控制与决策
Control and Decision

基于LNS-DEWKECA算法的多模态工业过程故障检测

顾幸生， 周冰倩

引用本文:
顾幸生, 周冰倩. 基于LNS-DEWKECA算法的多模态工业过程故障检测[J]. 控制与决策, 2020, 35(8): 1879-1886.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2018.1712

文章编号: 1001-0920(2020)08-1879-08 DOI: 10.13195/j.kzyjc.2018.1712

基于LNS-DEWKECA算法的多模态工业过程故障检测

顾幸生†, 周冰倩

(华东理工大学 化工过程先进控制和优化技术教育部重点实验室，上海 200237)

摘 要: 受市场需求主导,工业过程需要在多种工作模态下切换,数据往往呈现多模态复杂分布特性,研究多模态的故障检测技术对于保障工业过程的安全运行具有重要意义. 为此,提出一种基于局部近邻标准化(LNS)和方向熵加权核熵成分分析(DEWKECA)的故障检测算法. 利用LNS实现多模态数据的标准化,相比于全局标准化,LNS可以有效消除多模态特性;考虑到故障样本与正常样本在变化趋势上的差异,定义样本变化方向的信息熵为方向熵,用来衡量样本变化方向的无序程度,从而利用DEWKECA实现数据降维,可以更有效提取数据变化方向特征;考虑到多模态数据往往服从非高斯分布,采用局部离群因子(LOF)算法建立监控统计量,根据核密度估计确定其控制限. 最后,通过数值例子及TE过程仿真验证所提出算法的有效性.

关键词: 多模态；故障检测；局部近邻标准化；方向熵；核熵成分分析；局部离群因子

中图分类号: TP277 **文献标志码:** A

Multimodal industrial process fault detection based on LNS-DEWKECA

GU Xing-sheng†, ZHOU Bing-qian

(Key Laboratory of Advanced Control and Optimization for Chemical Process of the Ministry of Education，East China University of Science and Technology，Shanghai 200237，China)

Abstract: Driven by the market demand, industrial processes need to switch between multiple modes, resulting in multimodal and complex distributed data. Therefore, the study of multimodal fault detection technology is of great significance to ensure safe operation of industrial processes. Aiming at the fact, a fault detection algorithm based on local neighborhood standardization (LNS) and directional entropy weighted kernel entropy component analysis (DEWKECA) is proposed. Firstly, multimodal data are normalized with LNS, which can eliminate multimodal characteristics efficiently when compared with global standardization. Considering the differences in changing trends between fault samples and normal samples, the information entropy of samples' changing direction is defined as the directional entropy, which is used to describe the disorder degree of samples' changing direction. So DEWKECA is used to realize dimensionality reduction, with which the characteristics of data's changing direction can be extracted effectively. Given that multimodal data often obey non-Gaussian distribution, the local outlier factor (LOF) algorithm is adopted to establish monitoring statistics, whose control limit can be determined by kernel density estimation. Numerical example and the TE process simulation verify the effectiveness of the proposed algorithm.

Keywords: multimode；fault detection；local neighborhood standardization；directional entropy；kernel entropy component analysis；local outlier factor

0 引 言

随着大数据、人工智能等技术的飞速发展以及信息化与自动化的高度融合,工业过程日益复杂化、大型化和集成化,故障发生的概率随之增加. 而复杂系统各环节往往紧密相连,即使发生很小的故障也可能带来一系列危害,所以及时有效地检测到故障的发生对于保障企业安全生产和人民人身安全具有重要意义[1]. 实际工业过程中,企业往往需要根据市场需求在多个操作模态之间切换以获得最大效益,由于传统多元统计方法如主元分析(principal component analysis, PCA)[2]、偏最小二乘(partial least squares, PLS)[3]等都是以过程数据服从单一操作模态为前提,故对于多模态工业过程的故障检测效果较差.

监测多模态过程可通过建立全局模型或者多个局部模型来实现. Jiang等[4]采用高斯混合模型

收稿日期: 2018-12-17；**修回日期:** 2019-04-26.

基金项目: 国家自然科学基金项目(61573144, 61773165, 61773165).

责任编委: 方华京.

†通讯作者. E-mail: xsgu@ecust.edu.cn.

(Gaussian mixture model, GMM)来实现多模态过程的故障检测; Raveendran等[5]采用混合概率PCA和贝叶斯期望最大化的策略,建立了多模态过程的监控模型; Zhang等[6]采用基于模态子空间和残差子空间的贝叶斯监测策略监测多模态工业过程; Song等[7]采用递归局部离群因子算法划分过程为稳定模态和转换模态,分别建立相应模型,然后进行故障检测. 上述方法对于划分模块的精度要求较高,划分结果直接影响到后续的故障检测效果,且一般需要过程先验知识的支持,所以针对全局模型的多模态故障检测方法也有不少学者研究. Kano等[8]采用外部分析(external analysis, EA)消除了运行模式变化对过程数据的影响,继而对消除影响后的数据进行故障检测; Yu等[9]采用即时学习(just in time learning, JITL)的方法,建立局部输入输出模型,通过监测模型输出与实际输出的残差来判断是否产生故障. 二者都是基于建模来消除工况变化对过程数据产生的影响,建模精度对后续故障检测效果影响较大. Ma等[10]根据不同模态数据均值与方差往往不同的特点,采用局部近邻标准化(local neighborhood standardization, LNS)策略,解决了全局标准化后仍会保留数据多模态特性的问题.

