文章编号 :1002-0411 (2004)02-0145-06

可重入生产系统的平均报酬型强化学习调度

柳长春 , 沈志江 , 于海斌
(中国科学院沈阳自动化研究所 , 辽宁 沈阳 110016)

摘 要 :在可重入生产系统中 ,一个重要的问题就是对调度策略进行优化 ,以提高系统平均输出率 .本文采用了一种平均报酬型强化学习算法来解决该问题 ,直接从所关心的系统品质出发 ,自动获得具有自适应性的动态调度策略 .仿真结果表明 ,其性能优于两种熟知的优先权调度策略 .

关键词 :平均报酬型强化学习 ;可重入系统 ;调度 ;暂态差分
中图分类号 : *TP*13 文献标识码 : *B*

Average Reward Reinforcement Learning Scheduling of Closed Reentrant Production Systems

LIU Chang-chun , SHEN Zhi-jiang , YU Hai-bin
(*Shenyang Institute of Automation , Chinese Academy of Sciences , Shenyang* 110016 , *China*)

Abstract : How to schedule the closed reentrant queueing networks so as to maximize the system mean output is an intractable NP-hard problem . In this paper , a method of average reward reinforcement learning (RL) is applied to automatically find an adaptive scheduling policy by directly optimizing the mean output . Numerical study demonstrates that the RL scheduler consistently outperforms all the known priority policies .

Keywords : average reward reinforcement learning ; reentrant system ; scheduling ; temporal differences

1 引言 (Introduction)

随着半导体 .胶卷等高新技术产品生产的发展 ,一类特殊的制造系统 ——可重入生产系统正受到越来越广泛的关注 .在该系统中 ,由于某些设备非常精密和昂贵 ,不可能在所有需要的工序都加以配备 ,必须要求工件在加工过程中重复访问这些瓶颈设备 .为了避免加工冲突 ,提高系统平均输出率 ,有必要研究其调度问题 . Kumar 将可重入系统列为继 job-shop 和 flow-shop 之后的第三类生产系统[1] .它的显著特点是工件在加工过程中的不同阶段可能重复访问某些机器 ,不同加工阶段的工件可能竞争同一台机器 .这使得它的调度问题比其它系统更为复杂 .该系统工件流量大 ,加工路径确定 ,而且由于各个工序加工时间服从指数分布导致了系统的随机性 .最优调度策略必是非空闲策略 ,只要加工站前有缓冲区不空 ,机器就不能空闲[2,3] .本文只研究此类策略 .

虽然可重入生产系统调度问题可以抽象为马尔可夫决策过程(MDP) ,但该问题是一个 NP 问题 ,无法用动态规划法求解 .此外 ,由于可重入生产系统的强耦合性 ,针对状态空间较大的 MDP 的问题分解方法在这里也是不适用的 .因此目前的研究多集中于优先权调度策略 ,如最先缓冲区优先服务策略 FBFS(First Buffer First Serve) 和负载平衡策略 WBAL(Workload Balancing)[4] .其中 WBAL 在两站闭环系统中有最好的性能 .虽然优先权调度策略研究取得了一定进展[2,4] ,但它没有充分地利用系统的状态信息 .另外 ,策略的产生往往依赖于对系统的启发式知识 ,而不是直接由明确的优化指标来指导策略的自动生成 .

强化学习 RL(Reinforcement Learning) 采用试错(Trail-and-error) 法 ,同环境进行动态交互 ,根据经验学习得到最优策略 .其中 ,平均暂态差分算法 TD(λ) (Average Cost Temporal-Difference Learning) 是最近兴起的 ,由 Tsitsiklis 和 Roy 在 1997 年提出[5] ,可以直接优化平均类型指标 .本文采用该算法并结合函数近似技术对可重入生产系统调度策略寻优 ,以提高系统平均输出率 .

