基于性能预测的遗传强化学习动态调度方法

魏英姿[1]，谷侃锋[2]
（1.沈阳理工大学信息科学与工程学院，沈阳 110159；2.中国科学院沈阳自动化研究所沈阳现代装备研究设计中心，沈阳 110016）

摘 要：针对作业车间动态调度问题，**在模式驱动调度的框架下，提出遗传强化学习动态调度方法**。首先，采用优先规则编码的染色体表达问题的解，**将染色体分割成基因模式作为分阶段调度算法的状态模式**；其次，**设计性能预测变量，构建启发式立即回报函数**，引导和加快遗传强化学习算法的搜索进程；再次，**设置遗传算子、强化学习及其相关参数以实现搜索过程“开采”与“探索”之间的平衡**；最后，仿真实验结果验证了遗传强化学习调度方法的有效性。

关键词：强化学习；遗传算法；预测；生产周期；作业车间动态调度
中图分类号：TP18 **文献标识码**：A **文章编号**：1004-731X (2010) 12-2809-04

Genetic Reinforcement Learning Approach to Dynamic Scheduling Based on Performance Prediction

WEI Ying-zi[1], *GU Kan-feng*[2]
(1. School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China;
2. Advanced Equipment Research and Design Center, Shenyang Institute of Automation, Chinese Academy of Science, Shenyang 110016, China)

Abstract: *In the framework of pattern driven scheduling, a genetic reinforcement learning* (GRL) approach to schedule the job in the dynamical job-shop was proposed. First, the chromosome was coded by preference rules-based representation for the problem. *The chromosome was divided into gene schema as state patterns for the multi-phase scheduling system.* Secondly, *a performance predictive variable to construct instant reward function was designed* which was used to guide the learning system to progress rapidly. Thirdly, *genetic operators, RL and controlling parameters carried out the search strategy for the balance of “exploration” and “exploitation”*. Finally, the simulation results verify the efficiency of GRL scheduling approach.

Key words: reinforcement learning; genetic algorithm; prediction; makespan; dynamic job-shop scheduling

引 言

实际生产过程中的调度问题大多属于动态调度问题。动态调度方法针对生产现场的实际情况的变化产生更具可操作性的决策调度方案，动态调度比静态调度更强调算法的鲁棒性及算法对扰动的快速响应能力[1]。在研究动态调度问题的方法中，优先规则调度是最基本、最具影响力的方法。优先规则根据某个准则从机器的待加工队列中选择一个作业进行加工，这种局域决策会很快得出结果，因此适用于动态调度环境[2-3]。模式驱动的调度（Pattern Driven Scheduling，简称，PDS）是动态调度发展的一种高级形式，它通过抽取调度系统的特征模式，以刻画系统所处的状态，运用机器学习技术对调度环境采取有自适应能力的动态调度策略，从而提高生产系统的性能。

遗传算法(Genetic Algorithm, GA)是20世纪70年代提出的一种智能优化方法，由于GA求解复杂优化问题的巨大潜力，使其在工业工程、人工智能、生物工程、自动控制等领域得到广泛而成功的应用。与传统的启发式算法相比，GA不适于邻域最优解的微调结构，因此把传统的启发式算法（如局域搜索）嵌入到GA中构造一个更强的GA是必不可少的。强化学习(Reinforcement Learning, RL)是20世纪90年代以来发展起来的一种机器学习方法，在智能控制、机器人系统规划及分析预测等领域都有应用。强化学习是一种目标驱动的自适应优化控制方法，与环境之间的交互是强化学习获取智能的来源[4]。Zhang[5]把强化学习应用于作业车间调度问题，他们的方法是一种修复迭代方法，在其学习系统中，每一状态都是一个完整的调度，采用的算法是TD(λ)。Aydin[6]运用 QIII学习算法训练 agent 动态选择分配规则求解job-shop动态调度问题，该方法本质上是一种基于先验知识的调度，却忽略了强化学习机制的作用效果。Wang[7]研究单机调度问题的强化学习方法，讨论了 Q 学习应用于生产调度的参数设定问题。Pettinger[8]研究采用RL控制GA运算进程的控制参数，求解了对称旅行商问题。王本年等[9]提出基于强化学习机制的遗传算法，依据染色体的适应度，作为各行动策略的联合奖赏，求解了经典函数优化问题。潘燕春等[10]研究依据适应值指导强化学习过程，求解了同顺序flow-shop静态调度问题。

