控制与决策
Control and Decision

基于强化学习的倒立摆分数阶梯度下降RBF控制

薛晗， 邵哲平， 方琼林， 刘晓佳

引用本文:
薛晗, 邵哲平, 方琼林, 等. 基于强化学习的倒立摆分数阶梯度下降RBF控制[J]. 控制与决策, 2021, 36(1): 125−134.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2019.0816

基于强化学习的倒立摆分数阶梯度下降RBF控制

薛 晗†, 邵哲平, 方琼林, 刘晓佳

(集美大学 航海学院，福建 厦门 361021)

摘 要: 为了提高强化学习的控制性能,提出一种基于分数梯度下降RBF神经网络的强化学习算法. 通过评价神经网络和执行神经网络组成强化学习系统,利用神经网络记忆和联想,学会控制倒立摆,提高控制精度,使误差趋于零,直至学习成功,并证明闭环系统的稳定性. 通过倒立摆的物理实验发现,当分数阶阶数较大,微分的作用更显著,对角速度和速度的控制效果更好,角速度和速度的均方误差和平均绝对误差较小;当分数阶阶数较小,积分的作用更显著,对倾斜角和位移的控制效果更好,因此倾斜角和位移的均方误差和平均绝对误差较小. 仿真实验的结果表明,所提算法动态响应好,超调量小,调整时间短,精度高,泛化性能好. 它优于基于RBF神经网络的强化学习算法和传统强化学习算法,能有效地加快梯度下降法的收敛速度,提高其控制性能. 在引入适当的干扰后,所提算法能够快速地自我调节并恢复稳定状态,控制器的鲁棒性和动态性能满足实际要求.

关键词: 强化学习；径向基神经网络；倒立摆；分数阶；梯度下降；神经网络控制

中图分类号: TP18 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2019.0816 **开放科学(资源服务)标识码(OSID):**

引用格式: 薛晗, 邵哲平, 方琼林, 等. 基于强化学习的倒立摆分数阶梯度下降RBF控制 [J]. 控制与决策, 2021, 36(1): 125-134.

Reinforcement learning based fractional gradient descent RBF neural network control of inverted pendulum

XUE Han†, SHAO Zhe-ping, FANG Qiong-lin, LIU Xiao-jia

(Institute of Navigation，Jimei University，Xiamen 361021，China)

Abstract: In order to improve the control performance of reinforcement learning, a reinforcement learning algorithm based on the fractional gradient descent RBF neural network is proposed. Based on the evaluation neural network and action neural network, the reinforcement learning system uses neural network memory and association, and learns to control the inverted pendulum. The control accuracy is improved with the error tending to zero until the learning is successful. The stability of the closed-loop system is proved. The physical experiment of inverted pendulum is carried out. It is pointed that when the fractional order is large, the differential effect is more significant, the control effect of diagonal velocity and velocity is better, and the mean square error and mean absolute error of angular velocity and velocity are smaller. When the fractional order is small, the effect of integral is more significant, and the control effect on tilt angle and displacement is better. The results indicate that the algorithm has good dynamic response, small overshoot, short adjustment time, high precision and good generalization performance. It is superior to the reinforcement learning algorithm based on the RBF neural network and the traditional reinforcement learning algorithm. It can effectively accelerate the convergence speed of the gradient descent method and improve its control performance. After introducing appropriate disturbance, the controller can quickly self-adjust and recover the stable state. The robustness and dynamic performance of the controller meet the actual requirements.

Keywords: reinforcement learning; RBF neural network; inverted pendulum; fractional order; gradient descent; neural network control

0 引 言

非线性控制理论是21世纪控制理论的主旋律,其带来了更先进的控制系统,使自动化水平有很大的飞越[1]. 强化学习(reinforcement learning, RL)学习从环境状态到行为的映射,使得智能体选择的行为能够获得环境最大的奖赏,使得外部环境对学习系统

收稿日期: 2019-06-09；**修回日期:** 2019-08-09.

基金项目: 国家自然科学基金项目(51579114)；福建省自然科学基金项目(2018J05085).

责任编委: 王燕舞.

†通讯作者. E-mail: imlmd@163.com.

的运行性能最佳,已广泛应用于非线性控制领域. 强化学习不要求预先给定任何数据,而是通过接收环境对动作的奖励获得学习信息并更新模型参数. 由评估网络根据外部强化信号对环境建模,提供更有效的内部强化信号给执行神经网络,使它产生更恰当的行动. 内部强化信号使执行神经网络、评估网络在每一步都可以进行学习,而不必等待外部强化信号,从而大大加速了两个网络的学习.

神经网络能通过自身在线学习,有效实现控制. 然而,神经网络存在学习速度慢、容易陷入局部极值、学习记忆不稳定等缺点. 梯度下降法是一种有效的神经网络训练方法[2-4]. 分数阶微积分是数学的一个重要分支,诞生于1695年. 分数阶控制具有记忆效应,稳定性更好,参数选择范围更大、更灵活[5]. 分数阶梯度下降法由于其广泛的适用性而受到研究者的关注. Wang等[6]提出了分数梯度下降法用于神经网络的BP训练,采用Caputo导数对传统二次能量函数定义的误差分数阶梯度进行了估计; Khan等[7]为径向基函数神经网络提出了一种新的分数阶梯度下降学习算法,采用基于分数梯度下降法和改进的黎曼导数的凸组合; Yang等[8]提出了一种分数梯度下降法训练BP神经网络,用卡普托导数估计了传统二次能量函数定义的分数阶梯度的误差; Chen等[9]提出了一种自适应分数阶BP神经网络,结合了竞争进化算法种群极值优化和分数阶梯度下降学习机制来解决手写体数字识别问题.

倒立摆是多变量、高阶次、非线性、强耦合、自然不稳定系统,其稳定控制是控制理论中的典型问题. 近年来,强化学习在倒立摆控制领域得到了发展. Xu等[10]将基于核的近似动态规划应用于非线性系统实时在线学习控制,用于单连杆倒立摆系统和双连杆倒立摆系统; Li等[11]介绍了一种用于连续马尔可夫决策过程的流形正则化强化学习方案; Liu等[12]将在线最小二乘策略迭代法与经验回放法相结合,存储在线生成的样本,并用最小二乘法更新控制策略,应用于倒立摆系统; He等[13]研究了让智能体根据给定的最终目标自适应地内部奖励,并应用于三连杆倒立摆; Xu等[14]将核方法集成到自适应评价设计的评价中,提出了一种新的稀疏核机器自适应评价设计的框架,并应用于倒立摆问题和球板控制.

经典强化学习算法在面对状态空间控制问题时,容易出现维数灾难. 解决这个问题的方法之一是采用近似器来拟合值函数,例如RBF神经网络等. 在已有的RBF网络与强化学习结合的非线性系统中,大多数只是采用基于基本的RBF神经网络,当训练样本增多时,RBF的隐层神经元数增加,使得RBF的复杂度增加,结构庞大,从而运算量也有所增加. 本文将分数阶梯度下降RBF引入到强化学习中,利用分数阶微积分的记忆效应、稳定性好、参数选择范围更大更灵活的优点,能够提高神经网络的训练和计算的准确性,改善强化学习的学习效率,使得控制收敛更快,也更稳定.

1 倒立摆数学模型

设M为车体质量,m为摆质量,b为摩擦系数,l为摆长,I为摆的转动惯量,θ为摆的倾斜角,x为小车的位移,F为作用于小车的输入力,N为小车与摆之间的水平作用力,P为小车与摆之间的垂直方向作用力. 倒立摆的运动模型如图1所示.

