基于局部临近标准化的 FD-KNN 故障检测

李　元，李美萱，张　成，冯立伟
（沈阳化工大学 信息工程学院，辽宁 沈阳 110142）

摘　要：针对方差相差大的多模态故障诊断数据问题，提出一种基于局部临近标准化（local nearest neighborhood standardization，LNNS）的 k 近邻故障检测方法（fault detection -k-nearest neighbor rule，FD-KNN）。首先，计算每个样本的局部近邻，采用近邻特征实现标准化，克服传统标准化方法 Z-score 将多模态数据看成一个整体而使数据不准确问题；其次，计算每个样本间距离，建立局部临近标准化距离模型，通过临近距离确定控制限。最后，在半导体生产过程中进行仿真应用研究，通过实验结果的比较与分析表明了所提方法的有效性。

关键词：多模态；局部近邻；k 近邻；故障检测

中图分类号：TP277　　**文献标志码**：A　　**文章编号**：1672-3767(2017)05-0001-06

DOI：10.16452/j.cnki.sdkjzk.2017.05.001

FD-KNN Fault Detection Based on Local Nearest Neighborhood Standardization

LI Yuan, LI Meixuan, ZHANG Cheng, FENG Liwei
(College of Information Engineering, Shenyang University of Chemical Technology, Shenyang, Liaoning 110142, China)

Abstract: For large variance of multi-mode fault diagnosis data, this paper presents the k-nearest neighbor fault detection method (FD-KNN) based on local nearest neighborhood standardization (LNNS). Firstly, the local nearest neighbor of each sample was calculated and the local nearest neighborhood feature was used to achieve data standardization so as to overcome the inaccurate data produced by the traditional standard method of Z-score because it took the multi-mode data as a whole. Secondly, the distance between samples was calculated and a local nearest neighborhood standardization distance model was established to determine the control limits based on local distance. Finally, simulation study was conducted in the process of semiconductor manufacturing and the effectiveness of the proposed method was verified by analyzing comparing the experimental results with those of other methods.

Key words: multi-mode; local nearest neighborhood; k-nearest neighbor; fault detection

多模态生产过程是指由于外界环境等条件的变化和生产方案的变动或是过程本身固有特性等因素，导致生产过程具有多个稳定工况[1-3]。由于生产策略不同，使得实际工业过程有多个运行模态，在这类多模态过程中，过程数据的均值与协方差结构将随着模态的不同而发生变化。传统的主元分析法（principal component analysis，PCA）[4-6]在多模态过程故障诊断应用中，将所有的样本作为一个整体对待，忽略了类别属性，而它所忽略的投影方向有可能刚好包含了重要的可分性信息，并且 PCA 对非线性过程的过程监视具有局限性，在多模态过程中不能给出满意的监控效果。

收稿日期：2017-04-04
基金项目：国家自然科学基金项目（61673279，61490701）；辽宁省教育厅重点实验室项目（LZ2015059，510.99）；辽宁省教育厅一般项目（L2015432）
作者简介：李　元（1964—），女，辽宁沈阳人，教授，博士，主要从事过程控制、数据驱动的故障诊断方面的研究.
Email：li-yuan@mail.tsinghua.edu.cn.
李美萱（1989—），女，辽宁锦州人，硕士研究生，主要从事数据驱动的故障诊断方面的研究.

为解决多模态过程监控与故障检测问题，HE 等[7]提出了一种 k 近邻故障检测方法(FD-KNN)，通过计算各训练样本的距离分布，确定统计控制限，进行过程的监视与故障检测[8-9]。在过程数据样本中，假设输入样本被认为是正常的，但如果其距离小于控制限，则为故障数据点，这种方法可以很好的解决具有多模态的数据问题。但当不同模态数据分布密度相差很大时，可能会漏报一些故障数据。传统标准化方法 Z-score，是将不同模态作为一个整体对待，这样就忽略了各个模态方差相差很大的情况，从而造成检测故障不准确。对于方差相差较大的多模态情况，采用局部临近标准化，可以避免由于错误标准化对数据的影响和导致的故障漏报和错报问题。

