基于TD(λ)的自然梯度强化学习算法

陈圣磊[1] 谷瑞军[1] 陈 耿[1] 薛 晖[2]
(南京审计学院信息科学学院 南京 211815)[1] (东南大学计算机科学与工程学院 南京 210096)[2]

摘 要 近年来强化学习中的策略梯度方法以其良好的收敛性能吸引了广泛的关注。研究了平均模型中的自然梯度算法,针对现有算法估计梯度时效率较低的问题,在梯度估计的值函数逼近中采用了TD(λ)方法。TD(λ)中的资格迹使学习经验的传播更加高效,从而能够降低梯度估计的方差,提升算法的收敛速度。车杆平衡系统仿真实验验证了所提算法的有效性。

关键词 策略梯度,自然梯度,TD(λ),资格迹
中图法分类号 TP181 **文献标识码** A

Natural Gradient Reinforcement Learning Algorithm with TD(λ)

CHEN Sheng-lei[1] GU Rui-jun[1] CHEN Geng[1] XUE Hui[2]
(School of Information Science, Nanjing Audit University, Nanjing 211815, China)[1]
(School of Computer Science and Engineering, Southeast University, Nanjing 210096, China)[2]

Abstract In recent years, policy gradient methods arouse extensive interests in reinforcement learning with its excellent convergence property. Natural gradient algorithms were investigated in this paper. To resolve the problem of low efficiency when estimating the gradient in present algorithms, TD(λ) method was used to approximate the value functions when estimating the gradient. The eligibility traces in TD(λ) make the propagation of learning experience more efficient. As a result, the variance in gradient estimation can be decreased and the convergence speed can be improved. The simulation experiment in cart-pole balancing system demonstrates the effectiveness of the algorithm.

Keywords Policy gradient, Natural gradient, TD(λ), Eligibility trace

1 引言

强化学习(Reinforcement Learning)是一类求解序列决策问题的机器学习方法,通过试探-评价的逐步迭代获得关于问题的知识,从而产生该决策问题的最优策略。与监督学习(Supervised Learning)要求给出教师信号不同,强化学习仅需要环境的评价性反馈信号,需要的环境模型信息较少,因此被认为是一种求解复杂决策问题的有效手段[1,2]。

根据是否显式地表示行为策略,强化学习可以分为值函数方法和策略梯度方法。值函数方法不会显式地表示行为策略,而是在学习过程中更新状态-动作对(State-Aciton Pair)的值函数,从估计的值函数中采用贪婪方法改进策略,如经典的Q学习[3]和多步Q学习[4]。值函数方法在一些应用中取得了显著的效果,但是也存在以下问题:只能求解确定性策略,无法处理随机策略问题;值函数的微小变化会导致策略很大的变化,策略变化不连续,使得值函数方法在很多问题中不能保证收敛[5]。而策略梯度方法直接给出策略函数的显式表示,并且估计优化目标对于策略参数的梯度,采用梯度下降法逼近局部最优策略。相对于值函数方法,策略梯度方法既可以学习确定性策略,又可以学习随机策略,并且收敛性也容易证明,因此近年来受到广泛的关注[6,7]。

最早的策略梯度算法可以追溯到Williams的REINFORCE算法[8]。Sutton对其进行了改进,采用学习得到的值函数减小梯度估计中的方差[5]。Bagnell提出采用自然梯度代替常规梯度,以解决常规梯度算法收敛速度慢的问题,取得了较好的效果,因此其成为目前策略梯度研究中的热点[9]。Peters, Morimura等人研究了折扣模型中的自然梯度算法[10,11],Bhatnagar等研究了平均模型中的增量式自然梯度Actor-Critic算法[12]。然而,Bhatnagar的算法在值函数估计中仅采用了即时差分(Temporal difference)TD(0)方法,不能有效地传播学习经验,并且收敛速度较慢,导致这类算法在处理连续状态控制问题时无法满足实际应用的需要。本文在平均模型框架中,针对现有算法梯度估计时效率较低的问题,在梯度估计的值函数逼近中采用TD(λ)方法。TD(λ)方法中的资格迹能够更加高效地传播学习经验,从而能够降低梯度估计的方差,提高算法的性能。仿真实验表明,改进的算法取得

