)$,上上次采样误差;$\Delta u_i(t)$,本次控制量增量;

[0106] $\Theta_i(t)=[e_i(t)\ \ \Delta e_i(t)\ \ \Delta^2 e_i(t)]$,RBF网络的输入向量;

[0107] $\Delta e_i(t)=e_i(t)-e_i(t-1)$,$e_i(t)$的一次差分;

[0108] $\Delta^2 e_i(t)=e_i(t)-2e_i(t-1)+e_i(t-2)$,$e_i(t)$的二次差分。

[0109] 步骤四:选择RBF网络作为参数化手段,设置神经网络中心,随机配置初始Actor和Critic网络权值参数,状态向量$\Theta_i(t)$作为输入,经过隐含层与输出层的计算,对应网络输入$\Theta_i(t)=[e_i(t)\ \ \Delta e_i(t)\ \ \Delta^2 e_i(t)]^T$,输出为$Y=k_I'\ k_P'\ k_D'\ V(t)]^T$。参照图5,RBF神经网络同时实现策略函数和值函数的学习;

[0110] 第一层:输入层是整个RBF网络的输入。该层共有三个输入节点,这三个节点的输入分别是状态向量$\Theta_i(t)$的一个分量。$\Theta_i(t)=[e_i(t)\ \ \Delta e_i(t)\ \ \Delta^2 e_i(t)]^T$,分别代表系统输出误差、误差的一次差分、误差的二次差分。

[0111] 第二层:隐含层取5个节点,激活函数选用高斯型核函数。根据高斯核函数的形式,第j个隐节点的输出为:

[0112] $$\Phi_j(t)=exp\left(-\frac{\|x(t)-\mu_j(t)\|^2}{2\sigma_j^2}\right),j=1,\ldots,5$$

[0113] 其中:$\mu_j=[\mu_{1j}\ \mu_{2j}\ \mu_{3j}]^T$为第j个隐节点的中心向量,$\sigma_j$为第j个节点的宽度向量,隐含层是对激活函数的参数进行调整,采用非线性优化策略,学习速度较慢。

[0114] 第三层:为了减小计算量,提升计算速度,利用一个RBF网络同时实现策略函数和值函数的学习。Actor和Critic共享RBF网络的输入层和隐层的资源,故输出层由Actor和Critic两部分组成,共四个输出节点。其中前三个输出为Actor部分的输出K'(t)的三个分量,第四个节点的输出为Critic部分的值函数V(t)。输出分别根据以下公式来计算:

[0115] $$K_m'(t)=\sum_{j=1}^{5}w_{jm}(t)\Phi_j(t),\ m=1,2,3$$

[0116] $$V(t)=\sum_{j=1}^{5}w_{j4}(t)\Phi_j(t)$$

[0117] 其中,j=1,2,…,5为隐含层节点编号;m=1,2,3,为输出层节点编号;$w_{j1,2,3}$为隐含层第j个节点分别与输出层Actor第1,2,3个节点之间的权值;w_{j4}为隐含层第j个节点与输出层Critic(第四个节点)之间的权值。

[0118] 步骤五:利用Actor-Critic模型进行动作评价及修正;

[0119] Actor-Critic学习模型主要分为两个部分:执行器Actor和评价器Critic。Actor-Critic学习可对值函数和策略函数进行逼近,其中策略函数由Actor部分策略梯度估计方法进行梯度下降学习来实现;而值函数估计由Critic部分采用TD学习算法实现。参照图4,Actor-Critic学习模型的体系结构:对于状态s,执行器根据当前策略选择动作a,状态s接

受动作a的作用后，转移到状态s+1，同时产生一个回报信号r；状态s和回报信号r作为评价器的输入，其输出为值函数的估计，并产生一个TD误差信号，用于评价器和执行器网络的更新学习，对选择的动作进行评价，以修正执行器的动作选择策略。

[0120] (1)Actor网络

[0121] Actor的主要作用是学习策略。利用高斯干扰K_{ξ}对K'(t)进行修正，得到最终K(t)。高斯干扰K_{ξ}是一个期望为零，依赖值函数信息的方差为$\sigma = \frac{1}{1+e^{V(t)}}$的正态分布函数；修正公式为：

