DOI:10.19651/j.cnki.emt.2416239

基于 CVDA 与 LLE 算法的工业过程故障检测*

蒋磊峰[1] 张 成[2] 李 元[1]
(1.沈阳化工大学信息工程学院 沈阳 110142;2.沈阳化工大学理学院 沈阳 110142)

摘 要:针对工业过程数据中存在的非线性、高维度的问题,提出一种结合典型变量差异度分析与局部线性嵌入算法的故障检测方法。CVDA 算法构建的差异度矩阵能实现有效的故障监测,但其依赖于线性投影,仅对数据结构中线性特征的变化敏感。使用 LLE 算法通过保持样本间的局部关系,将高维的数据映射到低维空间,进行特征的再次提取,进一步挖掘数据的非线性特征和局部邻域信息。最后在低维流形空间中建立隔离森林模型,将得到样本点的异常分数作为故障检测评价标准。通过一组非线性数值实例和 TE 化工过程数据,将本文所提方法与传统的 KPCA、PPA 以及 CVDA 进行对比分析,验证所提算法的有效性及优越性。

关键词:典型变量差异度分析;非线性过程;局部线性嵌入;隔离森林
中图分类号:TN081 **文献标识码**:A **国家标准学科分类代码**:510.80

Industrial process fault detection based on CVDA and LLE algorithms

Jiang Leifeng[1] Zhang Cheng[2] Li Yuan[1]
(1. College of Information Engineering, Shenyang University of Chemical Technology, Shenyang 110142, China;
2. College of Science, Shenyang University of Chemical Technology, Shenyang 110142, China)

Abstract: To address the issues of nonlinearity and high dimensionality in industrial process data, a fault detection method combining Canonical Variate Dissimilarity Analysis and Locally Linear Embedding is proposed. The dissimilarity matrix constructed by the CVDA algorithm can effectively monitor faults, but it relies on linear projections and is only sensitive to changes in linear features of the data structure. The LLE algorithm is used to map high-dimensional data to a low-dimensional space by preserving local relationships between samples, further extracting features and uncovering nonlinear characteristics and local neighborhood information. Finally, an isolation forest model is established in the low-dimensional manifold space to obtain anomaly scores of sample points as the fault detection evaluation criterion. Through a set of nonlinear numerical examples and the Tennessee Eastman chemical process data, the proposed method is compared and analyzed with traditional KPCA、PPA and CVDA to verify its effectiveness and superiority.

Keywords: canonical variate dissimilarity analysis;nonlinear process;LLE;isolation forest

0 引 言

在大数据与人工智能兴起的时代背景下,现代工业系统组成越来越复杂,为适应不断提高的工业安全与质量保障的需要,基于数据驱动的故障检测技术成为保证整个生产过程安全性和稳定性的重要手段[1]。主要包括主元分析(principal components analysis,PCA)[2]、典型变量分析(canonical variate analysis,CVA)[3]和偏最小二乘法(partial least squares,PLS)[4]等。

在工业过程中所采集到的过程数据通常具有复杂的非线性关系,而 PCA 等方法通常假设过程是线性的,这限制了其处理非线性数据的能力。为了实现对非线性过程的检测,核主元分析(kernel principal component analysis,KPCA)[5]被提出,通过使用非线性核函数将具有非线性特征的数据映射到更高维的线性可分空间中然后在这个高维特征空间进行传统 PCA。典型变量差异度分析(canonical variate dissimilarity analysis,CVDA)[6]作为 CVA 方法的一种扩展通过进一步构造出差异度特征向量,增强了对早期故障的敏感性,提高了故障检测的准确率。但由于其基于线性投影,仅能提取数据中的线性特征,忽略了对非线性

收稿日期:2024-06-16
*基金项目:国家自然科学基金(62273242)项目资助

结构的处理。核规范变量残差分析(kernel canonical variate dissimilarity analysis, KCVDA)[7]将 CVDA 推广到非线性过程,使用核函数将数据映射到高维空间,同时考虑输入和输出数据构建了一个新的统计量。文献[8]提出一种结合加权 CVDA 与 KPCA 的办法来处理非线性动态过程,使用 KPCA 对加权 CVDA 构建的差异度矩阵进一步提取非线性特征,同时提取过程数据中包含的线性和非线性特征,提高了故障检测效果。以上基于 CVDA 的方法均依赖于核函数的优异性能来处理非线性问题,而核函数以及其中参数的选择会对检测结果造成很大影响,计算复杂度也很高[9]。同时,基于核函数的方法没有明确考虑数据的内在非线性几何结构可能驻留在流形上的可能性。流形学习的方法能够将复杂的高维空间转化为更易理解的低维空间,并保留高维空间中的局部邻域结构,从而在低维空间中处理原本的数据。局部线性嵌入(locally linear embedding,LLE)[10]是一种针对非线性数据的图嵌入式特征提取方法,由于模型自身参数少且不需要通过迭代求解,可以学习任意维的低维流形,因此被广泛应用。LLE 通过维持数据集各邻域内样本之间存在的线性关系,在维数约减的同时,保持各邻域中样本之间的原有的线性关系不变,从而利用局部的线性结构反映全局的非线性结构。

隔离森林(isolation forest,IF)[11]算法是一种基于隔离思想的算法,认为故障样本比正常样本更容易被隔离,不要求数据预先遵从任何分布,使用垂直于坐标轴的随机超平面对数据集划分,利用故障样本与正常样本在树结构中深度不同来进行判别。相比通过密度和聚类来进行检测的方法,隔离森林没有繁琐的计算过程,检测所需时间大大降低,同时也减少了异常的掩盖和淹没效应[12]。

