DOI: 10.7641/CTA.2015.41129

基于扩散K近邻距离的间歇过程故障诊断

李　元[1†], 刘亚东[1], 张　成[1,2]

(1. 沈阳化工大学 技术过程故障诊断与安全性研究中心, 辽宁 沈阳, 110142;

2. 东北大学 信息科学与技术学院, 辽宁 沈阳, 110819)

摘要: 针对间歇过程多模态、变量非线性、非高斯分布等特征, 提出一种基于扩散K近邻距离的故障诊断方法. 该方法首先在样本集完全图中应用马尔科夫随机游走定义带有分量权重的扩散距离, 可以有效提取数据样本的关联信息和统计特征, 然后应用K近邻规则方法对样本数据进行故障诊断. 这种应用扩散距离替换传统K近邻规则欧式距离的统计方法, 既可以提升对数据样本关联性信息的有效提取能力, 又可以使得K近邻规则处理非线性、多模态检测问题的性能得以保持. 通过在半导体蚀刻批次过程中的仿真应用, 与传统线性、非线性方法的对比分析, 实验结果验证了方法的有效性.

关键词: 扩散距离; K近邻规则; 故障诊断; 间歇过程

中图分类号: TP277　　**文献标识码**: A

Fault detection of batch process based on diffusion K-nearest neighbors distance

LI Yuan[1†], LIU Ya-dong[1], ZHANG Cheng[1,2]

(1. Research Center for Technical Process Fault Diagnosis and Safety, Shenyang University of Chemical Technology, Shenyang Liaoning 110142, China;

2. School of Information Science and Technology, Northeastern University, Shenyang Liaoning 110819, China)

Abstract: According to the multi-mode, variable nonlinear and non-Gaussian distribution characteristics of the batch process, we propose a new fault detection method based on diffusion K-nearest neighbor distance (FD–DDKNN). In this work, the diffusion distance with component weight which can effectively fetch correlation information and statistical characteristics in data samples is defined through Markovian random walk in complete graph of the samples set. Then, the adapted K-nearest neighbor rule (KNN) method is applied to data samples to detect faults. This method that replaces diffusion distance to conventional Euclidean distance in K-nearest neighbor rule, not only raises the ability of fetching relevance information in data samples, but also improves the performance of dealing with nonlinear and multi-mode characteristics based on KNN rule for detection problem. By the simulation application in the semiconductor etching batch process, compared with the traditional linear and nonlinear methods, the experimental results validate the effectiveness of the proposed method.

Key words: diffusion distance; K-nearest neighbors rule; fault diagnosis; batch process

1　引言(Introduction)

在现代工业制造业中, 产品生产广泛采用间歇生产过程. 由间歇生产过程的特点统计发现批次不等长、多工序、分布非高斯等已成为该过程的显著特征. 为保证生产质量和提高生产效率, 间歇过程的故障检测方法研究逐渐成为必备条件. 多元统计控制通过主元分析(principal component analysis, PCA)应用T^2和SPE统计量进行故障检测, 该方法在间歇过程中已经得到广泛应用[1–5], 但PCA在具有多工序、分布非高斯、非线性、多模态等特点批次生产过程中应用是相对困难的.

Jon和Zhao等人[6–7]针对多工序和多模态的影响, 提出多模型方法. 模型建立和维护的开销大是多模型方法所具有的缺点, 同时多模型系统一旦建立其可移植性具有较大的局限性. Choi等人[8–12]针对过程特征或操作条件改变等问题提出回归模型方法, 例如回归主元分析. 产品设备保养后过程特征发生变化, 直接影响该方法的适应适应性, 同时回归模型识别正常变

收稿日期: 2014−12−04; 录用日期: 2015−05−20.

†通信作者. E-mail: li-yuan@mail.tsinghua.edu.cn.

国家自然科学基金项目(61174119, 61490701), 辽宁省教育厅重点实验室项目(LZ2015059)资助.

Supported by National Natural Science Foundation of China (61174119, 61490701) and Liaoning Province Key Laboratory Foundation (LZ2015059).

化和故障缓慢漂移问题是困难的. 基于核多元分析技术主要解决非线性问题, 基本思想是通过非线性映射将原始的非线性输入空间变换到一个高维隐性的线性特征空间. 张等人[13]提出将非线性频谱与核主元分析相结合的方法, 能够大幅度降低频谱数据维数, 最终提高故障识别能力. Scholkopf和Cui等人[14–15]提出的核多元分析由于非线性映射后的数据形式不可见, 为故障检测带来不便.

针对间歇过程的非线性和不等长特性, 张等人[16]提出一种基于统计模量故障检测方法, 并将其应用在间歇生产过程中, 统计模量方法有效提升检测性能, 同时降低了批次不等长、多工况等特征对故障检测的影响. He等人[17]提出一种基于 K 近邻规则的故障检测方法 (fault detection using the K-nearest neighbor, FD–KNN), 该方法在故障检测过程中克服半导体数据非线性和多工况特点, 在应用中取得较好的效果. 为了降低FD–KNN的计算负担, 同时保持其处理非线性与多模态的能力, He 等人[18]提出基于PC–KNN的故障检测方法, 该方法应用PCA数据降维之后, 在低维主元空间应用K近邻技术进行检测, 可以显著提高检测效率. 由于FD–KNN应用欧式距离进行近邻的特征统计, 忽略变量对距离贡献的权重分析, 因此对较小特征的故障样本检测能力受到限制.

