

多元时间序列 PCA 分割及在天然气脱水装置 工况识别中的应用

宋伟¹, 熊伟¹, 董莎莎¹, 谭建¹, 彭波¹, 吴娇¹, 梁天佑², 尹爱军²

(1. 中国石油西南油气田分公司重庆气矿, 重庆 400021;

2. 重庆大学 a 机械传动国家重点实验室; b 机械工程学院, 重庆 400044)

摘要: **目的** 建立不同的工作状态下的设备状态预测评估模型。**方法** 利用主成分分析法将多元时间序列数据进行切割, 将分割后的数据段按照基于密度的方法, 依照定义的距离, 对多元时间序列进行聚类合并, 从而得到不同工况下的时间序列。**结果** 按照脱水装置工况状态, 将重沸器特征对应的多元时间序列划分为不同的数据段。**结论** 有效实现了脱水装置重沸器的参数数据分割, 并识别出不同工况。

关键词: 多元时间序列; PCA; 序列分割; 聚类; 工况分离

DOI: 10.7643/issn.1672-9242.2020.04.014

中图分类号: TB115

文献标识码: A

文章编号: 1672-9242(2020)04-0085-05

PCA Segmentation of Multivariate Time Series and Its Application in the Condition Identification of Natural Gas Dehydration Unit

SONG Wei¹, XIONG Wei¹, DONG Sha-sha¹, TAN Jian¹, PENG Bo¹, WU Jiao¹,
LIANG Tian-you², YIN Ai-jun²

(1. Southwest Oil and Gasfield Company Chongqing Gas District, Chongqing 400021, China; 2. a. State Key Laboratory of Mechanical Transmissions, b. School of Mechanical Engineering, Chongqing University, Chongqing 400044, China)

ABSTRACT: The work aims to establish the equipment state prediction and evaluation models in the different conditions. The PCA method was used to cut the multivariate time series data, and the segmented data segments were clustered and merged according to the density-based method and the defined distance to obtain the time series under different conditions. According to the operating conditions of the dehydration unit, the multivariate time series corresponding to reboiler characteristics were divided into different data segments. The parameter data segmentation of the reboiler of the dehydration unit is effectively realized, and different working conditions are identified.

KEY WORDS: multivariate time series; PCA; sequence segmentation; parameter clustering; separation of working conditions

收稿日期: 2019-07-05; 修订日期: 2019-09-05

Received: 2019-07-05; **Revised:** 2019-09-05

基金项目: 重庆市重大主题专项重点研发项目 (cstc2018jszx-cyztzxX0032); 中国石油重庆气矿科研项目 (k18-11)

Fund: Key Science and Technology Research Project of Chongqing under Grant (cstc2018jszx-cyztzxX0032); Petro China Chongqing Gas Mine Scientific Research Project (k18-11)

作者简介: 宋伟 (1968—), 男, 高级工程师, 主要研究方向为地面建设和油气储运。

Biography: SONG Wei (1968—), Male, Senior engineer, Research focus: ground construction, oil and gas storage and transportation.

通讯作者: 尹爱军 (1978—), 男, 博士, 教授, 主要研究方向为智能测试与仪器, 智能运维与健康管理。

Corresponding author: YIN Ai-jun (1978—), Male, Doctor, Professor, Research focus: intelligent test and instrument, intelligent operation and health management.

石化设备的工作过程非常复杂,检测数据特征数众多,且持续时间长。在长期工作中,工况复杂。在状态预测和评估中,若使用同一模型,可能无法取得精确的预测和诊断结果。因此按照工况,将多元时间序列进行分割,建立不同工况条件下的预测评估模型,可实现更为精确的状态检测。多元时间序列分割已在金融、医学、社会与行为科学、化工等领域得到研究和应用^[1],时间序列分割可以用于从给定的时间序列中提取稳定的序列段,以及识别变换点等。