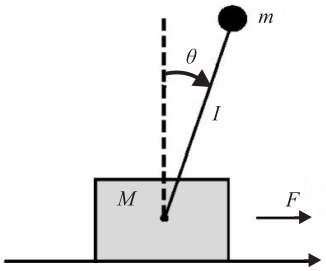


图1 倒立摆运动模型

取

$$\theta = \pi + \phi. \tag{1}$$

θ较小时,$\cos\theta \approx -1$, $\sin\theta \approx -\phi$, $\dot{\theta}^2 \approx 0$,可得

$$\ddot{x} = \frac{-b(I+ml^2)\dot{x} + (I+ml^2)F + m^2l^2\phi g}{I(M+m)+Mml^2}, \tag{2}$$

$$\ddot{\theta} = \frac{-mlb\dot{x} + mlF + (M+m)mgl\phi}{I(M+m)+Mml^2}. \tag{3}$$

令状态变量为

$$X = [x, \dot{x}, \phi, \dot{\phi}]^{\mathrm{T}}. \tag{4}$$

取输入为$u = F$,则倒立摆的动态方程如下:

$$\dot{X} = AX + Bu, \tag{5}$$

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & \dfrac{-b(I+ml^2)}{I(M+m)+Mml^2} & \dfrac{m^2l^2g}{I(M+m)+Mml^2} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & \dfrac{-mbl}{I(M+m)+Mml^2} & \dfrac{(M+m)mgl}{I(M+m)+Mml^2} & 0 \end{bmatrix}, \tag{6}$$

$$B=\frac{1}{I(M+m)+Mml^2}\begin{bmatrix}0\\ I+ml^2\\ 0\\ ml\end{bmatrix}. \tag{7}$$

2　基于分数阶梯度下降RBF神经网络的强化学习算法

基于分数梯度下降RBF神经网络的强化学习算法(fractional gradient descent radial basis function based on reinforcement learning, FGDRBF-RL)由评价神经网络和执行神经网络组成. 神经网络具有学习能力,通过网络记忆和联想,学会控制倒立摆,提高控制经度,使误差趋于零,直至学习成功. 算法流程如下.

step 1: 初始化评价神经网络和执行神经网络的权值以及倒立摆状态;

step 2: 根据倒立摆状态变量,判断回报函数,计算行动量;

step 3: 根据状态变量和行动量,由评价神经网络计算评价函数;

step 4: 将行动量作用于倒立摆,产生下一时刻倒立摆的状态量;

step 5: 更新评价神经网络和执行神经网络权值;

step 6: 循环,调整学习因子,直至学习成功或者学习失败.

FGDRBF-RL的系统结构如图2所示.

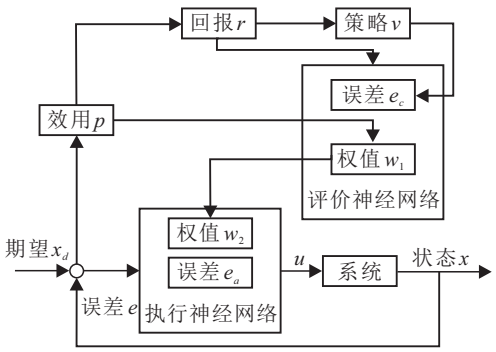


图 2　系统结构

2.1　RBF神经网络

1985年,径向基神经网络(radial basis function, RBF)被提出. 常见的径向基函数为高斯函数,有

$$\varphi_i(x)=\mathrm{e}^{-\frac{||x-x_i||^2}{2\sigma^2}}. \tag{8}$$

其中: x为输入向量, $||x||$为其欧几里得范数, x_i为中心向量, σ_j为径向基函数的宽度. 设输入层神经元数为m,隐含层神经元数为P. 神经网络输出为

$$y=\sum_{i=1}^{P}w_i\varphi_i||x-x_i||. \tag{9}$$

考虑如下系统:

$$x(k+1)=f(x(k))+u(k)+d(k). \tag{10}$$

其中: x为系统状态变量, f为系统函数, u为系统控制输入, d为扰动, x_d为系统期望状态. 系统跟踪误差为

$$e(k+1)=x(k+1)-x_d(k+1). \tag{11}$$

控制目标是使得系统跟踪误差e以及评价神经网络权值误差和执行神经网络的权值误差,能在有限的学习次数内趋近于零.

f由执行神经网络计算如下:

$$f(x(k))=w_2(k)\varphi_2(k)+\varepsilon(k). \tag{12}$$

其中: w_2为执行神经网络的理想权重, φ_2为执行神经网络的激活函数, ε为执行神经网络的估算误差. 设$\hat{w}_2$为执行神经网络权重的估计值,执行神经网络实际输出为

$$\hat{f}(x(k))=\hat{w}_2^{\mathrm{T}}(k)\varphi_2(k). \tag{13}$$

设计系统控制输入如下:

$$u(k)=x_d(k+1)-\hat{f}(x(k))+\Lambda e(k), \tag{14}$$

其中Λ为对角阵增益. 将式(10)代入(11)可得

$$e(k+1)=f(x(k))+u(k)+d(k)-x_d(k+1). \tag{15}$$

将式(14)代入(15),可得

$$e(k+1)=f(x(k))-\hat{f}(x(k))+\Lambda e(k)+d(k). \tag{16}$$

将式(12)代入(16),可得

$$e(k+1)=w_2^{\mathrm{T}}(k)\varphi_2(k)+\varepsilon(k)-\hat{f}(x(k))+\Lambda e(k)+d(k). \tag{17}$$

将式(13)代入(17),可得

$$e(k+1)=w_2^{\mathrm{T}}(k)\varphi_2(k)+\varepsilon(k)-\hat{w}_2^{\mathrm{T}}(k)\varphi_2(k)+\Lambda e(k)+d(k). \tag{18}$$

定义执行神经网络权重估计误差如下:

$$\widetilde{w}_2(k)=\hat{w}_2(k)-w_2. \tag{19}$$

将式(19)代入(18),可得

$$e(k+1)=\Lambda e(k)-\widetilde{w}_2^{\mathrm{T}}\varphi_2(k)+\varepsilon(k)+d(k). \tag{20}$$

2.2　分数阶微积分

Riemann-Liouville分数阶微积分定义如下.

定义1　对于在区间$[t_0,t]$上绝对可积的函数x,其Riemann-Liouville积分为

$${}_{t_0}D_t^{\alpha}x(t)=\frac{1}{\Gamma(\alpha)}\int_{t_0}^{t}(t-\tau)^{\alpha-1}x(\tau)\,\mathrm{d}\tau, \tag{21}$$

其中$\Gamma(x)$定义为

$$\Gamma(x)=\int_0^{\infty}\mathrm{e}^{-t}t^{x-1}\,\mathrm{d}t. \tag{22}$$

定义2　对于在$[t_0,t]$区间上绝对可积的函数x,其Riemann-Liouville微分为

$$ {}_{t_0}D_t^{\alpha}x(t)=\frac{\mathrm{d}^m}{\mathrm{d}t^n}\Big[\frac{\int_{t_0}^{t}(t-\tau)^{m-\alpha-1}x(\tau)\,\mathrm{d}\tau}{\Gamma(m-\alpha)}\Big], \tag{23}$$

其中$\alpha\in[m-1,m)$, m为正整数.

