

基于 POT 的多元统计过程核电数据异常检测方法

张立志¹, 罗晓域^{2,✉}, 郑胜³

(1. 中核武汉核电运行技术股份有限公司, 湖北 武汉 443074;

2. 三峡大学电气与新能源学院, 湖北 宜昌 443002;

3. 三峡大学理学院, 湖北 宜昌 443002)

摘要: [目的] 核电设备的安全运行对核电厂至关重要, 发生事故所带来的损失是不可估量的。因此, 对核电设备进行有效的异常检测十分必要。针对固定阈值和人为检测方法的局限性, 这些方法难以适应时序数据的动态变化, 文章提出一种基于 POT 的多元统计过程的异常检测方法。[方法] 文章采用主成分分析方法构建异常检测模型, 将模型的 SPE 统计量作为 POT 算法的初始阈值, 然后将超过初始阈值的部分进行广义帕累托分布拟合, 从而确定最终的动态阈值。当异常分数超过最终阈值则发出异常警告。通过将多元统计过程控制和极值理论相结合, 该方法利用多元统计过程控制快速发现核电厂运行数据中的异常情况, 并结合极值理论对极端事件的建模与分析来提高异常检测的灵敏度和可靠性, 能够快速发现核电厂高维运行数据中存在的异常情况。[结果] 在仿真实验结果中, 文章提出的方法相较于常规的多元统计方法和 POT 方法, 具有更高的准确率、召回率。在核电厂不同设备上的实际运行数据的实验中, 证明了该方法在异常检测上的有效性。[结论] 将多元统计过程控制和极值理论结合, 提出的异常检测方法不仅能检测到由数据相互关系改变引起的异常, 而且能利用 POT 方法确定最终阈值避免传统多元统计过程控制中出现的误检。该方法能处理核电厂高维时序运行数据, 提高异常发现的效率, 确保了核电厂安全高效地运行从而提高核电厂的经济效益。

关键词: 多元统计过程控制; 主成分分析; 动态阈值; 极值理论

中图分类号: TL67; TL48

文献标志码: A

文章编号: 2095-8676(2024)

OA: <https://www.energychina.press/>

An Anomaly Detection Method for Multivariate Statistical Process Based on POT

ZHANG Dazhi¹, LUO Xiaoyu^{2,✉}, ZHENG Sheng³

(1. China Nuclear Power Operation Technology Corporation, Ltd., Wuhan 443074, Hubei, China;

2. College of Electrical Engineering and New Energy, China Three Gorges University, Yichang 443002, Hubei, China;

3. College of Science, China Three Gorges University, Yichang 443002, Hubei, China)

Abstract: [Introduction] The safe operation of nuclear power equipment is crucial for nuclear power plants (NPPs), and the losses caused by accidents are immeasurable. Therefore, effective anomaly detection for nuclear power equipment is necessary. Considering the limitations of fixed thresholds and manual detection methods, which are difficult to adapt to the dynamic changes in time series data, this paper proposes an anomaly detection method based on POT for multivariate statistical processes. [Method] This paper adopted PCA to construct an anomaly detection model, where the SPE statistic of the model served as the initial threshold for the POT algorithm. Subsequently, the portion exceeding the initial threshold was fitted with a generalized Pareto distribution to determine the final dynamic threshold. An anomaly warning was issued when the anomaly score exceeded the final threshold. By combining multivariate statistical process control (MSPC) with extreme value theory (EVT), this method used MSPC to discover anomalies in the operating data of NPPs quickly and improved the sensitivity and reliability of anomaly detection by modeling and analyzing extreme events, so that it can quickly detect anomalies in high-dimensional operating data of NPPs. [Result] In the simulation experiment results, the proposed method has a higher accuracy and recall rate than conventional multivariate statistical and POT methods. In experiments with actual operating data from different equipment in NPPs, the method's effectiveness in anomaly detection has been demonstrated. [Conclusion] By

收稿日期: 2024-04-01 修回日期: 2024-05-08

基金项目: 中核集团核工业仿真技术重点实验室对外开放基金项目“基于半监督/自监督的深度学习异常检测方法研究”(B220631)

combining MPSC with EVT, the anomaly detection method proposed in this paper can not only detect anomalies caused by changes in data relationships but also avoid false detection in traditional MSPC by determining the final threshold using the POT method. This method can handle high-dimensional time series operating data of NPPs, improve the efficiency of anomaly detection, ensure the safe and efficient operation of NPPs, and improve their economic benefits.

Key words: MSPC; PCA; POT; EVT

0 引言

作为一种低碳、高效的清洁能源^[1-2],核能在应对全球气候变化中发挥着积极的作用^[3],2021年的核电厂发电占比达到了4.77%。核电厂所需要的安全性和可靠性使得异常检测技术^[4]在核电领域受到了更广泛的关注。随着监测技术的发展,工业过程中积累了大量高度相关的过程变量,如压力、振动、液位、温度等。通过观察大量的过程变量可发现或预测系统的故障,但对于操作人员而言是巨大的挑战,且对于过程变量间相关性^[5]发生变化而过程变量本身的均值、方差不发生变化的情况,操作人员更难以及时发现故障。随着生产过程的现代化和监测规模的增加,传统的基于模型的方法难以及时发现异常变化,而基于数据驱动的方法,尤其是多元统计过程控制(Multivariate Statistical Process Control, MSPC)方法^[6]能有效地处理庞大且高度相关的数据,而得以有效发展。

基于MSPC的异常检测方法通常遵循两阶段:1)利用历史正常数据训练多元统计模型;2)利用训练的模型对数据实现异常监测。在MSPC中,模型通常由两部分组成:降维模型和监控指标及其控制限^[7]。研究表明,基于主成分分析(Principal Component Analysis, PCA)的MSPC方法非常适合过程故障监测^[8-9],PCA用相互独立的变量来表示测量过程中的相关向量,将高维数据用低维特征表示,并在子空间中定义监控指标及其控制限^[7]。

异常检测^[10-11]依赖于用户定义的先验阈值,使得异常检测算法对阈值的选取非常敏感。设定较高的阈值会使得部分异常漏检,反之会增加误检率。MSPC方法通过设定置信度来确定监控指标的误差限,以避免虚假警报^[12]。然而,最终的检测效果依然受置信度选择影响。阈值的选取常见的有固定阈值和动态阈值^[13]。固定阈值由于无法适应变化^[14]的数据,使用范围有限。因而研究的重心更多在动态阈

