基于 Q-学习的底盘测功机自适应 PID 控制模型

乔 通[1,2],周 洲[1,2],程 鑫[1,2],郭兰英[1],王润民[1,2]
(1. 长安大学 信息工程学院,陕西 西安 710064;
2. 陕西省车联网与智能汽车测试技术工程研究中心,陕西 西安 710064)

摘 要:为了解决汽车底盘测功机控制系统在动态控制时出现延迟较高和误差大的问题,提出了一种基于强化学习的底盘测功机控制策略。以 PID 控制算法为基础,扭力偏差为控制器输入,调节电压控制量为输出,选择扭力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;在底盘测功机仿真试验中验证了控制器的调控性能,并与传统 PID 控制以及神经网络 PID 控制的结果进行了对比;实验结果表明,基于 Q 学习的自适应 PID 控制模型较传统 PID 算法控制周期缩减至 40.7%,相较于神经网络 PID 算法控制周期缩短至 27.9%。相对于传统 PID 控制模型与神经网络 PID 模型,基于 Q 学习的自适应 PID 控制模型输出力上升过程稳定且快速。提出的基于 Q 学习的自适应 PID 控制模型能够有效提升底盘测功机控制精度,满足其使用的工业要求。

关键词:强化学习;PID 控制;Q 学习;控制策略;底盘测功机
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2022)05-0117-06
doi:10.3969/j.issn.1673-629X.2022.05.020

Adaptive PID Control Model of Chassis Dynamometer Based on Q-Learning

QIAO Tong[1,2],ZHOU Zhou[1,2],CHENG Xin[1,2],GUO Lan-ying[1],WANG Run-min[1,2]
(1. School of Information Engineering,Chang'an University,Xi'an 710064,China;
2. Shaanxi Engineering Research Center of Internet of Vehicles and Intelligent Vehicle Testing Technology,
Xi'an 710064,China)

Abstract:In order to solve the problems of high delay and large error in dynamic control of chassis dynamometer control system,a chassis dynamometer control strategy based on reinforcement learning is proposed. Based on the PID control algorithm,the torque deviation is the input of the controller,the control quantity is the output,and the selection of the torque difference is the learning strategy of the intelligent body reward and the penalty,and the PID parameter is adjusted by the Q learning algorithm. In the simulation test of chassis dynamical machine,the control performance of the controller is verified,and the results of the traditional PID control and the neural network PID control are compared. The experimental results show that the control cycle of the adaptive PID control model based on Q learning is reduced to 40.7% compared with the traditional PID algorithm,and the control cycle is shortened by 27.9% compared with the neural network PID algorithm. Compared with the traditional PID control model and neural network PID model,the process of output force rising of the adaptive PID control model based on Q-Learning is stable and fast. The proposed adaptive PID control model based on Q learning can effectively improve the control accuracy of the chassis dynamometer and meet the industrial requirements of the chassis dynamometer.

Key words:reinforcement learning;PID control;Q learning;control strategy;chassis dynamometer

0 引 言

汽车底盘测功机(转鼓试验台)主要包含滚筒和加载装置,以电涡流机输出加载力来模拟汽车在道路上行驶的场景,能够在室内对汽车进行综合测试,且对测试所需要的环境要求较低[1]。目前底盘测功机中大都采用标定 PID 参数或模糊 PID 控制法对加载的力进

收稿日期:2021-05-29 **修回日期**:2021-09-29
基金项目:国家重点研发计划项目(2018YFB1600800);陕西省重点研发计划(2020GY-018);西安市科技计划项目(20RGZN0008);中央高校基本科研业务费专项资金项目(300102241305)
作者简介:乔 通(1997-),男,硕士,CCF 会员(H3303G),研究方向为人工智能与自动控制;通讯作者:周 洲(1981-),男,硕士,高级工程师,研究方向为智能交通与信息系统工程。

