文章编号：0253-374X(2022)02-0241-12

DOI：10.11908/j.issn.0253-374x.21186

基于改进典型相关分析的中低速悬浮系统异常检测方法

王　平[1,2]，梅　子[2]，龙志强[2]

（1. 中国空气动力研究与发展中心 设备设计及测试技术研究所，四川 绵阳 621000；2. 国防科技大学 智能科学学院，湖南 长沙 410073）

摘要：利用悬浮系统的多类监测数据，提出了一种基于改进典型相关分析（canonical correlation analysis，CCA）的中低速悬浮系统异常检测方法。运营线数据验证了该方法能获得较好的阈值，且与基于*K*-medoids的方法和基于支持向量数据域描述（support vector data description，SVDD）的方法相比，该方法能获得更高的检测率。

关键词：悬浮系统；异常检测；典型相关分析

中图分类号：TP29　　**文献标志码**：A

Anomaly Detection Method of Middle-low Speed Suspension System Based on Improved Canonical Correlation Analysis

WANG Ping[1,2], *MEI Zi*[2], *LONG Zhiqiang*[2]

(1. Facility Design and Instrumentation Institute, China Aerodynamics Research and Development Center, Mianyang 621000, China; 2. College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410073, China)

Abstract: This paper presents an anomaly detection method for middle-low suspension system based on an improved canonical correlation analysis (CCA) by using the multiple types of monitoring data from the suspension system. A better threshold and a higher detection rate can be obtined by the proposed method in comparison with the thresholds by *K*-medoids -based method and the support vector data description (SVDD)-based method.

Key words: suspension system; anomaly detection; canonical correlation analysis

中低速磁浮列车作为一种新型的城市轨道交通工具，越来越得到公众的关注和认可。悬浮系统作为中低速磁浮列车的关键系统之一，其安全性和可靠性对中低速磁浮列车的运行具有十分重要的影响。其中，异常检测技术作为一种提高系统运行安全的有效手段，受到了学术界和工业界的广泛关注和研究。因此，为提高中低速磁浮列车悬浮系统的安全性和可靠性，研究悬浮系统的异常检测技术具有十分重要的意义。

国外，Harrou 等提出了一种基于主成分分析（principal component analysis，PCA）的 MCUSUM 异常检测策略，能更好地检测出系统的微小异常[1]。Bi等提出了一种基于PCA的异常检测模型，能够准确有效地检测出异常，提高了算法的效率和稳定性[2]。Xie 等针对在线和准确的异常检测问题提出了一种基于双边PCA的异常检测方法，实现了快速准确的实时异常检测[3]。Pan 等提出了一种用于传感器数据集成核PCA和关联规则挖掘的数据驱动异常检测方法，可实现卫星电力系统的异常检测[4]。Yin等提出了一种改进的数据流聚类算法，并根据改进的算法设计了异常检测模型，所建立的模型能够随着数据流的变化进行修改，从而及时检测出异常的发生[5]。Gu提出了一种系统的方法来识别客流异常的性质并动态估计其警报级别，能够很好应用于预警管理和优化客流组织策略的实时决策[6]。Li提出了一种新的基于群集的异常检测方法来检测异常航班，能够在事故发生之前识别和减轻风险[7]。针对物联网应用程序中实时检测异常问题，Lyu提出了一种Fog-Empowered异常检测方法，不仅能够减少延迟和能耗还能实时检测异常，并且有较高的检测精度[8]。针对从异常环境中的训练集导出的主成分可能会被异常扭曲的问题，O'Reilly等提出了一种最小体积椭圆主成分分析的分布式异常检测方法，能够得到更为稳健的训练集主成分[9]。针对大多数异常检测算法无法满足有效性和实时能力的问题，Ding提出了一种基于长短期记忆（long short-term memory，LSTM）神经网络和高斯混合模型

收稿日期：2021-05-02
基金项目：国家"十三五"重点项目(2016YFB1200600)；国家自然科学基金(62003049，61801479)
第一作者：王平(1989—)，男，工学博士，主要研究方向是故障预测与健康管理。E-mail：wang13548607921@163.com

论文拓展介绍

(Gaussian mixture models, GMM)的实时异常检测算法[10]。Yang等提出了一个基于二部图和共聚的异常检测创新框架,能够在新浪微博数据集上高精度地检测个体和群体异常[11]。Yan提出了一种深度半监督异常检测方法,并将该方法应用于燃气轮机燃烧室异常检测方面,能够有效地检测出燃烧室的异常或故障[12]。Lu等提出了一种用于异常检测的半监督机器学习算法,解决了传统的基于关键质量指标的硬决策方法难以承担大数据环境下监控体验质量异常检测任务的问题[13]。针对无线通信中的频谱异常检测问题,Feng等采用深结构自编码神经网络对频谱异常进行检测,并以时频图作为学习模型的特征,同时用阈值来区分异常和正常数据[14]。为了解决无法获得异常的经验知识或历史数据完全没有标记而导致的传统故障识别方法不适用等问题,Li等提出了一种新颖的基于深度学习的机械设备异常检测方法[15]。Liang等提出了一种共享的连接深度神经网络,用于电力消耗时间序列异常预测[16]。

国内,对于车载灵活数据速率控制器局域网络,罗峰等提出了一种基于支持向量机的异常入侵检测算法[17]。针对多波束海底地质数据的异常问题,何书锋等提出了一种新的异常检测方法——深度支持向量检测算法[18]。王慧珍等提出了一种基于Logistic集成学习的列车MVB网络异常检测方法[19]。考虑无人机传感器易受网络攻击问题,充分利用数据的时间相关性,李晨等提出了针对无人机传感器数据的异常检测模型[20]。针对非规则采样且具有缺失值的多维航空时序数据,闫媞锦等提出了非规则采样多维时序数据异常检测算法[21]。韩昭蓉等提出了一种基于双向LSTM模型的轨迹异常点检测算法。该方法的检测性能显著优于恒定速度阈值法、不考虑数据时序性的经典机器学习分类算法和卷积神经网络模型[22]。为了解决训练样本数据集中正类、负类样本不平衡的问题,姚宇等提出一种考虑负类样本信息的加权超椭球体支持向描述方法[23]。针对传统异常检测模型在数据不平衡情况下对少数异常类样本识别效果较差的问题,王杰等提出了一种基于改进扩散映射的支持向量数据描述算法,构建新的模型并将其应用于工业异常检测[24]。针对核动力系统故障类型多、故障发生概率小、故障样本匮乏的特殊问题,基于夹角余弦距离计算理论,从正常历史运行数据出发,王雯珩等提出一种检测系统异常的算法,有效应对了故障样本不足的现实问题[25]。针对目前大多数方法仅从单一视角检测业务流程执行异常而导致的异常检测不全面问题,孙笑笑等提出了一种基于上下文感知的多角度业务流程在线异常检测方法[26]。

此外,虽然在实际工程应用中,根据《中低速磁浮交通车辆悬浮控制系统技术条件CJ/T458-2014》,目前悬浮系统已具备一套自诊断系统,且该自诊断系统根据上述经验阈值进行异常检测,但额定的悬浮间隙以及间隙波动的情况比较复杂。主要有:

(1)根据线路情况和列车的状况调整额定的悬浮间隙,如北京线设计为8mm,长沙线为9mm等,且悬浮控制器会根据速度的变化来调整额定的悬浮间隙。另外,由于每个传感器和结构安装的差异性使得每个点的额定悬浮间隙不一定是8mm。

