文章编号:1001 - 2060(2023)04 - 0172 - 09

基于互信息的自动聚类算法在故障诊断过程中的应用

何 康,任少君,司风琪
(东南大学 能源热转换及过程测控教育部重点实验室,江苏 南京 210096)

摘 要:随着热工建模过程中参数的增多,根据参数之间的相关性进行分块建模成为降低模型复杂度、提高模型监测效果的有效手段之一。因此提出了一种基于互信息的自动聚类、分块建模方法。首先,获取参数之间的互信息矩阵,在此基础之上以训练数据的平均平方预测误差最小为标准,使用谱聚类算法对参数进行自动聚类。然后,分别建立每个子块对应的主成分分析(Principle component analysis,PCA)模型,并将所有子块的建模结果通过贝叶斯理论进行融合来对多个子块模型进行统一监测。最后,采用基于最小角度回归(Least angle regressions,LARS)的故障诊断方法定位故障发生的方向和幅值。通过数学案例的验证和电厂高温再热器的实际应用,表明了所提方法在故障监测和诊断方面的有效性。

关 键 词:互信息;谱聚类;PCA;LARS;故障诊断

中图分类号:TK221 **文献标识码**:A DOI:10.16146/j.cnki.rndlgc.2023.04.022

[**引用本文格式**]何 康,任少君,司风琪. 基于互信息的自动聚类算法在故障诊断过程中的应用[J]. 热能动力工程,2023,38(4):172 - 180. HE Kang,REN Shao-jun,SI Feng-qi. Application of automatic clustering algorithm based on mutual information in fault diagnosis [J]. Journal of Engineering for Thermal Energy and Power,2023,38(4):172 - 180.

Application of Automatic Clustering Algorithm based on Mutual Information in Fault Diagnosis

HE Kang, REN Shao-jun, SI Feng-qi
(Key Laboratory of Energy Thermal Conversion and Process Control of Ministry of Education, Southeast University, Nanjing, China, Post Code: 210096)

Abstract: With the increase of parameters in thermal modeling, block modeling based on the correlation between parameters is one of the effective methods to reduce the model complexity and improve the monitoring effect of the model. Therefore, an automatic clustering and block modeling method based on mutual information is proposed. Firstly, the mutual information matrix of parameters is obtained, on this basis, the parameters are automatically clustered by spectral clustering algorithm with the minimum mean square prediction error of the training data as the criterion. Then, the principle component analysis (PCA) modeling corresponding to each sub-block is established, and the modeling results of all sub-blocks are fused by Bayesian theory to monitor the multi-sub-block models in a unified manner. Finally, a fault diagnosis method based on least angle regressions (LARS) is used to identity the direction and amplitude of the faults. The effectiveness of the proposed method in fault monitoring and diagnosis is demonstrated by verification of mathematical case and practical application of high temperature reheater in power plant.

Key words: mutual information, spectral clustering, PCA, LARS, fault diagnosis

收稿日期:2022 - 03 - 07; **修订日期**:2022 - 04 - 28

作者简介:何 康(1996 -),男,东南大学硕士研究生.

引　言

热工过程中的设备运行环境恶劣复杂,设备之间的关联耦合性强,运行参数多。在众多的监测方法中,基于多元统计学的方法如主成分分析(PCA)法、偏最小二乘(Partial Least Square,PLS)法等由于其算法简单、适用范围广,得到了大量的研究和应用[1-2]。PCA 建模方法将数据映射到低维空间来建立全局模型,并通过残差子空间的统计量来进行过程监测[3]。但是随着模型参数的增多,整个模型会越来越复杂,这就会造成模型的监测和诊断效果下降。因此,为了降低模型的复杂性,提高监测诊断的效果,分块建模成为新的研究热点[4]。

分块建模的主要思想是通过定量评估参数间的相关性对过程参数进行分类,从而针对具有强相关的子块参数进行分块建模。Tong 等人[5]通过将主成分空间和残差空间分成 4 个部分进行分块建模来提高模型的精度。Ge 等人[6]沿主成分方向对主元信息进行重构,寻找参数间的相关性,提出分布式主成分分析(Distributed Principle Component Analysis,DPCA)的分块建模方法,有效提升了模型的监测效果。但是当监测出故障之后,如何准确地进行故障隔离、定位故障发生的参数,仍是一个待解决的问题。Xu 等人[7]在划分参数时,综合考虑了变量之间的相关性和冗余性,提出了基于最小冗余最大相关性的分布式 PCA 建模监测方法,该方法可以在一定程度上提高模型的监测效果。但是实际热工过程中,参数之间的关系往往很复杂,仅仅依靠相关性有时难以对参数进行准确分类。

互信息(Mutual Information,MI)是一种成熟的统计分析方法,通过信息熵去度量两个变量之间的依赖。Ji 等人[8]提出了一种基于过程变量间互信息的非平稳过程监控方法,计算正常操作条件下互信息矩阵的特征值欧氏距离,以获得统计量并对过程进行监测。Hassani 等人[9]基于互信息对电力系统中的海量数据进行特征选择,并使用这些特征进行故障监测。现存的研究方法,在使用互信息对参数进行聚类时大部分都需要根据先验知识去确定聚类个数。但是在热工建模过程中,随着参数的增多,这种先验知识往往很难准确获得。而在分块建模中,参数聚类的准确与否会对后续模型监测效果产生重要影响。