行控制,PID 参数一经整定就不能改变。但电涡流机具有非线性、紧耦合的特点,所以上述两种策略的控制效果并不理想[2]。

随着机器学习的发展,强化学习已被广泛应用于 PID 在线调整等序列决策问题,取得了一定的效果[3-6]。在国内方面的相关研究中,张训等[7]采用积分分离 PID 算法,实现转速、励磁电流和转矩、励磁电流的两个双闭环控制器,满足了测功机的控制要求,但达不到现如今底盘测功机控制的工业要求;郭磊等[8]设计的模糊自适应 PID 算法有效提高了跟踪性能和调节速度,完成了对 PID 增益值的调整,此方法需要增益值从零开始调整,所需要的控制时间也相对较长;游博洋等[9]设计了基于神经网络 PID 控制器的外骨骼系统,有效的提高了外骨骼机器人的易用性和实用性;贾燕燕等[10]基于神经网络设计的自适应网络功率机制动态调整发射功率的大小,较好地解决了无线体域网中的传感器控制节能问题;赵明皓等[11]基于深度强化学习设计的无人艇自主航行控制算法,比传统的 PID 控制在稳定性以及抗干扰上具有优势。国外方面,V N Thanh 等[12]使用 Q 学习算法设计的自适应 PID 控制器对伺服机器人进行控制,并验证了其优越性;P Kofinas[13]为了处理连续的状态-动作空间,设计了模糊 Q 学习代替传统的 Q 学习算法,仿真表明了其有效性。上述研究都取得了许多积极的成果,对该文研究的开展具有较好的借鉴意义。

该文分析了底盘测功机的加载方式以及常见强化学习算法的特点,结合其规律进行分析,并研发对应的状态空间、动作空间和奖励等等,训练 Q 表完成对 PID 增益值的自动调节。主要研究基于强化学习的 PID 策略设计出来的 QPID 控制器,对底盘测功机输出扭矩的控制效果。

1 强化学习控制策略设计

1.1 强化学习

强化学习是通过与外部的环境进行交互,每次交互会获得奖赏,再通过该奖赏指导下一次的行为,其目标是使智能体能够取得最大累积奖赏[14]。强化学习的结果是寻找出一个策略 $\pi: S \rightarrow A$,能够让每个状态 s 的值函数 $V^{\pi}(s)$ 或者状态-动作值函数 $Q^{\pi}(s,a)$ 达到最大。$V^{\pi}(s)$ 与 $Q^{\pi}(s,a)$ 分别表示某个“状态”上或者是某个“状态-动作”上的累积奖赏[15]。

强化学习也在不断的发展,Q-Learning 算法被认为是其中最主要的进展之一。Q-学习算法考虑了状态作用值函数 Q,不考虑被控制系统确切的数学模型,通过时间差分对系统进行控制[16]。Q-Learning 是 RL 中 value-based 的算法,其中的 Q 意为在某个时刻的状态时,选择某个动作可以获得相应的收益,环境状态会依据此次智能体的动作,反馈出其所获得的立即奖赏 r,再依据 r 进行 Q 表的更新,公式如下:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',\pi(s')) - Q(s,a)] \quad (1)$$

其中,α 为学习率,$0 \leqslant \alpha \leqslant 1$。

算法 1:Q 学习算法。

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

Step2:循环每个 episode;

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(1)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。

1.2 底盘测功机自适应 PID 控制器设计

该文提出了一种基于 Q 学习算法的 PID 控制器,用于调整底盘测功机的扭矩输出,整个控制器的结构如图 1 所示。系统的直接控制由一个传统的 PID 完成,而参数的自适应调整是基于 Q-学习算法在训练过程中获得的 Q 表,传统的 PID 实现输入电压的调节。控制器的输入为人为设定的加载力的目标值 Fref,将每次调整之后的扭力值 $F_n(t)$ 与目标值的误差量输入到 PID 中,进而完成此次的调整。待调节完之后,获得此次调节的扭力值 $F_n(t)$,把这次的扭力值进行离散化,即可得到此次的状态 $n(t)$。之后开始本次的 Q 表更新,总共有 3 个 Q 表,对应于 PID 的三个参数,一个参数对应到一张 Q 表上。当 Q 学习算法更新完毕之后,Q 表最终会趋于稳定。此时在三张 Q 表中,选择某一个状态之后,每张 Q 表都会选择出此时 PID 控制器最优的增益值去调整。