针对工业工程数据故障检测中存在的非线性、高维度的问题,本文提出一种结合典型变量差异度分析与局部线性嵌入的故障检测方法。使用 LLE 算法有效弥补了 CVDA 所构建的差异度矩阵难以处理非线性特征的问题的同时,对数据局部邻域内线性关系的保持和维数约减,使得隔离森林算法进行故障检测时更加高效和准确;最后采用窗口半径为 3 的滑动窗口处理异常得分,减少噪声对得分的影响。

1　基础理论

1.1　典型变量差异度分析

假设 $\boldsymbol{X}=[x_1,x_2,\cdots,x_t,\cdots,x_n]^{\mathrm{T}}\in\boldsymbol{R}^{n\times m}$ 为采集到的正常工况下的数据,其中,$x_t\in R^m$,t 表示第 t 个样本,n 为样本个数,m 表示变量维数,由此构造 x_t 对应的历史和未来观测向量 $\boldsymbol{q}_{p,t}$,$\boldsymbol{q}_{f,t}$。

$$\boldsymbol{q}_{(p,t)}=[x_{(t-1)}^{\mathrm{T}},x_{(t-2)}^{\mathrm{T}},\cdots,x_{(t-b)}^{\mathrm{T}}]^{\mathrm{T}}\in\boldsymbol{R}^{m} \tag{1}$$

$$\boldsymbol{q}_{f,t}=[x_t^{\mathrm{T}},x_{t+1}^{\mathrm{T}},\cdots,x_{t+b-1}^{\mathrm{T}}]^{\mathrm{T}}\in\boldsymbol{R}^{\hat{m}} \tag{2}$$

式中:$\hat{m}=m\times b$,b 表示观测向量中包含的样本个数。构造历史和未来观测矩阵 $\boldsymbol{Q}_p$ 和 $\boldsymbol{Q}_f$,由 x_t 对应的 $\boldsymbol{q}_{p,t}$,$\boldsymbol{q}_{f,t}$ 向量分别组成,如式(3)和(4)所示。

$$\boldsymbol{Q}_P=[\boldsymbol{q}_{p,b+1},\boldsymbol{q}_{p,b+2},\cdots,\boldsymbol{q}_{p,b+m}]^{\mathrm{T}}\in\boldsymbol{R}^{M\times m} \tag{3}$$

$$\boldsymbol{Q}_f=[\boldsymbol{q}_{f,b+1},\boldsymbol{q}_{f,b+2},\cdots,\boldsymbol{q}_{f,b+M}]^{\mathrm{T}}\in\boldsymbol{R}^{M\times\hat{m}} \tag{4}$$

式中:矩阵的列数 $M=n-2b+1$。CVDA 算法通过寻找投影向量 $\boldsymbol{h}\in\boldsymbol{R}^{\hat{m}}$ 和 $\boldsymbol{l}\in\boldsymbol{R}^{\hat{m}}$ 使得典型变量对 $\boldsymbol{q}_{p,t}^{\mathrm{T}}\boldsymbol{l}$ 和 $\boldsymbol{q}_{f,t}^{\mathrm{T}}\boldsymbol{h}$ 之间具有最大相关性,对应的优化目标定义如式(5)所示。

$$\begin{cases}\underset{\boldsymbol{l},\boldsymbol{h}}{Max}\,\boldsymbol{l}^{\mathrm{T}}\boldsymbol{\Sigma}_{pf}h\\ s.t.\,\boldsymbol{l}^{\mathrm{T}}\boldsymbol{\Sigma}_{pp}\boldsymbol{l}=1,\boldsymbol{h}^{\mathrm{T}}\boldsymbol{\Sigma}_{ff}\boldsymbol{h}=1\end{cases} \tag{5}$$

上述优化问题的求解可通过式(6)中的奇异值分解进行。

$$\boldsymbol{\Sigma}_{pp}^{-1/2}\boldsymbol{\Sigma}_{pf}\boldsymbol{\Sigma}_{ff}^{-1/2}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}} \tag{6}$$

式中:$\boldsymbol{U}$ 和 $\boldsymbol{V}$ 分别表示左右奇异向量组成的矩阵;对角矩阵 $\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_{\hat{m}})$ 是由特征值 $\lambda_i(1\leqslant i\leqslant\hat{m})$ 降序排列构成的矩阵,特征值 λ_i 代表投影向量对 $\{l_i,h_i\}$ 的相关系数;$\boldsymbol{Q}_p$ 和 $\boldsymbol{Q}_f$ 的协方差和互协方差矩阵为 $\boldsymbol{\Sigma}_{pp}$、$\boldsymbol{\Sigma}_{ff}$ 和 $\boldsymbol{\Sigma}_{pf}$,其具体表达式如式(7)~(9)所示。

$$\boldsymbol{\Sigma}_{pp}=\frac{1}{M-1}\boldsymbol{Q}_p^{\mathrm{T}}\boldsymbol{Q}_p \tag{7}$$

$$\boldsymbol{\Sigma}_{ff}=\frac{1}{M-1}\boldsymbol{Q}_f^{\mathrm{T}}\boldsymbol{Q}_f \tag{8}$$

$$\boldsymbol{\Sigma}_{pf}=\frac{1}{M-1}\boldsymbol{Q}_p^{\mathrm{T}}\boldsymbol{Q}_f \tag{9}$$

