DOI: 10.19650/j.cnki.cjsi.J1905445

基于 EWT-LOF 的热工过程数据异常值检测方法*

董 泽[1,2],贾 昊[1,2]

(1. 华北电力大学 河北省发电过程仿真与优化控制技术创新中心 保定 071003;
2. 华北电力大学控制与计算机工程学院 北京 102206)

摘 要:异常数据检测是热工过程数据处理的重要组成部分,也是进行系统建模、优化、控制的基础。针对热工过程频繁变工况导致异常数据检测困难的情况,提出一种将信号分解方法与基于密度的检测方法相结合的热工过程异常值检测方法。首先利用经验小波变换方法提取热工过程时间序列的运行趋势,去除序列运行趋势后采用局部离群因子方法对各数据点求取其局部异常值,最后使用箱型图的方法确定序列异常点。通过使用某电厂 1 000 MW 机组的负荷数据作为实验数据,分别设置 0.5%、1%、2%、5%、10% 5 种误差验证方法的有效性。实验结果表明,所提异常检测方法除对动态过程和稳态过程均具有适用性外,在以上 5 种误差条件下均取得了较高的检测准确率。

关键词: 异常数据检测;经验小波变换;局部离群因子;数据预处理;热工过程
中图分类号: TP274 TH81 **文献标识码**: A **国家标准学科分类代码**: 470.20

Outlier detection method for thermal process data based on EWT-LOF

Dong Ze[1,2], Jia Hao[1,2]

(*1.Hebei Technology Innovation Center of Simulation & Optimized Control for Power Generation, North China Electric Power University, Baoding 071003, China; 2.School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China*)

Abstract: Outlier detection is an important part of data processing in thermal process, and is also the basis for system modeling, optimization and control. Aiming at the problem that the operational condition of the thermal process changes frequently, which causes the difficulty of outlier detection, this paper proposes a thermal process outlier detection method combining signal decomposition method and density-based detection method. Firstly, the empirical wavelet transform method is used to extract the operational trend of the thermal process time series. After removing the sequence operational trend, the local outlier factor method is used to obtain the local outlier values for the data points. Finally, the box plot method is used to determine the sequence outlier points. The load data of the 1000MW unit in a certain power plant was used as the experiment data, five errors of 0.5%, 1%, 2%, 5% and 10% were set respectively to verify the effectiveness of the method. The experiment results show that besides having applicability to both dynamic and steady state processes, the outlier detection method proposed in this paper achieves high detection accuracy under the above five error conditions.

Keywords: outlier detection; empirical wavelet transform; local outlier factor; data pre-processing; thermal process

0 引 言

热工生产过程中的过程数据是指火力发电生产过程中传感器采集到的生产数据,如流量、温度、浓度等数据。热工过程数据富含大量机组运行状态的有用信息,是实现机组状态监测、故障诊断、优化运行的基础。在热工实际生产过程中,数以万计的传感器被安装在火电机组的各个子系统中。随着火电机组不断向大容量、高参数的方向发展,热工过程对于传感器采集到的数据的要求也变得越来越高。由于传感器长期工作在高温、高压等极端恶劣条件下,其测量精度会受到很大影响,出现故障的

收稿日期:2019-08-04 Received Date:2019-08-04
*基金项目:中央高校基本科研业务费专项资金(2018QN096)、河北省自然基金(E2018502111)项目资助

概率也随时间的增长而不断增大。运行在异常或故障状态下的传感器采集到的数据会包含大量的异常值,如果在使用前不对这些数据进行处理,被异常值“污染”的运行数据会给后续的建模和控制任务带来巨大的挑战。正是由于数据异常值检测对后续基于数据的任务具有至关重要的作用,因此不仅应用于发电行业[1-3],还被广泛应用于化工[4]、水利[5]、航空航天[6-7]、电力系统[8-9]、钢铁[10]等诸多行业。

Chandola 等[11]和陈斌等[12]分别综述了异常值检测研究领域中的不同方法。根据上述文献,可以将异常检测方法归结于以下 4 类:基于统计预测的检测方法[13]、基于距离的检测方法[14]、基于密度的检测方法[15]和基于机器学习的检测方法[16-17]。基于统计预测的检测方法首先假设待检测的数据符合某种特定类型的分布,将异常点视为严重偏离该分布的数据点,并用分布的不一致性将其检测出来。如果一段数据本身存在多种数据分布,则该方法的检测结果可能会出现较大偏差。基于距离的检测方法通过设定某种距离函数,并计算数据点之间的距离,认为异常点为与其他对象存在较大距离的点。该方法对数据的分布没有硬性要求且检测速度快,但当数据密度发生明显变化时会产生较高的漏检率。基于密度的检测方法将待检测点与邻域内其他点的密度做比较,最终确定其是否为异常点。不同于基于统计和基于距离的异常检测方法选择适用于整个数据集的距离度量函数来对数据的异常进行计算,基于密度的检测方法仅考虑数据对象周围的局部数据分布就可以完成离群点检测。由于真实数据集中的分布模式往往不唯一,故采用基于密度的局部离群检测方法往往可以获得更高的准确度。基于机器学习的方法主要可以分为两大类:一类是人工神经网络,另一类是支持向量机。该方法通过提前建立待测变量的预测模型,根据预测值与实际值的偏离程度来判断是否存在异常。对于该方法而言,预测模型精度会对异常检测结果产生决定性影响。如果所建立模型不够准确,那么异常检测结果的准确性将无法得到保证。