(2)列车运行时由于速度、轨道不平顺等多种因素综合作用,或者列车悬浮静止时由于车轨共振等因素作用导致异常的间隙值低于经验阈值,从而导致漏报。

综上所述,虽然LSTM等深度学习方法取得了较好的结果,但实际工程中不能标记出足够多的异常数据。而多元分析技术能利用历史数据进行异常检测,且不需要大量的异常数据,甚至直接通过健康数据就能检测异常数据。常用的多元分析技术有基于PCA、基于偏最小二乘方法(partial least squares, PLS)和基于典型相关分析(canonical correlation analysis, CCA)的异常检测方法。虽然三种方法都有模型训练和异常检测两步,但又有一定的区别。基于PCA的检测方法在整个过程中只考虑一个数据集,基于PLS的检测方法考虑的是过程变量和质量相关变量,而基于CCA的检测方法面向的是系统中存在明确的输入输出关系且输入输出数据在线可测的情况,即基于CCA的检测方法可以视为基于PCA和基于PLS的检测方法的一种扩展[27]。因此,针对悬浮系统异常检测问题,利用悬浮系统的输入输出数据,提出了一种基于改进的CCA的多维时间序列异常检测方法。通过CCA处理悬浮系统的多维数据以获得故障检测指标,即二次统计量。由于悬浮系统中部分数据呈非高斯分布导致二次统计量也呈非高斯分布的问题,使用Box-Cox变换将每种类别下建立的二次统计量转换为高斯分布变量,并利用高斯分布的置信区间来确定异常阈值。

1　悬浮系统的异常数据

1.1　数据分类

由于悬浮系统在不同的运行场景下所产生的数据之间存在较大的差异,故需要对数据进行划分,以便在不同的运行场景下进行异常检测,这样有利于提高检测的可靠性。悬浮系统在运行中会产生大量的数据,主要包括悬浮间隙、电磁铁电流、悬浮电磁铁的垂向加速度、悬浮控制器的输入电压和车辆运行速度。

图1为悬浮监控单元中某一个悬浮控制单元在某一天的间隙数据,由车库内、出库、正线运行和回库4部分数据组成,其中,一天的数据约不到70万个样本点,而在正线运行过程中采集的悬浮数据有近50万个采样点。因此,正线运行过程的数据是本文的研究重点。由图可知,在正线运行的数据中,站内悬浮静止的间隙数据(如第1个虚线区域所示)与站间运行的间隙数据(如第2个虚线区域所示)的幅值有明显的差异。

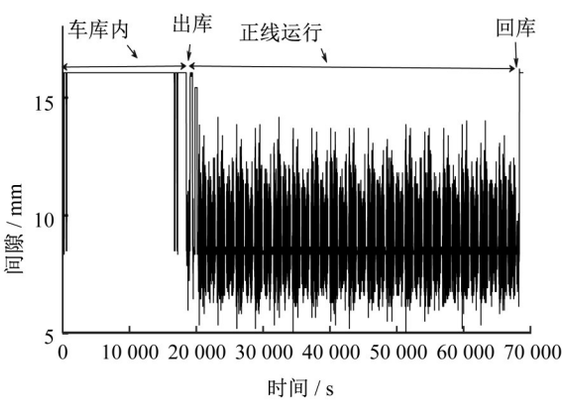

图1　某一个悬浮控制单元在某一天的间隙数据

Fig. 1　The gap data of a certain suspension control unit on a certain day

为了更为直观地反映两者之间的差异,从悬浮系统的历史数据中选择一段列车从始发站运行到终点站的悬浮系统历史数据,如图2所示。从图2可以看出,站内悬浮静止的间隙数据与站间运行的间隙数据之间的差异大且站间运行的间隙波动较为频繁。因此,本文将磁悬浮列车的数据分为2类(站内静止悬浮和站间行驶),其中,第1个虚线框中的数据属于站内静止悬浮,第2个虚线框中的数据属于站间行驶。

1.2　典型的异常类型

结合工程经验和运营数据,本文从站间行驶和站内静止悬浮中分别选择一些典型的异常。

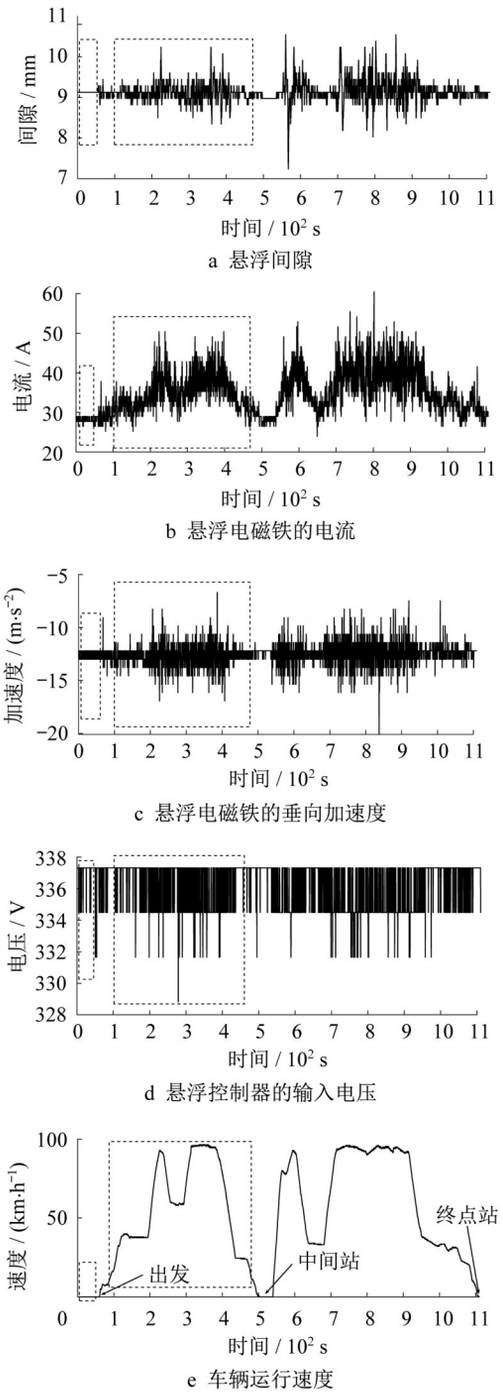

图2　悬浮系统的历史数据曲线

Fig. 2　The historical data curve of the suspension system

1.2.1　站间行驶时异常类型

本文从站间行驶的历史数据中选择含有3类异常的数据。图3~图5分别为第1类异常、第2类异常和第3类异常的数据,其中虚线框中的数据属于异常数据,图a中两条直线分别为基于经验的上、下

