

引用格式:刘元修,史峥,张培勇.一种基于随机森林的可寻址 WAT 良率诊断方法[J].微电子学与计算机,2019,36(9):94-98. [LIU Y X,SHI Z,ZHANG P Y. A yield diagnosis method based on random forest for addressable WAT[J]. Microelectronics & Computer,2019,36(9):94-98.]

一种基于随机森林的可寻址 WAT 良率诊断方法

刘元修,史 峥,张培勇

(浙江大学 超大规模集成电路设计研究所,浙江 杭州 310027)

摘 要: WAT(Wafer Acceptance Test,晶圆允收测试)是晶圆在完成制造后必须通过的一项电学测试.本文提出了一种基于随机森林的良率诊断方法,使用随机森林算法对来自可寻址 WAT 的测试数据建立分类模型,并从中提取关键规则.提取出的规则集合可以帮助分析人员快速、准确地对生产中导致低良率的根本原因进行定位、分析,对于良率提升有着重要意义.在真实可寻址 WAT 测试数据集上应用本文提出的方法,得到的规则集合在分类性能上较基于决策树的方法有着更好的表现.

关键词: 晶圆允收测试;良率提升;随机森林;规则提取

中图分类号: TP391

文献标识码: A

文章编号: 1000—7180(2019)09—0094—05

A yield diagnosis method based on random forest for addressable WAT

LIU Yuan-xiu,SHI Zheng,ZHANG Pei-yong

(Institute of VLSI Design, Zhejiang University, Hangzhou 310027, China)

Abstract: WAT(Wafer Acceptance Test) is an electrical test that a wafer must pass upon completion of fabrication. An addressable WAT yield diagnosis method is proposed in this paper, which uses the random forest algorithm to establish a classification model on test data and extracts key rules. The extracted rules can help analysts quickly and accurately locate the root cause of low yield in production, which plays a key role in yield enhancement. The proposed method has been experimented on dataset from addressable WAT, and the obtained rule set has better classification performance than the rule set obtained by method based on decision tree.

Key words: wafer acceptance test; yield enhancement; random forest; rules extraction

1 引言

半导体行业中,良率提升意味着生产成本降低与利润率提高,是半导体企业保持核心竞争力的关键因素之一.半导体生产流程繁复,并伴随有多项测试,每一项都会产生大量测试数据,从这些数据中快速、高效地挖掘出有价值的信息,是提升良率的重要手段.

WAT(Wafer Acceptance Test,晶圆允收测

试),又被称为电特性测试(E-Test),通常在晶圆完成所有加工步骤后进行,用于测试半导体器件的电特性;而电特性的波动是良率损失^[1]的主要原因,因此 WAT 在半导体制造良率控制中扮演着重要角色.

WAT 通过对晶圆上半导体器件的电特性进行测试来评估晶圆在生产中是否出现缺陷,但因为目标器件的电特性无法直接测量,为了达到测试目的,需要在晶圆上另外放置测试结构作为替

代.按照所希望监控的目标器件进行设计的测试结构一般会放置在晶圆的划片沟槽中.在晶圆完成制造后,测试机通过探针与晶圆上放置的测试结构进行电气连接,完成各项电特性测试.由于测试结构与目标器件近似,因此可以使用测试结构的电特性测试结果来评估目标器件的生产情况,以此达到监控工艺的目的.

传统测试芯片中,每个测试结构都需要独立的引脚与之相连,而引脚由于需要和外界进行电气连接,必然会占用较大的面积,从而使得传统测试芯片的面积利用率非常低,进而造成 WAT 测试覆盖率的低下.目前半导体生产工艺愈发复杂,需要在晶圆上部署更多的测试结构来提升对目标器件的覆盖率,传统测试芯片越来越难以满足需求.

文献[2]提出了一种应用于 WAT 的可寻址测试芯片,如图 1 所示,测试结构被组织为阵列,通过行/列地址译码电路进行寻址,信号选择电路用于选择被激活的测试结构,通入测试信号进行相应测试.译码电路的引入,使相同数量的引脚能完成寻址的测试结构数量是原先的成百上千倍,大幅提高了测试芯片的面积利用率,进而提高了 WAT 的测试覆盖率与测试效率.