消除工况变化带来的影响后,采用单模态工业过程故障检测算法对多模态工业过程进行故障检测. 基于核主元分析(kernel principal component analysis, KPCA)[11], Jenssen教授提出了核熵成分分析(kernel entropy component analysis, KECA)[12],相比KPCA而言, KECA在数据结构特征提取上具有一定的优势. KECA以信息熵损失最小为原则实现数据降维,降维后的数据具有一定的角结构特征[13]. 过程监控依据对处理后的数据建立的监控统计量来实现,由于实际工业过程难以满足过程变量服从高斯分布的条件,传统T^2统计量和SPE统计量故障检测效果受限. 局部离群因子(local outlier factor, LOF)算法[14]根据样本密度大小来判断该样本是否是离群点,对数据分布没有要求. 由于故障样本相对于正常样本即为离群点,可以通过LOF算法建立监控统计量. Lee等[15]建立LOF统计量监测工业过程,监控性能较好.

关于多模态工业过程故障检测方法的研究,如何提高故障检测的精度、降低故障的误报率和漏报率仍是学者们关注的热点. 本文提出一种基于局部近邻标准化(local neighborhood standardization, LNS)和方向熵加权核熵成分分析(directional entropy weighted kernel entropy component analysis, DEWKECA)的故障检测算法. 首先,采用LNS对数据进行标准化处理,消除过程数据的多模态特征;然后,采用DEWKECA算法对数据进行降维和提取特征信息,采用KECA将数据映射到低维空间,考虑到样本数据的变化趋势也包含特征信息,提出方向熵,利用方向熵对低维样本数据进行加权,对于一些故障样本和正常样本分布较接近而变化趋势不同的情况,可以有效减少漏报现象;最后,对于DEWKECA算法处理后的数据,构建LOF统计量,利用核密度估计确定其控制限. 数值仿真例子和TE过程仿真验证了本文方法的故障检测效果.

1 方向熵加权核熵成分分析算法

由于多模态过程数据中各模态数据特性不同,相比全局样本,由样本邻居构成的样本邻域更能描述样本本身的特征信息,故LNS比全局标准化更适用于处理多模态数据. LNS的基本思想就是在处理训练集样本过程中,对于每一个训练集样本$x_i \in \boldsymbol{R}^{1\times m}$($m$为变量维数),从剩余训练集数据中选取$x_i$的$k$个邻居$x_i^1, x_i^2, \ldots, x_i^k$组成邻域$N(x_i) = [x_i^1, x_i^2, \ldots, x_i^k]^{\mathrm{T}}$,根据下式进行标准化处理[10]:

$$\bar{x}_{ij} = \frac{x_{ij} - m(N_j(x_i))}{s(N_j(x_i))}. \tag{1}$$

其中: $j = 1, 2, \ldots, m$, $\bar{x}_i = [\bar{x}_{i1}, \bar{x}_{i2}, \ldots, \bar{x}_{im}]$为标准化后的样本数据,邻域$N_j(x_i)$表示$N(x_i)$的第$j$维变量,$m(N_j(x_i))$和$s(N_j(x_i))$分别表示$N_j(x_i)$的均值和标准差. 对于新样本$x_{\mathrm{new}}$,为避免故障样本位于不同模态中间时所找到的样本邻居来自不同模态,采用x_{new}的第1个邻居的邻域均值和标准差来对样本进行标准化处理[16],即

$$\bar{x}_{\mathrm{new},j} = \frac{x_{\mathrm{new},j} - m(N_j(x_{\mathrm{new}}^1))}{s(N_j(x_{\mathrm{new}}^1))}. \tag{2}$$

其中: $\bar{x}_{\mathrm{new}} = [\bar{x}_{\mathrm{new},1}, \bar{x}_{\mathrm{new},2}, \ldots, \bar{x}_{\mathrm{new},m}]$为标准化后的样本数据, $m(N_j(x_{\mathrm{new}}^1))$和$s(N_j(x_{\mathrm{new}}^1))$分别表示样本$x_{\mathrm{new}}$第1个样本邻居$x_{\mathrm{new}}^1$的邻域$N(x_{\mathrm{new}}^1)$的第$j$维向量的均值和标准差.

实际工业过程中,故障信息在样本数据的变化趋势中也会有所体现. 一般来说,正常样本在某个区域内随机变化,变化方向是无序的,不确定性程度高;发生故障时,某些受故障影响较大的变量将呈现一定趋势的变化,不确定性程度发生改变,所以衡量样本变化方向的不确定性也可以反映样本正常与否. 而KECA算法忽略了这一点,本文提出的DEWKECA算法在KECA算法基础上进一步提取样本的方向信息,使得故障样本更易被识别.

1.1 核熵成分分析

KECA算法降维的目的是在核特征空间保留数据均值矢量的最大欧式长度,不要求数据满足高斯分布,且依据熵贡献率的原则所确定的主元个数比KPCA更少. 下面介绍其基本理论[17].