引理1 对于函数$f(x)=(x-x_0)^v, 0\leqslant m\leqslant p<m+1$, 下式成立:

$$ {}_{x_0}D_x^{p}f(x)=\frac{\Gamma(v+1)}{\Gamma(v-p+1)}(x-x_0)^{v-p}. \tag{24}$$

2.3 评价神经网络

采用评价神经网络来估算策略效用. 效用函数定义如下:

$$ p_i(k)=\begin{cases}0,\ e_i^2<\xi;\\ 1,\ e_i^2\geqslant\xi.\end{cases} \tag{25}$$

其中ξ为门限. 当跟踪误差小于门限时, 跟踪性能好, 策略函数取0值; 当跟踪误差大于门限时, 跟踪性能较差, 策略函数取1值. 设V为评价函数, 取策略效用函数如下:

$$ V(k)=\hat{w}_1^{\mathrm{T}}(k)\varphi_1(k). \tag{26}$$

设r为回报函数, γ为折扣因子, N为时间跨度, 则取回报函数如下:

$$ r(k)=\gamma^N p(k). \tag{27}$$

定义评价神经网络的误差如下:

$$ e_c(k)=V(k)-\gamma[V(k-1)-r(k)]. \tag{28}$$

将式(26)和(27)代入(28), 可得

$$ \begin{aligned}&e_c(k)=\\&\hat{w}_1^{\mathrm{T}}(k)\varphi_1(k)-\gamma\hat{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)-\gamma^{N+1}p(k).\end{aligned} \tag{29}$$

评价神经网络的能量函数定义为

$$ E_c(k)=\frac{1}{2}e_c^{\mathrm{T}}(k)e_c(k). \tag{30}$$

对式(30)求偏导, 可得

$$ \frac{\partial^{\alpha}E_c(k)}{\partial\hat{w}_1^{\alpha}(k)}=\frac{\partial E_c(k)}{\partial e_c(k)}\frac{\partial^{\alpha}e_c(k)}{\partial\hat{w}_1^{\alpha}(k)}=\frac{\partial^{\alpha}e_c(k)}{\partial\hat{w}_1^{\alpha}(k)}e_c^{\mathrm{T}}(k). \tag{31}$$

对式(29)求偏导, 可得

$$ \frac{\partial^{\alpha}e_c(k)}{\partial\hat{w}_1^{\alpha}(k)}=\frac{\Gamma(2)}{\Gamma(2-\alpha)}\varphi_1(k)\hat{w}_1^{1-\alpha}(k). \tag{32}$$

将式(32)代入(31), 可得

$$ \frac{\partial^{\alpha}E_c(k)}{\partial\hat{w}_1^{\alpha}(k)}=\frac{\Gamma(2)}{\Gamma(2-\alpha)}\varphi_1(k)\hat{w}_1^{1-\alpha}(k)e_c^{\mathrm{T}}(k). \tag{33}$$

将式(29)代入(33), 可得

$$ \begin{aligned}\frac{\partial^{\alpha}E_c(k)}{\partial\hat{w}_1^{\alpha}(k)}=&\frac{\varphi_1(k)\hat{w}_1^{1-\alpha}(k)}{\Gamma(2-\alpha)}[\hat{w}_1^{\mathrm{T}}(k)\varphi_1(k)-\\&\gamma\hat{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)+\gamma^{N+1}p(k)]^{\mathrm{T}}.\end{aligned} \tag{34}$$

对于评价神经网络, 根据梯度下降方法可得权值更新方法为

$$ \hat{w}_1(k+1)=\hat{w}_1(k)-\eta_1\frac{\partial^{\alpha}E_c(k)}{\partial\hat{w}_1^{\alpha}(k)}, \tag{35}$$

其中η_1为自适应学习速率. η_1在标准RBF中为常数, 实际中很难确定一个从始至终都合适的最佳学习率. 梯度较小时, η_1太小会使训练次数增加, 更新缓慢; 在误差变化剧烈区域, η_1太大会因调整量过大而使训练出现振荡, 迭代次数增加, 甚至收敛过程不稳定. 为了加速收敛过程, 采用自适应学习速率使在整个训练过程中η_1得到合理调节, 适应不同特征数据. 梯度较小时学习率加快, 梯度较大时主动降低更新速度. 设$\eta_{1\max}$为η_1的最大值, 取

$$ \eta_1(k)=\eta_{1\max}\Big/\sqrt{1+\Big[\frac{\partial^{\alpha}E_c(k)}{\partial\hat{w}_1^{\alpha}(k)}\Big]^2}. \tag{36}$$

将式(34)代入(35), 可得

$$ \begin{aligned}&\hat{w}_1(k+1)=\\&\hat{w}_1(k)-\eta_1\frac{\varphi_1(k)\hat{w}_1^{1-\alpha}(k)}{\Gamma(2-\alpha)}\cdot[\hat{w}_1^{\mathrm{T}}(k)\varphi_1(k)-\\&\gamma\hat{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)+\gamma^{N+1}p(k)]^{\mathrm{T}}.\end{aligned} \tag{37}$$

评价神经网络的权重误差定义为

$$ \widetilde{w}_1(k)=\hat{w}_1(k)-w_1, \tag{38}$$

其中w_1为评价神经网络的理想权重.

2.4 执行神经网络

执行神经网络的误差定义如下:

$$ \begin{aligned}e_a(k)=&\hat{w}_1^{\mathrm{T}}(k)\varphi_1(k)+\hat{w}_2^{\mathrm{T}}(k)\varphi_2(k)-\\&w_2^{\mathrm{T}}(k)\varphi_2(k)-\varepsilon(k)-d(k).\end{aligned} \tag{39}$$

执行神经网络的能量函数定义如下:

$$ E_a(k)=\frac{1}{2}e_a^{\mathrm{T}}(k)e_a(k). \tag{40}$$

对式(40)求偏导, 可得

$$ \frac{\partial^{\alpha}E_a(k)}{\partial\hat{w}_2^{\alpha}(k)}=\frac{\partial E_a(k)}{\partial e_a(k)}\frac{\partial^{\alpha}e_a(k)}{\partial\hat{w}_2^{\alpha}(k)}=\frac{\partial^{\alpha}e_a(k)}{\partial\hat{w}_2^{\alpha}(k)}e_a^{\mathrm{T}}(k). \tag{41}$$

对式(39)求偏导, 可得

$$ \frac{\partial^{\alpha}e_a(k)}{\partial\hat{w}_2^{\alpha}(k)}=\frac{\Gamma(2)}{\Gamma(2-\alpha)}\varphi_2(k)\hat{w}_2^{1-\alpha}(k). \tag{42}$$

将式(42)代入(41), 可得

$$ \frac{\partial^{\alpha}E_a(k)}{\partial\hat{w}_2^{\alpha}(k)}=\frac{\Gamma(2)}{\Gamma(2-\alpha)}\varphi_2(k)\hat{w}_2^{1-\alpha}(k)e_a^{\mathrm{T}}(k). \tag{43}$$

将式(39)代入(43), 可得

$$ \begin{aligned}\frac{\partial^{\alpha}E_a(k)}{\partial\hat{w}_2^{\alpha}(k)}=&\frac{\varphi_2(k)\hat{w}_2^{1-\alpha}(k)}{\Gamma(2-\alpha)}[\hat{w}_1^{\mathrm{T}}(k)\varphi_1(k)+\\&\widetilde{w}_2^{\mathrm{T}}(k)\varphi_2(k)-\varepsilon(k)-d(k)]^{\mathrm{T}}.\end{aligned} \tag{44}$$

对于执行神经网络,根据梯度下降方法可得权值更新方法为

$$\hat{w}_2(k+1)=\hat{w}_2(k)-\eta_2\frac{\partial^\alpha E_a(k)}{\partial\hat{w}_2^\alpha(k)},\tag{45}$$

其中η_2为执行神经网络的自适应增益,取

$$\eta_2(k)=\eta_{2\max}\Big/\sqrt{1+\Big[\frac{\partial^\alpha E_a(k)}{\partial\hat{w}_2^\alpha(k)}\Big]^2}.\tag{46}$$

将式(44)代入(45),可得

$$\begin{aligned}&\hat{w}_2(k+1)=\\&\hat{w}_2(k)-\eta_2\frac{\varphi_2(k)\hat{w}_2^{1-\alpha}(k)}{\Gamma(2-\alpha)}\cdot[\hat{w}_1^{\mathrm{T}}(k)\varphi_1(k)+\\&\widetilde{w}_2^{\mathrm{T}}(k)\varphi_2(k)-\varepsilon(k)-d(k)]^{\mathrm{T}}.\end{aligned}\tag{47}$$

2.5 FGDRBF-RL收敛性分析

引理2[15-16]　对于系统

$$x(k+1)=f(x(k),k)+d(k),\tag{48}$$

若存在函数$L(x(k),k)$对在紧集$\boldsymbol{R}^n$的x正定,且有

$$L(x(k),k)>0,\tag{49}$$

$$\Delta L(x(k),k)<0,\ ||x||>R>0,\tag{50}$$

使得以R为半径的球在S内,则系统一致最终有界(uniformly ultimately bounded, UUB),x的范数在邻域R内有界.