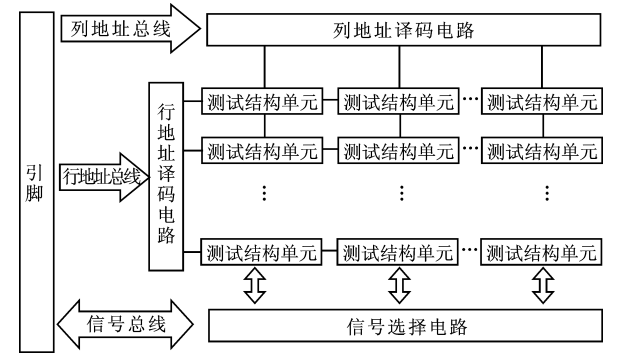


图 1 可寻址测试芯片

应用了可寻址测试芯片的 WAT 测试数据包含了测试结构的设计参数、部署位置等信息.当出现低良率问题时,通过对测试数据的分析找到与低良率相关的测试结构特征,这些信息可以帮助分析人员快速定位生产缺陷,进而找到导致低良率的根本原因.随着制造工艺日益复杂,测试数据量也急剧提升,如何快速、高效地在海量测试数据中进行良率诊断,成为了可寻址 WAT 应用场景中的关键问题.

文献[3]构建了一种集成良率管理系统,可以对包括 WAT 数据在内的多种集成电路生产测试数据

进行分析,通过趋势图、相关性分析等方法对低良率成因进行探索.但随着数据量的激增,简单的统计学方法在效率上已经难以满足需求.相应的,分析人员开始借助机器学习方法对集成电路生产测试数据进行数据挖掘,其中规则提取方法因为具有可理解性上的优势,在良率诊断场景中得到了广泛应用.文献[4-6]使用决策树算法对晶圆生产线数据建立分类模型并导出决策规则,分析人员根据这些规则定位到在晶圆制造过程中导致缺陷的阶段或工具,进而帮助完成良率诊断.

决策树算法具有很好的可解释性,其根节点到叶节点的每一条路径都可以表示为一条分类规则,因此成为规则提取最常用的模型之一.随着半导体工艺发展,其缺陷模式也表现地愈发复杂,在这种场景下,决策树算法在拟合能力上暴露出缺陷,提取出的规则难以在查全率与查准率上实现很好的平衡.针对该情况,本文提出了一种基于随机森林的可寻址 WAT 良率诊断方法,使用随机森林算法对测试数据建立分类模型,在训练完成的随机森林模型中选择部分决策树并提取其中的关键决策规则.相比传统的基于决策树的低良率诊断方法,该方法在导出规则集合的查准率与查全率上拥有更好的表现.

2 可寻址 WAT 良率诊断

应用了可寻址测试芯片的 WAT 所获得的典型测试数据如表 1 所示,每一条记录代表某一特定测试结构的某一次电特性测试是否合格,这里的测试结构参数主要包含两部分,一部分是测试结构的设计参数,例如晶体管的 W/L 尺寸等,另一部分是用于定位该测试结构的一些信息,例如该测试结构所属的晶圆编号、在晶圆上的部署位置等.测试项代表该条测试数据对应的测试类型,常见的有晶体管的 Ioff、Idsat 等.如果电特性测试得到的数值超过该工艺下预设的阈值,则认为该次测试未能通过,这里将不合格的记录的标签记为 1.

表 1 可寻址晶圆允收测试典型测试数据集

编号	测试结构参数	测试项	是否合格
0	$v_{0,0} \sim v_{0,m-1}$	Ioff	0
1	$v_{1,0} \sim v_{1,m-1}$	Ioff	1
...	...	...	...
$n-1$	$v_{n-1,0} \sim v_{n-1,m-1}$	Idsat	0

在 WAT 完成后,分析人员会在测试数据上进

行良率分析,某一部分测试数据合格率过低,会被认为是出现了低良率情况.分析人员通过测试数据去寻找与低良率相关的信息,例如失败的记录是否集中于晶圆的某个位置,或者低良率涉及到的测试结构设计参数之间是否具有一致的特征,这些信息可以帮助分析人员快速定位晶圆的生产缺陷,进而找到导致低良率的根本原因.