瑞利熵是一种信息熵的指标,计算公式为

$$H(p) = -\log \int p^2(x)\mathrm{d}x, \tag{3}$$

其中$p(x)$为中心化后的样本x的概率密度函数. 根据对数函数的单调性,将式(3)转化为

$$V(p) = \int p^2(x)\mathrm{d}x, \tag{4}$$

即通过对$V(p)$的估计来实现瑞利熵$H(p)$的求取. 引入Parzen窗概率密度估计算子

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{K}_\sigma(x, x_i). \tag{5}$$

其中: N为样本数; $\boldsymbol{K}_\sigma(x, x_i)$为Parzen窗,同时也是中心为$x_i$、宽为$\sigma$的核函数,这里的核函数常采用径向基函数,即

$$\boldsymbol{K}_\sigma(x, x_i) = \exp\Big(-\frac{\|x - x_i\|^2}{2\sigma^2}\Big). \tag{6}$$

将式(5)代入(4),利用样本均值对样本期望近似估计,并经过一系列变换,得

$$\hat{V}(p) = \frac{1}{N}\sum_{i=1}^{N}\hat{p}(x_i) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{K}_\sigma(x_i, x_j) = \frac{1}{N^2}\mathbf{1}^{\mathrm{T}}\boldsymbol{K}\mathbf{1}. \tag{7}$$

其中: $\boldsymbol{K}$为$N \times N$的核矩阵,其元素$\boldsymbol{K}_{i,j} = \boldsymbol{K}_{\sqrt{2}\sigma}(x_i, x_j)$, $\mathbf{1}$为元素均为1的$N \times 1$维向量. 式(7)实现了瑞利熵的核矩阵表达,则可通过核矩阵的特征值和特征向量来表示瑞利熵的估计值. 对核矩阵进行分解,有

$$\boldsymbol{K} = \boldsymbol{E}D\boldsymbol{E}^{\mathrm{T}}. \tag{8}$$

其中: $\boldsymbol{E} = [e_1, e_2, \ldots, e_N]$为特征向量矩阵, $\boldsymbol{D} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$为特征值矩阵. 此时式(7)转化为

$$\hat{V}(p) = \frac{1}{N^2}\sum_{i=1}^{N}(\sqrt{\lambda_i}e_i^{\mathrm{T}}\mathbf{1})^2. \tag{9}$$

式(9)表明,特征值大并不能保证对瑞利熵的贡献也大,只有$\xi_i = \sqrt{\lambda_i}e_i^{\mathrm{T}}\mathbf{1}$大才能保证对瑞利熵的贡献率也大. 将$\xi_i$从大到小排序为$\xi_1, \xi_2, \ldots, \xi_N$,选择最大的前$d$项对应的特征向量构成投影空间,标准化后化得到投影矩阵$\boldsymbol{P} = [e_1, e_2, \ldots, e_d]$,则降维后的样本矩阵为$\boldsymbol{Y} = \boldsymbol{K}\boldsymbol{P}$,相应的特征值向量为$\Lambda = [\lambda_1, \lambda_2, \ldots, \lambda_d]$.

1.2 方向熵

熵是信息论中用来表征变量不确定程度的指标[18]. 本文利用熵来衡量样本变化趋势的无序程度,熵值越大表明样本数据变化方向越无序,熵值越小表示样本变化方向越有序. 正常样本在每一维变量上都呈现随机变化的趋势,因而总体无序程度较高,熵值大;而一旦发生故障,受故障影响较大的变量将呈现一定趋势的变化,且变化幅度大于受故障影响小的变量. 从整体上看,发生故障时样本呈一定方向的变化,无序程度减小,熵值减小.

对于离散随机变量$\boldsymbol{X} = \{x_1, x_2, \ldots, x_n\}^{\mathrm{T}}$,其香农熵定义为

$$H = -\sum_{i=1}^{n} p(x_i)\log p(x_i), \tag{10}$$

其中$p(x_i)$表示x_i处的概率密度函数.

KECA消除了原始数据变量间的冗余,所以计算降维后矩阵的方向熵更能反映数据变化趋势. 发生故障时,各主成分分量对于故障的贡献程度不同,主成分分量所占比例越大,所包含故障信息越多,对故障的贡献程度越大;反之,主成分分量所占比例越小,包含故障信息越少,对故障的贡献程度也就越小. 为凸显对故障贡献程度较大的主成分分量y_i,首先以特征值λ_i对y_i进行加权,即

$$z_i = y_i \times \sqrt{\lambda_i},\ i = 1, 2, \ldots, n. \tag{11}$$

其中: λ_i为对应于y_i的特征值, $\boldsymbol{Z} = [z_1, z_2, \ldots, z_n]^{\mathrm{T}} \in \boldsymbol{R}^{n\times d}$为加权后的矩阵. 用差分向量$z_i - z_{i-a}(0 < a < i)$表示$z_i$的变化方向,以此类推,得到$(n-a)$个差分向量$z_{a+1} - z_1, z_{a+2} - z_2, \ldots, z_i - z_{i-a}, \ldots, z_n - z_{n-a}$,记为$\Delta z_i, i = a+1, a+2, \ldots, n$. 其中a为常数,表示两样本数据之间的间隔. 采用余弦距离描述Δz_i在方向上相对于Δz_{i-1}的变化,则Δz_i与Δz_{i-1}之间的余弦距离为

$$d_i = \cos(\Delta z_i, \Delta z_{i-1}) = \frac{\langle \Delta z_i, \Delta z_{i-1}\rangle}{\|\Delta z_i\|\|\Delta z_{i-1}\|}. \tag{12}$$

本文引入滑动窗口策略,一方面考虑到信息熵是一种累加求和的形式,另一方面若样本间隔大,则不管故障发生与否,变化方向差异都很大,无序程度总是很高,比较熵值没有意义. 设窗宽为w,可通过计算$\{d_{i-w+1}, d_{i-w+2}, \ldots, d_i\}$的信息熵$H_i$衡量样本$z_i$变化方向的无序程度,将$H_i$称为样本$z_i$的方向熵,不断滑动窗口,可得到数据矩阵$\boldsymbol{Z}$的方向熵矢量$\boldsymbol{H}$. 由于故障样本的方向熵值较正常样本的方向熵值减小,而后续判断工业过程是否发生故障是观察监测统计量是否超过相应控制限,若监测统计量增大至超过控制