定理1　对于FGDRBF-RL,评价神经网络的权重按式(37)更新,执行神经网络的权值按式(47)更新. 若假设下式成立,则系统跟踪误差e、评价神经网络权值误差$\widetilde{w}_1$和执行神经网络的权值误差$\widetilde{w}_2$一致最终有界:

$$0<\Lambda_{\max}<\frac{\sqrt{3}}{3},\tag{51}$$

$$\frac{\eta_2\varphi_2^2(k)\hat{w}_2^{1-\alpha}(k)}{\Gamma(2-\alpha)}<1,\tag{52}$$

$$\frac{\eta_1\varphi_1(k)\hat{w}_1^{1-\alpha}(k)}{\Gamma(2-\alpha)}<1,\tag{53}$$

$$\lambda_2<\frac{\Gamma(2-\alpha)}{2\varphi_1(k)\hat{w}_1^{1-\alpha}(k)\gamma^2},\tag{54}$$

$$\frac{\hat{w}_2^{1-\alpha}(k)}{\lambda_3\Gamma(2-\alpha)}>\frac{3}{\lambda_1},\tag{55}$$

$$\frac{\hat{w}_2^{1-\alpha}(k)}{\Gamma(2-\alpha)\varphi_1(k)}>\frac{1}{\lambda_2}+\frac{2\hat{w}_2^{1-\alpha}(k)}{\lambda_3\Gamma(2-\alpha)},\tag{56}$$

其中$\lambda_1>0$, $\lambda_2>0$, $\lambda_3>0$为系数.

证明　定义Lyapunov函数如下:

$$\begin{aligned}L(k)=&\frac{1}{\lambda_1}e^{\mathrm{T}}(k)e(k)+\frac{1}{\eta_1}\widetilde{w}_1^{\mathrm{T}}(k)\widetilde{w}_1(k)+\\&\frac{1}{\lambda_2}||\widetilde{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)||^2+\frac{\widetilde{w}_2^{\mathrm{T}}(k)\widetilde{w}_2(k)}{\lambda_3\eta_2}.\end{aligned}\tag{57}$$

其差分可按下式计算:

$$\Delta L(k)=L(k+1)-L(k).\tag{58}$$

将式(57)代入(58),可得

$$\begin{aligned}&\Delta L(k)=\\&\frac{e^{\mathrm{T}}(k+1)e(k+1)-e^{\mathrm{T}}(k)e(k)}{\lambda_1}+\\&\frac{1}{\eta_1}[\widetilde{w}_1^{\mathrm{T}}(k+1)\widetilde{w}_1(k+1)-\widetilde{w}_1^{\mathrm{T}}(k)\widetilde{w}_1(k)]+\\&\frac{||\widetilde{w}_1^{\mathrm{T}}(k)\varphi_1(k)||^2-||\widetilde{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)||^2}{\lambda_2}+\\&\frac{1}{\lambda_3\eta_2}[\widetilde{w}_2^{\mathrm{T}}(k+1)\widetilde{w}_2(k+1)-\widetilde{w}_2^{\mathrm{T}}(k)\widetilde{w}_2(k)]=\\&L_1+L_2+L_3+L_4,\end{aligned}\tag{59}$$

其中$L_1\sim L_4$分别为式(59)的4项. 将式(20)代入L_1,由$\forall a,b,c,(a+b+c)^2\leqslant 3a^2+3b^2+3c^2$,可得

$$\begin{aligned}&L_1=\\&\frac{||\Lambda e(k)-\widetilde{w}_2^{\mathrm{T}}\varphi_2(k)+\varepsilon(k)+d(k)||^2}{\lambda_1}-\frac{1}{\lambda_1}||e(k)||^2\leqslant\\&\frac{3\Lambda_{\max}^2-1}{\lambda_1}||e(k)||^2+\frac{3}{\lambda_1}||\widetilde{w}_2^{\mathrm{T}}\varphi_2(k)||^2+\\&\frac{3}{\lambda_1}||\varepsilon(k)+d(k)||^2,\end{aligned}\tag{60}$$

其中$\Lambda_{\max}$为Λ的最大特征值.

将式(37)和(38)化简,可得

$$\begin{aligned}&\widetilde{w}_1(k+1)=\\&\Big[1-\frac{\eta_1\varphi_1(k)\hat{w}_1^{1-\alpha}(k)}{\Gamma(2-\alpha)}\Big]\widetilde{w}_1(k)-\\&\eta_1\frac{\varphi_1(k)\hat{w}_1^{1-\alpha}(k)}{\Gamma(2-\alpha)}[w_1^{\mathrm{T}}(k)\varphi_1(k)-\\&\gamma\hat{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)+\gamma^{N+1}p(k)]^{\mathrm{T}}.\end{aligned}\tag{61}$$

计算L_2,化简可得

$$\begin{aligned}&L_2=\\&-2\frac{\varphi_1(k)\hat{w}_1^{1-\alpha}(k)}{\Gamma(2-\alpha)}||\widetilde{w}_1(k)||^2-\\&2\widetilde{w}_1^{\mathrm{T}}(k)\frac{\varphi_1(k)\hat{w}_1^{1-\alpha}(k)}{\Gamma(2-\alpha)}[w_1^{\mathrm{T}}(k)\varphi_1(k)-\\&\gamma\hat{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)+\gamma^{N+1}p(k)]^{\mathrm{T}}+\\&\frac{\eta_1\varphi_1^2(k)\hat{w}_1^{2-2\alpha}(k)}{\Gamma^2(2-\alpha)}||\widetilde{w}_1^{\mathrm{T}}(k)+w_1^{\mathrm{T}}(k)\varphi_1(k)-\\&\gamma\hat{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)+\gamma^{N+1}p(k)||^2.\end{aligned}\tag{62}$$

由$\forall a,b,2ab=(a+b)^2-a^2-b^2$,可得

$$\begin{aligned}&2\widetilde{w}_1(k)[w_1^{\mathrm{T}}(k)\varphi_1(k)-\\&\gamma\hat{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)+\gamma^{N+1}p(k)]^{\mathrm{T}}=\\&||\widetilde{w}_1^{\mathrm{T}}(k)+w_1^{\mathrm{T}}(k)\varphi_1(k)-\\&\gamma\hat{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)+\gamma^{N+1}p(k)||^2-\end{aligned}$$

$$||\widetilde{w}_1(k)||^2-||w_1^{\mathrm{T}}(k)\varphi_1(k)-\gamma\hat{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)+\gamma^{N+1}p(k)||^2. \tag{63}$$

将式(63)代入(62),化简可得

$$\begin{aligned}L_2=&-\frac{\varphi_1(k)\hat{w}_1^{1-\alpha}(k)}{\Gamma(2-\alpha)}||\widetilde{w}_1(k)||^2+\\&\frac{\varphi_1(k)\hat{w}_1^{1-\alpha}(k)}{\Gamma(2-\alpha)}||w_1^{\mathrm{T}}(k)\varphi_1(k)-\\&\gamma\hat{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)+\gamma^{N+1}p(k)||^2-\\&\frac{\varphi_1(k)\hat{w}_1^{1-\alpha}(k)}{\Gamma(2-\alpha)}\Big[1-\frac{\eta_1\varphi_1(k)\hat{w}_1^{1-\alpha}(k)}{\Gamma(2-\alpha)}\Big]\cdot\\&||\widetilde{w}_1(k)+w_1^{\mathrm{T}}(k)\varphi_1(k)-\\&\gamma\hat{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)+\gamma^{N+1}p(k)||^2.\end{aligned} \tag{64}$$