基于工艺监测参数的大数据分析技术在设备状态预测以及故障诊断中逐渐占据重要地位。Aksoy 等人^[2]提出基于动态规划算法的快速分割方法。Janos Abonyi 等人^[3]运用基于 PCA 的数据分割方法实现了对于乙烯生产的化工过程的工况分离。有很多学者对分割后的时间序列进行了聚类等问题的研究,如 Dong 和 Pedrycz 等人^[4]利用模糊 C-均值聚类对均分的时间序列片段进行序列聚类,但是该方法把多元时间序列的形状特征赋予到一元序列上,因此会出现信息丢失问题。Janos Abonyi 等人^[3]使用了层次聚类法进行聚类,但无法确定最终聚类簇数。文中针对天然气脱水装置长期工作、工况复杂、数据量大的特点,在使用基于 PCA 的多元时间序列分割的基础上,利用基于密度的聚类方法以及基于相似性的度量方法,将数据聚类,该方法可以自适应地选择聚类簇数。

1 基于PCA的多元时间序列分割原理

1.1 PCA 基本原理

PCA (Principal Components Analysis) 即主成分分析法。对于一个给定 $n \times n$ 的非奇异矩阵 X , 其特征值模最大的特征向量方向即为该矩阵的主要方向,即特征方向。PCA 通过坐标旋转,将 n 维矩阵投影到 k 维主方向上^[5]。对于对角阵,可以通过矩阵对角化得到特征值与特征向量,如式(1)所示。

$$X = U \Lambda U^T \quad (1)$$

式(1)中: U 为特征列向量按对应特征值模从大到小排列组成的矩阵; Λ 为对角线元素按特征值模从大到小排列的对角阵。特征向量对应特征值,特征值绝对值最大的特征向量成为矩阵的主方向。

1.2 基于PCA的数据分割

令多元时间序列 $T = \{x_k = [x_{1,k}, x_{2,k}, x_{3,k}, \dots, x_{n-1,k}, x_{n,k}]^T | 1 \leq k \leq N\}$, 其中, k 为采样的时间点, n 为特征序号。令 $S(a,b) = \{a \leq k \leq b\}$, 则 $x_{S(a,b)}$ 对应 $x_a, x_{a+1}, \dots, x_b$, 从而设 $S_i^c = \{S_i(a_i, b_i) | 1 \leq i \leq c\}$, 代表若 T 共分为 c 段, 那么第 i 段的位置。其中 $a_i = b_{i-1} + 1$ 。定义损失函数为:

$$\text{cost}(S_i^c) = \sum_{i=1}^c \text{cost}(S_i) \quad (2)$$

式(2)中, 如果设每个节点的坐标为 $S_1 < S_2 < \dots < S_c$, 则 $S_i = S(S_i, S_{i+1} - 1)$ 。

基于 PCA 的数据分割算法如下所述。

输入: 多元时间序列, 所希望的最终分段数 c

输出: 分段节点坐标

创建一个初始的精细分段, 每一段记为 $S(a_i, b_i)$, 设现有段数为 l

计算 $\text{cost}(S(a_i, b_i))$

While $l > c$

搜寻 $\min(\text{cost}(S(a_i, b_i)))$

将 $\arg \min(\text{cost}(S(a_i, b_i)))$ 与 $i+1$ 或 $i-1$ 融合

End

在该算法中, 基于重构误差以及统计量误差的损失函数定义为:

$$\text{cost} = \text{cost}_{T^2} + \text{cost}_Q \quad (3)$$

式(3)中: cost_Q 为矩阵重构误差; cost_{T^2} 为 Hotelling 统计损失^[6], 其为 T 统计量在多元序列下的扩展, 在各领域已有广泛运用^[7-9]。

对于多元时间序列数据, 矩阵重构误差为:

$$\text{cost}_Q(S_i(a_i, b_i)) = \frac{1}{b_i - a_i + 1} \sum_{k=a_i}^{b_i} Q_{i,k} \quad (4)$$

$$Q_{i,k} = \min_{\hat{x}_i} \sum \|x_i - \hat{x}_i\|_2 = \min_{\hat{x}_i} (x_k - \hat{x}_i)^T (x_k - \hat{x}_i) = x_k^T (I - U_{i,p} U_{i,p}^T) x_k \quad (5)$$

式(5)中: I 为单位向量; $U_{i,p}$ 为多元时间序列段的自协方差矩阵经 PCA 降维后的特征向量矩阵。

Hotelling 统计损失为:

$$\text{cost}_{T^2}(S_i(a_i, b_i)) = \frac{1}{b_i - a_i + 1} \sum_{k=a_i}^{b_i} T_{i,k}^2 \quad (6)$$