限阈值。当前数据来源于标准间隙为9mm的悬浮系统,则基于经验的上、下限阈值分别为5mm和13mm。

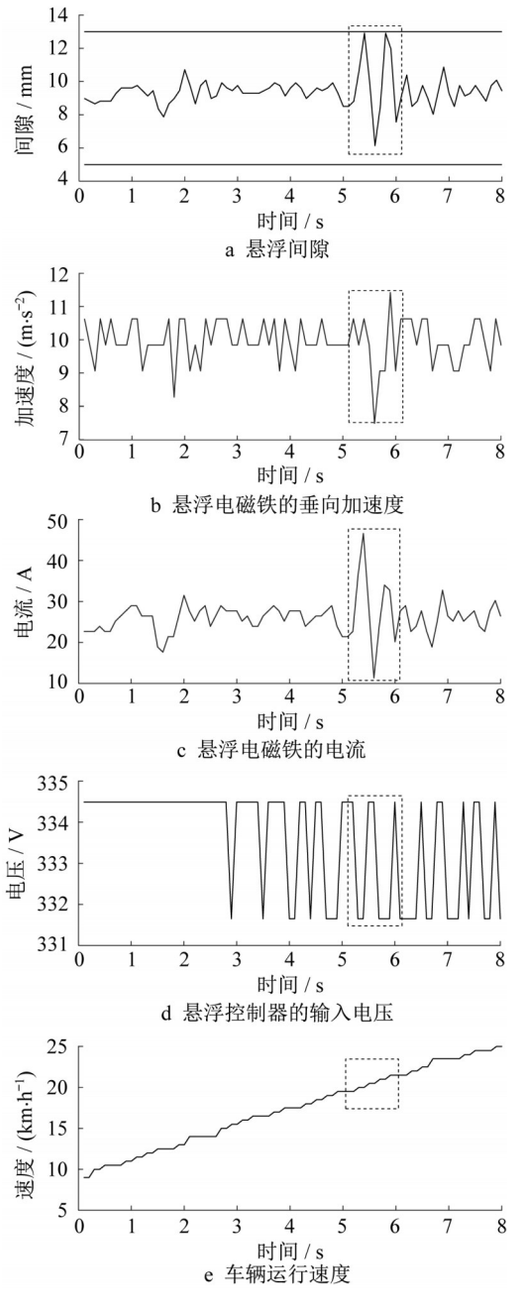

图3 第1类异常的数据

Fig. 3 The first type of abnormal data

图3中间隙有明显的波动,但没有超过经验阈值,此时加速度、电流和电压也对应有一定的波动。在实际工程中,当悬浮系统过三型接头或弯道时容易发生该类异常。对于这类异常,通过经验阈值不一定能检测出来。图4的这类异常的现象是间隙有很大的波动,且超过经验阈值,此时加速度、电流和电压也对应有巨大的波动。在实际工程中,当悬浮

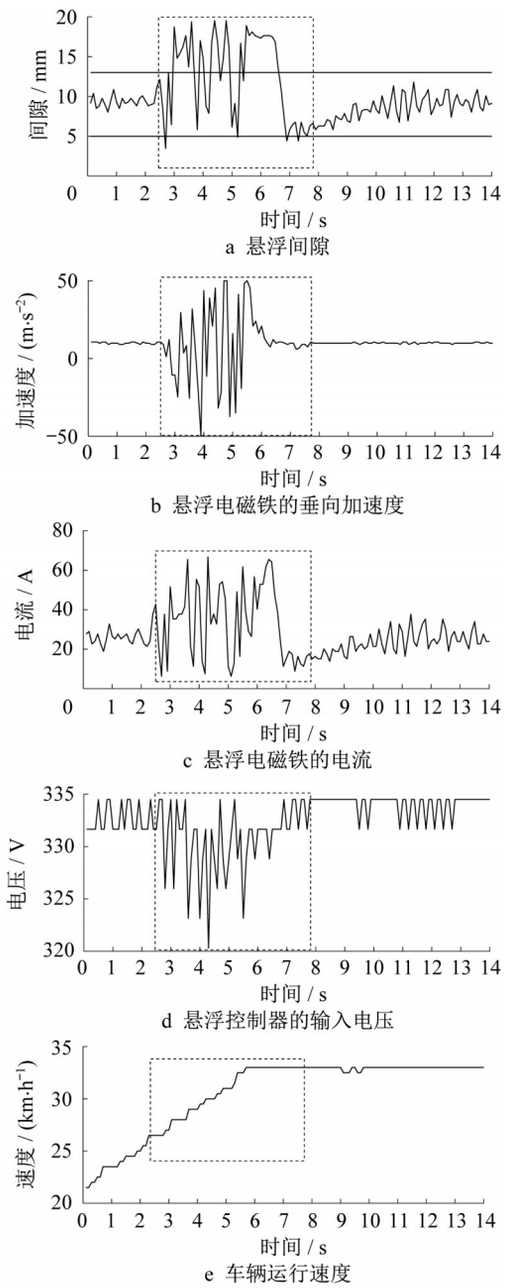

图4 第2类异常的数据

Fig. 4 The second type of abnormal data

系统多次与轨道发生碰撞时容易发生该类异常。对于这类异常,通过经验阈值很容易检测出来。图5的这类异常中,除了个别位置的间隙稍微大点外,其他位置的间隙没有明显的波动,且都没有超过经验阈值,此时电流和电压也没有明显的波动,但加速度的波动很明显。在实际工程中,当加速度传感器有异常或加速度传感器的灵敏度比间隙传感器高或控制器内部接插件出现问题时容易发生该类异常。对于这类异常,根据经验阈值方法无法检测出来。

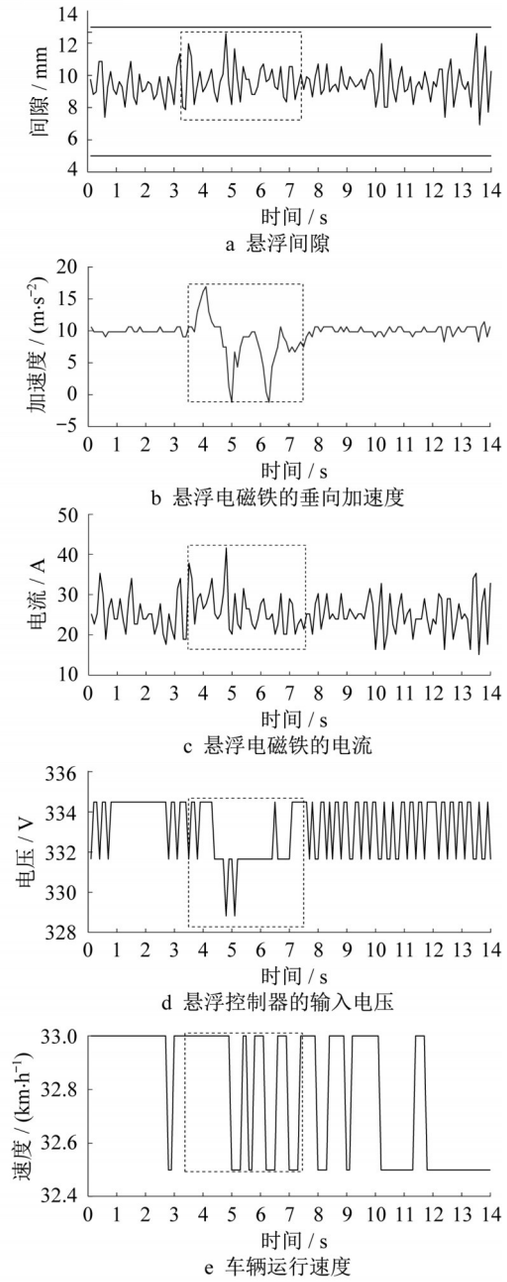

图5　第3类异常的数据

Fig. 5　The third type of abnormal data

1.2.2　站内静止悬浮时异常类型

本文从站内静止悬浮的历史数据中选择含有第4类异常的数据。图6为第4类异常的数据，其中虚线框中的数据属于异常数据。这类异常的现象是间隙没有明显的波动，此时电压不变，但电流有明显的波荡，加速度有剧烈的波动。在实际工程中，当列车与轨道产生共振时容易产生这类异常，但经验阈值方法无法检测出来。