求解式(5)的优化问题可以得到历史以及未来数据的投影矩阵 $\boldsymbol{L}=[l_1,l_2,\cdots,l_{\hat{m}}]$ 和 $\boldsymbol{H}=[h_1,h_2,\cdots,h_{\hat{m}}]$,其表达式如式(10)和(11)所示。

$$\boldsymbol{L}=\boldsymbol{\Sigma}_{pp}^{-1/2}\boldsymbol{U} \tag{10}$$

$$\boldsymbol{H}=\boldsymbol{\Sigma}_{ff}^{-1/2}\boldsymbol{V} \tag{11}$$

根据特征值大小,将投影向量对 $\{l_i,h_i\}$ 划分入状态子空间以及残差子空间,前 r 个最大特征值 $\lambda_1>\lambda_2>\cdots>\lambda_r$ 对应的投影向量 $\boldsymbol{L}_r=[l_1,l_2,\cdots,l_r]$ 确定状态子空间,剩余 $\hat{m}-r$ 个特征值对应的投影向量确定残差子空间。求得 CVDA 模型的差异度特征向量,其对应的计算公式如式(12)所示。

$$\boldsymbol{d}_t=\boldsymbol{H}_r^{\mathrm{T}}\boldsymbol{q}_{f,t}-\boldsymbol{\Lambda}_r\boldsymbol{L}_r^{\mathrm{T}}\boldsymbol{q}_{p,t} \tag{12}$$

1.2　局部线性嵌入

对给定高维数据集 $\boldsymbol{X}=[x_1,x_2,\cdots,x_m]^{\mathrm{T}}\in\boldsymbol{R}^{D\times m}$,其中 D 和 m 分别表示 X 的变量数和样本数,LLE 提取低维嵌入结果 $\boldsymbol{Y}=[y_1,y_2,\cdots,y_m]^{\mathrm{T}}\in\boldsymbol{R}^{d\times m}(d\ll D)$ 实现步骤如下:

1) 对于任意 $x_i\in R^D$,选取与 x_i 之间欧氏距离最小的 k 个样本点,形成 x_i 的邻域 $Q(i)$。

2) 计算权重矩阵 $\boldsymbol{W}$,使得误差函数 $J(w)$ 最小:

$$\begin{cases}J(\boldsymbol{W})=\sum\limits_{i=1}^{m}\|x_i-\sum\limits_{j\in Q(i)}w_{ij}x_j\|^2\\ s.t.\sum\limits_{j\in Q(i)}w_{ij}=1\end{cases} \tag{13}$$

其中,若 x_j 不是 x_i 的近邻点时,则对应的 $w_{ij}=0$。

3) 计算原始数据集 $\boldsymbol{X}$ 在低维空间的投影 $\boldsymbol{Y}$。利用式(13)求得的权重矩阵 $\boldsymbol{W}$ 最小化损失函数 $J(\boldsymbol{Y})$,损失函数和约束条件如下:

$$\begin{cases} J(\boldsymbol{Y})=\sum\limits_{i=1}^{m}\| x_i-\sum\limits_{j\in Q(i)}w_{ij}x_j\|^2 \\ s.t.\ \sum\limits_{i=1}^{m}y_i=0,\dfrac{1}{m}\sum\limits_{i=1}^{m}y_iy_i^{\mathrm{T}}=I \end{cases} \tag{14}$$

式(14)可进一步求解得:

$$J(\boldsymbol{Y})=tr(\boldsymbol{YMY}^{\mathrm{T}}) \tag{15}$$

其中,$\boldsymbol{I}$ 为单位矩阵,$\boldsymbol{M}=(\boldsymbol{I}-\boldsymbol{W})(\boldsymbol{I}-\boldsymbol{W})^{\mathrm{T}}$。$\boldsymbol{X}$ 在低维空间的投影 $\boldsymbol{Y}$ 是通过对矩阵 $\boldsymbol{M}$ 的分解,选取其中对应最小的 d 个非零特征值的特征向量构成。

LLE 作为一种流形学习算法,同样存在泛化问题,新样本的映射问题难以得到恰当地解决,只能依靠映射矩阵 $\boldsymbol{A}=[a_1,a_2,\cdots,a_d]^{\mathrm{T}}\in\boldsymbol{R}^{d\times D}$。可以通过式(16)求得映射矩阵[13],

$$a_j=\underset{a}{\operatorname{argmin}}\sum_{i=1}^{k}\| a^{\mathrm{T}}x_i-y_{ij}\|^2 \tag{16}$$

解得 $\boldsymbol{A}=\boldsymbol{YX}^{\mathrm{T}}(\boldsymbol{XX}^{\mathrm{T}})^{-1}$,新的样本 y_i 可以通过 $y_i=\boldsymbol{A}x_i$ 计算其新的嵌入坐标。

2 基于 CVDA 和 LLE 的故障检测

CVDA 基本思想是找到一组投影向量使得历史数据与未来数据的相关性最大化,通过求解这个最优问题,构建出差异度矩阵 $\boldsymbol{d}_t$ 有效放大故障样本与正常样本之间的差异。CVDA 方法所提出的差异度矩阵 $\boldsymbol{d}_t$ 能实现对微小故障的有效检测,但其主要依赖于线性投影和协方差矩阵的特征值分解,仅对数据结构中线性特征的变化敏感,而对于通常具有非线性特征的模型残差部分难以捕捉其变化,其影响不能与其他不确定性相分离[14],这将导致故障检测难以达到理想的结果,而通过进一步提取差异度矩阵 $\boldsymbol{d}_t$ 中的非线性特征可以解决上述问题。LLE 的加入将复杂的高维非线性特征数据映射为更容易被理解的低维局部线性关系,通过邻域内线性关系保持不变的方法,有效的在低维空间还原了 $\boldsymbol{d}_t$ 中的非线性特征。而降维后的数据,避免了隔离森林作为检测手段,进行特征选择时难以挑选合适的特征而导致的检测效果不稳定的问题,大大提高了检测率。相较于传统的工业过程故障检测统计量方法,隔离森林基于整体数据进行检测,对数据本身的变化更加敏感,能够更有效地处理非线性数据。在检测过程中,隔离森林对故障样本具有更好的鲁棒性和灵敏性。

