基于鲁棒加权模糊聚类的污水处理过程监测方法

张瑞垚 [1] 周 平 [1]

摘 要 针对非线性强、先验故障知识少、异常工况识别难的污水处理过程监测问题, 提出一种基于鲁棒加权模糊 c 均值 (Robust weighted fuzzy c-means, RoW-FCM) 聚类与核偏最小二乘 (Kernel partial least squares, KPLS) 的过程监测方法. 首先, 针对污水处理过程的高维非线性耦合特性, 采用核偏最小二乘对高维输入变量进行降维; 其次, 针对传统基于最近邻分配的模糊 c 均值算法对离群点敏感以及存在聚类不平衡簇的问题, 提出充分考虑样本间相互关系的基于鲁棒加权模糊 c 均值聚类算法. 通过引入可能性划分矩阵作为权值参数实现不同样本数据的区分加权, 提高了离群点数据聚类的鲁棒性, 同时引入聚类大小控制参数解决不平衡簇的问题. 进一步将基于鲁棒加权模糊 c 均值算法对核偏最小二乘降维后的得分矩阵进行聚类, 利用聚类得到的隶属度矩阵实现异常工况的检测; 最后, 建立隶属度矩阵与过程变量的回归模型, 并利用得到的变量贡献矩阵描述变量对各个簇的解释程度, 实现异常工况的识别. 数值仿真以及污水处理过程数据实验表明该方法具有更好的鲁棒性能, 在异常工况检测和识别上具有较好的效果.

关键词 污水处理, 鲁棒加权模糊 c 均值, 核偏最小二乘, 过程监测

引用格式 张瑞垚, 周平. 基于鲁棒加权模糊聚类的污水处理过程监测方法. 自动化学报, 2022, **48**(9): 2198−2211

DOI 10.16383/j.aas.c200392

Robust Weighted Fuzzy Clustering for Sewage Treatment Process Monitoring

ZHANG Rui-Yao[1] ZHOU Ping[1]

Abstract Aiming at the problems of strong nonlinearity, little prior knowledge of faults, and difficulty in identifying abnormal working-conditions in the sewage treatment process, this paper proposes a novel process monitoring method based on robust weighted fuzzy c-means (RoW-FCM) clustering and kernel partial least squares (KPLS). First, the KPLS algorithm is presented to reduce the dimensionality of the high-dimensional input variables for the sewage treatment process with complicated nonlinear coupling characteristics. Second, the fact that in view of the traditional fuzzy c-means algorithm based on nearest neighbor assignment is sensitive to outliers and there are unbalanced clusters in clustering, an RoW-FCM clustering algorithm is proposed, which fully considers the relationship between samples. For this RoW-FCM, by introducing the possibility partition matrix as the weight parameter to distinguish and weight different samples, the robustness of outlier data clustering is improved, and the problem of unbalanced cluster is solved by introducing the cluster size control parameter. By clustering the score matrix after dimension reduction with KPLS, the membership matrix can be obtained, which will be used for detecting the abnormal working-conditions. On this basis, the regression model between the membership matrix and the process variables is established, and the resulted variable contribution matrix, which describes the explanatory degree of each cluster, will be used to identify the abnormal working-conditions. At last, both numerical simulation and data experiments of sewage treatment process show that the proposed method has better robust performance and better effect in detecting and identifying the abnormal working-conditions.

Key words Sewage treatment, robust weighted fuzzy c-means, kernel partial least squares, process monitoring

Citation Zhang Rui-Yao, Zhou Ping. Robust weighted fuzzy clustering for sewage treatment process monitoring. *Acta Automatica Sinica*, 2022, **48**(9): 2198−2211

污水处理工业在中国水资源可持续发展中占据重要一环. 目前, 应用最广泛的污水处理工艺是活性污泥法[1]. 如图 1 所示, 活性污泥法污水处理工艺流程通常按照处理程度分为一级处理 (预处理)、二级处理 (生化处理) 和三级处理 (深度处理)[2]. 原污水首先经过格栅拦截较大悬浮物或漂浮杂质后进入沉砂池, 沉砂池将密度较大无机悬浮物从污水中分离, 然后进入初沉池. 完成一级处理的污水经初沉池出水, 并与回流的二沉池沉淀污泥按一定比例混合进入曝气池. 曝气池分为缺氧区和好氧区. 在缺氧区中, 内循环回流的硝态氮在异养菌无氧呼吸作

收稿日期 2020-06-09 录用日期 2020-09-07
Manuscript received June 9, 2020; accepted September 7, 2020
国家自然科学基金 (61890934, 61790572, 61991400), 辽宁省“兴辽英才计划”(XLYC1907132) 和中央高校基本科研业务费 (N180802003) 资助
Supported by National Natural Science Foundation of China (61890934, 61790572, 61991400), Liaoning Revitalization Talents Program (XLYC1907132), and Fundamental Research Funds for the Central Universities (N180802003)
本文责任编委 魏庆来
Recommended by Associate Editor WEI Qing-Lai
1. 东北大学流程工业综合自动化国家重点实验室 沈阳 110819
1. State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819

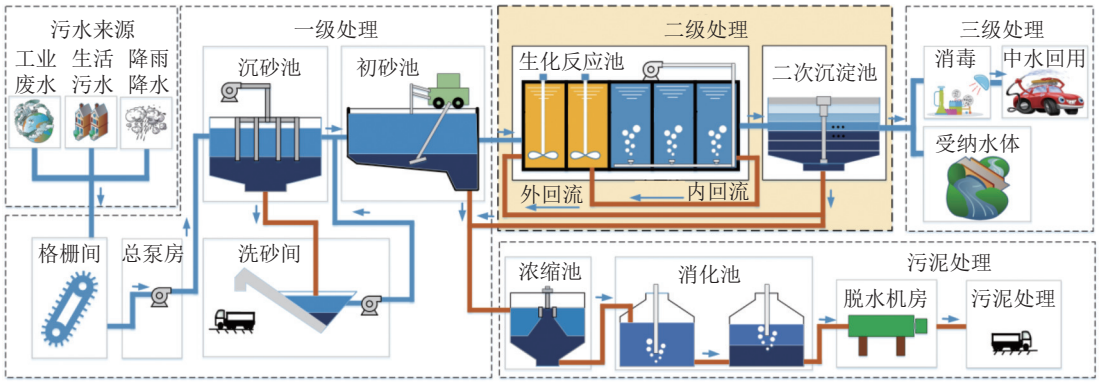


图 1　污水处理工艺流程示意图

Fig. 1　Schematic diagram of sewage treatment process

用下被还原为氮气; 在好氧区中, 氨氮在自养菌有氧呼吸作用下发生硝化反应, 有机物被进一步降解. 随后污水经曝气池出水进入二沉池, 将澄清水与活性污泥进行固液分离. 分离后, 澄清水排入受纳水体或经过物理、化学等技术进一步去除污染物后实现中水回用. 二沉池除回流污泥外的沉淀污泥与初沉池的污泥混合, 经过浓缩、消化、脱水等工艺后做最终处置及回收利用[2−4].

污水处理的根本目的是将城市生活、工业生产等产生的污水经过上述污水处理的各道工序后达到国家规定的出水指标. 目前, 污水处理出水质量指标主要包括生化需氧量、化学需氧量、总悬浮物、总磷、氨氮等. 在污水处理过程中, 由于进水流量、进水组分、污染物种类、天气变化等都是被动接受, 微生物种群、溶解氧浓度、污水 pH 值等多种因素对微生物的生命活动都会产生巨大的影响, 因此保持污水处理厂的长期稳定运行十分困难[4−5]. 由于污水处理时常处于非平稳状态运行, 因此容易引发异常工况的发生. 如果不能及时监测到污水处理过程异常工况, 导致不能正确判断且没有采取有效措施加以调整纠正, 会导致出水水质不达标、污水处理能力降低, 甚至会引发污水处理过程的崩溃, 导致不可逆的事故发生, 使得运行成本大大增加并且造成环境污染. 所以, 通过建立有效监测方法来监测污水处理过程, 对异常工况做出准确判断, 并及时准确地采取有效措施, 对保证污水处理过程安全稳定顺行以及出水水质的达标尤其重要.

由于污水处理过程是一个多变量、强耦合、大时滞、高度非线性的复杂动态非平稳生化反应过程[5], 机理模型很难完全考虑污水处理全流程的运行状态. 大部分机理模型都是基于局部过程建立的, 因此在描述污水处理过程特性时具有很大的局限性[6−7], 这就促进了数据驱动尤其是基于机器学习与多元统计分析的过程监测与故障诊断方法在污水处理过程中的应用[5]. 文献 [8] 提出了一种基于在线估计技术和反向传播神经网络的故障检测和诊断方法, 不仅具有鲁棒性, 而且能够避免阈值问题, 显示出较好的应用可靠性. 文献 [9] 提出的粗集支持向量机分类方法降低了样本属性并保留一定的冗余性, 对污水处理过程运行状态的监测实验验证了该方法的有效性. 文献 [10] 针对主元分析对于噪声和不确定信息描述能力不足的问题, 提出了因子分析故障诊断方法, 在污水仿真基准模型的验证表明该方法能够降低传统主元分析方法的故障误报率, 对不确定信息具有较好的描述能力. 近年来, 由于污水处理数据缺少分类标识, 且先验知识匮乏, 因此模糊聚类技术在污水处理过程监测中得到了越来越多的应用. 模糊聚类是一种无监督分类技术, 本身具有捕获数据非线性结构的能力, 可以充分挖掘污水处理过程的数据信息, 通过建立模糊相似关系对过程进行监测和诊断[11]. 文献 [12] 针对采样数据维度过高的问题, 采用了主元分析和可能性模糊 c 均值 (Possibilistic fuzzy c-means, PFCM) 聚类相结合的方法, 在田纳西−伊斯曼过程仿真实验中取得较好效果. 但是主元分析是一种线性降维技术, 对于污水处理这样的高维非线性系统, 其实际应用效果会有很大局限性. 文献 [13] 提出了偏最小二乘、可能性聚类 (Possibilistic c-means, PCM) 与模糊 c 均值 (Fuzzy c-means, FCM) 的组合方法, 并给出了一种递归原型更新算法. 偏最小二乘算法的使用抑制了与输出数据无关的噪声和变化, 促进了 PCM 和 FCM 的应用, 使其更容易找到簇和相应的原型, 但聚类算法 FCM 对离群点敏感, 因此其监测效果易受离群点影响, 鲁棒性差. 当监测到异常工况发生时, 需要及时识别出导致异常工况发生的异常变量. 目前, 贡献图方法是最为普遍的故障识别方法[14]. Zhou 等[15] 提出了基于主元分析的贡献图方法, 用于辨识与故障相关的关键变量. Dunia 等[16] 提出了

基于重构和平方预测误差方法, 即利用重构平方预测误差与实际平方预测误差的比值进行故障辨识. 文献 [17] 提出了一种基于核主成分分析的方法, 特别是在鲁棒重构误差的基础上, 提出了一种新的故障识别方法. 其基本思路是当重构的变量是故障变量时, 此变量的故障指标会比非故障变量的指标值偏小. 如今, 基于模糊聚类的故障识别方法的研究也得到越来越多专家学者的研究. 文献 [18] 提出了一种基于自回归滑动平均模型双谱分布特征与模糊 c 均值聚类分析的故障识别方法, 该方法通过 FCM 聚类构造类模板和最小距离模板的分类器, 实现了滚动轴承的故障识别. 文献 [19] 将模糊 c 均值算法和 Gustafson-Kessel 聚类算法用于燃气轮机故障的故障检测和识别, 仿真结果表明模糊聚类方法具有可接受的故障识别性能.

