控制与决策
Control and Decision

基于优化DBSCAN聚类算法的晶圆图预处理

陈寿宏， 易木兰， 张雨璇， 尚玉玲， 杨平

引用本文:
陈寿宏, 易木兰, 张雨璇, 等. 基于优化DBSCAN聚类算法的晶圆图预处理[J]. 控制与决策, 2021, 36(11): 2713−2721.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2020.0738

基于优化DBSCAN聚类算法的晶圆图预处理

陈寿宏[1,2]，易木兰[2]，张雨璇[2]，尚玉玲[2]，杨 平[1†]

(1. 江苏大学 机械工程学院，江苏 镇江 212013；2. 桂林电子科技大学 电子工程与自动化学院，广西 桂林 541004)

摘 要: 晶圆图是由半导体生产过程中对晶圆进行可测试性检测而得到的,通过对晶圆图进行分类可以为生产过程中出现的问题提供依据,从而解决问题,降低生产成本. 在对晶圆图进行分类之前,最重要的是特征提取,晶圆图除了本身拥有一定的空间图案以外,还存在着很多的噪声,影响着特征提取的过程. 传统的DBSCAN算法用于滤波,需要人为确定两个参数,最小邻域Eps和最小点数MinPts,参数的选择直接影响了聚类的准确性. 为此,提出一种基于优化DBSCAN聚类算法的滤波方式,自动确定DBSCAN的参数,以解决传统的手动设定参数的弊端. 该算法基于参数自动寻优策略,选取DBSCAN 聚类后簇内密度参数和簇间密度参数的综合指标来评定最优参数. 实验结果表明,该算法能自动并合理地选择较好的参数,具有很好的聚类效果,对后续的特征提取及分类也具有很大的帮助.

关键词: 晶圆图；DBSCAN；自动；聚类；密度；滤波

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.0738 **开放科学(资源服务)标识码(OSID):**

引用格式: 陈寿宏,易木兰,张雨璇,等. 基于优化DBSCAN聚类算法的晶圆图预处理[J]. 控制与决策, 2021, 36(11): 2713-2721.

Wafer map preprocessing based on optimized DBSCAN clustering algorithm

CHEN Shou-hong[1,2], *YI Mu-lan*[2], *ZHANG Yu-xuan*[2], *SHANG Yu-ling*[2], *YANG Ping*[1†]

(1. School of Mechanical Engineering，Jiangsu University，Zhenjiang 212013，China；2. School of Electronic Engineering and Automation, Guilin University of Electronic Technology，Guilin 541004，China)

Abstract: The wafer map is obtained by testing each die in the wafer during semiconductor production for defects and marking the defective die. The classification of the wafer map can provide a basis information for problems that occur in the production process, thereby solving the problems and reducing the cost. Before classifying the wafer map, the most important thing is feature extraction. In addition to having a certain spatial pattern, the wafer map also has a lot of noise, which affects the process of feature extraction. When the traditional density-based spatial clustering of applications with noise (DBSCAN) algorithm is used for filtering, it needs to manually determine the value of Eps and MinPts parameters, and the selection of the parameters directly affects the accuracy of the clustering. Therefore, this paper proposes a filtering method based on the optimized DBSCAN clustering algorithm to automatically determine the parameters of the DBSCAN, which can solve the traditional drawbacks of manually parameters setting. This method selects a comprehensive index of cluster intra-cluster density and inter-cluster density to evaluate the optimal parameters. The experimental results show that the proposed algorithm can automatically and reasonably select better parameters and has a good clustering effect, which is also very helpful for subsequent feature extraction and classification.

Keywords: wafer map；DBSCAN；automatic；clustering；density；filter

0 引 言

半导体产业发展迅速,在半导体制造业中不仅需要使用大量昂贵且精密的设备,还需要高标准的作业环境,因此,需要大量的资金投入. 正是由于半导体具有这样高成本、低生命周期的特性,如何提升良率,减少不良产品,进而使得企业获利是一项特别重要的课题[1].

随着半导体制造技术的不断进步,市场需求的提

收稿日期: 2020-06-11；**修回日期:** 2020-09-08.

基金项目: 国家自然科学基金项目(61661013)；广西自然科学基金项目(2018GXNSFAA281327)；桂林电子科技大学研究生教育创新计划项目(2019YCXS085, 2020YCXS095).

†通讯作者. E-mail: yangping1964@163.com.

高,生产环境、生产设备和从业人员差异性的增加,整个半导体产业划分为设计、制造、封装和测试4个阶段[2]. 晶圆作为集成电路中不可或缺的一部分[3-4],制造过程复杂、漫长且昂贵[5]. 它的生产过程需要经过氧化、光刻、蚀刻、离子注入和金属化等上百个步骤,并需要不断反复,任意一道工序出现问题,都会造成缺陷的产生. 例如,涂抹光刻胶涂抹不均匀,紫外线曝光不均匀,晶圆从一个步骤到下一个步骤之间被刮坏等等[6-7],都能造成晶圆最后不可用. 因此,找到产生晶圆缺陷的原因,提高生产良率,降低生产成本是每一个生产厂家都很注重的问题.

晶圆在封装测试之前,会对晶圆上的每个晶粒进行测试,对存在缺陷的晶粒进行标记,对晶粒进行有无缺陷标记以后就会在整个晶圆上形成一定的空间图案,称之为晶圆图. 根据晶圆图的特点可以将其分为8种缺陷类型,分别是Center、Donut、Edge-loc、Edge-ring、Local、Near-full、Random、Scratch和一种正常类型None. 每一种缺陷类型都可以为生产过程中存在的问题提供有用信息,使得工程师及时发现问题,增加生产良率,降低生产成本[8-9]. 例如,Scratch类型是由机械处理造成的,Edge-ring类型是由蚀刻问题引起的,Center类型一般是在薄膜的沉积步骤中造成的[10]. 目前主要的研究方法是通过对晶圆图准确地分类,从而快速地判断产生缺陷的原因. 常见的晶圆图分类一般需要经过以下几个步骤:预处理(包括滤波等)、特征提取、分类. 预处理作为晶圆图处理的第1步,对于后续的晶圆图分类的准确性也起着至关重要的作用.

