联合样本输出与特征空间的半监督概念漂移检测法及其应用

孙子健[1, 2] 汤 健[1, 2] 乔俊飞[1, 2]

摘 要 城市固废焚烧 (Municipal solid waste incineration, MSWI) 过程受垃圾成分波动、设备磨损与维修、季节交替变化等因素的影响而存在概念漂移现象, 这导致用于污染物排放浓度的建模数据具有时变性. 为此, 需要识别能够表征概念漂移的新样本对污染物测量模型进行更新, 但现有漂移检测方法难以有效应用于建模样本真值获取困难的工业过程. 针对上述问题, 提出一种联合样本输出与特征空间的半监督概念漂移检测方法. 首先, 采用基于主成分分析 (Principal component analysis, PCA) 的无监督机制识别特征空间内的概念漂移样本; 然后, 在样本输出空间采用基于时间差分 (Temporal-difference, TD) 学习的半监督机制对上述概念漂移样本进行伪真值标注后, 再用 Page-Hinkley 检测法确认能够表征概念漂移的样本; 最后, 采用上述步骤获得的新样本结合历史样本对模型进行更新. 基于合成和真实工业过程数据集的仿真结果表明所提方法具有优于已有方法的性能, 能够在加强模型漂移适应性的同时有效缩减样本标注成本.

关键词 城市固废焚烧, 概念漂移检测, 半监督机制, 特征空间, 样本空间

引用格式 孙子健, 汤健, 乔俊飞. 联合样本输出与特征空间的半监督概念漂移检测法及其应用. 自动化学报, 2022, **48**(5): 1259−1272

DOI 10.16383/j.aas.c200984

Semi-supervised Concept Drift Detection Method by Combining Sample Output Space and Feature Space With Its Application

SUN Zi-Jian[1, 2] TANG Jian[1, 2] QIAO Jun-Fei[1, 2]

Abstract The modeling data used for pollutant emission concentration in the municipal solid waste incineration (MSWI) is time-varying due to the concept drift phenomenon, which is caused by factors such as fluctuations in waste composition, equipment wear and repair, and seasonal changes. Thus, it is necessary to identify new samples that can represent the concept drift for pollutant measurement model updating. However, the existing methods are limited by the modeling samples′ true values, which are difficult to be effectively applied to industrial processes. Thus, a semi-supervised concept drift detection method by combining sample output space and feature space is proposed. Firstly, unsupervised mechanism based on principal component analysis (PCA) is used in the sample feature space to identify concept drift samples. Then, semi-supervised mechanism based on temporal-difference (TD) learning is used in the sample output space to label the pseudo-true value for the identified concept drift samples. Further, the Page-Hinkley detection method is used to confirm the concept drift samples. Finally, the new samples obtained by the above steps are combined with historical samples to update the measurement model. The simulation results based on synthetic and real industrial process data sets show that the proposed method has better performance than the existing methods. Moreover, the cost of sample annotation is effectively reduced and the drift adaptability of the measurement model is enhanced.

Key words Municipal solid waste incineration (MSWI), concept drift detection, semi-supervised mechanism, feature space, sample space

Citation Sun Zi-Jian, Tang Jian, Qiao Jun-Fei. Semi-supervised concept drift detection method by combining sample output space and feature space with its application. *Acta Automatica Sinica*, 2022, **48**(5): 1259−1272

城市固废 (Municipal solid waste, MSW) 的全球年增长率随城镇人口增加和居民消费水平提高而不断增加[1], 我国部分城市甚至陷入 “垃圾围城”困境[2]. 该现象直接危害环境卫生和生态平衡, 因此

收稿日期 2020-11-27 录用日期 2021-03-02
Manuscript received November 27, 2020; accepted March 2, 2021
国家自然科学基金 (62073006, 62021003, 61890930-5), 北京市自然科学基金 (4212032, 4192009), 科学技术部国家重点研发计划 (2018YFC1900800-5), 矿冶过程自动控制技术国家 (北京市) 重点实验室 (BGRIMM-KZSKL-2020-02) 资助
Supported by National Natural Science Foundation of China (62073006, 62021003, 61890930-5), Natural Science Foundation of Beijing (4212032, 4192009), National Key Research and Development Program of China (2018YFC1900800-5), and the National (Beijing) Key Laboratory of Automatic Control Technology for Mining and Metallurgical Process (BGRIMM-KZSKL-2020-02)
本文责任编委 魏庆来
Recommended by Associate Editor WEI Qing-Lai
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
1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124 2. Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing 100124

MSW 处理成为亟待解决的全球性问题. 具有无害化、减量化和资源化等特点的 MSW 焚烧 (Municipal solid waste incineration, MSWI) 是世界范围内广泛采用的措施[3], 但该过程的排放尾气中含有氮氧化物、二噁英等难以检测的有害污染物. 目前, MSWI 企业主要通过控制焚烧运行参数实现污染物排放浓度的控制. 显然, 实现 MSWI 过程污染物排放最小化的关键之一是实时、精准地测量这些难测参数的排放浓度[4]. 对此, 软测量模型因具有经济性和快速性等优点而成为当前最为常见的实时测量策略[5]. 但是, 由于工业过程多具有复杂性、随机性和时变性等特征, 这使得基于历史数据构建的软测量模型因不能覆盖新样本所表征的数据分布而导致泛化性能恶化, 导致这一现象的本质原因是概念漂移[6].

概念漂移可表述为数据分布随时间发生变化, 从软测量模型的视角可理解为样本输出空间与特征空间的映射关系发生了改变[7]. 该现象是由难以预知的工业生产环境改变、物料成分波动和设备磨损与维护等因素引起, 并难以避免地导致模型测量精度显著降低[6]. 例如, MSWI 过程中的炉膛温度变化可使烟气污染物生成关系改变, MSW 含水率的差异会导致炉内燃烧状态的变化[3], 这些现象均会引起概念漂移, 进而使得基于历史数据构建的污染物浓度测量模型的精度下降[8]. 因此, 如何采用漂移检测方法有效识别能够表征新概念的漂移样本并将其用于软测量模型的更新, 是提高模型泛化性能需要解决的首要问题[9].

有监督型漂移检测的代表性算法是漂移检测法 (Drift detection method, DDM)[10−11], 其根据新样本测量性能定义警告与漂移等级. 当测量误差超过警告等级时, 存储新样本; 当超过漂移等级时, 采用存储的新样本及历史样本构建新模型以代替旧模型. 类似地, 文献 [12] 计算模型在总体样本和最近样本中获得可接受测量误差的概率, 采用 Hoeffding 不等式判断概率差异后确认是否发生漂移; 文献 [13] 通过比较模型更新前后输出权重值的变化程度表征漂移; 文献 [14−15] 分别采用指数加权移动平均和 Page-Hinkley 检测法确认模型测量精度的变化, 以判断是否发生了概念漂移. 由上可知, 难测参数的测量误差变化能够表征概念漂移对测量模型的直接影响, 该类方法具有计算过程简便高效的优点; 但面向实际工业过程, 上述算法忽视了难测参数真值无法全部获取的实际现状. 例如, 在 MSWI 过程中, 氮氧化物的排放浓度采用人工采样分析方法时其真值获取周期过长, 采用烟气传感器检测时其易受恶劣工况影响而导致测量失真[16]; 二噁英的排放浓度因其采样与化验分析的复杂性导致其真值标注周期长且成本高昂[3]. 因此, 上述有监督型漂移检测方法难以在实际工业过程中直接使用.

无监督型漂移检测的代表性算法有: 文献 [17−19] 基于多元统计策略分别采用近似线性依靠 (Approximate linear dependence, ALD) 条件、主成分分析 (Principal component analysis, PCA) 和角度优化全局降维算法 (Angle optimized global embedding, AOGE) 分析样本特征空间的分布变化; 文献 [20−21] 基于距离度量策略采用马氏距离和领域熵度量特征空间的概念变化; 文献 [22−23] 基于假设检验策略提出基于重采样和累计区域密度的检测方法. 该类算法的特点是在漂移检测阶段不依赖难测参数真值, 但在模型更新阶段仍需采用标注真值的样本, 因此难以在短期内使得模型具有对漂移的适应能力[24].

此外, 复杂工业过程中概念漂移的影响会同时体现为模型测量误差和样本特征空间的综合变化. 因此, 仅基于样本特征空间的分布差异难以有效表征概念漂移现象[10]. 针对上述问题, 面向分类任务, 文献 [25] 提出半监督漂移学习框架, 通过监视分类器置信度变化初步筛选漂移样本, 再根据置信度得分估计漂移样本的伪标签, 最后进行模型更新. 类似地, 文献 [26] 提出基于密度估计的半监督漂移检测, 在少量有标注样本前提下采用增量估计器标注其余样本的标签而实现漂移检测. 但目前为止, 面向复杂工业过程回归建模领域的半监督概念漂移检测方法鲜有报道. 由于分类任务常具有明确且有限的类别标签用于划分样本概念, 其算法设计方式不适用于连续型变量, 因此上述方法难以直接用于回归建模领域[27].

综上, 本文充分考虑 MSWI 过程中的概念漂移现象和难测参数真值无法及时获取的问题, 提出联合样本输出与特征空间的半监督漂移检测方法. 首先, 采用高斯过程回归 (Gaussian process regression, GPR) 依据历史样本构建离线测量模型; 然后, 采用基于 PCA 的无监督机制检测特征空间漂移的样本并将其记录在待标注缓存窗口; 接着, 在样本输出空间中采用基于时间差分 (Temporal-difference, TD) 学习的半监督机制对上述缓存窗口内的样本进行伪真值标注, 并采用 Page-Hinkley 检测法确认能够表征概念漂移的新样本; 最后, 采用新样本与历史样本更新软测量模型.

1 城市固废焚烧 (MSWI) 过程概念漂移问题描述

1.1 城市固废焚烧过程描述

MSWI 过程主要由固废储运、固废焚烧、蒸汽

发电、烟气处理和烟气排放等系统组成, 其工艺流程如图 1 所示.

结合图 1, 针对固废焚烧阶段可描述如下[3].

MSW 由抓斗投放至进料器并送入炉排式焚烧炉. 经干燥炉排预热后, MSW 通过一次风机输送的助燃空气在燃烧炉排中着火燃烧, 在燃烬炉排内燃烧完毕, 产生的烟气经二次风机产生的高度湍流分解后进入烟气管道. 该阶段中, 难测参数氮氧化物的生成原因主要包括[28]: 1) MSW 本身含有的有机和无机含氮化合物在焚烧过程中与氧气发生化学反应; 2) 一次风和二次风中的氮气高温氧化; 3) 助燃燃料 (汽油等) 高温裂解. 因此, 炉膛温度、炉膛含氧量、烟气停留时间与湍流程度等因素改变均会使氮氧化物生成关系变化并产生概念漂移.