3 基于随机森林的规则提取方法

3.1 随机森林

随机森林^[7]是一种经典的集成学习算法,由 LeoBreiman 于 2001 年提出.随机森林算法使用决策树作为基学习器,采用 Bagging 方法^[8]进行集成.在此基础上,随机森林进一步在决策树的训练过程中引入了特征的随机选择.

随机森林对测试样本进行分类时,每一棵决策树均会产生一个对应的预测类别,这些预测结果会通过一次简单的投票,来确定该测试样本最终的预测类别.随机森林算法较为简单且易于并行化实现,能够处理大规模数据,同时拥有优异的分类性能.

3.2 混淆矩阵

在分类场景下,使用混淆矩阵作为评估分类器性能的统计量.对于每一个样本,其真实类别与预测类别共有四种组合,分别为真正例(TP)、假正例(FP)、真负例(TN)与假负例(FN),这四类结果组合得到如表 2 所示的混淆矩阵.

表 2 混淆矩阵

	预测正例	预测负例
实际正例	TP	FN
实际负例	FP	TN

以混淆矩阵为基础,可以导出常用的几种评估指标:

accuracy=(TP+TN)/(TP+TN+FP+FN)

TPrate=TP/(TP+FN)

precision=TP/(TP+FP)

Fmeasure=2 * TP/(2 * TP+FP+FN)

FPrate=FP/(FP+TN)

3.3 规则提取

这里所述的规则提取,指的是将分类模型转换为一组可以被理解的分​​类规则集合 R:

R = (r₁ ∨ r₂ ∨ ⋯ ∨ r_k)

这里的 r_i 代表一条形如“if⋯then⋯”形式的分类规则,表现为:

r_i:(condition_i)→y_i

箭头的左边称为规则前件,描述了规则成立的前提,是一组条件的集合.箭头右边称为规则后件,描述了规则前件成立时,所指向的预测类别 y_i.

随机森林算法在大部分应用场合中均被作为一种黑盒模型来使用,其可理解性并未得到太多的关注.事实上,随机森林是多棵决策树的集成,而决策树很容易转化为所需的规则集合,有着良好的可理解性,因此从随机森林模型中提取规则是可行的.

随机森林往往集成了几十上百棵决策树,直接将这​​些决策树转换为规则集合,得到的规则的数量会十分庞大,失去了可理解性.本文提出的随机森林规则提取方法,提取得到的规则集合所包含的规则数量大幅降低,同时维持了原有随机森林的分类性能,具体的算法如下:

输入:训练完成的随机森林模型

输出:提取得到的规则集合

Step1:从训练完成的随机森林模型中提取所有的决策树,记为 TreeSet_{original}.

Step2:将 TreeSet_{original} 中的每一棵树都看成独立的决策树分类器,在训练集上分别评估它们的分类性能,评估标准如下:

TreeScore=Fmeasure-FPrate

根据打分对所有决策树进行排序,选取分数最高的 k 棵进行保留,得到 TreeSet_{selected}.

Step3:从 TreeSet_{selected} 中提取所有决策树的决策规则,保留其中规则后件为正例的部分,得到初始决策规则集合 RuleSet_{original}.

Step4:更新 RuleSet_{original} 中的规则在训练集上的表现,从中选取表现最好的一条,加入到 RuleSet_{extracted} 中,并从 RuleSet_{original} 中移除.

为了评估规则发现未知正例样本的能力,在 TP 的基础上,增加了一种新的统计量 FirstTP,其含义是:样本被当前规则 r 分类为正例,且该样本未被 RuleSet_{extracted} 内的任何一条已存在规则所覆盖.

单条规则评价指标如下:

RuleScore=FirstTP-FP+a * TP

这里 a 是一个小于 1 的正常数,用于控制 TP 的权重.

Step5:使用 Fmeasure 作为评估指标,评估 RuleSet_{extracted} 在训练集上的分类性能.

重复 Step4 至 Step5,直到满足终止条件,或达到最大迭代次数,得到的结果即最终提取出的规则

集合RuleSet_{final}. 这里的终止条件指的是在连续几次迭代中,规则集合的 Fmeasure 未能提升.