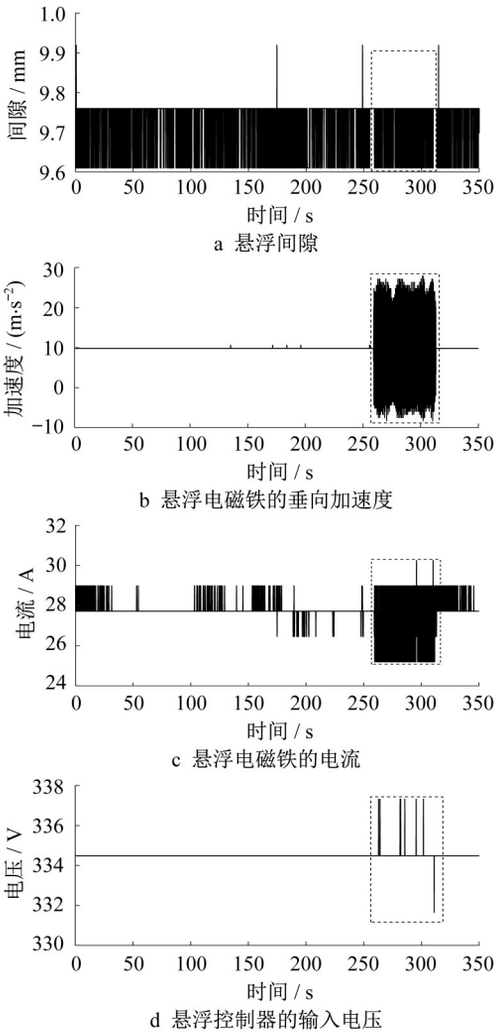

图6　第4类异常的数据

Fig. 6　The fourth type of abnormal data

2　异常检测方法

考虑多维时间序列进行异常检测时，一般会面临两方面的问题：检测异常的指标难以建立和数据呈非高斯分布难以处理。

由于经验阈值方法仅采用悬浮系统的间隙数据，并没有充分利用悬浮系统的其它数据，导致该方法对于第3类和第4类异常的检测能力存在一定的不足。对此，利用悬浮系统的间隙、电流、加速度、电压和速度等数据，采用CCA建立指标。

由于悬浮系统中间隙、电流、加速度、电压和速度等数据并不都服从高斯分布，导致当直接通过CCA处理悬浮数据并获得检测指标，即二次统计量后，二次统计量不服从高斯分布。此时，通过常规的阈值设置方法获得的异常阈值，很容易导致误报或漏报的情况。因此，需要将二次统计量的分布转换

成高斯分布。

对此,通过Box-Cox变换将每种类别建立的二次统计量转换为高斯分布变量,并利用高斯分布的特点来确定不同类别下的异常阈值。

2.1 传统的CCA算法

假设某一个类别下的N个过程数据样本可表示为

$$X_0=\left[x_0(1),x_0(2),\cdots,x_0(N)\right]\in\mathbf{R}^{l\times N} \tag{1}$$

$$Y_0=\left[y_0(1),y_0(2),\cdots,y_0(N)\right]\in\mathbf{R}^{m\times N} \tag{2}$$

式中:$x_0(i)$和$y_0(i)(i=1,\cdots,N)$是在相同类别下测得的过程输入和输出向量;l和m分别为输入和输出的变量个数。

通过式(3)和式(4)去掉平均值,即

$$x(i)=x_0(i)-\mu_x \tag{3}$$

$$y(i)=y_0(i)-\mu_y \tag{4}$$

其中,$\mu_x=\frac{1}{N}\sum_{i=1}^{N}x_0(i)$,$\mu_y=\frac{1}{N}\sum_{i=1}^{N}y_0(i)$。

用X和Y表示去均值后的输入和输出数据,即

$$X=\left[x(1),x(2),\cdots,x(N)\right]\in\mathbf{R}^{l\times N} \tag{5}$$

$$Y=\left[y(1),y(2),\cdots,y(N)\right]\in\mathbf{R}^{m\times N} \tag{6}$$

然后输入和输出的协方差和互协方差可以估算为

$$\boldsymbol{\Sigma}_x\approx\frac{1}{N-1}\sum_{i=1}^{N}\left(x_0(i)-\mu_x\right)\left(x_0(i)-\mu_x\right)^{\mathrm{T}}=\frac{XX^{\mathrm{T}}}{N-1} \tag{7}$$

$$\boldsymbol{\Sigma}_y\approx\frac{1}{N-1}\sum_{i=1}^{N}\left(y_0(i)-\mu_y\right)\left(y_0(i)-\mu_y\right)^{\mathrm{T}}=\frac{YY^{\mathrm{T}}}{N-1} \tag{8}$$

$$\boldsymbol{\Sigma}_{xy}\approx\frac{1}{N-1}\sum_{i=1}^{N}\left(x_0(i)-\mu_x\right)\left(y_0(i)-\mu_y\right)^{\mathrm{T}}=\frac{\mathrm{XY}^{\mathrm{T}}}{N-1} \tag{9}$$

根据CCA技术[28],相关矩阵E定义为

$$E=\boldsymbol{\Sigma}_x^{-\frac{1}{2}}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_y^{-\frac{1}{2}} \tag{10}$$

通过奇异值将相关矩阵E分解为

$$E=\tau\Sigma\mathrm{R}^{\mathrm{T}} \tag{11}$$

式中:$\tau=(\gamma_1,\cdots,\gamma_l)$为相关矩阵的左奇异向量;$\mathrm{R}=(r_1,\cdots,r_m)$为右奇异向量;$\Sigma=\begin{bmatrix}\Sigma_\varrho & 0\\ 0 & 0\end{bmatrix}$为典型相关系数,$\varrho$表示非零奇异值个数。$\mathrm{Rank}(\Sigma_\varrho)=\varrho$,$\Sigma_\varrho=\mathrm{diag}(\lambda_1,...,\lambda_\varrho),1\geqslant\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_\varrho\geqslant0$是典型的相关系数,$\gamma_i(i=1,\cdots,l)$和$r_j(j=1,\cdots,m)$是对应的奇异向量。

令

$$\mathrm{J}=\boldsymbol{\Sigma}_x^{-\frac{1}{2}}\tau(:,1:\varrho) \tag{12}$$

$$\mathrm{L}=\boldsymbol{\Sigma}_y^{-\frac{1}{2}}\mathrm{R}(:,1:\varrho) \tag{13}$$

$$\mathbf{M}^{\mathrm{T}}=\boldsymbol{\Sigma}_\varrho\mathrm{J}^{\mathrm{T}} \tag{14}$$

残差可以定义为

$$\boldsymbol{r}(k)=\mathrm{L}^{\mathrm{T}}\boldsymbol{y}(k)-\mathbf{M}^{\mathrm{T}}\boldsymbol{x}(k) \tag{15}$$

出于检测目的构造了二次统计量$Q(k)$[29]。$Q(k)$的表达式为

$$Q(k)=\boldsymbol{r}^{\mathrm{T}}(k)\,\boldsymbol{r}(k) \tag{16}$$

2.2 基于Box-Cox变换的异常阈值设定

2.2.1 传统阈值设定的不足

根据文献[30]进行阈值设置。

$$J_{\mathrm{th},\mathrm{Q}}=g\chi^2_{1-\alpha}(h) \tag{17}$$

其中,$g=\frac{s}{2\mu_0}$,$h=\frac{2\mu_0^2}{s}$,μ_0和s能被估计。

$$\mu_0=\frac{1}{N}\sum_{k=1}^{N}Q(k) \tag{18}$$

$$s=\frac{1}{N-1}\sum_{k=1}^{N}\left(Q(k)-\mu_0\right)^2 \tag{19}$$

当$Q(k)$为高斯分布时,通过公式(17)获得的阈值会比较理想。然而,当$Q(k)$是非高斯分布时,该方法确定的阈值将产生较大误差。

2.2.2 新阈值的设定

为了确定与系统的不同健康状况相对应的$Q(k)$的范围,可以通过Box-Cox变换将非高斯分布转换为高斯分布[31],然后利用高斯分布的性质来确定$Q(k)$的范围。

Box-Cox转换的过程为通过式(20)将$(p_1,p_2,\cdots,p_n)$转换为$(z_1,z_2,\cdots,z_n)$。

$$z_j(\lambda)=\begin{cases}\dfrac{(p_j^{\lambda}-1)}{\lambda}, & \lambda\neq0\\ \ln p_j, & \lambda=0\end{cases}\quad j=1,2,...,n \tag{20}$$