由$\forall a,\ b,\ (a+b)^2\leqslant 2a^2+2b^2$,根据式(38)可得

$$\begin{aligned}&||w_1^{\mathrm{T}}(k)\varphi_1(k)-\gamma\hat{w}_1^{\mathrm{T}}(k-\\&1)\varphi_1(k-1)+\gamma^{N+1}p(k)||^2=\\&||w_1^{\mathrm{T}}(k)\varphi_1(k)-\gamma w_1^{\mathrm{T}}\varphi_1(k-1)-\\&\gamma\widetilde{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)+\gamma^{N+1}p(k)||^2\leqslant\\&2||w_1^{\mathrm{T}}(k)\varphi_1(k)-\gamma w_1^{\mathrm{T}}\varphi_1(k-1)+\\&\gamma^{N+1}p(k)||^2+2\gamma^2||\widetilde{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)||^2.\end{aligned} \tag{65}$$

将式(65)代入(64),可得

$$\begin{aligned}&L_2\leqslant\\&-\frac{\varphi_1(k)\hat{w}_1^{1-\alpha}(k)}{\Gamma(2-\alpha)}\Big[1-\frac{\eta_1\varphi_1(k)\hat{w}_1^{1-\alpha}(k)}{\Gamma(2-\alpha)}\Big]\cdot||\widetilde{w}_1(k)+\\&w_1^{\mathrm{T}}(k)\varphi_1(k)-\gamma\hat{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)+\\&\gamma^{N+1}p(k)||^2-\frac{\varphi_1(k)\hat{w}_1^{1-\alpha}(k)}{\Gamma(2-\alpha)}||\widetilde{w}_1(k)||^2+\\&2\frac{\varphi_1(k)\hat{w}_1^{1-\alpha}(k)}{\Gamma(2-\alpha)}\cdot[||w_1^{\mathrm{T}}(k)\varphi_1(k)-\gamma w_1^{\mathrm{T}}\varphi_1(k-1)+\\&\gamma^{N+1}p(k)||^2+\gamma^2||\widetilde{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)||^2].\end{aligned} \tag{66}$$

将式(47)代入(19),可得

$$\begin{aligned}&\widetilde{w}_2(k+1)=\\&\widetilde{w}_2(k)-\frac{\eta_2\varphi_2(k)\hat{w}_2^{1-\alpha}(k)}{\Gamma(2-\alpha)}\cdot[\hat{w}_1^{\mathrm{T}}(k)\varphi_1(k)+\\&\widetilde{w}_2^{\mathrm{T}}(k)\varphi_2(k)-\varepsilon(k)-d(k)]^{\mathrm{T}}.\end{aligned} \tag{67}$$

计算L_4,由式(67)化简可得

$$\begin{aligned}&L_4=\\&\frac{\eta_2\varphi_2^2(k)\hat{w}_2^{2-2\alpha}(k)}{\lambda_3\Gamma^2(2-\alpha)}||\widetilde{w}_2^{\mathrm{T}}(k)\varphi_2(k)+\\&\hat{w}_1^{\mathrm{T}}(k)\varphi_1(k)-\varepsilon(k)-d(k)||^2-\\&\frac{2\hat{w}_2^{1-\alpha}(k)\widetilde{w}_2^{\mathrm{T}}(k)\varphi_2(k)}{\lambda_3\Gamma(2-\alpha)}[\widetilde{w}_2^{\mathrm{T}}(k)\varphi_2(k)+\\&\hat{w}_1^{\mathrm{T}}(k)\varphi_1(k)-\varepsilon(k)-d(k)]^{\mathrm{T}}.\end{aligned} \tag{68}$$

由$\forall a,b,-2ab=-(a+b)^2+a^2+b^2$,可得

$$\begin{aligned}&-2\widetilde{w}_2^{\mathrm{T}}(k)\varphi_2(k)[\hat{w}_1^{\mathrm{T}}(k)\varphi_1(k)-\varepsilon(k)-d(k)]^{\mathrm{T}}=\\&-||\widetilde{w}_2^{\mathrm{T}}(k)\varphi_2(k)+\hat{w}_1^{\mathrm{T}}(k)\varphi_1(k)-\varepsilon(k)-d(k)||^2+\\&||\widetilde{w}_2^{\mathrm{T}}(k)\varphi_2(k)||^2+||\hat{w}_1^{\mathrm{T}}(k)\varphi_1(k)-\varepsilon(k)-d(k)||^2.\end{aligned} \tag{69}$$

将式(69)代入(68),化简可得

$$\begin{aligned}L_4=&-\frac{\hat{w}_2^{1-\alpha}(k)}{\lambda_3\Gamma(2-\alpha)}\Big[1-\frac{\eta_2\varphi_2^2(k)\hat{w}_2^{1-\alpha}(k)}{\Gamma(2-\alpha)}\Big]\cdot\\&||\widetilde{w}_2^{\mathrm{T}}(k)\varphi_2(k)+\hat{w}_1^{\mathrm{T}}(k)\varphi_1(k)-\varepsilon(k)-\\&d(k)||^2-\frac{\varphi_2^2(k)\hat{w}_2^{1-\alpha}(k)||\widetilde{w}_2(k)||^2}{\lambda_3\Gamma(2-\alpha)}+\\&\frac{\hat{w}_2^{1-\alpha}(k)}{\lambda_3\Gamma(2-\alpha)}||\hat{w}_1^{\mathrm{T}}(k)\varphi_1(k)-\varepsilon(k)-d(k)||^2.\end{aligned} \tag{70}$$

由$\forall a,b,(a+b)^2\leqslant 2a^2+2b^2$,根据式(38),可得

$$\begin{aligned}&||\hat{w}_1^{\mathrm{T}}(k)\varphi_1(k)-\varepsilon(k)-d(k)||^2=\\&||\widetilde{w}_1^{\mathrm{T}}(k)\varphi_1(k)+w_1^{\mathrm{T}}\varphi_1(k)-\varepsilon(k)-d(k)||^2\leqslant\\&2||\widetilde{w}_1^{\mathrm{T}}(k)\varphi_1(k)||^2+2||w_1^{\mathrm{T}}\varphi_1(k)-\varepsilon(k)-d(k)||^2.\end{aligned} \tag{71}$$

将式(71)代入(70)，可得

$$\begin{aligned}L_4\leqslant&-\frac{\hat{w}_2^{1-\alpha}(k)}{\lambda_3\Gamma(2-\alpha)}\Big[1-\frac{\eta_2\varphi_2^2(k)\hat{w}_2^{1-\alpha}(k)}{\Gamma(2-\alpha)}\Big]\cdot\\&||\widetilde{w}_2^{\mathrm{T}}(k)\varphi_2(k)+\hat{w}_1^{\mathrm{T}}(k)\varphi_1(k)-\varepsilon(k)-\\&d(k)||^2-\frac{\varphi_2^2(k)\hat{w}_2^{1-\alpha}(k)||\widetilde{w}_2(k)||^2}{\lambda_3\Gamma(2-\alpha)}+\\&\frac{\hat{w}_2^{1-\alpha}(k)}{\lambda_3\Gamma(2-\alpha)}[2||\widetilde{w}_1^{\mathrm{T}}(k)\varphi_1(k)||^2+\\&2||w_1^{\mathrm{T}}\varphi_1(k)-\varepsilon(k)-d(k)||^2].\end{aligned} \tag{72}$$

记

$$\begin{aligned}&\Omega^2=\\&\frac{3}{\lambda_1}||\varepsilon(k)+d(k)||^2+\frac{2\varphi_1(k)\hat{w}^{1-\alpha}(k)}{\Gamma(2-\alpha)}\cdot\\&||w_1^{\mathrm{T}}(k)\varphi_1(k)-\gamma w_1^{\mathrm{T}}\varphi_1(k-1)+\gamma^{N+1}p(k)||^2+\\&\frac{2\hat{w}_2^{1-\alpha}(k)}{\lambda_3\Gamma(2-\alpha)}||w_1^{\mathrm{T}}\varphi_1(k)-\varepsilon(k)-d(k)||^2.\end{aligned} \tag{73}$$