晶圆图中通常伴有很多的噪声,这些噪声一般是生产环境中的灰尘及颗粒随机产生的,可能在晶圆的任意位置,其分布是随机的、无规律的,如图1(a)和图1(c)中存在的噪声,这些噪声的存在模糊了晶圆图的主要特征,淹没了晶圆图的缺陷模式,给特征提取造成困难,影响分类准确率.

为了去除这些噪声点,通常采用滤波处理,常见的滤波方法有中值滤波和均值滤波. 如图1所示,在图1(a)中是原始的Center类型晶圆图,图1(b)中是Center类型均值滤波处理后的晶圆图,可以看出晶圆图滤波处理以后,有效地凸显了晶圆图的主要特征. 因此,为了提高晶圆图缺陷模式的分类准确率,在提取特征之前去除晶圆图中的噪声点是很有必要的. 然而,在图1(c)中对Scratch类型的晶圆图采用均值滤波处理后,破坏了Scratch类型的主要特征,不利于特征提取,影响分类结果. 因此,均值滤波存在一定的局限性. Kim[11]等采用了连接路径滤波的方法,Wang等[12]使用了空间滤波器,由文献[5]可知,当参数设置不当时,使用空间滤波器会去除部分可用于帮助确定缺陷类型的重要信息.

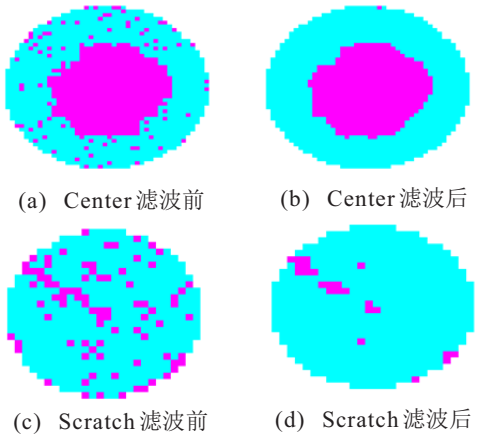


图1 Center、Scratch类型晶圆图均值滤波前后对比

造成上述结果的主要原因是在实际中各批次的晶圆图大小不相同,缺陷分布不均匀,形状也不规则,所以不能对所有的晶圆图采用相同的参数滤波. 因此,选择适当的预处理方法以及设置合适的参数是一项具有挑战性的任务.

本文提出一种基于自动寻优DBSCAN (density-based spatial clustering of applications with noise)的晶圆图预处理方法. 该方法基于传统的DBSCAN聚类,通过分析每个晶圆图缺陷分布特点,能够自动地选择DBSCAN的两个重要参数(Eps和MinPts),从而得到较优的聚类结果,提取更明显的特征,为晶圆图缺陷模式的识别和分类做好准备.

1 相关工作

针对晶圆图缺陷分布的特点,Jin等[5]首次提出将DBSCAN用于晶圆图的噪声点检测中.

1.1 固定参数DBSCAN

DBSCAN是一种经典的基于密度的聚类算法. 该算法从样本密度的角度来考察样本间的可连接性,并基于可连接的样本不断扩展聚类簇,最后确定聚类结构. 该方法能在具有噪声的空间数据库中发现任意形状的簇,可将密度足够大的相邻区域连接,能有效处理异常数据[13],主要用于对空间数据的聚类. DBSCAN 算法涉及到两个重要参数,邻域半径Eps 和聚类簇最小聚类点数MinPts.

Jin等[5]将晶圆图看作是一个Moore邻域结构,它由一个中心样本点和围绕它的8个样本点组成. 中心样本点被认为是核心点,中心样本点距东、北、西、南4个样本点的距离均为1个单元,与对角线方向的

样本点的距离均为$\sqrt{2}$,因此选择Eps参数为$\sqrt{2}$. 将孤立离群点(未连接到任何其他缺陷晶粒的缺陷晶粒)和孪生离群点(两个缺陷晶粒,其中每个缺陷晶粒仅连接到另一个缺陷离群点)认为是噪声点,因此,设置MinPts为3. 根据晶圆图结构的特点,上述方法中针对不同类型的晶圆图选取了相同的DBSCAN参数,而不同的参数对于不同的缺陷类型,处理效果不同. 两个参数$(\sqrt{2},3)$是最小的取值,这可能会导致不同簇集合之间被合并. 因此,研究DBSCAN算法,使其能根据特定晶圆图的空间结构特点自动选择最优的Eps和MinPts参数,去除易混淆的噪声点,突出晶圆图的主要空间图案特征是非常必要的.

1.2　自动参数DBSCAN

当前,许多国内外学者针对DBSCAN参数自动寻优进行了研究. 在文献[14-16]中,都可以根据数据集的特点确定Eps的值,然而对于MinPts的设置,仍需要手动设置,没有实现完全的自动化. Khan等[17]提出的ADDBSCAN自适应聚类算法需要提前指定簇的数量,无法自动识别簇类数目. 李文杰等[18]提出的KANN-DBSCAN算法,通过利用数据集自身分布特性生成候选Eps和MinPts参数,自动寻找聚类结果的簇数变化稳定区间,并将该区间中密度阈值最少时所对应的Eps和MinPts参数作为最优参数,该方法虽然精确度相对较高,但随着数据量的增大,时间消耗也逐渐增大,而且若数据密度分布不均匀,则可能会出现簇数不稳定的情况,也就无法得到参数最优的情况.

综上所述,本文针对晶圆图数据集分布特性,首先生成Eps和MinPts参数列表,然后通过采用簇间密度和簇内密度的综合指标来确定最优参数,并在此基础上增加特征簇和特征点,实现自动寻找DBSCAN最优参数的晶圆图预处理.