限即视为发生故障,这与发生故障时方向熵值减小是相违背的. 为统一发生故障时二者的变化趋势,考虑以方向熵的倒数进行加权,同时为使加权值分布于1附近,将方向熵倒数乘上方向熵矢量的均值,则变换后的方向熵值W_i为

$$W_i=\begin{cases}\dfrac{m(\boldsymbol{H})}{H_i},\ i>w+a;\\ 1,\ \text{otherwise.}\end{cases}\tag{13}$$

其中$m(\boldsymbol{H})$表示$\boldsymbol{H}$的均值. 主成分矩阵$\boldsymbol{Y}$经方向熵加权后为

$$\boldsymbol{Q}=\boldsymbol{W}Y.\tag{14}$$

其中:$\boldsymbol{W}=\text{diag}(W_1,W_2,\dots,W_n)$为变换后的方向熵矩阵,$\boldsymbol{Q}$为方向熵加权后的主成分矩阵.

2 局部离群因子

LOF是由Breunig等[19]提出的一种基于密度的离群点检测算法,用样本的局部离群因子衡量样本自身的离群程度. 假设存在样本集$\boldsymbol{X}\in\boldsymbol{R}^{n\times m}$,其中$n$为样本个数,$m$为样本维数,则LOF的基本思想是对于样本$x_i$,首先在剩余样本中找出该样本的$K$个邻居,构成样本$x_i$的邻域$N(x_i)=\{x_i^1,x_i^2,\dots,x_i^f,\dots,x_i^K\},f=1,2,\dots,K$,定义$x_i$与邻域中最远邻居的欧氏距离为$k_\text{dist}(x_i)$,$x_i$与其第$f$邻居之间的欧氏距离为$d(x_i,x_i^f)$,则$x_i$的局部离群因子计算步骤[20]如下:

计算样本x_i与其邻居x_i^f之间的可达距离

$$\text{reach\_dist}(x_i,x_i^f)=\max\{k\_\text{dist}(x_i^f),d(x_i,x_i^f)\},\tag{15}$$

则样本x_i的局部可达密度为

$$\text{lrd}(x_i)=\frac{K}{\sum\limits_{f=1}^{K}\text{reach\_dist}(x_i,x_i^f)},\tag{16}$$

从而根据局部可达密度计算出样本x_i的局部离群因子

$$\text{LOF}(x_i)=\frac{1}{K}\sum_{f=1}^{K}\frac{\text{lrd}(x_i^f)}{\text{lrd}(x_i)},\tag{17}$$

其中$\text{lrd}(x_i^f)$表示x_i^f的局部可达密度. 样本x_i的局部离群因子即定义为其邻域内所有样本的平均局部可达密度与其自身局部可达密度的比值. 当该值与1接近时,说明样本自身局部可达密度与其邻居局部可达密度值相近,该样本不是离群点;当该值大于1时,说明样本自身局部可达密度小于其邻居局部可达密度值,判定为离群点.

3 故障检测步骤

总结本文所提出的多模态工业过程的故障检测算法,分为离线建模和在线检测两个步骤,分别如下.

step 1: 离线建模.

step 1.1: 采集多模态正常样本数据组成训练集$\boldsymbol{X}\in\boldsymbol{R}^{n\times m}$. 对于训练集中每个样本,确定其$k$邻域,按式(1)进行标准化处理.

step 1.2: 对于标准化后的样本数据,采用DEWKECA算法提取特征. 首先利用KECA算法进行数据降维和特征提取,得到降维矩阵$\boldsymbol{Y}$和投影矩阵$\boldsymbol{P}$;然后根据式(14)计算得到方向熵加权后矩阵$\boldsymbol{Q}$.

step 1.3: 对于矩阵$\boldsymbol{Q}$,采用LOF算法计算每个样本点的LOF值,从而建立监控统计量,采用核密度方法确定监控统计量的控制限LOF_limit.

step 2: 在线检测.

step 2.1: 采集当前数据样本x_{new},为x_{new}在训练集$\boldsymbol{X}$内寻找最近邻x_{new}^1,接着在训练集的剩余样本中寻找x_{new}^1的k个最近邻组成k邻域,并采用式(2)对当前样本进行标准化处理.

step 2.2: 对于标准化后的样本数据,采用DEWKECA算法进行处理. 首先采用投影矩阵$\boldsymbol{P}$进行降维,得到y_{new},则可求出特征值加权后的矩阵,继而得到差分向量$\Delta z_{\text{new}}=z_{\text{new}}-z_{\text{new}-a}$和当前样本余弦距离$d_{\text{new}}=\cos(\Delta z_{\text{new}},\Delta z_{\text{new}-1})$,根据$\{d_{\text{new}-w+1},d_{\text{new}-w+2},\dots,d_{\text{new}}\}$可求出当前样本对应方向熵$H_{\text{new}}$,继而求出变换后方向熵$q_{\text{new}}=\dfrac{m(\boldsymbol{H})}{H_{\text{new}}}y_{\text{new}}$.

step 2.3: 对于q_{new},为其在矩阵$\boldsymbol{Q}$中寻找邻居样本构成样本邻域,计算LOF统计量.

step 2.4: 根据控制限LOF_limit判断当前样本正常与否,若LOF值超过LOF_limit,则为故障样本,反之则为正常样本,回到step2.1中继续进行在线检测.