强化学习算法求解大规模复杂优化问题往往面临大状态空间搜索和长延迟奖励的困难，而遗传算法在运行过程中常常出现早熟现象，因此，本文从动态规划角度出发，将动态调度问题转化为分阶段预测的强化学习的状态空间搜索

收稿日期：2010-04-19 **修回日期**：2010-07-28
基金项目：辽宁省自然科学基金项目（20092060）
作者简介：**魏英姿**（1973-），女，辽宁，满族，博士，副教授，硕导，研究方向为智能调度、强化学习和遗传算法等；**谷侃锋**（1973-）男，辽宁，博士，副研究员，硕导，研究方向为机器人学、先进制造装备基础理论与应用。

问题，提出遗传算法和强化学习机制相结合的遗传强化学习（Genetic Reinforcement Learning，GRL）动态调度算法。

1 动态调度问题模型

作业车间调度是最一般的调度类型，属于典型的 NP-hard 问题，也是目前研究最广泛的一类调度问题，是指由 m 个不同的机器加工 n 个有特定加工路线(顺序)的工件，不同工件的工序间没有顺序约束，调度的任务是确定每台机器上的工件加工顺序，使调度目标最优。本文研究的作业车间动态调度问题模型：工件到达系统，在机器前的队列中等待，机器可以有空闲，然后被加工并送到下一台机器，工件到达时间 g_i、加工时间 p_{ij} 依据$[G_1,G_2]$范围内的某种随机分布规律确定。调度目标为最小化生产周期 makespan。

$$makespan = C_{\max} = \max\{C_1,\cdots,C_n\} \tag{1}$$

式(1)中 C_i 为工件 i 的加工完毕时间。

定义 1：调度步　每当系统有资源变化的时刻，就设此时为一个调度点，在某一调度点，完成一道加工工序，本文将其定义为完成一个调度步。

在某一调度点，完成调度步 t 的最大完成时间为 $C^t_{\max}$，如式(2)。

$$C^t_{\max} = \max\{C^f_1,\cdots,C^f_n\} \tag{2}$$

式(2)中 C^f_i 是工件 i 完成 f 道工序的加工完成时间，如式(3)。

$$C^f_i = g_i + \sum_{j=1}^{f} w_{ij} + p_{ij} \tag{3}$$

式(3)中 w_{ij}、p_{ij} 分别为工件 i 进行第 j 道工序的等待时间和加工时间，g_i 为工件 i 的到达时间，f 是已完成工序数目。假设工件需要由 m 道工序才能加工完成，那么当所有工件都完成工序数目 f=m，此时最大完成时间 $C^t_{\max}$ 即为生产周期 makespan。在调度步 t 的最大完成时间 $C^t_{\max}$ 作为状态空间提供的启发式信息，可以在调度中间过程中为学习系统选取动作的好坏提供评价依据。

为考察生产系统设备相关联的性能指标，给出如下概念：

定义 2：机器利用率(Machine Utilization percentage，MU)是指在给定时间内，机器的工作时间与机器可用时间的比率。

定义 3：平均机器利用率的计算，如式(4)：

$$\overline{MU} = \frac{1}{m}\sum_{j=1}^{m} MU_i \tag{4}$$

2 动态调度问题的遗传强化学习解决途径

采用规则表达法构建调度，对于 n 个工件、m 台机器的调度问题，一个完整的调度可表达为规则集 $\{dr_1, dr_2\ldots,dr_i,\ldots,dr_{n*m}\}$。$dr_i$ 表达的含义为：第 i 调度步中的冲突由优先规则 dr_i 来处理。上述规则集串直接作为遗传强化学习运算的染色体，执行交叉、变异、繁殖、强化学习等操作。

Q 学习是强化学习算法中应用最广泛的方法之一。Q 学习方法用 $Q(s,a)$函数表示在状态 s 下执行动作 a 的效果，$Q(s,a)$根据下式更新：

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma \max_{a'\in A} Q(s',a')) \tag{5}$$

其中，α 为学习率，γ 为折扣系数，s'为系统到达的新状态，r 为获得的立即回报，由回报函数确定。强化学习的最终目标是发现最优策略 π，使从状态集到动作集的映射，能达到总的最大折扣回报值。

