控制与决策
Control and Decision

基于动态融合LOF的城市污水处理过程数据清洗方法

鲁树武, 伍小龙, 郑江, 何政, 顾剑, 韩红桂

引用本文:
鲁树武,伍小龙,郑江,何政,顾剑,韩红桂. 基于动态融合LOF的城市污水处理过程数据清洗方法[J]. 控制与决策, 2022, 37(5): 1231−1240.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2020.1535

基于动态融合LOF的城市污水处理过程数据清洗方法

鲁树武[1,2], 伍小龙[1,2], 郑 江[3], 何 政[3], 顾 剑[4], 韩红桂[1,2†]

(1. 北京工业大学 信息学部，北京 100124；2. 计算智能与智能系统北京市重点实验室，北京 100124；
3. 北京城市排水集团有限责任公司，北京 100124；4. 北京北排水环境发展有限公司，北京 100122)

摘 要: 围绕城市污水处理过程数据存在连续噪声和缺失的问题,提出一种基于动态融合局部异常因子(dynamic fusion local outlier factor, DFLOF)的污水处理过程数据清洗方法. 首先,设计一种基于滑动窗口的数据动态分段方法,通过计算每个子段数据的均值、最大值和峰值区间信息获得数据异常属性值;其次,建立一种基于DFLOF的数据可信度评价模型,利用基于动态融合局部异常因子算法评估数据的可信度,保证异常数据检测和剔除的准确率;最后,提出一种基于径向基函数(radial basis function, RBF)神经网络的数据补偿方法对缺失数据进行补偿,实现污水处理过程数据的清洗. 将该数据清洗方法应用于实际污水处理过程,实验结果表明:基于动态融合局部异常因子的数据清洗方法能够实现污水处理过程中异常数据的清洗,从而提高数据质量.

关键词: 污水处理过程；数据清洗；动态融合LOF；径向基函数神经网络

中图分类号: TP181 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2020.1535

引用格式: 鲁树武,伍小龙,郑江,等. 基于动态融合LOF的城市污水处理过程数据清洗方法[J]. 控制与决策, 2022, 37(5): 1231-1240.

Data-cleaning method based on dynamic fusion LOF for municipal wastewater treatment process

LU Shu-wu[1,2], *WU Xiao-long*[1,2], *ZHENG Jiang*[3], *HE Zheng*[3], *GU Jian*[4], *HAN Hong-gui*[1,2†]

(1. Faculty of Information Technology，Beijing University of Technology，Beijing 100124，China；2. Beijing Key Laboratory of Computational Intelligence and Intelligent System，Beijing 100124，China；3. Beijing City Drainage Group Company Limited，Beijing 100124，China；4. Beijing Drainage Water Environment Development Company Limited，Beijing 100022，China)

Abstract: In order to reduce the impact of continuous data noise and loss, a dynamic fusion local outlier factor (DFLOF) method is proposed for data-cleaning of the municipal wastewater treatment process (WWTP). First, a data dynamic segmentation method based on sliding window is designed to obtain the abnormal attribute of each segment, including mean value, maximum value and peak interval. Then, a data reliability evaluation model based on the DFLOF is established to evaluate each data segment by using the dynamic fusion local outlier factor algorithm, which improves the accuracy of abnormal data detection and elimination. Finally, a data compensation method based on radial basis function neural network is proposed to compensate the missing data and further realize the data-cleaning of the WWTP. The proposed cleaning method is applied to a real WWTP, the experimental results show that the data-cleaning method based on the dynamic fusion local outlier factor is able to clear abnormal data and improve the data quality.

Keywords: wastewater treatment process；data-cleaning；dynamic fusion local outlier factor (DFLOF)；radial basis function neural network

0 引 言

城市污水处理过程数据是水质检测、过程控制、决策以及异常工况预警等重要依据. 然而,由于数据采集过程受传感器故障、扰动以及污水处理环境波动等因素影响,数据通常会出现缺失、跳变、漂移、噪声等问题. 一旦数据发生异常,将直接导致污水处理

收稿日期: 2020-11-06；**录用日期:** 2021-02-10.

基金项目: 国家自然科学基金重大项目(61890930-5, 61622301)；国家重点研发计划项目(2018YFC1900800-5)；国家自然科学基金创新群体项目(62021003)；北京高校卓越青年科学家项目(BJJWZYJH01201910005020).

责任编委: 阳春华.

†通讯作者. E-mail: rechardhan@bjut.edu.cn.

厂难以维持稳定状态运行,影响污水处理效果[1-4]. 因此,亟待研究有效的数据清洗方法,提高运行数据质量,保障污水处理过程达标运行[5]. 事实上,由于污水处理过程具有复杂性、滞后性等特征,采集的污水数据不仅存在延迟,同时由于厌氧池、缺氧池、曝气池等各操作单元的数据检测环境复杂,引起数据异常因素不明,导致异常数据难以精确识别、剔除和补偿[6-7],从而无法进行有效地清洗. 因此,如何设计一种有效的数据清洗方法,保证数据质量,仍是污水处理过程面临的难题[8-12].

围绕异常数据的识别,常见的方法主要包括:基于统计分布、距离、密度和聚类的方法[13-17]. 这类方法主要通过挖掘数据异常的显著特征,实现异常数据的识别. 例如,基于数理统计的异常数据清洗方法主要通过构建数据的特征分布模型,并计算出不符合正常分布的数据点,进而检测出异常数据. Kallummil等[18]提出了一种基于统计学判别分析的异常数据识别方法,该方法通过模拟实验获取数据的特征分布模型,然后根据采样值与特征模型值之间的距离判别该数据点是否离群,从而实现异常数据的识别. 然而,该方法建立特征分布模型过程中运用的样本数据具有一致性且波动范围小等特征,当实验环境变化导致数据样本发生剧烈波动时,正常数据易被误识别为异常数据. Zhao等[19]提出了一种基于最近邻距离(k-nearest neighbor, k-NN)的离群数据检测方法,克服了数理统计方法中数据分布模型难以适应数据波动的限制,该方法通过计算每个数据点与其邻域数据点的距离作为参照,将偏差大于平均距离的数据点判断为异常数据,从而实现离群点的检测. 但是,由于该方法在判别数据点时均需要反复计算数据点间邻域距离,导致该方法时间复杂度较高. 邓廷权等[20]提出了一种基于中值滤波的数据清洗方法,该方法通过数据快速排序来获得数据中值,并对偏离中值较大范围的异常数据进行滤波,从而实现了异常数据判别和剔除. 实验结果表明,该方法相比于k-NN最近邻距离算法提升了异常数据识别速率. 然而,该方法围绕全局数据设定阈值,导致局部离群数据难以识别和检测. 为了提高局部离群数据的检测精度,一些学者提出了基于密度的离群点检测算法. 例如,马贺贺等[21]提出了一种基于局部离群因子的异常数据检测方法,该方法通过计算局部数据点的离群因子,获得局部数据样本偏离正常值的程度,进而实现离群数据的判别;Jiang等[22]将k均值聚类算法与局部离群因子相结合,设计了一种基于聚类和局部信息的组合算法,改善了局部离群点检测效果;Geng等[23]提出了基于k近邻与局部异常因子聚合的方法,该方法利用k近邻算法获得数据的空间状态,再运用局部异常因子聚合算法获得数据的离群因子大小,提高了局部离群点的识别精度. 然而,以上方法仅仅利用局部样本数据距离或全局数据标准差、中位数等定义异常数据属性,导致在数据波动、突变与异常数据混合情况下,异常数据难以被识别和剔除.