[0122] $K(t)=K'(t)+K_{\xi}$

[0123] (2)Critic网络

[0124] Critic主要用来评估值函数。本专利应用TD算法来学习。AC学习中的误差δ_{TD}与状态转移中相邻状态的值函数和回报函数有关。回报函数r(t)，用来反映选择动作的好坏，定义为：

[0125] $$r(t)=\begin{cases}1, |e(t)/e(t-1)|<1\\ 0, |e(t)/e(t-1)|=1\\ -1, |e(t)/e(t-1)|>1\end{cases}$$

[0126] 步骤六：K(t)作为PID参数，将控制信号作用于被控系统，观测下一采样时间的系统输出和奖励函数值；

[0127] 控制作用为：

[0128] $\Delta u(k)=k_P\Delta e(t)+k_I e(t)+k_D\Delta^2 e(t)$

[0129] 步骤七：根据新的系统输出值构建新的系统状态；

[0130] 输入向量为：

[0131] $\Theta_i(t+1)=[e_i(t+1)\quad \Delta e_i(t+1)\quad \Delta^2 e_i(t+1)]$

[0132] 步骤八：预测下一采样时间的输出函数和回报函数，计算值函数和TD误差，更新网络参数，收集所有Agent参数以更新Global Network的参数；

[0133] (1)TD误差δ_{TD}为：

[0134] $\delta_{TD}=q_t-V(S_t,W'_v)$

[0135] $q_t=r_{t+1}+\gamma r_{t+2}+\ldots+\gamma^{n-1}r_{t+n}+\gamma^n V(S_{t+n},W'_v)$

[0136] 其中，γ为折扣因子，0＜γ＜1，用来确定延迟回报与立即回报的比例，此处取γ＝0.99。W'_v为Critic网络权值，δ_{TD}反映Actor网络所选动作的优劣程度。

[0137] (2)系统的学习性能指标为：

[0138] $$E(t)=\frac{1}{2}\delta_{TD}^2(t)$$

[0139] (3)梯度下降法进行权值迭代更新，具体公式为：

[0140] $$w_{j(t+1)}=w_j(t)+\alpha\frac{\partial E}{\partial w}$$

[0141] 其中，α为学习率，是可调节的常数；$\frac{\partial E}{\partial w}$为策略梯度，又有：

[0142] $$\frac{\partial E}{\partial \delta_{TD}}=\frac{\partial\left(\frac{1}{2}\delta_{TD}^{2}\right)}{\partial \delta_{TD}}=\delta_{TD}$$

[0143] 故RBF网络中Actor与Critic权重更新公式为：

[0144] $w_{j(t+1)}=w_{j}(t)+\alpha_{A}\delta_{TD}(t)Y_{j}(t)$，j=1,2,3

[0145] $w_{4(t+1)}=w_{4(t)}+\alpha_{c}\delta_{TD}(t)Y_{j}(t)$

[0146] 其中，α_{A}为Actor的学习率，α_{c}为Critic的学习率。

[0147] (4)隐含层节点的中心和宽度的更新的具体公式为：

[0148] $$\mu_{jm}(t+1)=\mu_{jm}(t)+\alpha_{\mu}\delta_{TD}(t)w_{j4}(t)\Phi_{j}(t)\frac{x_{i}(t)-\mu_{ij}(t)}{\sigma_{j}^{2}(t)}$$

[0149] $$\sigma_{j4}(t+1)=\sigma_{j4}(t)+\alpha_{\sigma}\delta_{TD}(t)w_{j4}(t)\Phi_{j}(t)\frac{\| x_{i}(t)-\mu_{ij}(t)\|^{2}}{\sigma_{j}^{3}(t)}$$

[0150] 其中，α_{μ}，α_{σ}分别为中心和宽度的学习率。

[0151] (5)Actor网络结构与Critic网络结构利用自身梯度更新中央大脑网络Global Net中存储的的网络参数，更新公式为：

[0152] $$dW_{a}=dW_{a}+\alpha_{a}\left(\nabla W_{a}^{'}ln\pi\left(a|s;W_{a}^{'}\right)\delta_{TD}\right)$$

[0153] $$dW_{v}=dW_{v}+\alpha_{c}\left(\frac{\partial {\delta_{TD}}^{2}}{W_{v}^{'}}\right)$$