到稿日期:2010-01-25 返修日期:2010-04-05 本文受国家自然科学基金项目(70971067,60905002),江苏省高校自然科学重大基础研究项目(08KJA520001),江苏省六大人才高峰项目(2007148)资助。

陈圣磊(1977—),男,博士,讲师,主要研究方向为机器学习与数据挖掘,E-mail:tristan_chen@126.com;谷瑞军(1979—),男,博士,讲师,主要研究方向为模式识别与数据挖掘;陈 耿(1965—),男,博士,教授,主要研究方向为数据挖掘、计算机审计;薛 晖(1979—),女,博士,讲师,主要研究方向为机器学习、模式识别。

了更好的效果。

2 策略梯度框架与自然梯度

2.1 策略梯度框架

通常采用马尔可夫决策过程(Markov Decision Process, MDP)模型来描述强化学习问题。MDP 模型可以定义为一个五元组:$\langle T,S,A,r(s,a),p(\cdot|s,a)\rangle$[13],$T=\{0,1,\cdots\}$为时刻集合,$S$ 为状态集合,A 为动作集合,回报函数 $r(s,a)=E[r_{t+1}|s_t=s,a_t=a]$,状态转移概率 $p(s'|s,a)=\Pr(s_{t+1}=s'|s_t=s,a_t=a)$,$\forall a\in A,\forall s,s'\in S$。系统在 t 时刻根据随机策略 $\pi(a|s)=\Pr(a_t=a|s_t=s)$选取动作。策略 π 下的平均回报函数定义为

$$J(\pi)=\lim_{N\to\infty}\frac{1}{N}E[\sum_{t=0}^{N-1}r_{t+1}|\pi]=\sum_{s\in S}d^{\pi}(s)\sum_{a\in A}\pi(a|s)r(s,a)$$

式中,$d^{\pi}(s)$是状态 s 在策略 π 下的平稳分布。

平均模型下的动作值函数和状态值函数分别定义为

$$Q^{\pi}(s,a)=\sum_{t=0}^{\infty}E[r_{t+1}-J(\pi)|s_0=s,a_0=a,\pi]$$

$$V^{\pi}(s)=\sum_{a\in A}\pi(a|s)Q^{\pi}(s,a)$$

策略梯度方法的基本思想就是参数化策略,即 $\pi=\pi(a|s,\theta)$,其中 $\theta\in\mathcal{R}^l$ 是参数向量。然后利用优化目标对于策略参数的梯度$\nabla_{\theta}J(\pi)$更新参数 θ。然而该梯度难以利用解析方法计算,只能通过其它途径进行估计。Sutton 给出了如下的策略梯度定理[5]。

定理 1 对于任意给定的 MDP,无论是折扣型回报还是平均型回报,均有下式成立。

$$\nabla_{\theta}J(\pi)=\sum_{s\in S}d^{\pi}(s)\sum_{a\in A}\nabla_{\theta}\pi(a|s)Q^{\pi}(s,a) \tag{1}$$

该梯度表达形式中没有出现策略变化对状态分布的影响$\nabla_{\theta}d^{\pi}(s)$,因此适宜通过采样估计梯度值。在 Williams 的 REINFORCE 算法中[8],采用实际的回报 $R_t=\sum_{k=1}^{\infty}r_{t+k}-J(\pi)$估计动作值函数 Q^{π},进而估计出策略梯度值,求得最优策略。然而该方法估计梯度时的方差较大,速度较慢。Sutton 提出采用逼近函数$\hat{Q}_{\omega}^{\pi}$来近似 Q^{π}[5],如果逼近函数与策略参数相容,即$\nabla_{\omega}\hat{Q}_{\omega}^{\pi}(s,a)=\nabla_{\theta}\log\pi(a|s)$,也就是有$\hat{Q}_{\omega}^{\pi}(s,a)=\psi(s,a)^{T}\omega$,其中 $\psi(s,a)=\nabla_{\theta}\log\pi(a|s)$,那么在最小均方误差意义下,采用梯度下降调整参数 ω,得到的最优函数值$\hat{Q}_{\omega^*}^{\pi}$即是 Q^{π} 的最佳逼近,因此可以在梯度估计中代替 Q^{π},由此给出了策略梯度估计的另一种方法,取得了更好的效果。