4 仿 真

4.1 数值例子

本文采用文献[21]提出的数值仿真例子进行测试,具体结构为

$$\begin{cases}x_1=0.576\,8s_1+0.376\,6s_2+e_1,\\ x_2=0.738\,2s_1^2+0.056\,6s_2+e_2,\\ x_3=0.829\,1s_1+0.400\,9s_2^2+e_3,\\ x_4=0.651\,9s_1s_2+0.207\,0s_2+e_4,\\ x_5=0.397\,2s_1+0.804\,5s_2+e_5.\end{cases}\tag{18}$$

其中包括5个变量x_1、x_2、x_3、x_4、x_5和5个相互独立且服从$N(0,0.01)$的白噪声e_1、e_2、e_3、e_4、e_5. N表示

高斯分布. 根据s_1、s_2的两种不同分布,模型有两种不同的模态,分别为

mode 1　$s_1: U(-10,-7), s_2: N(-15,1)$;

mode 2　$s_1: U(2,5), s_2: N(7,1)$.

其中U表示均匀分布. 以mode 1和mode 2下的各400组正常样本组成训练集. 测试集中前400个样本为正常数据,后400个样本为故障数据,设定:

fault 1: 系统运行在mode 1下,在第401个样本给x_5加上幅值为4的阶跃信号;

fault 2: 系统运行在mode 2下,在第401个样本给x_1加上$0.02(i-400)$的斜坡信号;

fault 3: 系统运行在mode 2下,在第401个样本给x_1加上幅度、偏移和频率均为1的正弦信号.

本例中在对数据进行LNS处理后,分别使用KECA-T^2、KECA-LOF、DEWKECA-T^2和DEWKECA-LOF进行故障检测. 为公平比较,统一设定主元数为4. 样本近邻数k值的确定要兼顾离群点对数据的影响和各模态下的样本数,若k值选取过小,则离群点对标准化的结果影响较大;若k值选取过大,则可能选取的邻居中包含与当前样本不属于同一模态的样本数据. 根据测试经验,本文LNS和LOF中k取值为50;核函数带宽σ取值为10 000;T^2和LOF统计量置信度均设为0.99. 计算方向熵时,参数a的选取如果过小,则样本方向变化的无序程度受离群点的干扰大;如果选取过大,则用差分向量衡量样本的变化方向没有意义. 兼顾二者,本例中选取$a=6$. 同样的,如果窗宽w设置过大,则检测出样本方向产生变化的延迟增大;如果过小,方向熵值在样本数据正常和故障时变化不大,本例中选择w为20.

采集mode 1和mode 2下各400个正常数据组成正常数据集. 图1为4种方法对于正常数据的检测结果,除少量误报现象外,4种方法均未检测出故障,说明了各方法的有效性. 表1列出了4种方法对于3种不同故障的检测率. fault 1中引入了阶跃信号,由于故障幅度较大,4种方法均可以及时有效地检测到故障的发生; fault 2中引入了斜坡信号,幅度随时间推移缓慢增加,在达到一定程度后,4种方法同样可以较好地检测到故障; fault 3中引入了正弦信号,故障幅度小,4种方法对于fault 3的检测效果如图1所示. 结合图1和表1可以看出,KECA算法的漏报率较高,而采用DEWKECA算法提取了数据变化方向信息后,漏报率降低,T^2统计量的故障检测率从66%提升到80.25%,LOF统计量的故障检测率从70%提升到99.2%,体现了方向熵加权的优势.

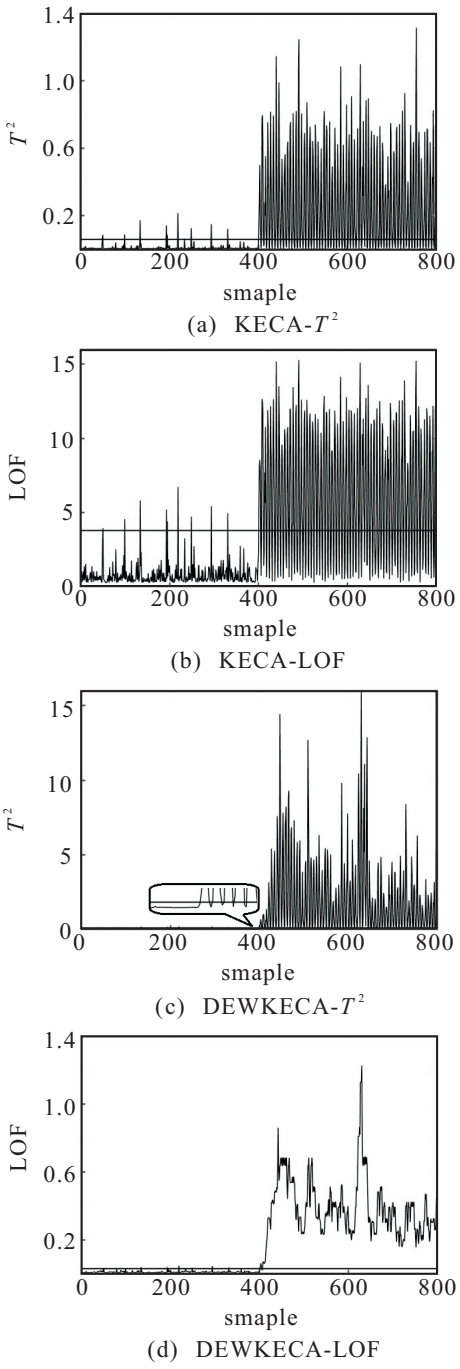


图1　fault 3检测结果

表1　4种方法的故障检测率

fault	KECA-T^2	KECA-LOF	DEWKECA-T^2	DEWKECA-LOF
1	100	100	100	100
2	92.75	92	93.75	94
3	66	70	80.25	99.2

以fault 3为例,图2通过熵值变化反映方向变化的无序程度,从第401个样本开始,熵值明显减小,说明故障样本变化趋势的无序程度降低. 对比KECA降维和DEWKECA降维后,第1主元PC1和第2主元PC 2分布散点图如图3所示. 由图3(a)可以看出,不少故障样本和正常样本集分布较近,故KECA算法漏报率高. 仔细观察向量分布范围,正常样本的PC1和