1.2.1　生成Eps和MinPts参数列表

DBSCAN算法中Eps和MinPts参数可以用于表示样本点间的紧密程度. 在以Eps为半径的邻域内,邻域内包含的点数越多,密度越大,即邻域内的样本点之间的关系越紧密,则可以认为这些样本点间的相似程度越大. 当Eps增大到一定程度时,邻域面积增大,邻域内包含的点数增长速率减小,导致邻域密度和Eps对应的MinPts减小. 在邻域面积内包含MinPts样本点的密度为MinPts/S,其中,S为以Eps为半径的圆的面积,$S = \pi \times \mathrm{Eps}^2$,因此,需要权衡Eps和MinPts两者之间的关系.对此,本文采用如下生成Eps和MinPts的方法:

1) K-平均最近邻算法生成Eps列表.

①假设数据集Y中有n个样本点,求数据集中各个样本点间的欧氏距离,组成距离矩阵$Y_{n\times n}$,$Y_{n\times n}$是实对称分布矩阵. $Y_{n\times n} = \{D(i,j)|1 \leqslant i \leqslant n, 1 \leqslant j \leqslant n\}$,$D(i,j)$为数据集$Y$中第$i$个样本点与第$j$个样本点间的欧氏距离. 欧氏距离反映了各样本点间的紧密联系程度,距离越大,联系越紧密.

②对$Y_{n\times n}$每一行的元素进行升序排列,得到向量$Y_k = (Y_0, Y_1, \ldots, Y_n)$,即某一个样本点,按照其他样本点与其关系的紧密程度进行升序排列,第1列元素Y_0表示其本身.

③对向量Y_K中的元素求平均,可得到向量Y_K的平均距离$\overline{Y}_K$,$\overline{Y}_K$反映了在数据集中针对任意样本点,第K个样本点与该样本点的紧密程度,将所有$\overline{Y}_K$作为候选Eps参数序列,组成候选集Y_{Eps},$Y_{\mathrm{Eps}} = \{\overline{Y}_K|1 \leqslant K \leqslant n\}$.

2)生成MinPts列表.

对于上述生成的Eps列表,若采用其中任意一个Eps值,则对于数据集中的每个样本点都可以计算出其Eps邻域包含的点数,每个样本点的Eps邻域中包含点数各不相同,对各个邻域内的点数求平均值,即可求出与Eps列表一一对应的MinPts列表,该列表反映了以Eps为邻域半径的平均密度. 该方法具体计算方式如下:以Eps为距离矩阵$Y_{n\times n}$的临界值,求出$Y_{n\times n}$每一行小于Eps参数的个数N,然后将每列求出的参数个数取平均值(不为整数则向上取整),该平均值即为MinPts 参数.

3)提取列表.

当$K = 1$时,是所有样本点到自身的距离,可以忽略. 当K值不断增大,可以很容易分析得出,Eps越大,包含的样本点数越多,当增加到某一个临界值时,晶圆图上的所有缺陷点都将被划分为同一个聚类簇. 再继续增大K值,不会改变聚类结果,因此,继续增大K值对于研究其特点没有意义,反而会增加计算量,浪费时间. 因此,假设当$K = a$时,聚类簇数为1,则a为该临界值,故只取K为$2 \sim a$对应的Eps和MinPts参数列表.

1.2.2　确定最优参数

1)密度.

①簇内密度.

Maria等[19]提出以平均散射Scat计算簇内密度$\mathrm{Scat}(C) = \dfrac{1}{C}\displaystyle\sum_{i=1}^{n}\|\sigma_{(\nu_i)}\| \Big/ \|\sigma_{(S)}\|$. 其中:$S$为数据集,$C$为聚类簇的个数,$\nu_i$为第$i$个簇的中心,$\sigma$为标准差. 各个簇标准差的平均值与数据集总标准差的比

值越小,簇内密度越高,说明聚类效果越好.

采用DBSCAN方式滤波,除了要考虑每个簇的标准差,还需要考虑平均距离[20]. 因此,本文对上述公式作了改进.

定义1 簇内密度

$$\mathrm{Com}=\frac{1}{C}\frac{\sum_{i=1}^{n}\|\sigma_{(\nu_i)\times\mathrm{mean}(\nu_i)}\|}{n_i}. \tag{1}$$

其中:ν_i为第i个簇的中心点,$\sigma(\nu_i)$为第i个簇的标准差,$\mathrm{mean}(\nu_i)$为第i个簇的平均距离,n_i为第i个簇内的点数,C为聚类簇的数量.

假设两个聚类簇具有相同的平均值和标准差,数据点更多的聚类簇紧凑性更好,簇内密度更高. 也就是说,平均值和标准差更小,则紧凑性更好. 将单个聚类簇推广得到晶圆图的所有聚类簇,则所有聚类簇的平均值和标准差越小,紧凑性越好,簇内密度越高. 因此,Com值越小,簇内密度越高.

②簇间密度.

假设数据集$D=\{\nu_i|i=1,2,\dots,C\}$,将该数据集划分为$C$个聚类簇,其中$\nu_i$是每个聚类簇的中心. 对DBSCAN的参数指标定义如下.

定义2 平均聚类簇距离

$$\mathrm{stdev}=\frac{1}{CK}\sum_{i=1}^{C}\sum_{j=1}^{K}(x_j-\nu_i). \tag{2}$$

其中:ν_i为簇C_i的核心点,K为第i个聚类簇中样本点数,x_j为样本点,C为聚类簇数量,stdev为每个样本点与该聚类簇核心点距离的平均值.

定义3 簇间密度

$$\mathrm{Den}=\frac{1}{C(C-1)}\times\sum_{i=1}^{C}\Big(\sum_{j=1,j\neq i}^{C}\frac{\mathrm{density}(u_{ij})}{\max(\mathrm{density}(\nu_i),\mathrm{density}(\nu_j))}\Big). \tag{3}$$

其中:ν_i、ν_j分别为簇C_i和簇C_j的中心,u_{ij}为簇C_i和簇C_j合并后的核心点. $\mathrm{density}(u)$的定义如下:

$$\mathrm{density}(u)=\sum_{K=1}^{n_{ij}}f(x_K,u), \tag{4}$$

n_{ij}属于簇C_i和簇C_j样本数量之和,即样本点$x_K\in(C_i\bigcup C_j)$.

$$f(u)=\begin{cases}0,\ d(x,u)>\mathrm{stdev};\\1,\ \mathrm{otherwise}.\end{cases} \tag{5}$$

故$\mathrm{density}(u)$表示以核心点u为圆心,以stdev为半径的圆内包含的样本点数之和.