值方面,L. Haan^[15]等基于极值理论(Extreme Value Theory, EVT)^[16]提出了(Peak Over Threshold, POT)方法来自动确定阈值,将超过初始阈值的极值进行广义帕累托分布(Generalized Pareto Distribution, GPD)拟合,将极值建模为参数化分布,根据给定的风险值得到最终的阈值。此外,GPD对数据分布不做任何的假设,因而在异常检测中被广泛应用^[17-18]。

POT方法能很好的处理单变量和单峰分布的极值,但是处理多变量或更丰富分布的数据时,特别是对于一个系统中的过程变量间相关性改变而过程变量本身的均值或方差不发生变化的情况,对该算法是一个严峻的挑战。目前通常采用将阈值设置为样本的分位数98%,存在求解困难、计算不稳定,有时无法获得合适的阈值等问题^[19]。

为适应多源数据且各变量相互关联的时变系统的故障监测和预警,文章将MSPC和EVT相结合,提出基于POT方法的多元统计异常检测方法(SPE-POT)。该方法首先基于PCA在残差空间上定义平方预报误差(Squared Prediction Error, SPE)统计量及其误差限。然后,对超过其误差限的SPE值的近似广义帕累托分布,通过极大似然估计对估计该分布的参数。最后,根据给定的风险值确定最终阈值实现异常检测。在核电厂仿真数据和历史数据的实验中,该算法实现了对异常的有效检测并降低了虚假警报。极值理论专注于极端事件的建模和分析,这些事件可能对系统的运行产生重大影响。通过将多元统计过程控制和极值理论相结合,可以综合考虑系统中的常规变化和极端事件,从而提高异常检测的灵敏度和可靠性。该方法能够更全面地捕获核电高维运行数据中的异常情况,为核电厂运行和安全提供更可靠的保障。

综上所述,文章的贡献如下:1)提出了一种基于多元统计和极值理论的异常检测方法,该方法能够适用于高维时序数据;2)对异常检测采用分级报警,超过SPE误差限时发出预警,超过阈值 z_q 的发出报

警; 3) 以一定的风险值来控制报警事故的比率, 允许操作人员根据不同级别的设备根据风险值动态调节阈值。

1 算法原理

1.1 基于 PCA 的 MSPC 异常检测方法

MSPC 通常分为 4 个步骤, 其流程如图 1 所示。

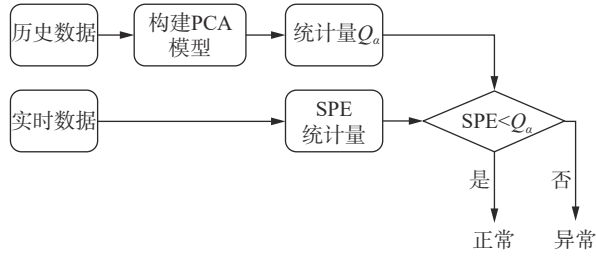


图 1 基于多元统计过程控制的异常检测流程图

Fig. 1 Flowchart of anomaly detection based on MSPC.

第 1 步, 利用历史数据构建 PCA 模型对数据进行主成分分析, 根据特征值的累计贡献率选出前 k 个主成分, 实现数据降维。记过程变量向量的符号为 $\mathbf{X} = [X_1, X_2, \dots, X_n]$, \mathbf{X}_M 中每一行是 \mathbf{X} 的一个样本。为避免过程变量不同量纲对结果的影响和便于数学上的处理, 首先对样本数据进行归一化处理。设 \mathbf{X} 的均值向量为 $\boldsymbol{\mu}_M = [\mu_1, \mu_2, \dots, \mu_n]$, 标准差向量为 $\boldsymbol{\sigma}_M = [\sigma_1, \sigma_2, \dots, \sigma_n]$, 则归一化后的过程变量为:

$$x_i = \frac{X_i - \mu_i}{\sigma_i}, \quad i = 1, 2, \dots, n$$

记归一化后的过程变量为 $\mathbf{x} = [x_1, x_2, \dots, x_n]$ 。同时, 计算出 \mathbf{x} 的均值向量和协方差矩阵分别为:

$$\boldsymbol{\mu} = [\mu_1, \mu_2, \dots, \mu_n]$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} e_{11} & e_{12} & \cdots & e_{1n} \\ e_{21} & e_{22} & \cdots & e_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{nn} \end{bmatrix}$$

其中, 易得到 $\mu_i = 0, e_{ii} = 1, i = 1, 2, \dots, n$ 。对方差矩阵 $\boldsymbol{\Sigma}$ 的 n 个特征值 λ_i 及其对应的单位正交特征向量 \mathbf{p}_i , 其中 $\lambda_{i-1} \geq \lambda_i \geq 0$ 。根据特征值的定义有:

$$\boldsymbol{\Sigma} = \mathbf{p} \boldsymbol{\Lambda} \mathbf{p}^T$$

\mathbf{x} 的 n 个主成分定义为 $t_i = \mathbf{x} \mathbf{p}_i (i = 1, 2, \dots, n)$ 。由于过程变量之间存在高度相关性, 用 \mathbf{x} 的前 k 个主成分就能表示 \mathbf{x} 中的大部分信息 (如 85% 以上)。确定主成分的个数如下所示:

$$\sum_{i=1}^k \lambda_i / \sum_{j=1}^n \lambda_j \geq \eta$$

其中, η 表示设定的累计贡献率。

第 2 步, 计算统计量 SPE。基于 PCA 模型可得到估计向量 $\hat{\mathbf{x}} = \mathbf{x} \mathbf{p}_i \mathbf{p}_i^T$, 其中 $\mathbf{p}_i = [p_{i1}, p_{i2}, \dots, p_{in}]$ 。而 $\mathbf{x}_e = \mathbf{x} - \hat{\mathbf{x}}$ 为 PCA 模型的预测误差向量, 也称为残差向量。SPE 描述了实时数据 \mathbf{X} 与建模数据集 \mathbf{X}_M 的统计差异, SPE 的定义为:

$$\text{SPE} = \mathbf{x}_e \mathbf{x}_e^T \quad (1)$$

第 3 步, 计算统计量控制限。统计量 SPE 的控制限计算公式如下:

$$Q_\alpha = \theta_1 \left[\frac{c_\alpha h_0 \sqrt{2\theta_2}}{\theta_1} + 1 + \frac{\theta_2 (h_0^2 - h_0)}{\theta_1^2} \right]^{\frac{1}{h_0}} \quad (2)$$