2.2 自然梯度强化学习

上述带有逼近函数的梯度估计方法提升了策略梯度强化学习算法的性能,但是收敛仍然较慢,在实际应用中效率不高。Kakade 和 Bagnell 等人提出采用自然梯度(Natural Gradient)代替常规梯度的思想[9,14]。自然梯度$\tilde{\nabla}_{\theta}J(\pi)$可以这样计算:$\tilde{\nabla}_{\theta}J(\pi)=G^{-1}(\theta)\nabla_{\theta}J(\pi)$,其中 $G(\theta)$表示 Fisher 信息矩阵。

当使用逼近函数$\hat{Q}_{\omega}^{\pi}$来代替 Q^{π} 时,策略梯度可以表示为

$$\begin{aligned}\nabla_{\theta}J(\pi)&=\sum_{s\in S}d^{\pi}(s)\sum_{a\in A}\nabla_{\theta}\pi(a|s)\hat{Q}_{\omega}^{\pi}(s,a)\\&=\sum_{s\in S}d^{\pi}(s)\sum_{a\in A}\nabla_{\theta}\pi(a|s)\nabla_{\theta}\log\pi(a|s)^{T}\omega\\&=\sum_{s\in S}d^{\pi}(s)\sum_{a\in A}\pi(a|s)\nabla_{\theta}\log\pi(a|s)\nabla_{\theta}\log\pi(a|s)^{T}\omega\\&=F(\theta)\omega\end{aligned} \tag{2}$$

式中,$F(\theta)=\sum_{s\in S}d^{\pi}(s)\sum_{a\in A}\pi(a|s)\nabla_{\theta}\log\pi(a|s)\nabla_{\theta}\log\pi(a|s)^{T}$。

Peters 证明了在强化学习模型中,$F(\theta)=G(\theta)$[10],因此自然梯度$\tilde{\nabla}_{\theta}J(\pi)=G^{-1}(\theta)F(\theta)\omega=\omega$。所以我们只要能估计出 ω,就可以得到平均回报函数的自然梯度。

3 基于 TD(λ)的自然梯度强化学习

3.1 动作值函数的参数更新

Sutton 等证明了当动作值函数的逼近函数与策略参数相容,并且由最小化均方误差

$$\varepsilon^{\pi}(\omega)=\sum_{s\in S}d^{\pi}(s)\sum_{a\in A}\pi(a|s)[Q^{\pi}(s,a)-\psi(s,a)^{T}\omega]^2$$

得到 ω^* 时,可以在梯度估计中使用$\hat{Q}_{\omega^*}^{\pi}$代替 Q^{π}[5]。并且,$\psi(s,a)^{T}\omega^*$ 既可以看作是优势函数 $A^{\pi}(s,a)=Q^{\pi}(s,a)-V^{\pi}(s)$的最小二乘解,也可以看作是动作值函数 $Q^{\pi}(s,a)$的最小二乘解。然而,由于$\sum_{a\in A}\pi(a|s)\psi(s,a)^{T}\omega=0$,$\forall s\in S$,最好将 $\psi(s,a)^{T}\omega^*$ 看作是优势函数的最小二乘逼近[12]。

基于上述结论,我们可以通过最小化均方误差

$$\varepsilon^{\pi}(\omega)=\sum_{s\in S}d^{\pi}(s)\sum_{a\in A}\pi(a|s)[A^{\pi}(s,a)-\psi(s,a)^{T}\omega]^2 \tag{3}$$

来求得 ω^*。运用梯度下降法,其梯度值为

$$\begin{aligned}\nabla_{\omega}\varepsilon^{\pi}(\omega)&=2\sum_{s\in S}d^{\pi}(s)\sum_{a\in A}\pi(a|s)[\psi(s,a)^{T}\omega-A^{\pi}(s,a)]\psi(s,a)\\&=2\sum_{s\in S}d^{\pi}(s)\sum_{a\in A}\pi(a|s)[\psi(s,a)\psi(s,a)^{T}\omega-A^{\pi}(s,a)\psi(s,a)]\end{aligned}$$