对于两个不同的聚类簇,分离度越高越好. 假设两个聚类簇合并后密度为$\mathrm{density}(u_{ij})$,$\mathrm{density}(u_{ij})$相对其中密度较大的聚类簇比值越小,即簇间密度Den越小,两个聚类的分离程度越大,聚类效果越好. 将此方法推广到晶圆图的所有聚类簇,不同聚类簇间分离度越高,簇间密度Den越小,聚类效果越好.

③最终评价指标

$$\mathrm{Score}=\mathrm{Com}+\mathrm{Den}. \tag{6}$$

经过上述分析,簇内密度Com和簇间密度Den都是越小越好,故采用两者之和Score作为评价指标. Score越小聚类效果越好.

选择Eps、MinPts参数使得Score最小时,就是DBSCAN聚类的最优参数Eps、MinPts. 以Score为评价指标寻找DBSCAN最优参数的算法步骤如下:

输入:晶圆图的二维矩阵,即数据集Y;

输出:DBSCAN聚类的Eps、MinPts最优参数.

step 1: 计算样本中各数据点间的欧氏距离,得到距离矩阵$Y_{n\times n}$;

step 2: 将$Y_{n\times n}$中每一行元素升序排列,再计算每列的平均值;

step 3: 将平均值$\overline{Y}_K$作为Eps的候选集Y_{Eps},即$Y_{\mathrm{Eps}}=\overline{Y}_K$;

step 4: 计算每行$Y_K\leqslant Y_{\mathrm{Eps}}$的个数$N$,将$N$作为MinPts的候选参数(MinPts与Eps是一一对应的);

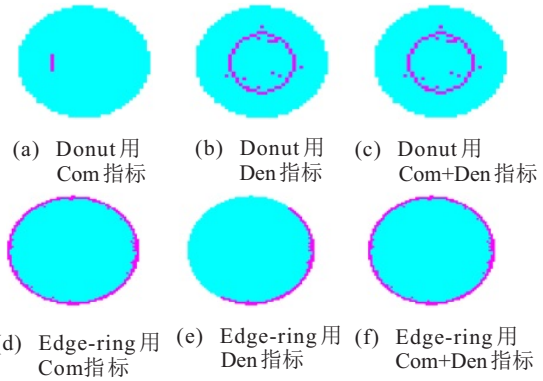
step 5: 使用每组Eps和MinPts参数进行DBSCAN聚类(当聚类簇数为1时不再继续),并使用式(6)计算指标Score;

step 6: 选取Score值最小时所对应的参数Eps和MinPts,即为最优参数;

step 7: 输出最优Eps和MinPts参数.

2)选用不同指标聚类的效果对比.

为了展示本文采用评价指标Score确定最优参数的效果,本文对单独使用簇内密度Com为指标和单独使用簇间密度Den为指标做了实验,用3种不同指标确定DBSCAN聚类的两个最优参数. 对于Donut



(a) Donut用Com指标 (b) Donut用Den指标 (c) Donut用Com+Den指标

(d) Edge-ring用Com指标 (e) Edge-ring用Den指标 (f) Edge-ring用Com+Den指标

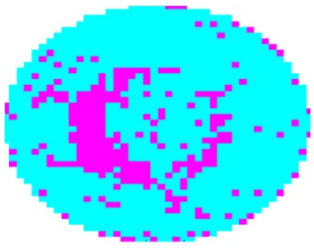
图2 不同指标聚类效果对比

类型的晶圆图,采用Den为指标与采用Com + Den为指标的效果基本相同,而采用Com为指标的效果明显要比另外两种差,如图2所示,(a)为采用Com为指标的晶圆图,(b)和(c)分别为采用Den和Com + Den为指标聚类的晶圆图. 对于Edge-ring类型,采用Com为指标的效果与采用Com + Den为指标的效果相差较小,如图2中(d)和(f)所示,但采用Den为指标时大部分的晶圆图都如图2(e)所示,效果要差很多.

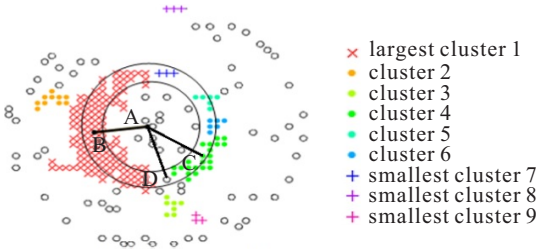
1.3　增加其他特征簇和特征点

由于在自动参数DBSCAN中只保留了最大聚类,而某些晶圆图聚类后只保留最大聚类簇可能会丢失一些重要数据,这些数据涵盖了晶圆图的重要特征. 因此,本文提出采取两种补偿方式:若是一个聚类簇为重要数据则称之为特征簇;若是一些单独的点为重要数据则称之为特征点.

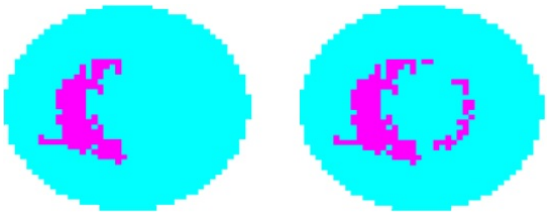
以图3为例,图3中(a)为Donut类型原图,(b)为Donut经过DBSCAN聚类后的结果,图中点A为晶圆图的中心点(坐标中心),点B为最大聚类簇的核心点(聚类簇中所有点到该点的距离之和最小),点C为聚类簇4的核心点,点D为噪声点. 由于不同的晶圆图有不同的大小尺寸,聚类簇形状各异,本文采用比值系数R和r来消除与晶圆图尺寸等相依赖的关系.