强化学习在求解动态调度问题时可看作是一种动态规划方法，按照动态规划求解问题的基本思路，本文将作业车间调度决策过程划分阶段，选取恰当的状态变量、决策变量及定义最优指标函数，把复杂调度问题化成一族同类型的子问题，逐个求解，以降低问题的复杂性，解决大状态空间搜索的困难。遗传强化学习系统的任务就是建立状态属性模式与优先规则间的匹配关系，形成一种模式驱动调度器。

2.1 基因模式的划分

模式定理（Schema Theory）是 GA 的理论基础，它说明高适应值、长度短、阶数低的模式在后代中至少以指数增长包含该模式的串的数目。在某种意义上讲，这样的模式成了问题的一部分解，又叫积木块（Building Blocks），这些积木块通过多次遗传操作和并置，形成潜在的适应性较高的染色体。然而，GA 对好的模式的搜索是间接的，它隐含地体现在群体的样本中。我们考虑用 RL 方法引导 GA 对优秀模式的搜索方向，为此，本文将基因空间转化为 RL 算法可直接处理的状态空间，将染色体划分为序列基因模式作为 RL 算法的状态模式(State Pattern)。按照调度步数先后顺序，采用均匀聚类方法划分染色体的状态模式。根据问题规模大小确定划分状态模式的数目，建立 Q 值表，如表 1 所示，其中 S_size 为状态模式数，d 为状态间隔参数 d=m*n/S_size。

状态空间分解的合理性目前还没有定量评价的判据，但一般学者认为，只要状态空间划分的区段足够精细，是可以得到满意的性能的。考虑到调度过程是在离散时间点上进行有限资源的“选择”决策，因此，调度性能在一定范围内对状态空间划分粒度的影响是不敏感的。

表 1　$Q(s,a)$值表

状态	状态确定标准	规则 0	规则 1	规则 2	规则 3	规则 4
0	if(t>=0 && t<d)　s=0;	Q(0,0)	Q(0,1)	Q(0,2)	Q(0,3)	Q(0,4)
1	if(t>=d && t<2*d)　s=1;	Q(1,0)	Q(1,1)	Q(1,2)	Q(1,3)	Q(1,4)
2	if(t>=2*d && t<3*d) s=2;	Q(2,0)	Q(2,1)	Q(2,2)	Q(2,3)	Q(2,4)
3	if(t>=3*d && t<4*d) s=3;	Q(3,0)	Q(3,1)	Q(3,2)	Q(3,3)	Q(3,4)
i	if(t>=i*d && t<(i+1)*d) s=i;	Q(i,0)	Q(i,1)	Q(i,2)	Q(i,3)	Q(i,4)
S_size-1	if(t>=(S_size-1)*d && t< S_size *d) s= S_size-1;	Q(S_size-1,0)	Q(S_size-1,1)	Q(S_size-1,2)	Q(S_size-1,3)	Q(S_size-1,4)

2.2 基于预测启发式的回报函数

强化学习算法建立回报函数往往采用定义等级办法：

$$r=\begin{cases}1 & \text{“好”状态}\\ -1 & \text{“坏”状态}\\ 0 & \text{其他}\end{cases} \tag{6}$$

作为评价调度系统性能好坏的标准，生产周期是在所有加工作业完成之后计算得到的，如果采用(6)式中的回报函数形式进行状态好坏的评价，那么，在调度过程的中间阶段，学习系统受到的立即奖励值只能为空，这种长延迟奖励将使非终点状态的调度步骤出现随机搜索的局面。

以调度目标为依据选取算法的性能预测变量。动态作业车间系统在调度步 t 状态下的最大完成时间为 $C^t_{\max}$，当调度完成 m*n 调度步的所有工序时，系统的最大完成时间即为调度系统的优化目标——生产周期 $C_{\max}$。在某阶段调度步 t，根据性能预测变量——最大完成时间 $C^t_{\max}$ 来评价学习系统当前状态下动作决策的好坏。在 Q 学习中，构建分阶段调度状态模式下的立即回报函数如(7)式，使最大化 Q 函数和最小化目标函数的优化方向一致。

$$r=R-C^t_{\max} \tag{7}$$

R 是一个较大正数常量，式(7)将 $C^t_{\max}$ 最小化问题转化为回报函数的最大化问题。在调度系统的非终止状态下，立即启发式回报函数能够较精确的评价动作的好坏，为学习系统直接及时地提供回报信息，从而引导强化学习算法更快的学会最优策略。