式(6)中: $T_{i,k}^2$ 为多元序列的总损失, $T_{i,k}^2 = y_{i,k}^T y_{i,k} \circ y_{i,k}$ 为:

$$y_{i,k} = \frac{1}{\sqrt{2}} U_{i,p}^T x_k \quad (7)$$

2 基于相似性的多元序列聚类

2.1 DBSCAN 聚类

DBSCAN^[10] (Density-Based Spatial Clustering of Applications with Noise) 即基于密度的聚类方法, 由于其出色的性能, 在各个领域都有广泛的运用^[11-13]。它有两个特点: 聚类时不需要预估簇的个数; 簇的个数由算法决定。给定邻域内, 由最小数据点数 MinPts, DBSCAN 将数据定义为以下三种类型, 如图1所示。

1) 核心点: 核心的 ε -邻域内, 含有超过给定邻域内最小数据点数 MinPts。

2) 边界点: 点在其 ε -邻域内的数据点数量小于 MinPts, 但位于核心点邻域内。

3) 噪音点: 不符合1)、2)两个定义点。

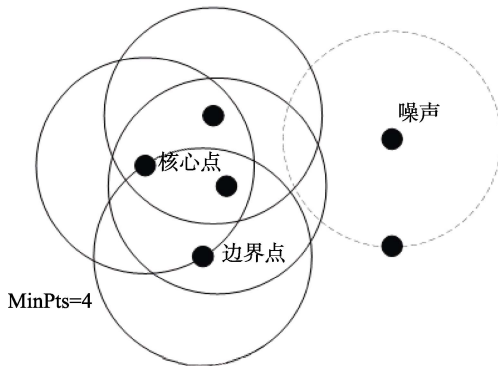


图 1 DBSCAN 聚类展示
Fig.1 Demonstration of DBSCAN

DBSCAN 如果给定参数 $\text{MinPts}=0$ ，那么将不删除噪声点，另外无需指定聚类数，这对于片段进行聚类是相对合适的。DBSCAN 聚类算法中，距离度量方式的选择尤为重要，这里使用了基于相似性的距离度量方式。

2.2 基于相似性的距离度量

基于相似性来衡量两个矩阵的距离，即利用特征值与特征向量，由于特征向量代表着矩阵的主次方向，显然用其来衡量矩阵距离在数学上是合理的，本质上是利用两个矩阵的线性相似性衡量两个矩阵的距离^[14]。Johannesmayer^[15]对 PCA 相似性因子进行了改进，用特征值的平方根，对特征向量进行加权，见式 (7)。

$$S_{\text{PCA}}^{\lambda}(X_n, Y_n) = \frac{\sum_{i=1}^p \sum_{j=1}^p (\lambda_i^{X_n} \lambda_j^{Y_n}) \cos^2 \theta_{i,j}}{\sum_{i=1}^p \lambda_i^{X_n} \lambda_i^{Y_n}} \quad (7)$$

式 (7) 中： $\lambda_i^{X_n}$ 与 $\lambda_j^{Y_n}$ 分别代表了矩阵 X_n 、 Y_n 的特征值； $\theta_{i,j}$ 代表 X_n 的第 i 个主方向和 Y_n 的第 j 个主方向之间的夹角。

由以上得出，用于 DBSCAN 聚类的距离公式为：

$$d(X_n, Y_n) = 1 - S_{\text{PCA}}^{\lambda}(X_n, Y_n) \quad (8)$$

基于相似性度量的 DBSCAN 聚类流程如下所述。

输入：数据集

ϵ ：邻域半径

MinPts：邻域内最小数据点数

距离度量方法

输出：分类

1) 搜索数据集中未被查询过的数据点 p ，若 p 未被归类，则检查其在文中叙述度量方法下定义的距离的 ϵ -邻域。若包含的对象数不小于 MinPts，则建立新簇 C ，将邻域内点归入候选集 N 。

2) 对候选集 N 中未被查询的对象 q ，检查其 ϵ -邻域。若至少包含 MinPts 个对象，则将这些对象加入 N ；如果 q 未归入簇，将 q 加入 C 。