根据现有文献,异常数据检测方法在不同的研究领域各有特点和应用前提,而国内对于火电厂热工过程数据的异常检测研究的文献较少,尚缺乏系统性。此外,热工过程数据的异常检测方法往往与数据预处理方法混合在一起。大部分数据预处理方法仅仅涉及特定的研究课题,未统一对此类问题进行归纳整理,很多关键性问题还有待进一步地深入研究[18]。热工过程常见异常数据检测方法有传统的基于阈值判断的单点检测法、数字滤波方法、统计方法以及基于机器学习方法的新型异常数据检测方法[17,19-21]。单点检测方法通过设定参数基准值和阈值作为判断数据是否异常的标准,一旦超出阈值范围即认为数据异常。该方法使用简单,但准确性完全取决于基准值和阈值的给定,人为影响因素大。数字滤波方法是通过使用如 Wiener、Kalman 等滤波器对数据进行滤波,以消除异常数据[22]。该方法并不注重具体异常值点的检测,且滤波的同时也会对原始数据造成一定影响。统计的方法可以发现偏差明显变大的测量,但对动态过程容易产生误判。对于实际热工过程而言,火电机组运行状态在稳态和动态中不断切换,数据分布并不能简单假设为某一标准分布,这使得该方法应用在热工过程异常数据检测时,检测结果可能会出现较大的偏差。机器学习方法可以对数据进行较好的拟合,缺点是算法复杂度较高且物理意义不明确。一旦拟合不准确,检测结果可能会产生较大偏差。

Breunig 等[23]提出了一种基于密度的局部离群点检测方法,即局部离群因子算法(local outliers factor, LOF)。该方法通过衡量待检测数据与其邻域数据的密度比值来判断其是否为孤立点。局部离群因子方法利用每个数据对象周围邻域的相对密度衡量异常因子,这样的相对密度反映了局部的数据分布,避免了全局数据对异常检测的影响。由于火电机组根据电网调度会频繁的进行升降负荷,机组处于动态-稳态的不断变换之中,直接对热工过程数据使用 LOF 方法进行异常检测会因为数据中含有趋势性而产生很大的误差,因此必须在使用前去除时间序列的运行趋势。经验小波变换(empirical wavelet transform, EWT)[24]是一种处理非平稳特性数据的有效方法。该方法自适应分割信号的傅里叶频谱,通过构造正交小波滤波器组获取到信号的不同模态。火电机组热工过程数据具有非平稳特性,数据特性随工况条件和变量种类的不同,表现出较大的差异性。由于采用常规方法对热工过程数据进行分析比较困难,而 EWT 方法对非平稳性数据的分解具有一定优势,故采用该方法提取热工过程时间序列的运行趋势[25]。

综合以上异常检测方法的优势和不足并结合热工过程数据的特点,通过使用热工过程历史数据进行仿真实验,验证了本文方法的有效性。实验结果表明,本文所提异常检测方法对热工过程动态过程和稳态过程均具有适用性,且检测结果具有较高的准确率。

1　火电机组运行数据特点描述

大型火电机组的结构十分复杂,包括了许多子系统,如燃烧系统、输煤系统、磨煤系统、风烟系统、脱硫脱硝系统、除尘系统、循环水系统等。许多参数对机组的安全性、经济性运行有十分重要的影响。

火电机组运行数据存在如下特点:1)监测点多,关键变量存在多个冗余测点,因此数据量庞大;2)测点所在环境复杂多样;3)数据类型多样,具有较强的动态特

性;4)数据测量和传输过程中往往夹杂了大量噪声。

异常值也被称为野值或离群值,目前尚无公认的准确定义。当前较为普遍的一种定义是:异常值是在数据集中与众不同的数据,使人怀疑是产生于完全不同的机制而非随机偏差下的数据。对于热工过程数据而言,过程数据的异常值可以分为随机误差和显著误差两大类。显著误差主要源于仪表的静差,仪表的精度减小或发生漂移、管线损耗、泄露,不易检测;随机误差为测量变量随机产生,一般服从一定的统计规律,容易去除。由于负荷的变动,显著误差和随机误差对数据异常检测和校正结果的影响也变得更加复杂。当机组处于变工况过程中时,随着负荷的变化会引发机组温度、流量、压力等参数的波动,从而影响整个机组的能量平衡状态。此时机组处于动态过程中,过程对象的特性呈强非线性,不易于使用模型描述。此外,在不同的工况下,机组各对象模型的参数变化也呈非线性,此特点增加了异常检测的复杂程度,同时也增加了数据校正和参数估计的难度。

综上所述,火电机组的状态处于动态-稳态的不断变换中,单一使用的动态和稳态异常值检测方法有时并不适用,一种可以同时应用于动态过程和稳态过程的异常值检测方法是十分必要的。

2 EWT-LOF 异常值检测方法描述

2.1 EWT 方法描述

EWT 方法基于傅里叶支撑选择一组正交小波滤波器组,其中包括一个低通滤波器和 $N-1$ 个带通滤波器分别对应近似和细节分量。原信号 $f(t)$ 可表示为:

$$f(t)=\sum_{i=0}^{N-1}f_i(t) \tag{1}$$

$$f_i(t)=F_i(t)\cos(\phi_i(t)) \tag{2}$$

式中:$f_i(t)$ 为调幅-调频信号,形式如式(2)所示。对 AM-FM 模态进行 Hilbert 变换,以获取有意义的瞬时频率和瞬时幅值,进而得到信号的 Hilbert 谱。式(2)中的 $F_i(t)$ 和 $\phi_i(t)$ 分别为调幅部分函数和调频部分函数。

然后对原始信号 $f(t)$ 的频谱进行连续自适应划分。为满足 Shannon 准则,将傅里叶支撑区间 $[0,\pi]$ 划分为 N 个连续的部分,即 $\Lambda_n=[\omega_{n-1},\omega_n]$(其中 $\omega_0=0,\omega_n=\pi$,ω_n 代表不同部分的边界,$n=1,2,\cdots,N$),即 $\cup_{n=1}^{N}\Lambda_n=[0,\pi]$。以 ω_n 为中心,定义宽度为 $2\tau_n$ 的过渡阶段。在每个 Λ_n 中,定义经验小波为该区间上的带通滤波器。然后以 Meyer 小波为基础构造经验小波,最终得到的经验小波函数和经验尺度函数如式(3)和(4)所示。

$$\hat{\psi}_n(\omega)=\begin{cases}1, & \omega_n+\tau_n\leqslant|\omega|\leqslant\omega_{n+1}-\tau_{n+1}\\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\tau_{n-1}}(|\omega|-\omega_{n+1}+\tau_{n+1})\right)\right] & \omega_{n+1}-\tau_{n+1}\leqslant|\omega|\leqslant\omega_{n+1}+\tau_{n+1}\\ \sin\left[\frac{\pi}{2}\beta\left(\frac{1}{2\tau_n}(|\omega|-\omega_n+\tau_n)\right)\right] & \omega_n-\tau_n\leqslant|\omega|\leqslant\omega_n+\tau_n\\ 0, & \text{其他}\end{cases} \tag{3}$$

$$\hat{\phi}_n(\omega)=\begin{cases}1, & |\omega|\leqslant\omega_n-\tau_n\\ \cos\left[\frac{\pi}{2}\beta\left(\frac{1}{2\tau_n}(|\omega|-\omega_n+\tau_n)\right)\right] & \omega_n-\tau_n\leqslant|\omega|\leqslant\omega_n+\tau_n\\ 0, & \text{其他}\end{cases} \tag{4}$$