(a)　Donut类型原图



(b)　最优参数DBSCAN聚类



(c)　保留最大聚类簇　　(d)　增加特征点和特征簇

图 3　增加特征点和特征簇示意

对于特征簇而言,比值系数定义如下:

$$R_i = \frac{\text{dis\_}C_i}{\text{dis\_max}}. \tag{7}$$

其中: dis_max为最大聚类簇$C_{\max}$核心点$\nu_{\max}$到晶圆图中心点o的欧氏距离, dis_C_i为除最大聚类簇的其他聚类簇C_i核心点ν_i到晶圆图中心点o的欧氏距离.

对于特征点,比值系数定义如下:

$$r_j = \frac{\text{dis\_}x_j}{\text{dis\_max}}. \tag{8}$$

其中: dis_max为最大聚类簇$C_{\max}$核心点$\nu_{\max}$到晶圆图中心点o的欧氏距离, dis_x_j为除形成聚类簇以外的其他缺陷点x_j到晶圆图中心点o的欧氏距离.

由以上定义可以看出,当比值系数越接近1时,该聚类簇或该点与最大聚类簇的相关性就越强. 因此,采用以下方式确定特征簇和特征点.

1) 确定特征簇. 以dis_max为单位长度,即设AB为单位长度,其他聚类簇C_i核心点ν_i到晶圆图中心点o的欧氏距离为dis_C_i = R × dis_max. 由于晶圆图的聚类簇本身具有一定的宽度,将比值系数设定在一定范围内,针对晶圆图的特点,将比值系数R范围设定为0.8～1.2. 则其他聚类簇的核心点在这个范围内,就将其归为特征簇,如图3(b)中点C为Cluster 4的核心点,则将该聚类簇划分为特征簇, Cluster 5、Cluster 6的核心点也在该范围内,同样也保留.

2) 确定特征点. 以dis_max为单位长度,即设AB为单位长度,缺陷点x_j到晶圆图中心点o的欧氏距离为dis_x_j = r × dis_max. 将比值系数范围r设定为0.8～1.2. 如果某个未形成聚类簇的缺陷点在这个比值系数范围内,则将该缺陷点归为特征点,如图3(b)所示,点D为未形成聚类簇的缺陷点,该点在比值系数范围内,故将点D归为特征点. 也就是说,聚类簇或者缺陷点在图3(b)中的两个环之内的聚类簇或点都要保留. 经过DBSCAN聚类保留最大聚类簇后,得到的晶圆图如图3(c)所示,在保留最大聚类簇的基础上增加特征点和特征簇的晶圆图如图3(d)所示,对比图3(c)和3(d)很明显能看出,增加特征簇和特征点后,晶圆图的特征保留得更加完整,更能准确地表达该晶圆图的缺陷模式.

2　实验过程及结果分析

2.1　数 据 集

本文所采用的晶圆图数据集为Wu等[21]提供的(WM-811k)数据库,该数据集包含811 457张从实际生产中收集的晶圆图,其中约有20 %的晶圆图经该

领域专家标记了缺陷类型. 总共有9种标签,分别为Center、Donut、Edge-loc、Edge-ring、Local、Near-full、Random、Scratch和None. 由于数据集是源于实际生产过程中,存在不均衡性,作为预处理阶段的分析,并不是最终的分类处理,本文只采用了训练集. 从WM-811k中标记的训练集来看,None的数量超出很多,为了保持数据的相对均衡,选None类型数据大致为其他8种类型数据之和. 每种类型具体使用的数据情况如表1所示.

表1 数据库

类型	数量
Center	3462
Donut	409
Edge-loc	2417
Edge-ring	8554
Local	1620
Near-full	54
Random	609
Scratch	500
None	18365

2.2 实验对比

在预处理阶段,对于None类型,期望所有的噪声点全部滤除,而对于Random 类型,没有特定的图案,所以难以通过聚类来评判,在研究预处理部分,只使用剩余的7种类型的数据作对比.

2.2.1 固定参数和自动参数对比

为了验证自动参数DBSCAN和固定参数DBSCAN的效果,本文采用轮廓系数S_i对两种算法进行评判. 轮廓系数[22]结合了聚类的紧凑度和分离度,可以用于评估聚类的效果,轮廓系数更大的,聚类效果更好. 对应单个样本i的轮廓系数公式为

$$S_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}. \tag{9}$$

其中:a_i是样本i在簇内到其他点的平均距离,描述簇内的紧凑度;b_i是样本i到不同簇中样本的平均距离,描述簇间的分离度. 整个数据集的轮廓系数就是所有样本轮廓系数的平均值,轮廓系数的范围为$-1\sim1$,当轮廓系数小于0时,说明到不同簇之间的平均距离小于簇内平均距离,聚类效果不好;当轮廓系数大于0时,说明到簇内平均距离小于不同簇之间的平均距离,聚类效果好. 最终通过实验及数据统计,轮廓系数较大的晶圆图数量如表2所示.

由表2可以看出,晶圆图通过自动参数DBSCAN滤波以后的轮廓系数总体上占91.75%,其中:Edge-ring最高,占95.97%,Scratch最低,占68.00%. 从以上数据可以看出:除Scratch以外,其他都在83%以上. 所以,以轮廓系数为指标对比,自动参数DBSCAN优于固定参数DBSCAN.

表2 轮廓系数最优的晶圆图数量

类型	总数	固定DBSCAN	自动DBSCAN
Center	3462	299 (8.64%)	3163 (91.36%)
Donut	409	44 (10.76%)	365 (89.24%)
Edge-loc	2417	278 (11.5%)	2139 (88.50%)
Edge-ring	8554	345 (4.03%)	8209 (95.97%)
Local	1620	269 (16.60%)	1351(83.4%)
Near-full	54	9 (16.67%)	45 (83.33%)
Scratch	500	160 (32.00%)	340 (68.00%)
总数	17016	1404 (8.25%)	15612 (91.75%)

从视觉直观上对比,如图4所示,其中(a)为Center类型的原图,(b)为Center类型固定参数DBSCAN滤波图,(c)为Center自动参数DBSCAN滤波图,两者对于Center类型处理效果相差不大. 而对于Donut类型,原图如图4(d)所示,经过固定参数DBSCAN和自动参数DBSCAN滤波处理结果分别如图4(e)和4(f)所示. 显然,自动参数DBSCAN处理效果更好,保留的特征更加明显.