3) 重复步骤 2)，继续检查 N 中未查询的对象，

直到当前候选集 N 为空。

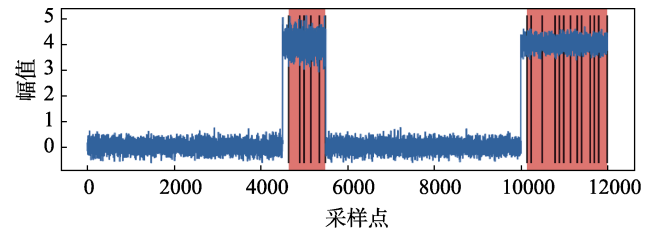
3 模拟数据分析

取模拟数据序列 $\{X_n\}$ ：

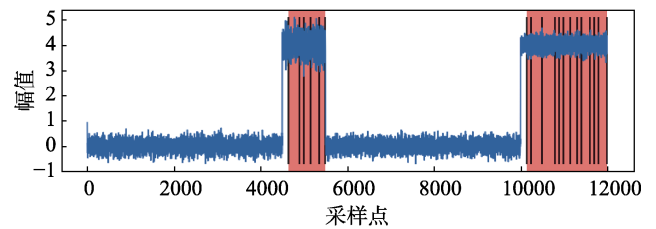
$$X_n = \begin{cases} \varphi_1 & n < 4500 \\ 4 + \varphi_2 & 4500 \leq n \leq 5500 \\ \varphi_3 & 5500 < n < 10\,000 \\ 4\varphi_2 & 10\,000 < n < 12\,000 \end{cases} \quad (9)$$

式 (9) 中： φ_1, φ_3 是符合 $N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$ 分布的

概率值的 0.2 倍； φ_2 是符合 $N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -3 \\ -3 & 1 \end{bmatrix}\right)$ 分布的概率值的 0.2 倍。分割和聚类结果如图 2 所示，其中不同颜色代表聚类结果，深色是一类，浅色是一类。



a 第一分量模拟值



b 第二分量模拟值

图 2 模拟数据分割

Fig.2 Segmentation of simulation data a) first component analog value; b) second component analog value

4 脱水装置多元工艺参数分割与工况识别

重庆气矿七桥中心站 50 万脱水装置现场和其实时监测系统如图 3 所示。该装置有重沸器、泵、吸收塔等 11 类设备，共有压力、温度、循环量等 22 个监测参数^[16]。脱水装置重沸器是重要组成部分，主要检测参数包括重沸器中部温度、精馏柱顶部温度、重沸器尾部温度。

对脱水装置 2015 年 12 月至 2018 年 11 月的重沸器中部温度等 3 个监测参数进行分割，采样周期为 5 min，共 309 353 个数据点。对数据滑动窗口滤波以及标准化后的分割结果如图 4 所示，其中 PCA 初始分割段数为 250 段，以及最终段数为 150 段。利用 DBSCAN 方法聚类的结果如图 5 所示，其中邻域内