式中:$\beta(x)$ 为 $[0,1]$ 区间内满足 K 阶导的任意函数;τ_n 为参数。τ_n 和 $\beta(x)$ 表示如下:

$$\begin{cases}\tau_n=\gamma\omega_n,0<\gamma<1\text{ 且 }\gamma<\min_n\left(\frac{\omega_{n+1}-\omega_n}{\omega_{n+1}+\omega_n}\right)\\ \beta(x)=x^4(35-84x+70x^2-20x^3)\end{cases} \tag{5}$$

定义近似系数和经验小波变换的细节系数分别为 $W_f^{\varepsilon}(0,t)$ 和 $W_f^{\varepsilon}(n,t)$。符号 FFT(·) 和 IFFT^{-1}(·) 分别表示傅里叶变换和傅里叶逆变换。近似系数 $W_f^{\varepsilon}(0,t)$ 和细节系数 $W_f^{\varepsilon}(n,t)$ 分别由信号与经验尺度函数 φ_1 和经验小波函数 ψ_n 的内积表示。

$$W_f^{\varepsilon}(0,t)=<f,\phi_1>=\int f(\tau)\,\overline{\phi_1(\tau-t)}\,\mathrm{d}\tau=\text{IFFF}^{-1}(\hat{f}(\omega)\hat{\phi}_1(\omega)) \tag{6}$$

$$W_f^{\varepsilon}(n,t)=<f,\psi_n>=\int f(\tau)\,\overline{\psi_n(\tau-t)}\,\mathrm{d}\tau=\text{IFFF}^{-1}(\hat{f}(\omega)\hat{\psi}_n(\omega)) \tag{7}$$

式(6)和(7)中的 $\hat{\phi}_1(\omega)$ 和 $\hat{\psi}_n(\omega)$ 分别是 $\phi_1(\omega)$ 和 $\psi_n(\omega)$ 的傅里叶变换,如式(3)和(4)所示;$\overline{\phi_1(\omega)}$ 和 $\overline{\psi_n(\omega)}$ 分别是 $\phi_1(\omega)$ 和 $\psi_n(\omega)$ 的共轭复数。得到原信号重构公式:

$$f(t)=W_f^{\varepsilon}(0,t)*\phi_1(t)+\sum_{n=1}^{N}W_f^{\varepsilon}(n,t)*\psi_n(t)=\text{IFFT}^{-1}(\hat{W}_f^{\varepsilon}(0,t)\times\hat{\phi}_1(t)+\sum_{n=1}^{N}\hat{W}_f^{\varepsilon}(n,t)\times\hat{\psi}_n(t)) \tag{8}$$

式中:$\hat{W}_f^{\varepsilon}(n,t)$ 和 $\hat{W}_f^{\varepsilon}(0,t)$ 分别代表 $W_f^{\varepsilon}(n,t)$ 和 $W_f^{\varepsilon}(0,t)$ 的傅里叶变换。根据式(8)、(2)中的经验模态 f_i 可定义为:

$$f_0(t)=W_f^{\varepsilon}(0,t)*\phi_1(t) \tag{9}$$

$$f_i(t)=W_f^{\varepsilon}(i,t)*\psi_i(t) \tag{10}$$

式中:“ * ”表示卷积。最终,通过使用 EWT,给定的真实信号可以分解为许多具有紧凑支持频谱的经验模态函数。

2.2　LOF 方法描述

LOF 算法主要涉及的概念有数据对象的 k-距离、k-距离邻域、数据对象的可达距离、可达密度和局部离群因子。相关概念定义如下:

定义 1(对象 p 的 k 距离 $d_k(p)$):

设 k 为一正整数,数据对象 p 的 k 距离记作 $d_k(p)$。在数据集 $\boldsymbol{D}$ 中,将两个数据对象 p 与 o 的距离记作 $d(p,o)$。

若使得 $d_k(p)=d(p,o)$,需满足:

1)数据集 $\boldsymbol{D}$ 中至少存在不包括 p 的 k 个点 $o'\in C/\{p\}$,满足 $d(p,o')\leqslant d(p,o)$。

2)数据集 $\boldsymbol{D}$ 中至多存在不包含 p 在内的 $k-1$ 个点 $o'\in C/\{p\}$,满足 $d(p,o')<d(p,o)$。

定义 2(对象 p 的第 k 距离邻域):

数据对象 p 的 k 距离邻域,就是所有与 p 的距离小于等于 k 的距离的数据对象 o 的集合,即:

$$N_k(p)=\{o\in \boldsymbol{D}\mid d(p,o)\leqslant d_k(p)\} \tag{11}$$

定义 3(可达距离):

数据对象 p 和 o 的可达距离记为:

$$reach-dist_k(p,o)=\max\{d_k(p),d(p,o)\} \tag{12}$$

其中 k 为一正整数。

定义 4(局部可达密度):

数据对象 p 的局部可达密度表示点 p 的第 k 邻域内到点 p 的平均可达距离的倒数,即:

$$lrd_k(p)=1\Big/\left(\frac{\sum\limits_{o\in N_k(p)} reach-dist_k(p,o)}{|N_k(p)|}\right) \tag{13}$$

定义 5(局部离群因子):

数据对象 p 的局部离群因子表示点 p 的邻域 $N_k(p)$ 的局部可达密度与点 p 的局部可达密度之比的平均数,即:

$$LOF_k(p)=\frac{\sum\limits_{o\in N_k(p)}\dfrac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|} \tag{14}$$

