文章编号: 1000−8152(2011)11−1665−06

小脑模型关节控制器网络在传送带给料生产加工站学习优化控制中的应用

周 雷[1], 孔 凤[1], 唐 昊[1,2], 张建军[1,2]

(1. 合肥工业大学 计算机与信息学院, 安徽 合肥 230009; 2. 安全关键工业测控技术教育部工程研究中心, 安徽 合肥 230009)

摘要: 研究单站点传送带给料生产加工站(conveyor-serviced production station, CSPS)系统的前视(look-ahead)距离最优控制问题, 以提高系统的工作效率. 论文运用半Markov决策过程对CSPS 优化控制问题进行建模. 考虑传统Q学习难以直接处理CSPS系统前视距离为连续变量的优化控制问题, 将小脑模型关节控制器网络的Q值函数逼近与在线学习技术相结合, 给出了在线Q学习及模型无关的在线策略迭代算法. 仿真结果表明, 文中算法提高了学习速度和优化精度.

关键词: 传送带给料生产加工站; 小脑模型关节控制器; Q学习; 在线策略迭代

中图分类号: TP202 **文献标识码**: A

Application of cerebellar model articulation controller network to learning optimization control in conveyor-serviced production station

ZHOU Lei[1], KONG Feng[1], TANG Hao[1,2], ZHANG Jian-jun[1,2]

(1. School of Computer and Information, Hefei University of Technology, Hefei Anhui 230009, China;
2. Engineering Research Center of Safety Critical Industry and Control Technology Ministry of Education, Hefei Anhui 230009, China)

Abstract: This paper is concerned with the optimization of the look-ahead distance for a conveyor-serviced production station(CSPS) to improve the efficiency of operations. The optimal control process for CSPS is modeled by a semi-Markov decision process(SMDP). Since the standard Q-learning is difficult to deal with the continuous variable optimal look-ahead control problem of CSPS directly, Cerebellar Model Articulation Controller(CMAC) for Q-values function approximation is combined with the online learning technology, and some online Q-learning and model-free online policy iteration algorithms are provided. Simulation results show that the proposed algorithms improve the learning speed and the precision of optimization.

Key words: conveyor-serviced production station; cerebellar model articulation controller; Q-learning; online policy iteration

1 引言(Introduction)

现实世界的生产加工企业中, 存在一些由加工站为生产加工主体的生产线, 例如先进制造业中的机器人装配线, 其中一类可以称为传送带给料生产加工站(conveyor-serviced production station, CSPS)系统[1~6], 如图1所示. 该系统优化目标是如何控制站点的前视(look-ahead)距离, 以协调加工站的工件下载和工件加工, 提高系统工作效率.

CSPS问题是自Ford生产线问题以来IE/OR领域的传统问题, 在系统模型、控制模式等方面已存在许多研究成果[1~9]. 其中, 排队理论、Markov或半Markov决策过程理论是分析CSPS系统的有效手段[4,6~9]. 文献[4]建立了CSPS优化控制问题的半Markov决策过程数学模型, 给出了一些性能值的理论计算方程, 但该方法依赖于精确地系统模型参数. 文献[6]建立了单站点CSPS系统基于性能势的在线策略迭代(online policy iteration, OPI)算法. 文献[9]讨论了单服务器系统的准入最优look-ahead策略. 近年来, 多站点CSPS系统的研究逐步成为热点, 主要涉及站点协作和平衡问题[7,8,10].

对于单站点CSPS系统look-ahead距离为连续变量的优化控制问题, 文献[6]采用简单离散化方法, 运用表格形式存储Q因子、性能势, 存在离散粒度有时难以控制及存储空间消耗大的问题, 且影响系统优化性能. 考虑小脑模型关节控制器(cerebellar model articulation controller, CMAC)神经网络的输入

收稿日期: 2010−07−20; 收修改稿日期: 2011−01−18.

基金项目: 国家自然科学基金资助项目(60873003, 61174186); 教育部留学回国人员科研启动基金资助项目(教外司留2008890); 安徽省自然科学基金资助项目(090412046); 安徽高校省级自然科学研究重点资助项目(KJ2008A058, KJ2011A230); 中日国际科技合作资助项目(2011FA10440).