针对数据缺失或异常数据剔除后的补偿问题,Lowe[24]提出了最近邻插值法以补偿缺失数据,该方法通过搜索样本点临近数据作为补偿数据,实现了缺失数据的插值补偿. 实验结果表明,该数据补偿方法具有计算量小、补偿速度快等优势. 然而,城市污水处理过程数据随进水流量的不同波动剧烈,临近数据与缺失数据差异较大,导致数据补偿精度较低. 为此,Pan等[25]提出了一种基于线性插值的方法,该方法通过计算缺失数据两端间的梯度判断待插入的缺失数据,提高了数据补偿的精度. 然而,当污水处理过程数据发生连续性或间歇性多个缺失时,线性插值法难以通过模拟数据规则获取准确数据补偿值. 蓝艇等[26]提出了一种非线性样条插值法进行数据补偿,该方法将缺失数据划分为多个子区域,并利用三次函数来拟合子区缺失值,提高了数据补偿的平滑性. 但是,当非线性补偿函数阶次较高时,无法准确拟合变量数据规律,补偿数据易出现漂移现象. 因此,如何更好地表达数据样本特征,提高补偿结果的精确性成为获取数据补偿值的关键. 近年来,一些学者提出了缺失数据的智能补偿方法,该类方法能够准确模拟数据规律或数据趋势预测,从而获得精确的数据补偿值. 例如,Han等[27]提出了一种基于模糊逻辑的数据补偿方法,利用模糊规则建立了基于局部历史数据与待补偿数据的插值模型,实现了缺失数据的准确补偿;Jiang等[28]提出了一种基于贝叶斯准则的数据补偿方法,该方法利用缺失值自身的分布信息计算补偿数据的最大后验概率,实现缺失值的补偿,实验结果表明,该方法能够预测污水数据的趋势,实现数据的插值补偿. 然而,上述方法在计算优化参数和惩罚项时,时间复杂度较高,难以快速计算最优的补偿数据. Chen等[29]提出了一种基于支持向量回归的数据补偿方法,该方法通过将低维输入变量映射到高维空

间,获得补偿数据的回归模型,实现了缺失数据的快速补偿. 上述方法主要依据变量自身数据的变化规律进行数据补偿,可适用于缺失数据较少的情况. 当数据自身缺失或异常数据量较多时,数据规律无法有效模拟,从而难以确保补偿数据的准确性.

为了解决上述问题,本文提出一种基于动态融合局部异常因子(dynamic fusion local outlier factor, DFLOF)的数据清洗方法. 首先,设计一种基于滑动窗口的方法,对污水处理数据进行动态分段和异常属性提取,获得全面的数据异常属性;其次,建立一种基于DFLOF的数据可信度评价模型,运用基于动态融合局部异常因子算法评估数据的可信度,提高异常数据检测和剔除的精度;最后,提出一种基于RBF神经网络的数据补偿方法,运用关联变量数据对缺失数据进行补偿,最终实现城市污水处理过程数据的清洗. 实际污水处理过程的应用效果表明,基于DFLOF的方法能够实现污水处理过程异常数据的清洗,提高数据的质量.

1　过程数据采集

在城市污水处理过程中,数据采集系统能有效提高数据获取的便捷性,可将大量的数据实时传输到中控室中,经过数据分析获取污水处理运行状态,为整个污水处理过程的管控提供有效的决策信息. 城市污水处理过程数据采集系统主要由计算机网络及相关的检测传感器、仪表等构成. 具体采集过程为: 首先,根据城市污水处理工艺要求选择变量数据采集仪表,并通过相应的数据传感器将检测数据传输到数据检测仪表,如氧化还原电位检测仪表、正磷酸盐分析仪、流量计仪表等;然后,通过污水处理厂局域网系统将实时数据采集到中控室服务器中;最后,在中控室上位机的组态软件中进行实时数据画面显示监测、数据存储以及远程传输,为城市污水处理过程系统建模、过程状态估计和性能分析提供可靠的数据.

根据城市污水处理厂建设需求,常见的厌氧-缺氧-好氧污水处理工艺所选择的数据采集仪表安装位置有: 厌氧池、缺氧池、好氧池和沉淀池. 具体数据采集位置及变量分别为: 进水池前端: 悬浮物浓度(SS)、化学需氧量(COD)、进水总磷(TP)、进水总氮(TN);厌氧池末端: ORP;缺氧池前端: 氧化还原电位(ORP);好氧池: 可溶性固体悬浮物(TSS)、溶解氧(DO);沉淀池前端: ORP,酸碱度(pH)、温度(T);沉淀池末端: pH、出水总磷(TP)、出水总氮(TN)等. 然而,在污水处理过程中,受进水流量的波动及水质成分的影响,进水端数据采集仪表易受污染和损坏;由于污泥浓度升高,温度变化和酸碱液腐蚀等影响,厌氧区和缺氧区传感器探头易受损坏,导致采集的部分污水数据严重偏离真实值. 此外,在数据传输过程中,现场数据传输局域网系统也易受电磁干扰,使得污水数据获取过程中会出现部分缺失的现象,难以为污水处理过程运行状态估计和系统建模提供准确数据. 因此,需要对污水处理过程数据进行异常值检测和清洗.

2　DFLOF数据清洗模型

2.1　动态融合LOF概述

DFLOF数据清洗模型架构如图1所示. 动态融合LOF算法是一种基于动态滑动窗口、异常因子以及神经网络融合的数据清洗算法. 通过所设计的算法精确定位污水处理过程异常数据发生的数据段,以及补偿连续缺失数据,最终实现采集污水变量数据的有效清洗.