收稿日期 :2002 - 12 - 24
基金项目 :国家自然科学基金资助项目(69974039)

2 可重入系统模型及优先权调度策略 (Model of reentrant systems and the priority scheduling policies)

2.1 闭环可重入生产系统的马尔可夫模型

在一个典型闭环可重入生产系统中,有 s 个加工站$\{1,2,\cdots,s\}$,每个加工站 $\sigma\in\{1,2,\cdots,s\}$包括一台机器.系统总共有 L 道工序,工序 l 的工件进入加工站 $\sigma(l)\in\{1,2,\cdots,s\}$,并在缓冲区 b_l 中等待,完工后进入加工站 $\sigma(l+1)$,在缓冲区 b_{l+1}中等待.加工站 $\sigma(L)$的缓冲区 b_L 是最后一个访问的缓冲区.顺序 $\sigma(1),\cdots,\sigma(L)$构成加工路径.由于存在某些阶段 $\sigma(i)=\sigma(j)$且 $i\neq j$,所以该系统被称为可重入系统.缓冲区 b_l 中的工件的加工时间服从以 $1/\mu_l=m_l<\infty$为均值的指数分布,系统生产单一类型的产品,所有工件的加工路径相同且确定,每台机器同时只能加工一个工件.当系统夹具有限时,采用闭环投料策略:系统每输出一个工件,就向系统输入一个工件,从而保证系统中工件个数一定.图1 即为一个典型可重入生产系统.

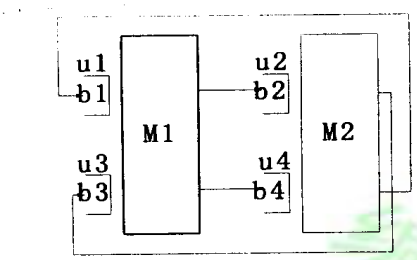

图1 典型可重入生产系统

Fig.1 Typical reentrant production systems

下面将该调度问题描述成为一个动态规划问题.引入如下表示:$x_1=[x_t(1),x_t(2),\cdots,x_t(L)]^{\mathrm{T}}\in X$ 为系统的状态,其中,X 表示状态空间;$x_t(l)$为 t 时刻缓冲区 b_l 中的工件个数,包括正被加工的工件,满足 $\sum_{l=1}^{L}x_t(l)=N$. 控制行动记为 $u_t=[u_t(1),u_t(2),\cdots,u_t(L)]^{\mathrm{T}}\in U$,$U$ 是所有控制动作的集合,一系列控制行动$\{u_t:t\geqslant 0\}$构成了一个调度策略 $\pi:X\to U$,它是从状态空间到控制动作集合的一个映射.对状态 $x\in X$, 允许控制动作可以描述为:$\pi(x)=a=[a_1,\cdots,a_L]^{\mathrm{T}}\in A(x)\subseteq U$.其中,$A(x)$为状态 x 的允许控制行动的集合.一个允许控制动作 a 必须满足可重入性约束、资源约束及非空闲条件.若缓冲区中的工件被加工,则 $a_t=1$,否则 $a_t=0$.以上定义了一个有限状态的马氏决策过程.通过适当选取时间单位,使 $\sum_{l=1}^{L}\mu_l=1$,可以得到相应的离散时间可控马尔可夫链,其状态转移概率如下:

$$\begin{cases} P_a(x,x-e_l+e_{l+1})=\mu_l a_l & 1\leqslant l\leqslant L-1\\ P_a(x,x-e_L+e_1)=\mu_L a_L & \\ P_a(x,x)=1-\sum_{l=1}^{L}\mu_l a_l & \end{cases} \tag{1}$$

其中,$e_l(l=1,\cdots,L)$ 为第 l 个元素为 1、其余元素全为 0 的列向量.优化目标是寻找一个调度策略 π^*,根据当前状态,确定一台机器前各缓冲区工件的加工顺序,最大化系统平均输出率:

$$\rho(\pi)=\lim_{T\to\infty}\frac{E\left(\sum_{t=0}^{T}g(x_t,\pi(x_t),x_{t+1})\right)}{T+1} \tag{2}$$

其中,g 为瞬时报酬函数,定义为:

$$g(x,a,y)=\begin{cases}1 & \text{if}\quad y_1-x_1=1\\ 0 & \text{otherwise}\end{cases} \tag{3}$$

它表示,在状态 x,采用策略 π规定的控制动作 $a=\pi(x)$,系统转移到状态 y,如果输出一个工件,RL 智能体将获得一个单位的奖励;否则报酬为 0.一个策略 π^* 是最优的,当且仅当对于所有允许控制策略 π都有:$\rho(\pi^*)\geqslant\rho(\pi)$,记 $\rho(\pi^*)$为 ρ^*.