传统 MSWI 过程常通过人工化验和烟气自动监控系统 (Continuous emission monitoring system, CEMS) 测定氮氧化物排放浓度. 其中, 人工化验主要包括在线采样和离线化验, 该方式测定周期较长且远滞后于实际过程, 因此无法向测量模型及时提供真值[3]; CEMS 常通过完全抽取或稀释抽取进行测量, 前者在正压环境或抽气量过大时易发生抽气口堵塞, 后者测量响应时间过长且对干燥压缩空气纯度要求高, 此外 CEMS 需要有资质的技术人员定期维护[16]. 上述方式均导致难测参数的真值获取困难. 因此, 需通过标注难测参数的伪真值, 以在无法获取全部真值的情况下分析过程中存在的概念漂移现象.

1.2　概念漂移问题描述

工业过程中通常根据概念漂移的产生原因将其分为过程漂移和传感器漂移[29]. 其中, 过程漂移包括过程内部结构变化 (机械元件磨损等) 和过程外部条件变化 (气候与工艺要求等); 传感器漂移常由传感器等硬件设施的测量精度改变导致, 不反映运行过程的真实参数变化. 本文主要研究 MSWI 过程中常见的概念漂移形式, 即由过程外部条件变化引起的过程漂移.

结合文献 [30] 中定义, 此处对工业过程中概念漂移问题描述如下:

给定 $[1,\ t]$ 时刻采样获得的训练样本集 $\boldsymbol{S}_{1,\,t}^{\text{train}} = \{\boldsymbol{d}_1^{\text{train}}, \cdots, \boldsymbol{d}_t^{\text{train}}\}$, $\boldsymbol{d}_t^{\text{train}} = (\boldsymbol{X}_t^{\text{train}}, y_t^{\text{train}})$ 是 $\boldsymbol{S}_{1,\,t}^{\text{train}}$ 内 t 时刻样本, $\boldsymbol{X}_t^{\text{train}}$ 为其特征空间 (包括炉膛温度、压力和蒸汽流量等可实时测量参数), y_t^{train} 为输出空间真值 (约定真值, 即通过化验分析等方法确定的工业难测参数最高基准值[31]). 假定 $\boldsymbol{S}_{1,\,t}^{\text{train}}$ 内样本均服从分布 $F_{1,\,t}^{\text{train}}(\boldsymbol{X}, y)$、新时刻样本 $\boldsymbol{d}_k^{\text{new}}(k \in [t+1, \infty))$ 服从分布 $F_k^{\text{new}}(\boldsymbol{X}, y)$, 当 $F_{1,\,t}^{\text{train}}(\boldsymbol{X}, y) \neq F_k^{\text{new}}(\boldsymbol{X}, y)$ 时, 认为样本 $\boldsymbol{d}_k^{\text{new}}$ 可表征概念漂移. 此处分布指 $\boldsymbol{X}$ 和 y 之间的联合概率分布.

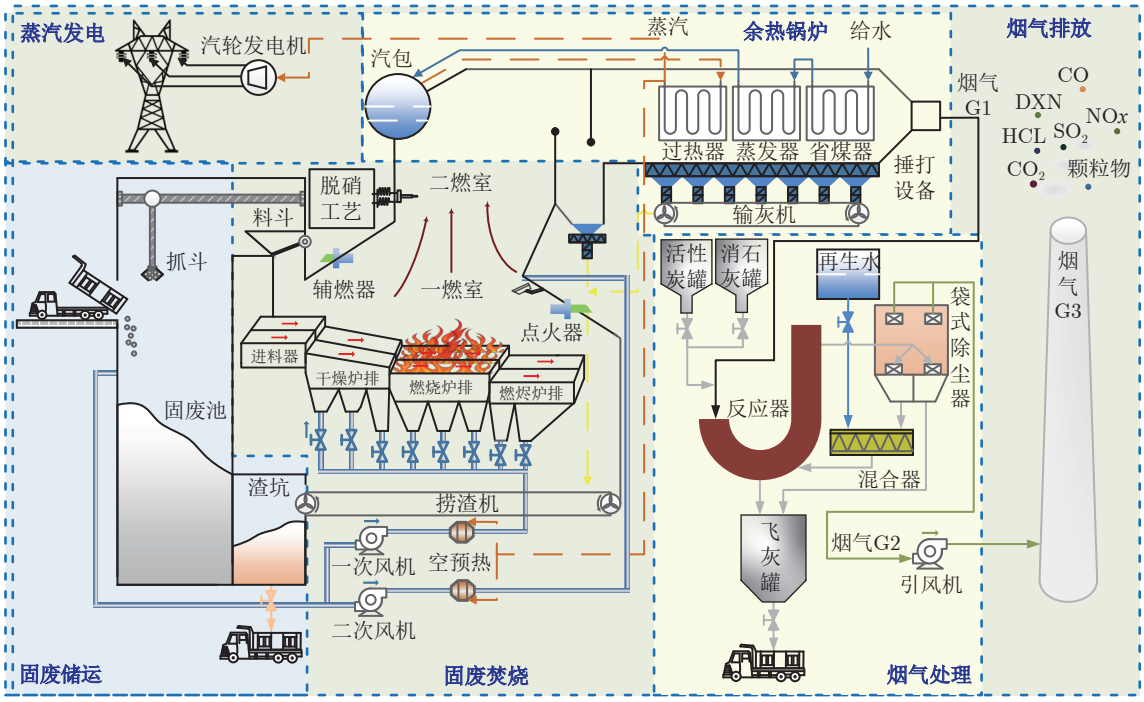


图 1　MSWI 工艺流程图

Fig. 1　The flow chart of MSWI process

根据描述, 常见概念漂移处理方式如图 2 所示.

图 2 中, 虚线框表示该部分内容并非始终可用(样本真值); 分布信息提取指通过测量误差、多元统计或假设检验等方式收集可表征样本分布特性的关键信息; 分布差异检测是针对已提取信息通过预设规则进行相似度量; 依据检测结果, 最终由具体算法判断新样本是否用于更新或舍弃[11].

2 概念漂移检测算法策略

依据上文分析, 本文提出联合样本输出与特征空间的半监督概念漂移检测算法, 其策略如图 3 所示.

图 3 中, $\boldsymbol{X}_{1,t}^{\text{train}}$ 和 $\boldsymbol{y}_{1,t}^{\text{train}}$ 分别表示初始训练集的特征空间与真值集合; $\boldsymbol{X}_{k}^{\text{new}}$ 和 $\hat{y}_{k}^{\text{new}}$ 分别是新样本特征空间与测量值; $\boldsymbol{X}_{1,w}^{n}$ 和 $\dot{\boldsymbol{y}}_{1,w}^{n}$ 分别代表缓存窗口第 n 次填满时, 窗口内样本的特征空间与伪真值集合; $\boldsymbol{X}_{n}^{\text{newtrain}}$ 和 $\dot{\boldsymbol{y}}_{n}^{\text{newtrain}}$ 是新训练集的特征空间与伪真值集合.

图 3 中各模块功能描述如下:

1) 软测量模型构建. 采用历史样本构建基础软测量模型, 并依据新样本的特征空间输出测量值.

2) 特征空间检测. 采用 PCA 对新样本的特征空间进行漂移检测, 当检测值超过 PCA 控制限时认为样本具有漂移可能性, 此时将该样本存入待标注缓存窗口, 当窗口内样本数量达到预设窗口容量时将这些样本送入输出空间检测模块.

3) 输出空间检测. 基于 TD 学习对待标注缓存窗口内样本的伪真值进行标注, 再采用 Page-Hinkley 检测法分析样本的伪真值与模型测量值差异, 以确认样本是否漂移.

4) 测量模型更新. 确认当前缓存窗口内样本发生概念漂移后, 将其结合历史样本共同构造为新训练集重新训练软测量模型, 同时重置待标注缓存窗口.

3 概念漂移检测算法实现

3.1 软测量模型构建模块

本文采用 GPR 构建基础软测量模型. GPR 通

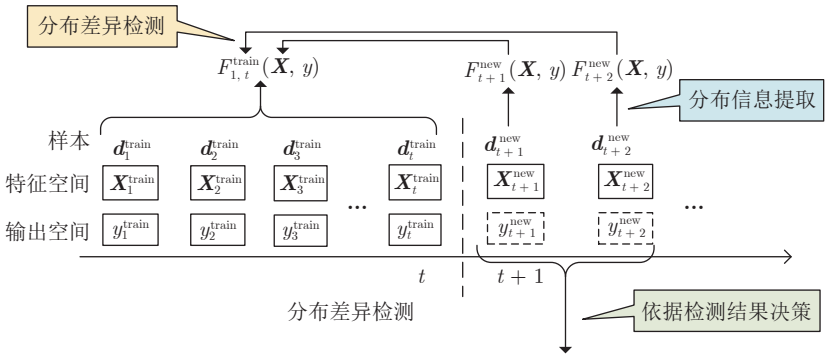


图 2 常见概念漂移处理方式

Fig. 2 The common way to deal with concept drift

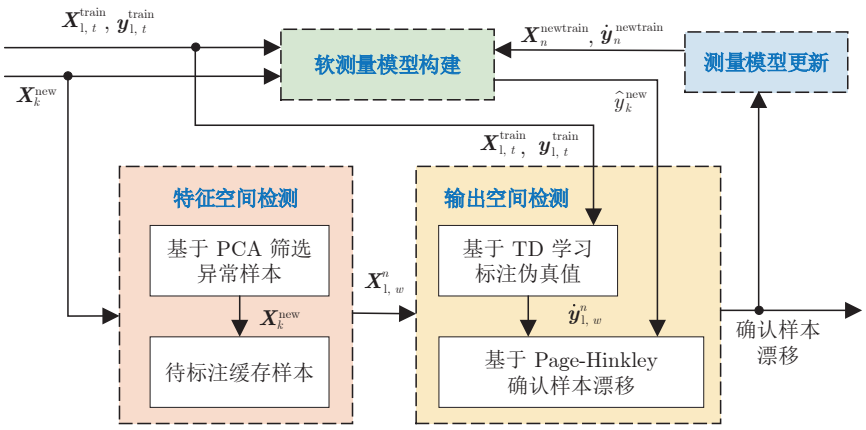


图 3 本文算法策略

Fig. 3 The strategy of the proposed algorithm

过贝叶斯推理确定样本复杂性水平并建立特征空间与输出空间的映射关系, 现已广泛应用于多种工业领域[32].