在设立分阶段状态模式的基础上，划分状态空间，不同调度阶段隶属于不同的状态子空间，而在同一阶段状态模式的子空间则相同，依据调度性能的预测变量来评价当前阶段的调度性能具有可比性，能够达到优化中间阶段调度的目的。

2.3 适应度函数

适应度是指导遗传算法搜索进程的重要指标，对种群中的个体进行评价，i 个体的适应度按照式(8)计算。

$$F_i=\frac{\text{makespan}_{\max}-0.9\text{makespan}_{\min}}{\text{makespan}_i-0.9\text{makespan}_{\min}} \tag{8}$$

其中，$\text{makespan}_{\max}$、$\text{makespan}_{\min}$ 分别为种群中最高、最低目标函数值。

2.4 搜索策略

“探索 exploration”和“开采 exploitation”是智能优化算法搜索策略的两个重要方面。在整个算法中，繁殖、交叉、变异、强化学习及其相关参数体现了开采与探索之间的平衡。繁殖的作用是开采搜索空间以便充分利用当前群体的已有信息。交叉和变异是探索搜索空间以便找寻那些可能最优的区域。交叉被认为是贡献于 GA 的全局搜索性能最重要的因素。强化学习根据优化目标学习统计的知识开采搜索区域，以找到局部最优解。强化学习操作在模式状态空间中搜索，也可看作是遗传算法的一个具有学习、自适应能力的交叉算子，它会明确引导 GA 搜索的方向。

为使强化学习的搜索能够根据现有的策略表进行局部寻优，同时又能够有较好的探索能力，实现 Q 学习“探索”与“开采”的平衡，本文采用可变探索率 ε-贪心搜索策略。在进行动作选择的过程中，学习系统有 $1-\varepsilon$ 机会执行趋向目标的贪心动作，如式(9)，有 ε 的机会随机选择动作作用于环境。

$$\pi_t(s,a)=\arg\max_{a'\in A}Q(s,a') \tag{9}$$

按照强化学习试错过程先后顺序，采用分段线性的探索率，如式(10)。

$$\varepsilon=\begin{cases}\varepsilon_0 & l\in[0,0.5Maxstep)\\ \dfrac{0.8Maxstep-l}{0.3Maxstep}\varepsilon_0 & l\in[0.5Maxstep,0.8Maxsteps]\\ 0 & l\in(0.8Maxstep,Maxstep)\end{cases} \tag{10}$$

其中，ε_0 为初始探索率，l 为迭代次数，Maxstep 为学习周期长度。

2.5 算法流程

图 1 为本文遗传强化学习算法的流程图。

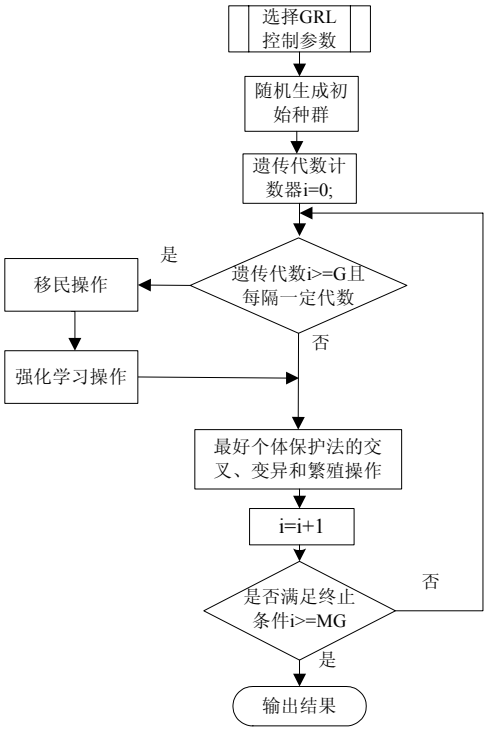

图 1 遗传强化学习算法流程图

其中强化学习算法的步骤如下：

步骤 1 初始化 确定算法控制参数。设定学习周期长度 Maxstep，任意初始化 $Q(s,a)$ 值，选取初始探索率 ε_0 和学习率 α 等。设置计数器 l=0。