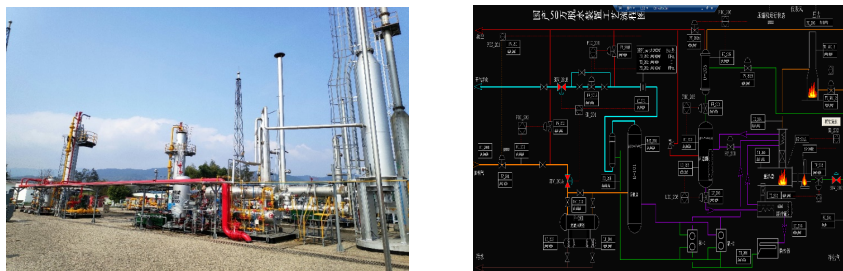


图3 脱水装置现场和实时监测系统

Fig.3 Field and real-time monitoring system of dehydration unit

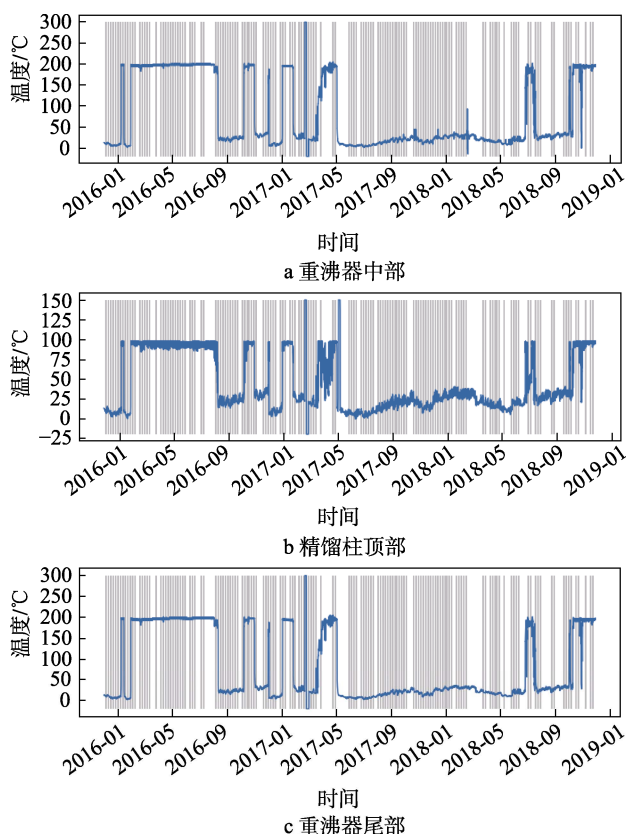


图4 多元时间序列分割

Fig.4 Segmentation of multivariate time series a) middle of reboiler; b) top of distillation column; c) reboiler tail

最小数据点数 MinPts 为 2, 搜索半径为 0.05。

由图 5 可知, 文中方法将温度时间序列分割为明显的二类: 黑色代表处于工作状态, 白色代表处于停机状态。结合生产日志, 重沸器工作过程主要为工作状态和停机状态。2016-01-07 至 2016-01-20 期间, 重沸器位于短暂的工作状态, 然后停止; 2016-02-05 至 2016-08-11, 重沸器处于一个长期的工作状态; 而后在 2016-10-07 至 2016-10-28、2016-12-31 至 2017-01-25、2017-03-23 至 2017-04-23 等时间都处于短暂的工作状态; 在 2015-12-01 至 2016-01-07, 重沸器位于停机状态, 随后工况改变; 在 2017-04-24 至 2018-06-27, 该装置处于长期停机状态, 其温度变化是由环境因素造成的。另外, 在 2017-03-13 至 2017-03-23, 重沸器处于一种过度状态, 推测可能是因为设备异常、设备

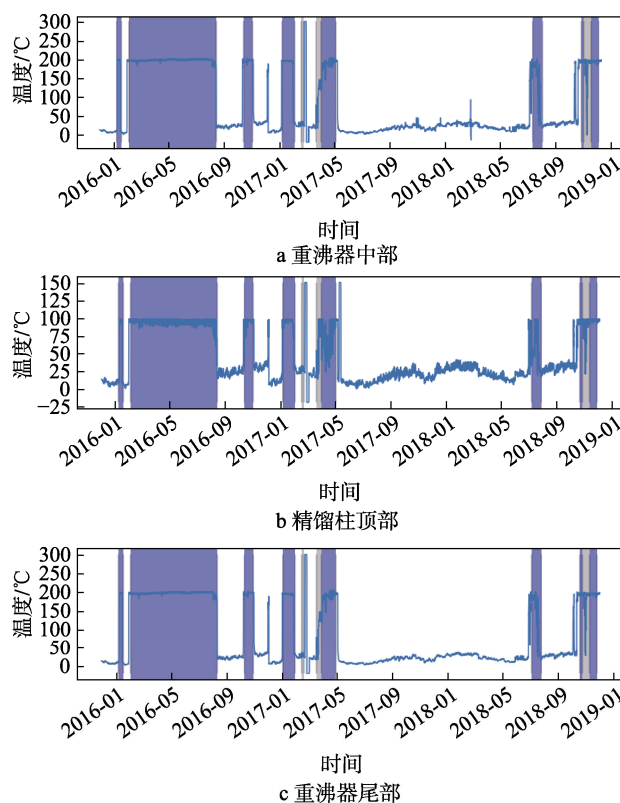


图5 多元序列片段聚类

Fig.5 Cluster of multivariate series fragments a) middle of reboiler; b) top of distillation column; c) reboiler tail

工况参数调试等原因造成的。

5 结语

基于工艺大数据的设备状态评估与检测已得到广泛关注。文中以重庆气矿 50 万脱水装置为研究对象, 结合 PCA 和 DBSCAN 方法, 将多元检测参数按照工况进行分割, 为基于不同的工况状态预测评估模型提供了前处理基础。分割聚类的结果受领域阈值、搜索半径等参数的影响, 后续需进一步研究。根据多元时间序列分割结果, 针对不同工况状态, 建立了不同工况条件下的设备状态预测模型, 消除了负载等工况参数变化对检测数据的影响, 实现更准确的预测和评估, 识别监测异常点。