综上, 本文针对非平稳污水处理工业过程的非线性强、先验故障知识少、异常工况识别难等问题, 提出了一种基于鲁棒加权模糊 c 均值 (Robust weighted fuzzy c-means, RoW-FCM) 与核偏最小二乘 (Kernel partial least squares, KPLS) 算法的新型过程监测方法. 首先, 采用 KPLS 对污水处理过程的高维输入过程变量进行降维, 同时解决了污水处理数据的非线性问题; 其次, 采用 RoW-FCM 聚类算法对通过 KPLS 算法降维得到的得分矩阵聚类, 通过聚类得到的隶属度矩阵进行污水处理过程异常工况检测分析; 再次, 建立隶属度矩阵与样本数据变量之间的回归模型, 通过解得的变量贡献矩阵进行异常工况识别; 最后, 对本文 RoW-FCM 算法进行数值仿真验证, 并基于污水处理过程数据进行实验验证和对比分析.

1 过程监测策略

提出的基于 RoW-FCM 聚类与 KPLS 的污水处理过程监测方法如图 2 所示, 主要包括高维数据降维、异常工况检测和异常工况识别 3 个部分.

1) 高维数据降维: 污水处理过程相应过程运行性能与出水水质的变量较多, 具有高维特性, 而且变量之间存在着很强的关联耦合特性. 如果把全部变量都用于模型的建立, 不仅会加大计算复杂度, 而且会由于冗余信息干扰影响建模与监测的性能, 因此需要对输入变量数据进行降维. 为此, 采用非线性的 KPLS 方法对高维数据进行降维. 首先将标准化后的过程变量投影到高维特征空间, 然后在高维特征空间建立过程变量与质量变量的偏最小二乘模型, 并采用交叉验证法确定主元数, 得到得分矩阵, 也即原始高维变量经过降维处理后的低维变量.

2) 异常工况检测: 针对常规 FCM 算法对于离群点敏感, 建立 RoW-FCM 聚类算法, 通过引入了权值参数对不同质量的样本数据的区分加权, 改善了聚类对离群点的鲁棒性, 同时引入聚类大小控制参数解决了不平衡簇问题. 由于传统基于欧氏距离的 FCM 算法是根据最近邻分配, 即对于球形数据集以外的如椭圆形类数据集不能有效聚类, 因此采用马氏距离, 可以充分考虑样本之间的相互关系. 将本文改进聚类算法对得分矩阵聚类, 得到隶属度矩阵, 通过所得隶属度矩阵对污水处理过程进行异常工况检测.

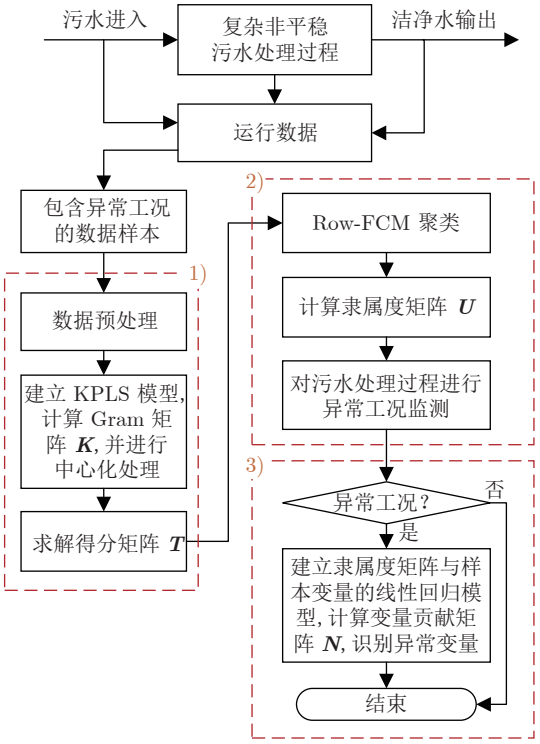


图 2 本文监测算法建模策略

Fig. 2 The monitoring algorithm modeling strategy in this paper

3) 异常工况识别: 为了识别导致异常工况的主导变量, 考虑变量对过程异常工况的解释程度. 基于此, 通过建立隶属度矩阵与过程变量的回归模型, 得到变量隶属度矩阵, 利用变量贡献矩阵描述变量对各个簇的解释程度, 即变量对各类工况的解释程度, 从而达到对异常工况识别的目的.

2 过程监测算法

2.1 高维数据降维的 KPLS 算法

设图 1 所示活性污泥污水处理过程的输入变量矩阵为 $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]^{\mathrm{T}} \in \mathbf{R}^{n \times m_1}$, 出水质量变量矩阵为 $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n]^{\mathrm{T}} \in \mathbf{R}^{n \times l}$, 式中 n 为样本数, m_1 为过程变量数, l 为质量变量数. 将输

入过程变量 $\{\boldsymbol{x}_i\}_{i=1}^n$ 通过非线性变换 ϕ 映射到高维特征空间 F, 如下所示:

$$\phi : \boldsymbol{x}_i \in \mathbf{R}^{m_1} \rightarrow \phi(\boldsymbol{x}_i) \in F \tag{1}$$

用 $\boldsymbol{\Phi}$ 表示输入矩阵 $\boldsymbol{X}$ 映射到高维特征空间 F 后的特征矩阵:

$$\boldsymbol{\Phi} = [\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_2), \cdots, \phi(\boldsymbol{x}_n)]^{\mathrm{T}} \in \mathbf{R}^{n\times M} \tag{2}$$

式中, M 为特征空间维数. 注意到在计算 Gram 矩阵 $\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} \in \mathbf{R}^{n\times n}$ 时, 通过使用核技巧 $\boldsymbol{K}_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle\phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j)\rangle$ 就无需确定 $\boldsymbol{\Phi}$ 的具体形式, 避免了在高维特征空间的复杂内积运算. 本文选用如下高斯函数作为核函数:

$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(\frac{-\|\boldsymbol{x}-\boldsymbol{y}\|^2}{c}\right) \tag{3}$$

式中, c 为高斯核函数宽度, 由 $5m_1$ 经验原则确定. Gram 矩阵 $\boldsymbol{K}$ 的中心化处理可按下式计算:

$$\bar{\boldsymbol{K}} = \left(\boldsymbol{E}_n - \frac{\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}}{n}\right)\boldsymbol{K}\left(\boldsymbol{E}_n - \frac{\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}}{n}\right) \tag{4}$$

式中, $\bar{\boldsymbol{K}}$ 为中心化后的 $\boldsymbol{K}$, $\boldsymbol{E}_n$ 为 $n\times n$ 的单位矩阵, $\mathbf{1}_n = [1, 1, \cdots, 1]^{\mathrm{T}} \in \mathbf{R}^n$.

使用 KPLS 方法可将 $\bar{\boldsymbol{K}}$ 和 $\boldsymbol{Y}$ 矩阵分解为:

$$\begin{cases} \bar{\boldsymbol{K}} = \hat{\boldsymbol{K}} + \boldsymbol{K}_r = \boldsymbol{T}\boldsymbol{P}^{\mathrm{T}} + \boldsymbol{K}_r \\ \boldsymbol{Y} = \hat{\boldsymbol{Y}} + \boldsymbol{Y}_r = \boldsymbol{T}\boldsymbol{Q}^{\mathrm{T}} + \boldsymbol{Y}_r \end{cases} \tag{5}$$

式中, $\boldsymbol{K}_r$ 和 $\boldsymbol{Y}_r$ 分别表示 $\bar{\boldsymbol{K}}$ 和 $\boldsymbol{Y}$ 的建模误差, $\boldsymbol{P} \in \mathbf{R}^{m_1\times A}$ 和 $\boldsymbol{Q} \in \mathbf{R}^{l\times A}$ 分别为 $\boldsymbol{\Phi}$ 和 $\boldsymbol{Y}$ 的负载矩阵, $\boldsymbol{T} = [\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_A] \in \mathbf{R}^{n\times A}$ 为 $\bar{\boldsymbol{K}}$ 和 $\boldsymbol{Y}$ 的得分矩阵, A 为交叉验证决定的主元个数, 并设 $\boldsymbol{t}_i^{row}$ 为得分矩阵 $\boldsymbol{T}$ 的第 i 个行向量.