扩散距离是源于扩散小波和扩散映射的概念与方法. 扩散映射的主要思想是在数据集上构造一个扩散图, 用扩散距离描述数据间的关联程度[19–22]. 扩散映射使用一系列的扩散核进行数据降维, 从而避免了高维矩阵进行特征分解时的不稳定性和不可行性. 扩散映射为揭示高维数据的复杂结构提供了一种重要的工具, 开辟了新的研究方向, 在聚类、分类、机器学习和降维等多个领域都有广泛的应用[23–24]. 扩散距离的关键在于它是基于扩散图上的多条路径, 因此较之测地距离, 扩散距离对噪声更具有鲁棒性, 能够准确提取数据样本的关联信息.

本文应用基于扩散距离的K近邻规则进行故障检测(fault detection based on K nearest neighbors using diffusion distance, FD–DDKNN). 该方法首先在训练样本完全图中应用马尔科夫随机游走, 通过定义权重矩阵得到t步转移概率, 进而定义出带有权重信息的扩散距离. 扩散距离可以有效衡量数据样本的关联信息, 同时使得数据分布信息更加清晰, 易于统计. 接下来, 应用扩散距离替换传统K近邻规则中距离度量方法, 可以有效提升对数据关联程度的分析, 同时使得传统K近邻方法处理数据非线性和多模态的能力得到保持. 通过模拟仿真实例说明FD–DDKNN可以有效处理具有非线性、非高斯、多模态等特点的批次过程检测问题. 本文方法在半导体蚀刻生产工艺中的应用进一步验证其有效性.

2 扩散距离(Diffusion distance)

本文目标是建立一种能够反映数据点间关联程度的距离度量方法. 首先假设处理的对象是数据集构成的完全图Ω, 样本点为图Ω中的节点. 为了区别图Ω中不同类别的点, 需要测量数据点间的相互作用. 如果图Ω中两个节点$\boldsymbol{x}$和$\boldsymbol{y}$间存在大量短距离路径, 则$\boldsymbol{x}$和$\boldsymbol{y}$之间是密集的, 同时可以认为$\boldsymbol{x}$和$\boldsymbol{y}$之间是存在高度相关性. 基于以上, 将引入扩散框架[20], 同时定义出扩散距离.

步骤 1 假设$G=\{\Omega, W\}$是一个包含n个节点附带权重的有限完全图, 其中$W=[w(\boldsymbol{x},\ \boldsymbol{y})]_{\boldsymbol{x},\boldsymbol{y}\in\Omega}\in\mathbb{R}^{n\times n}$为权重矩阵, 满足以下条件: ① 对称性: $W=W^{\mathrm{T}}$, $w(\boldsymbol{x},\boldsymbol{y})=w(\boldsymbol{y},\boldsymbol{x})$; ② 非负性: $w(\boldsymbol{x},\boldsymbol{y})\geqslant 0$, $\boldsymbol{x}$, $\boldsymbol{y}\in\Omega$. 这里, 定义权重$w(\boldsymbol{x},\boldsymbol{y})$完全应用数据驱动的方法, $w(\boldsymbol{x},\boldsymbol{y})$反映数据$\boldsymbol{x}$和$\boldsymbol{y}$关联度和密集度. 基于以上可以引用高斯核

$$w=\exp(-\|\boldsymbol{x}-\boldsymbol{y}\|^2/\beta), \tag{1}$$

数据点间的关联性取决于图G中点间权重的强度, 权重大, 关联性强, 否则, 关联性弱. 图1给出数据集Ω中标号为1−10的样本点结构分布, 节点3与其余点的关联性较弱, 或者认为其是一个离群点. 图2(a)中$w_{i\cdot}$表示节点i相应于其他节点的权重, 可以看出除了标签为1, 3节点外$w_{3\cdot}<w_{2\cdot}$, 进一步说明节点2处于一个密集度较高的分布区域, 由于节点2与3处于同一圆周之上, 故$w_{31}=w_{21}$.

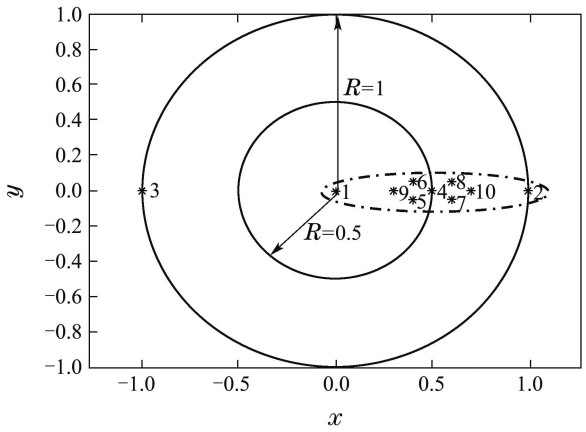


图 1 数据集G样本散点图

Fig. 1 Samples scatter diagram in data set G

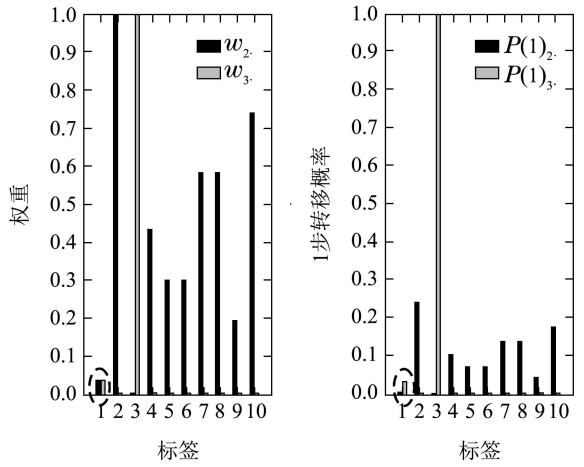
步骤 2 根据权重的定义, 图$G=\{\Omega, W\}$展现出数据集的位置结构信息, 接下来在图$G=\{\Omega, W\}$上定义一个马尔科夫随机游走, 定义转移概率.