若 LOF 值越接近于 1,表明 p 与其邻域对象密度越接近,越可能与邻域同属一簇;若 LOF 值越大于 1,则说明 p 的密度小于其邻域点密度越多,越可能是异常点。

2.3　EWT-LOF 异常值检测方法描述

1)算法流程

输入:数据集 $\boldsymbol{D}$,距离 K,箱型图截断点系数 β;

输出:异常值位置;

算法过程描述:

(1)使用 EWT 方法提取原始序列的最低频分量,作为序列的运行趋势;

(2)从原始序列中去除序列的运行趋势,得到新的序列;

(3)计算新序列的 LOF 值;

(4)使用箱型图方法自适应确定异常值位置。

2)算法思想

对于热工过程而言,由于机组受到负荷调度指令和自身煤质变化的影响,经常处于一种“动态-稳态”的波动状态。在这种情况下,通过建立对象模型来获取其运行趋势变得十分困难。而使用 EWT 方法从频域角度对机组运行数据进行分析,其低频分量可以反应机组的实际运行情况。另外由于异常值属于短时突变值,低频分量上受其影响很小,通过低频分量描述其运行趋势也更加准确。

对原始序列去除其运行趋势后再求其 LOF 值,并不会遗漏异常值信息,反而有利于消除在变工况时对变量求取 LOF 值的影响,提高异常值检测方法的准确性。

处于不同工况下的不同热工过程变量的特性并不完全相同,因此给出一个合适的异常值判断阈值是十分困难的,使用箱型图方法可以避免直接给出异常值判断阈值。箱型图方法是一种可以体现数据分散情况的统计图方法,可以描述数据的分布差异。该方法不需要事先对数据分布进行假定,对数据没有任何限制性要求,只是真实直观地表现数据形状的本来面貌。另外,箱型图以四分位数和四分位距作为判断异常值的基础,异常值不易对这个标准施加影响,异常值识别结果比较客观。

定义 IQR 为四分位距,即上下四分位数 Q_3 和 Q_1 的差值。

$$IQR=Q_3-Q_1 \tag{15}$$

定义异常截断点如式(16)所示。

$$\text{异常截断点}=\begin{cases}Q_3+\beta IQR\\ Q_1-\beta IQR\end{cases} \tag{16}$$

当取 $\beta=1.5$ 时,称其为内限;当取 $\beta=3$ 时,称其为外限。处于内限和外限之间的异常值称为温和异常值,处于外限之外的异常值称为极端异常值。对于其最终取值,要视数据的实际情况而定。

本文方法整体流程如图 1 所示。

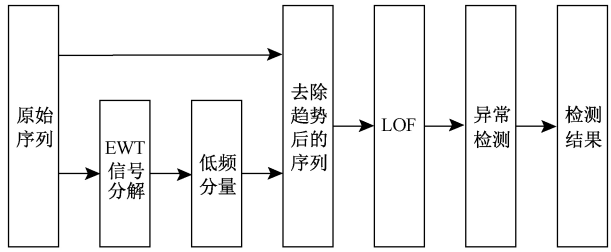


图 1　异常检测流程

Fig.1　Flow chart of outlier detection

3 热工过程历史数据异常值检测

本文实验采用某电厂 1 000 MW 机组的负荷数据验证所提方法的有效性，同时对实验结果进行分析。所用历史数据采样时间为 5 000 s，采样间隔设定为 5 s，共计 1 000 组。该段数据既包括动态，也包括稳态，用来验证方法对两种状态的适用性。分别在负荷原始数据的基础上在特定点处添加 0.5%、1%、2%、5%、10%的误差，用以模拟系统异常情况。在样本序列 80、110、860～863 处分别添加+0.5%、+1%、+2%、+5%、+10%的误差；在 200、300、500～503 处分别添加－0.5%、－1%、－2%、－5%、－10%的误差，用以模拟可能出现的误差情况。

采用误判率和漏检率来评判所提方法的检测效果，定义公式如下：

$$\text{误判率}=\frac{\text{误判个数}}{\text{检测结果个数}} \tag{17}$$

$$\text{漏检率}=\frac{\text{漏检个数}}{\text{真实异常值个数}} \tag{18}$$

在 1%误差下某电厂负荷历史运行数据如图 2 所示。从图 2 中可以看出机组负荷从 770 MW 左右升至 820 MW左右。其中包括 3 段稳定工况，即 770、815、820 MW。红色实线为采用 EWT 方法获取的原数据的低频分量，即负荷曲线的运行趋势。由图 2 可知，采用 EWT 方法得到的低频分量受到异常点的影响很小，可以很好的表示负荷曲线的运行趋势。

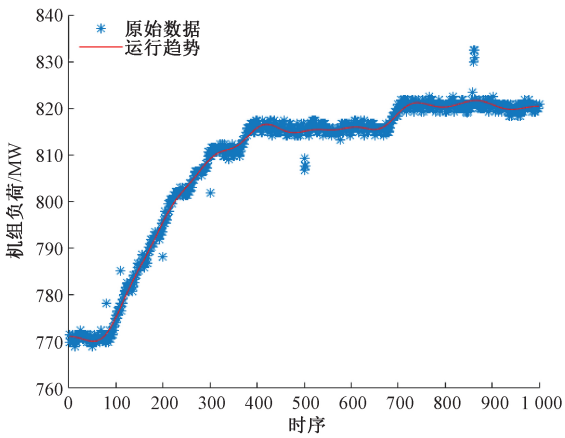


图 2 1%误差时负荷数据与运行趋势对比曲线

Fig.2 Comparison curves of load data vs. operational trend for the error of 1%

图 3～5 所示为在 1%误差下取 $K=10$，$\beta=3$ 时的异常值检测结果。去除运行趋势后数据的 LOF 值曲线如图 3(a)所示，可以看出异常值点处的 LOF 值明显比正常点处的 LOF 值大。原始序列与运行趋势的相对偏差曲线如图 3(b)所示，可以看出正常值点和异常值点存在明显的不同，很容易区分出正常值点和异常值点。