本文结合多种故障诊断方法在不同状态下进行的故障检测的表现及问题[10-12]，针对具有方差相差较大的多模态情况，提出了采用局部临近标准化样本，并基于 FD-KNN 方法建立故障检测模型。首先，对每一个样本采用局部临近的均值和标准差进行标准化，然后对标准化后的样本[13-14]，利用 FD-KNN 方法确定控制限，最后在半导体生产过程中进行仿真研究，实现在多模态过程中的故障检测。

1 基于 PCA 的故障检测

设 $\boldsymbol{X}$ 的原始数据矩阵的 n 个样本(行)和 m 个变量(列)。计算 $\boldsymbol{X}$ 的协方差矩阵 $\boldsymbol{R}$：

$$\boldsymbol{R}=\frac{1}{n-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}。\tag{1}$$

并且对 $\boldsymbol{R}$ 进行 SVD 分解：

$$\boldsymbol{R}=\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}。\tag{2}$$

其中 $\boldsymbol{\Lambda}$ 是特征值对角阵，且按降序排序的特征值($\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m\geqslant 0$)；$\boldsymbol{v}$ 是 $\boldsymbol{R}$ 的特征向量。选择负荷矩阵 $\boldsymbol{P}\in\boldsymbol{R}^{m\times k}$ 的列，将 $\boldsymbol{X}$ 降维。则样本在低维空间(主元空间)的投影包含在得分矩阵中。

对任意一个样本向量 $\boldsymbol{X}$ 都可以分解为

$$\boldsymbol{X}=\hat{\boldsymbol{x}}+\tilde{\boldsymbol{x}}=\hat{\boldsymbol{x}}+\boldsymbol{E}。\tag{3}$$

$$\hat{\boldsymbol{x}}=\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{x}\tag{4}$$

得分矩阵 $\boldsymbol{T}$ 的表达式为：

$$\boldsymbol{T}=\boldsymbol{X}\boldsymbol{P}。\tag{5}$$

由(3)和(4)可得：

$$\hat{\boldsymbol{x}}=\boldsymbol{T}\boldsymbol{P}^{\mathrm{T}}。\tag{6}$$

可以得到残差矩阵 $\boldsymbol{E}$：

$$\boldsymbol{E}=\boldsymbol{X}-\hat{\boldsymbol{X}}。\tag{7}$$

所以，原始样本 $\boldsymbol{X}$ 可以写成：

$$\boldsymbol{X}=\boldsymbol{T}\boldsymbol{P}^{\mathrm{T}}+\boldsymbol{E}=t_1\boldsymbol{p}_1^{\mathrm{T}}+t_2\boldsymbol{p}_2^{\mathrm{T}}+\cdots+t_k\boldsymbol{p}_k^{\mathrm{T}}+\boldsymbol{E}。\tag{8}$$

最主要的是选取主元 k 的值。用求取累计方差百分比(Cumulative Percent Vaviance，CPV)的方法，来确定主元的个数。

$$CPV=\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{A}\lambda_i}\geqslant 90\%。\tag{9}$$

其中 λ_i 是 $\boldsymbol{X}$ 的协方差矩阵的特征值，A 为总的特征值个数。当 CPV 值大于期望值时，所求的 k 值即为主元个数。当利用正常的 $\boldsymbol{X}$ 完成主元模型建立后，就可以对新的数据进行故障检测。

2 基于局部临近标准化的 FD-KNN 故障检测

2.1 基于 FD-KNN 的故障检测

HE 等[5]提出了 FD-KNN 算法，给出了一种根据计算训练样本间的距离来确定控制限，从而判断校验

数据的故障点。文献中将样本与其训练样本中的 k 个近邻之间距离的平方和作为监控统计量，并通过非中心直方分布来估计其控制限。

$$D_i^2 = \sum_{i=1}^{k} d_{ij}^2 。 \tag{10}$$

其中 D_i^2 是第 i 个样本和第 j 个近邻之间的欧式距离的平方。但是对于方差差距大的多模态来说，如果故障点距离密度大的模态比较接近时，FD-KNN 就很难检测出这类数据的故障点。