[0154] 其中，W_{a}为中央网络存储的的Actor网络权值，W'_{a}为每个Actor-Critic结构的Actor网络权值，W_{v}为中央网络存储的的Critic网络权值，W'_{v}为每个Actor-Critic结构的Critic网络权值，α_{A}为Actor的学习率，α_{c}为Critic的学习率。

[0155] 步骤九：参照图6的流程，运行A2C算法进行迭代，重复步骤三到七，直到学习完毕。

[0156] 实施例2

[0157] 对实施例1提供的基于并行优势动作评价的强化学习自适应PID控制方法进行实验评估，将实施例1的A2C算法在gym测试框架中训练，训练结果如图8所示。可以看出在200个回合左右，奖励函数收敛至最大值。

[0158] 训练不同回合数将参数代入PID控制器仿真，与同为深度强化学习PID控制器的DQN-PID做对比，得到的调节效果如图9所示，可以看出本发明提供的方法可以有效改善PID控制器因大超调，非线性和滞后性等问题带来的影响。

[0159] 综上所述，发明将参数整定问题近似为求解约束优化问题，通过结合多线程A2C的奖励机制，RBF的快速学习和控制系统的动态性能指标评价模块，对控制器参数进行在线自整定及优化。作为一种动态变参数PID控制算法，具有参数自整定及优化算法的可行性、有效性和普适性。本发明通过利用多线程并行的特性，异步训练多个智能体，经历不同的学习过程，打破样本间的关联性，作为一种动态变参数PID控制算法，解决传统定参数PID控制器的不足。

[0160] 应当认识到，本发明的实施例可以由计算机硬件、硬件和软件的组合、或者通过存储在非暂时性计算机可读存储器中的计算机指令来实现或实施。所述方法可以使用标准编程技术-包括配置有计算机程序的非暂时性计算机可读存储介质在计算机程序中实现，其

中如此配置的存储介质使得计算机以特定和预定义的方式操作——根据在具体实施例中描述的方法和附图。每个程序可以以高级过程或面向对象的编程语言来实现以与计算机系统通信。然而，若需要，该程序可以以汇编或机器语言实现。在任何情况下，该语言可以是编译或解释的语言。此外，为此目的该程序能够在编程的专用集成电路上运行。

[0161] 此外，可按任何合适的顺序来执行本文描述的过程的操作，除非本文另外指示或以其他方式明显地与上下文矛盾。本文描述的过程(或变型和/或其组合)可在配置有可执行指令的一个或多个计算机系统的控制下执行，并且可作为共同地在一个或多个处理器上执行的代码(例如，可执行指令、一个或多个计算机程序或一个或多个应用)、由硬件或其组合来实现。所述计算机程序包括可由一个或多个处理器执行的多个指令。

[0162] 进一步，所述方法可以在可操作地连接至合适的任何类型的计算平台中实现，包括但不限于个人电脑、迷你计算机、主框架、工作站、网络或分布式计算环境、单独的或集成的计算机平台、或者与带电粒子工具或其它成像装置通信等等。本发明的各方面可以以存储在非暂时性存储介质或设备上的机器可读代码来实现，无论是可移动的还是集成至计算平台，如硬盘、光学读取和/或写入存储介质、RAM、ROM等，使得其可由可编程计算机读取，当存储介质或设备由计算机读取时可用于配置和操作计算机以执行在此所描述的过程。此外，机器可读代码，或其部分可以通过有线或无线网络传输。当此类媒体包括结合微处理器或其他数据处理器实现上文所述步骤的指令或程序时，本文所述的发明包括这些和其他不同类型的非暂时性计算机可读存储介质。当根据本发明所述的方法和技术编程时，本发明还包括计算机本身。计算机程序能够应用于输入数据以执行本文所述的功能，从而转换输入数据以生成存储至非易失性存储器的输出数据。输出信息还可以应用于一个或多个输出设备如显示器。在本发明优选的实施例中，转换的数据表示物理和有形的对象，包括显示器上产生的物理和有形对象的特定视觉描绘。