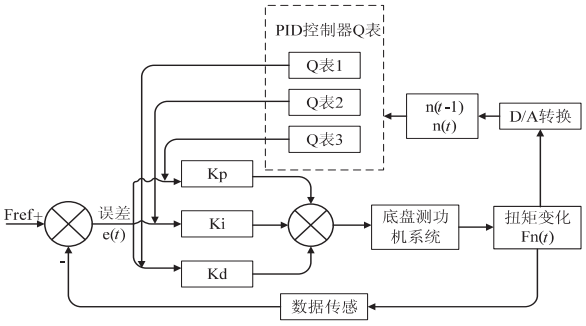


图 1 基于 QPID 的底盘测功机系统控制器结构

2 结合 Q 学习的 PID 控制算法

对于 Q 学习最重要的一个问题,就是如何训练 Q 表。该文设计的控制器,需要通过三张 Q 表使得底盘测功机不同扭矩输出的状态,对应到 PID 策略的各个

参数上。将 Q 学习策略与传统的 PID 策略进行结合,具体的训练过程如算法 2 所示。为了使得 Q 表可以快速收敛趋于稳定,实现了一种自适应学习率的算法——Delta-Bar-Delta[17]。在训练过程中,取得某个状态时的最佳参数之后,就根据公式计算出此次需要调整的输出量,输出量会通过 PID 控制器作用于底盘测功机,此时扭矩输出改变,进入到下一个状态。通过比较前后两个时刻的扭矩输出,就可以得到此次调整之后的立即奖赏 R_p,使用 R_p 更新 Q 表,开始下一次的训练,如此循环。当 Q 表趋于稳定之后,Q 表就含有了在每个状态下最优的 PID 参数,使用该参数即可控制底盘测功机的扭矩输出。

算法 2:结合 Q 学习的 PID 控制算法。

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

Step2:初始化学习率 ∂;

Step3:初始化 ε-greedy 策略的 ε;

Step4:当 episode<maxepisode 执行:

Step5:$t=0$;

Step6:初始化 $S_t(x(t),x'(t))$;

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

Step8:for $t=1$; ≤ maxtime;t ++

Step9:将状态 S_{t-1}, S_t 离散化,获得:$n_1(t-1)$ 和 $n_1(t)$;

Step10:for $i=1$;$i \leqslant 3$;i ++

Step11:遵循 ε-greedy 策略,根据 $n_1(t-1)$ 和 $n_1(t)$ 选择动作 A_i;

end

Step12:根据 PID 输出,获得完整的输出;

Step13:观察新状态 $S_{t+1}(x(t),x'(t))$;

Step14:获得的奖励 R_p;

Step15:将状态 S_{t+1} 离散化,获得:$n_1(t+1)$;

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

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

Step18:$S_t \leftarrow S_{t+1}$;

end

end

2.1　自适应学习率

为了使得 Q 表尽快达到稳定,使用了一种自适应学习率的算法,其定义为:

$$\Delta\alpha_t=\begin{cases}k & \text{if } \delta_{t-1}\delta_t>0\\ -\Phi\alpha_t & \text{if } \delta_{t-1}\delta_t<0\\ 0 & \text{if } \delta_{t-1}\delta_t=0\end{cases} \tag{2}$$

式中,$\Delta\alpha_t$ 是 t 增量;k 是提高学习率的正常数值;Φ 是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,$\delta_t=R_{t+1}+\gamma\max Q(S_{t+1},a)-Q(S_t,a)$;$\delta_t=(1-\Phi)\delta_t+\Phi\delta_{t-1}$。