2.2　局部近邻标准化(LNNS)方法

在训练集 $\boldsymbol{X}\in\boldsymbol{R}^{n\times m}$ 中，对训练集中的每个样本 x_i，$i=1,2,\cdots,n$ 计算样本间的欧式距离，确定 k 个近邻。这 k 个近邻组成 x_i 的一个新邻域样本，记为 $N(x_i)$。即：

$$N(x_i)=\{x_i^1,x_i^2,\cdots,x_i^k\}), \tag{11}$$

$$d(x_i,x_i^1)>d(x_i,x_i^2)>\cdots>d(x_i,x_i^k)。 \tag{12}$$

其中，$i=1,2,\cdots,n$。x_{ik} 表示 $\boldsymbol{X}$ 中样本到 x_i 的第 k 个近邻样本，$d(x_i,x_{ik})$ 表示样本 x_i 到第 k 个样本的距离。对训练样本 x_i 进行如下标准化得到：

$$x'_i=\frac{x_i-\mathrm{mean}(N(x_i))}{\mathrm{std}(N(x_i))}。 \tag{13}$$

对待检测的新样本 x_j 标准化时，从训练样本中确定前 k 个近邻，组成 x_j 的新的近邻样本 $M(x_j)$，进行如下标准化得到：

$$x'_j=\frac{x_j-\mathrm{mean}(M(x_j))}{\mathrm{std}(M(x_j))}。 \tag{14}$$

其中，$\mathrm{mean}(M(x_j))$ 表示样本 x_j 在训练样本中邻域均值，$\mathrm{std}(M(x_j))$ 表示样本 x_j 在样本训练样本中邻域标准差。

2.3　局部近邻标准化样本 FD-KNN 故障检测

模型建立和故障检测过程如图 1 所示。

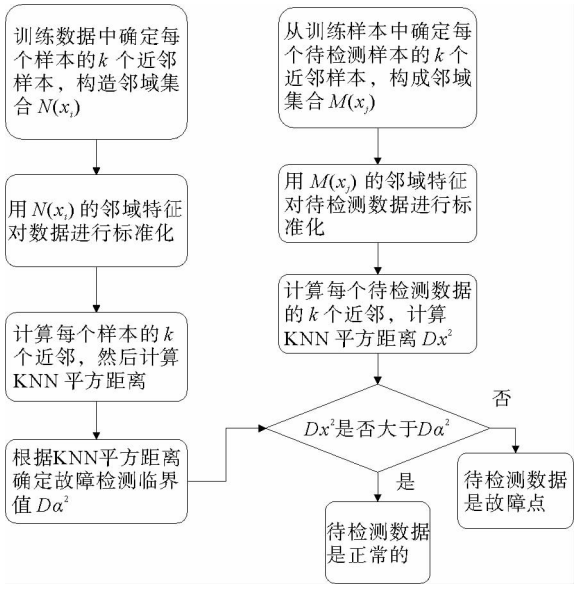

图 1　模型建立和故障检测过程

Fig. 1　Model building and the process of fault detection

2.3.1　模型建立

1）计算样本间的临近距离，找到每个样本的邻域，用局部近邻标准化方法对数据进行标准化；

2）采用 FD-KNN 算法，计算每个样本的前 k 个近邻距离；

3）计算每个样本的 KNN 平方距离，即 D_i^2。KNN 平方距离为第 i 个样本到其 k 个近邻距离的平方的和，即

$$D_i^2 = \sum_{j=1}^{k} d_{ij}^2 。 \tag{15}$$

其中 d_{ij} 表示第 i 个样本到第 j 个样本的欧式距离；

4）由于 D_i^2 近似符合偏 χ^2 分布，当计算出所有样本的 D_i^2 后，为了方便，利用核密度估计的方法求得 D_i^2 的序列累积分布函数，在设定置信区间的情况下，即可得到该置信度下的控制限 D_α^2。