其中,λ是一个使得每个独立的$p_j(\lambda)$服从正态分布$N(\mu,\sigma^2)$的常数。为了确定λ的值,定义将联合概率密度函数$(p_1(\lambda),p_2(\lambda),\cdots,p_n(\lambda))$为

$$f\left(p_1(\lambda),p_2(\lambda),\cdots,p_n(\lambda)\right)=\left(2\pi\sigma^2\right)^{-\frac{n}{2}}e^{-\frac{1}{2\sigma^2}\left(P(\lambda)-\mu1_n\right)^{\mathrm{T}}\left(P(\lambda)-\mu\mathrm{I}_n\right)} \tag{21}$$

其中,I_n是n阶单位向量。

λ固定时,将μ和σ^2的似然函数表示为

$$L(\mu,\sigma^2|\lambda)=\left(2\pi\sigma^2\right)^{-\frac{n}{2}}e^{-\frac{1}{2\sigma^2}\left(P(\lambda)-\mu I_n\right)^{\mathrm{T}}\left(P(\lambda)-\mu I_n\right)} \tag{22}$$

此外,μ和σ^2的最大似然函数为

$$\hat{\mu}(\lambda)=\frac{1}{n}\sum_{j=1}^{n}p_j(\lambda) \tag{23}$$

$$\hat{\sigma}^2(\lambda)=\frac{1}{n}\sum_{j=1}^{n}\left(p_j(\lambda)-\bar{p}(\lambda)\right)^2 \tag{24}$$

因此,似然函数的最大值为

$$L\left(\hat{\mu}(\lambda),\hat{\sigma}^2(\lambda)|\lambda\right)=\left(2\pi\hat{\sigma}^2(\lambda)\right)^{-\frac{n}{2}}e^{-\frac{n}{2}} \tag{25}$$

通过取式(25)的对数来获得式(26)。

$$\ln L\left(\hat{\mu}(\lambda),\hat{\sigma}^2(\lambda)|\lambda\right)=-\frac{n}{2}\ln\hat{\sigma}^2(\lambda)-\frac{n\ln(2\pi)}{2}-\frac{n}{2} \tag{26}$$

忽略式(26)右端的常数,并将等式记录为$l(\lambda)$,如公式(27)所示。

$$l(\lambda)=-\frac{n}{2}\ln\hat{\sigma}^2(\lambda)=-\frac{n}{2}ln\left[\frac{1}{n}\sum_{i=1}^{n}\left(p_j(\lambda)-\bar{p}(\lambda)\right)^2\right] \tag{27}$$

通过最大似然法确定λ的值。如果存在$\lambda=\lambda_0$,则导致$l(\lambda)=\max_{\lambda}(\lambda)$,则$\lambda=\lambda_0$是适用的。

式(27)仅适用于正数。但是,当存在$p_j<0$时,式(28)可用于p_j。

$$z_i(\lambda)=\begin{cases}\dfrac{\left(\left(p_j+a\right)^{\lambda}-1\right)}{\lambda}, & \lambda\neq0\\ \ln\left(p_j+a\right), & \lambda=0。\end{cases} \tag{28}$$

在经过式(28)处理后,其他步骤与式(20)~(27)相同。

通过Box-Cox变换将$Q(k)$的分布转换为高斯分布后,选择$[\mu-3\sigma,\mu+3\sigma]$作为异常阈值,这意味着在这个范围内约99.73%的样本是健康的[33]。

2.3　算法流程

异常检测的流程如图7所示,它由模型训练和异常检测两部分组成。左侧的虚线框是模型训练,右侧的虚线框表示异常检测。

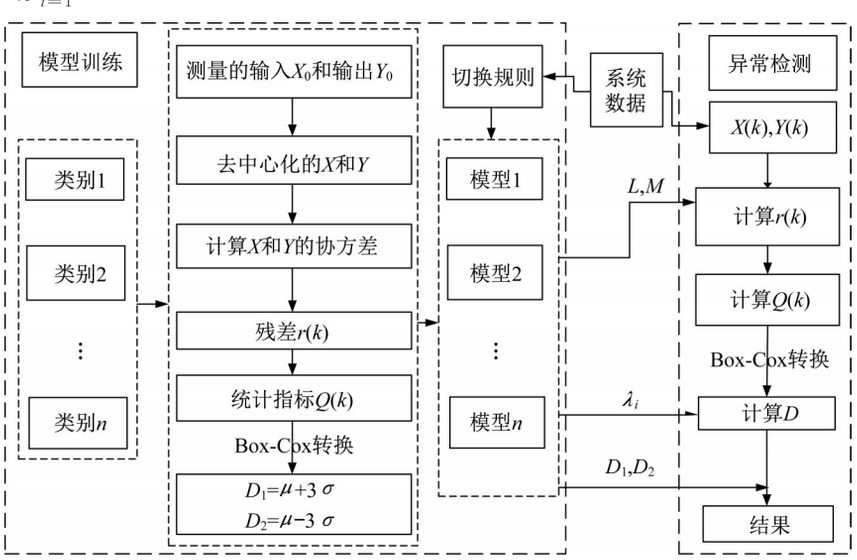

图7　异常检测流程图

Fig. 7　The flow chart of anomaly detection

模型训练主要是通过健康的历史数据获得每个类别下的异常检测模型和用于异常检测的μ_x、μ_y、$\boldsymbol{\Sigma}_x$、$\boldsymbol{\Sigma}_y$、$\boldsymbol{\Sigma}_{xy}$、γ、ρ、L、M^{T}、D_1和D_2。模型训练的步骤为

(1)获得n个类别;

(2)在第n_0个类别下获得N个健康样本,分别构建X_0和Y_0;

(3)根据式(3)~(9)计算$\mu_x,\mu_y,X,Y,\boldsymbol{\Sigma}_x,\boldsymbol{\Sigma}_y,\boldsymbol{\Sigma}_{xy}$;

(4)根据式(11)~(16)计算τ、R、L、M^{T}、r和Q;

(5)根据式(20)、式(21)和式(27)计算λ和z,并根据置信区间计算阈值;

(6)存储该类别下μ_x、μ_y、$\boldsymbol{\Sigma}_x$、$\boldsymbol{\Sigma}_y$、$\boldsymbol{\Sigma}_{xy}$、τ、R、L、M^{T};

(7)$n_0<n$?　是,返回到步骤2);否则,结束。

而异常检测主要是根据数据判断系统当前的类别,然后调取该类别下的μ_x、μ_y、$\boldsymbol{\Sigma}_x$、$\boldsymbol{\Sigma}_y$、$\boldsymbol{\Sigma}_{xy}$、τ、R、L和M^{T}用于计算当前的残差,再将残差与D_1和D_2进行比较。异常检测的步骤为

(1)获得k时刻的数据$x(k)$和$y(k)$;

(2)判断和切换当前的类别;

(3)选择当前类别下$\mu_x,\mu_y,\boldsymbol{\Sigma}_x,\boldsymbol{\Sigma}_y,\boldsymbol{\Sigma}_{xy}$,L和$\mathbf{M}^{\mathrm{T}}$;

(4)根据式(15)和式(16)计算$\mathbf{r}(k)$和$\mathbf{Q}(k)$;

(5)根据式(20)或式(28)计算$z(k)$;

(6)判断:

$\mu-3\sigma<z(k)<\mu+3\sigma\Rightarrow$系统是健康的;

其他$\Rightarrow$系统是异常的。

3 实验结果与分析

3.1 数据处理与分析

使用悬浮系统的监测数据,其中一段数据为健康数据,另一段数据为含有3类异常的数据。由于使用多维数据,因此,只通过长度为4个采样点的移动时间窗口获得140 000组训练数据。