可以为连续值, 且具有泛化能力强、收敛速度快及在线增量学习等优点. 而OPI中的在线学习技术能充分利用样本数据信息, 有效减少为学习Q因子值函数而频繁进行的随机行动选择, 可保证系统比较平稳、安全的运行. 因此, 本文在文献[6]的工作基础上, 将CMAC网络与在线学习技术相结合, 运用CMAC网络实现行动连续的Q值函数及性能势函数的学习逼近, 构造基于仿真样本数据学习的在线Q学习及相关OPI算法, 克服理论求解对精确模型参数的依赖及传统Q学习无法直接处理CSPS系统连续控制变量优化问题的缺点, 并提高算法学习速度和优化精度, 以能快速适应因需求变化引起的CSPS系统生产调整.

2 数学模型(Mathematical model)

如图1所示, CSPS系统look-ahead优化控制问题可建模为半Markov决策过程[4,6]. 缓冲库空余量为系统状态, 状态空间记为$\Phi=\{0,1,2,\cdots,N\}$, N为缓冲库最大容量. 记$v=(v(0),v(1),\cdots,v(N))$为控制策略, $v(i)$表示系统在状态i时的look-ahead距离, 其中$v(0)\equiv 0$, $v(N)\equiv\infty$[6]. 对于其他状态$v(i)\in D=[L_{\min},L_{\max}]$, D为行动空间, $L_{\min}$, $L_{\max}$分别为最小、最大前视距离, 一般令$L_{\min}=0$.

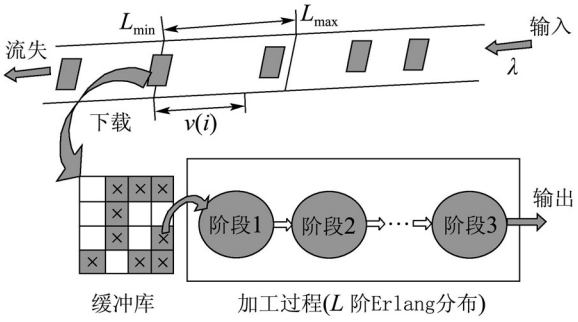

图 1 传送带给料生产加工站

Fig. 1 Conveyor-serviced production station

在策略v下, 记初始决策时刻T_0, $X_t(t\in\mathbb{R}^+)$表示t时刻系统状态. 在第n个决策时刻T_n, 系统状态$X_{T_n}=i$(简记$X_n=i$), 若在$v(i)$范围内至少有一个工件, 则等待第一个工件到达, 并将其捡取到缓冲库(捡取时间忽略不计), 系统进入下一个决策时刻T_{n+1}, 则决策间隔时间等于等待时间记为$\omega_n=T_{n+1}-T_n$, 且状态转移至$X_{n+1}=i-1$, 其时间轴如图2所示.

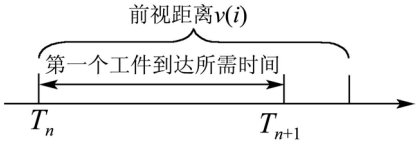

图 2 捡取工件时的时间轴

Fig. 2 The time-line of workpiece unloading

若在$v(i)$范围内没有工件, 且缓冲库不空, 加工缓冲库中的一个工件, 记服务时间为τ_n, 则$\omega_n=T_{n+1}-T_n=\max\{v(i),\tau_n\}$, 且状态转移至$X_{n+1}=i+1$, 其时间轴如图3所示.

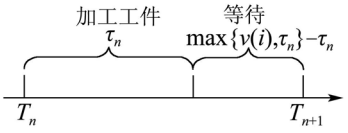

图 3 加工工件时的时间轴

Fig. 3 The time-line of workpiece processing

$f(X_n,v(X_n),X_{n+1},t)$为t时刻系统在状态X_n采取行动$v(X_n)$后, 转移到下一状态X_{n+1}之前单位时间期望代价函数, 其中$T_n\leqslant t<T_{n+1}$. 根据加工主体相关操作, $f(X_n,v(X_n),X_{n+1},t)$将由存储、服务、等待、完成工件加工和look-ahead等相应操作代价构成. 根据文献[11], 策略v下CSPS系统无穷时段折扣代价为

$$\eta_\alpha^v(i)=\mathrm{E}[\sum_{n=0}^{\infty}\int_{T_n}^{T_{n+1}}\alpha\mathrm{e}^{-\alpha t}f(X_n,v(X_n),X_{n+1},t)\mathrm{d}t|X_0=i],$$

其中折扣因子$\alpha\geqslant 0$. 当$\alpha>0$时, $\eta_\alpha^v(i)$表示状态i的无穷时段期望折扣代价; 当$\alpha\to 0^+$, 其极限$\eta_{0^+}^v(i)$表示策略v下无穷时段平均代价

$$\eta^v=\lim_{M\to\infty}\frac{1}{T_M}\times\mathrm{E}[\sum_{n=0}^{M-1}\int_{T_n}^{T_{n+1}}f(X_n,v(X_n),X_{n+1},t)\mathrm{d}t].$$

系统的优化目标就是寻找一个策略v^*, 使得系统的代价函数值$\eta_\alpha^{v^*}$或η^{v^*}达到最优.