参考文献:

- [1] 郭红月. 多元时间序列分割与预测方法及应用研究[D]. 大连: 大连理工大学, 2017.
GUO Hong-yue. Multivariate Time Series Segmentation and Prediction Approach and Application Research[D]. Dalian: Dalian University of Technology, 2017.
- [2] AKSOY H, GEDIKLI A, UNAL N E, et al. Fast Segmentation Algorithms for Long Hydrometeorological Time Series[J]. *Hydrological Processes*, 2008, 22(23): 301-304.
- [3] ABONYI J, FEIL B, NEMETH S, et al. Principal Component Analysis Based Time Series Segmentation: A New Sensor Fusion Algorithm[C]// *IEEE International Conference on Computational Cybernetics*. Folyamatmérnöki Intézet Tanszék, 2003.
- [4] DONG R, PEDRYCZ W. A Granular Time Series Approach to Long-term Forecasting and Trend Forecasting[J]. *Physica A: Statistical Mechanics and Its Applications*, 2008, 387(13): 3253-3270.
- [5] STEPHANOPOULOS G, HAN C. Intelligent Systems in Process Engineering: A Review[J]. *Comput Chem Eng*, 1996, 20: 743-791.
- [6] 任雪松, 于秀林. 多元统计分析[M]. 北京: 中国统计出版社, 2011.
REN Xue-song, YU Xiu-lin. Multivariate Statistical Analysis[M]. Beijing: China Statistics Press, 2011.
- [7] PARK S H, KIM S B. Multivariate Control Charts That Combine the Hotelling T-2 and Classification Algorithms[J]. *Journal of the Operational Research Society*, 2019, 70(6): 889-897.
- [8] ZHANG TONG, HUO Yi-xue, ZHANG Xin. Endogenous Third-degree Price Discrimination in Hotelling Model with Elastic Demand[J]. *Journal of Economics*, 2019, 127(2): 125-145.
- [9] SADEGHI Z, MASHHADI-HOSSAINALI M. Application of the T2-hotelling Test For Investigating Ionospheric Anomalies Before Large Earthquakes[J]. *Journal of Atmospheric and Solar-Terrestrial Physics*, 2019, 185: 7-21.
- [10] ESTER M, KRIEGER H P, JÖRG S, et al. A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]// *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Palo Alto: AAAI Press, 1996.
- [11] PARAL P, CHATTERJEE A, RAKSHIT A. Vision Sensor-based Shoe Detection for Human Tracking in a Human-Robot Coexisting Environment: A Photometric Invariant Approach Using DBSCAN Algorithm[J]. *IEEE Sensors Journal*, 2019, 19(12): 4549-4559.
- [12] KHATOON M, BANU W A. An Efficient Method to Detect Communities in Social Networks Using DBSCAN Lgorithm[J]. *Social Network Analysis and Mining*, 2019, 9(1): 554-562.
- [13] ZHANG Hang, LIU Peng, GUO Ying, et al. Blind Modulation Format Identification Using The DBSCAN Algorithm For Continuous-Variable Quantum Key Distribution[J]. *Journal of the Optical Society of America B-Optical Physics*, 2019, 36(3): 51-58.
- [14] ZOLTÁN B, JÁNOS A. Correlation Based Dynamic Time Warping[C]// *8th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics*. Budapest, Hungary, 2007.
- [15] ABONYI J, FEIL B, NÉMETH S, et al. Principal Component Analysisbased Time Series Segmentation—Application to Hierarchical Clustering of Multivariate Process Data[C]// *Proceedings of the IEEE International Conference on Computational Cybernetics*. IEEE, 2003.
- [16] 彭波, 张波, 谭健, 等. 基于工艺参数灰色关联度分析的天然气脱水装置异常检测[J]. *装备环境工程*, 2019, 16(5): 18-23.
PENG Bo, ZHANG Bo, TAN Jian, et al. Anomaly Detection of Natural Gas Dehydration Unit Based on Grey Correlation Analysis of Process Parameters[J]. *Equipment Environmental Engineering*, 2019, 16(5): 18-23.