$$\deg(\boldsymbol{x})=\sum_{\boldsymbol{z}\in\Omega}w(\boldsymbol{x},\boldsymbol{z}), \tag{2}$$

$$p_1(\boldsymbol{x},\boldsymbol{y})=\frac{w(\boldsymbol{x},\boldsymbol{y})}{\deg(\boldsymbol{x})}. \tag{3}$$

依式(2)−(3)可以得到一个$n\times n$的一步转移概率矩阵$P(1)=[p_1(\boldsymbol{x},\boldsymbol{y})]_{\boldsymbol{x},\boldsymbol{y}\in\Omega}\in\mathbb{R}^{n\times n}$, 其中: $p_1(\boldsymbol{x},\boldsymbol{y})$即为由$\boldsymbol{x}$ 到$\boldsymbol{y}$的一步转移概率, $p_1(\boldsymbol{x},\boldsymbol{y})$量值反映了图中节点一阶近邻结构分布, $p_1(\boldsymbol{x},\boldsymbol{y})$越大, 说明$\boldsymbol{x}$与$\boldsymbol{y}$越密集, 关联性越强. 在扩散映射方法中, 通过利用一步转

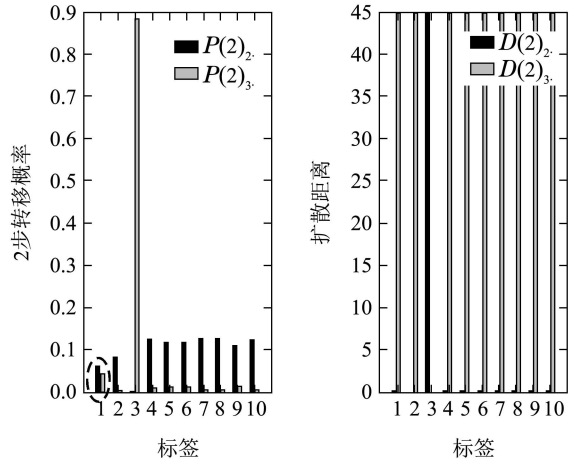
移概率矩阵$P(1)$获得更大范围近邻的结构信息. $P(t)=[p_t(\boldsymbol{x},\boldsymbol{y})]_{\boldsymbol{x},\boldsymbol{y}\in\Omega}\in\mathbb{R}^{n\times n}$是$P(1)$的$t$次迭代, $p_t(\boldsymbol{x},\boldsymbol{y})$是由$\boldsymbol{x}$到$\boldsymbol{y}$的$t$步转移概率, 随着$t$的适当增加,节点间近邻的关联影响可由$p_t(\boldsymbol{x},\boldsymbol{y})$精确表述. 换言之, $P(t)$反映了带有权重的结构图中数据节点的本质结构特征. 图2(b)与图3(a)分别给出马尔科夫随机游走中的1步转移和2步转移概率, 其中$P(t)_{i\cdot}$表示标号为i的节点向其余节点进行t步转移概率, 由$P(1)_{2\cdot}$和$P(1)_{3\cdot}$可见节点2相对于节点3而言处于一个关联性较强的区域($P(1)_{2k}>P(1)_{3k}, k=4,5,\cdots,10$), 如图2(b)所示, 但由一步转移概率并不能正确表述节点2, 3与节点1的关联程度($P(1)_{21}<P(1)_{31}$). 图3(a)中经过2步转移后上述关联性质得到保持, 同时节点间关联程度得到进一步加强, 数据分类更加明显($P(1)_{2k}>P(1)_{3k},\ k=1,4,5,\cdots,10$), 由图3(a)通过2步转移概率可以进一步验证节点3是一个离群点的事实.



(a) 权重分析　　(b) 1步转移概率分析

图 2　节点2与3权重与转移概率

Fig. 2　Weight and transition probability on node 2 and 3



(a) 2步转移概率分析　　(b) 与其他节点扩散距离

图 3　节点2与3转移概率和扩散距离

Fig. 3　Transition probability and diffusion distance on node 2 and 3

在关联图Ω中, 对于$\boldsymbol{x}$到$\boldsymbol{y}$的t步转移概率$p_t(\boldsymbol{x},\boldsymbol{y})$而言有

$$\lim_{t\to+\infty}p_t(\boldsymbol{x},\boldsymbol{y})=\varphi_0(\boldsymbol{y}),\tag{4}$$