3 CSPS系统基于CMAC的学习优化算法(Learning optimization algorithm of CSPS based on CMAC)

CSPS系统具有状态离散有限、行动连续的特点, 传统求解方法存在一些缺点, 如理论算法依赖精确模型参数、离散化方法的粒度难以控制且所需存储空间大等. 近年来, 强化学习、神经元动态规划等方法通过基于仿真样本的学习、值函数参数化逼近表示等技术或方法, 建立了许多模型无关、在线学习的解决Markov/半Markov决策过程的优化算法, 自然可以引入到CSPS系统优化控制中. 另外, 为适应需求变化而可能进行的生产调整, 算法还须具备在线学习、速度快等特点. 因此, 本文将强化学习与CMAC神经网络相结合, 并引入OPI充分利用样本信息的思想, 解决CSPS系统look-ahead优化控制问题, 充分发挥CMAC网络的学习能力和非线性映射能力强以及OPI在线学习的优点.

3.1 基于CMAC的Q学习(Q-learning based on CMAC)

CMAC神经网络是Albus于1975年根据小脑生物模型提出的一种人工神经网络[12], 具有结构简单、泛化能力强、学习速度快、逼近精度高等特点, 已被广泛应用于函数逼近、模式识别与机器人控制等多个领域[13~16]. 本文利用CMAC网络输入可为连续变量、泛化能力强等特点, 实现对CSPS系统状态–行动对Q因子值函数的逼近, 求解其最优控制问题.

文献[14]作业分配搬运系统中为每个离散行动设置了一个CMAC网络, 实现了混合状态的Q因子值函数逼近. 考虑到单站点CSPS系统缓冲库容量一般设置较小(状态空间简单), 设置过大对系统性能的改进影响不大[6]. 因此, 本文将为系统每个状态设置一个CMAC网络(状态0和N的行动唯一, 不需设置CMAC网络, 称为非探索状态, 其他状态为探索状态), 行动作为网络输入, Q因子则为输出.

在如图4所示状态i的CMAC网络中, 输入变量行动d将激活存储区AP中C个连续存储单元, 被激活的单元$a_{i,j}=1$, 未被激活的单元$a_{i,j}=0$. 因此, 输出$Q_\alpha(i,d,w_i)$计算公式为

$$Q_\alpha(i,d,w_i)=\sum a_{i,j}w_{i,j}=\sum_{j=h}^{h+C-1}w_{i,j},\qquad(1)$$

其中: w_i为状态i对应的网络权值向量, h为输入变量d所激活存储单元的首地址. 由图4可以看出, 输入空间相近的两个输入向量在存储区中有部分重叠单元, 因此其输出值也比较相近, 这种现象被称为CMAC神经网络的局部泛化特性, C又可称为泛化参数. 另外, CMAC网络权值的调整采用的是δ学习算法, 其权值的更新公式为

$$w_{i,j}:=w_{i,j}+\zeta\frac{c_n}{C},\qquad(2)$$

其中: ζ为学习率, $j=h,\cdots,h+C-1$, c_n为Q因子的即时差分值.

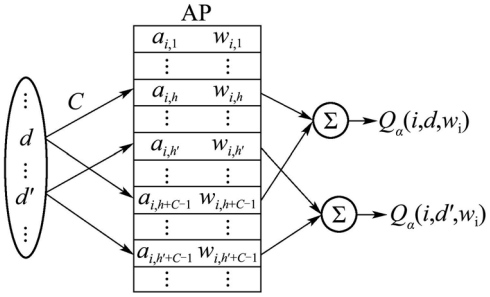

图 4 CMAC网络结构

Fig. 4 The structure of CMAC network

在T_n时刻, 设系统在状态$X_n(X_n\neq 0,N)$采取行动d_n, 得到一个样本$<X_n,d_n,X_{n+1},\omega_n,\tau_n>$. 根据文献[6], 即时差分$c_n$为

$$c_n=f'_\alpha(X_n,d_n,X_{n+1})-T_\alpha(\omega_n)\eta_n+\mathrm{e}^{-\alpha\omega_n}\cdot\min_{d\in D}Q_\alpha(X_{n+1},d,w_{X_{n+1}})-Q_\alpha(X_n,d_n,w_{X_n}),\qquad(3)$$