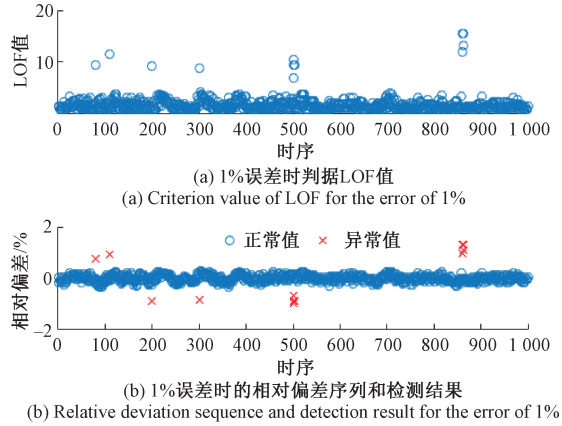


图 3 1%误差时的 LOF 值序列和相对偏差序列

Fig.3 LOF value sequence and relative deviation sequence for the error of 1%

未去除运行趋势数据的 LOF 值曲线如图 4 所示。与图 3(a)对比可知，图 4 所示的 LOF 值曲线明显带有变量的运行趋势，此时无法仅凭借各点的 LOF 值区分出异常值点和正常值点，这将导致基于 LOF 值的异常检测方法失效。因此，对于存在工况变化情况的数据，在检测前去除待检测变量的运行趋势是十分必要的。

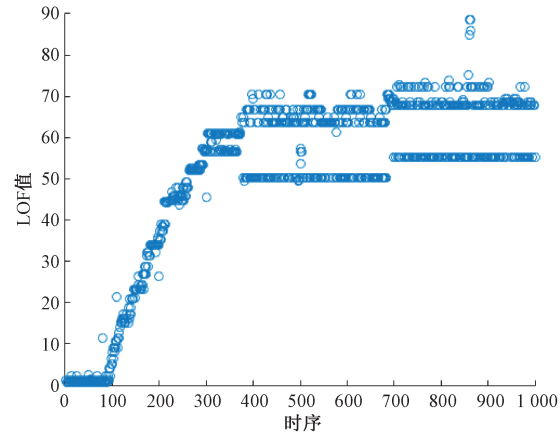
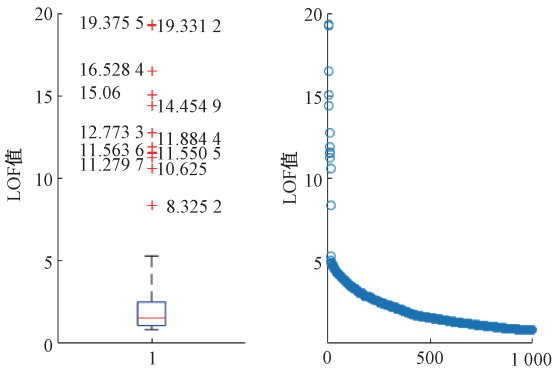


图 4 1%误差时原始序列的 LOF 值序列

Fig.4 LOF value sequence of original sequence for the error of 1%

1%误差时负荷数据最终的异常值检测结果如图 5 所示。图 5(a)所示为用箱型图展示的检测结果；图 5(b)所示为与箱型图对应的，按照各点 LOF 值由大到小依次重新排列的序列。从图 5 中可以看出，异常值点和正常值点的 LOF 值存在明显的区别，并且异常值点的 LOF 值均距离箱型图的外限较远，此时所有异常值点均可以被检测出来。



(a) 1%误差时箱型图检测结果　(b) 1%误差序列LOF值排序

(a) Box-plot test result for the error of 1%　(b) Sequence LOF value sorting for the error of 1%

图 5　1%误差时负荷数据异常检测结果

Fig.5　Outlier detection result of load data for the error of 1%

为了检验不同 K 值可能对检测结果带来的影响,在 1%、2%、5%、10% 4 种误差条件下,取 $\beta=3$ 时,根据不同 K 值得到的异常值检测结果如表 1 所示。

表 1　异常值检测结果

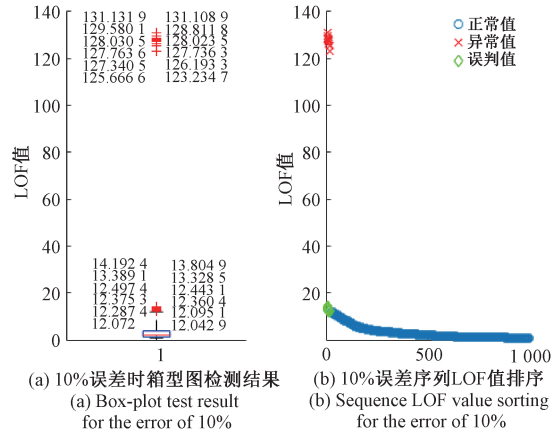
Table 1　Outlier value detection result

误差大小/%	参数	$K=5$	$K=10$	$K=15$	$K=20$
1	误判率	0/12	0/12	0/12	0/12
	漏检率	0/12	0/12	0/12	0/12
2	误判率	0/12	0/12	0/12	0/12
	漏检率	0/12	0/12	0/12	0/12
5	误判率	0/12	0/12	0/12	0/12
	漏检率	0/12	0/12	0/12	0/12
10	误判率	13/25	12/24	13/25	16/28
	漏检率	0/12	0/12	0/12	0/12

由表 1 可知,当取截断点系数 $\beta=3$ 时,1%、2%、5% 3 种误差在使用不同的 K 距离时得到了相同的检测结果,误判率和漏检率均为 0,这也再次证明了本文方法具有较好的检测能力。另外,结合 10% 误差时的检测结果,对于本文实验所用负荷数据,选取 $K=10$ 即可满足检测需求。

另外,由表 1 可知,当误差为 10% 时,本文方法除了可以检测出所有异常值外,同时也产生了一定程度的误判,且误判率与选择的 K 距离之间不存在明显的规律性。下面对 10% 误差、$K=10$ 时的异常检测结果进行分析。