其中: $\varphi_0(\boldsymbol{y})=\dfrac{\deg(\boldsymbol{y})}{\sum\limits_{\boldsymbol{z}\in\Omega}\deg(\boldsymbol{z})}$, $\varphi_0(\boldsymbol{y})$的数值与节点$\boldsymbol{y}$的度$\deg(\boldsymbol{y})$成比例增长. 显然$\varphi_0(\boldsymbol{y})$是节点$\boldsymbol{y}$的密度的一种测量. 依据马尔科夫链的可逆性, 可证明下列等式成立: $\varphi_0(\boldsymbol{x})p_1(\boldsymbol{x},\boldsymbol{y})=\varphi_0(\boldsymbol{y})p_1(\boldsymbol{y},\boldsymbol{x})$.

$$P(t)\underset{(t\to\infty)}{\to}\begin{bmatrix}\varphi_0(\boldsymbol{x}) & \varphi_0(\boldsymbol{y}) & \cdots & \varphi_0(\boldsymbol{z})\\ \varphi_0(\boldsymbol{x}) & \varphi_0(\boldsymbol{y}) & \cdots & \varphi_0(\boldsymbol{z})\\ \vdots & \vdots & \vdots & \vdots\\ \varphi_0(\boldsymbol{x}) & \varphi_0(\boldsymbol{y}) & \cdots & \varphi_0(\boldsymbol{z})\end{bmatrix}.\tag{5}$$

因此, t的选取不宜多大, 否则平稳分布将影响数据分布结构. 表1描述了节点2和3经t步转移后, 其转移概率之间的关系. $P(t)_{2k}/P(t)_{3k}$表示由节点2和3向节点k经t步转移概率比值, 可以看出一步转移概率并不能精确描述出节点3是一个离群点, 而经过2步转移后, 可以发现节点2向其余节点转移能力大于节点3相应的能力, 进而说明2步转移概率可以较精确地描述数据间的关联性. 随着t的增大, 转移概率变化平稳, 当$t=100$时, 出现平稳分布, 此时数据的结构信息不能通过转移概率精确体现.

表 1　t步转移概率对数据关联性分析

Table 1　Correlation analysis based on transition probability

	$P(1)_{2\cdot}/P(1)_{3\cdot}$	$P(2)_{2\cdot}/P(2)_{3\cdot}$	$P(10)_{2\cdot}/P(10)_{3\cdot}$	$P(100)_{2\cdot}/P(100)_{3\cdot}$
1	0.3	1.3	1.3	1
4	196	12.8	3.2	1
5	5.1	9.8	3	1
6	5.1	9.8	3	1
7	746	16	3.3	1
8	746	16	3.3	1
9	13.7	7.1	2.8	1
10	2830	19.6	3.4	1

步骤 3　考虑如下问题: 对于确定且有限的$t>0$, 寻求一种度量, 若$p_t(\boldsymbol{x},\cdot)$和$p_t(\boldsymbol{z},\cdot)$相应分布较近则$\Omega$中数据节点$\boldsymbol{x}$与$\boldsymbol{z}$较接近. 现定义节点$\boldsymbol{x}$与$\boldsymbol{z}$扩散距离(diffusion distance–DD)DD_t:

$$DD_t(\boldsymbol{x},\boldsymbol{z})=\sqrt{\|p_t(\boldsymbol{x},\cdot)-p_t(\boldsymbol{z},\cdot)\|^2_{1/\phi_0}}=\sqrt{\sum_{\boldsymbol{y}\in\Omega}\frac{\left(p_t(\boldsymbol{x},\boldsymbol{y})-p_t(\boldsymbol{z},\boldsymbol{y})\right)^2}{\phi_0(\boldsymbol{y})}},\tag{6}$$

其中$\dfrac{1}{\varphi_0(\boldsymbol{y})}$是节点密度不同区域相应的惩罚差异因子.

图3(b)分别给出图1中节点2, 3相应的扩散距离.

$D(t)_{i\cdot}$表示经t步转移后第i节点到其他节点的扩散距离, 便于区别, 相应的欧式距离记为$OD_{i\cdot}$. 由于欧式距离并没有考虑分量权重信息, 因此并不能较好表述数据间的关联性, 例如, $OD_{21}=OD_{31}$. 而引入扩散距离后, 数据的关联信息得到良好表现, 如图3(b), 可以看出节点3到其他节点的扩散距离均大于节点2的相应距离且$D(2)_{31}>D(2)_{21}$. 通过数据分析, 可以认为节点2较节点3而言分布在一个数据密集且关联性较强的区域, 即节点2与除节点3以外的点为同类别, 而节点3是一个离群点.

扩散距离反应的是来自扩散过程中数据点之间相关性的内在几何特征, 数据点间随机游走的步伐越多, 数据点间有更大的转移概率, 扩散距离就越小. 在一个特定的规模下数据的相关性, 如果在权重图里有更高的连接性, 数据点就越近, 因此扩散距离强调集群的概念; $DD_t^2(\boldsymbol{x},\boldsymbol{y})$包含了从$\boldsymbol{x}$到$\boldsymbol{y}$的所有步长, 因此对噪声的干扰具有较强的鲁棒性, 不同于测量学距离.

3 基于K近邻规则的故障检测方法(Fault detection using K-nearest-neighbor rule)

FD–KNN方法认为正常样本轨迹与训练样本的轨迹相似, 故障样本轨迹与之发生明显的偏离. FD–KNN已经成功被应用到半导体蚀刻工艺的故障检测中. FD–KNN方法由两部分组成: 模型建立和故障检测, 其流程如图4所示.