图8为悬浮间隙、电流、电压、加速度和速度的正态分布检验图。从图中可以看出,悬浮间隙、电流、电压、加速度和速度这5个量的分布并不都呈高斯分布。

由于多维数据中有部分数据不呈高斯分布,如果直接采用CCA处理多维数据并获得二次统计量Q,那二次统计量Q的分布也不理想,如图9所示。此时,根据二次统计量Q,利用传统的阈值设置方法所获得的阈值会很不合理,即,当使用式(17)~(19)来计算Q的阈值时,将不可避免地导致较大的误差。通过式(17)~(19)可得,$\mu_0=2.224\,6\times10^{-5}$,$s=7.530\,5\times10^{-10}$,$g=\dfrac{s}{2\mu_0}=1.692\,5\times10^{-5}$,$h=\dfrac{2\mu_0^2}{s}\approx2$,从而其阈值为

$$J_{\mathrm{th},Q}=g\chi^2_{1-\alpha}(h)=1.587\,6\times10^{-5} \tag{29}$$

图10为通过CCA获得的Q值,且这些值绝大部分都大于通过传统阈值设置方法获得的阈值$J_{th,Q}$。在图10中,大量样本明显分布在阈值以上。因此,该方法不可行。

图11为Q的正态分布检验图。从图11可以看出,Q的分布不遵循高斯分布。对此,可通过Box-Cox变换将Q变换为正态分布变量Q_1,即将Q代入式(20)~(28),并获得的参数λ为0.138 2。

图12和图13是Q_1的曲线和分布直方图。由图12和图13可知,Q_1的分布明显比Q的分布更接近高斯分布。为了进一步证明这点,可通过Q_1的正态分布检验图进行直观显示,如图14所示。与图11相比,图14中几乎所有的离散点都分布在直线附近。这表明Box-Cox变换可以有效地将非高斯分布数据转换为高斯分布数据。

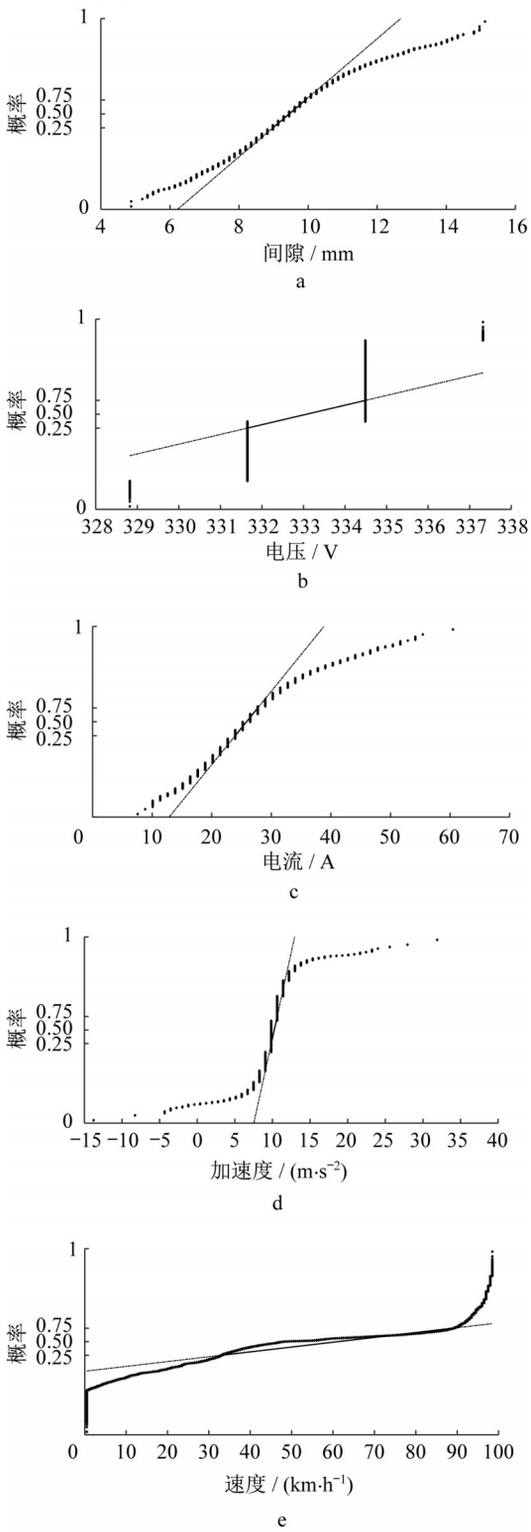

图8 悬浮系统数据的正态分布检验图

Fig. 8 The normal distribution test chart of suspension system data

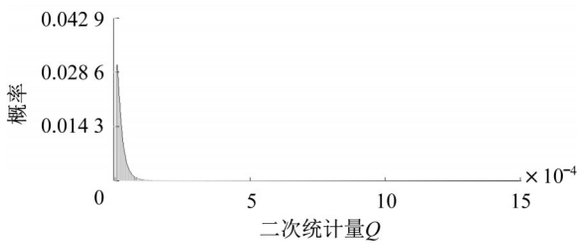

图9　Q的分布直方图

Fig. 9　The distribution histogram of Q

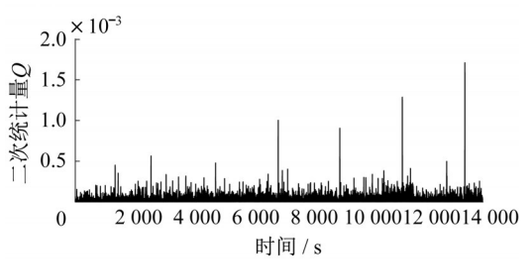

图10　Q值及传统方法的阈值

Fig. 10　Q value and the threshold of the traditional method

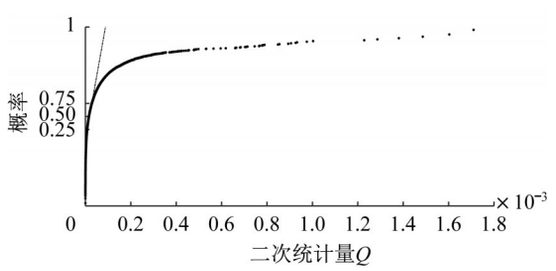

图11　Q的正态分布检验图

Fig. 11　The normal distribution test chart of Q

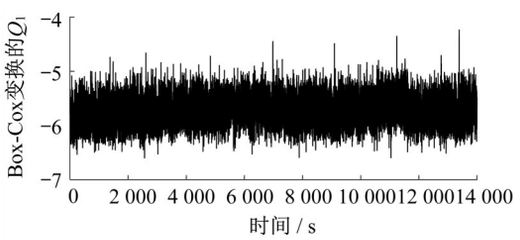

图12　Q_1曲线

Fig. 12　The Q_1 curves

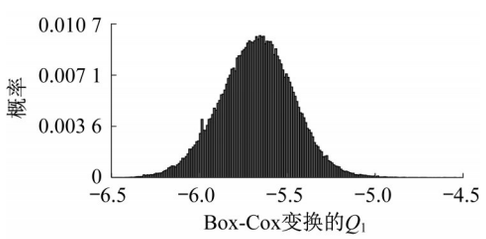

图13　Q_1分布直方图

Fig. 13　The distribution histogram of Q_1

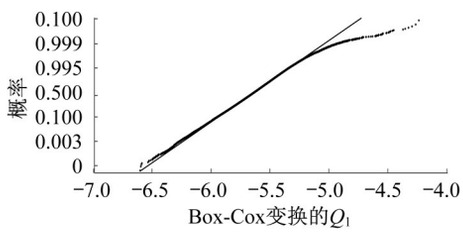

图14　Q_1的正态分布检验图

Fig. 14　The normal distribution test chart of Q_1

3.2　异常检测结果分析

图15为本文方法对站间行驶下3类异常的检测结果。图15的a～f分别为间隙、加速度、电流、电压、速度的测量数据和二次统计量及其对应的阈值，3类异常分别用3个虚线框圈出，从左往右分别是第1类、第2类和第3类异常，且图f中正方形为本文方法检测出的异常点。由图f可知，本文提出的方法能有效检测出3类异常。