其中:

$$T_\alpha(x)=\int_0^\infty \mathrm{e}^{-\alpha x}\mathrm{d}t=(1-\mathrm{e}^{-\alpha x})/\alpha,\ x\geqslant 0,$$

其极限$T_0(x)=x$; $f'_\alpha(X_n,d_n,X_{n+1})$为系统在状态$X_n$采取行动$d_n$并转移到$X_{n+1}$的累积折扣代价. 当$X_{n+1}=X_n-1$时,

$$f'_\alpha(X_n,d_n,X_{n+1})=T_\alpha(\omega_n)[k_1(N-X_n)+k_3]+k_5d_n\chi_N(X_n),\qquad(4)$$

其中: $\chi_N(N)=0$, $\chi_N(X_n)=1(X_n\neq N)$, k_1表示一个工件的单位时间存储代价, k_3表示单位时间等待代价, k_5表示单位时间look-ahead的即时代价. 当$X_{n+1}=X_n+1$时

$$f'_\alpha(X_n,d_n,X_{n+1})=T_\alpha(\omega_n)k_1(N-X_n-1)+k_2T_\alpha(\tau_n)+k_3[T_\alpha(\omega_n)-T_\alpha(\tau_n)]+k_4\mathrm{e}^{-\alpha\tau_n}+k_5d_n,\qquad(5)$$

其中: k_2表示单位时间服务代价, k_4表示加工完一个工件的即时代价(为负数, 表示实际报酬). 而η_n是平均代价的学习估计值, 其更新公式为

$$\eta_n:=\eta_n+\gamma_n(f'_{\alpha=0}(X_n,d_n,X_{n+1})-\eta_n\omega_n),\qquad(6)$$

其中: $f'_{\alpha=0}(X_n,d_n,X_{n+1})$表示平均准则下的转移代价, γ_n为学习步长.

状态0和N为非探索状态, 行动唯一, 不需要进行学习优化, 不设置CMAC网络, 只需简单记录其过程参数. CSPS系统基于CMAC的Q学习算法具体如下:

Step 1　初始化CMAC网络的权值、初始温度等参数, 令$s=0$, $n=0$, $T_n=0$, $\eta_n=0$;

Step 2　在T_n时刻, 观察得到状态X_n. 若X_n为探索状态, 转Step 3; 若X_n为非探索状态, 转Step 4.

Step 3　根据模拟退火算法选择行动d_n, 观测实际系统得到状态X_{n+1}, 记录决策间隔时间ω_n及服务时间τ_n, 按公式(6)(3)(2)更新权值向量w_{X_n}. 若t_s不满足终止条件, 令$s:=s+1$, 转Step 5.

Step 4　执行相应行动, 得到实际系统状态X_{n+1}, 记录决策间隔时间ω_n及服务时间τ_n, 转Step 5.

Step 5　若不满足终止条件, 令$n:=n+1$, 转Step 2, 否则学习结束.

3.2 OPI优化算法(OPI optimization algorithm)

文献[6]将策略迭代和Q学习相结合, 提出了基于性能势的OPI算法, Q因子和性能势采用传统的表格形式保存. 本文将CMAC网络值函数逼近引入

到OPI算法中, 给出两个基于CMAC网络的OPI优化算法. 其一利用CMAC网络逼近Q因子, 性能势采用表格形式保存; 另一个利用CMAC网络分别逼近Q因子和性能势. 为了便于区分, 前者称之为OPI-Q算法, 后者称之为OPI-Qg算法.