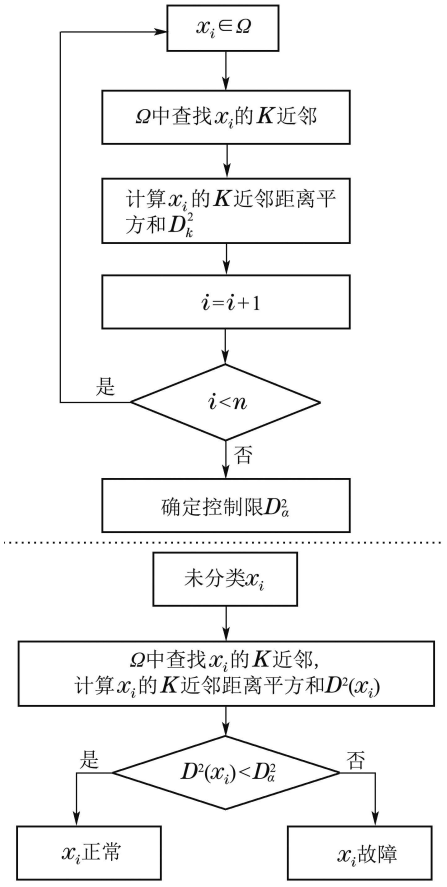


图 4 FD-KNN流程图

Fig. 4 Flowcharts on FD-KNN

4 基于DDKNN规则故障检测方法(Diffusion distance based K-nearest-neighbor rule approach for fault detection)

由于FD–KNN方法应用不带有权重的欧式距离进行关联性的度量, 对检测结果带来一定的影响. 本文引入更能表述数据特征的扩散距离进行故障检测, 进一步提高检测性能.

FD–DDKNN方法由两部分组成: 模型建立和故障检测.

I) 模型建立. 模型建立需要3步:

步骤 1 应用扩散距离查找训练集中$\boldsymbol{x}_i\in\Omega,\ i=1,2,\cdots,n$的$K$近邻$\{\boldsymbol{x}_i^1,\boldsymbol{x}_i^2,\cdots,\boldsymbol{x}_i^K\},\ \boldsymbol{x}_i^j\in\Omega,\ j=1,2,\cdots,K$;

步骤 2 为训练样本$\boldsymbol{x}_i\in\Omega,\ i=1,2,\cdots,n$计算$K$近邻扩散距离的平方和$DD_t^{(i)}$:

$$DD_t^{(i)}=\sum_{j=1}^{K}DD_t^2(\boldsymbol{x}_i,\boldsymbol{x}_i^j),$$

$DD_t^2(\boldsymbol{x}_i,\boldsymbol{x}_i^j)$为第$i$个样本与其第$j$近邻扩散距离的平方;

步骤 3 确定用于检测故障的控制限. 由于$DD_t^{(i)}$近似符合偏χ^2分布, 依据显著性水平α确定$CL, CL=\chi_\alpha^2(N)$, 为了方便CL也可以应用统计中核密度估计进行计算.

II) 故障检测. 对于未分类的新样本$\boldsymbol{x}_*$, 其检测过程由3步构成:

步骤 1 依扩散距离在训练集Ω中查找K近邻$\{\boldsymbol{x}_*^1,\boldsymbol{x}_*^2,\cdots,\boldsymbol{x}_*^K\}$;

步骤 2 计算$\boldsymbol{x}_*$的K近邻扩散距离平方之和: $DD_t^{(*)}=\sum_{j=1}^{K}DD_t^2(\boldsymbol{x}_*,\boldsymbol{x}_*^j)$;

步骤 3 比较$DD_t^{(*)}$与CL, 如果$DD_t^{(*)}\leqslant CL$, 则$\boldsymbol{x}_*$正常, 否则为故障.

FD–DDKNN采用具有能够反映数据关联性的扩散距离进行故障检测, 在保证充分提取数据关联信息的前提下, 使得FD–KNN处理非线性、多模态等问题的能力得到保持, 可以提高检测率. 以下仿真实验进一步说明这个问题.

5 工业实例(Industrial instance)

半导体蚀刻工艺数据来源于美国德州仪器公司的半导体生产过程[25]. 该数据是由3个工况的108个正常批次和21个故障批次组成. 第1工况有34个批次的正常样本和9个批次的故障样本; 第2工况有37个批次的正常样本和6个故障样本; 第3工况有37个批次的正常样本和6个故障样本. 由于第2工况的1个正常批次和1个故障批次大量缺失数据, 只有107个正常批次和20个故障批次是有效的. 本文从107批次中随机选取11个批次作为校验批次, 其余96个批次作为训练批次. 原始数据包括40个变量, 本文只使用其中的19个变量, 这些变量与产品的生产过程和最终状态密切相

关[18].

半导体蚀刻工艺中具有如下特点, 如图5 (变量Endpoint A散点图)所示: ① 批次宽度不等长: 在107批次中, 批次持续时间由95 s变化到112 s, 不同批次同一变量采样时间不相同; ② 工序宽度不等长: 从图5中可以看出, 不同工况中的阶段4持续时间不同, 由44 s变化至52 s. 且同一工况中阶段4不遵循相似的时间轨迹; ③ 进程漂移: 对于蚀刻过程, 由于材料的不同, 蚀刻的范围不同等原因都可以导致进程轨迹漂移; ④ 非高斯分布, 批次数据同一变量不服从高斯分布. 半导体数据以上特点严重影响传统故障检测方法.