联合式(59)、(60)、(66)、(72)和(73),可得

$$\begin{aligned}&\Delta L(k)\leqslant\\&-\frac{1-3\Lambda_{\max}^2}{\lambda_1}||e(k)||^2-\Big(\frac{\hat{w}_2^{1-\alpha}(k)}{\lambda_3\Gamma(2-\alpha)}-\\&\frac{3}{\lambda_1}\Big)||\widetilde{w}_2^{\mathrm{T}}(k)\varphi_2(k)||^2-\Big[\frac{\hat{w}_2^{1-\alpha}(k)}{\Gamma(2-\alpha)\varphi_1(k)}-\\&\frac{1}{\lambda_2}-\frac{2\hat{w}_2^{1-\alpha}(k)}{\lambda_3\Gamma(2-\alpha)}\Big]\cdot||\widetilde{w}_1^{\mathrm{T}}(k)\varphi_1(k)||^2-\end{aligned}$$

$$\Big[\frac{1}{\lambda_2}-\frac{2\varphi_1(k)\hat{w}_2^{1-\alpha}(k)\gamma^2}{\Gamma(2-\alpha)}\Big]\cdot||\widetilde{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)||^2-\frac{\varphi_1(k)\hat{w}_1^{1-\alpha}(k)}{\Gamma(2-\alpha)}\cdot\Big[1-\frac{\eta_1\varphi_1(k)\hat{w}_1^{1-\alpha}(k)}{\Gamma(2-\alpha)}\Big]||\widetilde{w}_1(k)+\gamma^{N+1}p(k)+w_1^{\mathrm{T}}(k)\varphi_1(k)-\gamma\hat{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)||^2-\frac{\hat{w}_2^{1-\alpha}(k)}{\lambda_3\Gamma(2-\alpha)}\Big[1-\frac{\eta_2\varphi_2^2(k)\hat{w}_1^{1-\alpha}(k)}{\Gamma(2-\alpha)}\Big]\cdot||\widetilde{w}_2^{\mathrm{T}}(k)\varphi_2(k)+\hat{w}_1^{\mathrm{T}}(k)\varphi_1(k)-\varepsilon(k)-d(k)||^2+\Omega^2. \tag{74}$$

记w_1的最大值为$w_{1\max}$,φ_1的最大值为$\varphi_{1\max}$,p的最大值为$p_{\max}$,d的最大值为$d_{\max}$,ε的最大值为$\varepsilon_{\max}$,则Ω的最大值$\Omega_{\max}$计算如下:

$$\Omega^2\leqslant\frac{6}{\lambda_1}(\varepsilon_{\max}^2+d_{\max}^2)+\frac{6\varphi_1(k)\hat{w}_2^{1-\alpha}(k)}{\Gamma(2-\alpha)}\cdot[(1+\gamma^2)w_{1\max}^2\varphi_{1\max}^2+\gamma^{2N+2}p_{\max}^2]+\frac{6\hat{w}_2^{1-\alpha}(k)}{\lambda_3\Gamma(2-\alpha)}(w_{1\max}^2\varphi_{1\max}^2+\varepsilon_{\max}^2+d_{\max}^2)=\frac{6\varphi_1(k)\hat{w}_1^{1-\alpha}(k)}{\Gamma(2-\alpha)}\gamma^{2N+2}p_{\max}^2+\Big[\frac{6}{\lambda_1}+\frac{6\hat{w}_2^{1-\alpha}(k)}{\lambda_3\Gamma(2-\alpha)}\Big](\varepsilon_{\max}^2+d_{\max}^2)+\frac{6\hat{w}_2^{1-\alpha}(k)}{\Gamma(2-\alpha)}\cdot\Big[\varphi_1(k)(1+\gamma^2)+\frac{1}{\lambda_3}\Big]w_{1\max}^2\varphi_{1\max}^2=\Omega_{\max}^2. \tag{75}$$

当式(51)～(56)成立时,若以下任一公式成立,则有$\Delta V(k)\leqslant 0$:

$$||e(k)||>\sqrt{\frac{\lambda_1}{1-3\Lambda_{\max}^2}}\Omega_{\max}, \tag{76}$$

$$||\widetilde{w}_1^{\mathrm{T}}(k)\varphi_1(k)||>\frac{\Omega_{\max}}{\sqrt{\dfrac{\hat{w}_2^{1-\alpha}(k)}{\Gamma(2-\alpha)\varphi_1(k)}-\dfrac{1}{\lambda_2}-\dfrac{2\hat{w}_2^{1-\alpha}(k)}{\lambda_3\Gamma(2-\alpha)}}}, \tag{77}$$

$$||\widetilde{w}_2^{\mathrm{T}}(k)\varphi_2(k)||>\frac{\Omega_{\max}}{\sqrt{\dfrac{\hat{w}_2^{1-\alpha}(k)}{\lambda_3\Gamma(2-\alpha)}-\dfrac{3}{\lambda_1}}}, \tag{78}$$

$$||\widetilde{w}_1^{\mathrm{T}}(k-1)\varphi_1(k-1)||>\frac{\Omega_{\max}}{\sqrt{\dfrac{1}{\lambda_2}-\dfrac{2\varphi_1(k)\hat{w}_2^{1-\alpha}(k)\gamma^2}{\Gamma(2-\alpha)}}}. \tag{79}$$

由引理2,跟踪误差e、评价神经网络权值估计误差$\widetilde{w}_1$和执行神经网络权值估计误差$\widetilde{w}_2$一致最终有界. □

3　仿真分析

3.1　实例介绍

为了验证控制效果,采用倒立摆进行实验. 实验中倒立摆的主要参数见表1.

表1　倒立摆参数

参数	数值
倒立摆参数/m	0.335
小车质量/kg	1.031
摆杆质量/kg	0.184
小车长度/m	0.24
小车宽度/m	0.185
总高度/m	0.46
角位移传感器供电电压/V	3.3～5
电机供电电压/V	7～13
编码器工作电压/V	5

3.2　实验结果

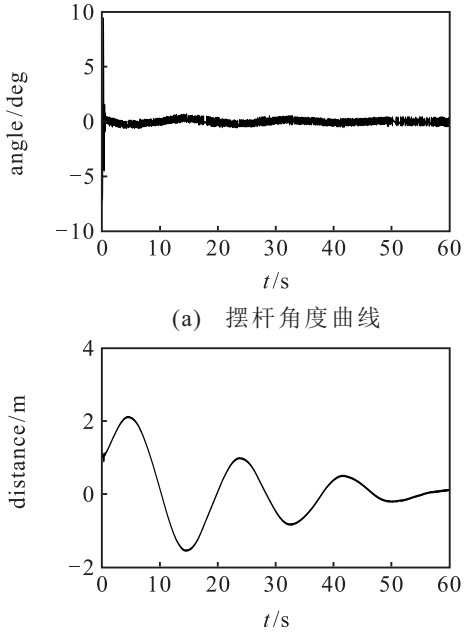
实验测试于Intel(R) Core(TM) i3-4150T CPU @ 3.00 GHz,内存4.00 GB的64位操作系统、基于x64的处理器上. 参数设置如下:折扣因子为0.9,评价神经网络隐含层神经元数量为5,执行神经网络隐含层神经元数量为4,初始倾斜角为1 deg,初始角速度为1 deg/s,初始位移为−0.5 m,初始速度为0 m/s. 期望的倾斜角设为0 deg,期望的角速度为0 deg/s,期望的位移为0 m,期望的速度为0 m/s. 系统参数如表2所示.

表2　强化学习控制系统参数

参数	数值
折扣因子	0.9
分数阶	0.9
初始评价神经网络学习率	0.5
初始执行神经网络学习率	0.5
位移跟踪误差门限/m	1
角度跟踪误差门限/deg	15
评价神经网络输入层神经元数	4
评价神经网络隐含层神经元数	6
执行神经网络输入层神经元数	5
执行神经网络隐含层神经元数	6
最大学习次数	1 000
每次最大平衡步数	20 000

评价神经网络和执行神经网络同时在线更新. 每次实验,当倒立摆的学习次数超过最大学习次数1 000次或者每次学习的平衡步数超过每次最大平衡步数20 000时,结束倒立摆的本轮学习,重新开始下一轮的学习过程. 若倒立摆在一次学习过程中能保持20 000步不倒下,认为本次学习能成功控制倒立摆运动平衡. 统计20轮学习实验成功所需的平均学习次数. 结果表明,采用FGDRBF-RL算法,平均25次学习后能成功控制倒立摆平衡.