2.2　过程监测的鲁棒加权模糊 c 均值 (RoW-FCM) 聚类算法

2.2.1　FCM 与 PCM 算法简介

聚类算法中, 比较有影响的重要工作就是 Dunn 将常规硬聚类目标函数推广到了模糊情形, 而 Bezdek 等[20] 又将 Dunn 的目标函数做了推广, 给出了如下基于目标函数的模糊聚类分析更一般的描述:

$$\begin{cases} \min\ J_{\mathrm{FCM}} = \displaystyle\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m d^2(\boldsymbol{t}_k^{row}, \boldsymbol{v}_i) \\ \text{s.t.} \\ \displaystyle\sum_{i=1}^{c} u_{ik} = 1,\ u_{ik} \in [0, 1], \quad 1 \le i \le c,\ 1 \le k \le n \end{cases} \tag{6}$$

式中, $\boldsymbol{U} = [u_{ik}]_{c\times n}$ 表示隶属度矩阵, u_{ik} 表示样本 $\boldsymbol{t}_k^{row}$ 对第 i 个聚类中心 $\boldsymbol{v}_i$ 的隶属度; $m \in [1, +\infty]$ 为模糊指数, 影响隶属度矩阵的模糊程度; $d^2(\boldsymbol{t}_k^{row}, \boldsymbol{v}_i)$ 表示样本 $\boldsymbol{t}_k^{row}$ 与第 i 个聚类中心 v_i 之间的欧氏距离.

$$u_{ik} = \left\{\sum_{j=1}^{c}\left(\frac{d^2(\boldsymbol{t}_k^{row}, \boldsymbol{v}_i)}{d^2(\boldsymbol{t}_j^{row}, \boldsymbol{v}_i)}\right)^{\frac{1}{m-1}}\right\}^{-1} \tag{7}$$

$$\boldsymbol{v}_i = \frac{\displaystyle\sum_{k=1}^{n}(u_{ik})^m \boldsymbol{t}_k^{row}}{\displaystyle\sum_{k=1}^{n}(u_{ik})^m} \tag{8}$$

Krishnapuram 等[21] 在 FCM 算法的基础上放松了对隶属度的概率约束, 提出了可能性聚类 (PCM) 算法, 该算法的目标函数及约束条件如下:

$$\begin{cases} \min\ J_{\mathrm{PCM}} = \displaystyle\sum_{i=1}^{c}\sum_{k=1}^{n}(w_{ik})^p d^2(\boldsymbol{t}_k^{row}, \boldsymbol{v}_i) + \\ \qquad\qquad \displaystyle\sum_{i=1}^{c}\tau_i\sum_{k=1}^{n}(1-w_{ik})^p \\ \text{s.t.} \\ \displaystyle\sum_{i=1}^{c} w_{ik} > 0,\ w_{ik} \in [0, 1], \quad 1 \le i \le c,\ 1 \le k \le n \end{cases} \tag{9}$$

式中, w_{ik} 表示样本 $\boldsymbol{t}_k^{row}$ 对第 i 个聚类中心 $\boldsymbol{v}_i$ 的可能性, 定义可能性划分矩阵 $\boldsymbol{W} = [w_{ik}]_{c\times n}$; $p \in [1, +\infty]$ 为可能性划分指数; $d^2(\boldsymbol{t}_k^{row}, \boldsymbol{v}_i)$ 表示样本 $\boldsymbol{t}_k^{row}$ 与第 i 个聚类中心 $\boldsymbol{v}_i$ 之间的欧氏距离; τ_i 为惩罚因子; $\sum_{i=1}^{c}\tau_i\sum_{k=1}^{n}(1-w_{ik})^p$ 为惩罚项, 避免可能性矩阵 $\boldsymbol{W}$ 为 $\mathbf{0}$ 的情况.

通常使用拉格朗日乘子法求解 PCM 目标函数极值对应的 w_{ik} 和 $\boldsymbol{v}_i$, 如下所示:

$$w_{ik} = \left\{1 + \left(\frac{d^2(\boldsymbol{t}_k^{row}, \boldsymbol{v}_i)}{\tau_i}\right)^{\frac{1}{m-1}}\right\}^{-1} \tag{10}$$

$$\boldsymbol{v}_i = \frac{\displaystyle\sum_{k=1}^{n}(w_{ik})^p \boldsymbol{t}_k^{row}}{\displaystyle\sum_{k=1}^{n}(w_{ik})^p} \tag{11}$$

惩罚因子 τ_i 的计算公式如下:

$$\tau_i = K_1\frac{\displaystyle\sum_{k=1}^{n}(w_{ik})^p d^2(\boldsymbol{t}_k^{row}, \boldsymbol{v}_i)}{\displaystyle\sum_{k=1}^{n}(w_{ik})^p} \tag{12}$$

式中, $K_1 > 0$, 通常取值为 1.

2.2.2 本文 RoW-FCM 算法及异常工况检测

聚类的鲁棒性是指所实现分区的稳定性和可再现性, 以及对噪声和离群点的不敏感性[22−23]. FCM 算法由于对隶属度的约束, 使得聚类结果对离群点特别敏感. 为了解决这个问题, 已有学者提出了多种解决方案. Barni 等[24] 提出的 PCM算法放松了对隶属度的概率约束, 使其对离群点具有较强的鲁棒性, 但容易导致重合聚类. Timm 等[25] 在所有的 PCM 集群原型之间建立一个排斥力, 其强度随着距离的增加而降低. 该方法有效避免了重合聚类, 但在两个聚类之间非常接近的情况下却不能准确处理. 针对 FCM 和 PCM 存在的上述问题, Pal 等[26] 提出了 PFCM 聚类算法, PFCM具有 FCM 与 PCM 的优点, 具有较好的鲁棒性, 但对参数设置有很大的依赖性. 基于此, 针对现有方法存在的上述问题, 提出鲁棒加权模糊 c 均值 (RoW-FCM) 聚类算法. 首先引入可能性划分矩阵作为权值参数, 同时考虑到欧几里德距离在聚类时的局限[27], 因此采用马氏距离. FCM 等算法的另一个主要缺点是它们倾向于使集群的大小相等. 也就是说, 如果一个大集群的数量不平衡, 那么它的一部分就会被错误地分类为另一个小集群, 考虑到这个问题, 本文进一步利用变量控制簇大小的方法来解决[28]. 综上, 本文 RoW-FCM 算法的聚类目标函数如下:

$$\begin{cases}\min J_{\text{RoW-FCM}}=\displaystyle\sum_{i=1}^{c}\sum_{k=1}^{n}(\alpha_i)^{1-m}(u_{ik})^m(w_{ik})^p\cdot \\ \qquad D(\boldsymbol{t}_k^{row},\boldsymbol{v}_i;\boldsymbol{S}_i)+\displaystyle\sum_{i=1}^{c}\eta_i\sum_{k=1}^{n}(1-w_{ik})^p \\ \text{s.t.} \\ \begin{cases}\displaystyle\sum_{i=1}^{c}u_{ik}=1,\ u_{ik}\in[0,1], & 1\le i\le c,\ 1\le k\le n \\ 0<\displaystyle\sum_{i=1}^{c}w_{ik}<c,\ w_{ik}\in[0,1], & 1\le i\le c,\ 1\le k\le n \\ \displaystyle\sum_{i=1}^{c}\alpha_i=1,\ a_i\in[0,1], & 1\le i\le c \\ |\boldsymbol{S}_i|=1, & 1\le i\le c\end{cases}\end{cases} \tag{13}$$

式中, $D(\boldsymbol{t}_k^{row},\boldsymbol{v}_i;\boldsymbol{S}_i)=(\boldsymbol{t}_k^{row}-\boldsymbol{v}_i)^{\mathrm{T}}\boldsymbol{S}_i^{-1}(\boldsymbol{t}_k^{row}-\boldsymbol{v}_i)$, $\boldsymbol{S}_i$ 是一个正定矩阵, 表示变量的模糊化协方差矩阵; η_i 为惩罚因子; $\sum_{i=1}^{c}\eta_i\sum_{k=1}^{n}(1-w_{ik})^p$ 为惩罚项; $\boldsymbol{A}=[\alpha_i]_{1\times c}$ 为聚类大小控制矩阵, α_i 为聚类大小控制因子.

引入拉格朗日乘子 λ、γ、ξ, 构造如下函数:

$$\begin{aligned}L=&\sum_{i=1}^{c}\sum_{k=1}^{n}(\alpha_i)^{1-m}(u_{ik})^m(w_{ik})^pD(\boldsymbol{t}_k^{row},\boldsymbol{v}_i;\boldsymbol{S}_i)+\\&\sum_{i=1}^{c}\eta_i\sum_{k=1}^{n}(1-w_{ik})^p+\sum_{k=1}^{n}\lambda_k\left(1-\sum_{i=1}^{c}u_{ik}\right)+\\&\sum_{i=1}^{c}\gamma_i|\boldsymbol{S}_i|+\xi\left(1-\sum_{i=1}^{c}\alpha_i\right)\end{aligned} \tag{14}$$

对函数 L 分别求 w_{ik}、u_{ik}、v_{ik}、α_i 的偏导数, 可得:

$$\begin{aligned}\frac{\partial L}{\partial w_{ik}}=0\Rightarrow&(\alpha_i)^{1-m}(u_{ik})^mp(w_{ik})^{p-1}D(\boldsymbol{t}_k^{row},\boldsymbol{v}_i;\boldsymbol{S}_i)-\\&\eta_ip(1-w_{ik})^{p-1}=0\end{aligned} \tag{15}$$

进一步解得:

$$w_{ik}=\left[1+\left(\frac{(\alpha_i)^{1-m}(u_{ik})^mD(\boldsymbol{t}_k^{row},\boldsymbol{v}_i;\boldsymbol{S}_i)}{\eta_i}\right)^{\frac{1}{p-1}}\right]^{-1} \tag{16}$$

$$\begin{cases}\dfrac{\partial L}{\partial u_{ik}}=0\\ m(\alpha_i)^{1-m}(u_{ik})^{m-1}(w_{ik})^pD(\boldsymbol{t}_k^{row},\boldsymbol{v}_i;\boldsymbol{S}_i)=\lambda_k\end{cases} \tag{17}$$

$$u_{ik}=\left(\frac{\lambda_k}{m}\right)^{\frac{1}{m-1}}\left[(\alpha_i)^{1-m}(w_{ik})^pD(\boldsymbol{t}_k^{row},\boldsymbol{v}_i;\boldsymbol{S}_i)\right]^{\frac{-1}{m-1}} \tag{18}$$

由 $\sum_{i=1}^{c}u_{ik}=1$, 有:

$$\begin{aligned}1=\sum_{j=1}^{c}\Bigg\{&\left(\frac{\lambda_k}{m}\right)^{\frac{1}{m-1}}\Big[(\alpha_j)^{1-m}(w_{jk})^p\cdot\\&D(\boldsymbol{t}_k^{row},\boldsymbol{v}_i;\boldsymbol{S}_j)\Big]^{\frac{-1}{m-1}}\Bigg\}\end{aligned} \tag{19}$$