10% 误差时负荷数据的异常检测结果如图 6 所示。结合图 6 和表 1 可知,当取 $K=10$ 时,共检测出异常值点 24 个,其中前 12 个点为真正的异常值点,后 12 个点实际为误判点。从图 6 还可以明显看出,前 12 个真正异常值点的 LOF 值远远大于其他点。之所以产生误判,主要是由于真实异常值点的 LOF 值过大,而其他正常值点的 LOF 值较小,这导致上四分位数 Q_3 较小,进而导致四分位距较小,极端异常值上限较小,最终导致了检测阈值较小,使得一部分正常值点被误判为异常值点。



(a) 10%误差时箱型图检测结果　(b) 10%误差序列LOF值排序

(a) Box-plot test result for the error of 10%　(b) Sequence LOF value sorting for the error of 10%

图 6　10%误差时负荷数据异常检测结果

Fig.6　Outlier detection result of load data for the error of 10%

另外,为了进一步验证本文方法对于较小幅度波动的异常值的检测效果,还对 0.5% 误差情况下的数据进行了异常值检测实验。由于异常值幅度较小,实验除了对多种 K 距离的结果进行对比外,还对采用了不同的箱型图截断点系数 β 的结果进行了对比和讨论,最终实验结果如表 2 所示。

表 2　0.5%误差下异常值检测结果

Table 2　Outlier value detection result for the error of 0.5%

截断点系数	参数	$K=5$	$K=10$	$K=15$	$K=20$
$\beta=1.5$	误判率	5/16	13/23	15/25	29/40
	漏检率	1/12	2/12	2/12	1/12
$\beta=3$	误判率	0/6	0/5	0/5	0/6
	漏检率	6/12	7/12	7/12	6/12

由表 2 可知,在 0.5% 误差条件下,当 $K=5$、$\beta=1.5$ 时,方法得到最好的检测结果。这主要是由于当误差幅值较小时,异常点与正常点在数值上十分接近,此时采用较小的 K 距离和截断点系数对异常值进行局部“细选”,可以得到很好的效果。但是对于热工过程实际情况来说,由于数据存在波动且波动幅度不一,有时会忽略 0.5% 的误差,将其归类为数据的正常波动。另外,为了与表 1 中所列的 4 种不同误差实验保持一致性,在 0.5% 误差实验中仍选择 $K=10$ 时的检测结果进行分析。

0.5% 误差时,去除运行趋势后数据的 LOF 值序列如图 7 所示。由图 7 可知,由于误差较小,除了在 110、860~863 处的 5 个异常值点的 LOF 值较为突出外,其他

7 个异常值点与正常点的 LOF 值相差并不十分明显。

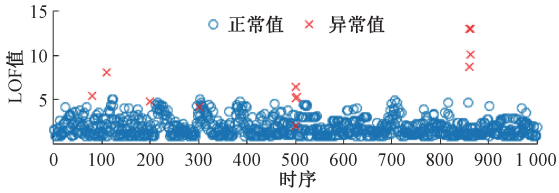


图 7 0.5% 误差时的 LOF 值序列

Fig.7 LOF value sequence for the error of 0.5%

0.5% 误差、$\beta=1.5$ 时的相对偏差序列和检测结果如图 8 所示。由图 8 可知，由于此时的异常值检测阈值较小，导致点 300 和 500 被漏检，同时也导致了 13 个正常值点被误判为异常值点。0.5% 误差、$\beta=1.5$ 时数据的异常检测结果箱型图如图 9 所示。从图 9 中可知，异常值点和误判值点的 LOF 值十分接近，且均高于箱型图的内限；而两个漏检值点和正常值点的 LOF 值十分接近，无法区分。

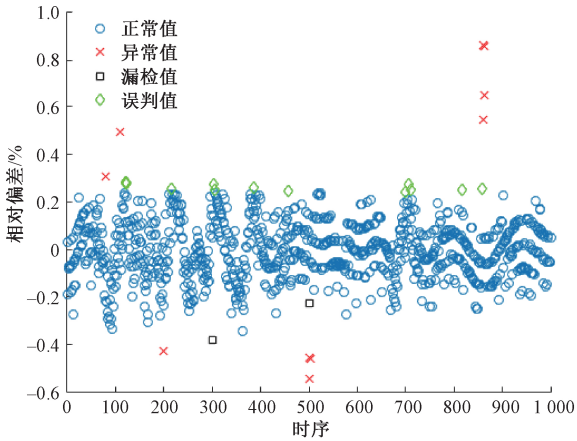
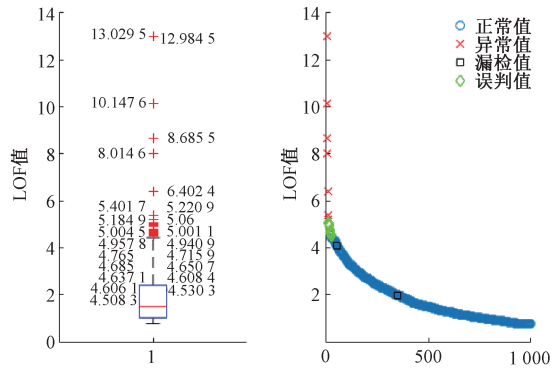


图 8 0.5% 误差时的相对偏差序列和检测结果($\beta=1.5$)

Fig.8 Relative deviation sequence and detection result for the error of 0.5% ($\beta=1.5$)

0.5% 误差、$\beta=3$ 时的相对偏差序列和检测结果如图 10 所示。由图 10 可知，由于此时的异常值检测阈值较大，只有 110、860～863 处的 5 个异常值点被检测出来，另外 7 个点被漏检。对应的异常检测结果箱型图如图 11 所示。由图 11 可知，只有上述 5 个异常值点的 LOF 值超过了箱型图外限，漏检值点中除有 1 个点接近箱型图的外限，其他点均与正常值点近似且距离外限较远。

综上所述，在 0.5% 误差时，由于误差幅度较小，导致了异常值点和真实值点混杂在一起，不易分辨。因此检测结果与 1%、2%、5%、10% 误差时相比，有一定的差距。



(a) 0.5%误差时箱型图检测结果 (a) Box-plot test result for the error of 0.5%

(b) 0.5%误差序列LOF值排序 (b) Sequence LOF value sorting for the error of 0.5%

图 9 0.5% 误差时负荷数据异常检测结果($\beta=1.5$)

Fig.9 Outlier detection result of load data for the error of 0.5% ($\beta=1.5$)