图16为本文方法对第4类异常的检测结果，其中直线为本文的阈值。

第4类异常持续的时间为538个采样点，本文的方法能检测出531个点，这说明本文方法能有效检测出第4类异常。

为验证所提方法的有效性，从现有的运营线数据中提取了32个第1类异常数据、104个第2类异常数据、41个第3类异常数据、208个第4类异常数据，分别采用本文的方法、基于K-medoids的方法[32]和基于SVDD的方法[33]计算异常的检测率，计算结果如表1所示。

由表可知，对于第2类异常，3种方法的检测率都为100%；对于第1类和第3类异常，本文方法的检测率明显比另外两种高；对于第4类异常，3种方法的检测率都为100%。综上所述，与基于K-medoids的方法和基于SVDD的方法相比，本文的方法能更有效地检测出4类异常。

4　结语

针对悬浮系统异常检测问题，为进一步提高异常检测率，提出了一种基于改进的CCA的多维时间序列异常检测方法。通过CCA处理悬浮系统的多维数据以获得故障检测指标，即二次统计量。由于悬浮系统中部分数据呈非高斯分布导致二次统计量也呈非高斯分布的问题，使用Box-Cox变换将每种类别下建立的二次统计量转换为高斯分布变量，并

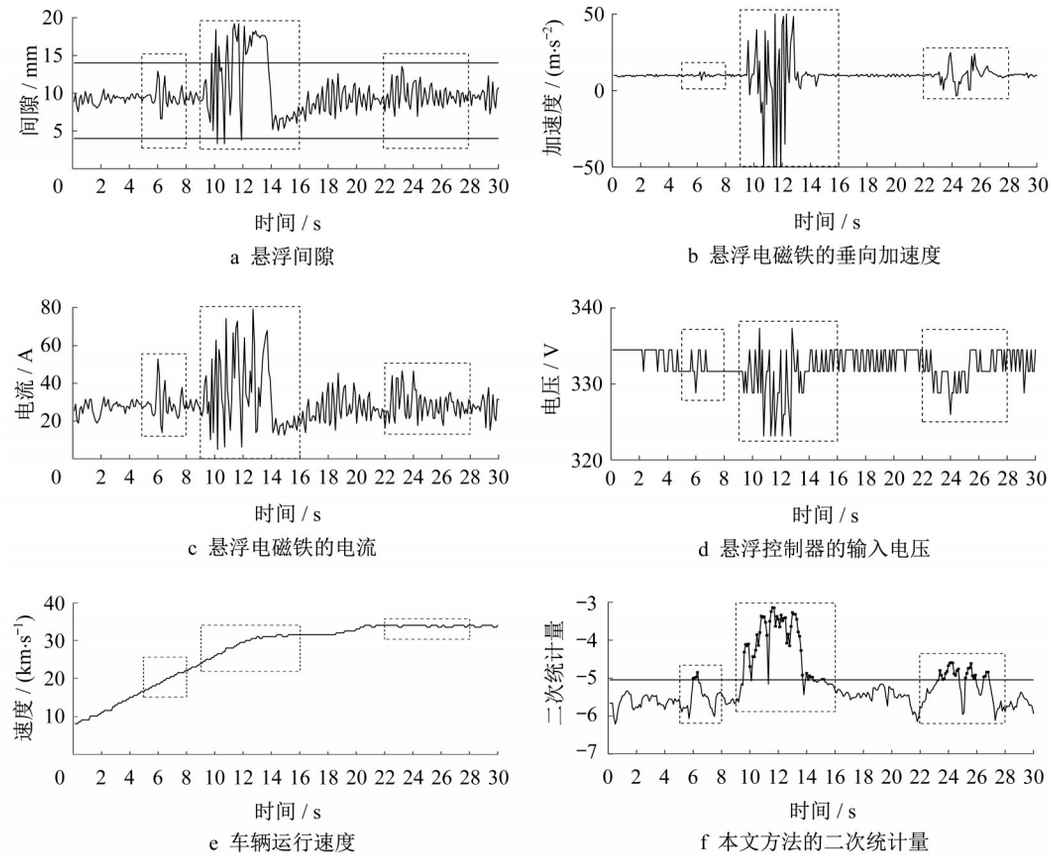

图15 站间行驶下3类异常的检测结果

Fig. 15 The detection results for three types of abnormalities under driving between stations

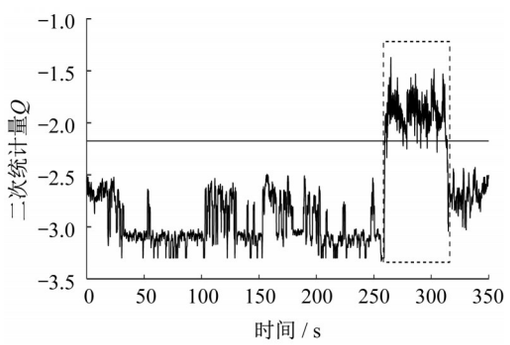

图16 本文方法对第4类异常的检测结果

Fig. 16 The detection result for the fourth type of anomaly by the method in this paper

利用高斯分布的置信区间来确定异常阈值。实验结果表明,本文方法能有效检测出4类异常,而且与基于*K*-medoids的方法和基于SVDD的方法相比,能更有效地检测出4类异常。

本文方法的贡献主要有:

(1) 利用悬浮系统的多维健康数据,提出了一种基于改进的CCA的多维时间序列异常检测方法,该方法对4类异常的检测率都在96%以上。

(2) 使用Box-Cox变换将每种类别下建立的二次统计量转换为高斯分布变量,解决了呈非高斯分布的二次统计量导致传统阈值设定不合理的问题。

表1 检测结果对比

Tab. 1 Comparison of test results

异常的类别	数量	本文方法		基于*K*-medoids的方法		基于SVDD的方法	
		检测出的数量	检测率/%	检测出的数量	检测率/%	检测出的数量	检测率/%
第1类	32	31	96.9	25	71.8	28	87.5
第2类	104	104	100	104	100	104	100
第3类	41	40	97.6	34	82.9	37	90.2
第4类	208	208	100	208	100	208	100

作者贡献声明：

王平：算法研究的执行人，构造新的算法，完成数据分析和实验验证、论文初稿的写作。

梅子：数据分析，论文写作与修改。

龙志强：研究的构思者及负责人。

参考文献：

[1] HARROU F, KADRI F, CHAABANE S, *et al*. Improved principal component analysis for anomaly detection: Application to an emergency department [J]. Computers & Industrial Engineering, 2015, 88: 63.

[2] BI M, XU J, WANG M, *et al*. Anomaly detection model of user behavior based on principal component analysis[J]. Journal of Ambient Intelligence and Humanized Computing, 2016, 7 (4): 547.

[3] XIE K, LI X, WANG X, *et al*. On-Line anomaly detection with high accuracy [J]. IEEE/ACM Transactions on Networking, 2018, 26(3): 1222.

[4] PAN D, LIU D, ZHOU J, *et al*. Anomaly detection for satellite power subsystem with associated rules based on Kernel Principal Component Analysis[J]. Microelectronics Reliability, 2015, 55(9/10): 2082.

[5] YIN C, ZHANG S, YIN Z, *et al*. Anomaly detection model based on data stream clustering [J]. Cluster Computing, 2019, 22(S1): 1729.