其中 $h_0 = 1 - 6\theta_1\theta_3\theta_2^2$ 为常数, $\theta_r = \sum_{j=k+1}^n \lambda_j^r (r = 1, 2, 3)$, k 为主成分分析过程中确定下来的主成分个数。 c_α 为标准正态分布的置信极限, 满足如下公式:

$$P\{N(0, 1) > N_{c_\alpha}(0, 1)\} = c_\alpha$$

第 4 步, 异常判断。如系统正常运行, 则样本的 SPE 值满足 $\text{SPE} < Q_\alpha$, 反之, 可认为出现异常。

1.2 基于极值理论的 POT 异常检测方法

基于极值理论的 POT 方法可以自动为时间序列选择合适的阈值。此外, 它不做任何关于数据分布的假设。通过带参数的广义帕累托分布 GPD 拟合概率分布的尾部。这种极端规律称为极值分布 EVD, 其形式如下:

$$G_\gamma : x \mapsto \exp(-(1 + \gamma x)^{-\frac{1}{\gamma}})$$

其中, $\gamma \in \mathbb{R}, 1 + \gamma x > 0$ 。通过将拟合到未知输入分布尾部, 可以评估潜在极端事件的概率。特别是, 从给定的概率 q 可以计算 z_q , 使得 $P(\mathbf{X} > z_q) < q$ 。要解决这个问题, 自然的方法是估计 γ 。POT 方法的基本思想是: 记 F_t 为超过阈值 t 的 \mathbf{x} 分布, 则条件变量分布函数可表示为:

$$F_t(x) = P(\mathbf{X} - t > x \mid \mathbf{X} > t)$$

其中, \mathbf{X} 是随机变量。对于 $x > t$, 即随机变量超过阈值 t 时, 近似遵循参数 γ, σ 的广义帕累托分布 GPD。当 t 极大时, $F_t(y) \approx G_{\gamma, \sigma}(y)$ 。由于 POT 方法尝试将 GPD 拟合到 $\mathbf{X} - t$, GPD 分布的位置 t 参数可以省去。在得到估计值 $\hat{\gamma}$ 和 $\hat{\sigma}$ 的情况下, 阈值 z_q 可以通过如下公式计算得到:

$$z_q \approx t + \frac{\hat{\sigma}}{\hat{\gamma}} \left(\left(\frac{qn}{N_t} \right)^{-\hat{\gamma}} - 1 \right) \quad (3)$$

其中 n 是总的样本数, N_t 是超过阈值 t 的样本数。一些经典方法可以用来进行 γ 和 σ 的估计, 如矩量法或概率加权矩, 以及效率和鲁棒性更好的极大似然估计。极值理论提供了一种估计阈值 z_q 的方法, 使得 $P(X > z_q) < q$, 而对样本 X 的分布没有任何的假设, 也没有任何关于其分布的先验知。

1.3 基于 POT 的多元统计过程异常检测算法

对于 n 个观测值 X_1, \dots, X_n 和确定的风险值 q , 需要计算阈值 z_q 满足 $P(X > z_q) < q$ 。POT 算法通过是设置一个高阈值 t , 得到超过阈值 t 的样本数据并对其拟合广义帕累托分布, 从而得到极值的分布, 并计算阈值 z_q 。文章提出的 SPE-POT 算法尝试将 MSPC 方法中基于置信度的误差限引入到 POT 算法中的阈值 t , 对超出误差限的统计量拟合 GPD 分布, 确定最终的阈值 z_q , 实现异常检测, 算法的流程图如图 2 所示。

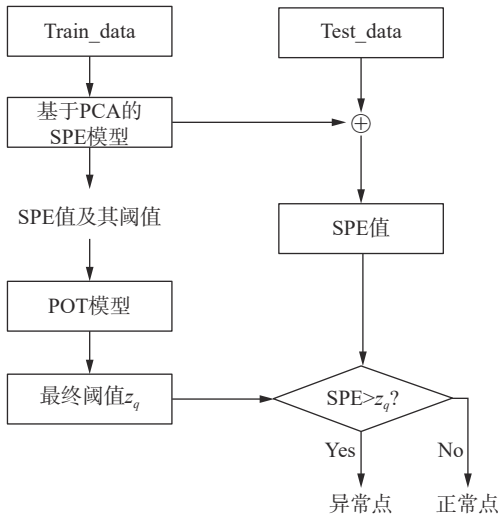


图 2 异常检测算法流程图

Fig. 2 Flowchart of anomaly detection algorithm

算法首先通过历史正常数据构建 PCA 模型, 根据设定的置信度 c_α 和累计贡献率 η 计算出 SPE 值及其误差限, 对超过 Q_α 的 SPE 值拟合其 GPD 分布, 最后根据风险值 q 确定最终阈值 z_q 。在实时数据异常检测阶段, 利用建立的 PCA 模型计算出其 SPE 值, 根据该值与最终阈值 z_q 的大小关系, 给出数据是否异常。文章提出的方法包含三个参数, 详细信息如表 1 所示。

表 1 SPE-POT 方法的参数设置

Tab. 1 Parameter setting for the SPE-POT method

编号	参数	设定值	描述
1	c_α	0.05	标准正态分布的置信极限
2	η	0.95	特征值累计贡献率
3	q	0.000 1	广义帕累托分布的概率

2 实验与分析

2.1 实验设置及评价指标

对方法的检测精度的评价指标^[20] 召回率 R 、准确率 P 和综合指标 F_1 如公式 4-6 所示:

$$R = \frac{TP}{TP + FN} \quad (4)$$

$$P = \frac{TP}{TP + FP} \quad (5)$$

$$F_1 = \frac{2 * P * R}{P + R} \quad (6)$$

式中:

TP —— 正确检测的异常点的个数;

FN —— 将异常点检测为正常点的个数;

FP —— 将正常点检测为异常的点个数。

2.2 仿真数据实验

仿真实验数据 1 由 3 个温度传感器的数据构成, 分别为 Data 1、2、3。如图 3 所示, 其中前 5000 组数据作为训练数据, 后 2000 组数据作为测试数据。测试数据中包含了人为添加的异常片段数据。添加异常数据的方式如下: Data 1 的 [500, 750] 段在原有值增加了 5, [1500, 1700] 段在原有基础上减少 5, Data