通过使用上面的方法,将当前的 TD 误差与前面步骤中的累计 TD 误差进行比较,从而更新学习速率。当学习率较大时,改变符号,从而使其在下一次调整时调低。如果学习率太小,学习率会按照之前的变化趋势不断增加,使得收敛速度加快,所以时间步骤 $t+1$ 中的学习速率为 $\alpha_{t+1}=\alpha_t+\Delta\alpha_t$。三个 Q 表都将采用该算法,但对于每张 Q 表的参数设置会有不同。

2.2　离散化

由于加载力的状态值连续,且过于繁多,所以对于加载效果一样的情形,可选择同一组 PID 参数进行控制,因此可以把连续的加载力变量分成几个区间,同一个区间内的加载力值作为一个相同的状态。区间的设置使用与定义使用相同的规则,其定义为:

$$n=\begin{cases}0 & \text{if } x_{con}<x_{min}\\ \left[\dfrac{x_{con}}{x_{max}-x_{min}}\times N\right]+1 & \text{if } x_{min}\leqslant x_{con}\leqslant x_{max}\\ 20 & \text{if } x_{con}>x_{max}\end{cases} \tag{3}$$

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 表示加载力被分成的区间数,文中 $N=20$。N 取决于模拟性能。扭矩 F_n 通过公式(3)区间划分,离散化设置的值如表 1。

表 1　设定离散化值

变量	下限	上限
扭矩 F_n / N	1	2 000

2.3　ε-greedy 策略

当给定当前状态之后,三个 Q 表都将根据 ε-greedy 方法选择每次的动作,此方法的定义如下:

$$A=\begin{cases}\text{随机动作 if } \zeta<\varepsilon\\ \arg_a^{\max}Q(s,a) \text{ other}\end{cases} \tag{4}$$

其中,$\zeta \in [0,1]$ 是一个正态分布的随机数。

为了加快收敛的速度,ε 的值会随着训练次数的增大而减小,在迭代次数达到某个数值后设为零,而具体的次数会根据训练表现来决定。在 ε-greedy 策略中,ε 的值比较大,表示选取一个随机动作的概率也比较大。具体 ε 定义为:

$$\varepsilon(\text{eps})=\begin{cases}\dfrac{1}{1+\text{eps}}+0.001 & \text{if eps}<0.6\times\text{maxep}\\ 0 & \text{other}\end{cases} \tag{5}$$

其中,eps 表示当前的 episode,maxep 是 episode 的最大值。

2.4　奖励策略

该文根据测功机系统的情况将立即奖赏分为三种情况:调节后加载力趋于设定力值,加载力远离设定力值和调节之后加载力无变化。

调控后扭矩趋于设定值。根据 a_t 收到的参数进行调节,所获得的扭矩 $F_n(t)$ 与目标值 Fref 的相差结果,若是远小于 $t-1$ 扭矩 $F_n(t-1)$ 与 Fref 的相差结果,意为此次的调控有效,设定此次调整的奖赏为相邻两次扭矩输出的差值。

调控后扭矩远离设定值。根据 a_t 得到的参数进行调节,所获得的扭矩 $F_n(t)$ 与设定值 Fref 的相差结果,若是远大于 $t-1$ 扭矩 $F_n(t-1)$ 与 Fref 的相差结果,意为此次的调节为错误调节,奖赏为负值。

调控后扭矩无变化。根据 a_t 得到的参数进行调节,所获得的扭矩 $F_n(t)$ 与设定值 Fref 的相差结果,若是与 $t-1$ 扭矩 $F_n(t-1)$ 与 Fref 的相差结果,二者相差不超过 20 N,意为此次的调节无效果,即奖赏值为 0。综上,奖励计划如下:

$$r(t)=\begin{cases}|F_n(t)-F_n(t-1)| & \text{if } |F_{nref}-F_n(t-1)|-|F_{nref}-F_n(t)|>20\\ 0 & \text{if } \|F_{nref}-F_n(t-1)|-|F_{nref}-F_n(t)\|\leqslant 20\\ -|F_n(t)-F_n(t-1)| & \text{if } |F_{nref}-F_n(t)|-|F_{nref}-F_n(t-1)|>20\end{cases} \tag{6}$$