将式 (17) 代入上式, 得:

$$u_{ik}=\frac{\left[(\alpha_i)^{1-m}(w_{ik})^pD(\boldsymbol{t}_k^{row},\boldsymbol{v}_i;\boldsymbol{S}_i)\right]^{-\frac{1}{m-1}}}{\displaystyle\sum_{j=1}^{c}\left[(\alpha_j)^{1-m}(w_{jk})^pD(\boldsymbol{t}_k^{row},\boldsymbol{v}_j;\boldsymbol{S}_j)\right]^{-\frac{1}{m-1}}} \tag{20}$$

$$\begin{cases}\dfrac{\partial L}{\partial v_{ik}}=0\\ \displaystyle\sum_{k=1}^{n}\Big\{(\alpha_i)^{1-m}(u_{ik})^mp(w_{ik})^p\Big[\boldsymbol{S}_i^{-1}(\boldsymbol{t}_k^{row}-\boldsymbol{v}_i)+\\ \qquad(\boldsymbol{S}_i^{-1})^{\mathrm{T}}(\boldsymbol{t}_k^{row}-\boldsymbol{v}_i)\Big]\Big\}=0\end{cases} \tag{21}$$

$$\left[\boldsymbol{S}_i^{-1}+(\boldsymbol{S}_i^{-1})^{\mathrm{T}}\right]\boldsymbol{v}_i=\frac{\left[\boldsymbol{S}_i^{-1}+(\boldsymbol{S}_i^{-1})^{\mathrm{T}}\right]\sum\limits_{k=1}^{n}(u_{ik})^m(w_{ik})^p\boldsymbol{t}_k^{row}}{\sum\limits_{k=1}^{n}(u_{ik})^m(w_{ik})^p} \tag{22}$$

可知$\left[\boldsymbol{S}_i^{-1}+(\boldsymbol{S}_i^{-1})^{\mathrm{T}}\right]$可逆, 解得:

$$\boldsymbol{v}_i=\frac{\sum\limits_{k=1}^{n}(u_{ik})^m(w_{ik})^p\boldsymbol{t}_k^{row}}{\sum\limits_{k=1}^{n}(u_{ik})^m(w_{ik})^p} \tag{23}$$

$$\begin{cases}\dfrac{\partial L}{\partial \boldsymbol{S}_i}=0\\ \sum\limits_{k=1}^{n}\Big[(\alpha_i)^{1-m}(u_{ik})^m(w_{ik})^p\times\\ \qquad\boldsymbol{S}_i^{-1}(\boldsymbol{t}_k^{row}-\boldsymbol{v}_i)(\boldsymbol{t}_k^{row}-\boldsymbol{v}_i)^{\mathrm{T}}\boldsymbol{S}_i^{-1}\Big]+\gamma_i\boldsymbol{S}_i^{-1}=0\end{cases} \tag{24}$$

$$\boldsymbol{S}_i=\frac{1}{\gamma_i}\sum_{k=1}^{n}\Big[(\alpha_i)^{1-m}(u_{ik})^m(w_{ik})^p\times\boldsymbol{S}_i^{-1}(\boldsymbol{t}_k^{row}-\boldsymbol{v}_i)(\boldsymbol{t}_k^{row}-\boldsymbol{v}_i)^{\mathrm{T}}\Big] \tag{25}$$

令 $\theta_i=(\alpha_i)^{1-m}/\gamma_i$, 得:

$$\boldsymbol{S}_i=\theta_i\sum_{k=1}^{n}(u_{ik})^m(w_{ik})^p(\boldsymbol{t}_k^{row}-\boldsymbol{v}_i)(\boldsymbol{t}_k^{row}-\boldsymbol{v}_i)^{\mathrm{T}} \tag{26}$$

为了消除拉格朗日乘子 γ_i, 令:

$$\hat{\boldsymbol{S}}_i=\sum_{k=1}^{n}(u_{ik})^m p(w_{ik})^p(\boldsymbol{t}_k^{row}-\boldsymbol{v}_i)(\boldsymbol{t}_k^{row}-\boldsymbol{v}_i)^{\mathrm{T}} \tag{27}$$

由 $|\boldsymbol{S}_i|=1$ 得:

$$\theta_i=\frac{1}{|\hat{\boldsymbol{S}}_i|} \tag{28}$$

$$\begin{cases}\dfrac{\partial L}{\partial \alpha_i}=0\\ \sum\limits_{k=1}^{n}(1-m)(\alpha_i)^{-m}(u_{ik})^m(w_{ik})^p D(\boldsymbol{t}_k^{row},\boldsymbol{v}_i;\boldsymbol{S}_i)=\xi\end{cases} \tag{29}$$

由 $\sum_{i=1}^{c}\alpha_i=1$, 得:

$$1=\sum_{j=1}^{c}\left[\frac{\sum\limits_{k=1}^{n}[(1-m)(u_{jk})^m(w_{jk})^p D(\boldsymbol{t}_k^{row},\boldsymbol{v}_j;\boldsymbol{S}_j)]}{\xi}\right]^{\frac{1}{m}} \tag{30}$$

将式 (29) 代入上式, 得:

$$\alpha_i=\frac{\left(\sum\limits_{k=1}^{n}(u_{ik})^m(w_{ik})^p D(\boldsymbol{t}_k^{row},\boldsymbol{v}_i;\boldsymbol{S}_i)\right)^{-m}}{\sum\limits_{j=1}^{c}\left(\sum\limits_{k=1}^{n}(u_{jk})^m(w_{jk})^p D(\boldsymbol{t}_k^{row},\boldsymbol{v}_j;\boldsymbol{S}_j)\right)^{-m}} \tag{31}$$

η_i 为惩罚因子, 采用下式计算:

$$\eta_i=K_2\frac{\sum\limits_{k=1}^{n}(u_{ik})^m(w_{ik})^p D(\boldsymbol{t}_k^{row},\boldsymbol{v}_i;\boldsymbol{S}_i)}{\sum\limits_{k=1}^{n}(u_{ik})^m(w_{ik})^p} \tag{32}$$

式中, $K_2>0$, 通常取值为 1.

算法 1. 本文 RoW-FCM 算法

输入数据. $\boldsymbol{T}=[\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_A]$, 设定聚类数目 c、模糊指数 m、可能性划分指数 p, 设置算法终止限 ε、算法最大迭代次数 $count_1$, 初始化迭代次数 $k=1$, 初始化隶属度矩阵 $\boldsymbol{U}^{(1)}=\left[u_{ik}^{(1)}\right]_{c\times n}$、聚类中心 $\boldsymbol{V}^{(1)}=\left[\boldsymbol{v}_i^{(1)}\right]_{c\times A}$、协方差矩阵 $\boldsymbol{S}^{(1)}=\left[\boldsymbol{S}_i^{(1)}\right]_{A\times A\times c}$ 以及聚类大小控制矩阵 $\boldsymbol{A}=[\alpha_i]_{1\times c}$;

1) 利用式 (17) 计算 $\boldsymbol{W}^{(k+1)}=\left[w_{ik}^{(k+1)}\right]_{c\times n}$;

2) 利用式 (20) 计算 $\boldsymbol{U}^{(k+1)}=\left[u_{ik}^{(k+1)}\right]_{c\times n}$;

3) 利用式 (23) 计算 $\boldsymbol{V}^{(k+1)}=\left[\boldsymbol{v}_i^{(k+1)}\right]_{c\times A}$;

4) 利用式 (25) 计算 $\boldsymbol{S}^{(k+1)}=\left[\boldsymbol{S}_i^{(k+1)}\right]_{A\times A\times c}$;

5) 利用式 (31) 计算 $\boldsymbol{A}=[\alpha_i]_{1\times c}$;

6) 如果 $\|\boldsymbol{U}^{(k+1)}-\boldsymbol{U}^{(k)}\|<\varepsilon$ 或算法迭代次数 $k>count_1$ 则算法终止; 否则, $k\leftarrow k+1$, 执行步骤 1.

注 1. 本文聚类算法对于隶属度矩阵和聚类中心的初始化并不敏感, 因此在迭代开始前, 即在算法 1 中的输入数据过程, 隶属度矩阵以及聚类中心的初始值采用随机初始化给出 .

2.3　基于 RoW-FCM 的异常工况识别算法

实际污水处理运行过程中, 当异常工况发生时, 及时识别造成异常工况发生的异常变量对指导操作人员做出有效操作决策具有重要意义. 聚类算法中, 隶属度矩阵描述了样本属于聚类中心的程度. 为了识别与异常工况相关的变量, 本文进一步提出一种新的基于变量贡献矩阵的识别方法. 该方法的基本思想就是: 每个变量对各种工况都有一个贡献值, 并且限定每个变量对所有工况的贡献值之和为 1. 如果某个变量对某个工况的贡献值最大, 即表明此变量是与此工况相关联的变量, 也就认为该变量是造成该工况的关键变量. 变量贡献矩阵通过建立隶属度矩阵与过程变量的线性回归模型得到, 其描述

了样本变量对各个簇的解释程度, 回归模型如下:

$$u_{ik}=\sum_{a=1}^{m_1}x_{ka}\eta_{ai}+\varepsilon_{ik},\quad i=1,\cdots,c,\ \ k=1,\cdots,n \tag{33}$$

式中, ε_{ik} 为误差项, 其期望、方差和协方差分别满足 $\mathrm{E}(\varepsilon_{ik})=0$, $\mathrm{Var}(\varepsilon_{ik})=\delta^2$ (常数), $\mathrm{Cov}(\varepsilon_{ik},\varepsilon_{ij})=0$, $j\neq k$; x_{ka}表示第 k 个样本的第 a 个变量; $\boldsymbol{N}=[\eta_{ai}]_{m_1\times c}$ 为变量贡献矩阵, η_{ai} 表示聚类 i 被变量 x_{ka} 的解释程度.