综上,为了提高模型的监测和诊断效果,提出了一种基于互信息自动聚类的分块建模故障诊断算法(Mutual Information Multi-least Angle Regressions Principle Component Analysis,MI-MLARSPCA):(1)在参数之间互信息矩阵的基础之上,以训练数据的平均平方预测误差最小作为最佳聚类个数的评判标准,使用谱聚类算法完成对参数变量的自动聚类;(2)当监测到故障发生之后,首先对故障发生的子块进行定位,然后使用基于 LARS 的重构算法对故障参数进行定位并计算出对应的故障幅值。使用数学仿真案例和电厂实际高温再热器的过程数据对所提方法进行验证。

1　基于互信息的自动聚类算法

1.1　互信息理论

在信息论和概率论中,互信息是衡量随机变量之间相互依赖程度的度量,反映两个变量直接的相关性:

$$\boldsymbol{I}(Y,Z) = \sum_{y\in Y}\sum_{z\in Z} p(y,z)\log\left(\frac{p(y,z)}{p(y)p(z)}\right) \quad (1)$$

式中:$I(Y,Z)$—Y 和 Z 的互信息值;$p(y)$—Y 的边缘概率密度函数;$p(z)$—Z 的边缘概率密度函数;$p(y,z)$—Y 和 Z 的联合概率密度函数。

对于连续随机变量,式(1)中的求和被替换成了二重积分:

$$\boldsymbol{I}(Y,Z) = \int_Y\int_Z p(y,z)\log\left(\frac{p(y,z)}{p(y)p(z)}\right)\mathrm{d}y\mathrm{d}z \quad (2)$$

可以看出,不同于相关系数,互信息不局限于实值随机变量,取决于联合分布 $p(y,z)$ 和边缘分布 $p(y)p(z)$ 乘积的相关性。值越大表明 Y 和 Z 之间相关性越高,为 0 则表明 Y 和 Z 相互独立。并且通过互信息的定义可以看出,其具有对称性和非负性,即:

$$\boldsymbol{I}(Y,Z) = \boldsymbol{I}(Z,Y) \geqslant 0 \quad (3)$$

式中:$\boldsymbol{I}(Z,Y)$—Z 和 Y 的互信息值。

对于给定的样本 $\boldsymbol{X} = (x_1,x_2,\cdots,x_n)$,样本 X 的

互信息矩阵 M：

$$\boldsymbol{M}(i,j) = \boldsymbol{I}(x_i, x_j) \tag{4}$$

式中：$\boldsymbol{M}(i,j)$— 样本x_i 和x_j 之间的互信息值。

1.2 谱聚类

谱聚类[10-12]是从图论中演化出的一种聚类算法，相比于 K-Means 等算法，对数据分布的适应性更强。其基本思想是把所有的数据看作空间中的点，这些点之间可以用边连接起来。距离较远点之间权重低，距离较近的点之间权重高。通过对所有数据点组成的图进行切图，让切图后不同子图之间的边权重尽可能低，而子图内的边权重尽可能高，从而达到聚类的目的。对于给定的样本 $\boldsymbol{X} = (x_1, x_2, \cdots, x_n)$ 和聚类数目 k，其计算步骤为：

（1）计算样本的相似度矩阵 S。

（2）根据相似度矩阵 S 构建度矩阵 D：

$$d_i = \sum_{j=1}^{n} s_{ij} \tag{5}$$

式中：d_i— 矩阵 D 对角线的第 i 个元素。

（3）根据式(5)计算拉普拉斯矩阵 L：

$$L = D - S \tag{6}$$

根据式(6)构建标准化的拉普拉斯矩阵L_{std}：

$$L_{\text{std}} = D^{-1/2} L D^{-1/2} \tag{7}$$

之后，计算 L_{std} 最小的 k 个特征值所对应的特征向量。

（4）将特征向量按行进行标准化，最终组成 $n \times k$ 维的特征矩阵 F。

（5）对 F 中的每一行使用 K-Means 聚类方法进行聚类，聚类的数目为 k，并最终得到新的类族 $C_1, C_2, \cdots, C_k$。

因此，如果将样本 $\boldsymbol{X}$ 的互信息矩阵 $\boldsymbol{M}$ 作为相似度矩阵 S 进行谱聚类，那么最终得到的聚类结果就能够使每个子块参数之间的互信息最小，子块内参数之间的互信息最大，从而达到对样本参数进行聚类的目的。

1.3 基于互信息的自动聚类算法

传统的谱聚类方法需要根据先验知识去确定聚类个数，但是热工过程参数多，参数之间关系复杂，往往很难获取到完备的先验知识去确定最佳的聚类个数。对于故障监测模型来说，评价其精度的标准之一就是平方预测误差（Square Prediction Error，SPE），其定义为：

$$\text{SPE} = (x_{\text{pre}} - x)^{\text{T}} (x_{\text{pre}} - x) \tag{8}$$

式中：x_{pre}— 样本 x 的预测值。

SPE 越小，表明模型的预测误差越小，训练模型的精度越高。因此，以训练模型的平均 SPE 最小为依据，自动确定谱聚类的最佳聚类个数，算法流程为：

（1）输入：要进行聚类的训练样本 $X \in R^{m \times n}$，其中 m 为样本个数，n 为参数个数。

（2）输出：聚类的类族 C。计算样本 X 的互信息矩阵 M；初始化聚类个数集合 $K = \{2, \cdots, n\}$，初始 SPE 为空集合。

$\forall k \in K$，以 M 作为相似度矩阵，k 作为聚类个数，使用谱聚类对参数进行聚类并得到对应的聚类结果。对于每个类族中的参数样本，使用 PCA 方法进行建模，得到对应的 SPE 统计指标，根据式(9)计算对应的平均 SPE，并将其加入 SPE 集合之中。

$$\overline{\text{SPE}} = \frac{\sum \text{SPE}}{k} \tag{9}$$

式中：$\overline{\text{SPE}}$— 平均 SPE 值；k— 聚类个数。

（3）计算 SPE 集合中的最小值，将其对应的聚类个数 k 作为谱聚类的最佳聚类个数，并输出对应的聚类结果 C。

2 LARS-PCA 建模诊断方法

2.1 LARS 回归算法

最小角度回归算法 LARS，是为了解决稀疏回归问题[13]而提出来的一种算法：

$$\begin{cases} \min\limits_{\beta} \| y - X\beta \|^2 \\ \sum\limits_{j=1}^{m} | \beta_j | \leqslant t \end{cases} \tag{10}$$