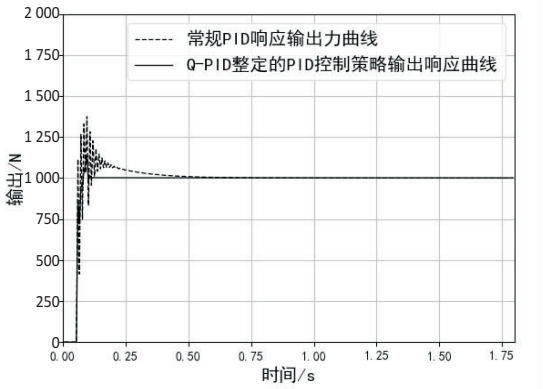
3 算法实验研究

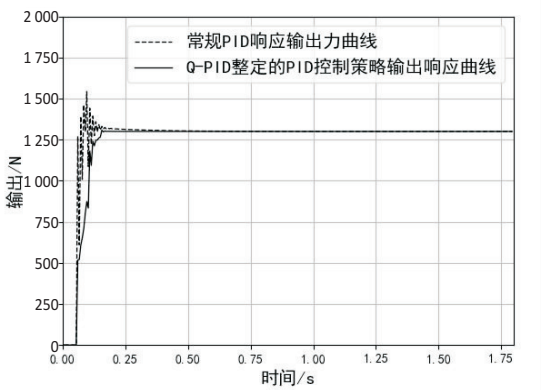
PyCharm 是一款系统模型库的功能十分丰富的仿真平台,该文使用 PyCharm 建立仿真系统,使用模拟的数据进行实验,验证使用 QPID 策略的可行性。选择相同的初始条件针对底盘测功机的恒力运行状态进行仿真控制,分别使用传统 PID 策略、BP-PID 策略以及文中提出的 QPID 策略进行系统仿真,根据结果进行对比分析。

(1)QPID 控制策略与传统 PID 控制策略的对比。

图 2 为分别使用两种控制策略,输出力从 0 N 分别到 1 000 N、1 300 N 和 1 500 N 的加载力响应曲线。

在仿真中,对比传统的 PID 控制策略,基于 QPID 控制策略加载力响应曲线的波动较小,一般在 120 ms 左右就可以实现加载力的响应过程,146 ms 后趋于稳定。传统PID策略下扭矩输出响应曲线的波动较大,





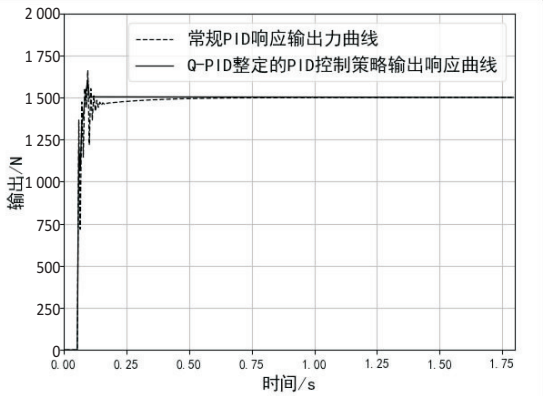


图 2 QPID 控制器与 PID 控制器的输出力响应曲线

一般在 249 ms 左右实现扭矩输出的响应,在 358 ms 后才达到设定值。基于 QPID 策略下的调整周期相较于传统的 PID 策略缩短至 40% 。

加载至 1 000 N 的响应曲线特征如表 2 所示。

表 2 QPID 控制器与 PID 控制器响应曲线特性

变量	PID	QPID
超调量/N	542.6	135.6
上升时间/ms	312	102
稳定时间/ms	372	126

在加载力目标值为 1 000 N 时,与 QPID 控制器(135.6 N)相关的曲线的超调远低于传统 PID 控制器(542.6 N)。除此之外,QPID 控制器(126 ms)的稳定时间比 PID 控制器(372 ms)的稳定时间短。