为求解 η_{ai}, 引入如下损失函数:

$$J=\sum_{i=1}^{c}\sum_{k=1}^{n}\left(u_{ik}-\sum_{a=1}^{m_1}x_{ka}\eta_{ai}\right)^2 \tag{34}$$

从 η_{ai} 的实际意义考虑, 类比隶属度, 对上述损失函数引入约束如下:

$$\sum_{i=1}^{c}\eta_{ai}=1,\ \eta_{ai}\in[0,1] \tag{35}$$

采用拉格朗日乘子法求解变量贡献矩阵 $\boldsymbol{N}$, 引入拉格朗日乘子 ζ, 构造目标函数如下:

$$L=\sum_{i=1}^{c}\sum_{k=1}^{n}\left(u_{ik}-\sum_{a=1}^{m_1}x_{ka}\eta_{ai}\right)^2+\sum_{a=1}^{m_1}\zeta_a\left(1-\sum_{i=1}^{c}\eta_{ai}\right) \tag{36}$$

对函数 L 求 η_{ai} 的偏导数, 可得:

$$\begin{cases}\dfrac{\partial L}{\partial\eta_{ai}}=0\\ \displaystyle\sum_{k=1}^{n}(u_{ik}-x_{ka}\eta_{ai})(-2x_{ka})=\zeta_a\end{cases} \tag{37}$$

$$\eta_{ai}=\frac{\zeta_a+2\displaystyle\sum_{k=1}^{n}x_{ka}u_{ik}}{\displaystyle\sum_{k=1}^{n}x_{ka}^2} \tag{38}$$

由于 $\sum_{a=1}^{c}\eta_{ai}=1$, 并将式 (37) 代入上式, 得:

$$\sum_{j=1}^{c}\eta_{aj}=\sum_{j=1}^{c}\frac{\displaystyle\sum_{k=1}^{n}\left(2x_{ka}^2\eta_{ai}-2x_{ka}u_{ik}\right)+2\sum_{k=1}^{n}x_{ka}u_{jk}}{\displaystyle 2\sum_{k=1}^{n}x_{ka}^2}=1 \tag{39}$$

进一步可得:

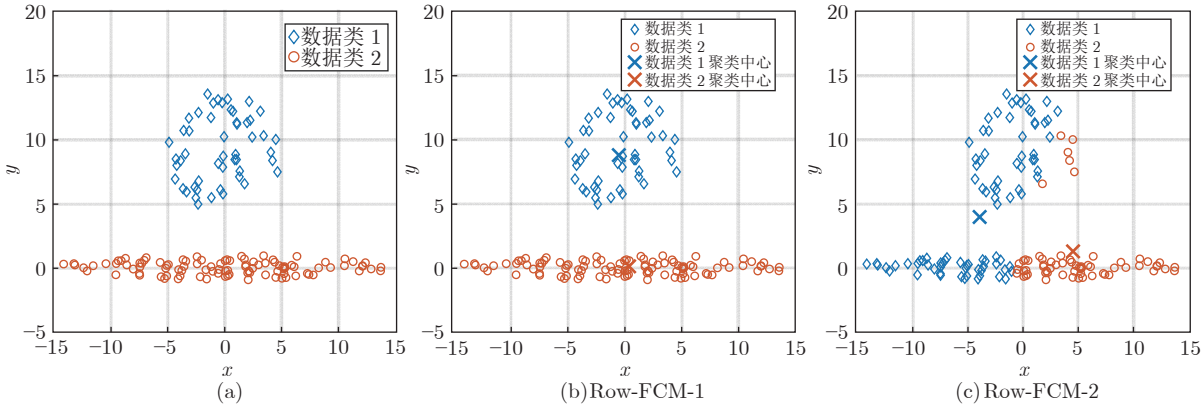
$$\eta_{ai}=\left(1-\frac{2\displaystyle\sum_{k=1}^{n}x_{ka}u_{jk}-c\times\sum_{k=1}^{n}x_{ka}u_{ik}}{\displaystyle\sum_{k=1}^{n}x_{ka}^2}\right)\times\frac{1}{c} \tag{40}$$

最后, 基于式 (40), 根据变量贡献矩阵 $\boldsymbol{N}^{(k+1)}=\left[\eta_{ai}^{(k+1)}\right]_{m_1\times c}$ 对污水处理过程进行异常工况识别, 规则如下: 若第 a 个变量对所有聚类的贡献 $\{\eta_{a1},\cdots,\eta_{ac}t\}$ 中的最大值为 η_{ag}, 则第 a 个变量为与第 g 种异常工况相关的过程变量, 其中 $g\in\{2,\cdots,c\}$, 且第 1 个聚类为正常工况样本的聚类.

3 RoW-FCM 聚类算法验证实验

首先, 采用图 3(a) 所示数据测试基于欧氏距离与马氏距离的聚类方法的性能. 实验数据集分为两组: 数据类 1 在一个半径为 5 的圆中随机生成 50 个样本点, 数据类 2 在一个长轴为 15、短轴为 1 的椭圆中随机生成 100 个样本点, 两组数据聚类中心之间的距离为 9. 本实验在目标函数式 (13) 的基础上分别采用马氏距离与欧氏距离作为对比. 为便于区分, 将采用马氏距离的方法记作 RoW-FCM-1, 将采用欧氏距离的算法记作 RoW-FCM-2. 两种方法聚类效果分别如图 3(b) 和图 3(c)所示. 可以看出, 采用马氏距离可以将椭圆数据集与圆形数据集很好地分开, 而基于欧氏距离的算法则不能将其有效分开.

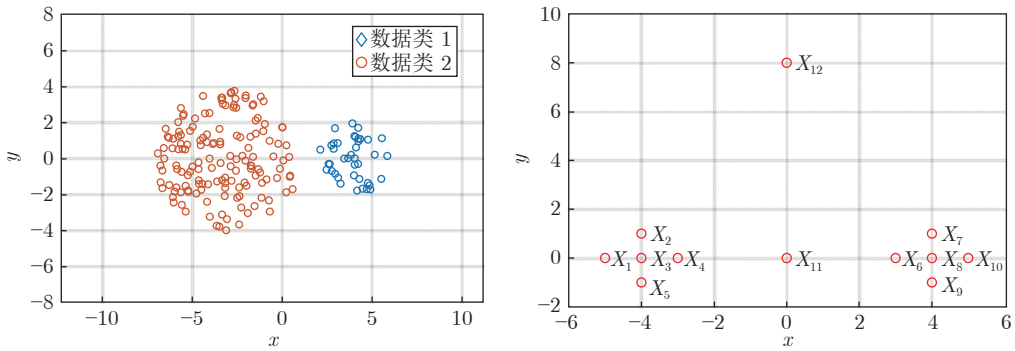
然后, 采用图 4(a) 所示数据集来测试本文方法对于不平衡集群的聚类性能. 图 4(a) 中的数据集分为两类数据: 数据类 1 在一个半径为 4 的圆中随机生成 150 个样本点, 数据类 2 在一个半径为 2 的圆中随机生成 40 个样本点, 两类数据聚类中心之间的距离为 7. 图 5 为分别采用 FCM、PCM、PFCM 和本文 RoW-FCM 对图 4(a) 数据集 A 进行聚类的结果. 由图 5 可知, FCM、PCM、PFCM 三种方法都将大集群的部分数据错误的分类为较小集群的部分, 其中 PCM 的聚类效果最差, 产生了重合聚类, 即聚类中心重合, 而本文 RoW-FCM 算法对两类集群有很好的划分. 为了测试本文 RoW-FCM 算法在聚类时对离群点的鲁棒性, 进一步采用图 4(b) 所示包含离群点的数据集 B 进行鲁棒性的测试. 在数据集中共有 12 个样本点, 其中, 数据类 1: $\{X_1, X_2, X_3, X_4, X_5\}$ 和数据类 2: $\{X_6, X_7, X_8, X_9, X_{10}\}$ 分别为 y 轴对称的聚类, 聚类中心分别为 $v_1^*=(-4,0)$ 和 $v_2^*=(4,0)$, X_{11} 和 X_{12} 为 2 个离群点, 它们距离 2 个聚类中心的距离相等. 在图 4(b)所示数据集上将 FCM、PCM、PFCM 和本文 RoW-FCM 进行



(a)　(b) Row-FCM-1　(c) Row-FCM-2

图 3　仿真实验数据及聚类效果图

Fig. 3　Simulation experiment data and clustering effect diagrams



(a) 不平衡簇测试数据集 A　(b) 离群点数据集 B

(a) Unbalanced cluster test data set A　(b) Outlier data set B

图 4　测试数据集

Fig. 4　Test data sets

数据对比实验. 不同方法的聚类结果参数及聚类效果分别见表 1 与图 6. 表 1 中 $\boldsymbol{U}$ 代表隶属度矩阵, $\boldsymbol{U}_1^{\mathrm{T}}$ 和 $\boldsymbol{U}_2^{\mathrm{T}}$ 分别表示矩阵 $\boldsymbol{U}$ 的第 1 行和第 2 行; $\boldsymbol{W}$ 代表可能性划分矩阵, $\boldsymbol{W}_1^{\mathrm{T}}$ 和 $\boldsymbol{W}_2^{\mathrm{T}}$ 分别表示矩阵 $\boldsymbol{W}$ 的第 1 行和第 2 行. 由聚类结果可知, PCM 算法产生重合聚类, 4 种算法中, PFCM 算法得到的聚类中心的偏移有所减小. 但是, 对比 4 种算法的聚类效果, 本文 RoW-FCM 算法聚类中心偏移距离最小, 受离群点影响最小, 具有最好的鲁棒性.

由本节 3 个数值实验可以看出, 在对不平衡数据集聚类时: 本文 RoW-FCM 算法通过引入控制距离尺寸的参数, 很好地解决了不平衡簇问题, 相比于 FCM、PCM、PFCM 算法有良好聚类性能. 在离群点数据实验中, RoW-FCM 算法比 FCM 和 PFCM 算法的鲁棒性更好. 而且相比于 PFCM, 本文 RoW-FCM 算法聚类性能对参数设置并不敏感. 最后, 相比于基于欧氏距离的聚类算法, RoW-FCM 算法采用马氏距离能够充分考虑样本间关系, 对于非球形数据集也有较好的聚类效果.