[0163] 如在本申请所使用的，术语“组件”、“模块”、“系统”等等旨在指代计算机相关实体，该计算机相关实体可以是硬件、固件、硬件和软件的结合、软件或者运行中的软件。例如，组件可以是，但不限于是：在处理器上运行的处理、处理器、对象、可执行文件、执行中的线程、程序和/或计算机。作为示例，在计算设备上运行的应用和该计算设备都可以是组件。一个或多个组件可以存在于执行中的过程和/或线程中，并且组件可以位于一个计算机中以及/或者分布在两个或更多个计算机之间。此外，这些组件能够从在其上具有各种数据结构的各种计算机可读介质中执行。这些组件可以通过诸如根据具有一个或多个数据分组(例如，来自一个组件的数据，该组件与本地系统、分布式系统中的另一个组件进行交互和/或以信号的方式通过诸如互联网之类的网络与其它系统进行交互)的信号，以本地和/或远程过程的方式进行通信。

[0164] 应说明的是，以上实施例仅用以说明本发明的技术方案而非限制，尽管参照较佳实施例对本发明进行了详细说明，本领域的普通技术人员应当理解，可以对本发明的技术方案进行修改或者等同替换，而不脱离本发明技术方案的精神和范围，其均应涵盖在本发明的权利要求范围当中。

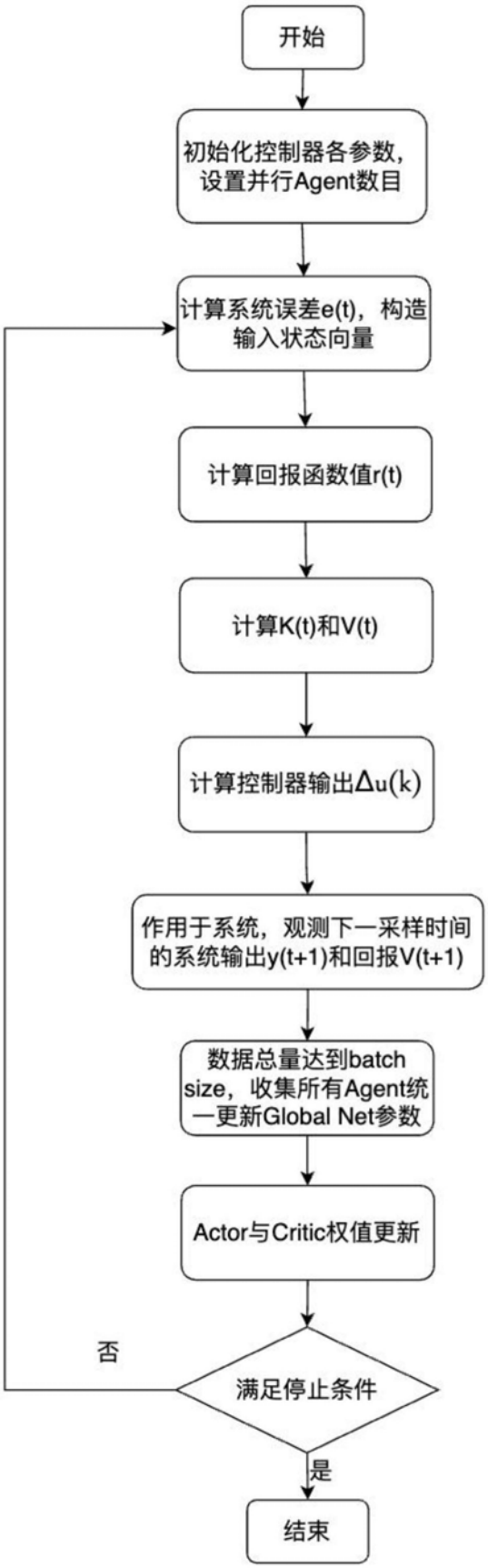

开始
初始化控制器各参数，设置并行Agent数目
计算系统误差e(t)，构造输入状态向量
计算回报函数值r(t)
计算K(t)和V(t)
计算控制器输出Δu(k)
作用于系统，观测下一采样时间的系统输出y(t+1)和回报V(t+1)
数据总量达到batch size，收集所有Agent统一更新Global Net参数
Actor与Critic权值更新
否
满足停止条件
是
结束