隔离森林是隔离树的集合,对给定的数据集 $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]^{\mathrm{T}}\in\boldsymbol{R}^{n\times m}$,$n$ 和 m 分别表示数据集的变量维度和样本个数,构建隔离森林的细节如下:

1) 从 $\boldsymbol{X}$ 中随机抽取 M 个样本作为构建当前树的子集 $\boldsymbol{N}\in\boldsymbol{R}^{n\times M}$。再从 $\boldsymbol{N}$ 中的 n 个变量中随机选取一个变量 y_j。

2) 从 y_j 的最大最小值之间随机选取一个分隔值 q。

3) 根据 $\boldsymbol{N}$ 中每个样本的 y_j 值大于或小于 q,将 $\boldsymbol{N}$ 划分到左右子节点 T_l 和 T_r。

4) 分别在 T_l 和 T_r 中重复以上步骤构造一棵 $iTree$,直到子节点只有一个实例点,或者子节点中的所有数据具有相同值,或者 $iTree$ 达到了限制高度。

5) 重复上述步骤构建出 T 棵树,形成完整的隔离森林 $\boldsymbol{IF}=[t_1,t_2,\cdots,t_T]$,$T$ 表示所构建树的数量。

构建好完整的隔离森林后,用测试数据 x 遍历每棵树,计算 x 的异常得分。异常分数的定义如下:

$$s(x,n)=2^{\frac{-E(h(x))}{c(n)}} \tag{17}$$

其中,为 $h(x)$ 表示 x 在隔离林中被隔离次数,$E(h(x))$ 表示 x 在 T 棵树中平均被隔离次数,$c(n)$ 为规范 $h(x)$ 的标准化因子,是将一个样本与其他 n 个样本隔离所需的平均步骤数,用作其他 n 个样本的平均路径长度基准,定义为:

$$c(n)=\begin{cases} 2H(n-1), & n>2 \\ 1, & n=2 \\ 0, & \text{其他} \end{cases} \tag{18}$$

其中,$H(i)$ 为调和数,$H(i)=\ln n+\varepsilon$,ε 为欧拉常数,其近似为 0.577 215 664 9[11]。

整个模型有 2 个独立阶段,即离线训练阶段和在线检测阶段。其详细步骤如下。

离线建模过程如下:

步骤 1)采集正常运行数据,并进行标准化。

步骤 2)构造历史和未来观测矩阵 $\boldsymbol{Q}_p$ 和 $\boldsymbol{Q}_f$,窗口长度 $b=3$。计算两者之间的协方差矩阵和互协方差矩阵 $\boldsymbol{\Sigma}_{pp}$、$\boldsymbol{\Sigma}_{ff}$ 和 $\boldsymbol{\Sigma}_{pf}$。

步骤 3)计算差异度矩阵 $\boldsymbol{d}_t$。

步骤 4)用 LLE 算法对 $\boldsymbol{d}_t$ 进行降维,得到映射矩阵 $\boldsymbol{A}$ 和降维后的数据 $\widehat{\boldsymbol{d}_t}$。

步骤 5)利用 $\widehat{\boldsymbol{d}_t}$ 构建隔离森林模型,并得到正常运行数据的异常得分,使用核密度估计计算控制限 s_{ucl}。

在线检测过程如下:

步骤 1)使用正常运行数据的均值和方差对待检测数据进行标准化。

步骤 2)构造历史和未来观测矩阵 $\boldsymbol{testQ}_p$ 和 $\boldsymbol{testQ}_f$。计算两者之间的协方差矩阵和互协方差矩阵 $\boldsymbol{test\Sigma}_{pp}$、$\boldsymbol{test\Sigma}_{ff}$ 和 $\boldsymbol{test\Sigma}_{pf}$。

步骤 3)计算测试样本的差异度矩阵 $\boldsymbol{testd}_t$。

步骤 4)用离线建模阶段得到的映射矩阵 $\boldsymbol{A}$ 将 $\boldsymbol{testd}_t$ 映射到低维空间得到 $\widehat{\boldsymbol{test\ d}_t}$。

步骤 5)使用构建好的隔离森林模型对 $\widehat{\boldsymbol{test\ d}_t}$ 进行检测,得到待检测数据的异常得分 s_{test},通过与控制限 s_{ucl} 比较,判断样本属于正常样本或故障样本。