该过程中, 首先根据训练样本集 $\boldsymbol{S}_{1,t}^{\text{train}}$ 获得样本真值集合 $\boldsymbol{y}_{1,t}^{\text{train}}$ 的先验概率分布:

$$\boldsymbol{y}_{1,t}^{\text{prior}} \sim \text{N}(0,\, \boldsymbol{K}_{1,t}^{\text{cov}}) \tag{1}$$

式中, $\boldsymbol{K}_{1,t}^{\text{cov}}$ 是 $\boldsymbol{S}_{1,t}^{\text{train}}$ 中样本的特征空间 $\boldsymbol{X}_{1,t}^{\text{train}}$ 对应的协方差矩阵, 其计算方式可详见文献 [33].

据此, GPR 对新样本 $\boldsymbol{d}_k^{\text{new}}$ 测量值 $\hat{y}_k^{\text{new}}$ 的估计可表示为:

$$\hat{y}_k^{\text{new}} \sim \text{N}\left(\boldsymbol{\mu}^*,\, \boldsymbol{K}^*\right) \tag{2}$$

$$\begin{cases} \boldsymbol{\mu}^* = \boldsymbol{K}_{(1,\,t)k}^{\text{cov T}} \cdot \boldsymbol{K}_{1,\,t}^{\text{cov}-1} \cdot \boldsymbol{y}_{1,\,t}^{\text{train}} \\ \boldsymbol{K}^* = \boldsymbol{K}_k^{\text{cov}} - \boldsymbol{K}_{(1,\,t)k}^{\text{cov T}} \cdot \boldsymbol{K}_{1,\,t}^{\text{cov}-1} \cdot \boldsymbol{K}_{(1,\,t)k}^{\text{cov}} \end{cases} \tag{3}$$

式中, $\boldsymbol{K}_{(1,\,t)k}^{\text{cov}}$ 和 $\boldsymbol{K}_{(1,\,t)k}^{\text{cov T}}$ 分别表示由 $\boldsymbol{y}_{1,\,t}^{\text{train}}$ 与 $\hat{y}_k^{\text{new}}$ 联合概率分布求得的协方差矩阵及其转置矩阵; $\boldsymbol{K}_{1,\,t}^{\text{cov}-1}$ 表示 $\boldsymbol{X}_{1,\,t}^{\text{train}}$ 对应协方差矩阵的逆矩阵; $\boldsymbol{K}_k^{\text{cov}}$ 是新样本的特征空间 $\boldsymbol{X}_k^{\text{new}}$ 对应的协方差矩阵. 模型测量输出通常取测量值 $\hat{y}_k^{\text{new}}$ 估计范围内的均值.

3.2　特征空间检测模块

本文采用 PCA 对新样本的特征空间进行概念漂移检测. PCA 可有效地从高维特征中提取关键变化信息, 因此被广泛应用于工业过程监控和故障诊断[34]. 采用历史样本的特征空间 $\boldsymbol{X}_{1,\,t}^{\text{train}}$ 建立 PCA 模型后, 对新样本 $\boldsymbol{d}_k^{\text{new}}$ 的检测流程如下[8].

首先, 将 $\boldsymbol{d}_k^{\text{new}}$ 的特征空间 $\boldsymbol{X}_k^{\text{new}}$ 分解为:

$$\begin{cases} \boldsymbol{X}_k^{\text{new}} = \hat{\boldsymbol{X}}_k^{\text{new}} + \tilde{\boldsymbol{X}}_k^{\text{new}} \\ \hat{\boldsymbol{X}}_k^{\text{new}} = \hat{\boldsymbol{X}}_k^{\text{new}} \hat{\boldsymbol{P}}_{1,\,t} \hat{\boldsymbol{P}}_{1,\,t}^{\text{T}} \\ \tilde{\boldsymbol{X}}_k^{\text{new}} = \boldsymbol{X}_k^{\text{new}} \left(\boldsymbol{I} - \hat{\boldsymbol{P}}_{1,\,t} \hat{\boldsymbol{P}}_{1,\,t}^{\text{T}}\right) \end{cases} \tag{4}$$

式中, $\hat{\boldsymbol{X}}_k^{\text{new}}$ 和 $\tilde{\boldsymbol{X}}_k^{\text{new}}$ 分别是 $\boldsymbol{X}_k^{\text{new}}$ 在 PCA 模型主元子空间和残差子空间中的投影; $\hat{\boldsymbol{P}}_{1,\,t}$ 为 $\boldsymbol{X}_{1,\,t}^{\text{train}}$ 对应的载荷矩阵.

然后, 计算 $\boldsymbol{X}_k^{\text{new}}$ 的 PCA 统计量 $\text{SPE}_k^{\text{new}}$ 和 $T_k^{2\text{new}}$:

$$\begin{cases} \text{SPE}_k^{\text{new}} \equiv \left\| \tilde{\boldsymbol{X}}_k^{\text{new}} \right\|^2 = \left\| \boldsymbol{X}_k^{\text{new}} \left(\boldsymbol{I} - \hat{\boldsymbol{P}}_{1,\,t} \hat{\boldsymbol{P}}_{1,\,t}^{\text{T}}\right) \right\|^2 \\ T_k^{2\text{new}} = \boldsymbol{X}_k^{\text{new}} \hat{\boldsymbol{P}}_{1,\,t} \hat{\boldsymbol{\Lambda}}_{1,\,t}^{-1} \hat{\boldsymbol{P}}_{1,\,t}^{\text{T}} \boldsymbol{X}_k^{\text{new}^2} \\ \hat{\boldsymbol{\Lambda}}_{1,\,t} = \dfrac{\hat{\boldsymbol{T}}_{1,\,t}^{\text{T}} \hat{\boldsymbol{T}}_{1,\,t}}{t-1} = \text{diag}\left\{\lambda_1,\, \lambda_2,\, \cdots,\, \lambda_c\right\} \end{cases} \tag{5}$$

式中, $\hat{\boldsymbol{\Lambda}}_{1,\,t}$ 是由 $\boldsymbol{X}_{1,\,t}^{\text{train}}$ 中前 c 个特征值组成的特征向量; $\hat{\boldsymbol{T}}_{1,\,t}$ 是 PCA 模型得分矩阵. 当满足如下条件时:

$$\text{SPE}_k^{\text{new}} > \text{SPE}_\alpha \text{ 或 } T_k^{2\text{new}} > T_\alpha^2 \tag{6}$$

样本 $\boldsymbol{d}_k^{\text{new}}$ 被认为特征空间漂移, 并记录在待标注缓存窗口中. 式中, SPE_α 和 T_α^2 表示 PCA 统计量控制限, 其定义可详见文献 [35].

3.3　输出空间检测模块

3.3.1　基于时间差分 (TD) 学习的伪真值标注

伪真值标注是实现半监督漂移检测的前提. 现有研究中, 文献 [36−37] 证明 TD 学习对特征空间漂移的样本具有良好的测量性能. TD 学习通过分析样本输出与特征空间的一阶差分量变化实现新样本测量[38], 其思路描述如下.

首先, 计算历史样本集 $\boldsymbol{S}_{1,\,t}^{\text{train}}$ 内各样本输出与特征空间的一阶差分量, 以样本 $\boldsymbol{d}_t^{\text{train}}$ 为例计算如下:

$$\Delta \boldsymbol{X}_t^{\text{train}} = \boldsymbol{X}_t^{\text{train}} - \boldsymbol{X}_{t-1}^{\text{train}} \tag{7}$$

$$\Delta y_t^{\text{train}} = y_t^{\text{train}} - y_{t-1}^{\text{train}} \tag{8}$$

然后, 建立关于一阶差分量的回归测量模型:

$$\Delta y = f^{\text{regression}}(\Delta \boldsymbol{X}) \tag{9}$$

新时刻样本的特征空间 $\boldsymbol{X}_{t+1}^{\text{new}}$ 被采集后, 计算其一阶差分量:

$$\Delta \boldsymbol{X}_{t+1}^{\text{new}} = \boldsymbol{X}_{t+1}^{\text{new}} - \boldsymbol{X}_t^{\text{train}} \tag{10}$$

据此计算其输出空间差分量:

$$\Delta \hat{y}_{t+1}^{\text{new}} = f^{\text{regression}}(\Delta \boldsymbol{X}_{t+1}^{\text{new}}) \tag{11}$$

最终, 新样本测量输出可表示为:

$$\hat{y}_{t+1}^{\text{new}} = \Delta \hat{y}_{t+1}^{\text{new}} + y_t^{\text{train}} \tag{12}$$

由于 TD 学习在面对特征空间漂移的样本时具有较好的鲁棒性, 因此本文将其用于标注缓存窗口内样本的伪真值. 经过 PCA 筛选且待标注缓存窗口已被填满后, 记窗口内样本集为 $\boldsymbol{S}_{\text{window}} = \{\boldsymbol{d}_1^{\text{window}}, \cdots, \boldsymbol{d}_w^{\text{window}}\}$, 其中 w 为预设缓存窗口样本容量.

具体标注策略为: 根据式 (7) 、式 (8), 计算历史样本输出与特征空间的一阶差分量集合分别为 $\Delta \boldsymbol{y}^{\text{train}}$ 和 $\Delta \boldsymbol{X}^{\text{train}}$, 并请求现场人员标注窗口内第一个样本 $\boldsymbol{d}_1^{\text{window}}$ 的真值. 原因是: 1) 实际工业过程存在成本高昂、检测延迟和维护困难等问题, 导致难以对全部样本进行真值标注; 2) 新样本发生概念漂移时, 其输入输出关系相较历史样本有较大改变, 此时仅依据历史样本难以推断漂移样本的伪真值. 综上, 仅标注窗口内第一个样本的真值, 可在缩减标注成本的同时提高后续伪真值标注工作的准确

性. 据此, 构建新一阶差分量集合为:

$$\Delta \boldsymbol{y}^{\text{train}'} = \left\{ \begin{array}{c} \Delta \boldsymbol{y}^{\text{train}} \\ \Delta y_1^{\text{window}} \end{array} \right\} \tag{13}$$

$$\Delta \boldsymbol{X}^{\text{train}'} = \left\{ \begin{array}{c} \Delta \boldsymbol{X}^{\text{train}} \\ \Delta \boldsymbol{X}_1^{\text{window}} \end{array} \right\} \tag{14}$$

式中, y_1^{window} 和 $\boldsymbol{X}_1^{\text{window}}$ 分别为 $\boldsymbol{d}_1^{\text{window}}$ 的真值与特征空间; $\Delta y_1^{\text{window}}$ 和 $\Delta \boldsymbol{X}_1^{\text{window}}$ 为 $\boldsymbol{d}_1^{\text{window}}$ 与当前训练集中最后时刻样本共同计算获得的一阶差分量.