滑动窗口采用数据分段的方式将连续异常点进行分离,并提取数据段内的异常属性,为异常数据的识别提供分布信息. 滑动窗口w设定为w_1与w_2之间,根据采集变量个数、污水处理过程周期变化率以及异常因子计算参数k,有

$$w_2 = l/k_{\min}, \tag{1}$$

$$w_1 < w < w_2. \tag{2}$$

其中: l为污水处理变量总个数;$k_{\min}$为异常因子中邻域最小迭代次数;w_2为最大滑动窗口长度;w_1为最小滑动窗口,依据数据变量周期初始化为10;w为数据段的最优窗口长度.

在异常因子计算中,根据异常属性信息进一步分析城市污水处理过程发生异常数据段,识别异常值. 其中: 邻域参数k依据最少邻域数和最近邻距离进行确定;局部异常因子通过计算异常数据段的第k邻域距离、局部可达距离、局部可达密度以及平均密度与对象密度之比来进行计算,实现连续异常数据的识别与剔除.

采集异常污水变量数据被剔除后,基于RBF神经网络进行缺失数据的补偿,将缺失数据相关变量作为输入,依据实验构造法进行网络结构设计,最终实现城市污水处理过程单变量连续异常数据的清洗. 具体数据清洗过程如下.

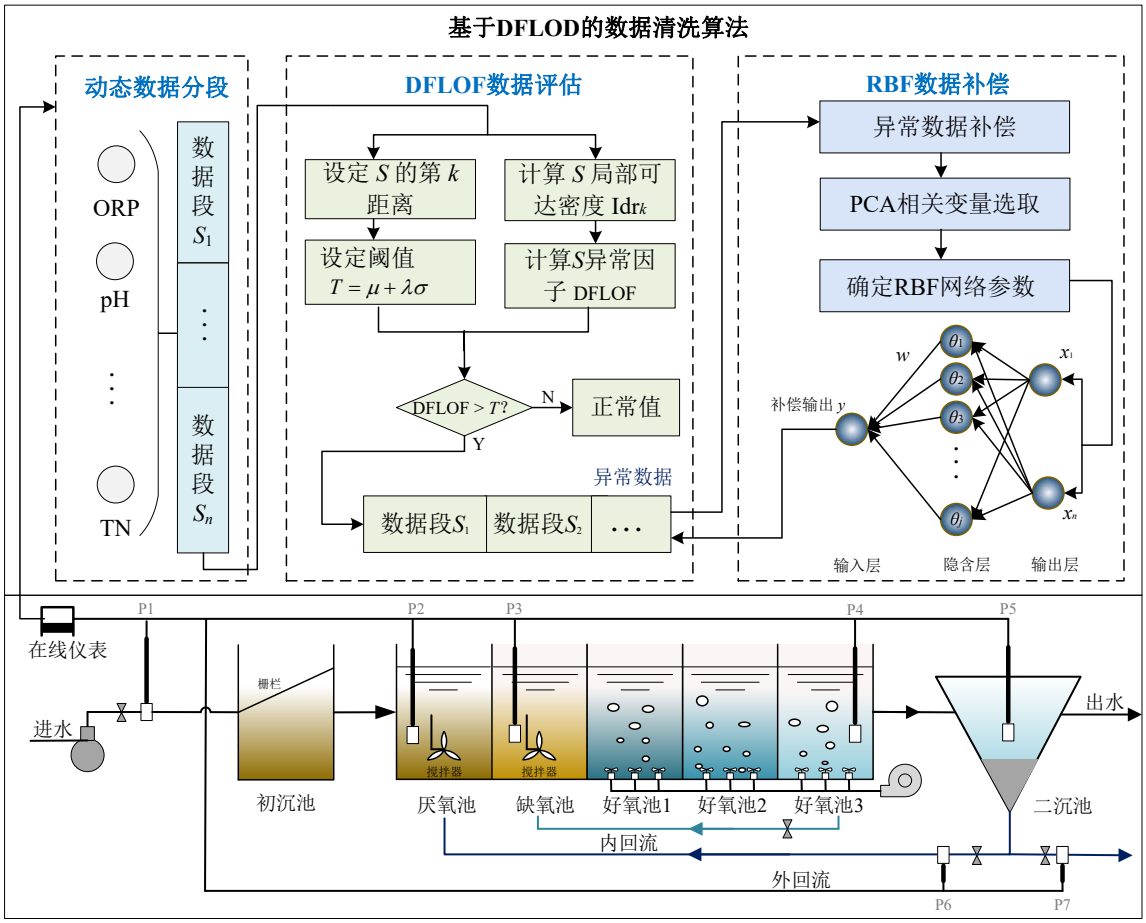

图 1 基于DFLOF的城市污水数据清洗架构

2.2 数据预处理

2.2.1 数据去量纲化

为了减少数据清洗过程中具有不同量纲、数量级的变量造成的影响,采用0均值归一化操作对数据进行压缩处理,计算公式为

$$x_{ij}=\frac{z_{ij}-\bar{z}_j}{b_j}. \tag{3}$$

其中:$i=1,2,\ldots,p$,p为采集的数据样本个数;$j=1,2,\ldots,q$,q为污水处理过程中采集的变量个数;z_{ij}为原始数据集中第j个变量的第i个数据;$\bar{z}_j$为第j个变量数据x_j的均值;b_j为变量数据z_j的标准差;x_{ij}为归一化后的标准数据.

2.2.2 窗口数据分段

采集污水数据集存在连续性缺失现象,并且具有一定的周期性,此时,采用滑动窗口对原始数据进行动态分段处理,可以降低数据周期性带来的影响,提高异常数据的检测精度. 具体污水数据分段步骤如下:

1) 设定污水数据集$D=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_q\}$,每类变量数据的数据向量为$\boldsymbol{X}_j=\{x_{j1},x_{j2},\ldots,x_{jp}\}$,$p$为数据集的数据个数. 滑动窗口大小为$w$个数据,并将滑动窗口放置在数据变量$\boldsymbol{X}_j$的起点$x_{j1}$,形成一个数据长度为$w$的子数据段.

2) 移动滑动窗口,以数据集的第2个点x_{q2}为起点位置获得第2个长度为w的子数据段,依次向后滑动,获得$p-w+1$个长度为w的子段数据$S_1,S_2,\ldots,S_{p-w+1}$,每个数据段记为$S_\tau=\{x_\tau,\ldots,x_{\tau+w-1}\}$,子段数据集设定为$D_{(s)}=\{S_1,S_2,\ldots,S_\tau\},\tau=1,2,\ldots,p-w+1$.