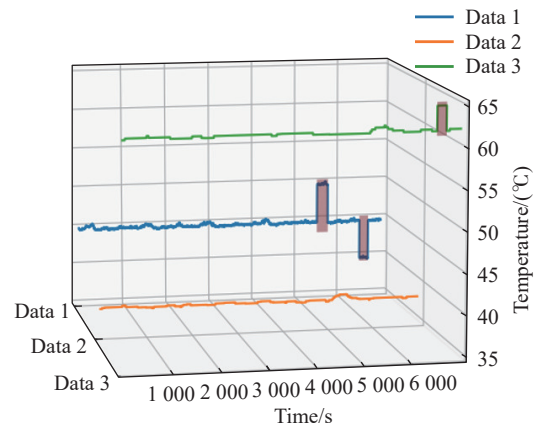


图 3 仿真数据 1 的 3 个传感器数据

Fig. 3 Data from three sensors for simulated data 1

3 的 [1 500, 1 700] 增加 3, 人为添加异常片段的数据的局部放大图如图 4 所示。

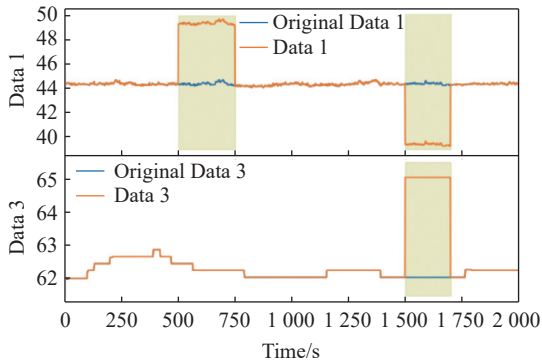


图 4 仿真数据 1 中添加异常的测试数据

Fig. 4 Test data with anomalies added in simulated data 1

仿真实验数据 2 由 3 个温度传感器的数据构成, 分别为 Data 4、5、6。如图 5 所示, 前 5 500 组数据作为训练数据, 后 2 500 组数据作为测试数据。测试数据中, 在三个传感器数据的 [1 500, 1 650] 段叠加平稳变化的“梯形”作为人为添加的异常, 局部放大图如图 6 所示。

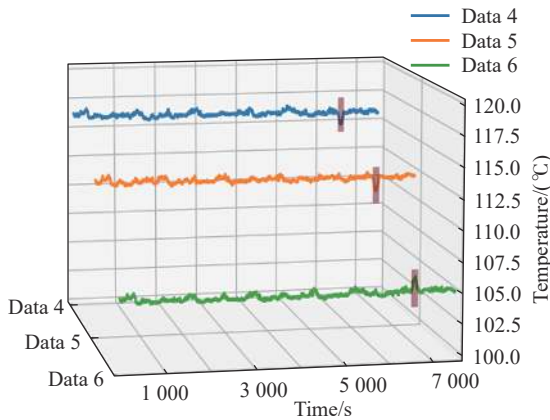


图 5 仿真数据 2 的 3 个传感器数据

Fig. 5 Data from three sensors for simulated data 2

利用训练数据进行建模, 并将模型应用于仿真数据 1 的测试数据, 检测结果如图 7 所示。蓝线表示对测试数据计算得到的统计量 SPE 值取对数的结果, 橙色虚线表示 MSPC 方法的误差限 Q_α , 红色虚线 SPE-POT 方法的阈值 z_q , 红色点为方法检出的异常点。

将上述检测到的异常点在测试数据 1 的原始数据标记, 结果如图 8 所示。其中, 上子图为 SPE-POT 方法的检测结果, 中间子图为 MSPC 方法的检

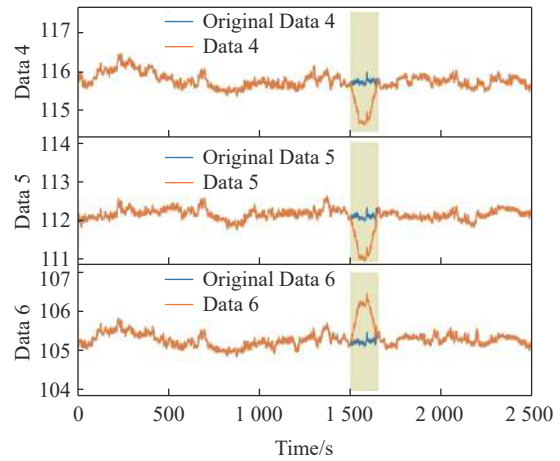


图 6 仿真数据 2 中添加异常的测试数据

Fig. 6 Test data with anomalies added in simulated data 2

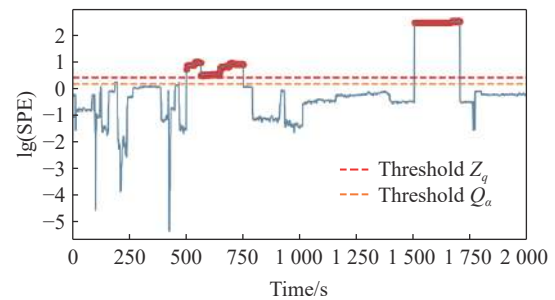


图 7 对仿真数据 1 的数据检测结果

Fig. 7 Data detection results of simulated data 1

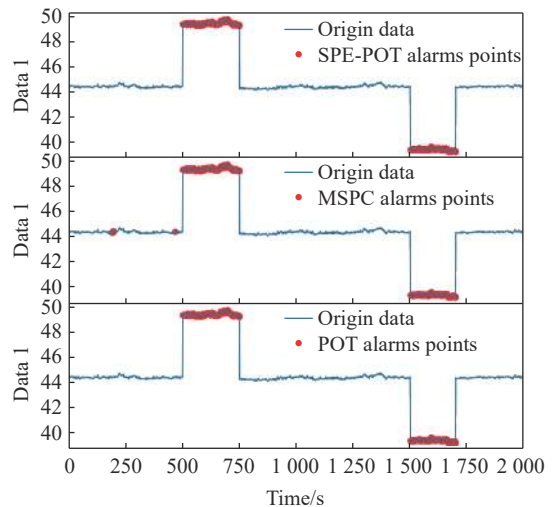


图 8 三种方法检测的结果还原到仿真数据 1, 上子图为 SPE-POT 方法的检测结果, 中间子图为 MSPC 方法的检测结果, 下子图展示的是 POT 方法的检测结果