从 $\boldsymbol{d}_2^{\text{window}}$ 起, 计算其与 $\boldsymbol{d}_1^{\text{window}}$ 的特征空间 $\boldsymbol{X}_1^{\text{window}}$ 的一阶差分量为:

$$\Delta \boldsymbol{X}_2^{\text{window}} = \boldsymbol{X}_2^{\text{window}} - \boldsymbol{X}_1^{\text{window}} \tag{15}$$

基于最近邻思想, 通过欧氏距离从 $\Delta \boldsymbol{X}^{\text{train}'}$ 中选取与 $\Delta \boldsymbol{X}_2^{\text{window}}$ 距离最小的 ε 个特征空间差分量, 并结合其对应的输出空间差分量共同记为:

$$\begin{aligned} \boldsymbol{\Omega}_{\text{nearest}} = \Big\{ & (\Delta \boldsymbol{X}_{\text{nearest\_1}}^{\text{train}'}, \Delta y_{\text{nearest\_1}}^{\text{train}'}), \cdots, \\ & (\Delta \boldsymbol{X}_{\text{nearest\_}\varepsilon}^{\text{train}'}, \Delta y_{\text{nearest\_}\varepsilon}^{\text{train}'}) \Big\} \end{aligned} \tag{16}$$

采用 $\boldsymbol{\Omega}_{\text{nearest}}$ 建立新的 GPR 软测量模型, 对 $\boldsymbol{d}_2^{\text{window}}$ 的输出空间差分量 $\Delta \hat{y}_2^{\text{window}}$ 进行测量:

$$\boldsymbol{\Omega}_{\text{nearest}} \Rightarrow \text{GPR}_{\text{nearest}} \tag{17}$$

$$\text{GPR}_{\text{nearest}}(\Delta \boldsymbol{X}_2^{\text{window}}) \Rightarrow \Delta \hat{y}_2^{\text{window}} \tag{18}$$

进而将 $\boldsymbol{d}_2^{\text{window}}$ 的伪真值 $\dot{y}_2^{\text{window}}$ 标注为:

$$\dot{y}_2^{\text{window}} = \Delta \hat{y}_2^{\text{window}} + y_1^{\text{window}} \tag{19}$$

此时 $\boldsymbol{d}_2^{\text{window}}$ 可表示为 $\boldsymbol{d}_2^{\text{window}} = (\boldsymbol{X}_2^{\text{window}}, \dot{y}_2^{\text{window}})$. 重复上述过程至窗口内样本均完成伪真值标注.

3.3.2 基于 Page-Hinkley 检测法的漂移样本确认

合理分析样本伪真值和测量值间的差异, 是确认样本最终概念漂移情况的关键. 现有研究表明, 基于累积和思想推导的 Page-Hinkley 检测法具有对分布漂移敏感、计算简便等特点, 因此可有效用于输出空间漂移检测[24]. 该方法中, 给定一系列观测值 $[l_1, l_2, \cdots, l_m]$, 计算备择假设 (观测值中存在漂移点 θ, 即 $1 < \theta < m$) 对原假设 (观测值中不存在漂移, 即 $\theta > m$) 的似然比统计量为[39]:

$$L_{m,\theta} = \frac{\prod\limits_{i=1}^{\theta} f_D(l_i) \prod\limits_{i=\theta+1}^{m} f_D(l_i - \delta)}{\prod\limits_{i=1}^{m} f_D(l_i)} \tag{20}$$

$\prod_{i=\theta+1}^{m} f_D(l_i) = 1$, $\sum_{i=m+1}^{m} l_i = 0$; $f_D(\cdot)$ 表示标准正态分布 N(0, 1) 的分布密度函数; δ 表示漂移样本服从数学期望为 δ 的正态分布.

式 (20) 以对数表示为:

$$Z_{m,\theta} = \ln L_{m,\theta} = \delta \sum_{i=\theta+1}^{m} \left(l_i - \frac{\delta}{2} \right) \tag{21}$$

据此, 备择假设 (有漂移) 对原假设 (无漂移) 的对数似然比统计量为:

$$Z_m = \max_{1 \le \theta < m} Z_{m,\theta} = \max \left\{ \delta \sum_{i=\theta+1}^{m} \left(l_i - \frac{\delta}{2} \right) \right\} \tag{22}$$

通过设置阈值与 Z_m 进行比较, 即可判断当前系列观测值内是否存在概念漂移.

当待标注缓存窗口内样本均完成伪真值标注后, 本文采用 Page-Hinkley 检测法对这些样本的输出空间进行概念漂移检测. 以 T 时刻的观测值 $Obs(T)$ 为例, 检测流程如下[24].

首先, 计算关于 $Obs(T)$ 的累计变量 φ_T:

$$\overline{Obs}_{T-1} = \frac{1}{T-1} \sum_{m=1}^{T-1} Obs(m) \tag{23}$$

$$\varphi_T = \sum_{m=1}^{T} \left(Obs(m) - \overline{Obs}_{m-1} \right) \tag{24}$$

其中, $\overline{Obs}_{T-1}$ 表示此前 T–1 时刻所有历史观测值的均值; 累计变量 φ_T 表示当前观测值 $Obs(T)$ 与历史观测值均值之差.

然后, 通过计算变化指标 PH_T 判断当前观测值 $Obs(T)$ 是否异常:

$$\phi_T = \min_{m=1,\cdots,T} \varphi_m \tag{25}$$

$$PH_T = \varphi_T - \phi_T \tag{26}$$

式中, ϕ_T 表示当前所有时刻中记录的最小累计变量值; PH_T 表示当前 T 时刻累计变量 φ_T 与最小累计变量值之差. 当满足条件 $PH_T > \lambda$ 时, 认为观测值 $Obs(T)$ 异常, 其中 λ 是经验阈值.

记待标注缓存窗口第 n 次填满且样本均被标注时, 窗口内样本集为 $\boldsymbol{S}_{\text{window}}^n = \{\boldsymbol{d}_1^n, \cdots, \boldsymbol{d}_w^n\}$. 计算当前窗口内样本平均测量误差 AveEro_n 如下:

$$\text{AveEro}_n = \frac{1}{w} \cdot \sum_{m=1}^{w} |\hat{y}_m^n - \dot{y}_m^n| \tag{27}$$

式中, $\hat{y}_m^n$ 和 $\dot{y}_m^n$ 分别表示窗口内第 m 个样本的测量值与伪真值.

在此基础上, 本文将观测值 $Obs(T)$ 选取为窗口第 n 次填满时窗口内样本的累积平均测量误差, 即:

$$Obs(T)|_{T=n}=\frac{Obs(T-1)\cdot(n-1)+\text{AveEro}_n}{n},\ n\geq 1 \tag{28}$$

此时, 累计变量 φ_T 表示当前累计平均测量误差与历史累计平均测量误差均值之差; ϕ_T 表示当前记录的最小 φ_T 值.

此外, 根据式 (26), 缓存窗口第一次被填满即 $n=1$ 时, $\phi_T=\varphi_T$, 此时样本输出空间中缺乏漂移判断依据, 因此本文将 ϕ_T 表示为:

$$\phi_T|_{T=n}=\begin{cases}\min\limits_{m=1,\cdots,T}\varphi_m, & n\geq 1\\ \phi_0, & n=0\end{cases} \tag{29}$$

式中, ϕ_0 为基准累计平均测量误差, 将依据验证样本平均测量误差获得. 同时, 本文设置 $\lambda=0$, 即当 $\varphi_T>\phi_T$, 代表当次窗口内累计平均测量误差相较历史样本明显升高时, 认为窗口内样本可表征概念漂移, 并将其用于构建新训练集.

3.4　测量模型更新模块

当缓存窗口内样本被确认漂移后, 本文根据历史样本和当前窗口内样本共同构建新训练集对测量模型进行更新. 以缓存窗口被第 n 次填满时窗口内样本 $\boldsymbol{S}^n_{\text{window}}$ 为例, 构造新训练集 $\boldsymbol{S}^{\text{newtrain}}_n$ 如下:

$$\boldsymbol{X}^{\text{newtrain}}_n=\begin{Bmatrix}\boldsymbol{X}^{\text{train}}_{1,\,t}\\ \boldsymbol{X}^{n}_{1,\,w}\end{Bmatrix} \tag{30}$$

$$\boldsymbol{y}^{\text{newtrain}}_n=\begin{Bmatrix}\boldsymbol{y}^{\text{train}}_{1,\,t}\\ \dot{\boldsymbol{y}}^{n}_{1,\,w}\end{Bmatrix} \tag{31}$$

式中, $\boldsymbol{X}^{\text{newtrain}}_n$ 和 $\boldsymbol{y}^{\text{newtrain}}_n$ 分别是新训练集 $\boldsymbol{S}^{\text{newtrain}}_n$ 的特征空间与真值集合; $\boldsymbol{X}^{n}_{1,\,w}$ 和 $\dot{\boldsymbol{y}}^{n}_{1,\,w}$ 分别是当前窗口内样本的特征空间与伪真值集合.

4　仿真分析

4.1　数据集

本文采用合成数据集验证所提方法的有效性, 并通过真实 MSWI 过程数据集验证其实际应用效果.

1) 合成数据集

合成数据集采用文献 [40] 所提方法构建. 正常样本生成依据为:

$$y=10\cdot\sin(\pi\cdot x_1\cdot x_2)+20\cdot(x_3-0.5)^2+10\cdot x_4+5\cdot x_5+\sigma(0,\,1) \tag{32}$$

式中, x_1、x_2、x_3、x_4 和 x_5 均服从 $[0,\,1]$ 区间内均匀分布, $\sigma(0,\,1)$ 是服从正态分布的随机数.