4 实验验证

4.1 问题定义

半导体制造流程复杂,晶圆在生产过程中出现的任何缺陷,都可能会导致良率降低. 为了评估晶圆的生产情况,应用可寻址 WAT 实现对晶圆上半导体器件电特性的测量. 从得到的测试数据中挖掘出与低良率相关的信息,能帮助分析人员快速定位晶圆生产缺陷,完成良率诊断.

在海量测试数据中完成上述的良率诊断,对于良率提升有着重要意义. 本文使用随机森林算法在可寻址 WAT 测试数据上建立分类模型并提取出一组规则,将测试结构信息与良率之间的关联转换为易于理解的规则的形式,以此帮助分析人员实现高效的良率诊断.

4.2 数据描述

本文所使用的数据集为来源于真实生产测试流程的可寻址 WAT 数据. 这部分数据属于同一片晶圆,用于监控晶圆上晶体管的 IoffD 测试项. 该数据集包含 191 052 条测试记录,类别比例(合格:不合格)约为 189:1,特征数量为 56.

4.3 评价指标

accuracy 是分类器性能最常用的评价指标之一,但是该指标在评价不平衡数据集时有很大的局限性. WAT 数据集是典型的不平衡数据集,正例(不合格)样本要远少于负例,为了能更好的对规则集合的分类性能做出评价,主要关注 TP rate 与 precision,TP rate 又称为查全率,表示所有正例中被正确分类的比例,precision 又被称为查准率,表示被分类为正例的所有样本中,实际为正例的比例. F measure 是 TP rate 与 precision 的调和平均数,是对查全率与查准率综合考察的一种指标. FP rate 表示所有实际负例中,被错误分类为正例的比例,该值越小,表示误报率越低.

4.4 实验验证与结果分析

对测试数据集进行随机切分,其中 80%数据作为训练集,剩余的 20%数据作为测试集.

首先使用随机森林算法在测试数据集上建立分类模型,在本文的场景中,训练得到的随机森林拥有 100 棵决策树,每棵树的最大深度为 6. 选择其中在训练集上表现最好的 10%,也就是 10 棵决策树进行保留,组成新的子随机森林,将这些决策树转换为

决策规则,并仅保留其中规则后件为正例的规则,一共得到由 284 条决策规则组成的原始规则集合. 使用本文提出的规则提取方法,从中提取出共 9 条规则,作为最终的关键规则集合.

图 2 是训练得到的原始随机森林与最终提取出的规则集合在测试集上的分类性能表现. 可以看到,提取出的规则集合在测试集上的表现,在查准率上出现了小幅度的下滑,其余指标基本与原始的随机森林保持一致,这说明该规则集合较好地保持了原随机森林的分类性能.

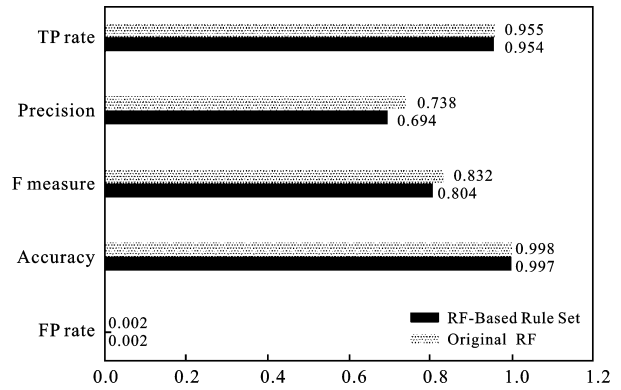


图 2 原始随机森林与提取出的规则集合分类性能对比

图 3 是本文提出的基于随机森林的规则提取方法与传统的基于决策树的规则提取方法之间的比较,可以看到,凭借随机森林本身优异的分类性能,基于随机森林的规则提取方法所提取出的规则集合表现出了明显的优势,查全率与 accuracy 两项的表现基本一致,而在查准率与 FP rate 两项上则是大幅领先基于决策树的规则提取方法. 这说明本文提出的规则提取方法所提取出的规则集合能够更准确地覆盖到低良率样本,且有效降低了误报率.

