

一种基于 LPP 子空间的热工过程多模式聚类方法研究

袁照威¹, 司凤琪², 孟 磊¹, 谷小兵¹

(1. 大唐环境产业集团股份有限公司, 北京 100097;

2. 东南大学 能源热转换及其过程测控教育部重点实验室, 江苏 南京 210096)

摘 要:针对热工多模态过程的模式识别和聚类问题,提出了一种基于局部保留投影(Local Preserving Projection, LPP)子空间的混合聚类方案。首先,将高维的多模式过程数据通过局部保留投影方法投射到低维子空间中,在剔除噪声、提高计算效率的同时保留局部结构;其次,在 LPP 子空间中,结合传统的分层和非分层聚类算法的优点,使用凝聚 k -means 算法,为多模式过程数据生成最佳的集成聚类解决方案。以某 600 MW 机组脱硫系统的多模式过程数据的识别与聚类过程为例,证明了该方法的有效性和实用性。

关 键 词:多模式过程;局部保留投影; k -means;凝聚分层聚类;湿法脱硫

中图分类号:TM621

文献标识码:A

DOI:10.16146/j.cnki.rndlgc.2022.02.014

[引用本文格式]袁照威,司凤琪,孟 磊,等.一种基于 LPP 子空间的热工过程多模式聚类方法研究[J].热能动力工程,2022,37(2):100-106. YUAN Zhao-wei, SI Feng-qi, MENG Lei, et al. A multi-mode clustering method for thermal process based on locality preserving projection subspace[J]. Journal of Engineering for Thermal Energy and Power, 2022, 37(2):100-106.

A Multi-mode Clustering Method for Thermal Process based on Locality Preserving Projection Subspace

YUAN Zhao-wei¹, SI Feng-qi², MENG Lei¹, GU Xiao-bing¹

(1. Datang Environment Industry Group Co., Ltd., Beijing, China, Post Code:100097; 2. Key Laboratory of Energy Thermal Conversion and Control of Ministry of Education, Southeast University, Nanjing, China, Post Code:210096)

Abstract: Considering the pattern recognition and clustering problems of thermal multi-mode processes, a hybrid clustering scheme based on the local preserving projection (LPP) subspace was proposed. Firstly, the data for the high-dimensional multi-mode process is projected into the low-dimensional subspace through the LPP method, which eliminates the noise, improves computational efficiency and also retains the local structure. Secondly, combining the advantages of the traditional hierarchical and non-hierarchical clustering algorithms in the LPP subspace, the agglomerative k -means algorithm is used to generate the best clustering solution for multi-mode process data. Taking the recognition and clustering processes of a certain 600 MW unit desulfurization system with the data of multi-mode processes as an example, the effectiveness and practicability of this method are proved.

Key words: multi-mode process, locality preserving projection, k -means, agglomerative hierarchical clustering, wet desulfurization

引言

随着计算机、传感器、数据存储和通信等技术的快速发展,基于统计理论和机器学习的数据驱动建模技术发展迅速,已被广泛应用于热工过程建模和故障检测。Zhu等人^[1]考虑到工业数据存在的离群点及缺失值等问题,提出了一种稳健的数据挖掘方法,建立更稳健的数据模型;任少君^[2]提出了一种基于MITNN方法的热工过程非线性建模方法,并针对建模过程中的过拟合问题和残差污染问题,分别提出了一种新的融合过程先验知识和基于重构神经网络的非线性系统故障诊断方法;吕游等人^[3]基于 NO_x 排放特性的多个热工变量,采用PLS-LSSVM方法建立了燃煤电站锅炉 NO_x 排放模型;许裕栗等人^[4]针对锅炉运行状态,提出了一种基于数据挖掘方法的锅炉在线运行状态的检测方法。