2.3.2　故障检测

1）对测试数据按照局部近邻标准化方法标准化，求标准化后的测试数据与训练数据之间的前 k 个距离；

2）计算测试数据的 KNN 平方距离 D_X^2。比较 D_X^2 和 D_α^2。如果 $D_X^2\leqslant D_\alpha^2$，则为正常数据，否则是故障点。

3 模拟实验

选取两个数值变量服从正态分布的简单例子。训练样本中包含两个模态，两个模态的方差差距很大，从而使得其中一个模态的样本间距很大，而另一个模态的样本间距很小。设置了 4 个正常分布的故障点和 1 个距离样本间距很小的故障点。分别用 PCA、FD-KNN 和 LNNS-FD-KNN 三种方法对待测点进行检测。设置的原始样数据点如图 2 所示。

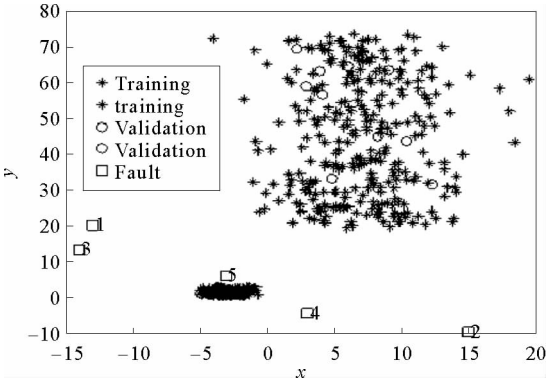

图 2 原始样本数据图

Fig. 2 Original sample chart

对于非线性多模态情况，PCA 检测不出其中的故障点（如图 3 所示）。

根据原始样本图，可以看出模态 1 的样本比较密集，而模态 2 的样本相对稀疏。如果按照 FD-KNN 算法，得出的临界点会根据两个模态的平均值来计算，这样就相当于拉大了模态 1 样本间的距离。由于故障点到模态 1 的距离太近，故障 5 检测不出来，如图 4 所示。

但同一组测试数据，本文提出的 LNNS-FD-KNN 的方法可以很好的避免这种情况（如图 5），故障点 5 可以被检测出来。这也说明了本文提出的方法好于 FD-KNN 方法。