PC2分别大致分布在[0.99, 1]和[−0.05, 0.1]内,区间长度相差不大,说明变化趋势在PC1和PC2上均有体现,且相差较小;而故障样本的PC1和PC2分别分布在[0.975, 1]和[−0.05, 0.1]内,PC1的分布区间长度明显大于PC2的分布区间长度,说明变化趋势主要体现在PC1上,且由于PC1对应的特征值较PC2更大,经过特征值加权后,该变化趋势进一步得到加强. 在计算相邻样本变化方向的夹角时,对于正常样本,夹角是在0° ∼ 180°随机变化,无序程度高;而故障样本在PC1上的变化幅度远大于其他主元上的变化幅度,夹角接近0°或者180°,无序程度降低. 利用信息熵对降维数据加权后,前两个主元分布如图3(b)所示,故障样本和正常样本的可区分性增强,说明了所提出算法的有效性.

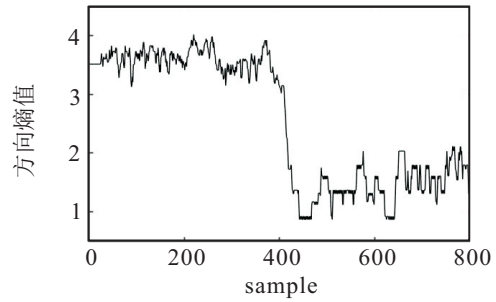


图 2 不同窗口数据的方向熵值

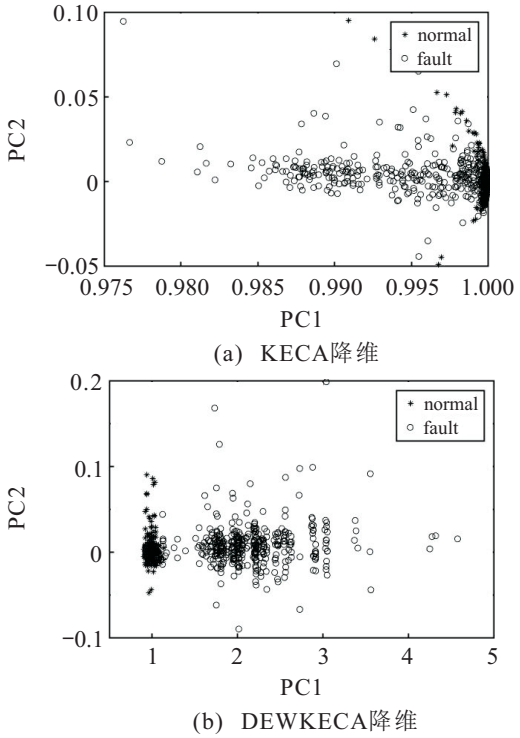


图 3 PC1和PC2散点图

4.2 TE过程仿真

Tennessee-Eastman(TE)过程是由Downs和Vogel提出的一个标准实验仿真平台,由反应器、冷凝器、气液分离塔、气提塔和压缩机5个单元构成,共包括12个操作变量和41个过程变量,以及6种不同的运行模态[22]. 本文采用http://depts.washington.edu/control/LARRY/TE/down-load.html提供的TE仿真平台采集得到mode 1和mode 2的过程数据,以除仿真过程中不变的操作变量5、9、12外的剩余9个操作变量和22个过程变量构造数据集. TE过程设有21种故障,其中故障3、9、15中过程观测变量的统计量变化不明显,检测效果较差[23],故本文不考虑这3种故障,对剩余12种已知故障和4种未知故障(故障17 ∼ 故障20),合计共16种故障进行分析.

训练集包含mode 1和mode 2下的500个正常样本;测试集包含1 000个样本,从第501个样本开始为故障样本,分别使用KECA-LOF、DEWKECA-T^2、KECA-T^2和DEWKECA-LOF进行故障检测. 为方便比较,所有主元数均设置为10. 由于对过程变量影响较大的故障采用KECA算法即可有效检测,参数a和滑动窗宽w的设定不考虑那些故障. 本例中主要综合考虑两种模态下的故障10、故障12、故障17、故障18,以平均故障检测率作为确定二者的标准. 依据4.1节中的说明,将参数a和w分别设置在$[1, 10]$和$[10, 30]$范围内. 图4为DEWKECA-LOF在不同的a和w下4种故障的平均检测率. 如图4所示,窗宽尺寸w大于20后故障检测率的增长趋势趋于平缓. 综合考虑计算量、故障检测延迟和故障检测率等因素,设置w为20,此时a为6的平均检测率最高,故设置参数$a = 6$.