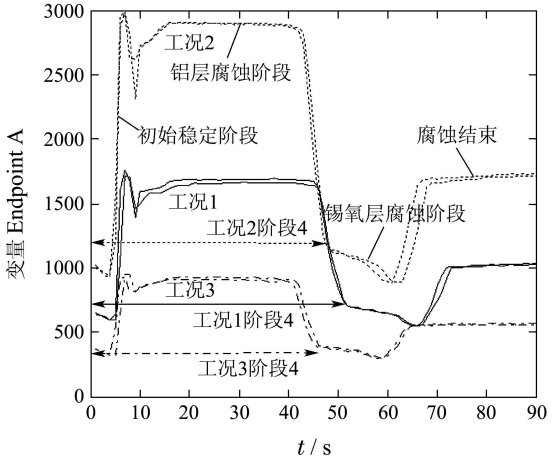


图 5 变量Endpoint A散点图

Fig. 5 Scatter plot of Endpoint A

本文采用文献[26]的方法将数据由3D展开成2D, 记为X, 如图6所示. 接下来, 应用PCA, KPCA, FD–KNN和FD–DDKNN方法对该数据集进行故障检测与分析.

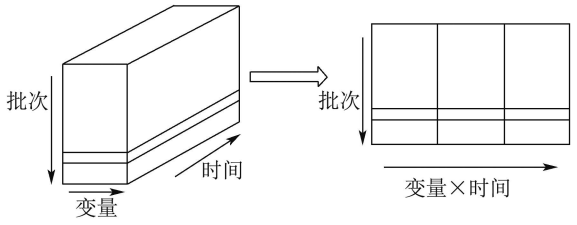
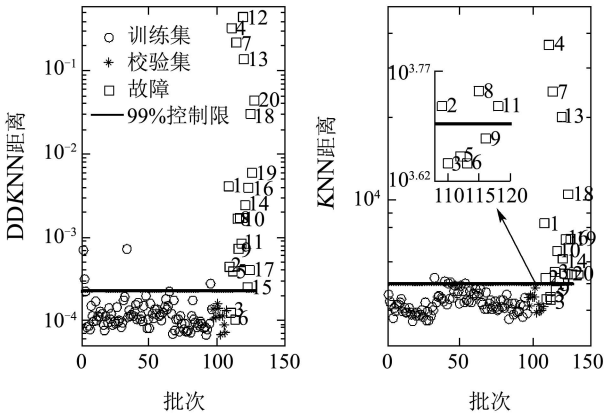


图 6 批次3D展开2D

Fig. 6 Unfold batch from 3D to 2D

如图7所示, FD–DDKNN方法选取99%的控制限可以有效检测出18个故障批次, 其中近邻数$k = 3$, 转移步数$t = 2$. 应用PCA, 选取2个主元和99%的控制限, SPE统计量共检测出12个故障批次, T^2统计量检测效果不佳, 仅检测出3个故障, 如图8所示.

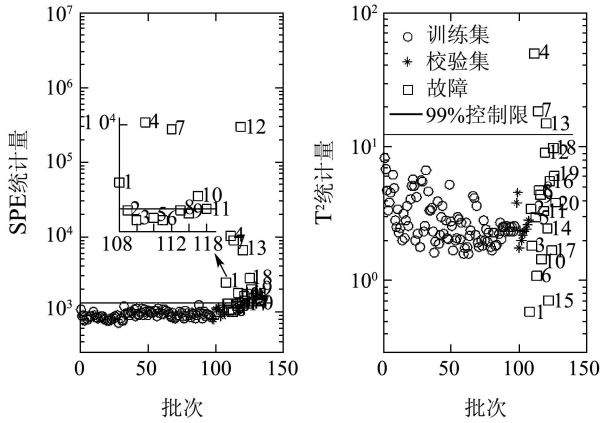
而应用PC–KNN和KPC–KNN, 其检测结果虽优于PCA, 但通过PC或KPCA方法降维后再用KNN检测的效果不如直接用KNN方法检测效果好, 如图9所示. 由于KPCA方法通常产生一个局部最优, 在保证全局最优上具有一定困难, 从而KPCA不能明确考虑数据的内在几何结构, 如图 10 所示, 用 KPC–SPE和KPC–T^2 检测, 多个故障批次未能检测出来. FD–DDKNN方法不受半导体数据特点的影响, 检测效果优于传统的PCA, KPCA方法, 训练集中数据并不符合高斯分布同时变量间具有非线性特征, 这是影响PCA检测性能的主要原因.



(a) FD−DDKNN　　(b) FD−KNN

图 7 FD–DDKNN和FD–KNN检测结果分析

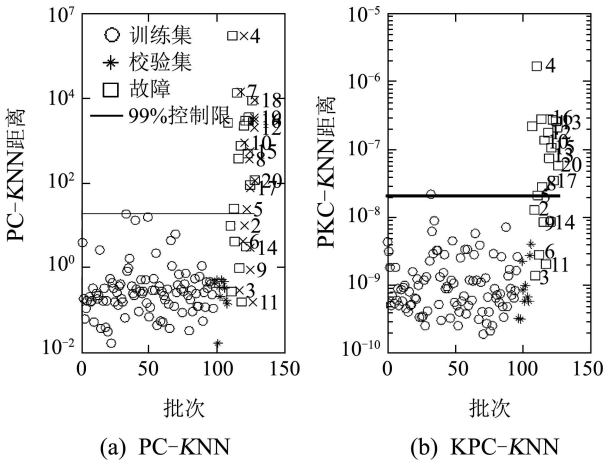
Fig. 7 Simulation results based on FD–DDKNN and FD–KNN