式中：X— 训练样本；y— 训练标签；β— 稀疏回归系数；t—β 向量中不为 0 的元素个数；m— 样本个数。

稀疏回归问题的本质就是进行高维数据的特征选择，在尽量保留数据原始特征的基础上，使得回归系数β 尽可能稀疏，即有尽可能多的项值为 0。这样在提高模型精度的同时也可以大幅度降低计算量。LARS 作为求解稀疏回归问题的经典算法，其

思想与前向选择方法[14]类似,都是逐步进行。但是对于 m 维的数据最多只需要 m 步就可以完成整个算法的迭代过程,原理如图 1 所示。

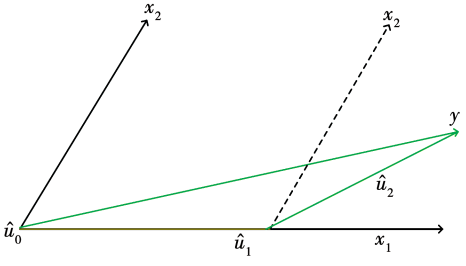

图 1　LARS 算法的原理示意图

Fig. 1 Schematic diagram of LARS algorithm

假设训练数据 $X = (x_1, x_2)$, 当前稀疏回归的预测值为 $\hat{u}$, 根据式定义相关系数 $C(\hat{u})$。

$$C(\hat{u}) = X^{\mathrm{T}}(y - \hat{u}) \tag{11}$$

算法初始时 $\hat{u}_0 = \vec{0}$,从图 1 可以看出,此时 $y - \hat{u}_0$ 更靠近x_1,即 $C_1(\hat{u}_0) > C_2(\hat{u}_0)$,于是 LARS 算法会选择沿着$x_1$ 方向更新 $\hat{u}$:

$$\hat{u}_1 = \hat{u}_0 + \hat{\gamma}_1 x_1 \tag{12}$$

式中:$\hat{\gamma}_1$— 步长;$\hat{u}_1$— 更新后的预测值。

而步长的选择则是使得 $y - \hat{u}_1$ 可以平分x_1 与x_2 的夹角,即 $C_1(\hat{u}_1) = C_2(\hat{u}_2)$。在完成第 1 次选择之后,第 2 次选择会沿着 $y - \hat{u}_1$ 的方向,直到残差 $y - \hat{u}$ 足够小为止,至此完成第 2 次的选择过程。此时对应的步长 $\hat{\gamma}$ 集合即为所求的稀疏回归系数 β。

对于高维数据,假设第 1 次选择之后,算法下一次的选择方向位于特征 X_k 和X_t 的角平分线上, 则 LARS 算法在下一次特征选择时都会探索更多的可能方向,以使得下一个特征 X_p 和预测值 $\hat{u}$ 的相关系数 $C_p(\hat{u})$ 与 $C_k(\hat{u})$ 和 $C_t(\hat{u})$ 相等,然后再更新对应的稀疏回归的预测值 $\hat{u}$。依次循环,直到残差足够小或者所有的变量已经选择完毕,算法终止。LARS 算法的具体求解过程可以参考文献[15]。

2.2　PCA 建模方法

对于给定的样本 $X \in \mathbb{R}^{m \times n}$, PCA 建模的流程:

(1) 求解 X 的协方差矩阵 C_{ov}。

$$C_{\mathrm{ov}} = \frac{1}{m-1} X^{\mathrm{T}}\mathrm{X} \tag{13}$$

式中:m—样本个数。

(2) 对 C_{ov} 进行特征值分解,得到对应的特征值 λ 和特征向量。并根据累计贡献率选择主成分个数 l,再将对应的特征向量组成新的矩阵 P。

$$\xi = \lambda_i / \sum_{i=1}^{n} \lambda_i \tag{14}$$

式中:ξ— 累计贡献率;λ_i— 第 i 个特征值;n —参数个数。

(3) 计算样本 X 的 SPE 控制线。

$$\delta_\alpha^2 = \theta_1 \left(\frac{C_\alpha \sqrt{2\theta_2 h_o^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right)^{h_0^{-1}} \tag{15}$$

式中:$\theta_i = \sum_{j=1+1}^{n} \lambda_j^i$;$h_0 = 1 - \frac{2\theta_1\theta_2}{3\theta_2^2}$;$C_\alpha$— 置信度为 α 的正态分布控制上限。

(4) 对于给定的监测样本 x_{m}, 计算其对应的 SPE 值,如果 SPE $> \delta_\alpha^2$ 则认为此时有故障发生。

2.3　基于 LARS 的故障诊断方法

当故障发生之后,就需要找到故障发生的方向 Ξ_i 以及故障幅值 f_i。故障发生之后,对应的重构监测指标[16]可以写成:

$$\min_{\Xi_i f} \mathrm{SPE}(x) = \min_{\Xi_i f} \| Rx - R\Xi_i f_i \|^2 \tag{16}$$

式中:$\boldsymbol{R} = I - \boldsymbol{P}\boldsymbol{P}^{\mathrm{T}}$,$\boldsymbol{I}$— 单位矩阵。