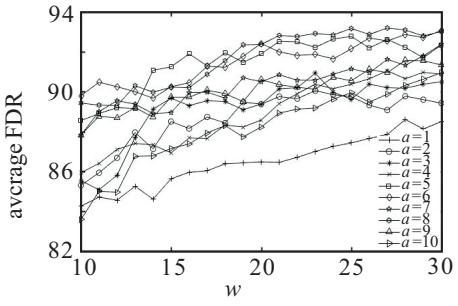
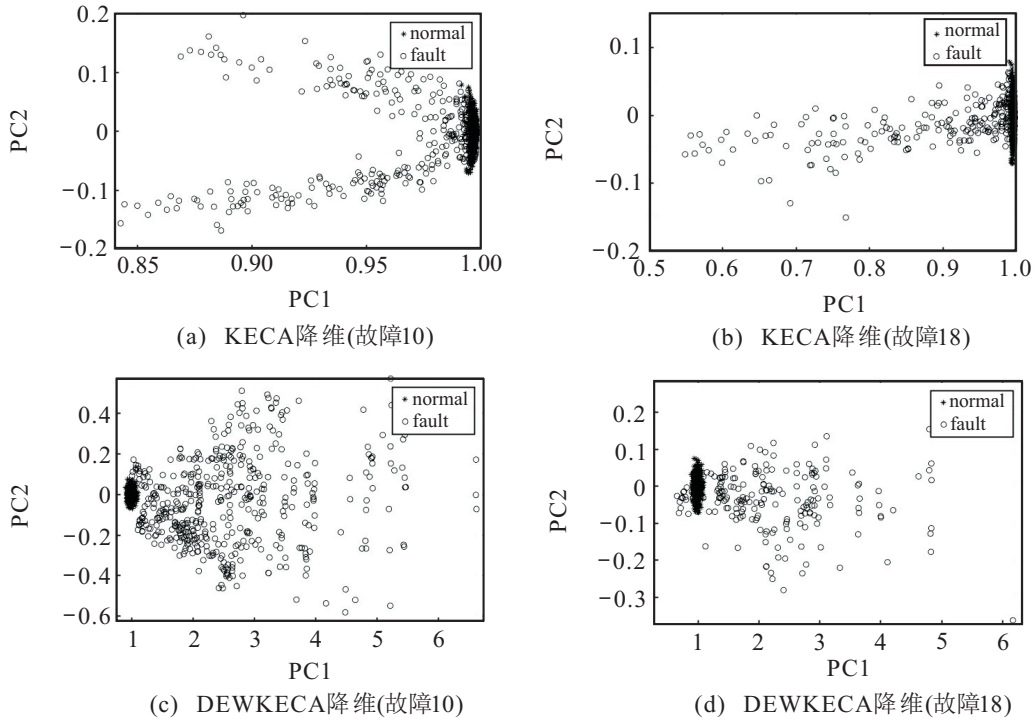


图 4 不同a和w下4种故障的平均检测率

表2中记录了4种方法分别在mode 1和mode 2下的故障检测率. T^2统计量和LOF统计量在KECA算法下和DEWKECA算法下的平均故障检测率分别为77.81 %和84.33 %, 80.9 %和95.31 %,可以看出LOF统计量故障检测效果明显高于T^2统计量,说明其对于复杂多模态数据具有更好的适用性. 对于影响过程较大的故障,方向熵加权并不能体现其优势,而对于某些故障样本与正常样本难以区分的情况,采用方向熵加权可以增加故障样本的可区分性. 以

表2　4种不同方法应用于TE过程mode 1和mode 2的故障检测率

故障	mode 1				mode 2			
	KECA-T^2	KECA-LOF	DEWKECA-T^2	DEWKECA-LOF	KECA-T^2	KECA-LOF	DEWKECA-T^2	DEWKECA-LOF
1	92.6	99.8	99	100	83.8	99.8	90	99.8
2	99.4	99.4	99.2	99.2	95.4	95.4	97.6	95.4
4	100	100	100	100	100	100	100	100
5	0	0.2	0	1.6	38.4	100	97	100
6	100	100	100	100	100	100	100	100
7	50.8	100	99.8	100	43.4	100	98.2	100
8	98.4	98.8	98.2	98	73	99	76.4	99
10	63.2	68	94.2	97.2	87	85.6	97	97
11	96	96	99	98.8	93.6	93.4	100	99.8
12	33.4	34.6	44.4	59.4	63.2	99.4	75	99.4
13	87.8	97.8	92.6	97.6	82.8	84	89	89.8
14	99.6	99.8	99.8	99.8	99.6	99	99.2	99.2
17	61.6	97.2	69.4	99.2	59.4	93.2	69.6	95.6
18	69	64.8	94.4	94.2	94	94	94.4	94.4
19	95.6	95.2	99.2	99	95.8	99.4	98.8	99.2
20	97.6	97.6	97.6	97.6	84.8	82.8	97.2	97
average	77.81	84.33	86.88	90.1	80.9	95.31	92.46	97.79



(a) KECA降维(故障10)　(b) KECA降维(故障18)

(c) DEWKECA降维(故障10)　(d) DEWKECA降维(故障18)

图5　故障10((a)(c))和故障18((b)(d))下的PC1和PC2散点图

mode 1下的故障10和故障18为例, 经KECA算法降维后前两个主元分布散点如图5(a)和图5(b) 所示,其中均存在不少故障样本和正常样本难以区分,直接采用KECA算法的检测效果不好,而采用方向熵加权后PC1和PC2分布散点如图5(c)和图5(d)所示,故障样本的可区分性明显增强,LOF统计量下故障10和故障18的检测率均达到90 %以上,说明DEWKECA 算法对于故障检测率的提升是有效的.