(a) PC−SPE　　(b) PC−T^2

图 8 PC–SPE和PC–T^2检测结果分析

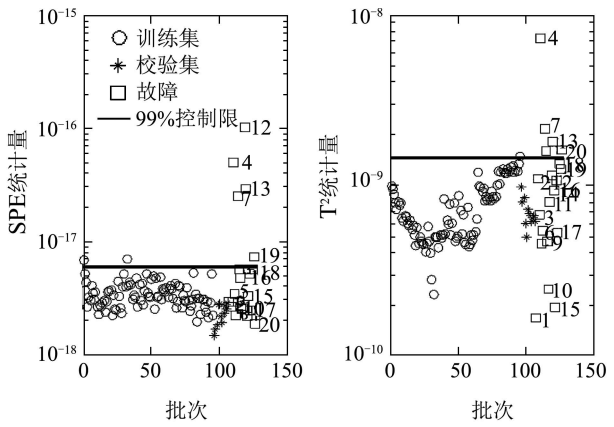
Fig. 8 Simulation results based on PC–SPE and PC–T^2



(a) PC−KNN　　(b) KPC−KNN

图 9 PC–KNN和KPC–KNN检测结果分析

Fig. 9 Simulation result based on PC–KNN and KPC–KNN



(a) KPCA–SPE (b) KPCA–T^2

图 10 KPC–SPE和KPC–T^2检测结果分析

Fig. 10 Simulation results based on KPC–SPE and KPC–T^2

如图11所示, 半导体数据集通过应用扩散映射方法, 将原始数据降维到二维空间中, 其样本散点图清晰表现出故障数据与校验数据均成功映射到不同的类别, 且多数故障数据成功分离出来. 本文方法相对FD–KNN方法而言, 后者有4个故障批次未被检出. 各种方法检测结果见表2.

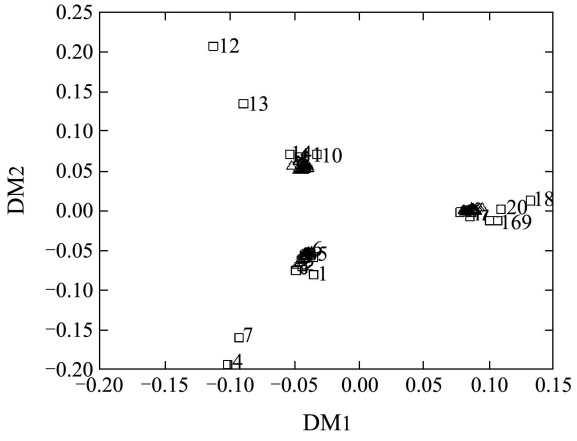


图 11 半导体数据集样本散点图

Fig. 11 Samples scatter diagram in semiconductor data set

表 2 检测结果分析

Table 2 The result analysis

方法	未检出批	检测率/%
本文方法	3,6	90
KNN	3,5,6,9	80
PCA–T^2	1,2,3,5,6,8,9,10,11,12, 14,15,16,17,18,19,20	15
PCA–SPE	2,3,5,6,8,9,11,17	60
KPCA–T^2	1,2,3,4,5,6,9,10,11,12, 14,15,16,17,19,20	20
KPCA–SPE	1,2,3,5,6,9,10,11,14,15,16,17,20	35
PC–KNN	2,3,6,9,11,14	70
KPC–KNN	2,3,5,6,9,11,14	65

6 结论(Conclusions)

本文提出一种基于扩散距离与KNN相结合的故障检测方法, 应用扩散距离近邻之和衡量样本的关联程度进行故障检测. 通过引入基于扩散距离的KNN方法可以有效处理非线性、非高斯、多模态等数据特征的检测问题. 将本文方法应用到数据特征复杂的半导体批次生产过程中, 检测效果优于传统的方法, 进一步验证本文方法的有效性. 由于在检测过程中频繁计算扩散距离及其分布, 计算量较大, 下一步工作在提高本文方法性能的前提下, 研究如何降低计算负载.

参考文献(References):

[1] 周东华, 李钢, 李元. 数据驱动的工业过程故障诊断与预测技术–PCA与PLS的方法 [M]. 北京: 科学出版社, 2011: 22 – 30.
(ZHOU Donghua, LI Gang, LI Yuan. *Data Driven Industrial Process Fault Diagnosis Technology-Based on PCA and PLS Methods* [M]. Beijing: Science Press, 2011: 22 – 30.)

[2] 李荣雨, 荣冈. 基于故障映射向量和结构化残差的主元分析(PCA)故障隔离 [J]. 控制理论与应用, 2008, 25(6): 1099 – 1104.
(LI Rongyu, RONG Gang. Principal component analysis(PCA) of fault isolation based on fault mapping vector and structured residual [J]. *Control Theory & Applications*, 2008, 25(6): 1099 – 1104.)

[3] LI Y, ZHANG X M. Diffusion maps based k-nearest-neighbor rule technique for semiconductor manufacturing process fault detection [J]. *Chemometrics and Intelligent Laboratory Systems*, 2014, 136(1): 47 – 57.