3) 提取子段数据S_τ的属性,包括最大值、最小值、均值和均方差Θ维属性.

此时,污水数据$\boldsymbol{X}_j$为一系列$m\times\Theta$维的数据集合,$m=p-w+1$. 计算每个数据点间的欧氏距离并作为距离函数来判断数据段密度大小.

2.3 基于DFLOF的异常数据识别

由于受污水处理过程噪声以及水质检测环境的影响和数据采集过程中易出现异常数据,会导致数据可靠性下降. 因此,需要对污水处理过程中的异常数据进行检测,并评估可信度,剔除异常数据. 为此,本

文选用基于DFLOF的数据异常识别方法,根据设定的污水数据集,计算每一组数据的第k距离大小;然后计算邻域密度与样本点密度之比;最后依据局部异常因子的大小判断数据段的异常程度. 具体检测过程如下.

1) 计算污水数据段S的第k距离(k-distance),其中,第k距离定义为:假设S和O为数据集$D_{(s)}$中的点,对于任意正整数k,S的第k距离为S与O之间的距离,记为$k<\text{distance}(O)$,数据段S满足如下条件:

① 至少存在k个数据段$O'\in D_{(s)}$,使$d(S,O')\leqslant d(S,O)$成立;

② 至多存在$k-1$个数据段$O'\in D_{(s)}$,使$d(S,O')<d(S,O)$成立.

数据S与数据O的距离为$d(S,O)$,其计算公式为

$$d(S,O)=\sqrt{\sum_{\nu=1}^{\Theta}(f(S_\nu)-f(O_\nu))^2}. \tag{4}$$

其中:Θ为每个数据段S_τ的异常属性个数;$f(S_\nu)$和$f(O_\nu)$是污水数据的第v维属性值,包括均值、最大值、峰值区间和均方差等,$v=1,2,\ldots,\Theta$.

2) 计算数据段S的第k距离邻域$N_{k\text{-distance}}(S)$,数据段S的第k距离邻域定义为:数据集$D_{(s)}$中与S的距离不超过其第k距离k-distance(S)的所有数据段集合,即

$$N_{k\text{-distance}}(S)=\{Z|d(S,Z)\leqslant k\text{-distance}(S)\}, \tag{5}$$

其中$N_{k\text{-distance}}(S)$记为$N_k(S)$.

3) 计算数据段S与其所有邻域数据$N_k(S)$的可达距离. 对象S相对于对象O的可达距离为

$$\begin{aligned}&\text{reach-distance}(S,O)=\\&\max\{k\text{-dinstance}(O),d(S,O)\}.\end{aligned} \tag{6}$$

其中:reach-distance(S,O)表示S与O的可达距离,max{k-distance$((O)$, $d(S,O)$}表示第k距离与直线距离中的最大值,$O\in N_{k\text{-distance}}(S)$.

4) 计算数据段S局部可达密度$\text{Idr}_k(S)$. 局部可达密度$\text{Idr}_k(S)$定义为:S与其所有第k距离邻域数据段的平均可达距离的倒数,计算公式如下:

$$\text{Idr}_k(S)=\frac{1}{\sum\text{reach-distance}(S,O)/N_k(S)}. \tag{7}$$

5) 数据段S的异常因子大小$\text{DFLOF}_k(S)$,可计算为

$$\text{DFLOF}_k(S)=\frac{\sum\limits_{o\in N_k(s)}\text{Idr}_k(O)/\text{Idr}_k(S)}{|N_k(S)|}. \tag{8}$$

数据段S的异常因子$\text{DFLOF}_k(S)$反映了该数据段的异常程度,异常因子值$\text{DFLOF}_k(S)$越大,说明该段污水数据发生异常可能性越大.

为了准确确定异常污水数据段,设定阈值参数T为

$$T=\mu+\lambda\sigma. \tag{9}$$

其中:μ为异常因子均值;σ为标准差;λ为数据异常程度的控制变量,λ值越大表示数据段的异常程度越高.

依据式(8)和(9)评估每个子数据段的可信度,将局部异常因子值DFLOF大于阈值T的数据视为异常值,并进行剔除. 因此,采集污水变量中异常值被剔除后会造成数据空缺,影响污水处理系统建模和控制精度. 结合原始数据中的缺失值,定义缺失数据段为

$$\boldsymbol{X}_\gamma=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_\gamma,\ldots,\boldsymbol{x}_q\}. \tag{10}$$

其中:$\boldsymbol{X}_\gamma$为含有缺失数据的变量数据段;$\boldsymbol{x}_\gamma$为第γ个变量数据集存在缺失,且缺失数据集为

$$\boldsymbol{x}_\gamma=\{x_{1\gamma},\ldots,x_{h\gamma},\underbrace{0,\ldots,0}_{g},x_{(h+g)\gamma},\ \ldots,\ x_{p\gamma}\}. \tag{11}$$

h为缺失数据的起点;g为缺失数据的个数,其大小等于滑动窗口长度w,且$g\leqslant p/2$.

2.4　基于RBF神经网络数据补偿

为了实现缺失数据的补偿,设计一种基于径向基函数(radial basis function, RBF)神经网络的数据补偿模型. RBF神经网络能够以任意精度逼近函数,精确补偿污水处理过程中的缺失值. 为了降低网络的计算复杂度,本文采用主成分分析方法筛选缺失数据的相关变量,具体计算步骤如下.

2.4.1　特征变量选取

1) 重构污水处理过程中采集的数据矩阵$\boldsymbol{X}_\gamma$,重构矩阵定义为

$$\bar{\boldsymbol{X}}_\gamma=\{\bar{\boldsymbol{x}}_1,\ \ldots,\bar{\boldsymbol{x}}_\gamma,\ \ldots,\bar{\boldsymbol{x}}_q\}, \tag{12}$$

$$\overline{\boldsymbol{x}}_\gamma=\{x_{1\gamma},\ \ldots,x_{h\gamma},x_{(h+g)\gamma},\ \ldots,\ x_{p\gamma}\}. \tag{13}$$

其中:$\bar{\boldsymbol{X}}_\gamma$为不包含异常和缺失数据的重构变量矩阵,$\overline{\boldsymbol{x}}_\gamma$为不包含异常和缺失值的数据段.