5　结　论

本文提出了一种基于LNS-DEWKECA算法的多模态故障检测算法. 该算法不仅能够有效提取数据分布特征,而且所构造的方向熵也能够反映样本数据的变化趋势信息,有效降低故障样本的漏检率;采用LOF作为监控统计量,更适用于复杂多模态数据,能够取得较好的故障检测效果. 首先,利用LNS对多模态数据进行标准化处理,消除多模态特性;然后,

采用KECA算法将原始数据降维,继而在低维空间计算每个样本的方向熵,利用方向熵加权相应样本数据,增强故障样本和正常样本的可区分性;最后,建立LOF统计量监测工业过程,提升故障检测能力. 通过数值例子和TE过程仿真说明了所提出方法在多模态工业过程故障检测中的有效性和优越性.

参考文献(References)

[1] 刘强, 柴天佑, 秦泗钊, 等. 基于数据和知识的工业过程监视及故障诊断综述[J]. 控制与决策, 2010, 25(6): 801-807.
(Liu Q, Chai T Q, Qin S Z, et al. Progress of data-driven and knowledge-driven process monitoring and fault diagnosis for industry process[J]. Control and Decision, 2010, 25(6): 801-807.)

[2] Dong Y N, Qin S J. A novel dynamic PCA algorithm for dynamic data modeling and process monitoring[J]. Journal of Process Control, 2018, 67: 1-11.

[3] Dong J, Zhang K, Huang Y, et al. Adaptive total PLS based quality-relevant process monitoring with application to the Tennessee Eastman process[J]. Neurocomputing, 2015, 154: 77-85.

[4] Jiang Q C, Huang B, Yan X F. GMM and optimal principal components-based Bayesian method for multimode fault diagnosis[J]. Computers & Chemical Engineering, 2016, 84(1): 338-349.

[5] Raveendran R, Huang B. Mixture probabilistic PCA for process monitoring-collapsed variational bayesian approach[J]. Proc of the 11th IFAC Symposium on Dynamics and Control of Process Systems. Trondheim Elsevier, 2016, 49(7): 1032-1037.

[6] Zhang S M, Zhao C H. Stationarity test and Bayesian monitoring strategy for fault detection in nonlinear multimode processes[J]. Chemometrics and Intelligent Laboratory Systems, 2017, 168: 45-61.

[7] Song B, Tan S, Shi H B. Key principal components with recursive local outlier factor for multimode chemical process monitoring[J]. Journal of Process Control, 2016, 47: 136-149.

[8] Kano M, Hasebe S, Hashimoto I, et al. Evolution of multivariate statistical process control: Application of independent component analysis and external analysis[J]. Computers and Chemical Engineering, 2004, 28(6/7): 1157-1166.

[9] Yu H, Yin S, Luo H. Robust just-in-time learning approach and its application on fault detection[J]. Proc of 20th IFAC World Congress. Toulouse: Elsevier, 2017, 50(1): 15277-15282.

[10] Ma H H, Hu Y, Shi H B. Fault detection and identification based on the neighborhood standardized local outlier factor method[J]. Industrial & Engineering Chemistry Research, 2013, 52(6): 2389-2402.

[11] Cui P, Zhan C J, Yang Y P. Improved nonlinear process monitoring based on ensemble KPCA with local structure analysis[J]. Chemical Engineering Research and Design, 2019, 142: 355-368.

[12] Jenssen R. Kernel entropy component analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(5): 847-860.

[13] Qi Y S, Wang Y, Lu C X, et al. Improved batch process monitoring and diagnosis based on multiphase KECA[J]. Proc of 10th IFAC Symposium on AdvancedControl of Chemical Processes ADCHEM. Shengyang: Elsevier, 2018, 51(18): 827-832.

[14] Ding H X, Ding K, Zhan J W, et al. Local outlier factor-based fault detection and evaluation of photovoltaic system[J]. Solar Energy, 2018, 164: 139-148.

[15] Lee J, Kang B Y, 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.

[16] Ma H H, Hu Y, Shi H B. A novel local neighborhood standardization strategy and its application in fault detection of multimode processes[J]. Chemometrics and Intelligent Laboratory Systems, 2012, 118: 287-300.

[17] Jenssen R. Kernel entropy component analysis: New theory and semi-supervised learning[C]. 2011 IEEE International Workshop on Machine Learning for Signal Processing. Santander: IEEE, 2011: 1-6.

[18] Yuan Z, Zhang X Y, Feng S. Hybrid data-driven outlier detection based on neighborhood information entropy and its developmental measures[J]. Expert Systems with Applications, 2018, 112: 243-257.

[19] Breunig M M, Kriegel H, Ng R T, et al. LOF: Identifying density-based local outliers[C]. Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2000: 93-104.

[20] 王培良, 叶晓丰, 杨泽宇. 基于Block-RPLS模型自适应更新的质量预测方法[J]. 控制与决策, 2018, 33(3): 455-462.
(Wang P L, Ye X F, Yang Z Y. Quality prediction method based on adaptive updating of Block-RPLS model[J]. Control and Decision, 2018, 33(3): 455-462.)

[21] Ge Z Q, Song Z H. Multimode process monitoring based on Bayesian method[J]. Chemometrics and Intelligent Laboratory Systems, 2009, 23(12): 636-650.

[22] Downs J J, Vogel E F. A plant-wide industrial process control problem[J]. Computers and Chemical Engineering, 1993, 17(3): 245-255.

[23] Chiang L H, Russel R, Braatz R D. Fault detection and diagnosis in industrial systems[M]. London: Springer, 2001.

作者简介

顾幸生(1960−), 男, 教授, 博士生导师, 从事工业过程生产计划与调度、复杂工业过程建模等研究, E-mail: xsgu@ecust.edu.cn;

周冰倩(1995−), 女, 硕士生, 从事工业过程故障检测与诊断的研究, E-mail: zbq975236325@163.com.

(责任编辑: 孙艺红)