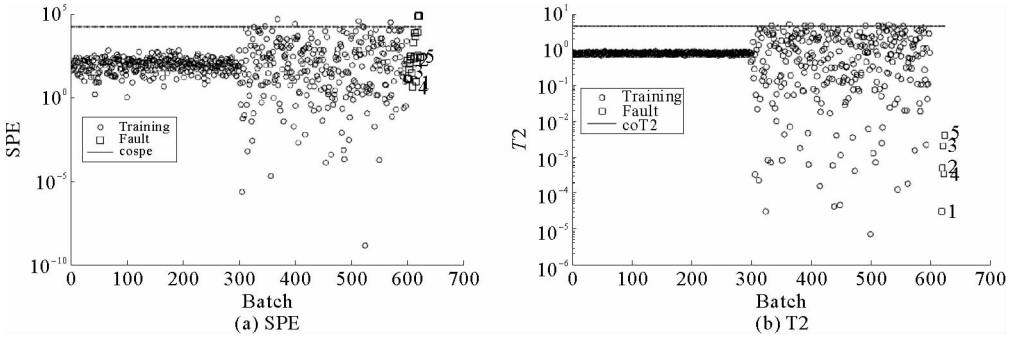

图 3 基于 PCA 的故障检测

Fig. 3 Fault detection based on PCA

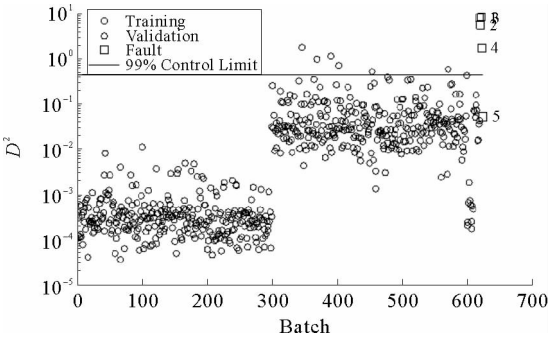

图 4 基于 FD-KNN 的故障检测

Fig. 4 Fault detection based on FD-KNN

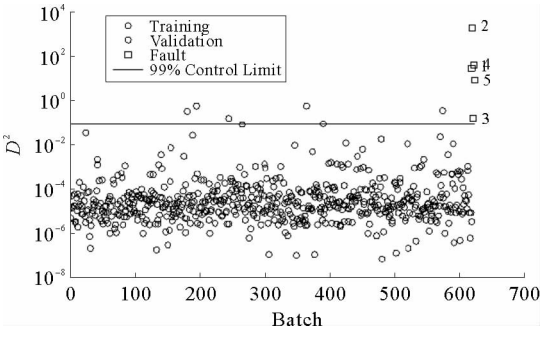

图 5 基于 LNNS-FD-KNN 的故障检测

Fig. 5 Fault detection based on LNNS-FD-KNN

4　半导体生产过程的仿真研究

4.1　数据采集

采集数据来自半导体生产中铝堆蚀刻工艺过程。数据集来自于 3 个不同试验，包括 128 个晶片数据。其中包括 107 个正常晶片数据和 21 个故障晶片数据。在 107 个正常晶片数据中随机选择 6 个和 21 个故障晶片组成校验集，剩下的 101 个正常晶片构成训练集[7]，本文只使用其中的 38 个变量，这些变量与产品生产过程的最终状态密切相关。

4.2　仿真实验结果

首先将半导体数据进行标准化处理后，再对这些数据运用各种算法。本文只对比 PCA、FD-KNN 以及 LNNS-FD-KNN 三种算法。实验结果如图 6～8 所示。