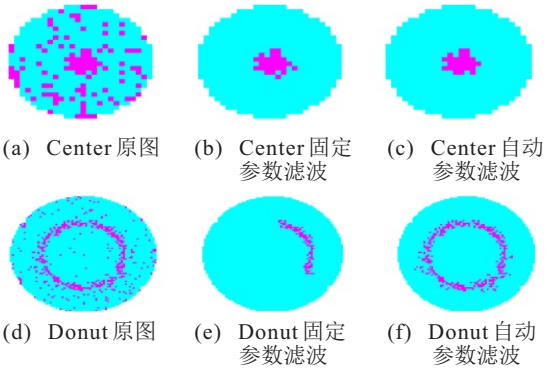


图4 Center和Donut类型滤波对比

2.2.2 自动参数DBSCAN与增加特征簇和特征点对比

通过自动参数DBSCAN聚类后,选择保留最大聚类簇. 然而,对于某些晶圆,尤其是Donut和Edge-ring类型的晶圆,只保留最大的聚类簇,会影响到晶圆的整体特征,所以在保留最大聚类的基础上,增加了特征点和特征簇.

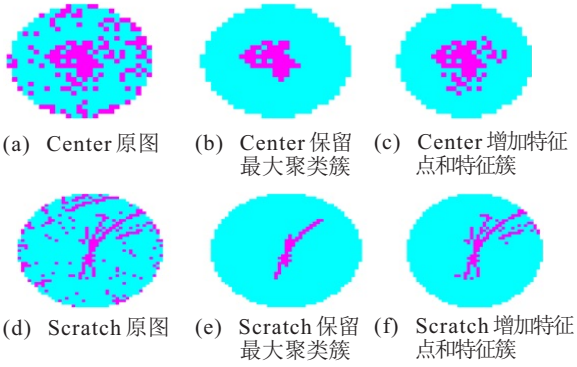


图5 保留最大聚类簇与增加特征簇和特征点对比

从视觉直观上对比. 对于Center类型, 如图5(a)～(c)所示, 其中(a)为Center的原图, (b)为保留最大聚类簇后的晶圆图, (c)为增加特征簇和特征点后的晶圆图, 由(b)和(c)来看, 虽然没有影响主观上的判断, 但增加了特征簇和特征点后的晶圆图, 特征更加完整. 对于Scratch类型, 如图5(d)～(e)和(f)所示, 增加特征簇和特征点以后, Scratch特征更加完整, 更有利于后续的特征提取和分类.

2.2.3　分类对比

为了说明本文提出的预处理方法的有效性, 在固定参数DBSCAN和自动参数DBSCAN的预处理方法下, 使用支持向量机对两两类型的晶圆进行二分类. 固定参数DBSCAN和自动参数DBSCAN分类准确率如表3～表12所示, Other表示除该类型外的其他所有类型, 加粗标记为准确率高的那一项, 两者准确率相同则不标记.

Center与其他类型的分类准确率如表3所示, 除了与Donut和Nearfull分类准确率相同以外, 其余分类结果, 自动参数DBSCAN都要比固定参数DBSCAN的准确率高.

Donut与其他类型的分类准确率如表4所示, 对于训练集, 与None和Random的分类, 使用自动参数

表3　Center的二分类准确率　%

类型1	类型2	训练集		验证集	
		固定参数	自动参数	固定参数	自动参数
Center	Donut	99.4	99.4	98.3	**98.7**
	Edge-loc	99.4	**99.8**	99.6	**99.8**
	Edge-ring	99.9	**100**	100	100
	Local	97.7	**98.9**	98.0	**99.0**
	Near-full	100	100	100	100
	None	99.6	**99.7**	99.6	**99.7**
	Other	99.0	**99.3**	98.8	**99.2**
	Random	99.0	**99.2**	99.0	**99.6**
	Scratch	99.7	**99.9**	99.7	99.7

表4　Donut的二分类准确率　%

类型1	类型2	训练集		验证集	
		固定参数	自动参数	固定参数	自动参数
Donut	Center	99.4	99.4	98.3	**98.7**
	Edge-loc	99.8	99.8	99.1	**99.3**
	Edge-ring	100	100	**99.9**	99.8
	Local	**96.7**	96.4	**95.9**	95.1
	Near-full	100	100	100	100
	None	99.9	**100**	99.8	**99.9**
	Other	99.6	99.6	99.4	**99.5**
	Random	98.7	**99**	95.9	**97.3**
	Scratch	**99.4**	99.0	98.2	98.2

表5　Edge-loc的二分类准确率　%

类型1	类型2	训练集		验证集	
		固定参数	自动参数	固定参数	自动参数
Edge-loc	Center	99.4	**99.8**	99.6	**99.8**
	Donut	99.8	99.8	99.1	**99.3**
	Edge-ring	98.5	**98.6**	**98.5**	98.4
	Local	96.7	**97.3**	95.9	**96.6**
	Near-full	**100**	99.9	100	100
	None	99.1	**99.2**	99.0	99.0
	Other	97.6	97.6	97.2	97.2
	Random	98.2	**99.1**	96.5	**98.2**
	Scratch	97.9	**98.3**	98.2	98.2

表6　Edge-ring的二分类准确率　%

类型1	类型2	训练集		验证集	
		固定参数	自动参数	固定参数	自动参数
Edge-ring	Center	99.4	**100**	100	100
	Donut	100	100	**99.9**	99.8
	Edge-loc	98.5	**98.6**	**98.5**	98.4
	Local	**100**	99.9	99.8	**99.9**
	Near-full	100	100	100	100
	None	99.8	99.8	99.8	99.8
	Other	99.4	99.4	99.3	99.3
	Random	100	100	99.9	99.9
	Scratch	99.6	**99.7**	99.7	99.7