(2)QPID 控制策略与 BP-PID 控制策略的对比。

图 3 为分别使用 QPID 与 BP-PID 控制策略,输出力从 0 N 分别到 1 000 N、1 300 N 和 1 500 N 的加载力响应曲线。

在仿真中,基于 QPID 的策略比 BP-PID 策略更快达到稳定,在 120 ms 左右就可以实现加载力的响应过程,在 146 ms 后趋于稳定。而 BP-PID 策略下扭矩输出的曲线上升时间与稳定时间较慢,在 425 ms 左右实现扭矩输出的响应,在 524 ms 后达到设定值。基于 QPID 控制策略下的调整周期相较于 BP 控制策略的调整周期缩短至 27.9% 。

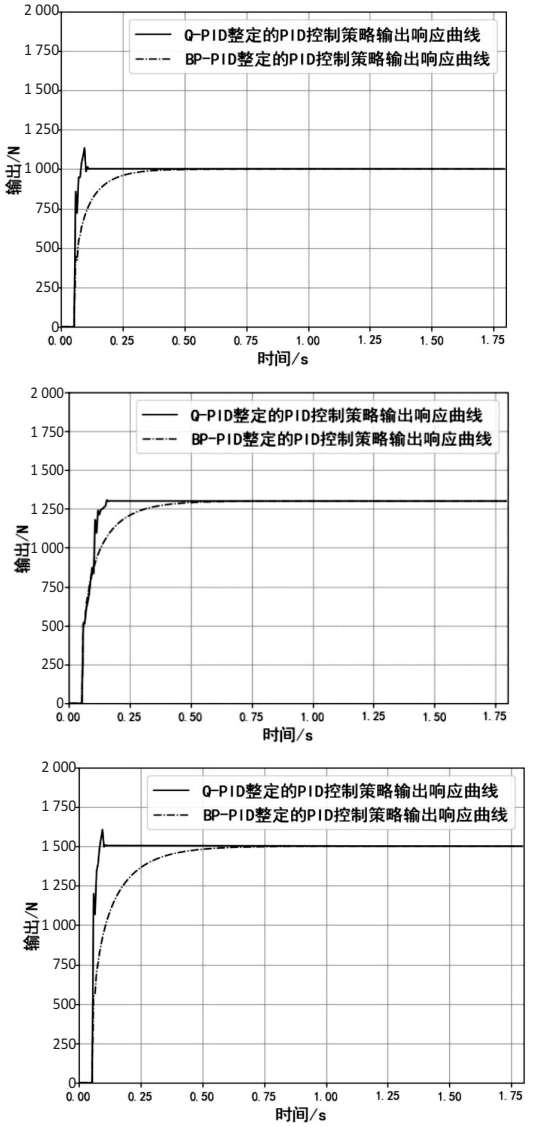


图 3　QPID 控制器与 BP-PID 控制器的输出力响应曲线

加载至 1 300 N 的响应曲线特征如表 3 所示。

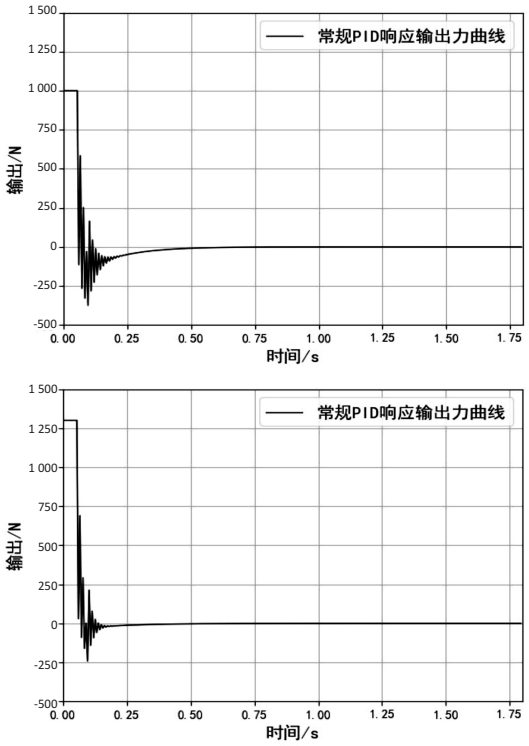
表 3　QPID 控制器与 PID 控制器响应曲线特性

变量	BP-PID	QPID
超调量/N	0	14.9
上升时间/ms	420	156