图1

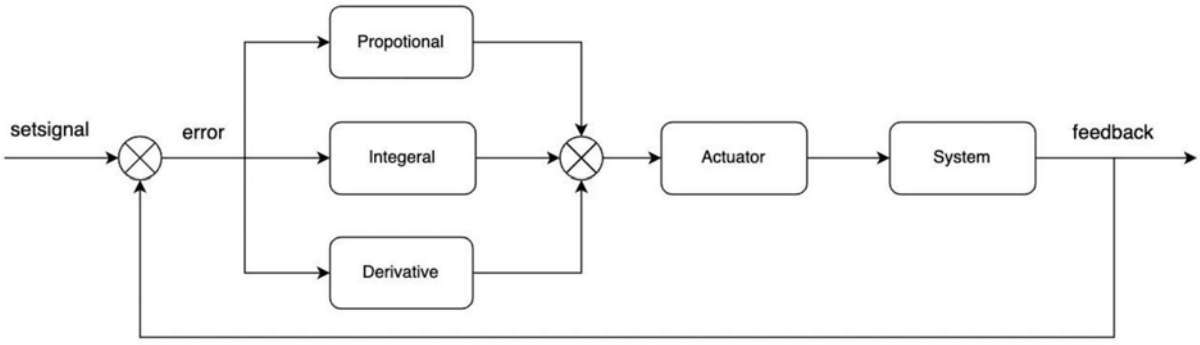

Propotional
setsignal
error
Integeral
Actuator
System
feedback
Derivative

图2

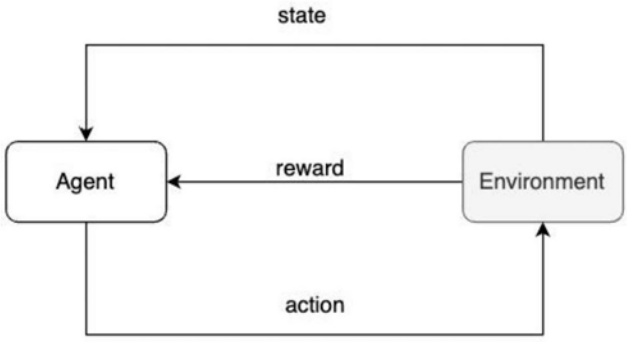

state
Agent
reward
Environment
action

图3

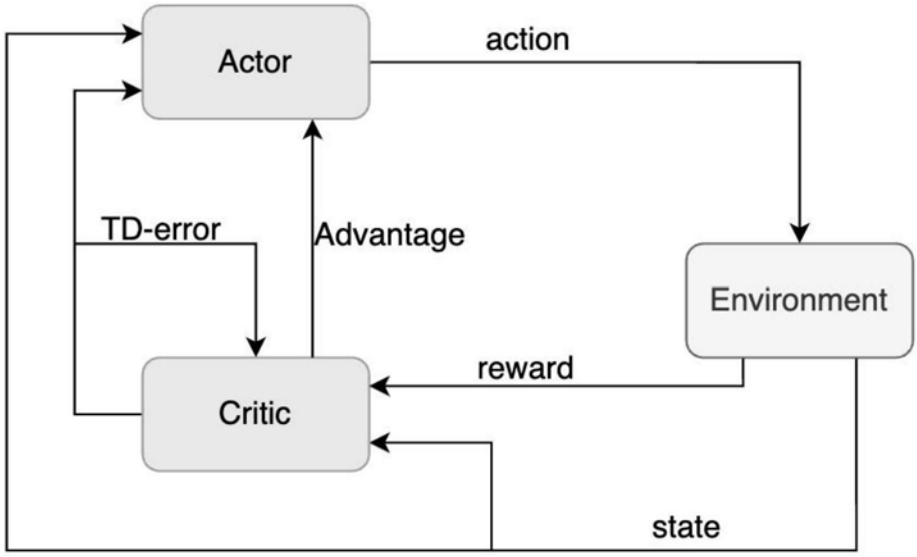

Actor
action
TD-error
Advantage
Environment
reward
Critic
state

图4