步骤 2 随机选择调度规则集串个体。

步骤 3 当 $l<Maxstep$，做

(1) 按照式(10)计算探索率 ε，按照 ε-贪心动作选择策略选取并执行动作 a。

(2) 判断下一决策期的状态模式，计算最大完成时间 $C^t_{\max}$、瞬时启发式回报值 r_t，按公式(5)更新 Q 值。

(3) 更新系统状态，判断是否达到调度的终止状态，若是，进行步骤 3(4)，否则，回转到步骤 3(1)。

(4) 令计数器 l=l+1。

步骤 4 判断 l>=Maxstep 是否成立，若是，学习过程终止，输出结果，否则转步骤 3。

2.6 算法分析

强化学习算法是一种局部寻优方法，但由于 Q 函数统计运算过程是个渐近优化的过程，加上“探索”与“开采”搜索策略的制衡作用，强化学习搜索过程的稳定性难以保证，我们引入最好个体保护策略的遗传算法，运行多个策略表，形成多峰并行搜索局面。采用二级运算方式，一级是运用遗传算法并行搜索，确保优化的正确方向，避免算法陷入局优；二级主要是局部寻优，用强化学习算法得到尽量好的解。进化过程中，每隔一定代数，移民操作向种群加入新的个体，引进新的遗传物质，以满足动态调度环境变化的要求。随机生成的个体的适应性并非必须高于被替换个体的性能，移民操作以牺牲种群的总体性能为代价换取种群的多样性，可以帮助算法预防早熟收敛。算法增加了移民操作也可以弥补种群信息量的不足，允许算法采用较小的初始种群规模，从而减少问题的搜索空间。

3 仿真实验结果与分析

本文设计一个动态调度问题算例，18 个工件在动态作业车间系统的 9 台机器进行加工，工件在每台机器上的加工时间是[2, 9]区间内按均匀分布规律确定的数据，工件到达时间是[2, 9]区间内按正态分布规律确定的数据。

遗传强化学习算法采用的控制参数为：种群规模取为 40，最大遗传代数取为 20，学习率 α=0.85，初始探索率 ε_0=0.05，学习周期长度 Maxsteps=40。运用文献[11]中的 5 种调度规则进行强化学习调度。表 2 列出了各个调度规则独立调度、随机组合调度、简单遗传算法强化学习和遗传强化学习调度方法的运算结果，其中所列数据均是经过 10 次运行得到结果的平均值。遗传强化学习算法得到的调度结果：调度目标最优值为 117，其调度规则集中的 5 种规则所占的比例分别为：5%、22%、60%、6%、7%。

强化学习调度方法的优化性能曲线如图 2 所示，算法的性能曲线波动较大，对于组合优化问题，即使有一次不恰当的“探索”都会严重影响其结果，但目标函数值的总体趋势是下降的，这是采用了“探索”和“开采”平衡的搜索策略的结果。使学习周期后期的 ε 取值为 0，就会使优化过程迅速收敛于一个较好解，表明 Q 学习算法具有较好的趋优性和鲁棒性。图 3 给出简单遗传算法(SGA)和遗传强化学习方法(GRL)性能对比曲线图。根据在 PentiumⅣ CPU2.6GHz、RAM512MHz 计算机 Visual C++ 6.0 环境上对算例的计算，遗传强化学习在 5 秒计算时间内即可得到令人满意的调度结果，完全可以满足动态调度的要求。

表 2 调度性能对比表

调度方法	生产周期	平均机器利用率	耗费机时(秒)
规则 0	169	78.1%	1
规则 1	121	94.4%	1
规则 2	124	93.4%	1
规则 3	179	71.8%	1
规则 4	147	74.4%	1
上述规则的随机组合	145	78.5%	1
遗传算法	126	86.8%	10
强化学习方法	119	91.2%	2
遗传强化学习方法	118	93.5%	5