进一步,可以将其写成:

$$\min_{\Xi_i f} \mathrm{SPE}(x) = \min_{\Xi_i f} \| \tilde{y} - \tilde{x}\beta \|^2 \tag{17}$$

式中:$\tilde{y} = \boldsymbol{R}x$;$\tilde{x} = \boldsymbol{R}\Xi_i$;$\beta = f_i$。

而数学表达式形式和稀疏回归问题的目标函数完全一致。因此,可以将故障诊断问题转化为稀疏回归系数的求解问题,即可以用 LARS 算法来快速求解稀疏回归系数 β,即故障幅值 f。需要说明的是,在使用 LARS 算法求解式时,根据参考文献[16],如果找到了正确的故障参数和对应的故障幅值,就可以使得监测指标 SPE 降到控制线 δ_α^2 以下。即在每个迭代计算完成之后,首先计算对应的 SPE 值,如果满足 SPE $< \delta_\alpha^2$,则终止计算,将得到的稀疏回归系数 β 作为故障幅值 f 的输出。

3　基于互信息的分块模型故障监测与诊断

随着建模参数的增多,对应的聚类子块数目和

子块模型也会增多,无法得到一个直观的最终决策。因此,采用贝叶斯融合策略[17],将所有子块的统计量组合成一个新的 B_{SPE}(Bayesian Square Prediction Error)统计量来进行统一监测,下标 SPE 表示平方预测误差。

对于监测样本 x_m,首先计算x_{m} 属于子块 c 的部分x_c 故障数据的后验概率。

$$P_{\mathrm{SPE}}(F \mid x_{\mathrm{c}}) = \frac{P_{\mathrm{SPE}}(x_{\mathrm{c}} \mid F)P_{\mathrm{SPE}}(F)}{P_{\mathrm{SPE}}(x_{\mathrm{c}})} \tag{18}$$

式中:F— 故障;$P_{\mathrm{SPE}}(F \mid x_{\mathrm{c}})$—$x_{\mathrm{c}}$ 故障的概率。

对于给定的置信度 α,分别计算故障和正常的条件概率。

$$\begin{cases} P_{\mathrm{SPE}}(F) = 1 - \alpha, \\ P_{\mathrm{SPE}}(N) = \alpha \end{cases} \tag{19}$$

$$\begin{cases} P_{\mathrm{SPE}}(x_c \mid F) = e^{-\frac{\mathrm{SPE}_{c,x_c}}{\mathrm{SPE}_c}}, \\ P_{\mathrm{SPE}}(x_c \mid N) = e^{-\frac{\mathrm{SPE}_c}{\mathrm{SPE}_{c,x_c}}} \end{cases} \tag{20}$$

式中:SPE_c—x_c 所属的子块 c 的 SPE 限值;SPE_{c,x_c}—监测数据为x_c 时计算得到的限值;N—正常。

根据全概率公式,可以得到 $P_{\mathrm{SPE}}(x_c)$的计算式:

$$P_{\mathrm{SPE}}(x_c) = P_{\mathrm{SPE}}(x_c \mid F)P_{\mathrm{SPE}}(F) + P_{\mathrm{SPE}}(x_c \mid N)P_{\mathrm{SPE}}(N) \tag{21}$$

对于 x_m 最终的贝叶斯融合 SPE 指标 B_{SPE} 值为:

$$B_{\mathrm{SPE}} = \sum_{i=1}^{k} B_{\mathrm{SPE},i} \tag{22}$$

式中:$B_{\mathrm{SPE},i}$— 第 i 个分组的贝叶斯 SPE 值。计算式为:

$$B_{\mathrm{SPE},i} = \frac{P(x_{c_i} \mid F)P(F \mid x_{c_i})}{\sum_{j=1}^{k} P(x_{c_i} \mid F)} \tag{23}$$

如果 $B_{\mathrm{SPE}} < 1 - \alpha$,则认为过程是正常的,否则认为过程是异常的。当监测到故障发生之后,先计算各子块的 $B_{\mathrm{SPE},i}$ 值,其最大值对应的子块为故障发生的子块。之后对这个子块的 PCA 模型使用 LARS 故障诊断方法进行故障隔离,找出故障发生的方向和幅值。

基于互信息的分块 MI-MLARSPCA 建模诊断算法的流程如图 2 所示。具有步骤为:

(1) 对训练数据 X 进行归一化,并计算 X 的互信息矩阵 M。

(2) 以 X 的平均 SPE 最小为依据,使用谱聚类获取 X 的聚类结果。对每个子块使用 PCA 算法进行建模,并计算对应的阈值。

(3) 对于监测数据x_m,计算其对应的 B_{SPE}值,并判断是否有故障发生。若有故障发生,使用 LARS 算法进行故障诊断来确定故障参数和对应的故障幅值。

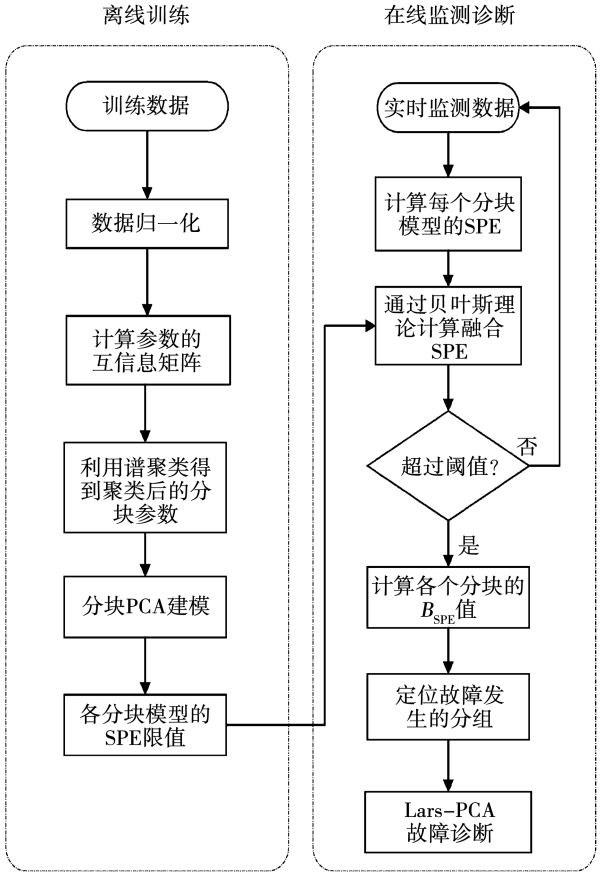

图 2 MI-MLARSPCA 算法流程图

Fig. 2 Flow chart of MI-MLARSPCA algorithm

4 数学仿真和实际案例应用

4.1 数学仿真

为了验证所提出算法的有效性,建立数学模型。

$$\begin{cases} X_1 = \boldsymbol{a}[s_1, s_2, s_3] + e_1 \\ X_2 = \boldsymbol{a}[s_4, s_5, s_6] + e_2 \\ X_3 = \boldsymbol{a}[s_7, s_8, s_9] + e_3 \end{cases} \tag{24}$$

式中:$s_1 \sim N(-5, 0.2)$,$s_2 \sim N(-5.5, 0.2)$,$s_3 \sim$

$N(-4.5,0.2)$,$s_4 \sim N(0,0.2)$,$s_5 \sim N(-0.5,0.2)$,$s_6 \sim N(0.5,0.2)$,$s_7 \sim N(5,0.2)$,$s_8 \sim N(5.5,0.2)$,$s_9 \sim N(4.5,0.2)$。

式中:s_i—服从高斯分布的随机数据;e_i—噪声;$\boldsymbol{a}$—系数矩阵。噪声$e_1,e_2,e_3 \sim N(0,0.01)$。

系数矩阵 $\boldsymbol{a}$ 为:

$$\boldsymbol{a} = \begin{bmatrix} 0.492\,899 & 0.328\,405 & 0.442\,6 \\ 0.407\,149 & 0.249\,226 & 0.285\,915 \\ 0.217\,613 & 0.417\,633 & 0.236\,637 \end{bmatrix} \tag{25}$$

生成 1 000 组样本并标准化,作为训练数据。图 3 分别展示了第 1 个、第 4 个和第 7 个参数与其他参数之间的互信息值。

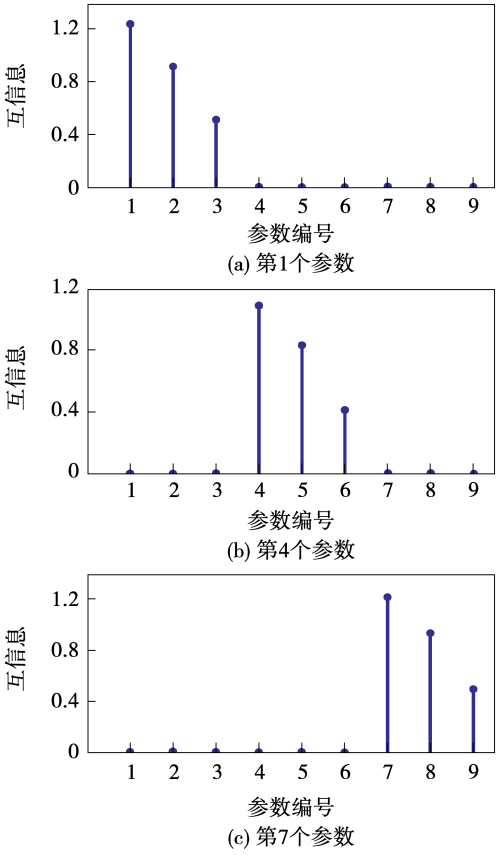

图 3　部分参数之间互信息值

Fig. 3 Mutual information values of partial parameters

以样本的互信息矩阵作为相似度矩阵,采用自动聚类算法对参数进行聚类,聚类的结果如图 4 所示。可以看出,本文的算法将原始数据分成了 3 类,并且将结构相似、相关性强的参数放在同一子块中,这与原始数据的分布特征也是完全吻合的。

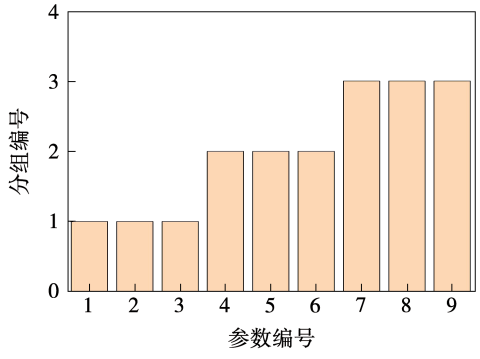

图 4　参数聚类结果

Fig. 4 Results of parameters clustering

针对训练数据,生成两类模拟故障:(1) 故障 1。对第 1 个参数,从第 250 个样本开始设置故障幅值大小为 0.018 的阶跃故障;(2) 故障 2。对第 2 个参数,从第 250 个样本开始设置故障幅值大小为 $0.001(i-250)$ 的线性故障。