Bhatnagar 等已经证明 TD 误差 $\delta_t=r_{t+1}-\hat{J}_{t+1}+\hat{V}(s_{t+1})-\hat{V}(s_t)$是$A^{\pi}(s_t,a_t)$的相合估计[12],因此$\nabla_{\omega}\hat{\varepsilon}^{\pi}(\omega)=2[\psi(s_t,a_t)\psi(s_t,a_t)^{T}\omega-\delta_t\psi(s_t,a_t)]$是$\nabla_{\omega}\varepsilon^{\pi}(\omega)$的相合估计。所以,只要求得 δ_t,就可以采用梯度下降法求解 ω。

3.2 资格迹的引入

在计算 δ_t 时需要用到平均回报和值函数的估计$\hat{J}$,$\hat{V}$。$\hat{J}$是比较容易估计的,然而估计$\hat{V}$就需要采用函数逼近的方式。Bhatnagar 等使用单步 TD 学习估计值函数[12],然而单步 TD 学习中传播学习经验的效率较低,从而导致算法收敛较慢。为解决这一问题,本文引入了资格迹。资格迹由 Sutton 提出,是强化学习中用于分配时域信度的技巧[15]。通过资格迹,TD(λ)可以把回报值沿当前路径反向传播,而无需显式地记录经历的路径。因此,资格迹的引入使得学习经验的传播更加高效,从而可以更加精确地估计值函数,降低梯度估计的方差。

假设采用线性函数$\hat{V}(s)=v^{T}\varphi(s)$来逼近值函数 V,$\varphi(s)$为基函数,v 为参数向量。t 时刻每个特征的迹向量依赖于当前路径以及 λ:$z_t=\sum_{i=1}^{t}\lambda^{t-i}\varphi(s_i)$,$0\leqslant\lambda\leqslant1$。由于 z_t 具有指数权重,很容易实现增量式计算。

$$\begin{aligned}z_t&=\sum_{i=1}^{t}\lambda^{t-i}\varphi(s_i)=\lambda\sum_{i=1}^{t-1}\lambda^{t-1-i}\varphi(s_i)+\varphi(s_t)\\&=\lambda z_{t-1}+\varphi(s_t)\end{aligned} \tag{4}$$

在实际计算中,首先计算一步 TD 误差 $\delta_t=r_{t+1}-\hat{J}_{t+1}+\hat{V}(s_{t+1})-\hat{V}(s_t)$,然后将该误差按各自的迹 z_t 分配给所有的

状态特征。因此，TD(λ)学习中逼近函数$\hat{V}$的更新规则为

$$v_{t+1}=v_t+\alpha_t\delta_t z_t \tag{5}$$

当λ=0时，TD(λ)就退化为单步TD算法。可以看出，由于资格迹z_t中包含了所有的历史状态信息，因此能够更加有效地分配误差，从而加快了学习经验的传播。

由此，就可以得到基于TD(λ)的自然梯度学习算法。下面是算法的完整描述。

初始化：

①初始化状态s_0

②初始化参数v_0,ω_0,θ_0

③$J_0=0,z_{-1}=0$

For $t=0,1,2\cdots$ do

1. 执行动作

①选择动作，$a_t\sim\pi(a_t|s_t;\theta_t)$

②执行a_t后观察下一个状态，$s_{t+1}\sim p(s_{t+1}|s_t,a_t)$

③观察回报r_{t+1}

2. 评价动作

①更新平均回报，$\hat{J}_{t+1}=(1-\xi_t)\hat{J}_t+\xi_t r_{t+1}$

②更新迹，$z_t=\varphi(s_t)+\lambda z_{t-1}$

③计算TD误差，$\delta_t=r_{t+1}-\hat{J}_{t+1}+v_t^T\varphi(s_{t+1})-v_t^T\varphi(s_t)$

④更新状态值函数参数，$v_{t+1}=v_t+\alpha_t\delta_t z_t$

⑤更新动作值函数的参数，$\omega_{t+1}=\omega_t-\alpha_t[\psi(s_t,a_t)\psi(s_t,a_t)^T\omega_t-\delta_t\psi(s_t,a_t)]$