稳定时间/ms	504	174

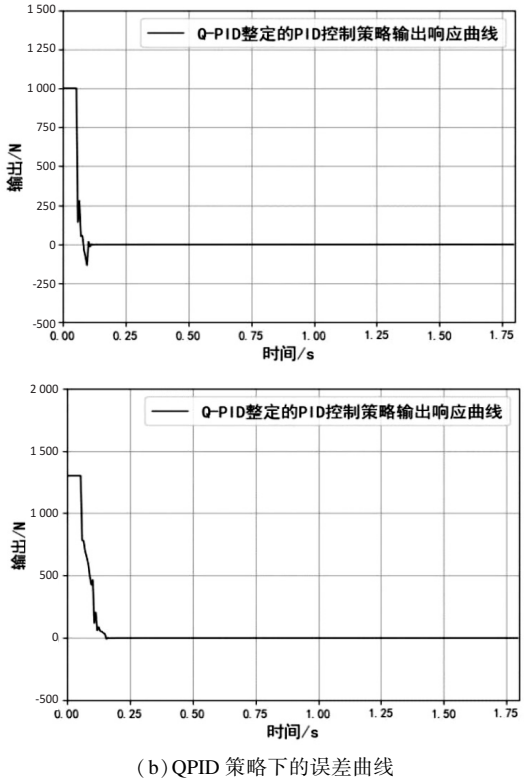
在加载力目标值为 1 300 N 时,与 QPID 控制器(14.9 N)相关曲线的超调大于 BP-PID 控制器(0 N)。另外,QPID 控制器(156 ms)的稳定时间比 BP-PID 控制器(504 ms)短。

根据国家质量监督检验检疫总局 2018 年发布的底盘测功机使用标准,底盘测功机运行状态的工业要求误差不大于 2.0%,加载响应需要在 300 ms 以内达到目标值的 90%。以上三种控制策略下的扭矩输出的误差曲线如图 4 所示。

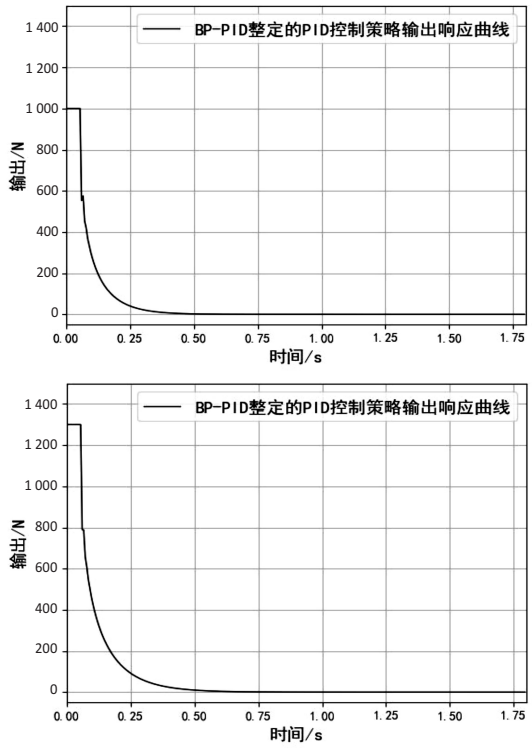
由图 4 可知,QPID 控制的系统加载力响应曲线的最大振幅 146 ms 后小于 10 N,达到工业要求;BP-PID 控制器的扭矩输出曲线的最大振幅 420 ms 后高达 50 N 左右;传统 PID 控制策略下的扭矩输出曲线的最大振幅 321 ms 后约为 27 N。基于 QPID 控制策略可以满足底盘测功机使用所需要达到的工业要求,其加载力的响应曲线正常,跟理论分析的结果保持一致。