如果在某策略下离散马尔可夫链是强遍历的,那么平均报酬 $\rho(\pi)$和起始状态无关[6].可重入生产系统是强遍历的[2,7],但是在策略 π下,系统从起始时刻到一个极大的时刻 t 得到的积累报酬却和起始状态 s 相关,记为 $\rho(\pi)t+e_t(s)$.其中,$e_t(s)$是一个与 t 相关的补偿.定义策略 π的相对值函数 $h_\pi(s)$为:

$$h_\pi(s)=\lim_{l\to\infty}\frac{1}{l}\sum_{t=1}^{l}e_t(s) \tag{4}$$

它表示系统在策略 π下从状态 s 出发到时刻 t 得到的积累报酬超过 $\rho(\pi)t$ 那部分的期望值.

假定系统在最优策略 π^* 下,从状态 i 用一个时间单位转移到状态 j,得到瞬时报酬为 $g(i,\pi^*(i),j)$,则可推导出平均代价问题的 Bellman 方程:

$$\rho^*+h^*(t)=\lim_{u\in A(i)}\sum_{j=1}^{n}p_{ij}(u)(g(i,u,j)+h^*(j))$$
$$h^*(x')=0 \tag{5}$$

其中,h^* 是最优相对值函数,$A(i)$ 表示状态 i 上的允许控制动作集合,$p_{ij}(u)$ 表示处于状态 i 的系统在控制动作 u 的作用下转移到后继状态 j 的概率,ρ^* 为最优策略对应的平均报酬,x' 为一个任意选定状态.

2.2　闭环可重入生产系统的优先权调度策略

虽然该调度问题可以描述成为一个平均类型的随机动态规划(Dynamic Programming, DP)问题,但由于问题的 NP 特性,很难通过动态规划方法获得动态调度策略.例如一个 L 道工序的封闭可重入生产系统,如果系统中有 N 个工件,则共有 C_{N+L-1}^{L-1} 个状态.当 N 和 L 增大时,系统的状态空间将出现"组合爆炸".另外,可重入生产系统是一个强耦合系统,一台机器对应的子问题的状态转移概率和报酬,不仅和本机器的调度策略有关,而且依赖于其它机器所采用的调度策略.这使得难以采用问题分解的方法来压缩状态空间求解该调度问题.

因此,目前的研究多集中于优先权调度策略.它是一种非抢占式的静态策略.优先级用矩阵 O 表示,第 i 行表示第 i 台机器前的缓冲区的优先级,按照从高到低顺序排列.比如,对图 1 所示的闭环两站系统,WBAL 的优先权矩阵为:

$$O_{\mathrm{WBAL}} = \begin{vmatrix} b_1 & b_3 \\ b_4 & b_2 \end{vmatrix} \tag{6}$$

它表示,在机器 1 上,b_1 优先级比 b_3 高,在机器 2 上,b_4 优先级比 b_2 高.只有当高优先权缓冲区为空的时候,才发生优先权转换,对低优先权缓冲区内的工件进行加工.

虽然 WBAL 策略比其它优先权策略有更好的性能,但它对系统状态信息的利用仍然是不充分的.对于每个缓冲区,只考虑了是"空"还是"非空",而没有对这个"非空"进行任何评价."非空"缓冲区内有多少工件,是 1 个还是 N 个,是没有区别的.此外,在优先权策略中,一台机器的控制动作完全取决于该机器前缓冲区内工件的分布情况,而忽略其它机器的影响.这不符合可重入系统的强耦合性.这种对缓冲区状态的粗糙评价和对系统强耦合性的忽视使得优先权策略无法针对当前状态选择更合理的控制动作,进一步提高系统平均输出率.

此外,从策略产生的方法和角度来看,优先权调度策略多是通过对系统的简单定性分析,有一定局限性,比如 WBAL 策略就只适用于两站系统.直接由品质指标出发自动生成调度策略的研究还不多见.

3　平均报酬型强化学习调度策略寻优 (Optimizing the scheduling policy based on average reinforcement learning)

本文基于函数近似,应用一种平均报酬型强化学习算法,直接优化系统平均输出率,进而自动获得性能优越的动态调度策略.