依据式(1)进行标准化计算,获得不包含异常和缺失值的标准化矩阵$\hat{\boldsymbol{X}}=(x_{ij})_{(p-g)\times q}$,计算与$\hat{\boldsymbol{X}}$对应的协方差矩阵为$\boldsymbol{V}$,$\boldsymbol{V}=(v_{id})_{(p-g)\times q}$,$d=1,2,\ldots,q$;待补偿污水数据与相关变量之间的相关系数为

$$v_{jd}=\frac{1}{p-g}\sum_{i=1}^{p-g}x_{ij}x_{id}. \tag{14}$$

2) 计算特征方程$|\lambda \boldsymbol{I}-\boldsymbol{V}|=0$,求解各特征值$\lambda_\delta(\delta=1,2,\ldots,q)$,并按由大到小的顺序进行排序,其中$\boldsymbol{I}$为与协方差矩阵$\boldsymbol{V}$相对应的单位矩阵.

3) 通过各特征值λ_δ求得相对应的特征向量e_δ $(\delta=1,2,\ldots,q)$.

4) 计算各相关变量和待补偿污水数据的相关性大小η_j,以及累计贡献率大小$G(l)$,即

$$\eta_j=\frac{\lambda_j}{\sum\limits_{s=1}^{q}\lambda_s}\times 100\,\%; \tag{15}$$

$$G(l)=\frac{\sum\limits_{s=1}^{l}\lambda_s}{\sum\limits_{s=1}^{q}\lambda}\times 100\,\%,\ 1\leqslant l\leqslant q. \tag{16}$$

将累计贡献率较大的前l个变量筛选为关键特征变量,作为神经网络的输入数据.

2.4.2 RBF神经网络

RBF神经网络具有3层结构,分别为输入层、隐含层和输出层.

网络的输入层依据筛选的关键污水变量设计为l个神经元,网络输入标准矩阵为

$$\hat{\boldsymbol{X}}=\{\hat{\boldsymbol{x}}_1,\hat{\boldsymbol{x}}_2,\ldots,\hat{\boldsymbol{x}}_l\}. \tag{17}$$

其中:$\hat{\boldsymbol{X}}$为输入矩阵,$\hat{\boldsymbol{x}}_l$为被选中的第l个变量的向量.

隐含层节点个数为J,传递函数设定为径向基函数,本文选用标准高斯函数作为核函数,即

$$\theta_j(x)=\mathrm{e}^{\frac{\|\hat{x}_l-c_j\|}{2\sigma_j^2}}. \tag{18}$$

其中:$\hat{x}_l$是污水处理过程数据补偿模型的第l个输入数据,代表影响数据补偿输出的关键污水变量;c_j为径向基函数中第j个隐含层节点的中心向量;σ为第j个隐含节点的宽度;θ_j为隐含层第j个神经元的输出值.

网络输出层采用线性加权法,该层输出公式为

$$y_i=\sum_{j=1}^{J}w_j\theta_j(x)+b, \tag{19}$$

$$\boldsymbol{x}_g=y_i,\ i=1,2,\ldots,g. \tag{20}$$

其中:$\boldsymbol{x}_g$表示补偿缺失数据段;y_i表示g个缺失样本数据$\boldsymbol{x}_\gamma$中第$\boldsymbol{x}_{(h+i)\gamma}$个补偿值;$w_j$为隐含层到输出层之间的权值;$b$为输出层阈值.

RBF神经网络隐含层节点个数会影响网络的非线性处理能力和网络的复杂程度,为了准确输出补偿数据,并降低网络复杂度,本文依据输入变量个数l,采用实验构造法确定隐含层节点个数为$l+2$.

为了保证网络参数能够在算法优化过程中获得较好的性能,本文运用梯度下降学习算法对径向基函数的中心、宽度和输出层权值进行训练,提高网络性能. 性能函数为误差平方和,即

$$\mathrm{SSE}=\sum_{i=1}^{N}(y_i-\bar{y}_i)^2. \tag{21}$$

其中:SSE为误差平方和,y_i为补偿数据段中第i个数据样本真实输出值,$\bar{y}_i$为第i个样本的期望输出值,N为训练样本的个数.

采用梯度下降法进行参数调优,使性能指标函数减小,直到达到期望误差为止. 高斯函数的中心c、宽度σ和隐含层到输出层连接权值w更新公式如下:

$$c_j(k+1)=c_j(k)-\eta_c\frac{\partial \mathrm{SSE}(k)}{\partial c_j(k)}, \tag{22}$$

$$\frac{\partial \mathrm{SSE}(k)}{\partial \sigma_j(k)}=-\frac{1}{\sigma_j^3(k)}\sum_{i=1}^{\tau}(y_i-\bar{y}_i)\times w_{jn}(\hat{x}_l-c_j(k))^2, \tag{23}$$

$$\sigma_j(k+1)=\sigma_j(k)-\eta_\sigma\frac{\partial \mathrm{SSE}(k)}{\partial \sigma_j(k)}, \tag{24}$$

$$\frac{\partial \mathrm{SSE}(k)}{\partial \sigma_j(k)}=-\frac{1}{\sigma_j^3(k)}\sum_{i=1}^{\tau}(y_i-\bar{y}_i)\times w_{jn}(\hat{x}_l-c_j(k))^2, \tag{25}$$

$$w_{jm}(k+1)=w_{jm}(k)-\eta_w\frac{\partial \mathrm{SSE}(k)}{\partial w_{jm}(k)}, \tag{26}$$

$$\frac{\partial \mathrm{SSE}(k)}{\partial w_{jm}(k)}=-\sum_{i=1}^{\tau}(y_i-\bar{y}_i)\varphi_j(\hat{x}_l), \tag{27}$$

其中η_c、η_σ、η_w分别为中心c、宽度σ和权值w的学习率. 通过梯度下降算法可以获得最优的数据补偿模型参数,从而保证算法的收敛性.

2.5 基于DFLOF-RBF的数据清洗

根据以上DFLOF-RBF算法设计,可以实现污水处理过程中某一变量数据的异常从诊断、剔除到获得补偿,具体步骤如下.

step 1: 获取原始污水数据,数据中存在连续噪声和缺失值.

step 2: 进行数据预处理,将原始数据进行归一化和窗口数据分段,获得数据段S_τ和Θ维异常属性值.

step 3: 按式(4)和(5)计算每一数据段S_τ的第k距离和邻域,获得污水处理过程数据初始可信度评估参数.

step 4: 按式(6)～(8)计算每一数据段S_τ的局部可达距离和可达密度,获得局部异常因子DFLOF.

step 5: 当局部异常因子值DFLOF大于阈值T时,当前数据段标记为异常数据段,并转向step 6;否则,标记为正常值,转step 2继续判断下一数据段.

step 6: 依据PCA主成分分析方法计算异常和缺失数据段的相关主元变量,获得RBF数据补偿模型的标准输入数据.

step 7: 采用梯度下降算法训练RBF神经网络模型,直到使其满足误差为止,获得补偿输出数据.