CVDA-LLE 的故障检测方法流程如图 1 所示。

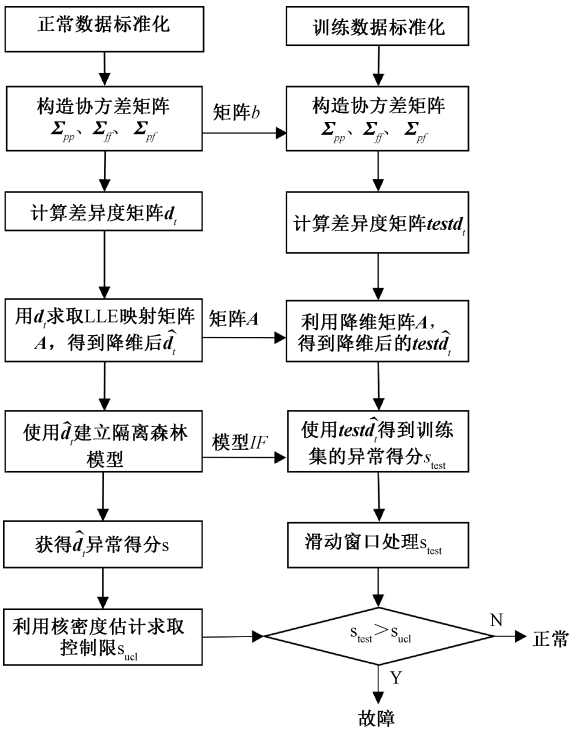

图 1　CVDA-LLE 方法的故障检测流程图

Fig. 1　Fault detection flowchart of CVDA-LLE method

3　仿真实验

3.1　非线性数值实例仿真应用

用一组非线性仿真实例来验证本方法的有效性，具体构成如式(19)、(20)所示。

$$\begin{cases} x_1 = 0.576\ 8s_1 + 0.376\ 6s_2 + \sigma_1 \\ x_2 = 0.738\ 2{s_1}^2 + 0.056\ 6s_2 + \sigma_2 \\ x_3 = 0.829\ 1s_1 + 0.400\ 9{s_2}^2 + \sigma_3 \\ x_4 = 0.651\ 9s_1s_2 + 0.207\ 0s_2 + \sigma_4 \\ x_5 = 0.397\ 2s_1 + 0.804\ 5s_2 + \sigma_5 \end{cases} \tag{19}$$

$$\begin{cases} s_1 \sim uniform(2,5) \\ s_2 \sim N(7,1) \end{cases} \tag{20}$$

其中，x_n 为观测变量，s_1 和 s_2 是控制变量，σ_n 是独立白噪声，服从均值为 0 标准差为 0.01 的高斯分布。

实验分为 2 个部分，第 1 部分离线建模：训练集由 400 个正常数据组成；第 2 部分在线检测：测试集的前 200 个数据作为正常样本，后 200 个为故障样本。一共设定 2 种故障类型：故障 1 设定为一种斜坡故障，具体添加在变量 x_1 从第 201 个故障开始到结束，i 为样本数；故障 2 作为一种阶跃故障，从变量 x_5 第 201 个样本开始增加 25%的幅值。

各个方法对此非线性数值仿真实例的检测结果如表 1 所示，KPCA 以及 CVDA 选择 85%的主元贡献率；PPA 提取主多项式成分数为 2，每个多项式阶数为 2；LLE 算法采用近邻近数为 12，降至 3 维；统计量控制限统一设置为 99%。选取每个方法各个统计量中检测效果最好的展示。

表 1　数值例子检测准确率

Table 1　Accuracy of numerical example detection　%

故障类型	KPCA	PPA	CVDA-LLE
F1	**99.5**	95	**99.5**
F2	91.5	2.5	**99**

从表中看出 CVDA-LLE 方法的有效性与优越性，检测率在两种方法中均为最高，特别是在故障 2 的检测结果上远远好于其他两种方法，在故障 2 的检测率上 CVDA-LLE 远远优于 PPA，检测结果如图 2 所示。

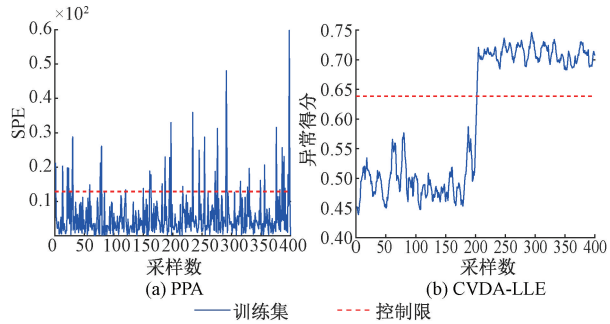

图 2　两种方法对故障 2 检测效果

Fig. 2　Detection effect of two methods on fault 2

相较于 KPCA，CVDA-LLE 在故障 2 上检测率仅提高了 7.5%，但是如图 3 所示，所展示的为分别经过 3 种方法处理后故障 2 在二维空间的散点分布图，蓝色部分表示正常样本数据，红色部分表示故障样本数据。图 3(a)为原始数据散点分布图。图 3(c)中可见原始数据在经过 PPA 处理后故障样本与正常样本依然呈现随机分布。图 3(b)中 KPCA 处理后的数据也能较好的分辨故障数据和正常数据，但图 3(d)中经过 CVDA-LLE 处理后的数据根据正常和故障样本分为了差异更加明显的 2 个簇，且数据在 2 个簇中的分布也更为密集，说明了 CVDA-LLE 方法在提取非线性故障数据的特征上更优于 KPCA 以及 PPA，能更好的实现对数据非线性特征的捕捉。

3.2　TE 过程仿真应用

TE(tennessee eastman)过程仿真应用平台是一个基于实际工业过程的仿真例子，用于测试过程系统工程中故障检测和诊断方法的效率[15]。过程主要包括 5 个操作单元：反应器、冷凝器、循环压缩机、气液分离器以及汽提塔。整个过程中共涉及 8 种物料成分，分别为主要参加反应的气体进料 U、C、D、E；惰性不可溶进料 B；反应副产品 F 以及反应液态主产物 G 和 H。在过程中共包含有 41 个测量变量和 12 个控制变量，其中 41 个测量变量又可以划分为 22 个过程测量变量以及 19 个成分测量变量[16-17]。其 21 种故障类型如表 2 所示。