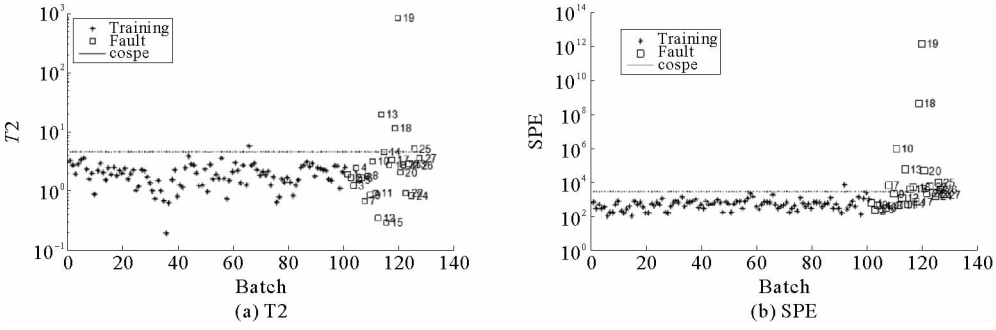

图 6　PCA 检测结果分析

Fig. 6　Analysis of PCA test results

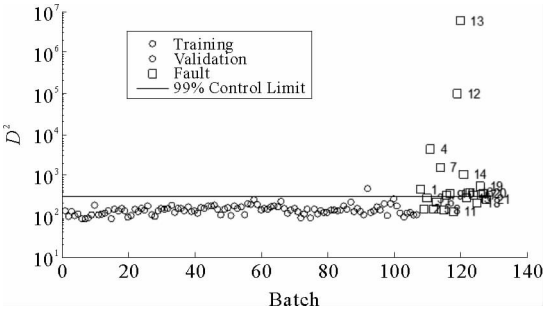

图 7　FD-KNN 检测结果

Fig. 7　FD-KNN test results

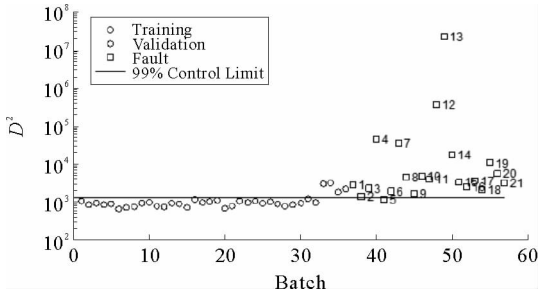

图 8　LNNS-FD-KNN 检测结果

Fig. 8　LNNS-FD-KNN test results

从图 7 可以发现，通过 FD-KNN 方法虽然可以检测出很多故障点，但是仍检测不出 3，6，2，5，8，11，15，18，21 故障点。

各种方法检测出的故障结果如表 1 所示。由于数据来自 3 个不同试验，所以本例是一个多模态的问题。首先是使用 PCA 方法，选择 3 个主元建立主元模型，21 批次的故障检测出 4 批次；使用 FD-KNN，检测出 9 批次；而用本文提出的 LNNS-FD-KNN 方法，只有 2 批次检测不到，大大提高了检

表 1　三种方法故障数目检测对比

Tab. 1　Comparison of three methods for fault detection

检测方法	未被检测出的故障点数目
PCA	16
FD-KNN	12
LNNS-FD-KNN	2

测准确率。主要原因是采用近邻标准化方法,使得样本间距离计算的更为精确。

5 结论

针对多模态非线性间歇过程,采用局部近邻的均值和标准差来替换全样本的均值和标准差对数据进行标准化处理,然后利用 FD-KNN 处理多模态问题来确定控制限,对半导体批次过程进行仿真研究,并与 PCA 和 FD-KNN 两种检测方法进行对比,验证所提方法具有更好的故障检测能力。

参考文献:

[1]周东华,李钢,李元. 数据驱动的工业过程故障诊断技术[M]. 北京:科学出版社,2011.

[2]KANO M,MIYAZAKI K,HASEBE S, et al, Inferential control system of distillation compositions using dynamic partial least squares regression[J]. Journal of Process Control,1998,10(2-3):157-166.

[3]MARTIN E, MORRIS A. Non-parametric confidence bounds for process performance monitoring charts[J]. Journal of Process Control,1996,6(6) 349-358.

[4]DONG D,MCAVOY T J. Nonlinear principal component analysis:Based on principal curves and neural networks[C//American Control Conference,IEEE,1994(2):1284-1288.

[5]DONG D,MCAVOY T J. Batch tracking via nonlinear principal component analysis[J]. AIChE Journal,1996b,42(8):2199-2208.

[6]QIN S J,VALLE S,PIOVOSO M. On unifying multi-block analysis with applications to decentralized process monitoring [J]. Journal of Chemometrics,2001,J15(15):715-742.

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

[8]WANG G Z,LIU J C,LI Y. Fault diagnosis using kNN reconstruction on MRI variables[J]. Journal of Chemometrics, 2015,29(7):399-410.

[9]WANG G Z,LIU J C,LI Y,et al. Fault detection based on diffusion maps and k-nearest neighbor diffusion distance of feature space[J]. Journal of Chemical Engineering of Japan,2015,48(9):756-765.

[10]HE X,WANG Z,LIU Y,et al. Least-squares fault detection and diagnosis for networked sensing systems using a direct state estimation approach[J]. IEEE Transactions on Industrial Informatics,2013,9(3):1670-1679.

[11]HE X,WANG Z,WANG X,et al. Networked strong tracking filtering with multiple packet dropouts:Algorithms and applications[J]. IEEE Transactions on Industrial Electronics,2013,61(3):1454-1463.

[12]郭小萍,姜芹芹,李元. 近邻标准化样本核特征量驱动的间歇过程故障检测[J]. 计算机与应用化学,2014,31(10):1157-1161.

GUO Xiaoping,JIANG Qinqin,LI Yuan. Local nearest neighborhood standardization sample about sample nuclear drive characteristics of intermittent process fault detection[J]. Computer and Applied Chemistry,2014,31(10):1157-1161.

[13]YU J,QIN S J. Multimode process monitoring with Bayesian inference:Based finite Gaussian mixture models,Aiche Journal,2008,54 (7):1811-1829.

[14]LEE J,KANG B,KANG S H. Integrating independent component analysis and local outlier factor for plant-wide process monitoring[J]. Journal of Process Control,2011,21(7):1011-1021.

(责任编辑:傅 游)