该算法将滑动窗口引入局部异常因子中,通过滑动窗口提取的异常信息计算每一污水数据的异常因子大小,从而提高了异常数据清洗的精度.

3　仿真实验与分析

3.1　实验设计

本实验所用数据为2020年1月北京市某污水处理厂真实数据,共包含3 000组污水样本数据. 为了验证本文基于DFLOF方法异常数据检测的精确性和算法的执行效率,将该方法与基于K-means聚类的异常数据检测方法、基于密度的异常检测方法进行对比实验,并进行实验结果分析. 其中:数据点测试集设定为1 000组数据,包含3 %异常样本;数据段测试集设定为3 000,包含3 %的异常样本. 使用异常数据点误检率作为检测精度的评价指标,误检率公式为

$$\mathrm{ER}=\frac{e+l}{e+l+t}. \tag{28}$$

其中:ER为异常数据的误检率,e为错误检测污水数据的个数,l为漏检个数,t为正确检测污水数据的个数.

通过异常数据检测实验和结果分析,对异常和缺失的溶解氧DO数据进行数据补偿实验. 同时,为了验证本文基于动态融合LOF方法数据补偿的精确性和有效性,将该方法与基于BP神经网络的数据补偿方法、线性插值方法、最近邻插值方法和中值滤波方法的数据补偿效果进行实验对比.

数据补偿实验共有370组正常数据样本,随机选取190组数据作为模型训练样本,180组数据作为测试样本. 模型输入为PCA主成分分析法筛选得到的DO关键特征变量:缺氧前端ORP,进水SS,进水COD,进水TP;模型输出为DO数据补偿输出. RBF神经网络数据补偿模型结构设计为4-6-1,并通过梯度下降算法对模型中心、权值和宽度进行调优,学习率设为0.02,步长为2 000. 基于BP神经网络的数据补偿模型结构设定为4-6-1,采用Sigmoid函数作为隐含层的传输函数,学习算法为最小二乘法,设定学习率为0.1,训练步数为1 000步.

通过数据补偿实验,分析两种方法的数据补偿精度和数据清洗效果. 然后,再随机选取100组污水样本,分别采用最近邻插值法、线性插值法和中值滤波法对污水数据清洗效果进行实验,设定均方根误差RMSE作为清洗效果评价指标,计算公式为

$$\mathrm{RMSE}=\sqrt{\frac{1}{100}\Big(\sum_{i=1}^{100}\|y_i-\overline{y}_i\|\Big)^2}. \tag{29}$$

其中:RMSE表示均方根误差,y_i为第i组样本中溶解氧浓度DO的实际输出值,$\overline{y}_i$为第i组样本中溶解氧浓度DO的期望输出.

3.2　实验结果与分析

图2、图3分别为基于DFLOF的污水数据点异常检测结果图和异常因子分析图. 其中:第k距离为30,阈值参数为1,窗口为10. 通过实验可以看出,基于DFLOF的方法能根据异常因子的大小评估数据的异常程度,精确地检测出二维污水数据中的异常点和离群点,但对于一维数据的连续异常值则难以实现有效检测.

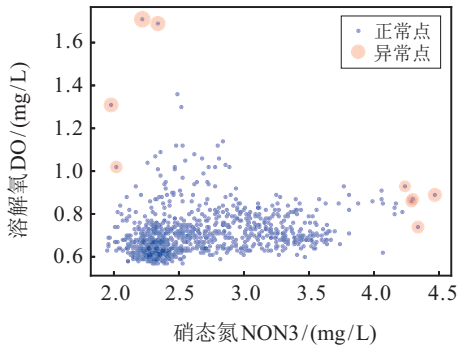

图2　基于DFLOF的数据点异常分析

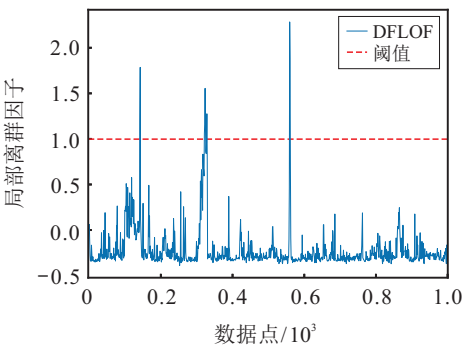

图3　基于DFLOF的数据点异常因子分析

图4、图5分别为污水数据段的异常检测结果图和异常因子分析图. 其中: k 为80, 阈值参数为1, 窗口为10. 可以看出, 污水数据扩展为一维数据段检测时, 能够根据异常属性值评估数据的可信度, 提高了单一污水数据异常检测的精度, 能够满足污水处理过程数据精确性和有效性要求.

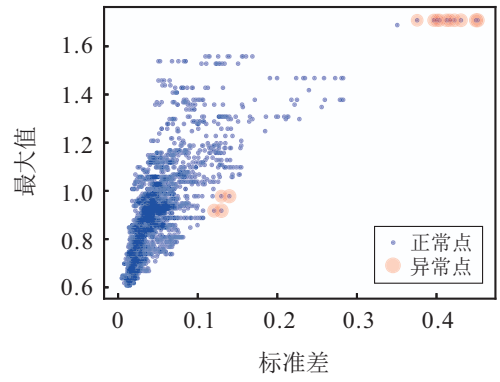

图 4 基于DFLOF的数据段异常分析

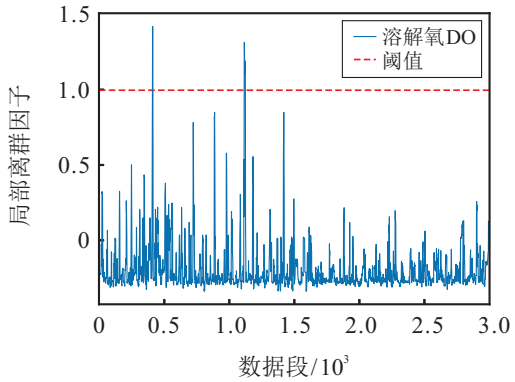

图 5 基于DFLOF的数据段异常因子分析

图6、图7分别给出了基于DFLOF的污水处理过程数据清洗方法和基于BP神经网络的污水数据清洗方法的清洗效果对比和清洗误差对比. 从图6和图7可以看出, 相比于BP神经网络的方法, 基于DFLOF的污水数据清洗方法获得的补偿数据与实际值误差较小, 补偿数据曲线可以更好地拟合目标曲线值. 结果表明, 基于DFLOF的污水数据清洗方法能够依据