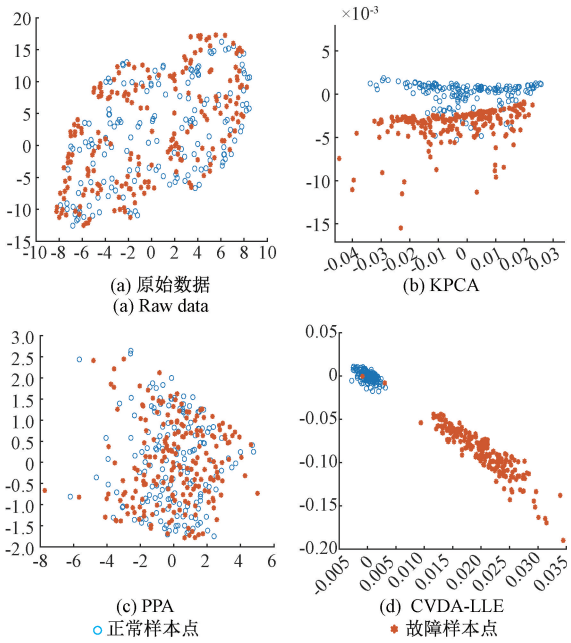

图 3　3 种方法故障 2 散点分布图

Fig. 3　Fault 2 scatter distribution map of three methods

表 2　TE 过程的 21 种故障

Table 2　21 types of faults in TE process

故障编号	性质描述	变化类型
F1	物料 U/C 进料比改变,物料 B 含量不变	阶跃
F2	物料 U/C 进料比改变,物料 B 含量改变	阶跃
F3	物料 D 进料温度改变	阶跃
F4	反应器冷却入口温度改变	阶跃
F5	冷凝器冷却入口温度改变	阶跃
F6	物料 U 进料损失	阶跃
F7	物料 C 压力损失	阶跃
F8	物料 U、B、C 的组成比例改变	随机变量
F9	物料 D 进料温度改变	随机变量
F10	物料 C 进料温度改变	随机变量
F11	反应器冷却水入口温度改变	随机变量
F12	冷凝器冷却水入口温度改变	随机变量
F13	反应动力学参数改变	慢偏移
F14	反应器冷却阀门	粘住
F15	冷凝器冷却阀门	粘住
F16	未知	未知
F17	未知	未知
F18	未知	未知
F19	未知	未知
F20	未知	未知
F21	物流 4 阀门固定在恒定位置	恒定位置

将 CVDA-LLE 与 KPCA、PPA 以及 CVDA 对 TE 过程的检测结果对比。在建立检测模型时,KPCA 以及 CVDA 的主元贡献率选择为 85%;PPA 提取主多项式成分数为 4,每个多项式阶数为 3;LLE 算法采用近邻近数为 $k=33$,降至 $d=12$ 维;隔离森林算法中构建隔离树 $T=100$ 棵,每棵树随机采样样本数 $M=256$,最大深度为 10;统在计算各个方法统计量的控制限时,均采用 99% 的置信度。

表 3 展示了 KPCA、PPA、CVDA 以及 CVDA-LLE 方法对 TE 过程各个的故障检测结果。21 类故障的检测准确率中,针对每一个故障用粗体突出表示 4 种方法中最高的检测率,表中可见本文所提方法对除故障 3、9、15 的其他故障类型均能实施有效检测,平均检测率为 95.28%,而故障 3、9、15 被认为是具有幅值低、易被噪声扰动和掩盖特点的微小故障,在众多方法中均难以实现检测。对于易被检测的故障 1、2、6、7、8、12、13、14、17、18,所有方法均能实现 90%以上有效监测,但本文所提方法对这些故障检测率平均为 98.4%。

表 3　各方法故障检测率

Table 3　Fault detection rate of each method　%

故障序号	KPCA		PPA		CVDA			CVDA-LLE
	T^2	SPE	T^2	SPE	T^2	SPE	D	
F1	99.4	**100**	99.1	99.5	**100**	99.6	**100**	**100**
F2	98.5	99	98	98.6	97.4	95.6	97.1	**99.1**
F3	2.3	4.8	12.5	**20.4**	6.9	8.3	4.3	13.8
F4	48.5	**100**	36.3	87.9	51.6	28.9	29.6	**99.7**
F5	26.4	28.1	29	40	**100**	99.7	**100**	**100**
F6	18.1	**100**	98.5	**100**	**100**	99.9	**100**	**100**
F7	**100**	**100**	81.9	**100**	57.4	28.3	47.1	**100**
F8	97.3	98.1	91.5	**99.1**	89.3	85.5	83.8	98.1
F9	3.3	5.1	9.6	18.1	7	6.1	5.3	**24.8**
F10	42.3	64.9	32.5	63.5	78.9	74.6	75.3	**90.6**
F11	51.1	66.4	38.5	69.8	60.4	19.1	47	**91**
F12	98.6	98.9	82.8	99.6	98.9	92.7	96.9	**100**
F13	95	95.5	92.5	95.8	94.4	89.4	91.5	**97.5**
F14	**100**	**100**	78	**100**	79.5	38.4	68.1	**100**
F15	3.8	8.6	10.6	23.6	9	6.9	5.9	**47**
F16	22.6	56.8	30.5	56.5	82	61.5	79.1	**98.8**
F17	80.4	96.8	71.8	90.9	94.4	78.5	90.8	**97.8**
F18	24.4	90.6	87.3	92.4	89.9	90.8	89.5	**91.5**
F19	7.4	29.9	7.4	14.5	78	15.3	73	**99.6**
F20	40.6	63.1	31.8	65.3	82.5	68.8	78	**92.6**
F21	41.6	52.7	**64.3**	52.3	53.6	15.8	47.6	61