在OPI-Q和OPI-Qg算法中, 每次迭代包括策略评估和策略更新两大步骤. 两个算法的基本过程一样, 即第k步迭代中, 记基准策略为v^k, 设有一样本轨道可以分成多个子样本轨道. 在每个子样本轨道的第一个探索状态i, 为由基准策略v^k确定的状态行动$v^k(i)$引入一个增量, 形成行动集$D_{k,i} = [v^k(i)-m, v^k(i)+m] \cap D, m \in \mathbb{R}^+$. 然后, 选择执行行动$d_n \in D_{k,i}$, 得到样本$< i, d_n, j, \omega_n, \tau_n >$, 直到下一个子样本轨道都将根据基准策略选择行动, 进行性能势学习和Q因子学习, 即为策略评估. 然后, 根据Q因子数值实现基准策略v^k的更新, 进入第$k+1$步迭代. 上述两个OPI算法中, 状态执行行动在一定范围内$v^k(i) \pm m$进行选择, 可以尽力保证系统平稳运行, 减少传统Q学习中为获得Q值函数而进行的频繁随机行动选择. 另外, 通过仿真数据样本可以充分利用历史信息, 加快算法学习速度.

在3.1小节中, 已给出基于CMAC网络的Q因子逼近公式, OPI-Q和OPI-Qg算法同样为每个探索状态各设置一个CMAC网络用来逼近Q因子, 输入为行动d, 输出记为$Q_\alpha^v(i,d,w_i)$. 而在性能势的学习表示上, OPI-Q利用基于TD学习逼近的表格形式保存, 而OPI-Qg则采用基于CMAC网络的参数化逼近方法. 根据文献[6]性能势逼近公式, 本文给出其基于CMAC网络的参数化表示. 在第k步迭代中, 记基准策略为v^k. CMAC网络输入为状态X_n, 输出表示性能势$\tilde{g}_\alpha^{v^k}(X_n, w')$, w'为网络权值向量, 更新公式为

$$w'_j := w'_j + \zeta' \frac{c'_n}{\mathrm{C}'}, \tag{7}$$

C′为泛化参数, w'_j为与输入X_n对应的连续权值中的一个, ζ'为学习率. 性能势即时差分c'_n参数化公式为

$$\begin{aligned} c'_n = {} & f'_\alpha(X_n, v^k(X_n), X_{n+1}) - T_\alpha(\omega_n)\tilde{\eta}^{v^k} + \\ & \mathrm{e}^{-\alpha\omega_n}\tilde{g}_\alpha^{v^k}(X_{n+1}, w') - \tilde{g}_\alpha^{v^k}(X_n, w'), \end{aligned} \tag{8}$$

$\tilde{\eta}^{v^k}$学习估计参考公式(6). 基于性能势的Q因子即时差分c_n参数化公式为

$$\begin{aligned} c_n = {} & f'_\alpha(X_n, d_n, X_{n+1}) - T_\alpha(\omega_n)\tilde{\eta}^{v^k} + \\ & \mathrm{e}^{-\alpha\omega_n}\tilde{g}_\alpha^{v^k}(X_{n+1}, w') - Q_\alpha^{v^k}(X_n, d_n, w_{X_n}), \end{aligned} \tag{9}$$

而Q因子参数化更新公式参考公式(2). 根据CMAC网络输出的Q因子数值, 可以获得第$k+1$步的基准策略

$$v^{k+1}(i) \in \arg\min_{d\in D} Q_\alpha^{v^k}(i,d,w_i), i \in \Phi \backslash \{0, N\}. \tag{10}$$

根据文献[6], 给出CSPS系统优化控制问题的OPI-Qg算法(OPI-Q算法类似, 此处不再赘述):

Step 1 选择初始状态i, 初始化基准策略v^k、Q因子网络权值w 、泛化参数C及C'; 初始化常数α, $m, p_{\max}, n_{\max}$及温度t_s, 令$s=0, k=0$.

Step 2 策略评估. 在基准策略v^k下, 初始化性能势网络权值w', 令$p=1, \tilde{\eta}^{v^k}=0$.

Substep 1 探索. 在状态i下, 根据模拟退火算法选择行动$d \in D_{k,i}$, 记录样本$< i, d, j, \omega, \tau >$; 并令$n=0, X_n=j$. 若$t_s$不满足终止条件, 由$t_s$退温至$t_{s+1}$, 令$s := s+1$.

Substep 2 性能势学习. 执行行动$v^k(X_n)$得到样本$< X_n, v^k(X_n), X_{n+1}, \omega_n, \tau_n >$, 由公式(6)∼(8)更新网络权值向量$w'$.

Substep 3 如果$n < n_{\max}$或X_n不是探索状态, 令$n := n+1$, 转Substep2; 否则, 若状态X_n为探索状态, 则根据公式(2)(6)(9)更新网络权值向量w.