平均报酬型强化学习,通过学习获得最优相对值函数 h^* 的某种逼近 $h_{\mathrm{app}}(\cdot, r)$ 来近似解决相应的动态规划问题.实际上,如果获得最优相对值函数 h^*,则最优策略 π^* 可以通过一步搜索得到:

$$\pi^*(i) = \arg\max_{u \in A(i)} E\{ g(i,u,j) + h^*(j) \} \tag{7}$$

函数近似方法采用近似函数 $h_{\mathrm{app}}(\cdot, r)$ 来逼近最优相对值函数 h^*,以获得一个性能较好的次优策略:

$$\pi(i) = \arg\max_{u \in A(i)} E\{ g(i,u,j) + h_{\mathrm{app}}(j, r) \} \tag{8}$$

在函数近似中,特征向量被用来对状态空间进行表达,并作为近似函数 $h_{\mathrm{app}}(\cdot, r)$ 数的输入:

$$\Phi(i) = (\phi_1(i), \phi_2(i), \cdots, \phi_K(i)) \quad \forall i \in X \tag{9}$$

其中,特征分量是一个从状态空间到实数集的映射:$\phi_K : X \to R$.

函数近似是一种隐式的聚类,它在学习过程中动态地构建状态空间的划分,进而压缩问题的状态空间[6].因此基于函数近似的强化学习适合解决有较大规模状态空间的问题.此外,函数近似把学习到的知识紧缩存储到参数向量 r,而且不是把状态空间中不同状态的相对值函数的值存储到表格中的不同位置,这样可以减少存储时间开销,同时还使知识在学习过程中在相似的状态中得到推广,提高学习效率.本文采用一种线性结构的近似函数,相对其它近似结构,它有较高的学习效率及收敛速度[6]:

$$h_{\mathrm{app}}(i, r) = \sum_{k=1}^{K} r(k)\, \phi_k(i) \tag{10}$$

其中,ϕ_k 是状态空间 X 上的特征分量,$r = (r(1), \cdots, r(k))^{\mathrm{T}}$ 是参数向量,$r(k)$ 是特征分量 ϕ_k 的权重.由(7)式、(8)式,策略寻优问题可以归结为:通过学习,调整参数向量 r,使 $h_{\mathrm{app}}(\cdot, r)$ 和 h^* 之间的误差在某个模的意义下最小.

平均报酬型暂态差分算法 $TD(\lambda)$,可以通过学习,调整 $h_{\mathrm{app}}(\cdot, r)$ 中的参数向量 r,使其逼近某一确定策略 π 相对值函数 h_π.在策略 π 下,得到一个

状态序列$\{i_t \mid t=1,2,\cdots\}$,其中系统状态转移矩阵由$\pi$确定.在时刻$t$,参数向量$r$的值为$r_t$,平均报酬为$\rho(\pi)$,其近似值估计记为$\rho$,则与从$i_t$到$i_{t+1}$这个状态转移相对应的暂态差分为:

$$d_t = g(i_t,u,i_{t+1}) - \rho + h_{\text{app}}(i_{t+1},r_t) - h_{\text{app}}(i_t,r_t) \tag{11}$$

$TD(\lambda)$调整参数向量及平均报酬估计:

$$r_{t+1} = r_t + \gamma_t d_t \sum_{k=0}^{t} \lambda^{t-k} \phi(i_k) \tag{12}$$

$$\rho_{t+1} = (1-\eta_t)\rho + \eta_t g(i_t,u,i_{t+1}) \tag{13}$$

其中,参数$\lambda \in [0,1)$表达了对时间信度的分配,γ_t和η_t是学习率,控制动作u为$\pi(i_t)$.

在可重入生产系统调度策略寻优中,希望从任一给定的初始策略出发,通过学习,最终获得最优相对值函数h^*的近似,进而得到最优策略.为此,本文将$TD(\lambda)$算法与值迭代(Value iteration)相结合来解决该问题.在时刻t状态i,系统根据当前r_t由(8)式选择控制动作,转移到状态i_{t+1},然后根据(11)(12)(13)式用$TD(\lambda)$算法对参数向量及平均报酬的估计进行调整,得到r_{t+1}和ρ_{t+1},然后再由(8)式选择下一个控制动作.这样通过学习,$h_{\text{app}}(\cdot,r)$将逼近于h^*,ρ将趋于最优平均输出率ρ^*.需要指出的是,在整个学习过程中,参数向量是不断变化的,这意味着调度策略也是不断变化的.下面给出平均报酬型强化学习调度优化算法:

算法1 用平均报酬型$TD(\lambda)$算法对r进行学习:

① 初始化参数向量r_0、平均输出率ρ_0和系统状态x_0,令时间$t=0$,输出工件个数$n=0$.

