Research on Semiconductor Scheduling Based on Multiway Trees Random Forest*

WANG Yu[1,2*]

(1. *Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang* 110016, *China*;
2. *Graduate University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract: In order to address the daily coarse release control problem in semiconductor scheduling, a novel release control algorithm based on multiway trees random forest is proposed. Firstly, taking weekly scheduling as input, the data classification are done by using random forest algorithm. According to the breed names, production capacity, the date of delivery, and classification, the feature information of semiconductor scheduling can be gotten. The results of data classification are achieved by the feature information. Secondly, according to the results of data classification and the time of change machine, the daily production capacity and production breeds can be estimated. Finally, by applying proven, the feasibility of the proposed algorithm can be proved. The experimental results show that proposed algorithm has a availability and superiority, and the time of change machine can be reduced drastically.

Key words: semiconductor scheduling; feature information; data classification; random forest; changing machine
EEACC: 6210C **doi:** 10.3969/j.issn.1005-9490.2014.03.001

基于多叉树随机森林的半导体排产研究*

王 玉[1,2*]

(1. 中国科学院沈阳自动化研究所, 沈阳 110016; 2. 中国科学院大学, 北京 100049)

摘 要: 针对半导体排产控制问题, 提出一种基于多叉树随机森林的数据分类综合排产算法。首先, 以周计划投产品种为输入, 采用多叉树随机森林数据分类方法, 以品种名称、投产量、交货期和所属类别作为半导体排产的特征信息进行数据分类; 其次, 根据分类结果, 以降低"改机"时间为目的, 进而确定日投产品种和数量; 最后, 通过应用研究验证算法的可行性。实验结果表明: 所提出的算法有效降低"改机"时间, 具有一定的有效性和优越性。

关键词: 半导体排产; 特征信息; 数据分类; 随机森林; 改机
中图分类号: TP315 **文献标识码:** A **文章编号:** 1005-9490(2014)03-0381-04

近年来, 排产控制研究是当前生产调度的热点问题, 主要体现在细日投料控制问题[1-2], 即主要对确定投料时刻问题进行研究, 提出了多种投料策略, 如固定时间投料策略[3]、恒定在制品投料策略[4](CONWIP)等。半导体排产控制的研究也主要集中在细日投料控制方面, Ryohei[5]等提出一种基于遗传算法和机器学习结果的的投料控制方法, 利用遗传算法计算合适的控制规则, 从而确定对产品的具体投料时刻; Liu[6]等提出一种新的多规则嵌入的投料控制策略, 考虑 Lot 优先规则、产能分配规则和机器加载规则, 确定投料品种的顺序和工作中心分配原则, 实际上是一种投料派工结合的控制策略; Chua[7]等提出基于多约束有限能力的智能投料控制系统。

半导体排产涉及到一台机器上对应多个生产品种的更换加工, 当不同加工条件的品种更换时, 往往伴随着设备上工夹具, 加工材料等的更换, 这些更换产生一定的时间成本代价, 生产车间称为"改机"现象。"改机"现象的存在导致设备利用率偏低, 生产周期较长。

现以半导体排产为研究背景, 以降低"改机"代价为目标, 针对半导体排产的粗日投料控制问题进行研究, 提出一种基于随机森林[8-13]的综合投料控制策略。该策略以周计划投产品种为输入, 采用基

项目来源: 辽宁省科技攻关项目(2010219010); 辽宁省自然科学基金项目(201102228)
收稿日期: 2013-08-28 修改日期: 2013-09-12

于多叉树随机森林数据分类方法，以半导体生产类型个数作为聚类类别数，依据生产过程中影响“改机”代价的品种属性信息作为分类依据，分类后保证属于同类别的品种“改机”代价相对较小。然后，依据分类结果采用基于品种平均和投产量平均的综合策略，保证生产效率。

1 多叉树随机森林模型

决策树是随机森林算法的基本单元，决策树的构造是由一个随机向量所决定。随机森林算法的本质是组合多个弱分类器(决策树)，使其误差减小的一种分类算法，一般采用二叉决策树作为基本模型，其模型如图1所示。