[4] YU J B. Fault detection using principal components-based Gaussian mixture model for semiconductor manufacturing processes [J]. *IEEE Transactions on Semiconductor Manufacturing*, 2011, 24(3): 432 – 444.

[5] HUNG H, WU P S, TU I P, et al. On multilinear principal component analysis of order-two tensors [J]. *Biometrika*, 2012, 99(3): 569 – 583.

[6] GUNTHER J C, CONNER J S, SEBORG D E. Process monitoring and quality variable prediction utilizing PLS in industrial fed-batch cell culture [J]. *Journal of Process Control*, 2009, 19(5): 914 – 921.

[7] ZHAO S J, ZHANG J, XU Y M. Monitoring of processes with multiple operating modes through multiple principle component analysis models [J]. *Industrial & Engineering Chemistry Research*, 2004, 43(22): 7025 – 7035.

[8] CHOI S W, MARTIN E B, MORRIES A J, et al. Adaptive multivariate statistical process control for monitoring time-varying processes [J]. *Industrial & Engineering Chemistry Research*, 2006, 45(9): 3108 – 3118.

[9] DAYAL B S, MACGREGOR J F. Recursive exponentially weighted pls and its applications to adaptive control and prediction [J]. *Journal of Process Control*, 1997, 7(3): 169 – 179.

[10] LEE D S, VANROLLEGHEM P A. Adaptive consensus principal component analysis for on-line batch process monitoring [J]. *Environmental Monitoring and Assessment*, 2004, 92(1/2/3): 119 – 135.

[11] LI W H, YUE H H, VALLECERVANTES V, et al. Recursive PCA for adaptive process monitoring [J]. *Journal of Process Control*, 2000, 10(5): 471 – 486.

[12] WANG X, KRUGER U, LENNOX B. Recursive partial least squares algorithms for monitoring complex industrial processes [J]. *Control Engineering Practice*, 2003, 11(6): 613 – 632.

[13] 张家良, 曹建福, 高峰, 等. 结合非线性频谱与核主元分析的复杂系统故障诊断方法 [J]. 控制理论与应用, 2012, 29(12): 1558 – 1564.

(ZHANG Jialiang, CAO Jianfu, GAO Feng, et al. Fault diagnosis complex system based on nonlinear spectrum and kernel principal component analysis [J]. *Control Theory & Applications*, 2012, 29(12): 1558 – 1564.)

[14] SCHOLKOPF B, MIKA S, BURGES C J C, et al. Input space versus feature space in kernel based methods [J]. *IEEE Transactions on Neural Networks*, 1999, 10(5): 1000 – 1017.

[15] CUI P L, LI J H, WANG G Z. Improved kernel principal component analysis for fault detection [J]. *Expert Systems with Applications*, 2008, 34(2): 1210 – 1219.

[16] 张成, 李元. 基于统计模量分析间歇过程故障检测方法研究 [J]. 仪器仪表学报, 2013, 34(9): 2103 – 2110.
(ZHANG Cheng, LI Yuan. Study on the fault-detection method in batch process based on statistical pattern analysis [J]. *Chinese Journal of Scientific Instrument*, 2013, 34(9): 2103 – 2110.)

[17] HE Q P, WANG J. Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes [J]. *IEEE Transactions on Semiconductor Manufacturing*, 2007, 20(4): 345 – 354.

[18] HE Q P, WANG J. Principal component based k-nearest-neighbor rule for semiconductor process fault detection [C] //*2008 American Control Conference*. Seattle, WA: IEEE, 2008: 1606 – 1611.

[19] COIFMAN R R, LAFON S. Diffusion maps [J]. *Applied and Computational Harmonic Analysis*, 2006, 21(1): 5 – 30.

[20] COIFMAN R R, LAFON S, LEE A B, et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps [J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2005, 102(21): 7426 – 7431.

[21] SINGER A, WU H T. Vector diffusion maps and the connection Laplacian [J]. *Communications on Pure and Applied Mathematics*, 2012, 65(8): 1067 – 1144.

[22] HUANG Y, ZHA X F, LEE J, et al. Discriminant diffusion maps analysis: A robust manifold learner for dimensionality reduction and its applications in machine condition monitoring and fault diagnosis [J]. *Mechanical Systems and Signal Processing*, 2013, 34(1): 277 – 297.

[23] COIFMAN R R, KEVREKIDIS I G, LAFON S, et al. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems [J]. *Multiscale Modeling & Simulation*, 2008, 7(2): 842 – 864.

[24] MAGGIONI M, MHASKAR H N. Diffusion polynomial frames on metric measure spaces [J]. *Applied and Computational Harmonic Analysis*, 2008, 24(3): 329 – 353.

[25] WISE B M, GALLAGHER N B, BUTLER S W, et al. A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process [J]. *Journal of Chemometrics*, 1999, 13(3-4): 379 – 396.

[26] SINGH K P, MALIK A, BASANT N. Multi-way partial least squares modeling of water quality data [J]. *Analytica Chimical Acta*, 2007, 584(2): 385 – 396.

作者简介:

李　元　(1964–), 女, 教授, 博士, 主要研究方向为过程控制、故障诊断, E-mail: li-yuan@mail.tsinghua.edu.cn;

刘亚东　(1989–), 男, 硕士研究生, 主要研究方向为故障诊断, E-mail: liuyadonglove@126.com;

张　成　(1979–), 男, 讲师, 博士研究生, 主要研究方向为系统监控、故障检测, E-mail: 85753141@qq.com.