Substep 4 如果$p < p_{\max}$, 则探索没有结束, 令$p := p+1, i = X_{n+1}$转Substep 1; 否则, 转Step 3.

Step 3 策略改进. 由式(10)可以获得新的基策略v^{k+1}.

Step 4 如果满足停止条件, 退出; 否则, 令$k := k+1, i = X_{n+1}$, 转Step 2.

上述OPI-Qg算法中, $n_{\max}$为学习性能势的最大步数, $p_{\max}$为进行探索的最大次数. 算法时间轴表示如图5, 图中圆圈处进行探索, $n_{\max}$处更新逼近Q因子的CMAC网络的权值向量, 黑点处改进策略.

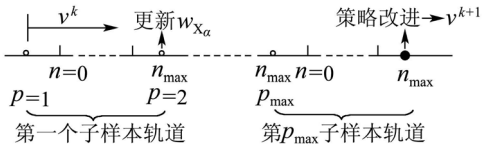

图 5 OPI-Qg算法的时间轴

Fig. 5 The time-line of OPI-Qg

4 仿真结果(Simulation results)

该CSPS系统模型源于日本松井正之教授的文献及文献[6], 为方便进行优化结果比较, 本文模型参数与文献[6]相同. 由策略迭代理论求解得到最优策略$v* = (0, 0.36, 0.48, 0.61, 0.79, \inf)$, 对应的平均代价$\eta^{v*} = 2.5210$(理论算法依赖于精确的模型参数). 表1为$\alpha = 0$时, 平均准则下4种算法的优化结果. 4种算法都为仿真学习优化算法, 其结果均为若干次独立实验结果的平均值. 其中, 传统Q学习中连续行动的离散化粒度为0.1, 优化结果源于文献[6]. 从表1可以看出, OPI-Qg算法结果最为接近理论求解数值, 而传统的Q学习算法结果最差(随着离散化

粒度的减小, 优化精度将会适当提高, 但算法运行时间及所需存储空间都将增大). 同时, OPI-Q和OPI-Qg算法结果要优于CMAC-Q学习优化算法, 说明了将OPI引入CSPS系统look-ahead优化控制问题可以提高优化精度.

表 1 平均准则下4种算法的优化结果

Table 1 Optimizaiton results of four algorithms under average criteria

算法	传统Q学习	CMAC-Q学习	OPI-Q	OPI-Qg
结果	2.5511	2.5312	2.5296	2.5253

图6为文献[6]Q学习的平均代价优化曲线. 图7为基于CMAC的Q学习的平均代价优化曲线. 从图可以看出, Q学习的优化曲线特别在早期波动比较大, 主要是学习优化过程中执行行动的随机选择造成的. 而基于CMAC的Q学习平均代价优化曲线下降得比较快, 曲线整体波动较小, 收敛速度较快, 体现了CMAC网络的泛化能力强、学习速度快等特点.