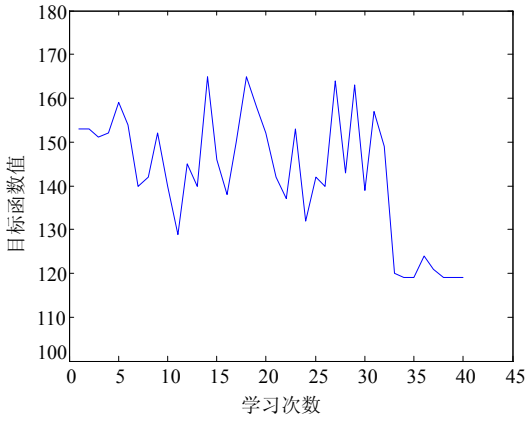

图 2 个体强化学习周期内的目标函数值变化曲线

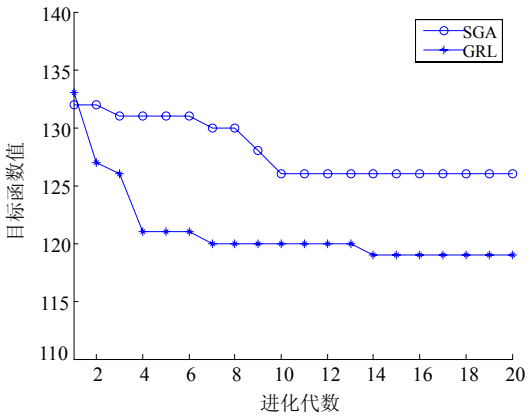

图 3 种群进化的最好解性能对比曲线图

4 结论

本文提出基于遗传强化学习机制的迭代优化调度方案，根据学习系统状态空间内的启发式信息预测调度系统的性能，构建立即启发式回报函数，避免了强化学习系统长延迟奖励的缺陷，采用二级运算模式，结合遗传算法和强化学习的各自特点，以确保算法搜索“探索”和“开采”的平衡。遗传算法对各类问题都有较好的适应能力，Q 学习方法是基于无模型基础上的，运用 Q 学习求解状态转移的搜索优化问题都有其实效，因此，本文工作对未来开发通用性的自适应在线学习控制调度器，具有一定的参考价值，也为动态不确定环境中的优化问题产生自适应的解决方案，提供一个新思路。

（下转第 2820 页）

放大效果来看，第三组滤波初值设置为大失准角，取值 $\boldsymbol{\phi}_0=[30^\circ \quad 10^\circ \quad 70^\circ]^T$，结束时刻 KF 姿态误差为 $\boldsymbol{\phi}=[7' \quad 8' \quad 400']^T$，而 UKF 为 $\boldsymbol{\phi}=[0.6' \quad 0.2' \quad 50']^T$；第四组初始方位失准角取值更大，初值 $\boldsymbol{\phi}_0=[30^\circ \quad 10^\circ \quad 170^\circ]^T$，此时方位误差估计的收敛时间更长，KF 结束时刻姿态误差约 $\boldsymbol{\phi}=[27' \quad -35' \quad 1500']^T$，可见，此时再采用线性模型滤波已经不再合适；采用非线性模型的 UKF 最终姿态误差为 $\boldsymbol{\phi}=[0.3' \quad -0.3' \quad 90']^T$，尽管方位误差仍然比较大，但是水平和方位失准角误差都 KF 精度要高的多，说明文中非线性模型更符合大失准角的真实情况。第三组和第四组 UKF 方位误差收敛至一定程度(2°)后很难进一步收敛，这是由于经过一段时间的滤波收敛和反馈校正，姿态误差逐渐减小，模型渐趋线性，初始时刻大失准角情况下设定的 UKF 滤波参数很难在小失准角情况下达到最优，如不作调整将很难进一步提高滤波精度。考虑到减小计算量的需要，一种值得推荐的方法是先采用大失准角 UKF 加反馈校正将误差收敛至小角度(5°)以内，然后换用 KF 对小失准角线性模型进行估计。

4 结论

在初始姿态信息完全未知的前提下建立起来的系统初始对准误差模型和量测方程将可能呈现强非线性，此时若仍然采用线性模型和 KF 将会因模型与实际情况不符导致收敛时间过长或无法收敛至极限值，从而失去应用价值。文中提出的任意大失准角情况下简化非线性对准方法对于解决某些强非线性条件下，如船舰等因海况恶劣无法采用解析式自对准算法的系统具有重要的实用意义。

参考文献:

[1] 魏春岭, 张洪钺. 捷联惯导系统粗对准方法比较[J]. 航天控制, 2000, 18(3): 16-21.