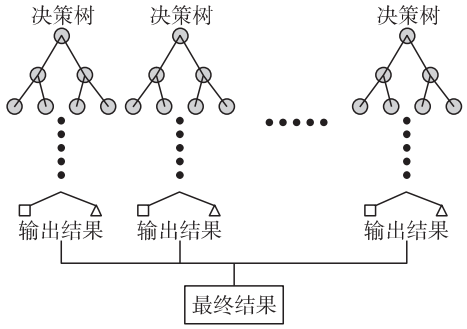


图1 随机森林模型图

由于二叉决策树只能对数据进行2分类，针对多类数据需对二叉决策树进行节点多分叉，形成每个节点多次分叉的过程，从而构造了多叉决策树模型。

随机森林的生成过程分以下4步：

Step 1(Bagging过程)：假设每类训练集中有 N 个样本，有放回地随机抽取 n 个样本，作为一棵决策树的训练样本。

Step 2(分裂属性选择过程)：假设特征向量是 m 维，选取 m_1 维作为子集指定给每个节点，从 m_1 中选择分类效果最佳的一维特征作为接点的分类属性，且保证在随机森林的生长过程中 m_1 保持不变。采用信息熵作为判断节点分裂属性选择的依据，设数据集种类为 m，任意一个数据集的分类概率为 P_i，则信息熵表达式 $H(X)$ 为

$$H(X)=\sum_{i=1}^{m}P_i\log_2(P_i) \tag{1}$$

Step 3(决策树的生长过程)：当每个节点的分类纯度达到期望比例或者生长层数达到给定值时，则停止决策树的生长，保证每个决策树都有最大限度的生长，且没有剪枝情况。

Step 4(生成随机森林过程)：重复Step 1～Step 3，生长出多颗决策树，从而生成森林。

从以上步骤可以看出，随机森林算法的误差更为稳定，克服了单一决策树的不足，体现了多个弱分类器合成强分类器的优势。

2 算法的实现

粗日排产控制的主要任务是根据周计划投产品种和数量信息，进行日投产品种和数量的确定。企业中常用的粗日投料策略主要有两种：基于投产品种平均分配的投料策略和基于品种投产量平均分配的投料策略。基于投产品种平均分配的投料策略，对周计划以品种为单位进行拆分，确定日投产品种；基于投料量平均分配的投料策略，将周计划中的各个品种的投产数量平均分配到每日。该两种方法都未考虑品种更换对生产的影响，导致实际生产“改机代价”较大。本文提出采用基于随机森林的排产控制策略，以降低改机代价为目标，首先对周计划投产品种进行聚类分析，然后在此基础上，采用基于品种平均和投产量平均的综合策略，确定每日投产品种和各个品种的投产数量。

2.1 分类因素提取与确定

(1)分类属性及权重确定。分析半导体生产中，“改机”影响因素有：圆片尺寸、装片胶、框架型号、模具、塑封料、等。不同属性影响品种更换的代价不同，如模具更换需要约4 h的时间，而塑封料更换仅需要15 min左右的时间，利用赋权的方式对各个属性的“改机”代价影响程度进行给定，假设单位改机时间代价为 t，各个因素的权重因子为 ω_1，则各因素的“改机”代价如式(2)所示：

$$C_i=\omega_i t \tag{2}$$

(2)分类类别个数 m 的确定。我们将周计划中投产的不同半导体作为分类依据，根据不同的半导体种类进行类间分类，设半导体的类别个数为 m。

(3)投产量对分类结果的影响。一般来说，车间中各个类型的产能基本均衡。本文假设半导体生产工序的各个类型的产能均衡，考虑品种投产量对品种划分的影响，对每一类别内品种投产量总和进行限定，保证聚类后，各个类别的投产量也基本均衡，则每个类别中的总产量约为：

$$n_c\approx\frac{\sum_i n_i}{k} \tag{3}$$

式中，n_i 表示品种 i 的投产量，n_c 表示类别 c 的总产量。

2.2 基于随机森林的排产模型

我们以品种名称、投产量、交货期和所属类别作为半导体排产的特征信息，设 X_i 是每类半导体的特

征向量,则表示为

$$X_i=(a_i^1,a_i^2,a_i^3,a_i^4) \tag{4}$$

其中,a_i^1 表示品种名称,a_i^2 表示投产量,a_i^3 表示交货期,a_i^4 表示所属类别。

我们将半导体的生产类别个数 m 作为分类数量,将不同种类的半导体特征向量作为训练随机森林的训练集,具体步骤如下:

Step 1: 根据半导体生产种类确定分类数,进而确定随机森林的分叉数 m,对每类半导体选取 n_x 作为训练集,总共 mn_x 个训练样本。

Step 2: 将不同的半导体训练样本分别标记模式类别($1\sim m$)。

Step 3: 从训练样本中随机抽取 $0.7mn_x$ 个训练样本,按照第 2 节所述构建半导体分类决策树。

Step 4: 重复 step3,构建多颗决策树,生成随机森林。

Step 5: 将待分类的半导体排产数据通过训练完成的随机森林进行完全分类,确定每个半导体排产数据的模式类别。

Step 6: 对分类后各个类别的投产品种分别进行排序,交货期越早,排序越靠前,需进行优先生产。

Step 7: 针对交货期不紧张的生产订单,则根据半导体数据分类结果进行合理的投产。

3　实验评估

3.1　分类准确性实验

实验选取 7 种不同型号的半导体进行研究,投产信息如表 1 所示。

表 1　投产品种信息

品种	投产量	交货日期	圆片尺寸	装片胶	框架型号	模具	塑封料
MC20976P-1	30000	24/03/2013	12	UN-6885	SDIP32A	TA7698	EME6300H
MC20976P-2	15000	25/03/2013	8	UN-6888-I	SDIP32A	TA7698	EME6300H
MC20976P-3	30000	26/03/2013	6	UN-6888-I	SDIP32A	TA7698	7300HX
TDA6782M-1	20000	25/03/2013	6	UN-6885	JH-100630	T0263E	EME6300H
TDA6782M-2	45000	29/03/2013	8	UN-6888-I	JH-100630	T0263E	7300HX
LS028212M	28000	26/03/2013	12	UN-6885	MHT-A194	T0263E	7300HX
D396650-e3	21000	27/03/2013	8	UN-6888-I	JH-100630	TA7698	EME6300H

按照 2.2 节所述,将实验的每种型号的半导体信息转变为特征向量,将 7 种型号的半导体特征向量作为分类依据,从而完成随机森林的构建。设每类半导体的投产量为 n_z^i,算法准确分类数为 n_t^i,平均分类准确率为 p_t,则将其定义为

$$p_t=\frac{1}{m}\sum_{i=1}^{m}\frac{n_t^i}{n_z^i}\times 100\% \tag{5}$$

按照式(5)定义,计算分类准确率,所得结果如表 2 所示。

表 2　分类准确率

品种	分类准确率%
MC20976P-1	98.6
MC20976P-2	98.1
MC20976P-3	98.4
TDA6782M-1	98.5
TDA6782M-2	97.9
LS028212M	98.4
D396650-e3	98.8

表 2 表明: 本文算法对不同品种的半导体的分类较为准确,其平均分类准确率高达 98.4%,从而验证了算法在半导体排产中的数据分类可行性。

3.2　"改机"时间比较实验

按照 2.2 节所述的算法流程,对待排产的半导体进行合理安排,为了方便比较不同类型的半导体粗日投料对生产的影响,假设生产车间只有一道工序,每种类型的产能各有一台机器,改机单位时间为 10 min,分布对基于品种平均分配的投料策略、基于投产量平均分配的投料策略和本文提出的基于品种分类的综合投料策略控制下的生产过程进行比较,其实验结果如表 3 所示。

表 3　各种投料策略比较

分配方式	改机频率	改机时间/min
基于品种平均分配	6	445
基于投产量平均分配	42	3115
本文算法	4	152

由实验结果可得: 本文所提出的基于随机森林对半导体品种进行分类算法能够很好的减少改机时间代价,从而缩短生产周期,提高生产速率。为了方

便计算,实验给出的规模较小,一般在半导体实际生产中,周投产品种达上千种,本文算法在规模庞大的实际应用中更能发挥其优越性,假设背景企业中,采用该算法后,生产周期缩短了约27 h,即改机时间代价降低了约27 h,大大提高了生成效率。