受到火电机组参与调峰调频等因素影响,在实际的热工过程中,负荷会呈现较大范围的变化且有一定的周期性,存在多种工况。热工过程的多模式化或多阶段化会导致采集到的过程数据符合不同模式分布的情况^[5],这会限制传统的单模态数据驱动模型和过程故障检测方法的使用,导致模型的精度下降,出现故障检测漏报和误报的情况^[6-9]。因此,针对热工过程的多模式特性,首先需要研究的是多工况的模式识别问题,即如何通过对复杂热工过程中采集的多模式数据的研究,将多模式过程数据进行聚类识别^[10]。聚类算法通过做出一系列的多元统计决策,将相关的过程数据进行划分,是模式识别领域的主要手段^[11]。目前,主流的聚类算法主要有两类:分层聚类算法和非分层聚类算法^[12]。分层聚类易于执行并提供直观的图形结果(树状图),但是这些算法的启发式特性通常会导致数据集的划分结果欠佳。非分层聚类试图优化得到簇的最佳划分的目标函数,能够产生质量较高的解决方案,但是算法优化过程结果容易陷入局部最小值。同时,该方法执行时需要进行初始化,随机性较大,很难保留最佳的解决方案。非分层聚类的另一个复杂之处在于,必须将聚类的数量指定为算法的参数。

因此,本文采用一种混合聚类方案,解决热工多

模式过程的模式识别和聚类问题。该方案既保留了两种传统聚类算法的优点,又克服了各自的缺点。这种混合聚类的核心思想首先是在一定数量的聚类范围内重复进行多次随机初始化运行的非分层聚类(如 k -means聚类),其次是将得到的随机初始化的运行结果以分层的方式进行汇总。此外,该方案针对非分层聚类方面还提出了一种简单的方法来确定簇的数量的适当范围,最终得到的分层聚类结果是图形化的,允许用户以所需的簇的数量选择一组聚类结果。

为验证混合聚类方法的实际效果,采用某600 MW火电机组脱硫系统的多模式过程数据进行试验。由于原始过程数据处于高维特征空间,而直接在该空间进行聚类操作比较繁琐和耗时,因此利用LPP子空间结合混合聚类方案以提高算法的运行效率^[13]。

1 基于LPP子空间的多模式过程聚类

1.1 多模式过程定义

如果一个或多个变量不满足稳态模式条件,则将过程定义为多模式^[14-15]。这意味着至少存在一个变量 $x(t)$,不满足^[10,14]:

$$\left| \frac{x(t) - x(t_0)}{t - t_0} \right| < T_x, \forall t \in [t_0 - \Delta t, t_0 + \Delta t] \quad (1)$$

式中: $x(t)$ —过程的任一变量; T_x —根据测量单位和过程噪声方差定义的阈值,可以通过过程数据调整 T_x ^[16]。

由于不同模式之间的过程统计信息有一定差别,因此可以将多模式特征视为一种特定类型的瞬态或时变行为。这种模式的可变性意味着过程历史数据根据特定的运行模式具有不同的特征。

1.2 LPP降维

LPP是一种基于几何思想的流形学习方法,用于减少采集的过程变量的维度。由于多模式过程数据在其原始空间中通常会分布在不同的簇中,并且LPP能够保留局部结构,同时减小维度,因此采用LPP作为多模式数据的预处理工具是合理的。

与主成分分析方法(Principal Component Analysis, PCA)类似,LPP通过找到一个转换矩阵 A 将高

维数据 $\mathbf{X} = [x_1, x_2, \dots, x_n]^T \in \mathbf{R}^m$ 降维到低维矩阵 $\mathbf{T} = [t_1, t_2, \dots, t_n]^T \in \mathbf{R}^p (p \ll m)$ 中。因此, t_i 代表 x_i , 其中 $t_i = \mathbf{A}^T x_i, i = 1, 2, \dots, n$ 。

为了找到转换矩阵 $\mathbf{A}$ 并保存局部信息, LPP 首先构造一个邻接图。令 G 表示具有 n 个节点的图。如果 x_i 位于 x_j 的 k 个近邻之间, 或者同样地, x_j 位于 x_i 的 k 个近邻之间, 则节点 x_i 和 x_j 通过一条边连接。

之后, 定义对称矩阵 $\mathbf{W}$ 对边进行加权。令 W_{ij} 代表节点 x_i 和 x_j 的边的权重。对于有边连接的两个点 x_i 和 $x_j, W_{ij} = 1$; 对于没有边连接的两个点 x_i 和 $x_j, W_{ij} = 0$ 。

接着, 解决广义特征值分解问题:

$$\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{a} = \lambda\mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{a} \quad (2)$$