表7　Local的二分类准确率　%

类型1	类型2	训练集		验证集	
		固定参数	自动参数	固定参数	自动参数
Local	Center	97.7	**98.9**	98.0	**99.0**
	Donut	**96.7**	96.4	**95.9**	95.1
	Edge-loc	96.7	**97.3**	95.9	**96.6**
	Edge-ring	**100**	99.9	99.8	**99.9**
	Near-full	100	100	100	100
	None	98.9	**99.0**	99.0	**99.1**
	Other	96.2	**96.2**	96.1	**96.5**
	Random	99.0	**99.2**	97.6	**98.0**
	Scratch	96.8	**96.9**	97.1	97.1

表8　Near-full的二分类准确率　%

类型1	类型2	训练集		验证集	
		固定参数	自动参数	固定参数	自动参数
Near-full	Center	100	100	100	100
	Donut	100	100	100	100
	Edge-loc	**100**	99.9	100	100
	Edge-ring	100	100	100	100
	Local	100	100	100	100
	None	100	100	100	100
	Other	100	100	100	100
	Random	100	100	99.3	**100**
	Scratch	100	100	100	100

表9 None的二分类准确率 %

类型1	类型2	训练集		验证集	
		固定参数	自动参数	固定参数	自动参数
None	Center	99.6	**99.7**	99.6	99.7
	Donut	99.9	**100**	99.8	**99.9**
	Edge-loc	99.1	**99.2**	99.0	99.0
	Edge-ring	99.8	99.8	99.8	99.8
	Local	98.9	**99.0**	99.0	**99.1**
	Near-full	100	100	100	100
	Other	98.2	**98.3**	**98.1**	98.0
	Random	99.9	**100**	**100**	99.9
	Scratch	99.4	99.4	99.4	99.4

表10 Other的二分类准确率 %

类型1	类型2	训练集		验证集	
		固定参数	自动参数	固定参数	自动参数
Other	Center	99.0	**99.3**	98.8	**99.2**
	Donut	99.6	99.6	99.4	**99.5**
	Edge-loc	97.6	97.6	97.2	97.2
	Edge-ring	99.4	99.4	99.3	99.3
	Local	96.2	**96.7**	96.1	**96.5**
	Near-full	100	100	100	100
	None	98.2	**98.3**	**98.1**	98.0
	Random	99.2	**99.6**	99.0	**99.6**
	Scratch	98.7	**98.9**	99.0	**99.2**

表11 Random的二分类准确率 %

类型1	类型2	训练集		验证集	
		固定参数	自动参数	固定参数	自动参数
Random	Center	99.0	**99.2**	99.0	**99.6**
	Donut	98.7	**99.0**	95.9	**97.3**
	Edge-loc	98.2	**99.1**	96.5	**98.2**
	Edge-ring	100	100	99.9	99.9
	Local	99.0	**99.2**	97.6	**98.0**
	Near-full	100	100	99.3	**100**
	None	99.9	**100**	**100**	99.9
	Other	99.2	99.6	99.0	**99.6**
	Scratch	**99.7**	99.6	99.1	**100**

表12 Scratch的二分类准确率 %

类型1	类型2	训练集		验证集	
		固定参数	自动参数	固定参数	自动参数
Scratch	Center	99.7	**99.9**	99.7	99.7
	Donut	99.4	99.4	98.2	98.2
	Edge-loc	97.9	**98.3**	98.2	98.2
	Edge-ring	99.6	**99.7**	99.7	99.7
	Local	96.8	**96.9**	97.1	97.1
	Near-full	100	100	100	100
	None	99.4	99.4	99.4	99.4
	Other	98.7	**98.9**	99.0	**99.2**
	Random	**99.7**	99.6	99.1	**100**

DBSCAN的准确率高,其余分类准确率基本相同. 而对于验证集,除了与Edge-ring和Local分类准确率较低以外,其余分类结果,自动参数DBSCAN要优于固定参数DBSCAN.

Edge-loc与其他类型的分类准确率如表5所示,除了与Edge-ring和Nearfull分类准确率自动参数低0.1 %,其他分类准确率,自动参数DBSCAN都要比固定参数DBSCAN高.

Edge-ring与其他类型的分类准确率如表6所示,对于训练集,Edge-ring与Local分类结果,固定参数DBSCAN准确率比自动参数DBSCAN准确率高0.1 %. 对于验证集,自动参数DBSCAN分类准确率要高于固定参数DBSCAN分类准确率0.1 %. 对于Edge-ring与Local的分类效果,两种预处理方法的优劣很难评估. 其他分类效果相差不大.

剩余表7、表8、表9、表10、表11、表12分别代表Local、Near-full、None、Random、Scratch与其他类型的分类准确率. 可以看出,使用自动参数DBSCAN的分类准确率,明显优于固定参数DBSCAN的分类准确率.

通过表3～表12的分类准确率的分析对比可以得出: 本文所提出的自动参数DBSCAN方法比传统固定参数DBSCAN方法好.

3 结 论

本文提出的基于自动寻找DBSCAN最优参数的晶圆图预处理方法,对于晶圆图的滤波效果好. 首先是在传统的DBSCAN的基础上,使用K-平均最近邻算法和数学期望法生成Eps和MinPts的参数列表,然后采用聚类簇的簇内密度、簇间密度的综合指标来选择Eps和MinPts的最优参数. 利用最优参数对晶圆图进行DBSCAN聚类,保留最大聚类簇,再根据其他聚类簇或缺陷点与晶圆中心的相对位置关系,保留特征点和特征簇.

本文采用的自动寻优DBSCAN的预处理方法从直观上分析要比其他的滤波方法好,保留的特征更明显且更具代表性,更能突出晶圆图的缺陷模式. 从轮廓系数这个评价指标来看,本文采用的方法轮廓系数较高,不同滤波方法在相同分类方法下的分类效果,本文提出的方法分类准确率更高,因此本文使用的滤波效果更好.

本文提出的自动寻优DBSCAN的预处理方法对于大部分的聚类问题,自动寻找最优参数的方法可以适用,但由于晶圆图缺陷模式的固有特性,增加特征点和特征簇的方法只适用于晶圆图分类的预处理,对于其他的数据集不一定适用.