漂移样本生成依据为:

$$y_R=10\cdot x_1\cdot x_2+20\cdot(x_3-0.5)+10\cdot x_4+5\cdot x_5+\sigma(0,\,1) \tag{33}$$

式中, 各特征取值范围满足:

$$(0\leq x_1\leq 1)\cap(x_2<0.3)\cap(x_3<0.3)\cap(x_4>0.7)\cap(x_5<0.3)$$

合成数据集共有样本 1 500 个, 其中前 1 000 个为正常样本, 后 500 个为漂移样本. 在正常样本中, 又划分前 500 个为建模样本, 后 500 个为验证样本. 验证样本设置目的是获得式 (29) 中基准累计平均测量误差 ϕ_0 值.

2) MSWI 过程数据集

MSWI 过程数据来自北京市某 MSWI 发电厂, 数据中包含的缺失值和异常值均根据现场经验以人工方式去除. 实验中选择氮氧化物的排放浓度作为测量目标, 考虑其生成和吸收过程, 选取炉膛温度、一次风量、二次风量、炉膛剩余氧量、尿素喷入量等相关性较强的 18 个变量作为样本特征. 过程数据集中具有样本 1 500 个, 其中前 1 000 个为正常样本, 后 500 个为漂移样本. 在正常样本中, 又划分前 500 个为建模样本, 后 500 个为验证样本. 其中, 正常样本在炉膛温度为 900 ℃ ~ 950 ℃ 时的对应工况中采集; 漂移样本在炉膛温度为 950 ℃ ~ 1000 ℃ 时的对应工况中采集.

上述数据集的详细参数及各特征在概念漂移环境中的变化情况, 如表 1 和图 4 所示.

表 1　各数据集参数介绍

Table 1　Detailed introduction of each data set

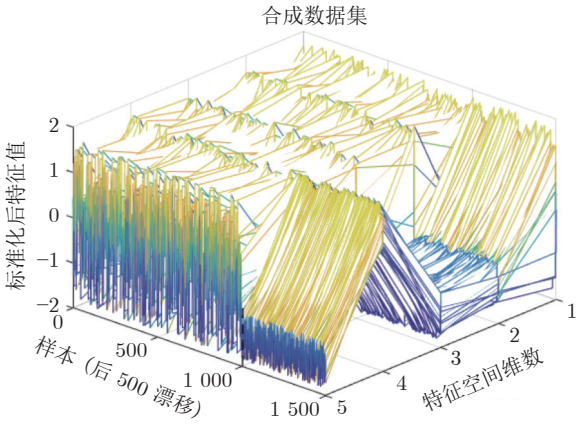
数据集	样本总数	建模样本数	验证样本数	漂移样本数	特征空间维数
合成	1 500	500	500	500	5
过程	1 500	500	500	500	18

由图 4 可知, 两数据集中建模样本与漂移样本间的特征空间分布情况具有明显差异, 间接反映了数据集中存在的概念漂移现象.

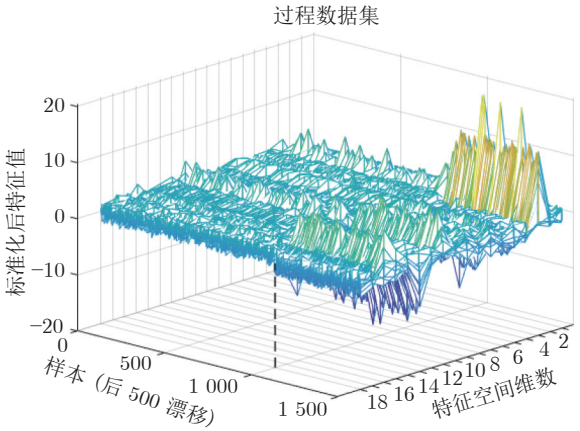
4.2　仿真结果

实验中各参数设置如表 2 所示. 其中, $Conf_{\text{SPE}}$ 和 $Conf_{T^2}$ 分别为 PCA 统计量控制限 SPE 和 T^2 的置信度; ϕ_0 为验证样本平均测量误差. 上述参数通过实验确定.

原始测量模型在各数据集中的测量结果如图 5 所示. 由图 5 可知, 原始测量模型在两个数据集的漂移发生时刻 (第 500 个样本) 均产生较大的测量误差, 并对此后的漂移样本均无法有效拟合.



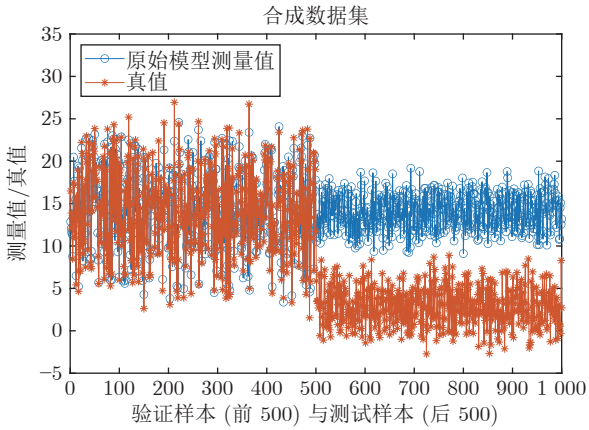
(a) 合成数据集
(a) The synthetic data set



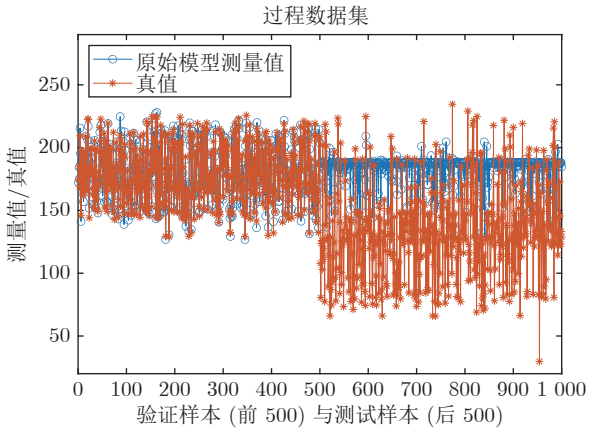
(b) 过程数据集
(b) The process data set

图 4 各特征在概念漂移环境中的变化情况

Fig. 4 Changes of each feature in the concept drift environment



(a) 合成数据集
(a) The synthetic data set



(b) 过程数据集
(b) The process data set

图 5 原始模型测量结果

Fig. 5 Measurement results of the original model

表 2 仿真参数设置

Table 2 Simulation parameter setting

参数名称	数据集	
	合成	过程
GPR 核函数	径向基核函数	径向基核函数
核函数宽度	0.5967	1.5116
核函数特征长度	0.7939	1.4734
待标注样本窗口容量 (w)	8	50
PCA 控制限置信度 ($Conf_{\mathrm{SPE}}$, $Conf_{T^2}$)	0.8, 0.8	0.9, 0.9
TD 学习最近邻数量 (ε)	6	5
Page-Hinkley 检测法基准累计平均测量误差 (ϕ_0)	2.2919	16.8846

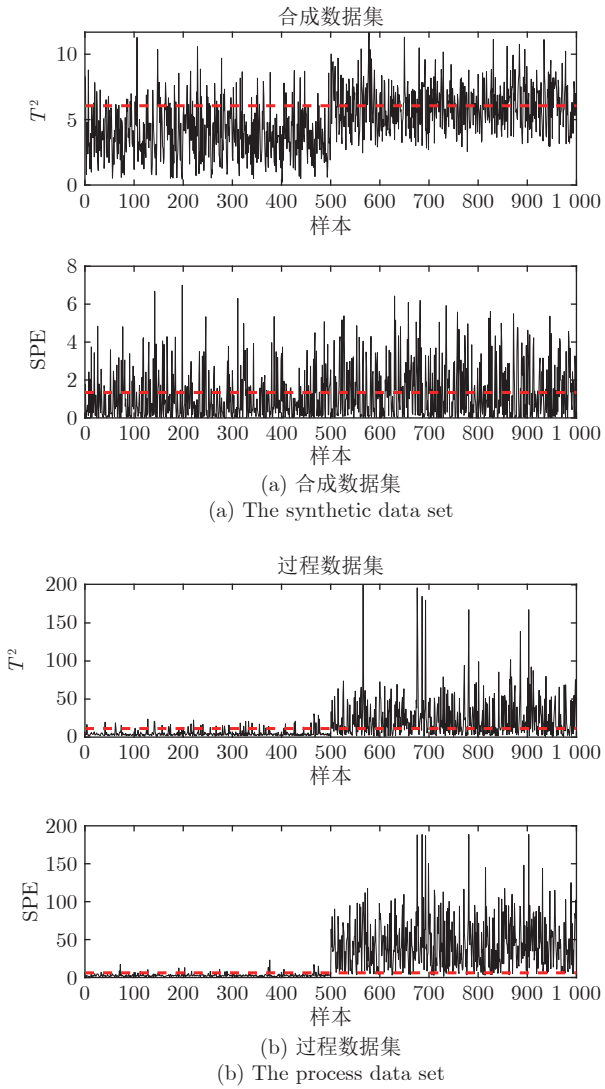
1) 特征空间漂移检测

针对数据集中存在的概念漂移现象, 采用 PCA 对验证样本和漂移样本特征空间的漂移检测结果如图 6 所示. 图中实线代表 PCA 统计量, 虚线代表统计量控制限.

图 6 显示了验证样本和漂移样本特征空间的 PCA 统计量与 PCA 统计量控制限的大小关系. 其中, 在合成数据集中共测得特征空间漂移样本 400 个; 在过程数据集中共测得特征空间漂移样本 450 个. 从图 6 可看出, 过程数据集中样本特征空间分布对工况变化较为敏感, 因此采用 PCA 可有效测出漂移时刻对应样本.