式中: λ —广义特征值, 对应的特征向量为 $\mathbf{a}$; $\mathbf{M}$ —对角矩阵, 其对角线上的值为 $\mathbf{W}$ 矩阵对应的某行或列数值之和, 即 $M_{ii} = \sum_j W_{ji}$; $\mathbf{L} = \mathbf{M} - \mathbf{W}$ —拉普拉斯矩阵。

令列向量 $(a_1, \dots, a_m)$ 为式(2)的解, 且根据其对应的特征值排序, $\lambda_1 < \dots < \lambda_m$ 。则投影可表示为:

$$t_i = \mathbf{A}^T x_i, (\mathbf{A} = a_1, \dots, a_p), p \ll m \quad (3)$$

式中: t_i — p 维向量; $\mathbf{A}$ — $m \times p$ 矩阵。

在生成的 LPP 子空间中, 不仅大大降低了建模的计算复杂度, 而且灵敏度也得到了显著提高。作为预处理工具, 这些特性有助于模式聚类 and 建模过程。

由于多模式过程的多样性, 不同的算法可能更适合于不同的情况^[17-19]。模式聚类的独特之处在于与 LPP 子空间的结合, 简化了聚类任务, 因为它仅需要 p 维数据。

1.3 凝聚 k -means 聚类

假设原始过程数据矩阵 $\mathbf{X}$ 是在 C 种不同的运行模式下采集得到。非分层 k 均值聚类解决的优化问题:

$$J = \sum_{i=1}^n r_{ic} \|x_i - \mu_c\|^2 \quad (4)$$

式中: μ_c —簇 c 的均值向量 ($c = 1, 2, \dots, C$), 且 r_{ic} 是对应的隶属度指标, 其定义为:

$$r_{ic} = \begin{cases} 1, & x_i \in \text{簇 } c \\ 0, & \text{其他} \end{cases} \quad (5)$$

k 均值运算对每个样本进行分类, 以便将其精确分配给一个簇。参数 k 控制簇的数量, 且必须提前给定。此外, k 均值算法的任何单次运行 (无论 k 值为多少) 都可能会在局部最小值处终止。为了解决这些局限性, 采用一种层次凝聚的非层次聚类算法, 以得到整体的聚类解决方案。尽管该方法不能严格保证全局最优, 但提供了一种切实可行的方法来显著提高得到全局最优解的可能性。

将 k 均值算法的某次运算结果存储在 $n \times k$ 的矩阵 $\mathbf{B}^r$ 中。如果在某次运算 r 中, 样本 i 是属于簇 c , 则元素 B_{ic}^r 为 1, 否则为 0。对于 k 值从 2 增加到 $k_{\max}$, 从 k 均值算法运行中生成 $N_{\min}$ 个结果, 则 $N = (k_{\max} - 1) \times N_{\min}$ 个结果可以用矩阵 $\mathbf{B}$ 来表示 ($\mathbf{B}$ 为 $\mathbf{B}^r$ 的联合):

$$\mathbf{B} = [\mathbf{B}^1, \mathbf{B}^2, \dots, \mathbf{B}^N] \quad (6)$$

式中: n —样本数量; r —某次运算的索引值; $\mathbf{B}_{ic}^r$ —矩阵 $\mathbf{B}^r$ 中的第 i 行第 c 列; $k_{\max}$ — k 取到的最大值; $N_{\min}$ —取每个 k 值, k -means 算法运算产生的结果数; N —取 k 从 2 到最大值产生的结果总数。 $n \times n$ 的凝聚距离矩阵 $\mathbf{D}$ 是不同样本之间差异性的有效度量:

$$\mathbf{D} = \mathbf{1} - \frac{1}{N}\mathbf{B}\mathbf{B}^T \quad (7)$$

式中: $\mathbf{1}$ —全 1 矩阵; $\mathbf{D}$ 用于在层次聚类生成的情况下生成树状图。

使用平方误差和作为获得最终分类情况的损失函数:

$$\Delta\text{SSE}(k) = \sum_{i \in j} \{D_{ij}(k+1) - D_{ij}(k)\}^2 \quad (8)$$

簇的最终数目为损失函数的拐点, 即 k 的进一步增加不会导致 ΔSSE 显著降低。

2 仿真实例

以石灰石-石膏湿法脱硫系统为研究对象, 如图 1 所示。其脱硫过程为: 来自锅炉的烟气经烟气系统和引风机增压后到达烟气换热器降温, 被降温后的烟气自下而上进入吸收塔, 而塔内的石灰石浆液经由浆液循环泵的调度自上而下喷淋, 烟气与浆液通过逆流混合的方式, 进行酸碱中和反应, 经过一系列物理和化学反应并伴随着持续的热交换, 从而