4　污水处理过程监测实验

本文基于污水处理过程的基准模型 BSM1 进行数据仿真实验. BSM1 是由欧盟科学技术合作组织与国际水协共同合作开发的一个独立仿真平台, 它能够较为合理地反应污水处理过程的反应机理, 其设备布局由一个生化反应池和一个二次沉淀池组成[29], 具体如图 1 二级处理工艺设备布局图所示. 选取二沉池出水中的生化需氧量、化学需氧量、悬浮物、氨氮作为出水质量指标. 同时, 根据工艺机理, 确定影响出水水质指标的 28 个关键过程变量如表 2 所示. 根据 $5m_1$ 原则, KPLS 的高斯核函数宽度选为 140, 同时通过交叉验证确定 KPLS 主元个数为 3 个. 所用测试数据包含进水流量异常和毒性冲击 2 种异常工况. 其中, 毒性冲击故障是由于来自工业、农业或医院等的有毒物质造成的. 毒性冲击会使活性污泥中的微生物出现 “中毒”现象, 破坏活性污泥系统, 导致污水处理效率下降, 造成生化需氧量、化学需氧量、总氮和总磷等出水指标异

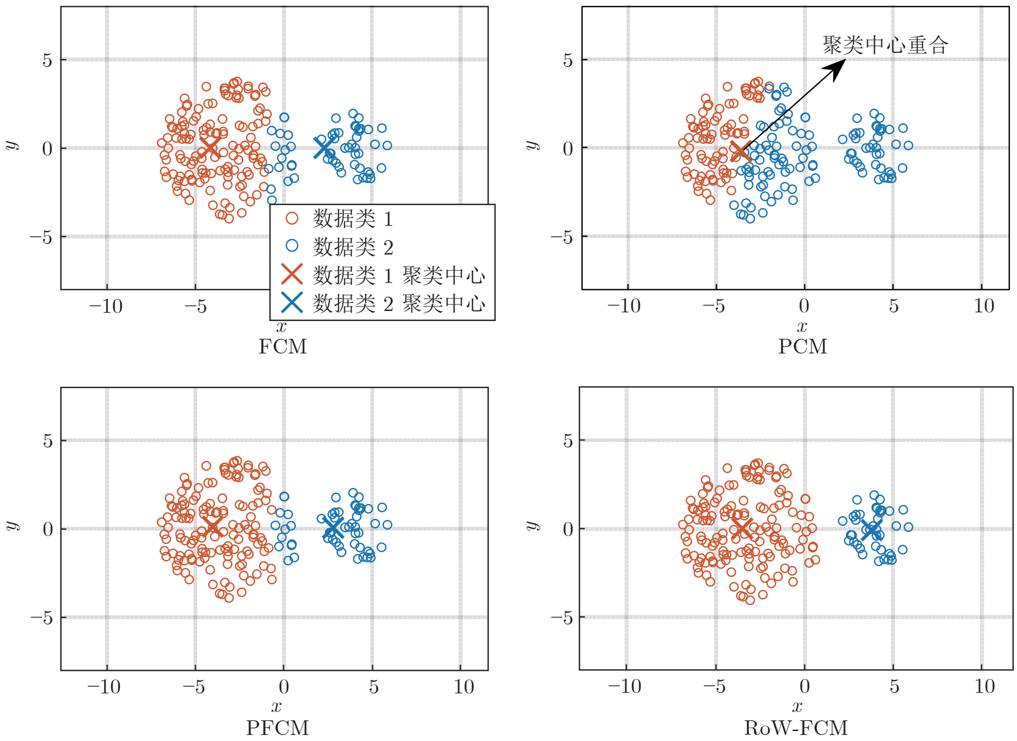


图 5　不平衡簇实验聚类效果图

Fig. 5　Experimental clustering effect of unbalanced clusters

表 1　FCM、PCM、PFCM、RoW-FCM 聚类参数

Table 1　FCM, PCM, PFCM, RoW-FCM clustering parameters

编号	FCM		PCM		PFCM				RoW-FCM			
	$\boldsymbol{U}_1^{\mathrm{T}}$	$\boldsymbol{U}_2^{\mathrm{T}}$	$\boldsymbol{W}_1^{\mathrm{T}}$	$\boldsymbol{W}_2^{\mathrm{T}}$	$\boldsymbol{U}_1^{\mathrm{T}}$	$\boldsymbol{U}_2^{\mathrm{T}}$	$\boldsymbol{W}_1^{\mathrm{T}}$	$\boldsymbol{W}_2^{\mathrm{T}}$	$\boldsymbol{U}_1^{\mathrm{T}}$	$\boldsymbol{U}_2^{\mathrm{T}}$	$\boldsymbol{W}_1^{\mathrm{T}}$	$\boldsymbol{W}_2^{\mathrm{T}}$
1	0.973	0.027	0.799	0.798	0.021	0.979	0.026	0.547	0.991	0.009	0.833	0.999
2	0.991	0.009	0.859	0.858	0.010	0.989	0.032	0.755	0.989	0.011	0.839	0.999
3	0.995	0.005	0.861	0.860	0.002	0.998	0.032	0.940	1.00	0.000	1.000	1.000
4	0.967	0.033	0.848	0.848	0.026	0.975	0.032	0.555	0.989	0.011	0.834	0.999
5	0.988	0.012	0.916	0.916	0.013	0.987	0.042	0.770	0.986	0.014	0.840	0.998
6	0.012	0.988	0.916	0.917	0.987	0.013	0.770	0.042	0.012	0.988	0.999	0.861
7	0.009	0.991	0.859	0.860	0.989	0.011	0.755	0.032	0.011	0.989	0.999	0.835
8	0.005	0.995	0.861	0.862	0.998	0.002	0.940	0.032	0.000	0.999	1.000	0.998
9	0.033	0.967	0.848	0.849	0.975	0.026	0.555	0.032	0.011	0.989	0.999	0.835
10	0.027	0.973	0.799	0.800	0.979	0.021	0.547	0.026	0.010	0.990	0.999	0.811
11	0.500	0.500	0.997	0.997	0.500	0.500	0.125	0.125	0.069	0.931	0.985	0.274
12	0.500	0.500	0.632	0.632	0.500	0.500	0.026	0.026	0.997	0.004	0.060	0.999
聚类中心	$\boldsymbol{v}_1=(-3.616, 0.383)$		$\boldsymbol{v}_1=(0.001, 0.369)$		$\boldsymbol{v}_1=(-3.736, 0.240)$				$\boldsymbol{v}_1=(-3.989, 0.010)$			
	$\boldsymbol{v}_2=(3.616, 0.384)$		$\boldsymbol{v}_2=(0.007, 0.369)$		$\boldsymbol{v}_2=(3.736, 0.240)$				$\boldsymbol{v}_2=(3.910, 0.000)$			
偏移距离	$r_1=0.543$		$r_1=4.016$		$r_1=0.357$				$r_1=0.010$			
	$r_2=0.543$		$r_2=4.010$		$r_2=0.357$				$r_2=0.090$			

常. 本文通过降低异养菌最大比生长速率 μ_H, 增大异养菌衰减系数 b_H 来模拟毒性冲击[30]. 因此, 本文主要就进水流量异常和毒性冲击两种异常工况的检测和识别问题进行分析.

4.1　污水处理过程异常工况检测结果

首先对测试数据集进行故障检测实验. 分别采

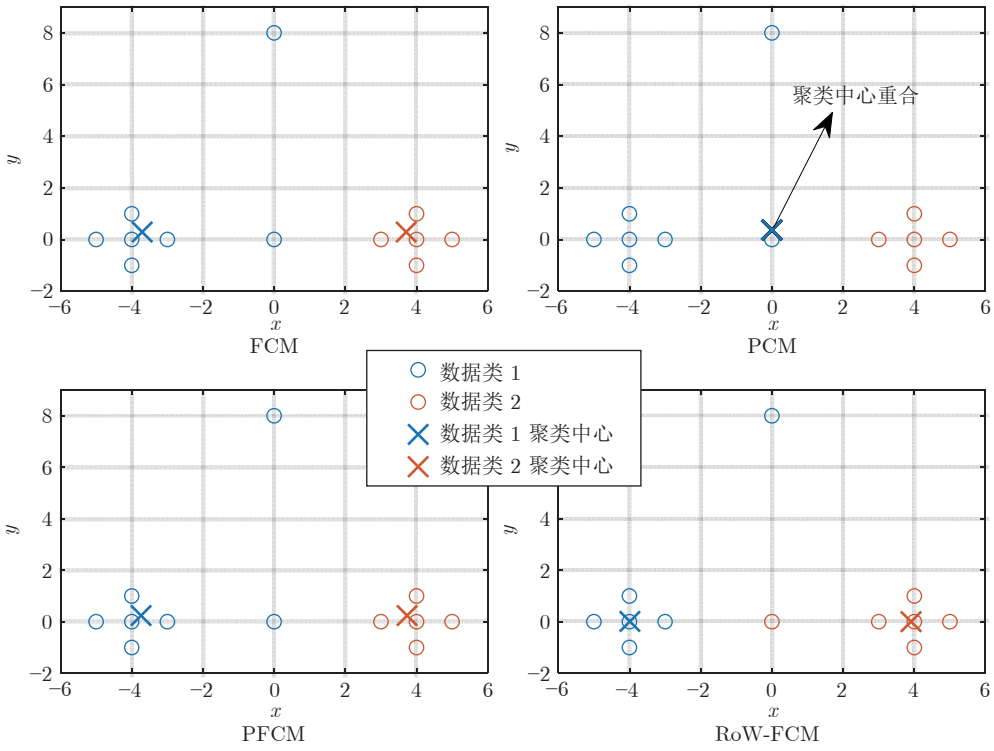


图 6　离群点实验聚类效果图

Fig. 6　Experimental clustering effect of outlier points

表 2　影响污水处理过程出水水质的主要过程变量

Table 2　The main process variables that affect the effluent quality of the sewage treatment process

编号	符号	变量物理含义	编号	符号	变量物理含义
1	Q_{in}	进水流量	15	$S_{S,3}$	反应池 3易生物降解有机底物量
2	$S_{NH,in}$	进水氨浓度	16	$S_{ALK,3}$	反应池 3 池碱度
3	$X_{BH,1}$	反应池 1活性异养菌生物量	17	$X_{BH,4}$	反应池 4 活性异养菌生物量
4	$S_{NO,1}$	反应池 1 硝氮浓度	18	$X_{BA,4}$	反应池 4 活性自养菌生物量
5	$S_{S,1}$	反应池 1 易生物降解有机底物量	19	$S_{O,4}$	反应池 4 溶解氧浓度
6	$S_{ALK,1}$	反应池 1 池碱度	20	$S_{NH,4}$	反应池 4 氨氮浓度
7	$X_{BH,2}$	反应池 2 活性异养菌生物量	21	$S_{S,4}$	反应池 4 易生物降解有机底物量
8	$S_{NO,2}$	反应池 2 硝氮浓度	22	$S_{ALK,4}$	反应池 4 池碱度
9	$S_{S,2}$	反应池 2 易生物降解有机底物量	23	$X_{BH,5}$	反应池 5 活性异养菌生物量
10	$S_{ALK,2}$	反应池 2 池碱度	24	$X_{BA,5}$	反应池 5 活性自养菌生物量
11	$X_{BH,3}$	反应池 3 活性异养菌生物量	25	$S_{O,5}$	反应池 5 溶解氧浓度
12	$X_{BA,3}$	反应池 3 活性自养菌生物量	26	$S_{NH,5}$	反应池 5 氨氮浓度
13	$S_{O,3}$	反应池 3 溶解氧浓度	27	$S_{S,5}$	反应池 5 易生物降解有机底物量
14	$S_{NH,3}$	反应池 3 氨氮浓度	28	$S_{ALK,5}$	反应池 5 池碱度