3. 更新策略参数

①$\theta_{t+1}=\theta_t+\beta_t\omega_{t+1}$

4 仿真实验与分析

为了验证本文算法的有效性，采用车杆平衡系统(Cart-Pole Balancing System)进行平衡控制仿真研究。车杆平衡系统包括一个可以在轨道上左右移动的小车，以及铰接在小车上、可以在轨道平面内自由转动的摆杆[16]，如图1所示。车杆平衡系统的目的在于通过改变施加在小车的力使摆杆尽可能保持平衡。

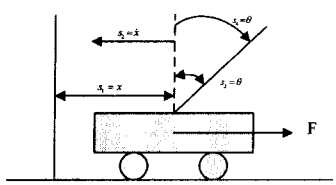

图1 车杆平衡系统

系统的状态由小车的水平位移x、水平速度$\dot{x}$、摆杆与竖直方向的夹角θ以及角速度$\dot{\theta}$表示，即$s=[x,\dot{x},\theta,\dot{\theta}]^T$。动作即为施加在小车上的力$F$，即$a=F$。在每个时间步中，要求小车水平位移必须保持在离轨道中心±4m范围内，同时摆杆必须保持在与垂直方向夹角±45°以内。如果车杆的状态不符合上述要求，则给予回报值−1，否则为0。车杆平衡系统的状态转移规律由如下方程描述。

$$\ddot{\theta}=\frac{g\sin\theta-\alpha ml\dot{\theta}^2\sin\theta\cos\theta-\alpha\cos\theta F}{4l/3-\alpha ml\cos^2\theta}$$

$$\ddot{x}=\alpha(F+ml\dot{\theta}^2\sin\theta-ml\ddot{\theta}\cos\theta)$$

式中，g为重力加速度常数，$g=9.8\text{m/s}^2$，m为摆杆的质量，$m=0.1\text{kg}$，M为小车的质量，$M=1.0\text{kg}$，l为摆杆长度的一半，$l=0.5\text{m}$，$\alpha=1/(m+M)$。仿真频率为50Hz。

假定施加在小车上的力符合正态分布，其均值为状态向量的线性函数，即力的选择策略为$\pi(a|s)=N(Ks,\sigma^2)$。为确保方差在参数更新中始终为正值，令$\sigma=0.1+1/(1+\exp(\eta))$。因此策略参数为$\theta=[K^T,\eta]$。实验中为防止参数更新过快，设定更新条件为$\measuredangle(\omega_{t+1},\omega_t)\leqslant\pi/90$。

在选择值函数的基函数时，我们在径向基函数中广泛应用的高斯函数。对每个状态s，取分布在四维状态空间中的405个点的高斯函数和常数1作为基函数，因此值函数的基向量为

$$\varphi(s)=(1,e^{-\frac{\|s-\mu_1\|^2}{2\sigma^2}},e^{-\frac{\|s-\mu_2\|^2}{2\sigma^2}},\cdots,e^{-\frac{\|s-\mu_{405}\|^2}{2\sigma^2}})^T$$

式中，$\mu_i(i=1,2,\cdots,405)$为四维状态空间上的网格点，各分量取值分别为$x\in\{-4,-2,0,2,4\}$，$\dot{x}\in\{-1,0,1\}$，$\theta\in\{-45,-33.75,-22.5,-11.25,0,11.25,22.5,33.75,45\}$，$\dot{\theta}\in\{-10,0,10\}$。方差$\sigma$取值为1。

将本文算法(NAC(λ))与Bhatnagar在文献[12]中提出的算法1和算法3进行了对比。算法1采用常规梯度，我们称之为RGAC，而算法3采用自然梯度，称之为NAC。学习率参照文献[17]设置，λ取值为0.3。实验结果如图2所示，图中数据是10次运行结果的平均值。从图中可以看出，RGAC经过300次仿真还未收敛，而NAC的时间步数达到370左右之后又急剧下降。分析发现，NAC在多数运行中可以收敛，但是个别情况中无法维持平衡态，影响了算法的整体性能。NAC(λ)经过大约150次仿真后时间步数就收敛到410左右，呈现出良好的收敛性与稳定性。