脱除烟气中的 SO_2 。脱硫后烟气中的液滴经过除雾器去除,烟气再经烟气换热器增加温度,达标后通过烟囱排入到大气之中;反应后吸收 SO_2 的石灰石浆

液(CaSO_3)流入脱硫塔底部的浆液池中,而增压风机鼓入的空气对其进行强制氧化,生成可二次利用的石膏(CaSO_4)。

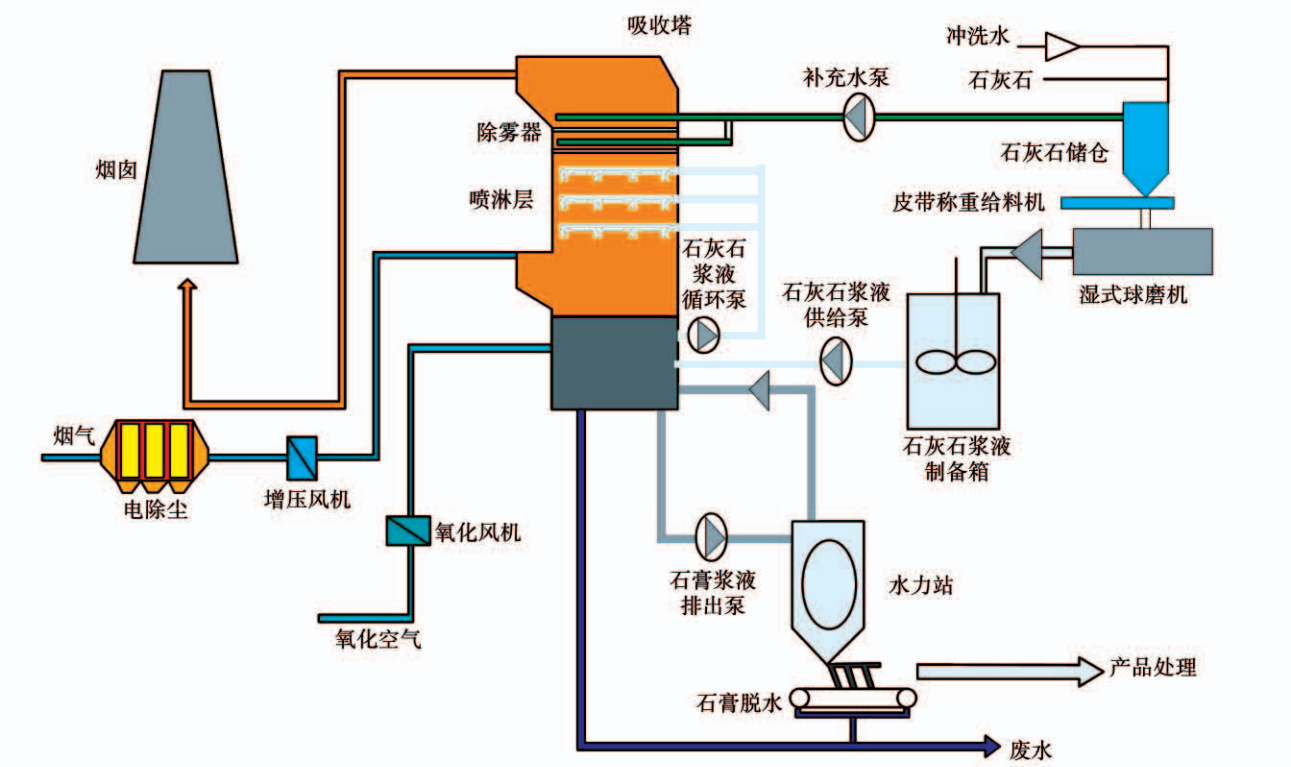


图 1 湿法烟气脱硫系统工艺流程图

Fig. 1 Process flow chart of wet flue gas desulfurization system

如图 1 中所示,在烟气脱硫的过程中,采用不同的浆液循环泵组合形成石灰石浆液的分层喷淋方式,不同分层喷淋方式下烟气与浆液的逆流混合过程随之发生改变, SO_2 的脱除过程也随之发生变化,即整个脱硫系统由于浆液循环泵组合方式的变化呈现了多模式运行的状态;在机组实际运行负荷范围内不同浆液循环泵组合方式运行下,以脱硫系统的多模式运行过程为对象,验证多模式聚类方法的效果。