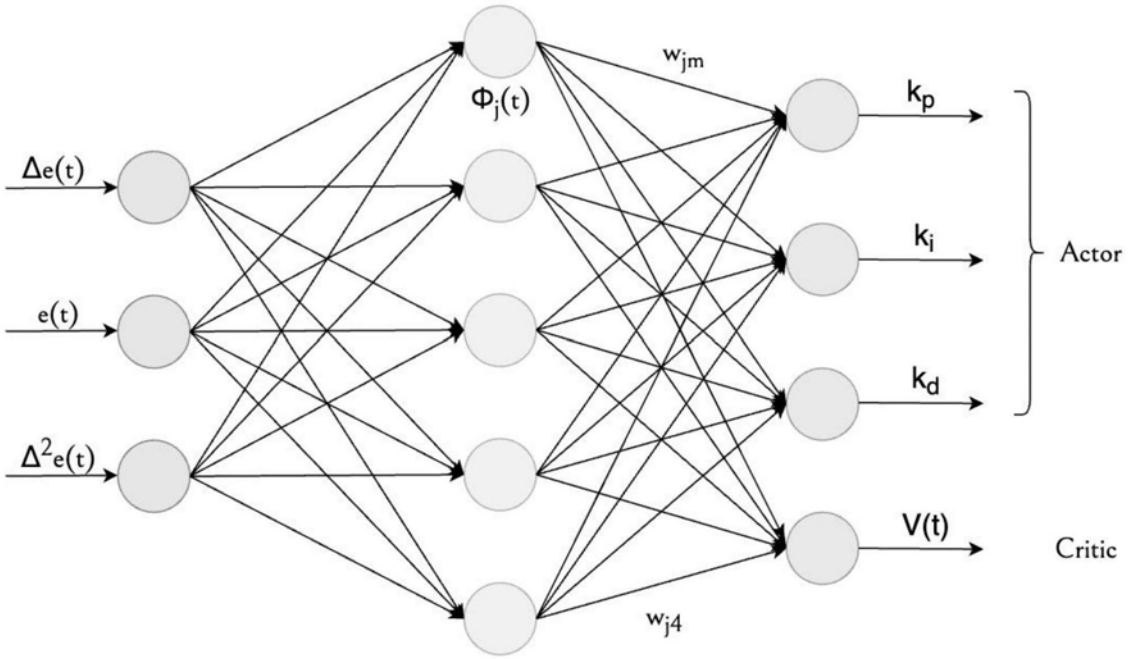

Δe(t)
e(t)
Δ²e(t)
Φj(t)
wjm
wj4
kp
ki
kd
Actor
V(t)
Critic

图5

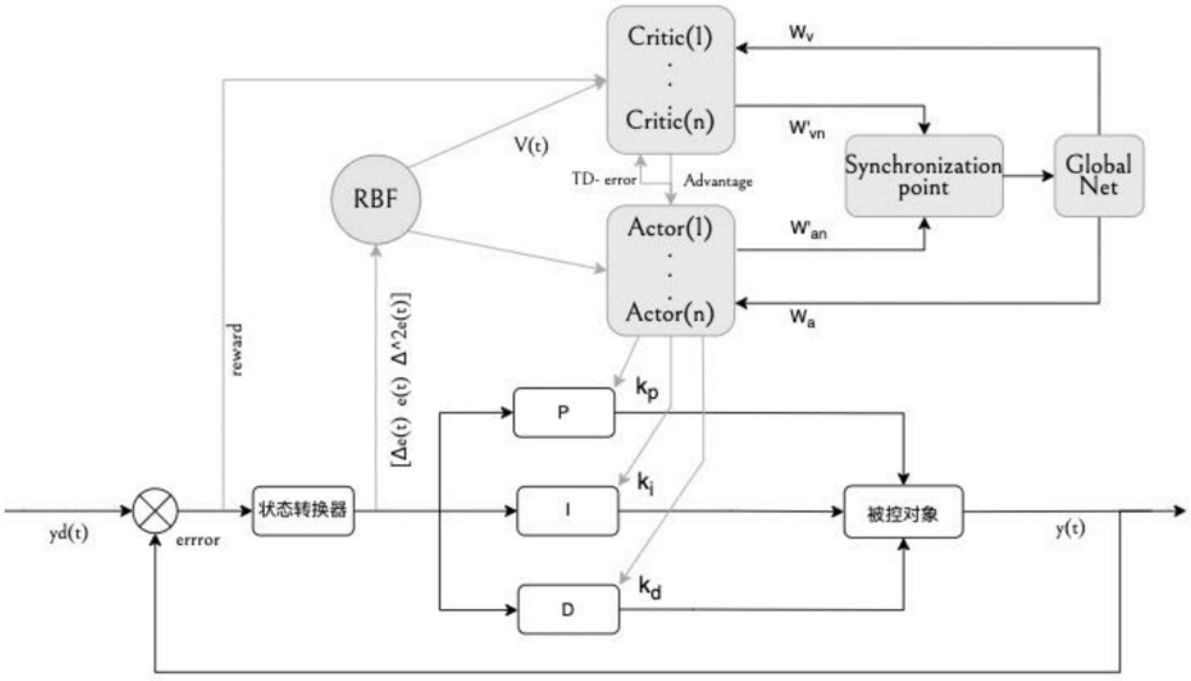

Critic(l)
Critic(n)
Wv
W'vn
Synchronization point
Global Net
V(t)
TD- error
Advantage
RBF
Actor(l)
Actor(n)
W'an
Wa
reward
[Δe(t) e(t) Δ^2e(t)]
kp
ki
kd
P
I
D
yd(t)
errror
状态转换器
被控对象
y(t)