② t时刻,观测系统当前状态x_t,得到所有该状态下的允许控制动作集合$A(x_t)$;根据某探索策略,选择状态x_t的控制动作u_t.实验采用如下方式:以一个概率β随机选择一个允许控制动作,而不是根据当前知识依照(8)式选择当前认为最优的控制动作.开始的时候设β为1,每选择完一个控制动作,β都以$\Delta\beta$递减.

③ 采用控制动作u_t使系统进入状态x_{t+1},根据(3)式定义的报酬函数得到瞬时报酬.如果系统输出一个工件,则$n=n+1$.

④ 根据(11)式和(12)式调整参数向量r_t,根据(13)式调整平均输出率ρ.

⑤ 判断是否满足终止条件,若n小于某个给定正数,令$t=t+1$;返回②,否则算法结束.

需要注意的是,虽然整个学习过程需要大量计算和试错,但根据"离线学习,在线应用"的思想,通过离线仿真学习得到的$h_{\text{app}}(\cdot,r)$可以在线产生调度策略,并能满足生产的实时性要求.

4 实验仿真与结果分析 (Experimental simulation and results analysis)

这里对图1所示的典型闭环可重入生产系统的调度问题进行仿真研究.首先,将算法1分别应用到工件个数为20和60的两个系统,获得动态策略SRLS(Scalable Reinforcement Learning Scheduler)和RLS(Reinforcement Learning Scheduler),并在各个系统中同优先权调度策略进行了比较;然后,考察通过学习得到的动态调度策略的推广性质:将SRLS直接应用到工件个数为60的系统中,进行了策略性能比较.同时还记录了获得RLS的学习趋势.

根据Kumar对可重入系统稳态分布的研究[2],选取两类信息在预处理后作为特征分量输入近似函数$h_{\text{app}}(\cdot,r)$:前$L-1$缓冲区内工件个数及其二次项.其中,二次项的引入可以获得对系统耦合性的某种近似.各特征分量定义为:

$$\phi_i(x) = \frac{x(i)}{\sum_{i=1}^{L-1} x_0(i)}, \quad i=1,2,\cdots,L-1$$

$$\phi_k(x) = \frac{x(i)\,x(j)}{\sum_{i=1}^{L-1}\sum_{j=1}^{L-1} x_0(i)\,x_0(j)} \tag{14}$$

x_0表示某一给定的初始状态.采用预处理是为了缩小各个特征的差别.这是因为,在采用线性拟合的函数近似中,特征和对应的权参数的乘积的正负和大小,表明了这个特征对近似函数,即对调度策略,影响的性质和程度.而在学习起始阶段,我们对此是没有任何知识的.实验表明,预处理过程有利于提高学习效率,避免振荡.

由于可重入系统是一个随机系统,所以对每个参数向量,即与之对应的调度策略进行评价的时候,需要多次仿真后取平均.评价过程需要大量计算,同时,为了获得学习趋势曲线,需要多次采样,为此引入评价参数:$EV=[ev\_win, ev\_num, ev\_out, cp\_num, cp\_out]$,其中,每$ev_win$次对$r$迭代后,对其进行一次采样,即选取当前调度策略,进行评价.在每次评价中,系统输出ev_out个工件为结束条件,并计算出系统平均输出率ρ,共实验ev_num次后再对ρ取平均.当学习结束,需选用较大的实验次数cp_num和每次实验结束前输出的工件个数cp_out,以便对最终获得的调度策略进行更严格的评价,进行策略性能比较.对工件个数为60

的系统进行调度策略优化的主要实验参数见表 1.

从表 2 .表 3 可以看到,对于封闭两站典型可重入系统,无论是系统平均输出率还是机器利用率,通过算法 1 得到的动态调度策略 RLS 和 SRLS,都优于最好的优先权策略 WBAL.同时,随着工件个数增大,性能提高更加明显.这是因为优先权调度策略对系统状态信息利用不充分,而且随着工件个数增加而加剧.动态调度策略 RLS 和 SRLS 则考虑了缓冲区中工件的个数和系统耦合性,所以获得了更好的系统性能.从图 2 的学习趋势曲线可以看出,从性能极其一般的与 r_0 对应的策略出发,算法 1 通过学习不断改善调度策略,自动有效地发现了性能优越的动态调度策略 RLS.