参考文献(References)

[1] Hong C W, Hong Y F, Liu S F. Research on process parameter design by molecular heuristic particle swarm optimization in semiconductor test vehicle[D]. Taiwan: National Tsinghua University, 2011: 1-10.

[2] 陈超. 晶圆制造中焊盘结晶缺陷的检测方法和工艺流程改善研究[D]. 上海: 上海交通大学, 2017.
(Chen C. Research on detection method and process flow improvement of pad crystal defects in wafer manufacturing[D]. Shanghai: Shanghai Jiaotong University, 2017.)

[3] Liu H, Shi S, Yang P, et al. An improved genetic algorithm approach on mechanism kinematic structure enumeration with intelligent manufacturing[J]. Journal of Intelligent & Robotic Systems, 2018, 89(3): 343-350.

[4] Chen S H, Wang Z, Hou X N, et al. A general boundary scan test system based on EDIF netlist file transfer to Protel netlist file[J]. International Journal of Materials and Structural Integrity, 2016, 10(1/2/3): 70-80.

[5] Jin C H, Na H J, Piao M H, et al. A novel DBSCAN-based defect pattern detection and classification framework for wafer Bin map[J]. Transactions on Semiconductor Manufacturing, 2019, 32(3): 286-292.

[6] Jeong Y S, Kim S J, Jeong M K. Automatic identification of defect patterns in semiconductor wafer maps using spatial correlogram and dynamic time warping[J]. IEEE Transactions on Semiconductor Manufacturing, 2008, 21(4): 625-637.

[7] Yang P, Qin X N. A hybrid optimization approach for chip placement of multi-chip module packaging[J]. Microelectronics Journal, 2009, 40(8): 1235-1243.

[8] Hansen M H, Nair V N, Friedman D J. Monitoring wafer map data from integrated circuit fabrication processes for spatially clustered defects[J]. Technometrics, 1997, 39(3): 241-253.

[9] Kang S. Joint modeling of classification and regression for improving faulty wafer detection in semiconductor manufacturing[J]. Journal of Intelligent Manufacturing, 2018, 31(2): 1-8.

[10] Yuan T, Kuo W, Bae S J, Detection of spatial defect patterns generated in semiconductor fabrication processes[J]. IEEE Transactions on Semiconductor Manufacturing, 2011, 24(3): 392-403.

[11] Kim J, Lee Y, Kim H. Detection and clustering of mixed type defect patterns in wafer bin maps[J]. IISE Transactions, 2018, 50(2): 99-111.

[12] Wang C H, Wang S J, Lee W D. Automatic identification of spatial defect patterns for semiconductor manufacturing[J]. International Journal of Production Research, 2006, 44(23): 5169-5185.

[13] Chakraborty S, Nagwani N K. Analysis and study of incremental DBSCAN clustering algorithm[J]. Computer Science, DOI: 10.1007/BF02948834.

[14] Yue S H, Ping L I, Guo J D, et al. A statistical information-based clustering approach in distance space[J]. Journal of Zhejiang University Science: Science in Engineering, 2005, 6(1): 71-78.

[15] Jahirabadkar S, Kulkarni P. Algorithm to determine - distance parameter in density based clustering[J]. Expert Systems with Applications, 2014, 41(6): 2939-2946.

[16] 秦佳睿, 徐蔚鸿, 马红华, 等. 自适应局部半径的DBSCAN聚类算法[J]. 小型微型计算机系统, 2018, 39(10): 2186-2190.
(Qin J R, Xu W H, Ma H H, et al. DBSCAN clustering algorithm with adaptive local radius[J]. Journal of Chinese Mini-Micro Computer Systems, 2018, 39(10): 2186-2190.)

[17] Khan M M R, Siddique M A, Arif R B, et al. ADBSCAN: Adaptive density-based spatial clustering of applications with noise for identifying clusters with varying densities[C]. The 4th International Conference on Electrical Engineering and Information and Communication Technology (ICEEICT). Bangladesh: IEEE, 2018: 445-451.

[18] 李文杰, 闫世强, 蒋莹, 等. 自适应确定DBSCAN算法参数的算法研究[J]. 计算机工程与应用, 2019, 55(5): 1-7.
(Li W J, Yan S Q, Jiang Y, et al. Research on method of self-adaptive determination of DBSCAN algorithm parameters[J]. Computer Engineering and Applications, 2019, 55(5): 1-7.)

[19] Maria H, Michalis V. Clustering validity assessment: Finding the optimal partitioning of a data set[C]. International Conference on Data Mining. San Jose, 2001: 187-194.

[20] Hu L, Zhong C. An internal validity index based on density-involved distance[J]. IEEE Access, 2019, 7: 40038-40051.

[21] Wu M J, Jang J-S R, Chen J L. Wafer map failure pattern recognition and similarity ranking for large-scale data sets[J]. IEEE Transactions on Semiconductor Manufacturing, 2015, 28(1): 1-12.

[22] 朱连江, 马炳先, 赵学泉. 基于轮廓系数的聚类有效性分析[J]. 计算机应用, 2010, 12: 139-141.
(Zhu L J, Ma B X, Zhao X Q. Clustering validity analysis based on silhouette coefficient[J]. Journal of Computer Applications, 2010, 12: 139-141.)

作者简介

陈寿宏 (1981–), 男, 博士生, 从事机器学习、集成电路测试等研究, E-mail: cshgl@guet.edu.cn;

易木兰 (1995–), 女, 硕士生, 从事机器学习、TSV测试的研究, E-mail: 1070749788@qq.com;

张雨璇 (1996–), 女, 硕士生, 从事机器学习、TSV测试的研究, E-mail: 1217695438@qq.com;

尚玉玲 (1977–), 女, 研究员, 博士生导师, 从事信号完整性、TSV非接触测试等研究, E-mail: syl@guet.edu.cn;

杨平 (1964–), 男, 教授, 博士生导师, 从事微/纳米机械系统、神经网络优化等研究, E-mail: yangping1964@163.com.

(责任编辑：孙艺红)