2) 基于 TD 学习的伪真值标注

针对特征空间漂移的样本, 基于 TD 学习对其伪真值标注结果与实际真值的比较如图 7 所示. 其中, 在合成数据集中共标注伪真值 350 个, 伪真值与真值间平均误差为 3.2760 (实际真值标准差为 2.2606); 在过程数据集中共标注伪真值 441 个, 伪真值与真值间平均误差为 35.9429 (实际真值标准差为 36.3831), 两个数据集中伪真值平均标注误差与实际真值自身离散程度相似. 此外, 从图 7 可看

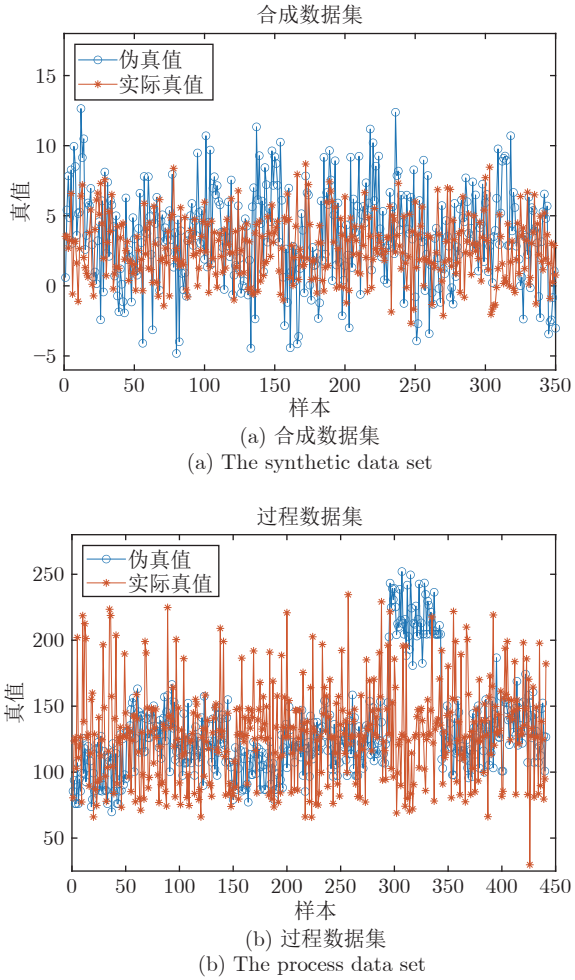


(a) 合成数据集
(a) The synthetic data set

(b) 过程数据集
(b) The process data set

图 6　针对特征空间的漂移检测结果

Fig. 6　Drift detection results in the feature space



(a) 合成数据集
(a) The synthetic data set

(b) 过程数据集
(b) The process data set

图 7　针对特征空间漂移样本的伪真值标注结果

Fig. 7　Pseudo-true value labeling results for samples with concept drift in the feature space

出, 伪真值变化趋势与样本真值相近, 因此在样本真值难以完全获取时可采用伪真值对样本输出空间漂移情况近似分析.

3) 输出空间检测结果

对特征空间漂移的样本完成伪真值标注后, 采用 Page-Hinkley 检测法对样本输出空间的漂移检测结果如图 8 所示.

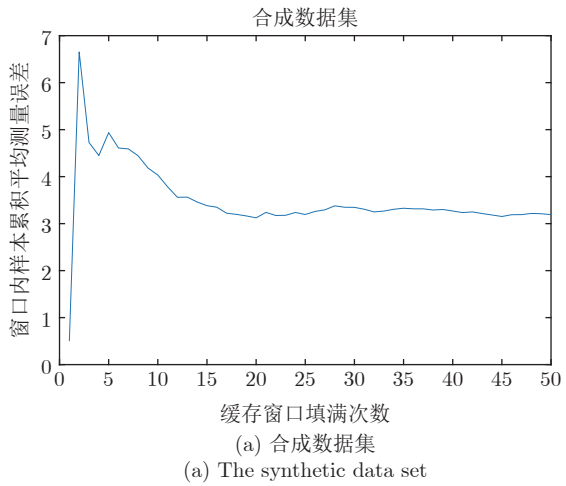
图 8 为每次待标注缓存窗口被填满且其中样本均被标注伪真值后, 窗口内样本累计平均测量误差的变化情况. 其中, 在合成数据集中待标注缓存窗口填满 50 次; 在过程数据集中待标注缓存窗口填满 9 次. 从图 8 可看出, 窗口内样本累计平均测量误差在漂移发生时刻明显升高, 随模型不断更新而趋于平稳, 表明所提算法可有效检测样本输出空间中存在的概念变化.

4) 测量模型更新

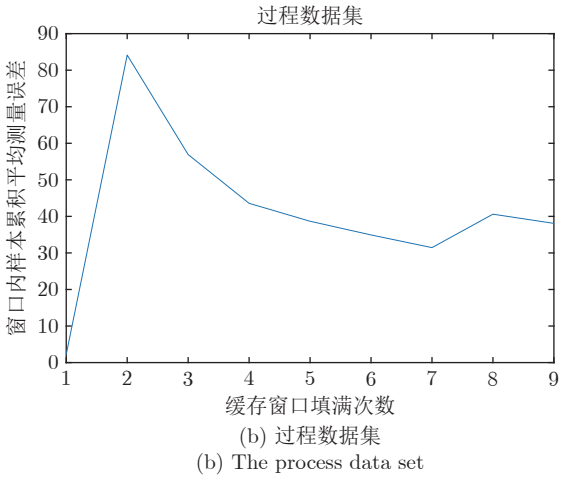
依据上述检测结果, 模型采用由概念漂移样本和历史样本组成的新训练集更新后, 在各数据集中的测量性能变化如图 9 所示.

由图 9 可知, 测量模型采用所提漂移检测算法后, 其测量误差相较原始模型明显下降, 详细更新信息及模型均方根测量误差 (Root mean squared error, RMSE) 变化情况如表 3 所示.

由表 3 可知: 1) 合成数据集中, 算法在 500 个漂移样本环境下, 共标注样本伪真值 350 个, 更新后使模型 RMSE 降低 66.2%, 相较原始模型真值需求量降低 99.2%; 2) 过程数据集中, 算法在 500 个漂移样本环境下, 共标注样本伪真值 441 个, 更新后使模型 RMSE 降低 45.5%, 相较原始模型真值需求量降低 98.2%. 上述结果表明: 所提算法可在大部分漂移样本真值未标注情况下, 显著提升模型面对概念漂移样本的测量性能, 可有效提高 MSWI 过



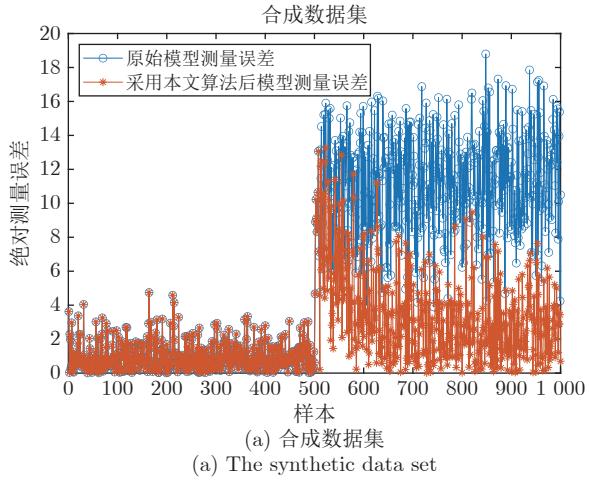
(a) 合成数据集
(a) The synthetic data set



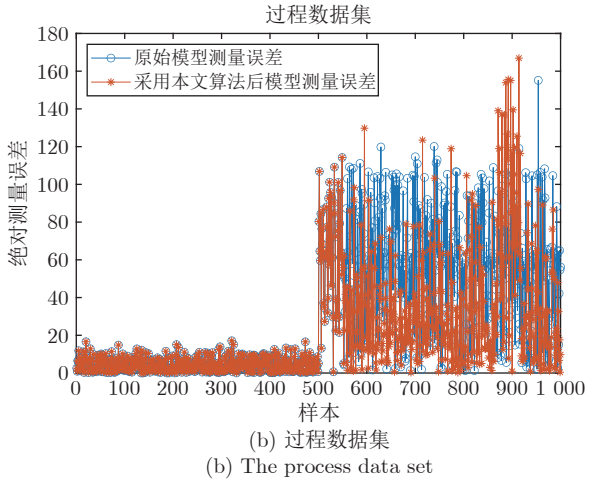
(b) 过程数据集
(b) The process data set

图 8 针对输出空间的漂移检测结果

Fig. 8 Drift detection results in the output space



(a) 合成数据集
(a) The synthetic data set



(b) 过程数据集
(b) The process data set

图 9 采用所提漂移检测算法后模型测量误差变化

Fig. 9 Changes of model measurement error after adopting the proposed drift detection algorithm

程氮氧化物浓度软测量模型在漂移环境中的测量精度.

4.3 方法比较

1) 漂移检测性能比较

为验证所提漂移检测算法具有优于已有方法的性能, 此处与仅基于特征空间的无监督型算法和仅基于输出空间的有监督型算法进行比较, 前者基于 PCA 检测样本特征空间漂移状况[19], 后者采用模型测量误差检测样本输出空间漂移状况[41]. 比较结果如表 4 和图 10 所示.

由上述结果分析可知: 1) 相较无监督型算法, 本文算法在两个数据集中均使模型更新后具有更低的测量 RMSE 值, 更新过程中真值需求量缩减 50.5% (合成)、98.0% (过程); 2) 相较有监督型算法, 本文算法具有更低的更新次数, 且在真值需求量分别缩减 55.6% 和 98.0% 的情况下, 仍使模型更新后具有与其接近的测量 RMSE 值. 综上表明: 所提算法可有效提升无监督型算法的更新效率, 并在仅少量真值标注情况下保持与有监督型算法相近的更新性能.

表 3 所提算法检测信息

Table 3 Detection information of the proposed algorithm

	合成数据集	过程数据集
缓存窗口填满次数	50	9
模型更新次数	44	8
标注漂移样本伪真值数	350	441
原始模型 RMSE	7.6478	53.0210
采用本文算法后模型 RMSE	**2.5840**	**28.8785**

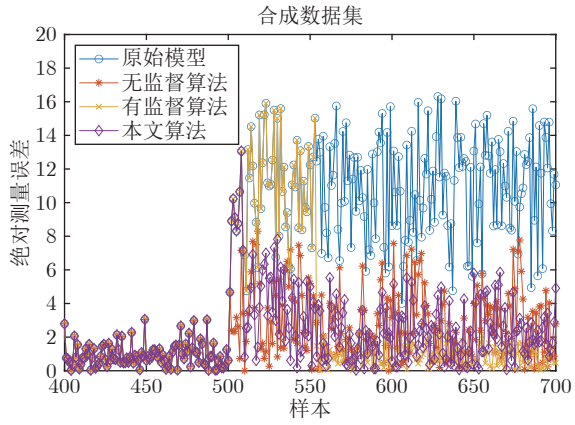
2) 建模策略比较

为验证 GPR 模型的高效测量性能, 此处与两种常用机器学习模型: 支持向量回归 (Support vec-

表 4　不同算法检测性能比较

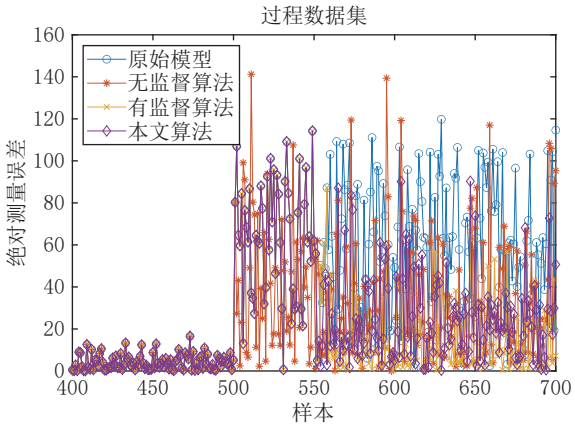
Table 4　Comparison of detection performance of different algorithms

数据集	检测算法	模型更新次数	更新所需真值数	模型测量 RMSE	其他
合成	无监督型	101	101	2.5846	需采用真值更新
	有监督型	99	990	2.2943	需采用真值检测与更新
	本文算法	44	50	**2.5840**	采用伪真值更新
过程	无监督型	463	463	35.8261	需采用真值更新
	有监督型	19	450	28.4729	需采用真值检测与更新
	本文算法	8	9	**28.8785**	采用伪真值更新



(a) 合成数据集 (局部)

(a) Part of the synthetic data set



(b) 过程数据集 (局部)

(b) Part of the processdata set

图 10　采用不同算法时模型测量误差变化

Fig. 10　Changes in model measurement errors when using different algorithms

tor regression, SVR) 和回归树 (Regression tree, RT) 进行比较. 除模型外其余参数均与上文实验中保持一致, 比较结果如表 5 所示.