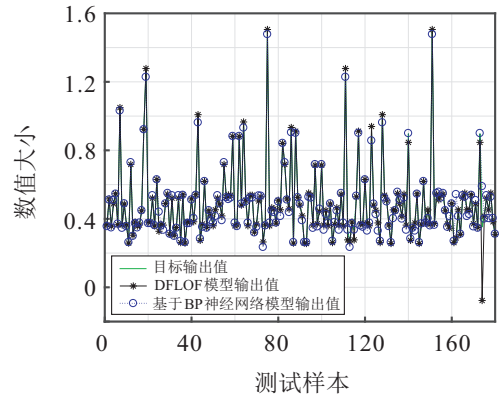

图 6 污水数据清洗结果

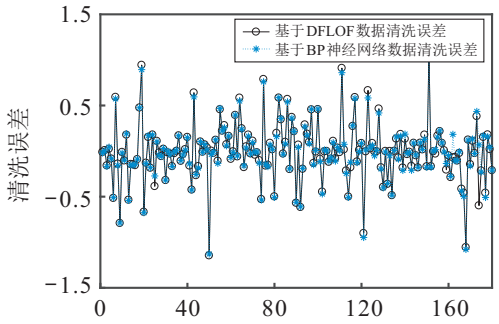

图 7 DO数据清洗误差

污水处理过程中数据间强耦合关系, 通过RBF神经网络径向基函数非线性映射获得待补偿污水数据(如: 溶解氧浓度DO)与采集相关特征变量之间数据补偿模型, 比基于BP神经网络的数据清洗方法具有更好的缺失数据补偿精度, 可以获得更加准确的数据清洗效果.

表1给出了不同异常数据检测方法的结果对比. 从表1可以看出, 基于DFLOF的方法具有更小的误检率和检测时间, 能够满足污水处理过程对数据精确性和可靠性的要求.

表 1 不同方法的异常数据检测结果对比(1)

检测方法	检测时间/s	误检率/%
K-means聚类	3.5643	21.3
DBSCAN聚类	4.4575	14.2
DFLOF法	**3.1244**	**9.3**

表2给出了基于DFLOF的污水处理过程数据清洗方法与BP神经网络的数据清洗方法、最近邻插值方法、线性插值方法和中值滤波法的均方根误差对比. 表2对比结果显示, 在污水数据清洗过程中, 基于动态融合LOF的方法相比于BP神经网络法、最近邻插值法、线性插值法以及中值滤波的方法, 具有更小的清洗误差, 从而表明, 本文基于动态融合LOF的污水处理过程异常数据清洗方法能够实现异常数据的清洗, 满足污水处理控制系统对数据可靠性和实时性的要求.

表 2 不同方法的异常数据检测结果对比(2)

清洗方法	检测时间/s	均方根误差
BP神经网络法	3.4294	1.3106×10^{-3}
最近邻插值法	3.2890	7.9837×10^{-3}
线性插值法	4.3584	7.3868×10^{-3}
中值滤波法	5.7564	8.3729×10^{-3}
DFLOF法	**3.4477**	$\mathbf{0.6531\times10^{-3}}$

通过与不同的数据清洗方法的对比,基于动态融合LOF的数据清洗方法的精确性和实时性得到了验证. 综合以上分析,基于动态融合LOF的污水异常数据清洗方法能够实现对污水噪声数据的剔除和缺失数据的补偿,提高了数据的质量.

4 结　论

针对城市污水处理过程中采集数据存在噪声和缺失的问题,本文提出了一种基于DFLOF的污水处理过程数据清洗方法,并使用污水处理厂真实数据进行了仿真实验,实验结果表明了该方法的有效性. 通过与其他方法进行对比,可以得到如下结论:

1) 采用基于动态融合局部异常因子的方法评估污水处理过程中的数据的准确度,可以实现对污水异常数据的识别和剔除.

2) 与K均值聚类(K-means)、密度分析等方法相比,基于DFLOF的数据清洗方法具有更高的异常数据检测精度,提高了数据质量.

3) 利于RBF神经网络对缺失数据进行补偿,具有更高的数据补偿精度,保证了数据的完整性和准确性,在城市污水处理过程中能够满足数据实际应用的需要.

参考文献(References)

[1] Claros J, Jiménez E, Aguado D, et al. Effect of pH and HNO2 concentration on the activity of ammonia-oxidizing bacteria in a partial nitritation reactor[J]. Water Science and Technology, 2013, 67(11): 2587-2594.

[2] Bai X Z, Wang Z D, Sheng L, et al. Reliable data fusion of hierarchical wireless sensor networks with asynchronous measurement for greenhouse monitoring[J]. IEEE Transactions on Control Systems Technology, 2019, 27(3): 1036-1046.

[3] 乔俊飞, 卢超, 王磊, 等. 城市污水处理过程模型研究综述[J]. 信息与控制, 2018, 47(2): 129-139.
(Qiao J F, Lu C, Wang L, et al. Models of urban wastewater treatment process: An overview[J]. Information and Control, 2018, 47(2): 129-139.)

[4] 韩红桂, 伍小龙, 张璐, 等. 城市污水处理过程异常工况识别和抑制研究[J]. 自动化学报, 2018, 44(11): 1971-1984.
(Han H G, Wu X L, Zhang L, et al. Identification and suppression of abnormal conditions in municipal wastewater treatment process[J]. Acta Automatica Sinica, 2018, 44(11): 1971-1984.)

[5] Zhang J, Sheng V S, Li T, et al. Improving crowdsourced label quality using noise correction[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(5): 1675-1688.

[6] 梁绍一, 韩德强. 基于邻域链的数据异常点检测[J]. 控制与决策, 2019, 34(7): 1433-1440.
(Liang S Y, Han D Q. Outlier detection based on neighborhood chain[J]. Control and Decision, 2019, 34(7): 1433-1440.)

[7] 周颖, 何磊. 具有控制时滞和测量数据丢失的直线电机迭代学习控制[J]. 控制与决策, 2017, 32(8): 1434-1438.
(Zhou Y, He L. Iterative learning control for linear motor system with control delay and measurement dropout[J]. Control and Decision, 2017, 32(8): 1434-1438.)