图3为摆杆角度和小车位移曲线. 横轴代表时间,单位为s;上图纵轴代表摆杆倾斜角,单位为deg;下图纵轴代表小车位移,单位为m.

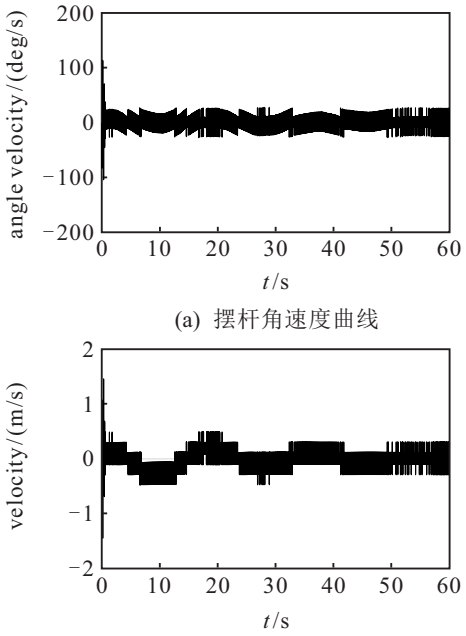


(a) 摆杆角度曲线

(b) 摆杆位移曲线

图3 摆杆角度和小车位移曲线

图4为摆杆角速度和小车速度曲线. 横轴代表时间,单位为s;上图纵轴代表摆杆倾斜角速度,单位为deg/s;下图纵轴代表小车速度,单位为m/s.

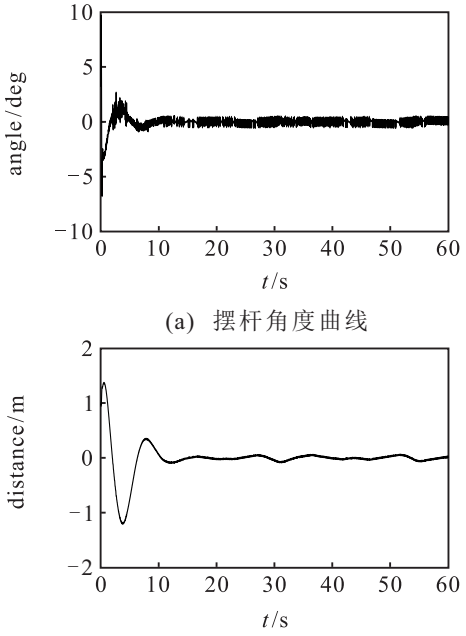


(a) 摆杆角速度曲线

(b) 小车速度曲线

图4 摆杆角速度和小车速度曲线

图5为施加了外界扰动情况下摆杆角度和小车位移曲线. 图中横轴代表时间,单位为s;上图纵轴代表摆杆倾斜角,单位为deg;下图纵轴代表小车位移,单位为m.



(a) 摆杆角度曲线

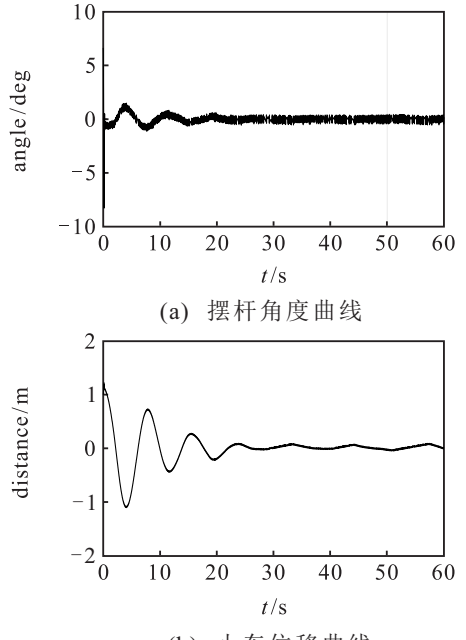
(b) 小车位移曲线

图5 施加外界扰动下摆杆角度和小车位移曲线

由图5可以看出:基于分数阶梯度下降RBF的强化学习,能够使得倒立摆在有外界扰动的情况下实现自主平衡.

3.3 控制算法性能比较

为了验证所提算法的有效性,将本文算法与基本强化学习算法、基于RBF的强化学习算法的控制效果进行比较. 实验环境与3.2节情况相同. 图6为基于RBF的强化学习算法控制的倒立摆摆杆角度和小车位移曲线. 图中横轴代表时间,单位为s;上图纵轴代表摆杆倾斜角,单位为deg;下图纵轴代表小车位移,单位为m.

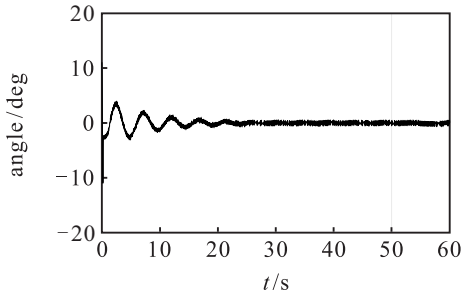


(a) 摆杆角度曲线

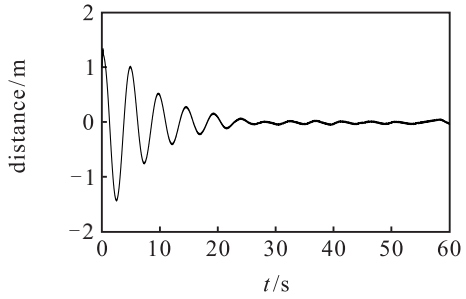
(b) 小车位移曲线

图6 RBF-RL的摆杆角度和小车位移曲线

图7为基于强化学习算法控制的倒立摆摆杆角度和小车位移曲线. 图中横轴代表时间,单位为s;上图纵轴代表摆杆倾斜角,单位为deg;下图纵轴代表小车位移,单位为m.



(a) 摆杆角度曲线



(b) 小车位移曲线

图7　RL的摆杆角度和小车位移曲线

在相同的仿真条件下,表3对比了不同强化学习系统的控制性能,列出了20轮学习实验成功所需的平均学习次数.

表3　不同强化学习系统性能比较

强化学习系统	学习次数
RL	192
RBF-RL	143
FGDRBF-RL	25

由此可见,本算法的学习效率高,收敛速度快,在控制性能和学习速度上优于基本强化学习和基于神经网络的强化学习算法,能够在常规神经网络的基础上进一步改善控制性能,得到更快的速度和更高的控制精确度,使系统在有限时间内收敛.

3.4　分数阶参数影响分析

为了验证分数阶对算法计算性能的影响,针对3.2节运动情况,对分数阶参数取不同的数值,其余参数不变,分别比较其对控制效果的影响. 不同分数阶微积分阶数下摆杆倾斜角、角速度、小车位移、小车速度的均方误差见表4.

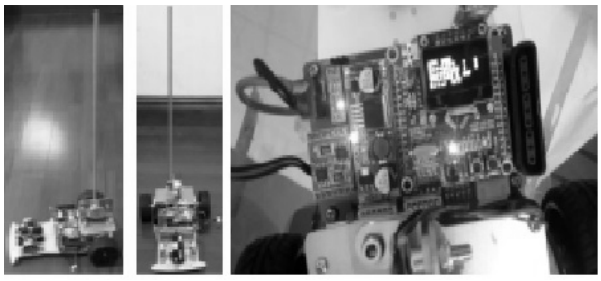
由表4可见: 当分数阶阶数较大时,系统倾斜角和位移的均方误差和平均绝对误差较大,角速度和速度的均方误差和平均绝对误差较小. 当分数阶阶数较小时,倾斜角和位移的均方误差和平均绝对误差较小,角速度和速度的均方误差和平均绝对误差较大. 当分数阶阶数较大,微分的作用更显著,对角速度和速度的控制效果更好,因此角速度和速度的均方误差和平均绝对误差较小. 当分数阶阶数较小,积分的作用更显著,对倾斜角和位移的控制效果更好,因此倾斜角和位移的均方误差和平均绝对误差较小. 不同分数阶情况下,调节过程的性能有所不同. 可以根据实际情况,选取不同分数阶,使系统满足不同的动态和静态,有更好的控制效果. 由于整数阶微积分是分数阶微积分的特例,分数阶微积分具有参数选择范围更大、更灵活的优点.