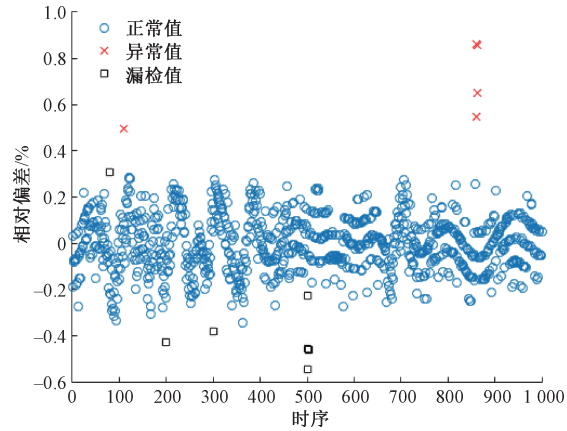
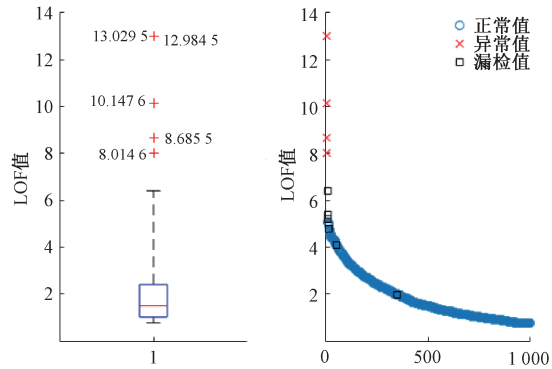


图 10 0.5% 误差时的相对偏差序列和检测结果($\beta=3$)

Fig.10 Relative deviation sequence and detection result for the error of 0.5% ($\beta=3$)



(a) 0.5%误差时箱型图检测结果 (a) Box-plot test result for the error of 0.5%

(b) 0.5%误差序列LOF值排序 (b) Sequence LOF value sorting for the error of 0.5%

图 11 0.5% 误差时负荷数据异常检测结果($\beta=3$)

Fig.11 Outlier detection result of load data for the error of 0.5% ($\beta=3$)

4 结　论

异常数据检测是热工过程数据处理的重要组成部分,也是后续对系统进行建模、优化和控制的基础,具有十分重要的作用。鉴于机组工况和煤质变化导致热工过程数据经常发生波动的实际情况,本文提出了一种热工过程异常值检测方法。该方法在使用 EWT 方法对原始热工过程时间序列的运行趋势进行去除后,使用 LOF 方法得到该序列的局部异常值序列;通过使用箱型图确定序列中的异常值点,从而避免了判断阈值不易直接给出的问题。通过使用某电厂 1 000 MW 机组的负荷数据作为实验数据,分别设置 0.5%、1%、2%、5%、10% 5 种误差验证方法的有效性。实验结果表明,本文所提异常检测方法除对动态过程和稳态过程均具有适用性外,在以上 5 种误差条件下均取得了较高的检测准确率。

但是,需要注意的是,本文所提方法虽然在一定程度上解决了变工况数据的异常值检测问题,但仍然需要完善。因此,下一步的研究工作主要集中在以下两个方面:

1)由于 K 距离和箱型图截断点系数 β 的选择会对最终的检测结果造成影响,在下一步工作中要尝试引入某种自适应机制,使得可以根据数据实际情况选取合适的参数。

2)本文方法适用于一维数据的异常值检测,不适用于高维数据。因此,在下一步工作中还要尝试将其扩展为高维数据的异常值检测方法。

参考文献

[1] 赵悦,方彦军,董政呈.基于状态识别的经验模态分解法火电厂运行数据预处理[J].热力发电,2019,48(1):49-54.
ZHAO Y, FANG Y J, DONG ZH CH. Operating data preprocessing using EMD method with state recognition for thermal power plants[J]. Thermal Power Generation, 2019, 48 (1):49-54.

[2] 司风琪,徐治皋.基于自联想神经网络的测量数据自校正检验方法[J].中国电机工程学报,2002(6):153-156.
SI F Q, XU ZH G. Self-verifying data validation method based on the autoassociative neural network (AANN)[J]. Proceedings of the CSEE, 2002(6):153-156.

[3] QI M, FU Z, CHEN F, et al. Outliers detection method of multiple measuring points of parameters in power plant units [J]. Applied Thermal Engineering, 2015, 85: 297-303.

[4] 苏卫星,朱云龙,胡琨元,等.基于模型的过程工业时间序列异常值检测方法[J].仪器仪表学报,2012,33(9):2080-2087.
SU W X, ZHU Y L, HU K Y, et al. Model-based outlier detection method for time series of process industry[J]. Chinese Journal of Scientific Instrument, 2012, 33(9): 2080-2087.

[5] 张峰,薛惠锋,WANG W,等.水资源监测异常数据模态分解-支持向量机重构方法[J].农业机械学报,2017,48(11):316-323.
ZHANG F, XUE H F, WANG W, et al. Methods of abnormal data detection and recovery for water resources monitoring based on EEMD and PSO-LSSVM [J]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(11):316-323.

[6] 彭喜元,庞景月,彭宇,等.航天器遥测数据异常检测综述[J].仪器仪表学报,2016,37(9):1929-1945.
PENG X Y, PANG J Y, PENG Y, et al. Review on anomaly detection of spacecraft telemetry data [J]. Chinese Journal of Scientific Instrument, 2016, 37(9): 1929-1945.

[7] 彭宇,何永福,王少军,等.飞行数据异常检测技术综述[J].仪器仪表学报,2019,40(3):1-13.
PENG Y, HE Y F, WANG SH J, et al. Flight data anomaly detection: a survey [J]. Chinese Journal of Scientific Instrument, 2019, 40(3):1-13.

[8] 赵天辉,王建学,马龙涛,等.基于非参数回归分析的工业负荷异常值识别与修正方法[J].电力系统自动化,2017,41(18):53-59.
ZHAO T H, WANG J X, MA L T, et al. Outlier detection and correction method for industrial loads based on nonparametric regression analysis[J]. Automation of Electric Power Systems, 2017, 41(18):53-59.