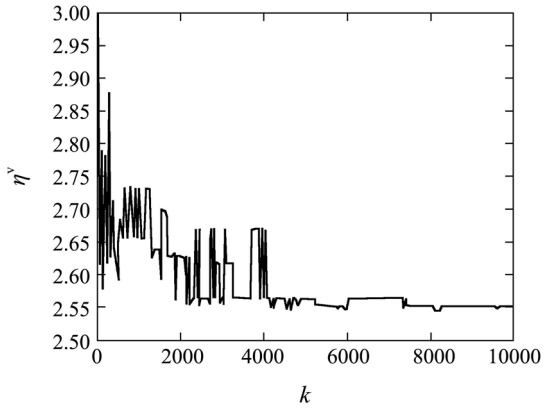

图 6 Q学习平均代价优化曲线

Fig. 6 The average cost optimization plot of Q-learning

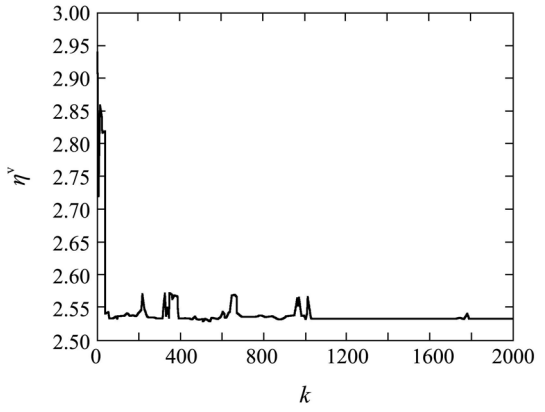

图 7 基于CMAC的Q学习平均代价优化曲线

Fig. 7 The average cost optimization plot of Q-learning based on CMAC

图8和图9给出的是OPI, OPI-Q和OPI-Qg优化算法的平均代价优化曲线, 它们具有相同的初始策略$v^0 = (0, 1, 1, 1, 1, \inf)$. 与文献[6]OPI算法相比, OPI-Q算法的最终优化结果更接近理论求解的最优值. 再结合图6与图7的比较结果, 说明将CMAC网络引入CSPS系统look-ahead优化控制问题是有效的, 较好的实现了连续行动变量性能值的逼近, 且学习速度快、优化精度高. 同时, 相关OPI算法优化曲线波动要小于基于CMAC的Q学习和一般Q学习的优化曲线波动, 体现了引入OPI相关技术的优势.

图9为OPI-Qg算法的平均代价优化曲线, 与图8的OPI-Q算法相比, 两者最终优化结果相差不大, 都非常接近理论求解的最优值. 然而, OPI-Q算法的优化曲线波动不断, OPI-Qg算法只在迭代开始时有较小的波动, 后面则非常平稳. 另外, 由图8和图9中曲线在迭代初期的下降速度可看出, Q因子和性能势全部采用CMAC网络逼近的OPI-Qg算法学习速度最快, 而Q因子和性能势全部采用表格形式存储的OPI算法学习速度最慢, OPI-Q算法学习速度介于两者之间, 体现了CMAC神经网络的局部泛化、学习速度快、非线性逼近能力强等特点.

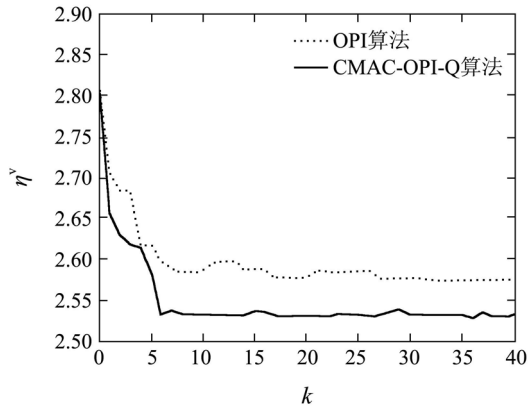

图 8 OPI-Q和OPI算法的平均代价优化曲线

Fig. 8 The average cost optimization plots of OPI-Q and OPI

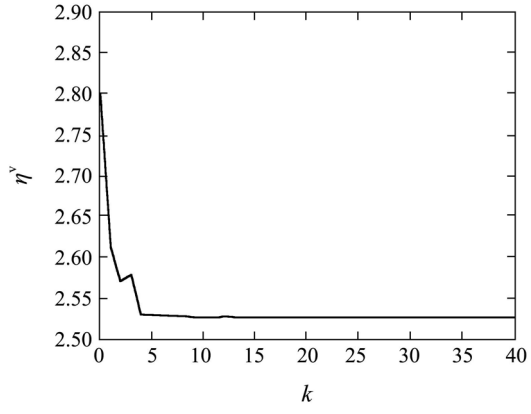

图 9 OPI-Qg算法的平均代价优化曲线

Fig. 9 The average cost optimization plot of OPI-Qg

5 结论(Conclusions)

针对CSPS系统控制行动为连续变量及生产调整对学习速度的要求, 本文将CMAC神经网络与强化

学习相结合, 并引入OPI充分利用样本信息的思想, 解决CSPS系统的look-ahead最优控制问题. 通过运用CMAC网络来逼近具有连续行动值的Q值函数或性能势函数, 充分发挥CMAC网络局部逼近、泛化能力强、学习速度快等特点, 给出了在线CMAC-Q学习及模型无关的OPI-Q, OPI-Qg优化算法. 仿真结果表明, 文中所提算法具有良好的优化效果, 学习速度快、优化精度高, 对于CSPS系统的设计和优化控制及为适应需求变化而进行的快速生产调整具有重要意义. 另外, 多站点CSPS系统的设计和优化控制、基于小样本数据的学习优化算法等将是进一步的研究课题.

参考文献(References):

[1] MORRIS W T. *Analysis for Material Handling Management*[M]. Richard D: Irwin inc., 1962.