2.1 数据处理

石灰石-石膏湿法脱硫系统有 4 台离心式浆液循环泵,其主要设计参数如表 1 所示,运行方式为连续运行。结合脱硫系统的多模式过程特性,选取过程数据参数如表 2 所示。从某 600 MW 机组火电厂

厂级监控信息系统(Supervisory Information System, SIS)中采集不同浆液循环泵组合方式下的实际运行数据进行试验。

表 1 浆液循环泵主要性能参数

Tab. 1 Main performance parameters of slurry circulating pump

离心式浆液循环泵	额定功率/kW	流量/ $\text{m}^3 \cdot \text{h}^{-1}$	扬程/m
A	900	9 880	25.3
B	1 000	9 880	27.3
C	1 120	9 880	29.3
D	1 200	9 880	31.3

电厂 SIS 采集的数据往往受到通信和传感器故障等影响,因此需要对采样的数据进行清洗。采用应用较为成熟的 iforest 孤立点检测算法^[20]对异常

数据进行剔除。

表 2 脱硫系统聚类参数

Tab.2 Clustering parameters of the desulfurization systems

参数	单位
发电机功率	MW
进口原烟气 SO ₂ 质量浓度	mg/m ³
进口原烟气流量	m ³ /h
净烟气 SO ₂ 质量浓度	mg/m ³
吸收塔浆液密度	kg/m ³
吸收塔石灰石浆液供给流量	m ³ /h
吸收塔浆液 pH 值	-
液气比	L/m ³
入口氧体积分数	%
入口烟尘质量浓度	mg/m ³
进口烟气温度	℃
吸收塔液位	m

采集到的过程数据需要进行数据标准化预处理,以消除不同参数的测量量纲对聚类结果的影响。采用 z-score 标准化,即数据的中心化和方差归一化方法对数据进行处理:

$$\begin{cases} \hat{x}_i = \frac{x_i - m(x)}{s(x)} \\ m(x) = \frac{1}{n} \sum_{i=1}^n x_i \\ s(x) = \sqrt{\sum_{i=1}^n \frac{(x_i - m(x))^2}{n - 1}} \end{cases} \quad (9)$$

式中: $\hat{x}_i, x_i$ —第 i 个输入参数的原始值和归一化后的值; $m(x)$ —参数的均值; $s(x)$ —参数的标准差。

此外,机组运行工况发生较大变化时,模型变量的统计特性也会发生变动,因此数据预处理还包括了对稳态数据的筛选。选用机组功率为特征变量,对采集的数据进行了稳态判定和筛选^[21]。从图 2 和图 3 可以看出,通过对过程数据进行稳态筛选,剔除编号为 35~40 的升负荷数据和 55~60 的降负荷数据,而其他稳态工况的数据点被完整地保留下来。经过数据处理后,一共筛选出 1 344 组稳态数据。主要包含了 3 种浆液循环泵组合运行方式,如表 3