4 结束语

本文针对半导体排产问题进行深入研究分析,提出一种基于多叉树随机森林的半导体排产算法,该算法首先将半导体的品种名称、投产量、交货期和所属类别转化为特征信息,将其输入到构建完成的随机森林,从而进行数据的完全分类,利用得到的数据分类法则进行排产的评估,确定日投产的半导体种类和数量。

实验结果表明:该算法能够有效地降低半导体排产过程中的改机时间代价,从而提高设备的利用率,大大缩短生产周期,提高生产效益。今后的工作是对现有的随机森林纯度分类依据进行深入研究,优化属性选择原则,从而提高算法对各类数据的适应能力。本文创新点在于首次将随机森林算法应用到工业排产研究中,开拓了机器学习的应用领域。

参考文献:

[1] 李友,江志斌,李娜,等. 晶圆制造系统投料策略综述[J]. 工业工程与管理,2011,16(6):108-114.

[2] 吴启迪,乔非,李莉,等. 半导体制造系统调度[M]. 北京:电子工业出版社,2006:102-104.

[3] Aurand S S,Miller P J. The Operating Curve:A Method to Measure and Benchmark Manufacturing Line Productivity[C]//Proceedings of IEEE Conference on SEMI Advanced Semiconductor Manufacturing,1997:391-397.

[4] Glassey C R,Mauricio G C. Closed-Loop Job Release Control for VLSI Circuit Manufacturing[J]. IEEE Transactions on Semiconductor Manufacturing,1988,1(1):36-46.

[5] Ryohei T,Nobutada F,Kanji U. Lot Release Control Using Genetics Based Machine Learning in a Semiconductor Manufacturing System[J]. Intelligent Autonomous Systems,2006(9):497-506.

[6] Liu W,Chua T J,Cai T X,et al. Practical Lot Release Methodology for Semiconductor Back-End Manufacturing[J]. Production Planning and Control,2005,16(3):297-308.

[7] Chua T J,Liu M W,Wang F Y,et al. An Intelligent Multi-Constraint Finite Capacity-Based Lot Release System for Semiconductor Backend Assembly Environment[J]. Robotics and Computer-Integrated Manufacturing,2007,23(3):326-338.

[8] Ganesan P,Krishna K K,Chou K C,et al. RSARF:Prediction of Residue Solvent Accessibility from Protein Sequence Using Random Forest Method[J]. Protein and Peptide Letters,2012,19(1):50-56.

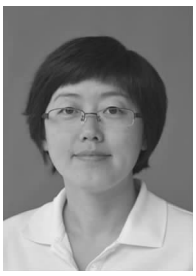
[9] Boulesteix A L,Bender A,Bermejo J L,et al. Random Forest Gini Importance Favours SNPs with Large Minor Allele Frequency:Impact,Sources and Recommendations[J]. Briefings in Bioinformatics,2012,13(3):292-304.

[10] Joshi N,George B,Vanajakshi L,et al. Application of Random Forest Algorithm to Classify Vehicles Detected by a Multiple Inductive Loop System[C]//Proceedings of IEEE Conference on Intelligent Transportation Systems,2012,1491-495.

[11] Adam E M,Mutanga O,Rugege D,et al. Discriminating the Papyrus Vegetation(Cyperus Papyrus L.) and Its Co-Existent Species Using Random Forest and Hyperspectral Data Resampled to HYMAP[J]. International Journal of Remote Sensing,2012,33(2):552-569.

[12] Boucekine M,Loundou A,Baumstarck K,et al. Using the Random Forest Method to Detect a Response Shift in the Quality of Life of Multiple Sclerosis Patients:A Cohort Study[J]. BMC Medical Research Methodology,2013,13(1):1-8.

[13] 蔡敏. 基于多特征组合优化的汉语数字语音识别研究[J]. 电子器件,2013,36(2):282-284.



王 玉(1977-),女,高级工程师,硕士研究生,主要研究领域生产调度、制造过程建模与仿真等的研究等,wangyu@sia.cn。