为了进一步说明 MI-MLARSPCA 算法的有效性,分别采用本文得到的最佳聚类个数 3(MI-MLARSPCA_3)和随机选取的聚类个数 2(MI-MLARSPCA_2)以及 PCA 算法对数据进行建模分析。首先进行故障的监测,在监测到故障的发生之后,采用 LARS 算法进行故障隔离。其中,诊出率(Fault detection rate, FDR)和误诊率(Fault alarm rate, FAR)的计算公式为:

$$\mathrm{FDR} = \left(\frac{1}{m}\sum_{i=1}^{m} o_i/r_i\right) \times 100\% \tag{26}$$

$$\mathrm{FAR} = \left(\frac{1}{m}\sum_{i=1}^{m} p_i/(n-r_i)\right) \times 100\% \tag{27}$$

式中:m— 样本个数;n— 参数个数;r_i— 第 i 个测试样本涉及的故障参数个数;o_i— 第 i 个测试样本实际被诊出的故障参数个数;p_i— 第 i 个测试样本被误诊的参数个数。

图 5 和图 6 分别展示了 3 种算法对两种故障的监测结果。图 7 展示了在第 500 个样本处不同类族的 B_{SPE} 值。表 1 展示了 3 种算法对两种故障的监测准确率、诊出率和误诊率。

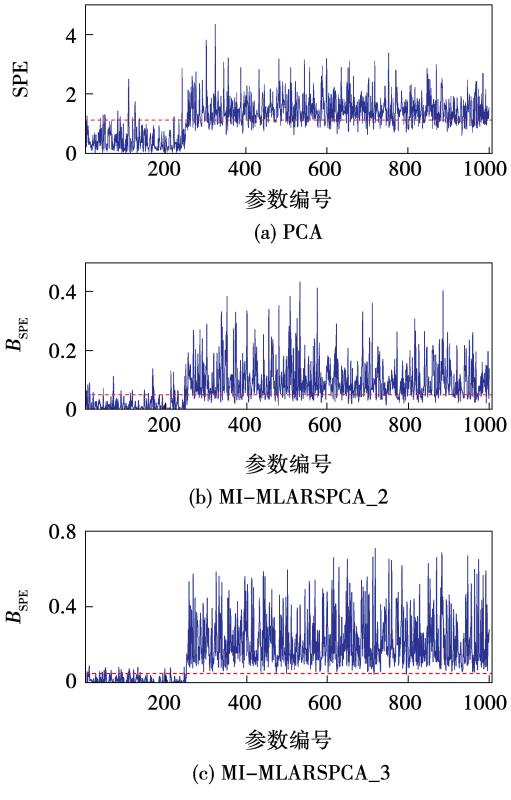

图 5 故障 1 监测结果

Fig. 5 Monitoring result of fault 1

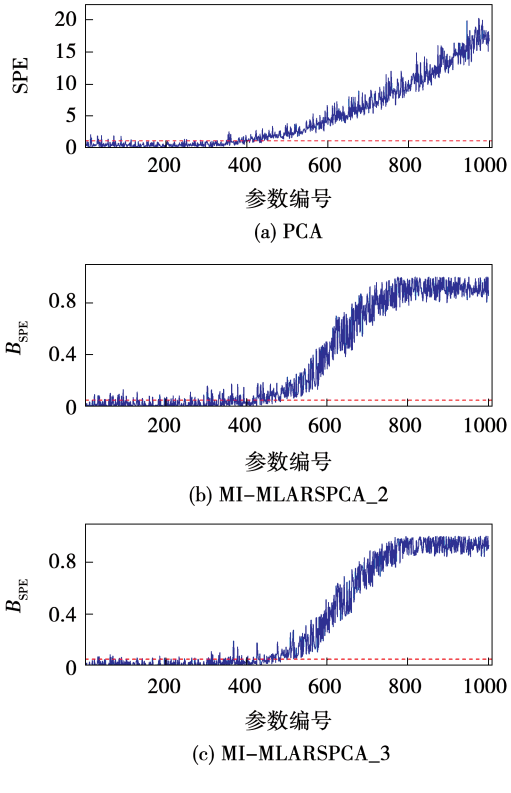

图 6 故障 2 监测结果

Fig. 6 Monitoring result of fault 2

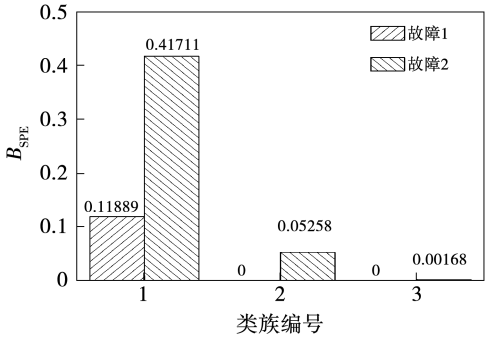

图 7 第 500 个样本处 3 个类族的 B_{SPE} 值

Fig. 7 B_{SPE} values of three blocks at 500th fault sample

表 1 3 种算法的建模结果(%)

Tab. 1 Results of modeling for three algorithms(%)

方法	故障 1		故障 2	
	FDR	FAR	FDR	FAR
PCA	67.2	3.61	79.6	3.38
MI-MLARSPCA_3	93.8	3.7	85.7	3.17
MI-MLARSPCA_2	73	3.6	82	3.21

由图 5 的监测结果和表 1 的诊断结果可以看出,对于故障 1,此时故障参数的故障幅值很小,PCA 算法已经难以捕捉到局部参数的异常,因此模型的监测效果较差,诊出率也低;相比较之下 MI-MLARSPCA 算法由于采用了分块建模的策略,能够更好地捕捉到局部参数的变化,监测和诊断效果好于 PCA 算法。同时,通过图 7 可以看出,本文的聚类算法将局部特征相似的参数自动划分到一个子块中,可以大幅度提高模型的监测精度以及故障隔离时的诊断效果。对于故障 2,由图 6 和表 1 可以看出,MI-MLARSPCA 算法在故障的监测和诊断方面也具有明显的优势。