(a)常规 PID 策略下的误差曲线



(b)QPID 策略下的误差曲线



(c)BP-PID 策略下的误差曲线

图 4 三种控制策略下的输出力误差曲线

4 结束语

针对底盘测功机的加载控制问题,提出了一种基于 Q 学习的 PID 控制策略,使用 QPID 对三个增益值进行调整,使其能够快速稳定达到加载目标值,最后完成了与另外两种策略的比对试验。通过分析对比试验的结果,证明在底盘测功机上使用 QPID 控制器,可以让加载力的响应时间缩小到 120 ms,在 146 ms 后稳定到工业要求的误差范围之内,控制周期缩短明显。说明基于 Q 学习的 PID 调节策略可以在底盘测功机上得到较好的应用。

参考文献:

[1] 田 颖,金振华,聂圣芳,等. 交流电力测功机控制系统的研究[J]. 汽车工程,2014,36(1):125-128.

[2] 尹 涛,郑 义,李 晶,等. 底盘测功机行驶阻力设定方法比较[J]. 小型内燃机与摩托车,2014,43(1):54-56.

[3] SHI Q,LAM H K,XIAO B,et al. Adaptive PID controller based on Q-learning algorithm[J]. CAAI Transactions on Intelligence Technology,2018,3(4):235-244.

[4] 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,2017,48(3):376-382.

[5] 段友祥,任 辉,孙歧峰,等. 基于异步优势执行器评价器的自适应 PID 控制[J]. 计算机测量与控制,2019,27(2):70-73.

[6] IGNACIO C,MARIANO D P,SEBASTIAN V,et al. Incremental Q-learning strategy for adaptive PID control of mobile robots[J]. Expert Systems with Applications,2017,80(5):183-199.

[7] 张 训. 汽车底盘测功机测控系统的研究[D]. 哈尔滨:哈尔滨理工大学,2009.

[8] 郭 磊,陈文会,刘小民. 模糊自适应 PID 在汽车底盘测功机中的仿真研究[J]. 电子设计工程,2013,21(7):76-79.

[9] 游博洋,王险峰,赵 玲,等. 基于神经网络 PID 控制器的外骨骼系统设计[J]. 计算机技术与发展,2021,31(5):1-6.

[10] 贾燕燕,谢志军. 基于 PID 神经网络的无线体域网功率控制算法研究[J]. 计算机应用研究,2018,35(9):2744-2747.

[11] 赵明皓,张翠萍,李宝安. 基于深度强化学习的无人艇控制应用研究[C]//中国指挥与控制学会(Chinese Institute of Command and Control). 第八届中国指挥控制大会论文集. 北京:兵器工业出版社,2020:50-57.

[12] THANH V,DANG P V,NAM L,et al. Reinforcement Q-learning PID controller for a restaurant mobile robot with double line-sensors[C]//ICMLSC 2020:The 4th international conference on machine learning and soft computing. New York:Association for Computing Machinery,2020:164-167.

[13] KOFINAS P,DOUNIS A I. Fuzzy Q-learning agent for online tuning of PID controller for DC motor speed control[J]. Algorithms,2018,11(10):148.

[14] SUTTON R S,BA R A G. Reinforcement learning:an introduction[J]. IEEE Transactions on Neural Networks,1998,9(5):1054.

[15] 周志华. 机器学习[M]. 北京:清华大学出版社,2016.

[16] 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.

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