表4　不同分数阶阶数的均方误差计算结果比较

α	MSE(θ)	MSE($\dot{\theta}$)	MSE(x)	MSE($\dot{x}$)
0.9	0.0795	0.0495	0.0164	0.0102
0.8	0.0793	0.0496	0.0164	0.0102
0.7	0.0790	0.0498	0.0163	0.0103
0.6	0.0786	0.0500	0.0162	0.0104
0.5	0.0780	0.0503	0.0160	0.0105
0.4	0.0772	0.0508	0.0158	0.0107
0.3	0.0761	0.0514	0.0155	0.0110
0.2	0.0748	0.0522	0.0151	0.0114
0.1	0.0732	0.0533	0.0146	0.0119
−0.1	0.0689	0.0562	0.0133	0.0134
−0.2	0.0662	0.0583	0.0125	0.0147
−0.3	0.0632	0.0608	0.0115	0.0165

4　实例验证

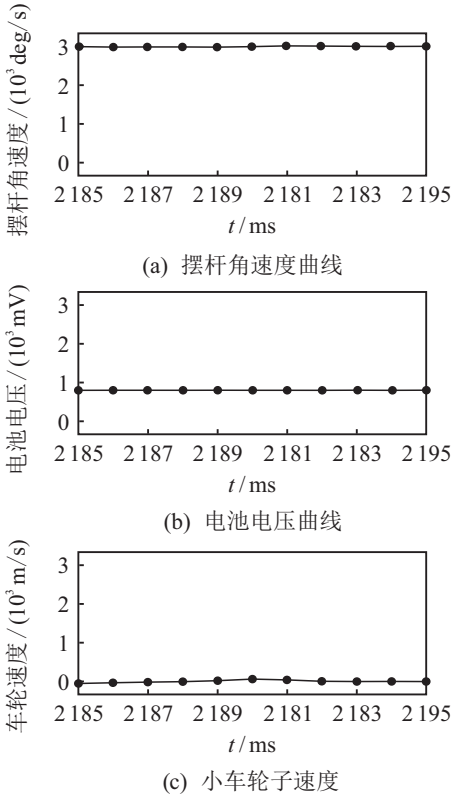
倒立摆实物实验所用倒立摆由stm32主板构成,主要硬件结构包括A4950驱动模块、电量测量模块、稳压模块、蓝牙模块和OLED显示屏. 小车和倒立摆可以在无人干预条件下实现自主平衡. 同时,在引入适量干扰情况下,小车倒立摆能够自主调整并迅速恢复稳定状态. 图8为倒立摆和小车系统的实物图.



(a) 主视图　(b) 左视图　(c) 俯视图

图8　倒立摆系统实物

图9为倒立摆系统输出曲线. 实验表明了该强化学习对小车倒立摆系统控制的有效性. 在引入适量干扰情况下,小车倒立摆能够自主调整并迅速恢复稳定状态.



(a) 摆杆角速度曲线

(b) 电池电压曲线

(c) 小车轮子速度

图 9 倒立摆系统输出曲线

5 结 论

本文提出了一种基于分数阶梯度下降RBF神经网络的强化学习算法,对倒立摆系统进行控制,证明了闭环系统的稳定性. 对控制参数和分数阶微积分的阶数等参数进行比较,分析其对控制效果的影响. 实物实验表明了该强化学习对小车倒立摆系统控制的有效性,控制器的鲁棒性和动态性能满足实际控制的指标要求. 在引入适量干扰情况下,小车倒立摆能够自主调整并迅速恢复稳定状态.

下一步将继续改进本控制算法,进一步提高强化学习系统的控制精度和鲁棒性能.

参考文献(References)

[1] Qiu J B, Sun K K, Wang T, et al. Observer-based fuzzy adaptive event-triggered control for pure-feedback nonlinear systems with prescribed performance[J]. IEEE Transactions on Fuzzy Systems, 2019, 27(11): 2152-2162.

[2] Yin P H, Zhang S, Lyu J C. et al. Blended coarse gradient descent for full quantization of deep neural networks[EB/OL]. (2018-08-15)[2019-01-06]. https://arxiv.org/abs/1808.05240.

[3] Kobayashi M. Gradient descent learning for quaternionic Hopfield neural networks[J]. Neurocomputing, 2017, 260: 174-179.

[4] Wang L N, Yang Y, Min R Q, et al. Accelerating deep neural network training with inconsistent stochastic gradient descent[J]. Neural Networks, 2017, 93: 219-229.

[5] Wang J, Shao C F, Chen Y Q. Fractional order sliding mode control via disturbance observer for a class of fractional order systems with mismatched disturbance[J]. Mechatronics, 2018, 53: 8-19.

[6] Wang J, Wen Y Q, Gou Y D, et al. Fractional-order gradient descent learning of BP neural networks with Caputo derivative[J]. Neural Networks, 2017, 89: 19-30.

[7] Khan S, Naseem I, Malik M A, et al. A fractional gradient descent-based RBF neural network[J]. Circuits, Systems, and Signal Processing, 2018, 37(12): 5311-5332.

[8] Yang G L, Zhang B J, Sang Z Y, et al. A caputo-type fractional-order gradient descent learning of BP neural networks[C]. International Symposium on Neural Networks. Sapporo: Springer, 2017: 547-554.

[9] Chen M R, Chen B P, Zeng G Q, et al. An adaptive fractional-order BP neural network based on extremal optimization for handwritten digits recognition[J]. Neurocomputing, 2020, 391: 260-272.

[10] Xu X, Lian C Q, Zuo L, et al. Kernel-based approximate dynamic programming for real-time online learning control: An experimental study[J]. IEEE Transactions on Control Systems Technology, 2014, 22(1): 146-156.

[11] Li H L, Liu D R, Wang D. Manifold regularized reinforcement learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(4): 932-943.

[12] Liu Q, Zhou X, Zhu F, et al. Experience replay for least-squares policy iteration[J]. IEEE/CAA Journal of Automatica Sinica, 2014, 1(3): 274-281.

[13] He H B, Zhong X N. Learning without external reward[J]. IEEE Computational Intelligence Magazine, 2018, 13(3): 48-54.

[14] Xu X, Hou Z S, Lian C Q, et al. Online learning control using adaptive critic designs with sparse kernel machines[J]. IEEE Transactions on Neural Networks and Learning Systems, 2013, 24(5): 762-775.

[15] Sun K K, Mou S S, Qiu J B, et al. Adaptive fuzzy control for nontriangular structural stochastic switched nonlinear systems with full state constraints[J]. IEEE Transactions on Fuzzy Systems, 2019, 27(8): 1587-1601.

[16] Jagannathan S. Neural network control of nonlinear discrete-time systems[M]. Boca Raton: Taylor and Francis, 2006: 99.

作者简介

薛晗 (1982−), 女, 副教授, 博士, 从事智能控制的研究, E-mail: imlmd@163.com;

邵哲平 (1964−), 男, 教授, 博士, 从事智能交通控制等研究, E-mail: zpshao@jmu.edu.cn;

方琼林 (1978−), 男, 副教授, 硕士, 从事智能交通等研究, E-mail: fql1437@163.com;

刘晓佳 (1979−), 女, 副教授, 博士, 从事智能交通等研究, E-mail: happylxj1314@163.com.

(责任编辑: 齐 霁)