4.2 实际案例应用

高温再热器作为锅炉汽水循环的核心部件长期处于高温条件下工作,很容易发生超温爆管的恶性事故[18]。选取某 600 MW 电厂的高温再热器作为研究对象,进行超温的故障监测和诊断。该电厂的高温再热器为屏式对流换热器,共有 26 屏,每个屏有进、出口 2 个温度测点。采集正常工况下的再热器的前 20 个屏的出口温度过程变量的 1 000 个样本作为训练样本。对于测试样本,前 250 个样本为正常运行工况下的样本,后 750 个样本为再热器第 2 屏末管壁温度在超温故障下运行的样本。图 8 为

部分参数和其他参数之间的互信息。

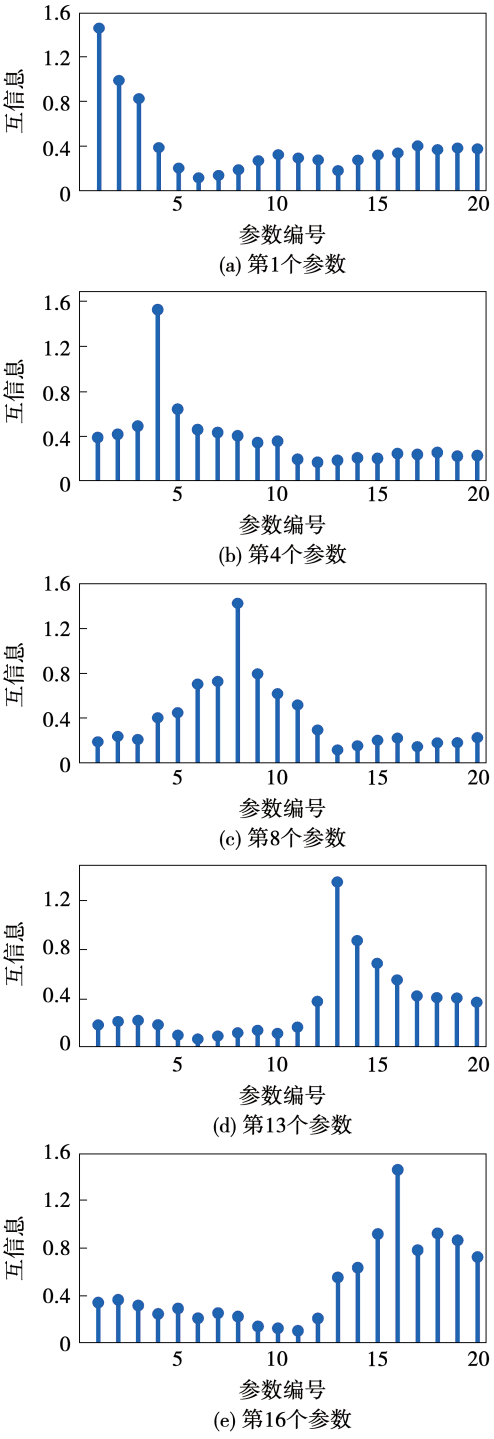

图 8　部分高温再热器参数间的互信息

Fig. 8 Mutual information of partial parameters of high temperature reheaters

首先利用自动聚类方法对 20 个高温再热器出口壁温参数进行聚类,结果如图 9 所示。可以看出,相邻的高温再热器出口壁温测点被分在同一个子块中,这与整个高温再热器出口壁温的特性相符合。

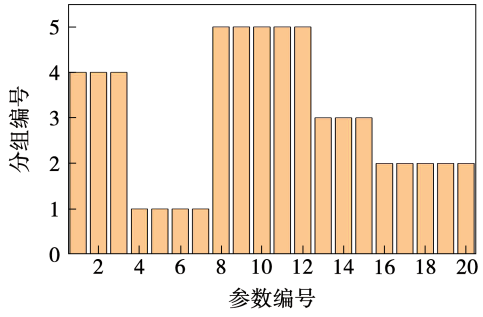

图 9　高温再热器出口壁温参数聚类结果

Fig. 9 Clustering result of outlet wall temperature parameters of high temperature reheaters

图 10 为高温再热器出口壁温超温的监测结果。可以看出,在第 250 个样本之后,MI-MLARSPCA 算法可以达到持续报警。图 11 的故障类族识别结果表明,在对测点参数进行合理的聚类之后,算法准确地识别到了超温的子块。结合 LARS 故障诊断算法,由表 2 可以看出,对故障测点的定位准确率可以达到 100%,很好地做到了故障参数的实时监测与智能预警。

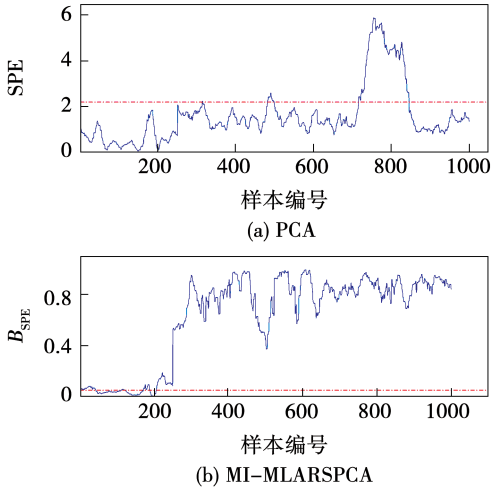

图 10　高温再热器故障监测结果

Fig. 10 Monitoring result of fault of high temperature reheaters

表 2　PCA 及 MI-MLARSPCA 高温再热器建模结果

Tab. 2 Results of modeling for high temperature reheaters by PCA and MI-MLARSPCA algorithms