用 FCM、PCM、PFCM 以及本文 RoW-FCM 四种算法在测试集上进行对比分析, 并将所有方法均仿真 30 次的平均结果作为最终结果, 如表 3 所示, 相关结果如图 7 ~ 10 所示. 可以看出, PCM 算法由于产生重合聚类, 其聚类效果差, 结合如图 8 的 PCM 隶属度矩阵值, 可知 PCM 算法不能监测到两种异常工况. 同时, 可以看到 FCM、PFCM 和本文 RoW-FCM 三种方法均能够监测到异常工况. 但是常规 FCM 和 PFCM 两种算法的隶属度矩阵值波动较大, 尤其在 0.5 附近区分度不明显, 导致聚类错误率

表 3 不同算法的聚类准确度与迭代次数

Table 3 Clustering accuracies and numbers of iterations of different algorithms

工况类型	聚类正确率 (%)				聚类收敛迭代次数 (收敛精度 10^{-5}, 30次仿真)			
	FCM	PCM	PFCM	RoW-FCM	FCM	PCM	PFCM	RoW-FCM
正常工况	92.3	80.8	93.9	97.5	—	—	—	—
异常工况 1	75.0	6.3	76.3	96.0	45.1	14	29.1	23.6
异常工况 2	80.3	3.5	77.5	97.0	—	—	—	—

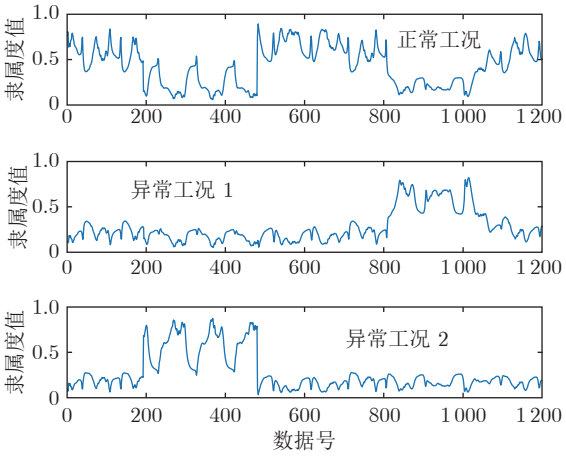


图 7 FCM 隶属度矩阵

Fig. 7 FCM membership matrix

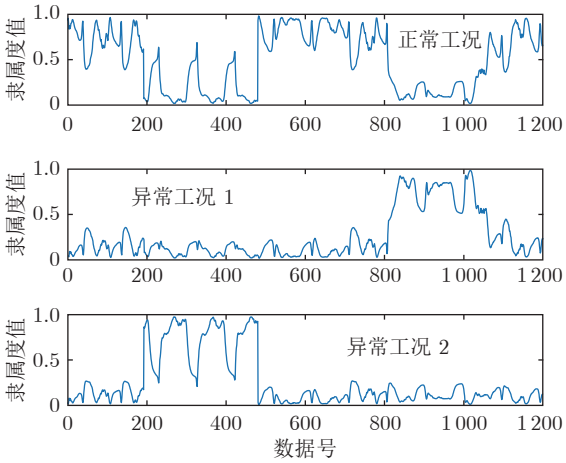


图 9 PFCM 隶属度矩阵

Fig. 9 PFCM membership matrix

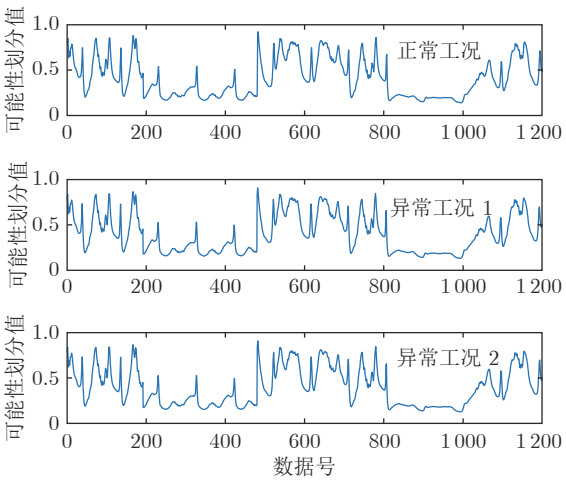


图 8 PCM 可能性矩阵

Fig. 8 PCM possibility matrix

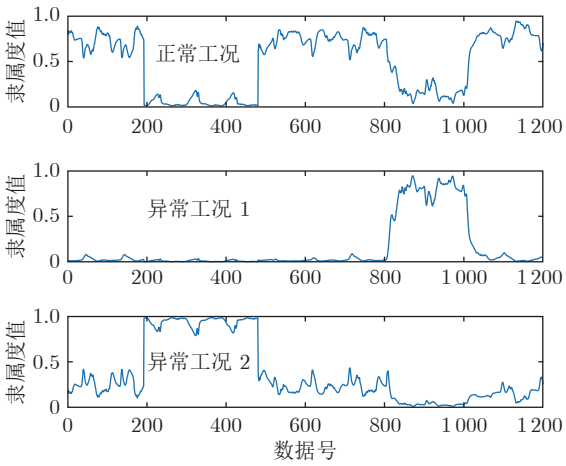


图 10 RoW-FCM 隶属度矩阵

Fig. 10 RoW-FCM membership matrix

升高. 而本文 RoW-FCM 算法的隶属度矩阵值平稳, 在 0.5 处区分度大, 能够将 2 类异常工况很好地进行聚类, 能够分别在 200、800 时刻附近监测到异常工况的发生. 从仿真的平均迭代次数来看, PCM 算法迭代次数最少, 但该算法由于产生重合聚类, 故不在考虑范围内. 另外, 在其余 3 种方法中, 本文算法具有最少的迭代次数. 综上, 本文 RoW-FCM 算法对 2 种异常工况的监测准确率最高, 迭代次数也最少, 所以 RoW-FCM 算法在实际污水处理过程监测中具有良好的异常工况检测性能.

4.2 污水处理过程异常工况识别结果

进一步采用第 3.3 节异常工况识别方法进行识别, 识别结果如表 4 和图 11 所示, 其中表 4 中的编

表 4　异常工况识别结果表

Table 4　Abnormal condition recognition result table

编号	正常工况	异常工况 1	异常工况 2	编号	正常工况	异常工况 1	异常工况 2
1	0.133	0.339	0.528	15	0.254	0.465	0.281
2	0.150	0.321	0.530	16	0.297	0.255	0.448
3	0.454	0.481	0.065	17	0.450	0.464	0.086
4	0.453	0.395	0.152	18	0.354	0.424	0.223
5	0.093	0.577	0.331	19	0.238	0.260	0.503
6	0.305	0.247	0.448	20	0.124	0.352	0.524
7	0.456	0.477	0.067	21	0.236	0.482	0.283
8	0.010	0.307	0.683	22	0.281	0.245	0.475
9	0.241	0.473	0.286	23	0.446	0.458	0.096
10	0.361	0.290	0.349	24	0.352	0.418	0.230
11	0.453	0.471	0.076	25	0.052	0.310	0.639
12	0.353	0.429	0.218	26	0.118	0.314	0.568
13	0.255	0.167	0.578	27	0.229	0.482	0.289
14	0.208	0.425	0.367	28	0.291	0.259	0.450

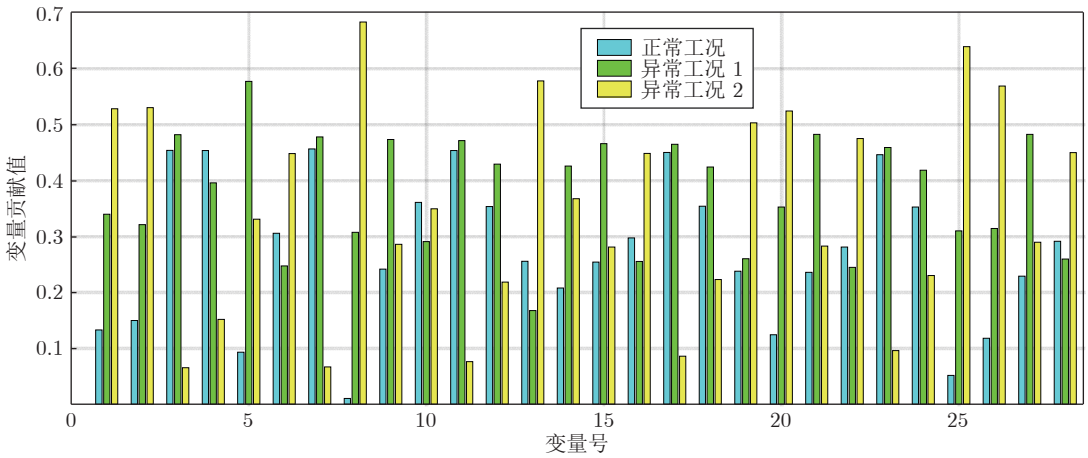


图 11　异常工况识别结果

Fig. 11　Recognition results of abnormal conditions

号与表 2 中的变量编号相对应, 加粗数值表示与异常工况关联变量的贡献值. 可以看出, 与异常工况 1 关联最大变量分别为: 3-$X_{BH,1}$、5-$S_{S,1}$、7-$X_{BH,2}$、9-$S_{S,2}$、11-$X_{BH,3}$、12-$X_{BA,3}$、14-$S_{NH,3}$、15-$S_{S,3}$、17-$X_{BH,4}$、18-$X_{BA,4}$、21-$S_{S,4}$, 23-$X_{BH,5}$、24-$X_{BA,5}$ 和 27-$S_{S,5}$. 这意味着异常工况 1 与活性异养菌生物量、活性自养菌生物量、易生物降解有机底物量有关, 据此可以判断异常工况 1 为毒性冲击, 即毒性冲击导致活性异养菌与活性自养菌数量减少, 从而导致易生物降解有机底物量等过程变量出现异常. 图 11 也显示与异常工况 2 关联最大的变量分别为: 1-Q_{in}、2-$S_{NH,in}$、6-$S_{ALK,1}$、8-$S_{NO,2}$、13-$S_{O,3}$、16-$S_{ALK,3}$、19-$S_{O,4}$、20-$S_{NH,4}$、22-$S_{ALK,4}$、25-$S_{O,5}$、26-$S_{NH,5}$ 和 28-$S_{ALK,5}$. 这意味着进水流量和进水氨浓度与异常工况 2 的发生有关, 这些变量的异常也同时导致反应池中碱度、硝氮浓度、氨氮浓度的变化, 故此判断异常工况 2 为进水量异常. 根据上述分析可知, 本文异常工况识别方法所得到的变量贡献矩阵对异常工况的解释符合实际情况, 能够识别出与异常工况相关的关键变量, 从而验证了本文方法在异常工况识别的有效性和实用性.