如果在小规模问题上得到的策略可以扩展到相似的大规模问题上,我们说该策略有相似扩展性.由表 2 ,在 $N=20$ 系统中得到的策略 SRLS 直接应用到 $N=60$ 这个较大系统中,其性能仍优越于 WBAL,只是比直接在该系统中学习得到的策略 RLS 性能略微下降.这是因为,基于函数近似的强化学习的本质,是通过仿真调整参数向量,将状态空间隐式地聚类,并使近似函数逼近最优值函数.通过对小规模问题的学习,获得的学习结果,即参数向量,蕴涵了如何对状态空间适当聚类和各个状态类的相对值函数的相对大小的知识.如果问题规模扩大,但函数近似的结构和特征向量的选取与小规模问题中相类似的话,那么,该学习结果在这个大规模的问题中也很有可能具有同样的性质.只不过由于状态空间规模增大,可能导致每个状态类的"粒度"增大,但相对值函数在状态类空间的形状可能和原来基本一致.这样,在小规模问题上学习得到的结果,就可以有效地扩展到这个相似的大规模的问题上去.仿真结果表明了通过平均报酬型强化学习得到的调度策略具有相似扩展性.需要注意的是,表 2 中 SRLS 性能比 RLS 略低,说明两个不同规模的问题虽然是相似的,但并不等价,所以针对该大规模问题本身直接学习的效果更好些.扩展性使得在较小的状态空间内通过离线学习得到的调度策略,能直接在线应用到大规模的实际问题中,这样能大大提高学习效率.

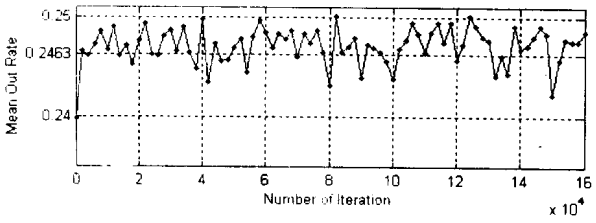

图 2　学习过程中输出率变化趋势图($N=60$)

Fig .2　Change trends of output during the learning process

表 1　主要实验参数 ($N=60$)

Tab .1　Main experiment parameters

实验参数	参数值
平均加工时间	$[m_1, m_2, m_3, m_4]=[1, 3.5, 3, 0.5]$
工件个数	$N=60$
初始参数向量	$r_0=e$
初始状态	$x_0=[15, 15, 15, 15]$
$[\gamma, \eta]$	[0 .005 ,0 .002]
EV	[2000 ,100 ,9999 ,500 ,9999]
x'	$x_0=[15, 15, 15, 15]$
λ	0 .5
$\Delta\beta$	0 .00005

表 2　调度策略性能比较 ($N=60$)

Tab .2　Performance comparison of the scheduling policies

调度策略	平均输出率	机器平均利用率	
		S1	S2
RLS	2 .4782e - 1	9 .8743e - 1	9 .9107e - 1
WBAL	2 .4630e - 1	9 .8386e - 1	9 .8958e - 1
FBFS	2 .4582e - 1	9 .7996e - 1	9 .8534e - 1
SRLS	2 .4757e - 1	9 .8602e - 1	9 .9076e - 1

表 3　调度策略性能比较($N=20$)

Tab .3　Performance comparison of the scheduling policies

调度策略	平均输出率	机器平均利用率	
		S1	S2
SRLS	2 .4127e - 1	9 .5842e - 1	9 .5991e - 1
WBAL	2 .3994e - 1	9 .5753e - 1	9 .5810e - 1
FBFS	2 .3602e - 1	9 .4520e - 1	9 .4658e - 1

5　结论 (Conclusion)

可重入生产系统调度是一个强耦合的状态空间巨大的复杂问题,目前的研究多集中于优先权调度策略.本文采用一种基于函数近似的平均报酬型暂态差分算法,通过学习调整参数向量,动态构建状态空间划分,获得了动态调度策略,其性能优于优先权调度策略.该策略是由系统品质指标直接出发自动获得的,是一种动态状态反馈自适应调度方法.此外,算法 1 的优势随着系统工件个数增大而增强,并且获得的调度策略具有相似推广性,这都预示了应用平均报酬型强化学习解决可重入生产系统调度问题的应用前景.