所示。

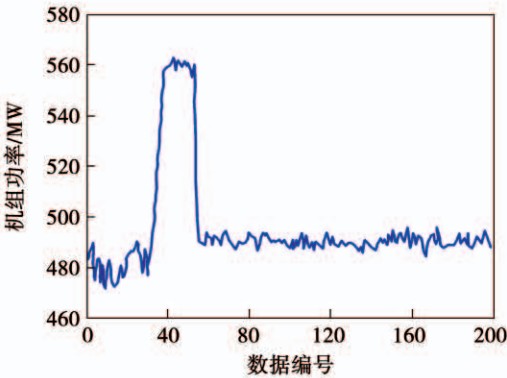


图 2 稳态筛选前负荷变化

Fig. 2 Load change before steady state screening

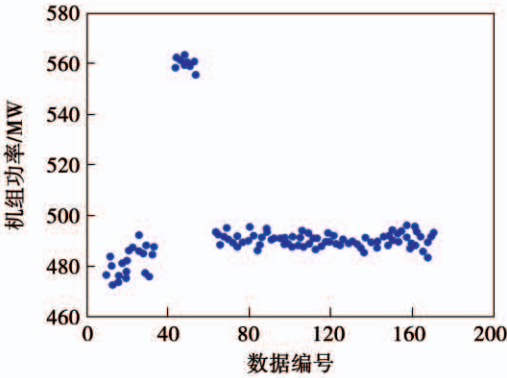


图 3 稳态筛选后负荷变化

Fig. 3 Load change after steady state screening

表 3 多模式样本构成

Tab.3 Multi-mode sample composition

浆液循环泵组合方式	样本编号
2 泵开启	1 ~ 364
3 泵开启	365 ~ 808
4 泵全开	809 ~ 1 345

2.2 LPP 降维

在数据处理和分析后,得到的脱硫系统数据通常是在原始的高维数据空间中。利用 LPP 算法对采集的过程数据进行降维,并保留聚类结构。降维后的过程数据分布如图 4 所示。

当脱硫过程运行于不同的模式时,过程数据的均值、方差的相关关系等统计特征会有明显变化。从各个维度变量自身密度估计图(图中的对角线上的子图)中可以看出,其并不服从单峰高斯分布,并

且明显分为多个模态,即过程数据呈现多模态特征。而局部保留投影(LPP)子空间最大化的保留了原始数据的聚类结构和数据分布情况。

2.3 模式聚类

脱硫系统多模式过程数据经过降维处理后,采用结合分层与非分层方法的混合聚类方案,即凝聚 k -means 聚类方法,在 LPP 子空间中进行模式聚类

与识别,并对结果进行分析。

在进行凝聚 k -means 算法的迭代运行时,设置参数 $N_{\text{run}} = 30, k_{\text{max}} = 7$ 。即当聚类个数 k 从 2 增加到 $k_{\text{max}} = 7$ 时,对应每个 k 值都运行 $N_{\text{run}} = 30$ 次 k -means 聚类算法。得到的聚类结果包括损失函数 ΔSSE 的变化关系,根据距离矩阵得到的分层聚类树状图,以及最终样本的聚类隶属关系,如图 5 所示。