方法	诊出率/%	误差诊/%
PCA	15.45	0.23
MI-MLRASPCA	100	0.95

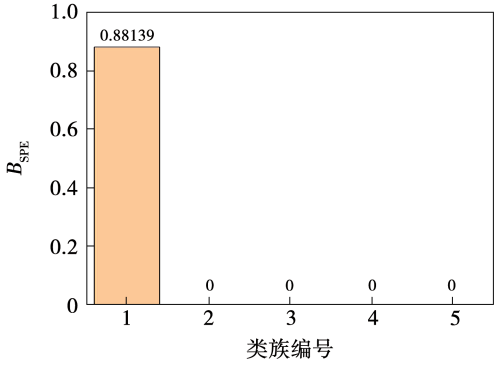

图 11 故障类族识别结果

Fig. 11 Results of fault cluster identification

5 结 论

对热工过程的分块建模方法进行了研究。首先获取参数之间的互信息矩阵,在此基础之上,以训练样本平均 SPE 最小为标准,使用谱聚类对参数进行自动聚类。对聚类得到的子块分别建立 PCA 模型,并使用贝叶斯理论对子块模型进行统一监测。当监测到故障发生时,首先定位到故障发生的子块,之后使用 LARS 算法对故障参数进行定位。数学仿真和现场实际案例应用都表明,该算法相比于传统的 PCA 算法在故障监测和诊断方面具有明显的优势,可以对现场运行的设备进行更准确地故障监测和预警。

参考文献:

[1] BOTRE C, MANSOURI M, NOUNOU M, et al. Kernel PLS-based GLRT method for fault detection of chemical processes [J]. Journal of Loss Prevention in the Process Industries, 2016, 43: 212 – 224.

[2] LI Wei, PENG Min-jun, WANG Qing-zhong. False alarm reducing in PCA method for sensor fault detection in a nuclear power plant [J]. Annals of Nuclear Energy, 2018, 118: 131 – 139.

[3] WESTERHUIS J A, GURDEN S P, SMILDE A K. Generalized contribution plots in multivariate statistical process monitoring [J]. Chemometrics and Intelligent Laboratory Systems, 2000, 51(1): 95 – 114.

[4] GE Zhi-qiang, CHEN Jung-hui. Plant-wide industrial process monitoring: a distributed modeling framework [J]. IEEE Transactions on Industrial Informatics, 2016, 12(1): 310 – 321.

[5] TONG Chu-dong, SONG Yu, YAN Xue-feng. Distributed statistical process monitoring based on four-subspace construction and Bayesian inference [J]. Industrial & Engineering Chemistry Research, 2013, 52(29): 9897 – 9907.

[6] GE Zhi-qiang, SONG Zhi-huan. Distributed PCA model for plant-wide process monitoring [J]. Industrial & Engineering Chemistry Research, 2013, 52(5): 1947 – 1957.

[7] XU Chen, ZHAO Shun-yi, LIU Fei. Distributed plant-wide process monitoring based on PCA with minimal redundancy maximal relevance [J]. Chemometrics and Intelligent Laboratory Systems, 2017, 169: 53 – 63.

[8] JI Cheng, WANG Jing-de, SUN Wei. A nonstationary process monitoring based on mutual information among process variables [J]. IFAC-PapersOnLine, 2021, 54(3): 451 – 456.

[9] HASSANI H, HALLAJI E, RAZAVI-FAR R, et al. Unsupervised concrete feature selection based on mutual information for diagnosing faults and cyber-attacks in power systems [J]. Engineering Applications of Artificial Intelligence, 2021, 100: 104150.

[10] JIA Hong-jie, WANG Liang-jun, SONG He-ping, et al. An efficient Nyström spectral clustering algorithm using incomplete Cholesky decomposition [J]. Expert Systems with Applications, 2021, 186: 115813.

[11] WANG Zhen-lei, ZHAO Su-yun, LI Zheng, et al. Ensemble selection with joint spectral clustering and structural sparsity [J]. Pattern Recognition, 2021, 119: 108061.

[12] ZHONG Guo, SHU Ting, HUANG Guo-heng, et al. Multi-view spectral clustering by simultaneous consensus graph learning and discretization [J]. Knowledge-Based Systems, 2022, 235: 107632.

[13] CARAIANI P. Using LASSO-family models to estimate the impact of monetary policy on corporate investments [J]. Economics Letters, 2022, 210: 110182.

[14] ZHANG Long, LI Kang. Forward and backward least angle regression for nonlinear system identification [J]. Automatica, 2015, 53: 94 – 102.

[15] ITURBIDE E, CERDA J, GRAFF M. A comparison between LARS and LASSO for initialising the time-series forecasting auto-regressive equations [J]. Procedia Technology, 2013, 7: 282 – 288.

[16] ALCALA C F, QIN S J. Reconstruction-based contribution for process monitoring [J]. Automatica, 2009, 45(7): 1593 – 1600.

[17] GE Zhi-qiang, GAO Fu-rong, SONG Zhi-huan. Two-dimensional Bayesian monitoring method for nonlinear multimode processes [J]. Chemical Engineering Science, 2011, 66(21): 5173 – 5183.

[18] 范大志. 600 MW 锅炉高温受热面爆管与寿命监测研究[D]. 保定:华北电力大学,2012.
FAN Da-zhi. Study of tube explosion and life monitoring of 600 MW power plant boiler [D]. Baoding: North China Electric Power University, 2012.

(丛 敏 编辑)