[9] 孙毅,李世豪,崔灿,等.基于高斯核函数改进的电力用户用电数据离群点检测方法[J].电网技术,2018,42(5):1595-1606.
SUN Y, LI SH H, CUI C, et al. Improved outlier detection method of power consumer data based on gaussian kernel function[J]. Power System Technology, 2018, 42(5):1595-1606.

[10] ZHAO J, LIU K, WANG W, et al. Adaptive fuzzy clustering based anomaly data detection in energy system of steel industry[J]. Information Sciences, 2014, 259: 335-345.

[11] CHANDOLA V, BANERJEE A, KUMAR V, et al. Anomaly detection: A survey [J]. ACM Computing Surveys, 2009, 41(3): 1-58.

[12] 陈斌,陈松灿,潘志松,等.异常检测综述[J].山东大学学报(工学版),2009,39(6):13-23.
CHEN B, CHEN S C, PAN ZH S, et al. Survey of outlier detection technologies[J]. Journal of Shandong University(engineering science), 2009, 39(6):13-23.

[13] 王希若,荣冈.高置信度的显著误差综合检测法[J].仪器仪表学报,2000,21(2):196-199.

WANG X R, RONG G. A compound test with high confidence level for gross error detection[J]. Chinese Journal of Scientific Instrument, 2000,21(2):196-199.

[14] 马贺贺,胡益,侍洪波.基于马氏距离局部离群因子方法的复杂化工过程故障检测[J].化工学报,2013,64(5):1674-1682.
MA H H, HU Y, SHI H B. Fault detection of complex chemical processes using Mahalanobis distance-based local outlier factor[J]. CIESC Journal, 2013, 64(5): 1674-1682.

[15] 周世波,徐维祥.一种基于偏离的局部离群点检测算法[J].仪器仪表学报,2014,35(10):2293-2298.
ZHOU SH B, XU W X. Deviation-based local outlier detection algorithm[J]. Chinese Journal of Scientific Instrument, 2014, 35(10):2293-2298.

[16] 窦珊,张广宇,熊智华.基于 LSTM 时间序列重建的生产装置异常检测[J].化工学报,2019,70(2):481-486.
DOU SH, ZHANG G Y, XIONG ZH H. Anomaly detection of process unit based on LSTM time series reconstruction[J]. CIESC Journal, 2019, 70(2): 481-486.

[17] 王雷,张瑞青,盛伟,等.基于支持向量机的回归预测和异常数据检测[J].中国电机工程学报,2009,29(8):92-96.
WANG L, ZHANG R Q, SHENG W, et al. Regression forecast and abnormal data detection based on support vector regression[J]. Proceedings of the CSEE, 2009, 29(8):92-96.

[18] 靳涛. 火电机组反向建模方法的研究[D]. 北京:华北电力大学,2011.
JIN T. Research on reversed modeling method for thermal power unit[D]. Beijing: North China Electric Power University,2011.

[19] 吴盈,司风琪,徐治皋.基于样条变换偏鲁棒 M-回归的电站热力过程数据检验[J].中国电机工程学报,2011,31(8):114-118.
WU Y, SI F Q, XU ZH G. Data validation of thermodynamic system in power plant based on the partial robust M-regression of splines[J]. Proceedings of the CSEE, 2011, 31(8):114-118.

[20] 周卫庆,司风琪,乔宗良,等.基于稳健估计的迭代型支持向量机及其在电站数据检验中的应用[J].中国电机工程学报,2011,31(11):113-118.
ZHOU W Q, SI F Q, QIAO Z L, et al. Iterative support vector machine based on robust estimation and its application in data validation in power plant[J]. Proceedings of the CSEE, 2011, 31(11):113-118.

[21] 司风琪,洪军,徐治皋.基于改进 Elman 网络的动态系统测量数据检验方法[J].东南大学学报(自然科学版),2005(1):50-54.
SI F Q, HONG J, XU ZH G. Dynamic system data validation method based on the improved Elman network[J]. Journal of Southeast University (Natural Science Edition), 2005(1):50-54.

[22] 王艳婷,史元浩,陈晓龙.基于无迹卡尔曼滤波预测的锅炉吹灰优化[J].电子测量与仪器学报,2019,33(3):51-57.
WANG Y T, SHI Y H, CHEN X L. Boiler soot blowing optimization based on unscented Kalman filter prediction[J]. Journal of Electronic Measurement and Instrumentation, 2019, 33(3):51-57.

[23] BREUNIG M M, KRIEGEL H, NG R T, et al. LOF: identifying density-based local outliers[J]. International Conference on Management of Data, 2000, 29(2): 93-104.

[24] GILLES J. Empirical wavelet transform[J]. IEEE Transactions on Signal Processing, 2013, 61(16): 3999-4010.

[25] 贾昊,董泽,闫来清.基于信号分解和统计假设检验的稳态检测方法[J].仪器仪表学报,2018,39(10):150-157.
JIA H, DONG Z, YAN L Q. Steady-state detection method based on signal decomposition and statistical hypothesis test[J]. Chinese Journal of Scientific Instrument, 2018, 39(10):150-157.

作者简介

董泽,1992 年、1995 年和 2001 年于华北电力大学分别获得学士、硕士和博士学位,现为华北电力大学教授,博士生导师,主要研究方向为大型火电机组建模理论与方法研究、智能控制理论及应用等。
E-mail:dongze33@126.com

Dong Ze received his B. Sc., M. Sc. and Ph. D. degrees all from North China Electric Power University in 1992, 1995 and 2001, respectively. Now, he is a professor and Ph. D. supervisor in North China Electric Power University. His main research interests include modeling theory and method of large thermal power unit, and the theory and application of intelligent control.

贾昊,2011 年在河北联合大学获得学士学位,2016 年在华北电力大学获得硕士学位,现为华北电力大学博士研究生,主要研究方向为大型火电机组历史数据挖掘与建模。
E-mail:Jiah_paper@163.com

Jia Hao received his B. Sc. degree in 2011 from Hebei United University, and received his M. Sc. degree in 2016 from North China Electric Power University. Now, he is a Ph. D. candidate in North China Electric Power University. His main research interests include historical data mining and modeling of large thermal power unit.