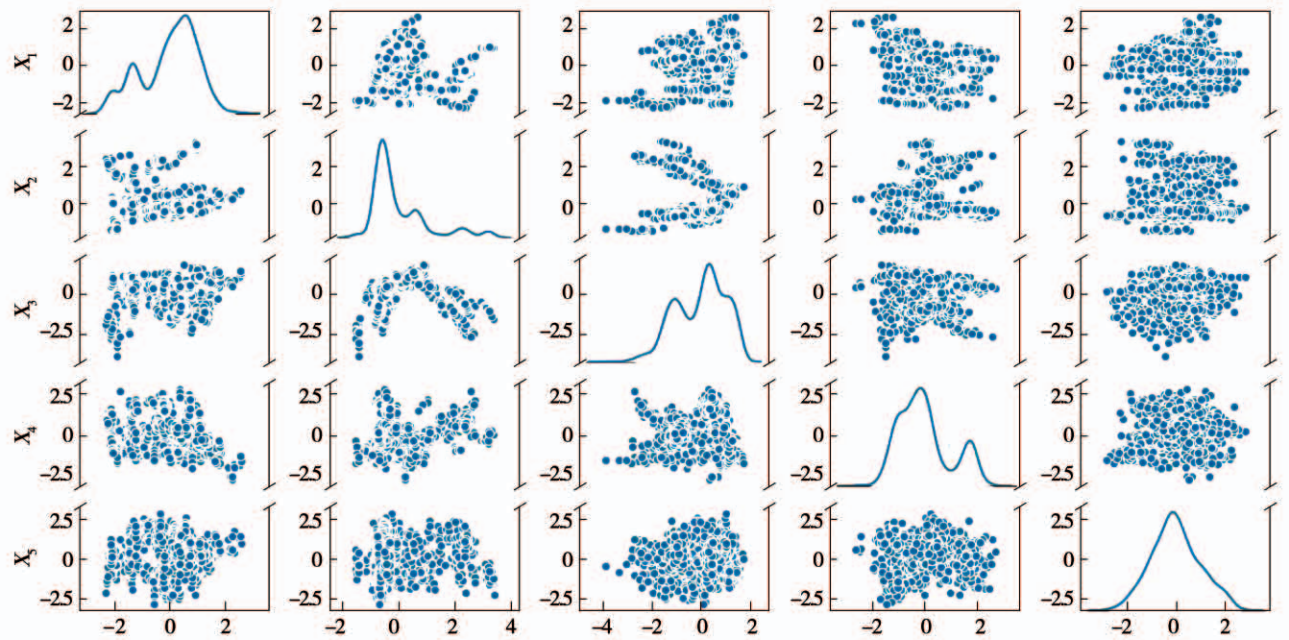


图 4 LPP 子空间中的数据分布
Fig.4 Data distribution in LPP subspace

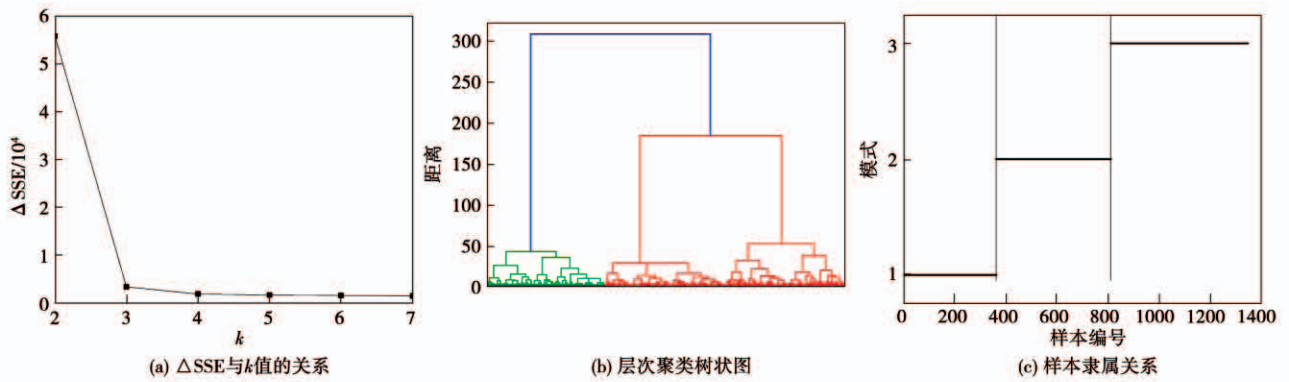


图 5 聚类结果
Fig.5 Clustering results

从图 5(a)中可以看出, $k = 3$ 为聚类损失函数的拐点。此后, k 值的进一步增加对 ΔSSE 降低影响较小。因此,簇的最终数目为 3 个,这也是测试样本里正确的模式数量。

从图 5(b)的树状图可以看出,所有基于 LPP 子空间中的多模式过程数据主要分成了 3 大类,通过调整树状图中截断距离的大小,在 3 个大类中可以根据需要进一步进行细分。

最后,测试样本的隶属关系如图 5(c) 所示,不同模式的过程数据可以被比较完整地识别和区分开。以上的仿真案例表明,基于 LPP 子空间的混合聚类方案可以解决热工过程中多模式场景下高维过程数据的模式识别和聚类问题。

3 结 论

针对热工过程的多模态特性,采用一种混合聚类方案,通过 LPP 子空间的凝聚 k -means 聚类方案,解决热工多模式过程的模式识别和聚类问题。传统的分层聚类方法只能产生一个单一的解决方案,聚类的效果往往很差。而另一方面,非分层聚类方法可以随机生成大量解决方案,其中许多方案可能是有效的。因此提出的混合聚类方法结合了两种传统聚类方法的优势,生成了 1 个单一的图形化高质量聚类结果。以某 600 MW 火电机组脱硫系统的多模式过程实际数据为例,验证了该方法的有效性。

参考文献:

- [1] ZHU J, GE Z, SONG Z, et al. Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data[J]. Annual Reviews in Control, 2018, 46:107 – 133.
- [2] 任少君. 热工过程海量数据流模型分析及诊断方法研究[D]. 南京:东南大学,2018.
REN Shao-jun. Model analysis and diagnosis of mass data flow in thermal process[D]. Nanjing:Southeast University, 2018.
- [3] 吕 游, 刘吉臻, 杨婷婷, 等. 基于 PLS 特征提取和 LS-SVM 结合的 NO_x 排放特性建模[J]. 仪器仪表学报, 2013, 34(11): 2418 – 2424.
LYU You, LIU Ji-zhen, YANG Ting-ting, et al. NO_x emission characteristic modeling based on feature extraction using PLS and LS-SVM[J]. Chinese Journal of Scientific Instrument, 2013, 34(11): 2418 – 2424.
- [4] 许裕栗, 张 静, 李 柠, 等. 基于数据挖掘的锅炉在线运行状态监测[J]. 热能动力工程, 2019, 34(2): 90 – 95, 123.
XU Yu-li, ZHANG Jing, LI Ning, et al. Online operational state monitoring of boiler based on data mining[J]. Journal of Engineering for Thermal Energy and Power, 2019, 34(2): 90 – 95, 123.
- [5] VAZQUEZ L, BLANCO J M, RAMIS R, et al. Robust methodology for steady state measurements estimation based framework for a reliable long term thermal power plant operation performance monitoring[J]. Energy, 2015, 93: 923 – 944.
- [6] YU J. A nonlinear kernel Gaussian mixture model based inferential monitoring approach for fault detection and diagnosis of chemical processes[J]. Chemical Engineering Science, 2012, 68(1): 506 – 519.
- [7] WANG X, WANG Z. A novel method for detecting processes with multi-state modes[J]. Control Engineering Practice, 2013, 21(12): 1788 – 1794.
- [8] XU Y, DENG X G. Fault detection of multimode non-Gaussian dynamic process using dynamic Bayesian independent component analysis[J]. Neurocomputing, 2016, 200: 70 – 79.
- [9] KODAMANA H, RAVEENDRAN R, HUANG B. Mixtures of probabilistic PCA with common structure latent bases for process monitoring, IEEE Trans. Control Syst. Technol. 2017, 99: 1 – 9.
- [10] CAO S, RHINEHART R R. An efficient method for on-line identification of steady state[J]. Process Control, 1995, 5(6): 363 – 374.
- [11] BISHOP C M. Pattern recognition and machine learning(information science and statistics)[M]. Springer-Verlag New York, Inc. 2006.
- [12] EISEN M B. Cluster analysis and display of genome-wide expression patterns. [J]. Proceedings of the National Academy of Sciences of the United States of America, 1998.
- [13] HE X. Locality preserving projections[J]. Advances in Neural Information Processing Systems, 2003, 16(1): 186 – 197.
- [14] SRINIVASAN R, WANG C, HO W K, et al. Dynamic principal component analysis based methodology for clustering process states in agile chemical plants[J]. Ind. Eng. Chem. Res, 2004, 43(9): 2123 – 2139.
- [15] SRINIVASAN R, VISWANATHAN P, VEDAM H, et al. A framework for managing transitions in chemical plants[J]. Comput. Chem. Eng, 2005, 29(2): 305 – 322.
- [16] KELLY J D, HEDENGREN J D. A steady-state detection(SSD) algorithm to detect nonstationary drifts in processes[J]. J Process Control, 2013, 23(3): 326 – 331.
- [17] MCLACHLAN G J, BASFORD K E. Mixture models: inference and applications to clustering[M]. New York:Marcel Dekker, 1988.
- [18] VESANTO J, ALHONIEMI E. Clustering of the self-organizing map[J]. IEEE Transactions on Neural Networks, 2000, 11: 586 – 599.
- [19] VISWANATH P, BABU V S. Rough-DBSCAN: a fast hybrid density based clustering method for large data sets[J]. Pattern Recognition Letters, 2009, 30: 1477 – 1488.
- [20] LIU F T, TING K M, ZHOU Z H. Isolation-based anomaly detection[J]. ACM Transactions on Knowledge Discovery from Data, 2012, 6(1): 1 – 39.
- [21] 毕小龙, 王洪跃, 司风琪, 等. 基于趋势提取的稳态检测方法[J]. 动力工程, 2006, 26(4): 503 – 506.
BI Xiao-long, WANG Hong-yue, SI Feng-qi, et al. A method based on tendency distillation for ascertaining state steadiness[J]. Journal of Power Engineering, 2006, 26(4): 503 – 506.