Fig. 8 Detection results of the three methods are restored to simulated data 1: the upper subplot shows the detection results using the SPE-POT method, and the middle subplot shows the detection results using the MSPC method, and the lower subplot shows the detection results using the POT method

测结果,下子图为 POT 方法的检测结果。从图中可以看出 MSPC 利用误差限检测的结果存在少量的误检, POT 方法在原始数据上检测如图 8 所示的阶跃异常能够完全正确地检测出来。同时, SPE-POT 方法在有效地把人为添加的异常检出的同时没有误检。实验的统计结果如表 2 所示,从表 2 中可以看出, SPE-POT 方法能够准确、完整地检测出人为添加的异常点,而 MSPC 方法把人为添加的异常检出的同时存在一部分误检。POT 方法在原始数据中进行检测,能够将添加的异常很好的检测出来。仿真数据 2 的测试数据上的检测结果如图 9 所示,将上述检测到的异常点在测试数据 2 的原始数据标记, SPE、MSPC 和 POT 方法检测的结果分别如图 10, 图 11 和图 12 所示。

表 2 SPE-POT 和 MSPC 方法检测的统计结果

Tab. 2 Statistical results of detection by SPE-POT and MSPC methods

评价指标	R	P	F_1
SPE-POT	100%	100%	100%
POT	100%	100%	100%
MSPC	100%	97.4%	98.7%

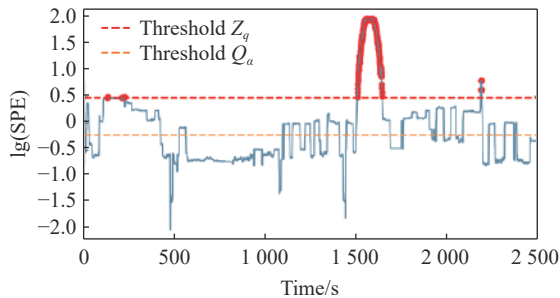


图 9 对仿真数据 2 的测试数据检测结果

Fig. 9 Detection results for the test data of simulated data 2

图 10 中的上、中、下子图展示的分别是 SPE-POT 方法检测的异常点在 Data 4、5、6 的测试数据上标记的结果。从图中可以看出, SPE-POT 方法将人为设置的异常区域有效地检出。图 11 中的上、中、下子图展示的分别是 MSPC 方法检测的异常点在 Data 4、5、6 的测试数据上标记的结果。从图中可以看出, MSPC 方法将人为设置的异常区域检出的同时误检了一些数据点,导致准确率和 F_1 的指标降低。利用 POT 方法分别对 Data 4、5、6 的检测结果如图 13

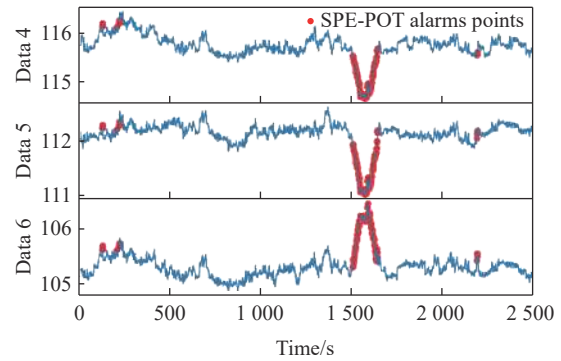


图 10 SPE-POT 方法异常检测结果,上、中、下子图分别为将检测结果在 Data 4、5、6 数据上标记的结果,红点表示检出的异常点

Fig. 10 Anomaly detection results of the SPE-POT method: the upper, middle, and lower subplots respectively display the results marked on data 4, data 5, and data 6. red dots represent the detected anomalous points

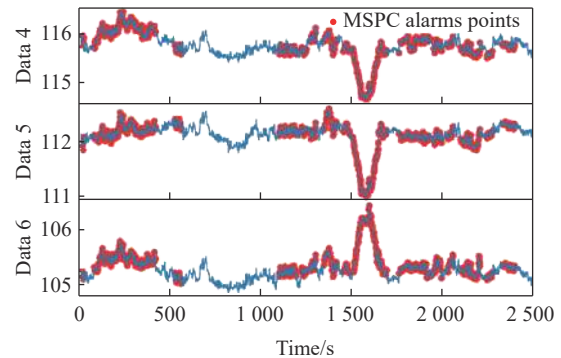


图 11 MSPC 方法异常检测结果,上、中、下子图分别为将检测结果在 Data4、5、6 数据上标记的结果,红点表示检出的异常点

Fig. 11 Anomaly detection results of the MSPC method: the upper, middle, and lower subplots respectively display the results marked on data 4, data 5, and data 6. red dots represent the detected anomalous points

所示,在人工添加的异常片段, POT 方法没有将其检测到,同时其他地方存在一些误检。

从表 3 中可以看出,由于 POT 方法未能将变化缓慢且几个数据间相关性发生变化的异常有效检出,导致该方法召回率、准确率和 F_1 的指标很低,而 MSPC 方法在检测过程容易误检一些数据点导致了准确率和 F_1 的降低。

2.3 真实数据实验

两个真实数据的描述见表 4 和表 5。实验数据 1 为某核电厂化学与容积控制系统 RCV 一回路中主泵及相关设备上 10 个的传感器数据,监测时间为

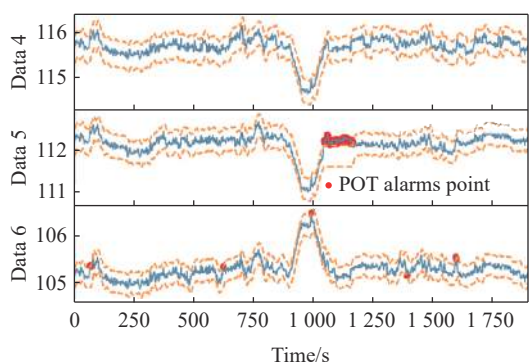


图 12 POT 方法检测结果还原到原始数据上的情况, 上、中、下子图分别为将检测结果在 Data 4、5、6 的测试数据上标记的结果, 红点表示检出的异常点

Fig. 12 Detection results of the POT method restored to the original data: the upper, middle, and lower subplots respectively display the results marked on data 4, data 5, and data 6 test data. red dots represent the detected anomalous points