[8] Kaegi R, Voegelin A, Sinnet B, et al. Behavior of metallic silver nanoparticles in a pilot wastewater treatment plant[J]. Environmental Science & Technology, 2011, 45(9): 3902-3908.

[9] McLellan S L, Huse S M, Mueller-Spitz S R, et al. Diversity and population structure of sewage-derived microorganisms in wastewater treatment plant influent[J]. Environmental Microbiology, 2010, 12(5): 1376.

[10] 韩改堂, 乔俊飞, 韩红桂. 基于递归模糊神经网络的污水处理控制方法[J]. 化工学报, 2016, 67(3): 954-959.
(Han G T, Qiao J F, Han H G. Wastewater treatment control method based on recurrent fuzzy neural network[J]. CIESC Journal, 2016, 67(3): 954-959.)

[11] 韩红桂, 林征来, 乔俊飞. 一种基于混合梯度下降算法的模糊神经网络设计及应用[J]. 控制与决策, 2017, 32(9): 1635-1641.
(Han H G, Lin Z L, Qiao J F. Design and application of hybrid gradient descent-based fuzzy neural network[J]. Control and Decision, 2017, 32(9): 1635-1641.)

[12] 冯宏伟, 姚博, 高原, 等. 基于边界混合采样的非均衡数据处理算法[J]. 控制与决策, 2017, 32(10): 1831-1836.
(Feng H W, Yao B, Gao Y, et al. Imbalanced data processing algorithm based on boundary mixed sampling[J]. Control and Decision, 2017, 32(10): 1831-1836.)

[13] Zhang S C, Li X L, Zong M, et al. Efficient kNN classification with different numbers of nearest neighbors[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(5): 1774-1785.

[14] Angiulli F, Basta S, Lodi S, et al. Distributed strategies for mining outliers in large data sets[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(7): 1520-1532.

[15] Ku W S, Chen H Q, Wang H X, et al. A Bayesian

inference-based framework for RFID data cleansing[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(10): 2177-2191.

[16] Zhao D S, Qiu L P, Song J Y, et al. Efficiencies and mechanisms of chemical cleaning agents for nanofiltration membranes used in produced wastewater desalination[J]. Science of the Total Environment, 2019, 65(2): 256-266.

[17] 王秋萍, 丁成, 王晓峰. 一种基于改进KH与KHM聚类的混合数据聚类算法[J]. 控制与决策, 2020, 35(10): 2449-2458.
(Wang Q P, Ding C, Wang X F. A hybrid data clustering algorithm based on improved krill herd algorithm and KHM clustering[J]. Control and Decision, 2020, 35(10): 2449-2458.)

[18] Kallummil S, Kalyani S. Noise statistics oblivious GARD for robust regression with sparse outliers[J]. IEEE Transactions on Signal Processing, 2019, 67(2): 383-398.

[19] Zhao J, Liu K, Wang W, et al. Adaptive fuzzy clustering based anomaly data detection in energy system of steel industry[J]. Information Sciences, 2014, 259: 335-345.

[20] 邓廷权, 董天祯, 谢巍, 等. 自适应中心加权的彩色图像中值滤波方法[J]. 控制与决策, 2013, 28(9): 1372-1376.
(Deng T Q, Dong T Z, Xie W, et al. Median filtering method based on adaptive central weighting for color images[J]. Control and Decision, 2013, 28(9): 1372-1376.)

[21] 马贺贺, 胡益, 侍洪波. 基于马氏距离局部离群因子方法的复杂化工过程故障检测[J]. 化工学报, 2013, 64(5): 1674-1682.
(Ma H H, Hu Y, Shi H B. Fault detection of complex chemical processes using Mahalanobis distance-based local outlier factor[J]. CIESC Journal, 2013, 64(5): 1674-1682.)

[22] Jiang F, Liu G Z, Du J W, et al. Initialization of K-modes clustering using outlier detection techniques[J]. Information Sciences, 2016, 332: 167-183.

[23] Geng Y L, Li Q Y, Zheng R, et al. RECOME: A new density-based clustering algorithm using relative KNN kernel density[J]. Information Sciences, 2018, 436/437: 13-30.

[24] Lowe D G. Similarity metric learning for a variable-kernel classifier[J]. Neural Computation, 1995, 7(1): 72-85.

[25] Pan F F, Stieglitz M, McKane R B. An algorithm for treating flat areas and depressions in digital elevation models using linear interpolation[J]. Water Resources Research, 2012, 48(6): 229-235.

[26] 蓝艇, 朱莹, 俞海珍, 等. 基于缺失数据的误差生成策略及其在故障检测中的应用[J]. 控制与决策, 2020, 35(2): 396-402.
(Lan T, Zhu Y, Yu H Z, et al. Missing data based method for residual generation and its application for fault detection[J]. Control and Decision, 2020, 35(2): 396-402.)

[27] Han H G, Qiao J F. Nonlinear model-predictive control for industrial processes: An application to wastewater treatment process[J]. IEEE Transactions on Industrial Electronics, 2014, 61(4): 1970-1982.

[28] Jiang Q C, Yan X F, Huang B. Neighborhood variational Bayesian multivariate analysis for distributed process monitoring with missing data[J]. IEEE Transactions on Control Systems Technology, 2019, 27(6): 2330-2339.

[29] Chen Y B, Xu P, Chu Y Y, et al. Short-term electrical load forecasting using the Support Vector Regression (SVR) model to calculate the demand response baseline for office buildings[J]. Applied Energy, 2017, 195: 659-670.

作者简介

鲁树武(1994−), 男, 硕士生, 从事智能信息处理、数据清洗、数据挖掘的研究, E-mail: shuwu_l@163.com;

伍小龙(1988−), 男, 讲师, 博士生, 从事智能特征建模、自组织模糊控制等研究, E-mail: lewis_wxl@sina.com;

郑江(1971−), 男, 教授级高工, 硕士, 从事污水处理、城市给排水等研究, E-mail: bdc@bdc.cn;

何政(1983−), 男, 工程师, 从事污水处理、数据分析、智慧水厂等研究, E-mail: hezheng@bdc.cn;

顾剑(1977−), 男, 高级工程师, 从事污水处理、污水质检等研究, E-mail: guj@ bdc.cn;

韩红桂(1983−), 男, 教授, 博士生导师, 从事神经元网络、过程控制、知识工程与数据挖掘等研究, E-mail: rechardhan@bjut.edu.cn.

(责任编辑：李君玲)