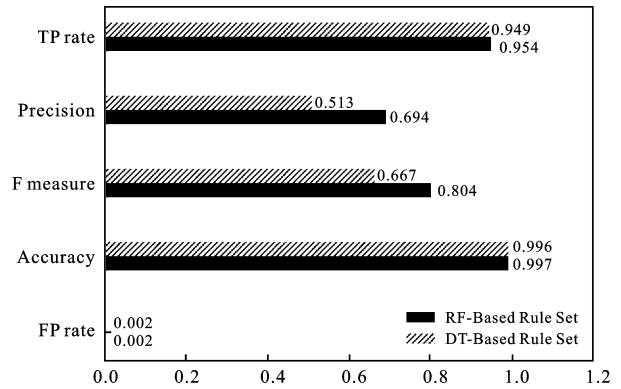


图 3 本文方法所提取的规则集合与基于决策树提取的规则集合分类性能对比

5 结束语

随着半导体工艺的发展,WAT 在半导体生产良率控制中扮演的角色也愈发重要. 可寻址测试芯片的应用大幅提升了 WAT 的测试效率,满足了日益复杂的测试需求. 但与此同时,如何从海量测试数据中高效地进行良率诊断,成为了亟待解决的问题.

本文提出的基于随机森林的可寻址 WAT 良率诊断方法,使用随机森林算法在测试数据上建立分类模型,并从中提取关键规则,得到的规则集合可以帮助分析人员快速定位缺陷,进而完成良率诊断. 在真实可寻址 WAT 数据集上进行验证,结果表明本文提出的方法,相较于基于决策树的规则提取方法,在提取出的规则集合的查准率与查全率上均有着更好的表现.

参考文献:

[1] CHIEN C F, CHEN Y J, WU J Z. Big data analytics for modeling WAT parameter variation induced by process tool in semiconductor manufacturing and empirical study [C] // Winter Simulation Conference (WSC), 2016. IEEE, Hyderabad, India, 2016: 2512-2522.

[2] ZHANG B, PAN W, ZHENG Y, et al. A fully automated large-scale addressable test chip design with high reliability[J]. european conference on circuit theory and design, 2011: 61-64. DOI: 10.1109/ECCTD.2011.6043609.

[3] LEE F. Advanced yield enhancement: integrated yield analysis[C] // Advanced Semiconductor Manufacturing Conference and Workshop, 1997. Cambridge, MA, USA, IEEE/SEMI. IEEE, 1997: 67-75.

[4] CHIEN C, HSU C, CHEN P, et al. Semiconductor fault detection and classification for yield enhancement and manufacturing intelligence[J]. Flexible Services and Manufacturing Journal, 2013, 25(3): 367-388.

[5] CHIEN C, WANG W, CHENG J, et al. Data mining for yield enhancement in semiconductor manufacturing and an empirical study[J]. Expert Systems With Applications, 2007, 33(1): 192-198.

[6] WEISS S M, BASEMAN R J, TIPU F, et al. Rule-based data mining for yield improvement in semiconductor manufacturing[J]. Applied Intelligence, 2010, 33(3): 318-329.

[7] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1):5-32.

[8] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24(2):123-140.

作者简介:

刘元修 男,(1994-),硕士研究生. 研究方向为集成电路测试数据分析.

史 峰(通讯作者) 男,(1967-),副教授,硕士生导师. 研究方向为集成电路 DFM 及 EDA.

E-mail: shiz@vlsi. zju. edu. cn.

张培勇 男,(1977-),副教授. 研究方向为集成电路设计.

(上接第 93 页)

[8] GUOYE CAI, ZHAO XUEQING, LIU ZHENXING, et al. A modified T/2 fractionally spaced coordinate transformation blind equalization algorithm[J]. International Journal of Communications Network & System Sciences, 2010, 3(2): 183-189.

[9] 饶伟, 郭业才, 汪胜前, 等. 基于坐标变换的常数模盲均衡新算法[J]. 电子学报, 2011, 39(1): 7-12.

作者简介:

魏海文 男,(1993-),硕士研究生. 研究方向为自适应盲均衡技术. E-mail: CJwei haiwen@163. com.

郭业才 男,(1962-),博士,教授. 研究方向为通信信号处理、自适应盲均衡技术.