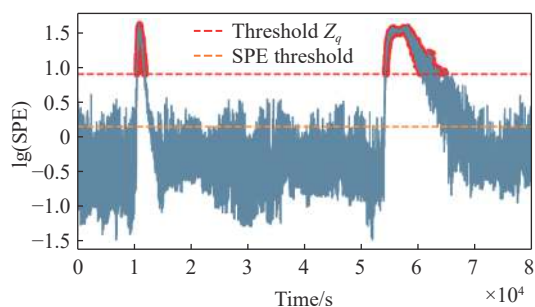


图 13 表 4 中数据的统计量 SPE 值及 SPE-POT 方法确定的阈值

Fig. 13 Statistical SPE values of the data in table 4 and the threshold determined by the SPE-POT method

表 3 SPE-POT 和 POT 方法检测的统计结果

Tab. 3 Statistical results of detection by SPE-POT and POT methods

评价指标	R	P	F ₁
SPE-POT	92.67%	89.68%	91.15%
POT	2.67%	2.86%	2.76%
MSPC	98.67%	11.80%	21.08%

2017年1月10日03:58:33—2017年2月14日18:11:13。实验数据2为汽轮机汽缸和转子的7个传感器数据, 监测时间为2020年3月10日06:44:12—2020年3月22日13:45:09。利用SPE-POT方法检测异常点的结果如图13所示, 蓝色的线表示利用方法对实验数据计算的统计量SPE值, 橙色虚线表示MSPC方法计算的误差限, 红色虚线为

SPE-POT方法确定的阈值。

表 4 真实 Data 1 的传感器信息

Tab. 4 Sensor information from real data 1

序号	传感器	描述
1	RCV200MT	电机非驱动端轴承温度
2	RCV201MT	电机驱动端轴承温度
3	RCV202MT	电机定子绕组温度
4	RCV206MT	增速箱LS轴承温度
5	RCV210MT	泵推力轴承温度
6	RCV223MT	电机定子绕组温度
7	RCV231MV	推力轴承箱振动(X方向)
8	RCV232MV	推力轴承箱振动(Y方向)
9	RCV233MV	径向轴承箱振动(X方向)
10	RCV234MV	径向轴承箱振动(X方向)

表 5 真实 Data 2 的传感器信息

Tab. 5 Sensor information from real data 2

序号	传感器	描述
1	GRE014MM	汽轮机轴向位移
2	GRE015MM	汽轮机轴向位移
3	GRE016MM	汽轮机轴向位移
4	GRE018MM	汽轮机转子偏心率
5	GRE019MY	高中压转子绝对胀差
6	GRE020MV	高中压缸绝对胀差
7	GRE315MV	低压转子绝对胀差

从图13中可以看出, MSPC确定的误差限 Q_a 能有效检出异常点, 同时引入了大量的误检, SPE-POT方法确定的阈值 z_q 则将误检率有效地降低。将异常检测结果还原到原始数据如图14所示, 上、下子图分别展示的是SPE-POT方法和MSPC方法的检测结果。图14中去温度曲线中的两“下凹”的尖峰为操作人员定期实验得到的异常数据, MSPC方法在检测出异常数据时存在着大量误检, 而SPE-POT方法能有效降低误检率。图15显示的是SPE-POT方法对表5列出的7个传感器数据的统计量SPE的检测结果。如图16所示, MSPC方法在检测出异常数据时存在着大量误检, 而SPE-POT方法能准确检出异常且没有误检。

图17显示的是SPE-POT方法对除氧器的21个传感器数据的统计量SPE的检测结果, 红色和橙色的虚线分别表示SPE-POT计算的阈值和MSPC方

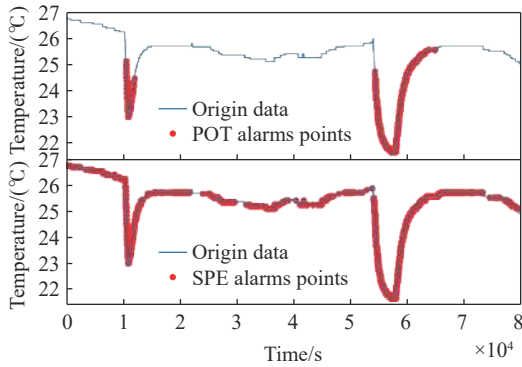


图 14 将检测结果分别还原到传感器 RCV200MT 的数据上, 上、下子图分别展示的是 SPE-POT 和 MSPC 方法的检测结果

Fig. 14 Detection results are restored to the data of sensor RCV200MT respectively: the upper and lower subplots display the detection results of the SPE-POT and MSPC methods, respectively

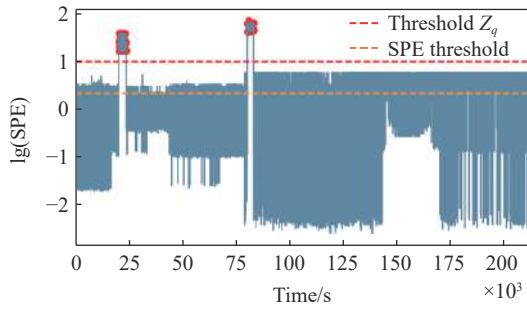


图 15 表 5 中数据的统计量 SPE 值及 SPE-POT 方法确定的阈值

Fig. 15 Statistical SPE values of the data in table 5 and the threshold determined by the SPE-POT method

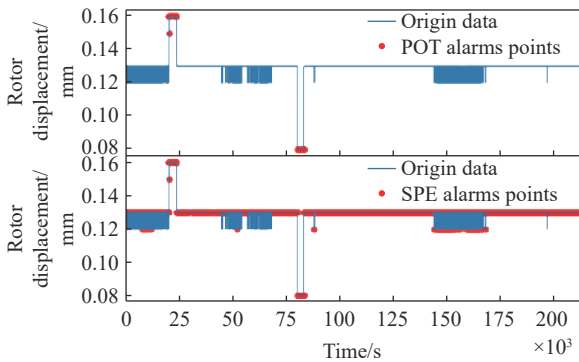


图 16 将两个方法的检测结果分别还原到传感器 GRE014MM 的数据上, 上、下子图分别展示的是 SPE-POT 方法和 MSPC 方法的检测结果

Fig. 16 Detection results of the two methods are respectively restored to the data of sensor GRE014MM: the upper and lower subplots display the detection results of the SPE-POT method and the MSPC method, respectively.