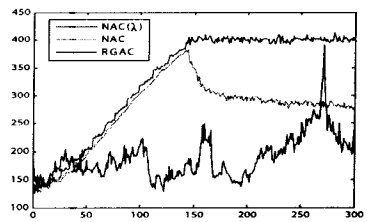

图2 3种算法的性能比较

结束语 强化学习中的策略梯度方法可以学习随机策略，并且收敛性也容易证明，因此近年来受到广泛的关注。本文研究了平均模型中的自然梯度算法，对于梯度估计中的值函数逼近问题，引入资格迹，使用TD(λ)来更新值函数的估计，从而降低了梯度估计的方差，提升了算法的收敛速度。车杆平衡系统仿真实验验证了本文算法的有效性。然而需要指出的是，尽管自然梯度方法比常规策略梯度算法性能有一定程度的提升，但是在所能维持的平衡次数上仍然难以与值函数方法相比。因此，如何在确保收敛的前提下寻求全局最优解，仍然是需要进一步研究的课题。

参 考 文 献

[1] 徐昕，贺汉根. 神经网络增强学习的梯度算法研究[J]. 计算机学报，2003，26(2)：227-233

[2] 周文云，刘全，李志涛. 一种大规模离散空间中的高斯强化学习方法[J]. 计算机科学，2009，36(8)：247-249

[3] Watkins J C H，Dayan P. Q-learning[J]. Machine Learning. 1992，8(1)：279-292

[4] 陈圣磊，吴慧中，韩祥兰，等. 一种多步Q强化学习方法[J]. 计算机科学，2006，33(3)：147-150

[5] Sutton R S, Mcallester D, Singh S, et al. Policy gradient methods for reinforcement learning with function approximation[C]// Proceedings of Advances in Neural Information Processing Systems 2000. 2000:1057-1063
[6] 王学宁,徐昕,吴涛,等. 策略梯度强化学习中的最优回报基线[J]. 计算机学报,2005,28(6):1021-1026
[7] 王学宁,陈伟,张锰,等. 增强学习中的直接策略搜索方法综述[J]. 智能系统学报,2007,2(1):16-24
[8] Williams R J. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning[J]. Machine Learning,1992,8(3/4):229-256
[9] Bagnell J A, Schneider J. Covariant Policy Search [C]// Proceedings of International Joint Conference on Artifical Intelligence. 2003
[10] Peters J, Schaal S. Natural Actor-Critic [J]. Neurocomputing, 2008,71(7-9):1180-1190
[11] Morimura T, Uchibe E, Doya K. Utilizing the natural gradient in temporal difference reinforcement learning with eligibility traces [C]// Proceedings of International Symposium on Information Geometry and its Application. 2005:256-263
[12] Bhatnagar S, Sutton R, Ghavamzadeh M, et al. Incremental Natural Actor-Critic Algorithms[C]// Proceedings of Advances in Neural Information Processing Systems. Vancouver, 2007
[13] 刘克. 实用马尔可夫决策过程[M]. 北京:清华大学出版社,2004
[14] Kakade S. A Natural Policy Gradient [C]// Proceedings of Advances in Neural Information Processing Systems. Vancouver, 2001
[15] Sutton R S. Learning to Predict by the Methods of Temporal Differences[J]. Machine Learning,1988,3:9-44
[16] Barto A G, Sutton R S, Anderson C W. Neurolike adaptive elements that can solve difficult learning control problems[J]. IEEE Transactions on System, Man and Cybernetics, 1983, 13(5):834-846
[17] Shalabh B, Richard S S, Mohammad G, et al. Natural actor-critic algorithms[J]. Automatica, 2009, 45(11):2471-2482

(上接第 177 页)

```
end if;
//选择最优货位
v_BestLocation:=f_GetBestLocation(v_BestVertical)
if (v_BestLocation is null)then
    return '#';
end if;
return v_BestLocation;
}
```