图6

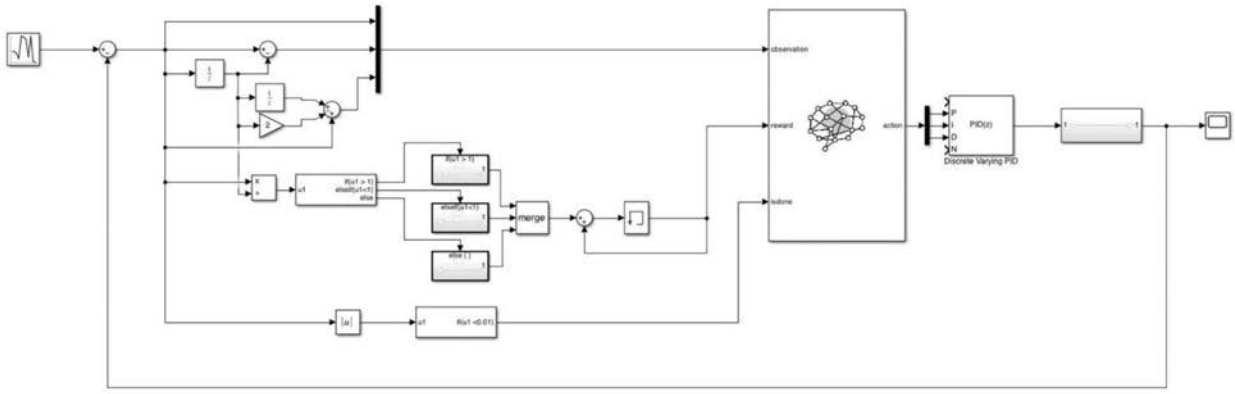

observation
reward
isdone
action
PID(z)
Discrete Varying PID
merge

图7

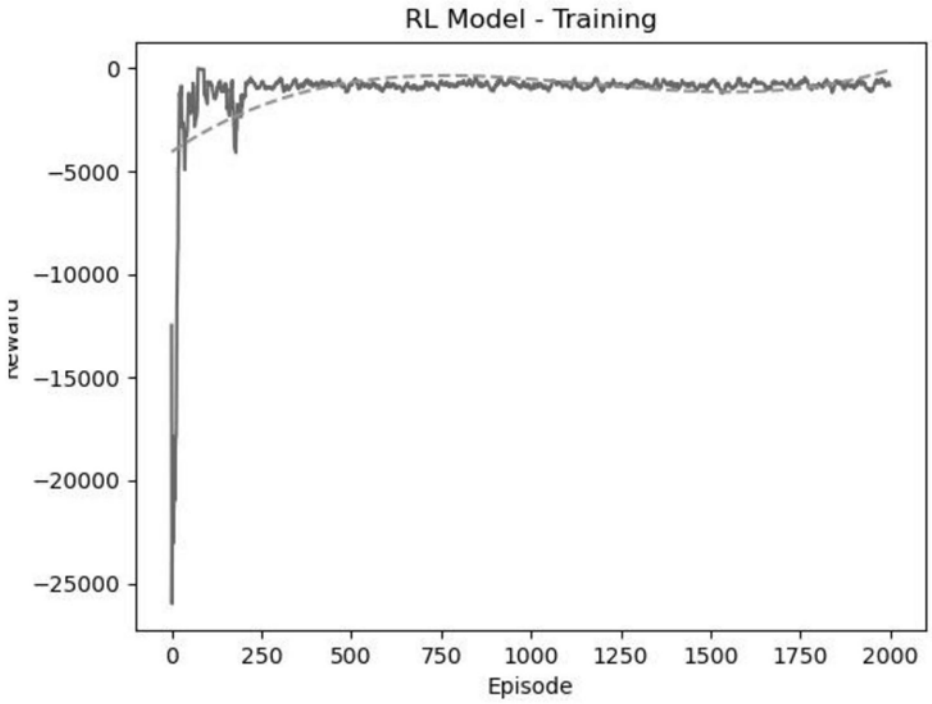
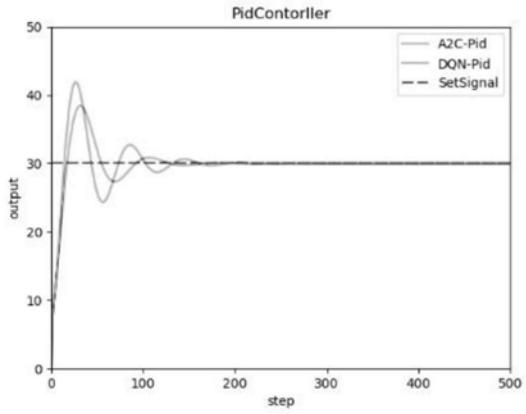
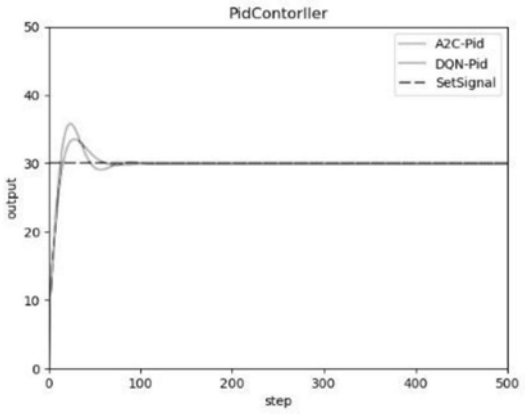

RL Model - Training