[6] GU J, JIANG Z, FAN W D, *et al*. Real-time passenger flow anomaly detection considering typical time series clustered characteristics at metro stations [J]. Journal of Transportation Engineering, Part A: Systems, 2020, 146(4): 04020015.

[7] LI L, DAS S, JOHN HANSMAN R, *et al*. Analysis of flight data using clustering techniques for detecting abnormal operations [J]. Journal of Aerospace Information Systems, 2015, 12(9): 587.

[8] LYU L, JIN J, RAJASEGARAR S, *et al*. Fog-empowered anomaly detection in IoT using hyperellipsoidal clustering [J]. IEEE Internet of Things Journal, 2017, 4(5): 1174.

[9] OREILLY C, GLUHAK A, IMRAN M A. Distributed anomaly detection using minimum volume elliptical principal component analysis [J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(9): 2320.

[10] DING N, MA H, GAO H, *et al*. Real-time anomaly detection based on long short-term memory and Gaussian mixture model [J]. Computers & Electrical Engineering, 2019, 79: 106458.

[11] YANG W, SHEN G W, WANG W, *et al*. Anomaly detection in microblogging via co-clustering [J]. Journal of Computer Science and Technology, 2015, 30(5): 1097.

[12] YAN W. Detecting gas turbine combustor anomalies using semi-supervised anomaly detection with deep representation learning [J]. Cognitive Computation, 2020, 12(2): 398.

[13] LU Y, WANG J, LIU M, *et al*. Semi-supervised machine learning aided anomaly detection method in cellular networks [J]. IEEE Transactions on Vehicular Technology, 2020, 69 (8): 8459.

[14] FENG Q, ZHANG Y, LI C, *et al*. Anomaly detection of spectrum in wireless communication via deep auto-encoders[J]. The Journal of Supercomputing, 2017, 73(7): 3161.

[15] LI Z, LI J, WANG Y, *et al*. A deep learning approach for anomaly detection based on SAE and LSTM in mechanical equipment [J]. The International Journal of Advanced Manufacturing Technology, 2019, 103(1/4): 499.

[16] LIANG P, YANG H D, CHEN W S, *et al*. Transfer learning for aluminium extrusion electricity consumption anomaly detection via deep neural networks [J]. International Journal of Computer Integrated Manufacturing, 2018, 31(4/5): 396.

[17] 罗峰，胡强，侯硕，等. 基于支持向量机的CAN-FD网络异常入侵检测[J]. 同济大学学报(自然科学版)，2020，48(12)：1790.
LUO Feng, HU Qiang, HUO Shuo, *et al*. Anomaly intrusion detection for CAN-FD bus by support vector machine [J]. Journal of Tongji University(Natural Science), 2020, 48(12): 1790.

[18] 何书锋，孙钿奇，王诏，等. 基于深度学习的多波束海底地质数据异常值检测方法[J]. 计算机应用与软件，2021，38(4)：95.
HE Shufeng, SUN Dianqi, WANG Zhao, *et al*. Anomaly detection method fot multibeam seabed geological data based on deeo learning[J]. Computer Applications and Software, 2021, 38(4): 95.

[19] 王慧珍，王立德，杨岳毅，等. 基于Logistic集成学习的列车MVB网络异常检测方法研究[J]. 机车电传动，2021(1)：138.
WANG Huizhen, WANG Lide, YANG Yueyi, *et al*. Anomaly detection for MVB network based on Logistic ensemble learning [J]. Electric Drive for Locomotives, 2021(1): 138.

[20] 李晨，王布宏，田继伟，等. 基于LSTM-OCSVM的无人机传感器数据异常检测[J]. 小型微型计算机系统，2021，42 (4)：700.
WANG Chen, WANG Buhong, TIAN Jiwei, *et al*. Anomaly detection method for UAV sensor data based on LSTM-OCSVM[J]. Journal of Chinese Computer Systems, 2021, 42 (4): 700.

[21] 闫媞锦，夏元清，张宏伟，等. 一种非规则采样航空时序数据异常检测方法[J]. 航空学报，2021，42(4)：558.
YAN Tijin, XIA Yuanqing, ZHANG Hongwei, *et al*. An anomaly detection method for irregularly sampled spacecraft time series data[J]. *Acta Aeronautica ET Astronautica Sinica*,

2021,42(4):558.

[22] 韩昭蓉,黄廷磊,任文娟,等.基于Bi-LSTM模型的轨迹异常点检测算法[J].雷达学报,2019,8(1):36.
HAN Zhaorong, HUANG Tinglei, REN Wenjuan, *et al*. Trajectory outlier detection algorithm based on Bi-LSTM model [J]. Journal of Radars, 2019, 8(1): 36.

[23] 姚宇,冯健,张化光,等.一种基于椭球体支持向量描述的异常检测方法[J].山东大学学报(工学版),2017,47(5):195.
YAO Yu, FENG Jian, ZHANG Huaguang, *et al*. Weighted hyper-ellipsoidal support vector data description with negative samples for outlier detection[J]. Journal of Shandong University (Engineering Science), 2017, 47(5): 195.

[24] 王杰,张雪英,李凤莲,等.改进DM-SVDD算法的异常检测研究及应用[J].太原理工大学学报,2021,52(5):764.
WANG Jie, ZHANG Xueying, LI Fenglian, *et al*. Research and application of anomaly detection based on improved DM-SVDD algorithm [J]. Journal of Taiyuan University of Technology, 2021, 52(5): 764.

[25] 王雯珩,于雷,王晓龙,等.基于夹角余弦的核动力系统异常检测算法设计[J].原子能科学技术,2021,55(S1):98.
WANG Wenheng, YU Lei, WANG Xiaolong, *et al*. Design of anomaly detection algorithm for nuclear power system based on included angle cosine [J]. Atomic Energy Science and Technology, 2021, 55(S1): 98.

[26] 孙笑笑,侯文杰,沈沪军,等.基于上下文感知的多角度业务流程在线异常检测方法[J].计算机集成制造系统,2021,27(9):2532.
SUN Xiaoxiao, HOU Wenjie, SHEN Hujun, *et al*. Multi-perspective online anomaly detection method of business processes based on context awareness[J]. Computer Integrated Manufacturing Systems, , 2021, 27(9), 2532.

[27] 陈志文,彭涛,阳春华, ,等.基于改进的典型相关分析的故障检测方法[J].山东大学学报(工学版),2017,47(5):44.
CHEN Zhiwen, PENG Tao, YANG Chunhua, *et al*. A fault detection method based on modified canonical correlation analysis [J]. Journal of Shandong University (Engineering Science), 2017, 47(5): 44.

[28] ANDERSON T W. An introduction to multivariate statistical analysis[R]. New York:Wiley, 1962.

[29] YIN S, DING S X, HAGHANI A, *et al*. A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process [J]. Journal of Process Control, 2012, 22(9): 1567.

[30] CHEN Z, DING S X, ZHANG K, *et al*. Canonical correlation analysis-based fault detection methods with application to alumina evaporation process[J]. Control Engineering Practice, 2016, 46: 51.

[31] SERDIO F, LUGHOFER E, PICHLER K, *et al*. Residual-based fault detection using soft computing techniques for condition monitoring at rolling mills [J]. Information Sciences, 2014, 259: 304.

[32] NG R, HAN J.CLARANS: a method for clustering objects for spatial data mining[J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(5): 1003

[33] 王振昊,王布宏.基于SVDD的ADS-B异常数据检测[J].河北大学学报(自然科学版),2019,39(3):323.
WANG Zhenhao, WANG Buhong. ADS-B anomaly data detection based on SVDD [J]. Journal of Hebei University (Natural Science Edition), 2019, 39(3): 323.