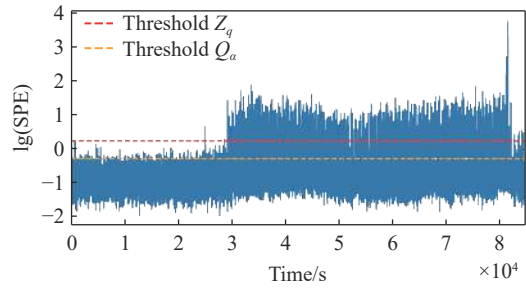


图 17 除氧器设备监测数据的 SPE 值及其阈值和 SPE-POT 方法确定的阈值

Fig. 17 SPE values of the deaerator equipment monitoring data and their thresholds, as well as the thresholds determined by the SPE-POT method.

法计算的阈值。将两种方法检测结果还原到该设备出现异常的传感器数据上, 同时利用 POT 方法对该传感器数据进行异常检测, 结果如图 18 所示。从图 18 可以看出, SPE-POT 方法能够将出现异常的区间准确检测出来, 同时避免了如图 18 中间子图所示的 MSPC 方法存在的误检。从图 18 下子图可以看出, POT 方法仅能检测出极少数的异常点。

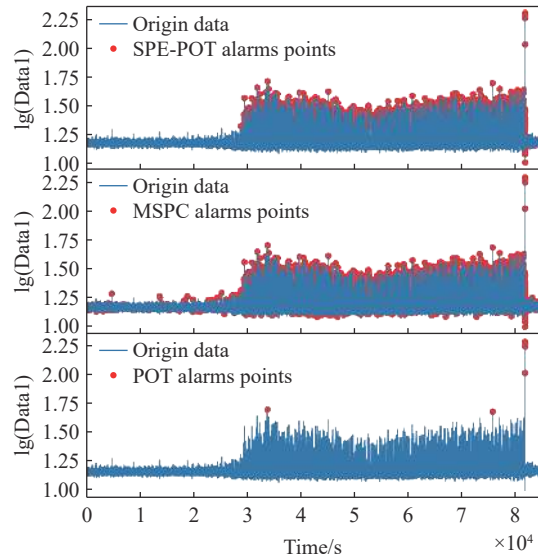


图 18 将 3 个方法的检测结果分别还原到传感器的数据上的情况, 上、中、下子图分别展示的是 SPE-POT 方法、MSPC 方法和 POT 方法的检测结果

Fig. 18 Detection results of the three methods are respectively restored to the situation of the sensor data: the upper, middle, and lower subplots display the detection results of the SPE-POT, MSPC, and POT method, respectively.

为了进一步验证文章提出的方法的稳定性, 将该方法推广应用到某核电厂二回路除氧器设备的运

行数据中。该设备中的运行数据经过核电工作人员确认, 在其中一个传感器中出现了异常, 异常跨越的区间如图 18 所示在 30000 到 81200 的区间内, 对除氧器设备监测的传感器一共有 21 个。此处, 我们利用 3 种异常检测方法进行检测, 仅对出现异常的传感器展示异常检测结果。

3 结论

文章从多元统计过程控制出发, 提出了基于极值理论的 SPE-POT 方法。该方法基于 PCA 模型对历史正常数据建模, 在残差空间中得到统计量 SPE 及其误差限, 并引入基于极值理论的 POT 方法对估计最终阈值进行异常检测, 从而提高了方法的性能。在仿真数据集和真实数据集上的实验均表明, 提出的方法在准确率误检率优于传统的 MSPC 方法。同时方法还具有对参数依赖较小的优点, 具有良好的实用性。

由于提出的 SPE-POT 方法利用 PCA 模型构建历史正常数据模型, 是通过发现数据间的相关关系来实现异常检测的, 故对数据中存在的噪声具有一定的抑制作用。同时, 当设备出现异常后, 反映到运行数据上的表现是传感器之间的耦合关系发生明显差异, 这是明显区别于噪声的情况, 通过多元统计过程控制能迅速捕获这样的异常。然而, 对核电厂中存在非线性过程的工况, 如设备状态切换、升降负荷等工况的适应性较不足。在后续工作中准备将 KPCA^[21] 引入本工作中, 对核电运行数据中的非线性过程建模, 将本方法应用于核电厂瞬态运行工况下, 进一步提升该方法的适应场景。

参考文献:

- [1] 王浩然, 冯天天, 崔茗莉, 等. 碳交易政策下绿氢交易市场与电力市场耦合效应分析 [J]. *南方能源建设*, 2023, 10(3): 32-46. DOI: 10.16516/j.gedi.issn2095-8676.2023.03.004.
WANG H R, FENG T T, CUI M L, et al. Analysis of coupling effect between green hydrogen trading market and electricity market under carbon trading policy [J]. *Southern energy construction*, 2023, 10(3): 32-46. DOI: 10.16516/j.gedi.issn2095-8676.2023.03.004.
- [2] 王鑫, 吴继承, 朴磊. “双碳”目标下核能发展形势思考 [J]. *核科学与工程*, 2022, 42(2): 241-245. DOI: 10.3969/j.issn.0258-0918.2022.02.001.
WANG X, WU J C, PU L. Consideration of the development situation of nuclear power under the goal of carbon peaking and carbon neutrality [J]. *Nuclear science and engineering*, 2022, 42(2): 241-245. DOI: 10.3969/j.issn.0258-0918.2022.02.001.
- [3] 蔡绍宽. 双碳目标的挑战与电力结构调整趋势展望 [J]. *南方能源建设*, 2021, 8(3): 8-17. DOI: 10.16516/j.gedi.issn2095-8676.2021.03.002.
CAI S K. Challenges and prospects for the trends of power structure adjustment under the goal of carbon peak and neutrality [J]. *Southern energy construction*, 2021, 8(3): 8-17. DOI: 10.16516/j.gedi.issn2095-8676.2021.03.002.
- [4] 吴铮, 张悦, 董泽. 基于改进高斯混合模型的热工过程异常值检测 [J]. *系统仿真学报*, 2023, 35(5): 1020-1033. DOI: 10.16182/j.issn1004731x.joss.22-0047.
WU Z, ZHANG Y, DONG Z. Outlier detection during thermal processes based on improved gaussian mixture model [J]. *Journal of system simulation*, 2023, 35(5): 1020-1033. DOI: 10.16182/j.issn1004731x.joss.22-0047.
- [5] 崔文浩, 郑胜, 秦雄杰, 等. 基于多尺度时间窗口的核电运行数据关联性分析方法研究 [J]. *南方能源建设*, 2023, 10(2): 143-150. DOI: 10.16516/j.gedi.issn2095-8676.2023.02.019.
CUI W H, ZHENG S, QIN X J, et al. Research on correlation analysis method for nuclear power operation data based on multi-scale time window [J]. *Southern energy construction*, 2023, 10(2): 143-150. DOI: 10.16516/j.gedi.issn2095-8676.2023.02.019.
- [6] YIN S, DING S X, XIE X C, et al. A review on basic data-driven approaches for industrial process monitoring [J]. *IEEE Transactions on industrial electronics*, 2014, 61(11): 6418-6428. DOI: 10.1109/TIE.2014.2301773.
- [7] JIN X H, FAN J C, CHOW T W S. Fault detection for rolling-element bearings using multivariate statistical process control methods [J]. *IEEE transactions on instrumentation and measurement*, 2019, 68(9): 3128-3136. DOI: 10.1109/TIM.2018.2872610.
- [8] ZHANG Y W, LI S, HU Z Y. Improved multi-scale kernel principal component analysis and its application for fault detection [J]. *Chemical engineering research and design*, 2012, 90(9): 1271-1280. DOI: 10.1016/j.cherd.2011.11.015.
- [9] JIANG Q C, YAN X F, HUANG B. Performance-driven distributed PCA process monitoring based on fault-relevant variable selection and bayesian inference [J]. *IEEE transactions on industrial electronics*, 2016, 63(1): 377-386. DOI: 10.1109/TIE.2015.2466557.
- [10] CHANDOLA V, BANERJEE A, KUMAR V. Anomaly detection: a survey [J]. *ACM computing surveys (CSUR)*, 2009, 41(3): 15. DOI: 10.1145/1541880.1541882.
- [11] WANG H, PENG M J, WESLEY HINES J, et al. A hybrid fault diagnosis methodology with support vector machine and improved particle swarm optimization for nuclear power plants [J]. *ISA transactions*, 2019, 95: 358-371. DOI: 10.1016/j.isatra.

- 2019.05.016.
- [12] NOMIKOS P. Detection and diagnosis of abnormal batch operations based on multi-way principal component analysis world batch forum, Toronto, May 1996 [J]. *ISA transactions*, 1996, 35(3): 259-266. DOI: [10.1016/S0019-0578\(96\)00035-3](https://doi.org/10.1016/S0019-0578(96)00035-3).
- [13] CHEN S Y, JIN G, MA X Y. Satellite on-orbit anomaly detection method based on a dynamic threshold and causality pruning [J]. *IEEE access*, 2021, 9: 86751-86758. DOI: [10.1109/ACCESS.2021.3088439](https://doi.org/10.1109/ACCESS.2021.3088439).
- [14] 卢培, 李小宝, 郑晨旭, 等. 350 MW 余热锅炉变工况运行特性分析 [J]. *南方能源建设*, 2022, 9(3): 41-49. DOI: [10.16516/j.gedi.issn2095-8676.2022.03.005](https://doi.org/10.16516/j.gedi.issn2095-8676.2022.03.005).
LU P, LI X B, ZHENG C X, et al. Analysis on operation characteristics of 350 MW waste heat boiler under variable working conditions [J]. *Southern energy construction*, 2022, 9(3): 41-49. DOI: [10.16516/j.gedi.issn2095-8676.2022.03.005](https://doi.org/10.16516/j.gedi.issn2095-8676.2022.03.005).
- [15] SIFFER A, FOUQUE P A, TERMIER A, et al. Anomaly detection in streams with extreme value theory [C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, Canada, August 13-17, 2017. Halifax: ACM, 2017: 1067-1075. DOI: [10.1145/3097983.3098144](https://doi.org/10.1145/3097983.3098144).
- [16] CHARRAS-GARRIDO M, DEVILLE Y, LEZAUD P. Corrective to the article: extreme value analysis - an introduction *Journal de la SFdS* Vol. 154 No2, 66-97 [J]. *Journal de la société française de statistique*, 2017, 158(3): 27-28.
- [17] YU X L, ZHAO Z B, ZHANG X W, et al. Deep-learning-based open set fault diagnosis by extreme value theory [J]. *IEEE transactions on industrial informatics*, 2022, 18(1): 185-196. DOI: [10.1109/TII.2021.3070324](https://doi.org/10.1109/TII.2021.3070324).
- [18] BEIRLANT J, GOEGEBEUR Y, TEUGELS J, et al. *Statistics of extremes: theory and applications* [M]. Hoboken: Wiley, 2004. DOI: [10.1002/0470012382](https://doi.org/10.1002/0470012382).
- [19] HUNDMAN K, CONSTANTINOU V, LAPORTE C, et al. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding [C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, August 19-23, 2018. London: ACM, 2018: 387-395. DOI: [10.1145/3219819.3219845](https://doi.org/10.1145/3219819.3219845).
- [20] LUO X Y, ZHENG S, HUANG Y, et al. Molecular clump extraction algorithm based on local density clustering [J]. *Research in astronomy and astrophysics*, 2021, 22(1): 015003. DOI: [10.1088/1674-4527/ac321d](https://doi.org/10.1088/1674-4527/ac321d).
- [21] WANG H, PENG M J, YU Y, et al. Fault identification and diagnosis based on KPCA and similarity clustering for nuclear power plants [J]. *Annals of nuclear energy*, 2021, 150: 107786. DOI: [10.1016/j.anucene.2020.107786](https://doi.org/10.1016/j.anucene.2020.107786).

作者简介:



张大志

张大志(第一作者)

1977-, 男, 总工程师, 博士, 主要研究方向核电仿真及数据同化及核电厂运行数据异常检测(e-mail)zhangdz02@cnp.com.cn。



罗晓域

罗晓域(通信作者)

1994-, 男, 博士研究生, 主要从事核电厂数据同化及核电厂运行数据异常检测工作(e-mail)lxu@ctgu.edu.cn。

郑胜

1965-, 男, 博士生导师, 教授, 主要研究方向为图像处理, 核电厂运行数据异常检测(e-mail)zsh@ctgu.edu.cn。

(编辑 徐嘉铨)