为体现基于 CVDA-LLE 检测方法的有效性和优越性,实验比较了 4 种方法在检测故障 11、19 的检测效果,如图 4、5 所示。

故障 11 是反应器冷却水入口温度改变引起的随机变

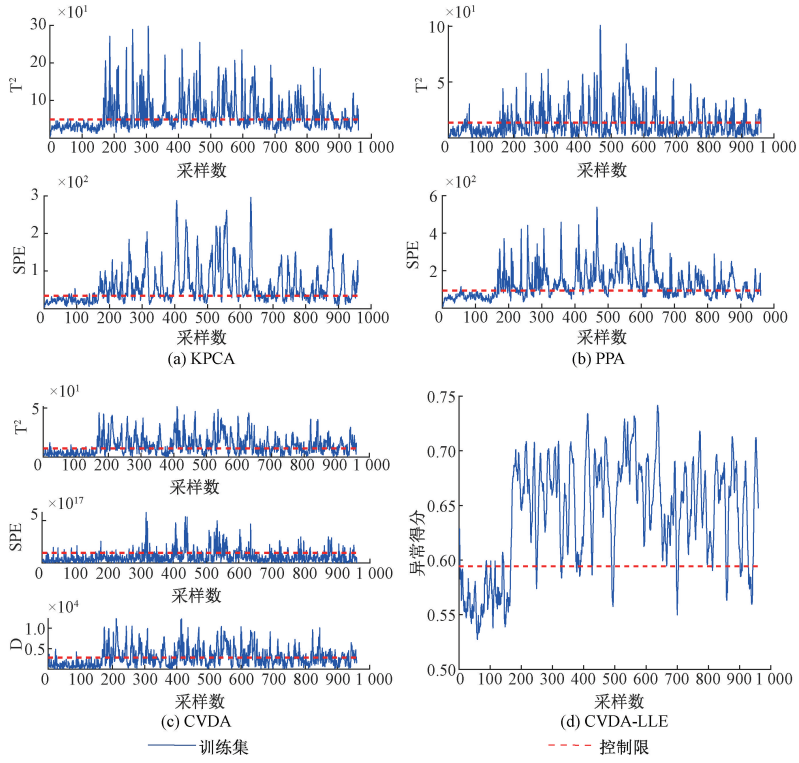

图 4　4 种方法对故障 11 的检测效果

Fig. 4　Detection effect of four methods on fault 11

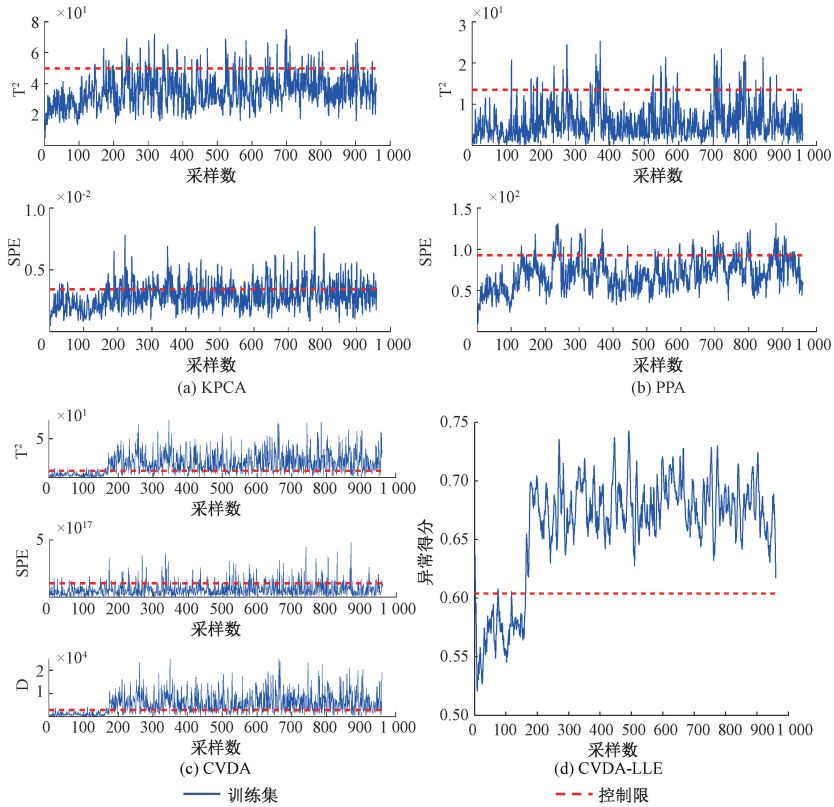

图 5　4 种方法对故障 19 的检测效果

Fig. 5　Detection effect of four methods on fault 19

量故障。从图 4(a)、(b)中可以看到,KPCA 以及 PPA 对故障 11 的检测不理想,检测率均维持在 65%左右,主要因为分布在各个区间的部分故障样本统计量低于决策边界。而图 4(d)中 CVDA-LLE 方法由于 LLE 维持了数据的局部邻域信息,放大了部分原本低于决策边界样本的统计量,成功实现了 91%的检测率。而图 4(c)所示 CVDA 方法由于对非线性特征处理能力的欠缺,对此故障的检测准确率低于 KPCA 和 PPA。从表 2 中故障 4、7、10、11、14、16、21 的检测率也可以看出,CVDA 方法的检测率均低于 KPCA 与 PPA 算法。而 CVDA-LLE 对各个故障的检测效果却是 4 种方法中最好的,这表明 LLE 通过对数据局部关系的保持,在弥补了 CVDA 方法对非线性问题处理缺陷的同时,对这些故障中存在的非线性特征的处理能力优于 KPCA 以及 PPA。

故障 19 是一种未知故障,如图 5(a)、(b)所示 KPCA 以及 PPA 对此故障均难以实现有效检测,说明这两种方法难以实现对此类故障非线性特征的捕捉。图 5(c)中 CVDA 的 T^2 统计量检测率为 78%,相较于 KPCA 以及 PPA 算法显著放大了故障样本与正常样本的差异,而图 5(d)中 CVDA-LLE 的检测率则高达 99.6%,从样本 161 开始,故障样本的异常得分明显高于正常样本,并且持续显著高于决策边界,说明 LLE 算法能够有效处理故障中的非线性特征,大大提高了 CVDA 算法的检测性能。

实验表明相较于传统针对非线性数据的检测算法 CVDA-LLE 方法的有效性和优越性。LLE 有效提取了数据中的非线性特征和局部邻域信息,成功弥补了 CVDA 对故障数据的非线性变化不够敏感的缺点,有效提升了检测率。