[2] MATSUI M. *A study on optimal operating polices in conveyor-serviced production system*[D]. Japan: Tokyo Institute of Technology, 1981.

[3] NAWIJN W M. *Stochastic conveyor systems*[D]. Netherlands: Twente University of Technology, 1983.

[4] MATSUI M. A generalized model of conveyor-serviced production station (CSPS)[J]. *Journal of Japan Industrial Management Association*, 1993, 44(1): 25 – 32.

[5] MATSUI M. CSPS model: look-ahead controls and physics[J]. *International Journal of Production Research*, 2005, 43(10): 2001 – 2025.

[6] TANG H, ARAI TAMIO. Look-ahead control of conveyor-serviced production station by using potential-based online policy iteration[J]. *International Journal of Control*, 2009, 82(10): 1917 – 1928.

[7] MATSUI M. *Manufacturing and Service Enterprise with Risk: A Stochastic Management Approach*[M]. New York: Springer, 2009.

[8] ABE K, YAMADA T, MATSUI M. A design problem of assembly line systems using genetic algorithm under the BTO environment[J]. *IEEJ Transactions on Electronics, Information and Systems*, 2004, 124(10): 2006 – 2013.

[9] NAWIJN W M. The optimal look-ahead policy for admission to a single server system[J]. *Operations Research*, 1985, 33(3): 626 – 643.

[10] 唐昊, 万海峰, 韩江洪, 等. 基于多Agent强化学习的多站点CSPS系统的协作look-ahead控制[J]. 自动化学报, 2010, 36(2): 289 – 296.
(TANG Hao, WAN Haifeng, HAN Jianghong, et al. Coordinated look-ahead control of multiple CSPS system by multi-agent reinforcement learning[J]. *Acta Automatica Sinica*, 2010, 36(2): 289 – 296.)

[11] CAO X R. *Stochastic Learning and Optimization: a Sensitivity-Based View*[M]. New York: Springer, 2007.

[12] ALBUS J S. A new approach to manipulator control: the cerebellar model articulation controller (CMAC)[J]. *Journal of Dynamic Systems, Measurement, and Control Transactions of ASME*, 1975, 1(9): 220 – 227.

[13] MILLER W T, GLANZ F H, KRAFT L G. Application of a general learning algorithm to the control robot manipulators[J]. *International Joint Robotic Research*, 1987, 6(2): 84 – 98.

[14] 唐昊, 丁丽洁, 程文娟, 等. 搬运系统作业分配问题的小脑模型关节控制器Q学习算法[J]. 控制理论与应用, 2009, 26(8): 884 – 888.
(TANG hao, DING Lijie, CHENG Wenjuan, et al. The cerebellar-model-articulation-controller Q-learning for the task assignment of a handling system[J]. *Control Theory & Applications*, 2009, 26(8): 884 – 888.)

[15] 朱大奇, 孔敏. 基于平衡学习的CMAC神经网络非线性滑模容错控制[J]. 控制理论与应用, 2008, 25(1): 81 – 86.
(ZHU Daqi, KONG Min. Fault-tolerant control of nonlinear system based on balanced learning CMAC neural network[J]. *Control Theory & Applications*, 2008, 25(1): 81 – 86.)

[16] 孙炜, 王耀南. 模糊CMAC及其在机器人轨迹跟踪控制中的应用[J]. 控制理论与应用, 2006, 23(1): 38 – 42.
(SUN Wei, WANG Yaonan. Fuzzy cerebellar model articulation controller and its application on robotic tracking control[J]. *Control Theory & Applications*, 2006, 23(1): 38 – 42.)

作者简介:

周 雷 (1981—), 男, 博士研究生, 讲师, 主要研究方向为离散事件动态系统、强化学习、无线网络, E-mail: zhouleizhl@163.com;

孔 凤 (1983—), 女, 硕士研究生, 研究方向为CSPS的在线优化方法, E-mail: kongfeng11@163.com;

唐 昊 (1972—), 男, 教授, 中国计算机学会高级会员, 目前研究方向为事离散事件动态系统、强化学习和神经元动态规划等仿真优化、鲁棒决策机制及智能优化理论和方法等, E-mail: htang@hfut.edu.cn;

张建军 (1963—), 男, 教授, 研究方向包括汽车电子、计算机集成制造技术研究, 制造业信息化等, E-mail: zjj@ialab.hfut.edu.cn.