[2] 陈令刚, 刘建业, 孙永荣, 等. 微小型捷联惯导系统解析式对准方法研究[J]. 航天控制, 2005, 23(4): 9-12.

[3] M J Yu, H W Park, C B Jeon. Equivalent Nonlinear Error Models of Strapdown Inertial Navigation System [C]// AIAA 1997 Guidance, Navigation and Control Conference. USA: AIAA, 1997: 581-587.

[4] S P Dmitriyev, O A Stepanov, S V Shepel. Nonlinear filtering Methods Application in INS Alignment [J]. IEEE Transactions on Aerospace and Electronic Systems (S0018-9251), 1997, 33(1): 260-272.

[5] H S Hong, J G Lee, C G Park. In-flight Alignment of SDINS under Large Initial Heading Error [C]// AIAA Guidance, Navigation, and Control Conference, Montreal Canada, 2001. USA: AIAA, 2001: 1-6.

[6] Y Kubo, S Fujioka, M Nishiyama, S Sugimoto. Nonlinear Filtering Methods for the INS/GPS In-Motion Alignment and Navigation [J]. International Journal of Innovative Computing, Information and Control (S1349-4198), 2006, 2(5): 1137-1151.

[7] E H Shin, N El-Sheimy. An Unscented Kalman Filter for In-motion Alignment of Low-cost IMUs [C]// IEEE PLANS, Monterey, CA, United States, 2004. USA: IEEE, 2004: 273-279.

[8] X Kong, E M Nebot, H D Whyte. Development of Non-linear Psi-angle Model for Large Misalignment Errors and Its Application in INS Alignment and Calibration [C]// IEEE International Conference on Robotics and Automation, Detroit, MI, USA, 1999. USA: IEEE, 1999: 1430-1435.

[9] 严恭敏, 严卫生, 徐德民. 简化 UKF 滤波在 SINS 大失准角初始对准中的应用[J]. 中国惯性技术学报, 2008, 16(3): 253-264.

(上接第 2812 页)

参考文献:

[1] Sun D, Lin L. A dynamic job shop scheduling framework: A backward approach [J]. International Journal of Production Research (S0020-7543), 1994, 32(4): 967-985.

[2] K M Mohanasundaram, K Natarajan, G Viswanathkumar, P Radhakrishnan, C Rajendran. Scheduling rules for dynamic shops that manufacture multi-level jobs [J]. Computers & Industrial Engineering (S0360-8352), 2003, 44(1): 119-131.

[3] 孙容磊, 熊有伦, 杜润生, 等. 规则调度迭代优化[J]. 计算机集成制造系统, 2002, 8(7): 546-550.

[4] Sutton R S, A Barto G. Reinforcement Learning: An Introduction [M]. Cambridge, MA, USA: MIT Press, 1998.

[5] Zhang Wei. Reinforcement Learning for Job-Shop Scheduling [D]. USA: Oregon State University, 1996.

[6] Aydin M E., Oztemel E. Dynamic Job-Shop Scheduling using Reinforcement Learning Agents [J]. Robotics and Autonomous Systems (S0921-8890), 2000, 33(2): 169-178.

[7] Wang Y-C, Usher J M. Application of reinforcement learning for agent-based production scheduling [J]. Engineering Applications of Artificial Intelligence (S0952-197), 2005, 18(1): 73-82.

[8] Pettinger J E, Everson R M. Controlling genetic algorithms with reinforcement learning [C]// Proceedings of the Genetic and Evolutionary Computation Conference. San Francisco, CA, USA: Morgan Kaufmann, 2002: 692-697.

[9] 王本年, 高阳, 陈兆乾, 等. RLGA：一种基于强化学习机制的遗传算法[J]. 电子学报, 2006, 34(5): 856-860.

[10] 潘燕春, 周泓, 冯允成, 等. 同顺序问题的一种遗传强化学习算法[J]. 系统工程理论与实践, 2007, (9): 115-122.

[11] 魏英姿, 曲艳丽, 胡玉兰. 基于合同网协议交互投标的动态调度方法研究[J]. 计算机科学, 2007, 34 (7): 124-127.