由表 5 分析可知, 上述模型均取最优测量结果时, GPR 表现仍优于其他模型. 在合成数据集中, GPR 具有最优的训练 RMSE、R^2 和测量 RMSE (分别为 0.1899、0.96 和 2.5840); 在过程数据集中, GPR 在训练阶段的拟合效果与 SVR 相近 (分别为 0.1348 和 0.98), 但在测量阶段具有最优泛化性能 (28.8785).

3) 近邻规则比较

为验证基于 TD 学习的伪真值标注过程中欧氏距离作为近邻规则的有效性, 此处与两种常用的相似性度量方式: 曼哈顿距离与切比雪夫距离进行比较. 比较过程中参数设置与实验部分保持一致, 其结果如表 6 所示.

由表 6 分析可知, 相较其他度量方式, 欧氏距离能够体现特征空间数值上的绝对差异, 而概念漂移样本相较历史样本常具有差异较大的特征值. 因此, 模型采用欧氏距离作为近邻规则时可较好捕获样本的相似性, 并在各数据集中均具有最优测量性能 (分别为 2.5840 和 28.8785).

4.4　参数分析

仿真过程中固定参数 (软测量模型核函数类型、核函数宽度、特征长度及基准累计平均测量误差 ϕ_0) 根据模型最小训练误差与最小验证样本测试误差选取, 可变参数 (待标注缓存窗口容量 w、PCA 控制限置信度 $Conf_{\mathrm{SPE}}$、$Conf_{T^2}$ 及 TD 学习最近邻数量 ε) 由实际仿真分析后选取. 以过程数据集为例, 不同可变参数对算法性能影响的分析结果如表 7 所示.

由表 7 可知:

1) 待标注缓存窗口容量 w 变化改变伪真值标注次数与模型更新次数, 进而对更新后模型 RMSE 产生影响. 当 w 偏小时缓存窗口易被填满, 更多样本被检测为特征空间异常并被确认漂移, 因此伪真值标注量与模型更新次数增加, 但由于单次更新模型的漂移样本数过少导致模型无法在每次更新时充分学习漂移特征, 易使更新后模型 RMSE 偏大. 当 w 偏大时缓存窗口难以填满, 伪真值标注量与模型更新次数随之降低, 但其较长的样本检索时间导致模型无法及时适应概念漂移, 同样易使更新后模型 RMSE 偏大.

表 5 不同模型测量性能比较

Table 5 Comparison of measurement performance of different models

数据集	测量模型	核函数 (核宽度)	最小叶尺寸	训练 RMSE	训练 R^2	测量 RMSE
合成	SVR	径向基 (0.5600)	—	0.2479	0.94	3.7900
	RT	—	4	0.3034	0.91	3.1241
	GPR	径向基 (0.5967)	—	**0.1899**	**0.96**	**2.5840**
过程	SVR	径向基 (1.1000)	—	0.1369	0.98	30.3916
	RT	—	4	0.1630	0.97	29.9548
	GPR	径向基 (1.5116)	—	**0.1348**	**0.98**	**28.8785**

表 6 不同距离函数对模型更新性能影响

Table 6 The influence of different distance functions on model updating performance

数据集	距离函数	伪真值标注平均误差	模型测量 RMSE
合成	曼哈顿距离	3.3434	3.1939
	切比雪夫距离	3.2382	3.2484
	欧氏距离	**3.2760**	**2.5840**
过程	曼哈顿距离	38.0043	28.9954
	切比雪夫距离	37.7392	28.9947
	欧氏距离	**35.9429**	**28.8785**

2) TD 学习中最近邻数量 ε 变化改变伪真值标注精度, 进而对更新后模型 RMSE 产生影响. 当 ε 偏小时被用于标注伪真值的历史样本数减少, 因此算法无法获取充足的历史差分量变化信息, 导致难以准确输出伪真值并易使更新后模型 RMSE 偏大. 当 ε 偏大时被用于标注伪真值的历史样本数增多, 此时算法易受相似度较低的历史差分量变化信息干扰, 同样导致更新后模型 RMSE 偏大.

3) 特征空间漂移检测过程中 PCA 控制限 ($Conf_{\mathrm{SPE}}$ 与 $Conf_{T^2}$) 的变化将改变算法在输出空间的检测样本数量, 进而使待标注缓存窗口填满次数、伪真值标注次数、模型更新次数及伪真值标注精度变化, 并对更新后模型 RMSE 产生影响. 其影响方式与可变参数 w、ε 变化所产生的影响相似, 即改变模型对漂移的学习程度与其更新效率.

上述分析表明, 可变参数的设置方式对软测量模型的最终性能具有一定影响. 在选择参数时需结合实际应用背景, 具体为: 1) 新样本概念变化缓慢或对模型测量影响程度较小时, 应设置较大缓存样本窗口容量以充分学习漂移特征, 从而获取最优测量性能; 反之则应设置较小缓存样本窗口容量以及时避免测量性能快速恶化; 2) 当新样本的特征空间分布与历史样本接近时, 应设置较小的最近邻数量以避免提取冗余差分量信息, 同时设置较低的 PCA 控制限有利于在输出空间区分新概念样本; 反之则应设置较大的最近邻数量和 PCA 控制限, 从而准确标注新样本伪真值并提前将其在特征空间中与历史样本区分, 提高输出空间检测效率. 实际上, 更新后模型 RMSE 变化不仅由算法中单一可变参数改变引起, 还体现为上述参数的综合变化. 因此, 所提漂移检测算法应用于工业过程时, 应设置可供交互的数据界面窗口, 实时调整可变参数以获取最优检测及模型更新效果.

5 结语

针对复杂工业过程存在概念漂移、部分难测参数的真值难以及时获取问题, 文中提出一种联合样本输出与特征空间的半监督概念漂移检测方法. 其策略是: 通过 PCA 筛选特征空间内存在概念漂移的样本后, 再结合 TD 学习算法和 Page-Hinkley 检测法, 在样本输出空间进行伪真值标注并识别能够表征概念漂移的新样本. 本文所提方法的创新性表现在: 1) 采用联合 PCA 和 Page-Hinkley 检测法的策略充分反映新样本在特征空间和样本输出空间的概念漂移行为; 2) 将基于 TD 学习的半监督机制用于特征空间漂移样本的伪真值标注, 为面向工业回归问题的半监督概念漂移检测提供了新方法; 3) 采用真实 MSWI 过程数据集验证了所提方法在实际应用中的可行性, 并表明其具有优于已有方法的性能.

目前, 面向工业回归测量领域的半监督漂移检测研究尚处于探索阶段. 进一步的研究方向包括: 1) 为避免凭借人工经验设定模型参数导致漂移检测过程的随意性和差异性, 研究模型参数的自适应选择算法; 2) 为提高标注的准确度, 对伪真值标注算法进行改进; 3) 为提高概念漂移检测算法的适应性, 研究针对实际工业过程的漂移理解和漂移处理策略.

表 7　不同可变参数对应算法性能变化

Table 7　Algorithm performance changes corresponding to different variable parameters

样本窗口容量 w	最近邻数量 ε	PCA 控制限 $Conf_{\mathrm{SPE}}$,$Conf_{T^2}$	缓存窗口填满次数	标注伪真值数	更新次数	伪真值标注平均误差	模型测量 RMSE
30	3	0.85, 0.85	16	464	13	38.9005	31.0823
		0.90, 0.90	16	464	15	48.2016	35.2513
		0.95, 0.95	16	464	12	37.7528	28.9876
	5	0.85, 0.85	16	464	15	40.0004	30.4071
		0.90, 0.90	16	464	15	47.6636	34.2694
		0.95, 0.95	15	435	13	39.0258	31.0078
	8	0.85, 0.85	16	464	12	40.1782	28.8912
		0.90, 0.90	16	464	15	46.5567	32.8323
		0.95, 0.95	15	435	14	38.4400	30.5321
50	3	0.85, 0.85	9	441	8	42.9923	30.1536
		0.90, 0.90	9	441	8	36.8999	29.7216
		0.95, 0.95	9	441	7	31.2822	29.3330
	5	0.85, 0.85	9	441	8	43.4483	29.8960
		0.90, 0.90	9	441	9	35.9429	**28.8785**
		0.95, 0.95	9	441	7	31.9674	29.9178
	8	0.85, 0.85	9	441	8	42.9759	29.4615
		0.90, 0.90	9	441	8	37.0338	29.2796
		0.95, 0.95	9	441	6	31.4267	29.3356
70	3	0.85, 0.85	6	414	5	44.7315	33.6308
		0.90, 0.90	6	414	5	46.9859	36.2573
		0.95, 0.95	6	414	5	33.4711	33.1686
	5	0.85, 0.85	6	414	5	41.9744	32.4663
		0.90, 0.90	6	414	5	44.4580	34.3495
		0.95, 0.95	6	414	5	33.6287	34.2660
	8	0.85, 0.85	6	414	5	42.3929	31.0446
		0.90, 0.90	6	414	5	45.8771	34.5003
		0.95, 0.95	6	414	5	33.2206	33.5950