reward
0
−5000
−10000
−15000
−20000
−25000
0
250
500
750
1000
1250
1500
1750
2000
Episode

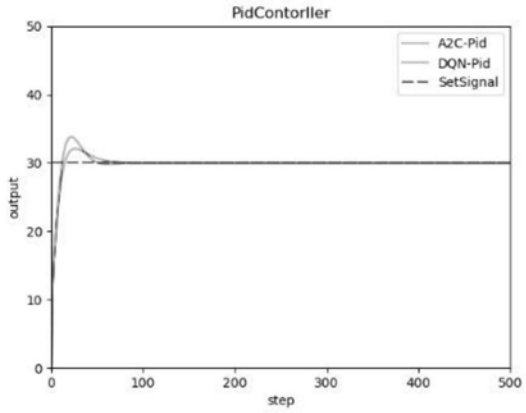
图8



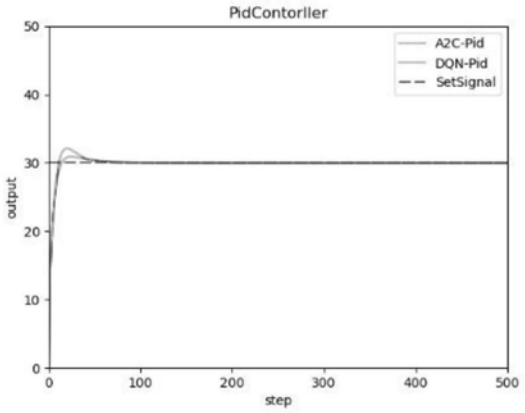
(a) 50 episodes



(b) 100 episodes



(c) 200 episodes



(d) 2000 episodes

图9