注 2. 本文异常工况识别算法思想源于 FCM 算法. 在 FCM 算法中, 每个样本对于所有聚类中心的隶属度之和为 1, 隶属度值最大表明这个样本属于其对应的某个聚类中心. 本文异常工况识别算法的思想是每个变量对各个工况都有一个贡献值, 并

且限定每个变量对所有工况的贡献值之和为 1. 如果某个变量对某个工况的贡献值最大, 即表明此变量是与此工况相关联的变量, 也就认为该变量是造成该工况的关键变量. 也就是说前文提到的 “最大” 指的是某个变量对某个工况的 “最大”贡献值, 即本文异常工况关联最大变量的选取标准是对工况贡献值最大的变量.

5 结束语

针对先验故障知识少的非平稳污水处理过程异常工况监测与识别的难题, 引入并改进了基于模糊 c 均值的聚类方法, 提出了一种基于 RoW-FCM 与 KPLS 的过程监测新方法. 该方法首先建立了质量变量与高维非线性污水处理过程变量的 KPLS 模型, 然后采用本文基于 RoW-FCM 的算法对污水处理过程进行监测. 数值仿真实验表明, 相比于 FCM、PCM、PFCM 算法, 本文 RoW-FCM 聚类算法对离群点具有更好的鲁棒性, 并解决了不平衡簇数据集聚类问题. 此外, 数值实验也表明本文算法采用马氏距离能够适应更多聚类数据结构, 明显优于基于欧氏距离的聚类算法. 基于污水处理过程的异常工况检测与识别数据实验表明, 本文方法在监测过程中准确率更高, 迭代次数少, 能够有效监测到污水处理过程中异常工况的发生, 并能够正确识别出异常工况相关的关键变量, 因此在污水处理过程监测和异常工况识别上具有较好的测试效果和应用前景.

References

1 Meng Xi, Qiao Jun-Fei, Han Hong-Gui. Soft measurement of key effluent parameters in wastewater treatment process using brain-like modular neural networks. *Acta Automatica Sinica*, 2019, **45**(5): 906−919
(蒙西, 乔俊飞, 韩红桂. 基于类脑模块化神经网络的污水处理过程关键出水参数软测量. 自动化学报, 2019, **45**(5): 906−919)

2 Qiao Jun-Fei, Han Gai-Tang, Zhou Hong-Biao. Knowledge-based intelligent optimal control for wastewater biochemical treatment Process. *Acta Automatica Sinica*, 2017, **43**(6): 1038−1046
(乔俊飞, 韩改堂, 周红标. 基于知识的污水生化处理过程智能优化方法. 自动化学报, 2017, **43**(6): 1038−1046)

3 Zhang Shuai, Zhou Ping. Recursive bilinear subspace modeling and model-free adaptive control of wastewater treatment. *Acta Automatica Sinica*, 2022, 48(7): 1747−1759
(张帅, 周平. 污水处理过程递推双线性子空间建模及无模型自适应控制. 自动化学报, 2022, 48(7): 1747−1759)

4 Cheng T, Dairi A, Harrou F, Sun Y, Leiknes T. Monitoring influent conditions of wastewater treatment plants by nonlinear data-based techniques. *IEEE Access*, 2019, **7**: 108827−108837

5 Han H G, Qiao J F. Hierarchical neural network modeling approach to predict sludge volume index of wastewater treatment process. *IEEE Transactions on Control Systems Technology*, 2013, **21**(6): 2423−2431

6 Han Hong-Gui, Wu Xiao-Long, Zhang Lu, Qiao Jun-Fei. Identification and suppression of abnormal conditions in municipal wastewater treatment process. *Acta Automatica Sinica*, 2018, **44**(11): 1971−1984
(韩红桂, 伍小龙, 张璐, 乔俊飞. 城市污水处理过程异常工况识别和抑制研究. 自动化学报, 2018, **44**(11): 1971−1984)

7 Liu H B, Zhang H, Zhang Y C, Zhang F S, Huang M Z. Modeling of wastewater treatment processes using dynamic Bayesian networks based on fuzzy PLS. *IEEE Access*, 2020, **8**: 92129−92140

8 Fuente M J, Vega P. Neural networks applied to fault detection of a biotechnological process. *Engineering Applications of Artificial Intelligence*, 1999, **12**(5): 569−584

9 Fan Xin-Wei, Du Shu-Xin, Wu Tie-Jun. Rough support vector machine and its application to wastewater treatment processes. *Control and Decision*, 2004, (5): 573−576
(范昕炜, 杜树新, 吴铁军. 粗SVM分类方法及其在污水处理过程中的应用. 控制与决策, 2004, (5): 573−576)

10 Liu Yi-Qi, Li Yan, Sun Zong-Hai, Huang Dao-Ping. Research on fault diagnosis of wastewater treatment process based on factor analysis. *Control Engineering of China*, 2015, **22**(3): 447−451
(刘乙奇, 李艳, 孙宗海, 黄道平. 面向污水处理过程因子分析故障诊断方法的研究. 控制工程, 2015, **22**(3): 447−451)

11 Ci Jia-Wei, Luo Jian-Xu. Fault detection in sewage treatment process based on weighted fuzzy clustering algorithm. *Journal of East China University of Science and Technology (Natural Science Edition)*, 2018, **44**(4): 504−510
(慈嘉伟, 罗健旭. 基于加权模糊聚类的污水处理过程故障检测. 华东理工大学学报 (自然科学版), 2018, **44**(4): 504−510)

12 Kang Wei-Xiao. Fault Diagnosis Method for Nonlinear System Based on PFCM Algorithm With Mahalanobis Distance[Master thesis], Harbin Institute of Technology, China, 2016
(康韦晓. 基于马氏距离的PFCM算法的非线性系统故障诊断方法[硕士论文], 哈尔滨工业大学, 中国, 2016)

13 Teppola P, Minkkinen P. Possibilistic and fuzzy c-means clustering for process monitoring in an activated sludge waste-water treatment plant. *Journal of Chemometrics*, 1999, **13**(3−4): 445−459

14 Qin S J. Statistical process monitoring: Basics and beyond. *Journal of Chemometrics*, 2003, **17**(8−9): 480−502

15 Zhou P, Zhang R Y, Xie J, Liu J P, Wang H, Chai T Y. Data-driven monitoring and diagnosing of abnormal furnace conditions in blast furnace ironmaking: An integrated PCA-ICA method. *IEEE Transactions on Industrial Electronics*, 2020, **68**(1): 622−631

16 Dunia R, Qin S J, Edgar T F, McAvoy T J. Identification of faulty sensors using principal component analysis. *AICHE Journal*, 2010, **42**(10): 2797−2812

17 Choi S W, Lee C, Lee J M, Park J H, Lee I B. Fault detection and identification of nonlinear processes based on kernel PCA. *Chemometrics & Intelligent Laboratory Systems*, 2005, **75**(1): 55−67

18 Xu H B, Chen G H, Wang X H. Fault identification of bearings based on bispectrum distribution of ARMA model and FCM method. *Journal of South China University of Technology*, 2012, **40**(7): 78−82, 89

19 Khormali, A O, Shoorehdeli, M A. Gas turbine fault detection and identification by using fuzzy clustering methods. In: Proceedings of the 2014 Second RSI/ISM International Conference on Robotics and Mechatronics. Tehran, Iran: 2014. 70−75

20 Bezdek J C, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 1984, **10**(2): 191−203

21 Krishnapuram R, Keller J M. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, 1993, **1**(2): 98−110

22 Zhang X, Pan W, Wu Z, Chen J, Mao Y, Wu R. Robust image segmentation using fuzzy c-means clustering with spatial inform-

ation based on total generalized variation. *IEEE Access*, 2020, **8**: 95681−95697

23 Krinidis S, Chatzis V. A robust fuzzy local information *c*-means clustering algorithm. *IEEE Transactions on Image Processing*, 2010, **19**(5): 1328−1337

24 Barni M, Capellini V, Mecocci A. Comments on a possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, 1996, **4**(3): 393−396

25 Timm H, Borgelt C, Döring C, Kruse R. An extension to possibilistic fuzzy cluster analysis. *Fuzzy Sets and Systems*, 2004, **147** (1): 3−16

26 Pal N R, Pal K, Keller J M, Bezdek J C. A possibilistic fuzzy *c*-means clustering algorithm. *IEEE Transactions on Fuzzy Systems*, 2005, **13**(4): 517−530

27 Miyamoto S, Ichihashi H, Honda K. *Algorithms for Fuzzy Clustering-methods in c-means Clustering With Applications*. Berlin: Springer-Verlag, 2008.

28 Komazaki Y, Miyamoto S. Variables for controlling cluster sizes on fuzzy *c*-means. In: Proceedings of the Modeling Decisions for Artificial Intelligence. Berlin, Heidelberg: Springer, 2013. 192−203

29 Qiao J F, Zhang W, Han H G. Self-organizing fuzzy control for dissolved oxygen concentration using fuzzy neural network1. *Journal of Intelligent & Fuzzy Systems*, 2016, **30**(6): 3411−3422

30 Garcia-Alvarez D, Fuente M J, Vega P, Sainz G. Fault detection and diagnosis using multivariate statistical techniques in a wastewater treatment plant. *IFAC Proceedings Volumes*, 2009, **42**(11): 952−957

张瑞垚 东北大学硕士研究生. 2018 年获东北大学学士学位. 主要研究方向为数据驱动质量监测.
E-mail: zryao_neu@163.com
(**ZHANG Rui-Yao** Master student at Northeastern University. He received his bachelor degree from Northeastern University in 2018. His main research interest is data-driven quality monitoring.)

周 平 东北大学教授. 分别于 2003、2006、2013 年获东北大学学士、硕士和博士学位. 主要研究方向为工业过程运行反馈控制和数据驱动建模与控制. 本文通信作者.
E-mail: zhouping@mail.neu.edu.cn
(**ZHOU Ping** Professor at Northeastern University. He received his bachelor, master and Ph.D. degrees from Northeastern University in 2003, 2006 and 2013, respectively. His research interest covers operation feedback control of industrial process, data-driven modeling and control. Corresponding author of this paper.)