需要指出的是,平均暂态差分算法是最近兴起的,总体来说,应用该算法解决问题的成功范例非常少见[6].当与函数近似相结合的时候,在学习过程中,参数向量和对平均报酬的估计的调整相互影响,有时实验参数的选择会影响算法性能.因此,有必要进一步对基于函数近似的平均暂态差分算法及其参数敏感性分析进行进一步研究.

参 考 文 献 (References)

[1] Kumar P R. Reentrant queueing networks [J]. Special Issue Queuing Systems. 1993,13:87~110.

[2] Kumar P R. Scheduling manufacturing systems of re-entrant lines [A]. Yao D D. Stochastic Modeling and Analysis of Manufacturing Systems [M]. New York:Springer-Verlag, 1994.325~360

[3] Meyn S P. Stability and optimization of multiclass queueing networks and their fluid models [A]. Proceedings of the Summer Seminar on "The mathematics of Stochastic Manufacturing Systems"[C]. American Mathematical Society, 1997.

[4] Harrison J M, Kumar P R. Scheduling network of queues: heavy traffic analysis of a two-station closed network[J]. Operations Research, 1990,38(6):1052~1064.

[5] Tsitsiklis J N, Roy B V. Average cost temporal difference learning [J]. Automatica, 1999,35(11):1799~1808.

[6] Bertsekas D P, Tsitsiklis J N. Neuro-Dynamic Programming [M]. Athena Scientific, 1996.

[7] Jin H, Ou J, Kumar P R. The throughput of irreducible closed markovian queueing networks: functional bounds, asymptotic loss, efficiency, and Harrison-wei conjectures [J]. Mathematics of Operations Research, 1997, 22(4):886~920.

作者简介

柳长春(1975-),男,硕士生.研究领域为机器学习,强化学习,多智能体系统,智能调度等.

沈志江(1976-),男,硕士生.研究领域为机器学习,强化学习,多智能体系统等.

于海斌(1964-),男,研究员.研究领域为智能生产调度,分布式控制系统,离散时间动态系统等.

(上接第144页)

响程度不同的因素对实际对象的影响,实际应用结果表明该方法能够简化对象模型,提高模型精度.

参 考 文 献 (References)

[1] 于静江, 周春晖. 过程控制中的软测量技术[J]. 控制理论与应用, 1996, 13(2):137~144.

[2] Chen S, Billings S A. Neural networks for nonlinear dynamic system modeling and identification[J]. International Journal of Control, 1992, 56(2): 359~366.

[3] 王雅琳, 桂卫华, 阳春华, 等. 自适应监督式分布神经网络及其工业应用[J]. 控制与决策, 2001, 16(5): 549~552.

[4] Su H B, Fan L T, Schlup J R. Monitoring the process of curing of epoxy/graphite fiber composites with a recurrent neural network as a soft sensor[J]. Engineering Applications of Artificial Intelligence, 1998,11(2): 293~306.

[5] Park S Y, Han C H. A nonlinear soft sensor based on multivariate smoothing procedure for quality estimation in distillation columns [J]. Computers and Chemical Engineering, 2000,24(2-7): 871~877.

[6] Baum E B, Haussler D. What size net gives valid generalization [J]. Neural Computation, 1998,1(1):151~160.

[7] 王旭东, 邵惠鹤. 基于神经网络的通用软测量技术[J]. 自动化学报, 1998, 24(5): 702~706.

[8] 汪树玉, 刘国华, 李富强, 等. 因素分析法在观测数据处理上的应用[J]. 水利学报, 1998, 5(1): 28~32.

[9] Riedmiller M, Braun H. A direct adaptive method for faster backpropagation learning: the RPROP algorithm[A]. Proceedings of the IEEE International Conference on Neural Networks [C]. New York:IEEE Press, 1993.586~591.

作者简介

李勇刚(1974-),男,博士生.研究领域为复杂生产过程的建模及应用.

桂卫华(1950-),男,教授,博士生导师.研究领域为大系统理论,复杂过程建模及优化等.

陈 峰(1977-),男,博士生.研究领域为复杂生产过程的建模及应用.