References

1 Kolekar K A, Hazra T, Chakrabarty S N. A review on prediction of municipal solid waste generation models. *Procedia Environmental Sciences*, 2016, **35**: 238−244

2 Li X, Zhang C, Li Y, Zhi Q. The status of municipal solid waste incineration (MSWI) in China and its clean development. *Energy Procedia*, 2016, **104**: 498−503

3 Qiao Jun-Fei, Guo Zi-Hao, Tang Jian. Dioxin emission concentration measurement approaches for municipal solid wastes incineration process: A survey. *Acta Automatica Sinica*, 2020, **46**(6): 1063−1089
(乔俊飞, 郭子豪, 汤健. 面向城市固废焚烧过程的二噁英排放浓度检测方法综述. 自动化学报, 2020, **46**(6): 1063−1089)

4 Tang Jian, Qiao Jun-Fei, Xu Zhe, Guo Zi-Hao. Soft measuring approach of dioxin emission concentration in municipal solid waste incineration process based on feature reduction and selective ensemble algorithm. *Control Theory and Applications*, 2021, **38**(1): 110−120
(汤健, 乔俊飞, 徐喆, 郭子豪. 基于特征约简与选择性集成算法的城市固废焚烧过程二噁英排放浓度软测量. 控制理论与应用, 2021, **38**(1): 110−120)

5 Tang Jian, Xia Heng, Qiao Jun-Fei, Guo Zi-Hao. Deep ensemble forest regression modeling method with its application research [Online], available: http://kns.cnki.net/kcms/detail/11.2286.T.20200723.1048.002.html, July 23, 2020
(汤健, 夏恒, 乔俊飞, 郭子豪. 深度集成森林回归建模方法及应用研究 [Online], available: http://kns.cnki.net/kcms/detail/11.2286.T.20200723.1048.002.html, July 23, 2020)

6 Wang S, Schlobach S, Klein M. What is concept drift and how to measure it? In: Proceedings of the 2010 International Conference on Knowledge Engineering and Knowledge Management. Lisbon, Portugal: Springer, 2010. 241–256

7 Widmer G, Kubat M. Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 1996, **23**(1): 69−101

8 Tang Jian, Chai Tian-You, Liu Zhuo, Yu Wen, Zhou Xiao-Jie. Adaptive ensemble modelling approach based on updating sample intelligent identification. *Acta Automatica Sinica*, 2016, **42**(7): 1040−1052
(汤健, 柴天佑, 刘卓, 余文, 周晓杰. 基于更新样本智能识别算法的自适应集成建模. 自动化学报, 2016, **42**(7): 1040−1052)

9 Žliobaitė I. Learning under concept drift: An overview [Online], available: http://arxiv.org/abs/1010.4784, October 22, 2010

10 Lu J, Liu A, Dong F, Gu F, Gama J, Zhang G. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 2018, **31**(12): 2346−2363

11 Gama J, Medas P, Castillo G, Rodrigues P. Learning with drift detection. In: Proceedings of the 17th Brazilian Symposium on Artificial Intelligence. São Luís, Brazil: Springer, 2004. 286–295

12 Pesaranghader A, Viktor H L. Fast hoeffding drift detection method for evolving data streams. In: Proceedings of the 2016 Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Riva Del Garda, Italy: Springer, 2016. 96–111

13 Yang Z, Al-Dahidi S, Baraldi P, Zio E, Montelatici L. A novel

concept drift detection method for incremental learning in non-stationary environments. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, **31**(1): 309−320

14 Frías B I, Campo A J, Ramos J G, Morales B R, Ortiz D A, Caballero M Y. Online and non-parametric drift detection methods based on Hoeffding′s bounds. *IEEE Transactions on Knowledge and Data Engineering*, 2014, **27**(3): 810−823

15 Mahdi O A, Pardede E, Ali N, Cao J. Diversity measure as a new drift detection method in data streaming. *Knowledge-Based Systems*, 2020, **191**: Article No. 105227

16 Korpela T, Kumpulainen P, Majanne Y, Häyrinen A, Lautala P. Indirect NO_x emission monitoring in natural gas fired boilers. *Control Engineering Practice*, 2017, **65**: 11–25

17 Tang J, Yu W, Chai T Y, Zhao L J. Online principal component analysis with application to process modeling. *Neurocomputing*, 2012, **82**: l67−168

18 Han X, Tian S, Romagnoli J A, Lic H, Suna W. PCA-SDG based process monitoring and fault diagnosis: Application to an industrial pyrolysis furnace. *IFAC-PapersOnLine*, 2018, **51**(18): 482−487

19 Liu S, Feng L, Wu J, Hou G, Han G. Concept drift detection for data stream learning based on angle optimized global embedding and principal component analysis in sensor networks. *Computers & Electrical Engineering*, 2017, **58**: 327−336

20 Toubakh H, Sayed-Mouchaweh M. Hybrid dynamic data-driven approach for drift-like fault detection in wind turbines. *Evolving Systems*, 2015, **6**(2): 115−129

21 Xu S, Feng L, Liu S, Qiao H. Self-adaption neighborhood density clustering method for mixed data stream with concept drift. *Engineering Applications of Artificial Intelligence*, 2020, **89**: Article No. 103451

22 Wang X S, Kang Q, Zhou M C, Yao S Y. A multiscale concept drift detection method for learning from data streams. In: Proceedings of the 14th International Conference on Automation Science and Engineering. Munich, Germany: IEEE, 2018. 786–790

23 Liu A, Lu J, Liu F, Zhang G. Accumulating regional density dissimilarity for concept drift detection in data streams. *Pattern Recognition*, 2018, **76**: 256−272

24 Lughofer E, Weigl E, Heidl W, Eitzinger C, Radauer T. Recognizing input space and target concept drifts in data streams with scarcely labeled and unlabelled instances. *Information Sciences*, 2016, **355**: 127−151

25 Haque A, Khan L, Baron M, Thuraisingham B, Aggarwal C. Efficient handling of concept drift and concept evolution over stream data. In: Proceedings of the 32nd International Conference on Data Engineering. Helsinki, Finland: IEEE, 2016. 481–492

26 Tan C H, Lee V, Salehi M. Online semi-supervised concept drift detection with density estimation [Online], available: https://arxiv.org/abs/1909.11251, November 11, 2019

27 Zhou Z H, Li M. Semi-supervised regression with co-training. In: Proceedings of the 2005 International Joint Conference on Artificial Intelligence. Scotland, UK: AAAI, 2005. 908–913

28 Miller J A, Bowman C T. Mechanism and modelling of nitrogen chemistry in combustion. *Progress in Energy and Combustion Science*, 1989, **15**(4): 287−338

29 Kadlec P, Gabrys B, Strandt S. Data-driven soft sensors in the process industry. *Computers & Chemical Engineering*, 2009, **33**(4): 795−814

30 Schlimmer J C, Granger R H. Incremental learning from noisy data. *Machine Learning*, 1986, **1**(3): 317−354

31 Yang Jun-Zhi. Full analysis on accuracy and related terms. *Science of Surveying and Mapping*, 2011, **36**(01): 75−76
(杨俊志. 测量准确度及相关术语辨析. 测绘科学, 2011, **36**(01): 75−76)

32 Wang B, Mao Z. Outlier detection based on gaussian process with application to industrial processes. *Applied Soft Computing*, 2019, **76**: 505−516

33 Schulz E, Speekenbrink M, Krause A. A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, 2018, **85**: 1−16

34 Yin S, Ding S X, Xie X, Luo H. A review on basic data-driven approaches for industrial process monitoring. *IEEE Transactions on Industrial Electronics*, 2014, **61**(11): 6418−6428

35 Tang J, Yu W, Chai T Y, Liu Z, Zhou X. Selective ensemble modeling load parameters of ball mill based on multi-scale frequency spectral features and sphere criterion. *Mechanical Systems & Signal Processing*, 2016, **66**: 485−504

36 Kaneko H, Funatsu K. Classification of the degradation of soft sensor models and discussion on adaptive models. *Aiche Journal*, 2013, **59**(7): 2339−2347

37 Yuan Xiao-Feng, Ge Zhi-Qiang, Song Zhi-Huan. Adaptive soft sensor based on time difference model and locally weighted partial least squares regression. *Journal of Chemical Industry and Engineering* (*China*), 2016, (3): 724−728
(袁小锋, 葛志强, 宋执环. 基于时间差分和局部加权偏最小二乘算法的过程自适应软测量建模. 化工学报, 2016, (3): 724−728)

38 Kaneko H, Funatsu K. Maintenance-free soft sensor models with time difference of process variables. *Chemometrics and Intelligent Laboratory Systems*, 2011, **107**(2): 312−317

39 Pu Xiao-Long, Improvement of CUSUM test. *Acta Mathematicae Applicate Sinica*, 2003, (2): 225−241
(濮晓龙. 关于累积和 (CUSUM) 检验的改进. 应用数学学报, 2003, (2): 225−241)

40 Ikonomovska E. Algorithms for Learning Regression Trees and Ensembles on Evolving Data Streams [Ph.D. Dissertation], Jožef Stefan International Postgraduate School, The Republic of Slovenia, 2012

41 Channoi K, Maneewongvatana S. Concept drift for CRD prediction in broiler farms. In: Proceedings of the 12th International Joint Conference on Computer Science and Software Engineering. Songkhla, Thailand: IEEE, 2015. 287–290

孙子健 北京工业大学信息学部硕士研究生. 主要研究方向为概念漂移检测, 城市固废焚烧过程难测参数软测量.
E-mail: sunzj@emails.bjut.edu.cn
(**SUN Zi-Jian** Master student at the Faculty of Information Technology, Beijing University of Technology. His research interest covers concept drift detection and soft measurement of difficulty-to-measure parameters in municipal solid waste incineration process.)

汤 健 北京工业大学信息学部教授. 主要研究方向为小样本数据建模, 城市固废处理过程智能控制.
E-mail: freeflytang@bjut.edu.cn
(**TANG Jian** Professor at the Faculty of Information Technology, Bei jing University of Technology. His research interest covers small sample data modeling and intelligent control of municipal solid waste treatment process.)

乔俊飞 北京工业大学信息学部教授. 主要研究方向为污水处理过程智能控制, 神经网络结构设计与优化. 本文通信作者.
E-mail: junfeiq@bjut.edu.cn
(**QIAO Jun-Fei** Professor at the Faculty of Information Technology, Beijing University of Technology. His research interest covers intelligent control of wastewater treatment process, and structure design and optimization of neural networks. Corresponding author of this paper.)