4 结 论

本文提出了一种基于 CVDA-LLE 的故障检测算法。CVDA 构造的差异度矩阵 $\boldsymbol{d}_t$ 有效放大了故障和正常样本之间的差异,但是其在对具有非线性特征的数据实施检测时效果不理想。实验表明加入 LLE 后,将复杂的高维非线性数据映射到更容易被理解的低维空间,通过局部邻域的线性关系在低维空间反映原非线性关系,有效的提取了 $\boldsymbol{d}_t$ 中的非线性特征,同时其维数约减以及局部结构信息的保留有效提升了隔离森林算法作为检测手段的灵敏度和鲁棒性,提高模型整体对于故障数据的敏感性。本文通过将提出的方法与其他现有方法在一组非线性数值例子以及 TE 过程数据中的检测效果进行比较,证实了所提方法的有效性和优越性。

参考文献

[1] 商亮亮,陆智林,文传博,等.基于规范变量残差的化工过程微小故障检测与诊断[J].控制理论与应用,2021,38(8):1247-1256.
SHANG L L, LU ZH L, WEN CH B, et al. Small fault detection and diagnosis in chemical processes based on canonical variate residuals [J]. Control Theory and Applications, 2021, 38(8): 1247-1256.

[2] 卞华棋,王育辉,王栋,等.基于 PCA 的空战机动高阶重构与评估[J].电子测量与仪器学报,2023,34(6):191-197.
BIAN H Q, WANG Y H, WANG D, et al. High-order reconstruction and evaluation of aerial combat maneuvers based on PCA[J]. Journal of Electronic Measurement and Instrumentation, 2023, 34(6): 191-197.

[3] CHEN ZH W, YANG CH H, PENG T, et al. A cumulative canonical correlation analysis-based sensor precision degradation detection method [J]. IEEE Transactions on Industrial Electronics, 2019, 66(8): 6321-6330.

[4] 付乐天,李鹏,高莲.考虑样本异常值的改进最小二乘支持向量机算法[J].仪器仪表学报,2021,42(6):179-190.
FU L T, LI P, GAO L. Improved least squares support vector machine algorithm considering sample outliers[J]. Chinese Journal of Scientific Instrument, 2021, 42(6): 179-190.

[5] 严帅,熊新.基于 KPCA 和 TCN-Attention 的滚动轴承退化趋势预测[J].电子测量技术,2022,45(15):28-34.
YAN SH, XIONG X. Rolling bearing degradation trend prediction based on KPCA and TCN-Attention [J]. Electronic Measurement Technology, 2022, 45(15): 28-34.

[6] PILARIO K E S, CAO Y. Canonical variate dissimilarity analysis for process incipient fault detection [J]. IEEE Transactions on Industrial Informatics, 2018, 14(12): 5308-5315.

[7] XIAO S J. Research on fault detection method based on kernel canonical variate analysis [D]. Beijing: Beijing University of Chemical Technology, 2019.

[8] 秦玉峰,史贤俊.基于递推规范变量残差和核主元分析的微小故障检测[J].控制理论与应用,2022,39(9):1716-1724.
QIN Y F, SHI X J. Small fault detection based on recursive canonical variate residuals and kernel principal component analysis[J]. Control Theory and Applications, 2022, 39(9): 1716-1724.

[9] LI H K, CHEN J H, SHENG D, et al. Thermal parameters modeling method and sensor fault diagnosis based on KPCA-RBF network[J]. Journal of Vibration Measurement and Diagnosis, 2016, 36(6):

1044-1049.

[10] 李元,耿泽伟.基于 LLE 与 K 均值聚类算法的工业过程故障诊断[J].系统仿真学报,2021,33(9):2066-2073.
LI Y, GENG Z W. Industrial process fault diagnosis based on LLE and K-means clustering algorithm[J]. Journal of System Simulation, 2021, 33(9): 2066-2073.

[11] HARIRI S, MATIAS C K, et al. Extended isolation forest[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(4): 1479-1489.

[12] 周杭,蒋瑜.基于高对比度子空间的改进孤立森林方法[J].计算机应用研究,2023,40(2):388-393.
ZHOU H, JIANG Y. Improved isolation forest method based on high contrast subspace [J]. Application Research of Computers, 2023, 40(2): 388-393.

[13] LIU Y H, HU Z B, ZHANG Y SH. Bearing feature extraction using multi-structure locally linear embedding [J]. Neurocomputing, 2021, 428: 280-290.

[14] DENG X G, TIAN X M, CHEN SH, et al. Nonlinear process fault diagnosis based on serial principal component analysis[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 29(3): 560-572.

[15] ZHU J L, GE ZH Q, SONG ZH H, et al. Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data[J]. Annual Reviews in Control, 2018, 46: 107-133.

[16] JIANG L, SONG ZH H, GE ZH Q, et al. Robust self-supervised model and its application for fault detection[J]. Industrial & Engineering Chemistry Research, 2017, 56(26): 7503-7515.

[17] 李元,姚宗禹.基于邻域保持嵌入的主多项式非线性过程故障检测[J].上海交通大学学报,2021,55(8):1001-1008.
LI Y, YAO Z Y. Principal polynomial nonlinear process fault detection based on neighborhood preserving embedding[J]. Journal of Shanghai Jiao Tong University, 2021, 55(8): 1001-1008.

作者简介

蒋磊峰,硕士研究生,主要研究方向为基于数据驱动的复杂过程故障诊断。

E-mail:3079397876@qq.com

张成,副教授,博士,主要研究方向为过程故障诊断分析。

E-mail:zhangcheng@syuct.edu.cn

李元(通信作者),教授,博士生导师,博士,主要研究方向为基于数据驱动的复杂过程故障诊断。

E-mail:li-yuan@mail.tsinghua.edu.cn