当通过 WMS 系统的"货位分配综合优化算法"计算得到最优货位后,把备件的入库信息写进入库单,通过调度系统向堆垛机下发作业、通知堆垛机把备件送到指定的货位地址上,作业完成后,系统会参照入库单上的记录修改货位上的备件信息。但这里要强调的是,为了提高立体仓库的工作效率,当今大多数立体仓库管理系统都具有批量下发作业的功能,也就是说,仓库管理员可以一次操作把多个货位作业信息写进作业表,实现一次下发多个作业,堆垛机批量执行的处理方式。所以,当通过此算法获得目标货位地址下发作业后,备件的信息还在入库单中,在作业未完成时,管理系统不会修改货位上的库存信息,因此在算法实际应用过程中,计算重量的函数还要包括未完成的入库单中备件的重量信息。

5 实际运行效果

下面提取了秦皇岛港务集团公司自动化立体仓库中的仓库、货架、备件等实际数据为依据,通过 44 次入库后,随机提取部分数据:

提取第 1 号堆垛机数据。表 1 列出了第 1 号堆垛机范围内的 4 排重量,体现了在整体货架若干排上重量相对均匀的分布。

表 1 第一号堆垛机各排重量

排	1	2	3	4
重量(千克)	177.36	224.50	215.97	200.10

提取第 3 排数据。表 2 所列从第 3 区到第 10 区的重量。

表 2 第 3 排 3—10 区域重量

备件名称	区域	货位	数量	单位重量(千克)	重量(千克)
圆钢//A3//φ36//7.99KG/m.	3	2-1-11-2	3	7.99	23.97
灰口铸铁//HT15-33	4	2-1-16-4	1	6.53	6.53
		2-1-19-3	2	6.53	13.06
圆钢//A3//φ36//7.99KG/m.	5	2-1-22-3	3	7.99	23.97
圆钢//A3//φ35//7.55KG/m.	6	2-1-28-8	3	7.55	22.65
生铁//z14	7	2-1-34-6	3	5.90	17.7
铸铁棒//φ	8	2-1-37-9	4	4.70	18.8
生铁//z14	9	2-1-42-4	3	5.90	17.7
圆钢//A3//φ35//7.55KG/m.	10	2-1-50-5	3	7.55	22.65

可见在目标排的基础上,选择把列划分成若干个区域,可以更为有效地让每排货架受力均匀。区域的划分有助于整体重量的均匀分布,优于仅从最轻列来确定目标列。测试结果体现了货架承重在入库选择货位过程中的重要性,同时兼顾了其他原则的有效应用。

本文主要介绍了以重量均匀分布为目标的货位分配策略,提出了一个从堆垛机控制范围、排、区域、列、层等多个角度进行计算的级联式的货位优化分配的算法及其实现,基本满足了用户提出的重量均匀分布、备件品种均匀分布、就近存取等的库存管理需求,对于仓库货架的整体安全性、出入库工作的高效性都是非常有意义的。本文所提出的算法已经在秦皇岛港务集团公司备件中心智能化立体仓库中进行了应用实施,根据实际仓库、货架、货物的数据信息进行分析,基本满足了用户提出的重量均匀分布和品种均匀分布的要求,仓库的整体和局部货架承重受力比较均匀,保证了仓库的安全。整个系统已运行一年,效果良好,验证了本文所提策略的有效性。

参 考 文 献

[1] 常发亮,刘增晓,等. 自动化立体仓库拣选作业路径优化问题研究[J]. 系统工程理论与实践,2007(2)
[2] 王恒山,浦志华. 一种基于实时控制的立体仓库出入库算法[J]. 系统工程理论与实践,1997(6)
[3] 赵鹤君,张月芳. 微型计算机在自动化立体仓库管理中的应用
[4] 刘安宇,张仰森. "单仓库多用户"自动化立体仓库管理系统的设计[J]. 北京机械工业学院学报,2009,24(1)
[5] 王琦. 江门冷冻厂有限公司立体仓库管理系统设计实现[D]. 重庆:重庆大学,2007(5)