面向工业过程软测量建模的概念漂移检测综述

乔俊飞†, 孙子健, 汤 健

(北京工业大学 信息学部, 北京 100124; 计算智能与智能系统北京市重点实验室, 北京 100124)

摘要: 基于数据驱动的软测量模型广泛用于工业过程中产品质量与环保指标等难测参数的在线测量, 该过程中存在的概念漂移问题易导致模型精度下降. 如何有效识别过程概念变化并精准检测漂移样本是提高模型测量性能的关键. 本文总结并分析目前漂移检测的研究思路与进展, 为面向工业过程软测量的漂移检测算法提供设计指导. 首先, 介绍了概念漂移的通常定义与其在工业过程中的表现形式; 然后, 从检测依据与检测对象两个视角分析了目前具有代表性的检测方法; 接着, 讨论了这些算法的技术特点和当前工业领域的研究难点; 最后, 展望了未来的研究方向.

关键词: 工业过程; 软测量; 概念漂移; 过程变量; 样本分布

引用格式: 乔俊飞, 孙子健, 汤健. 面向工业过程软测量建模的概念漂移检测综述. 控制理论与应用, 2021, 38(8): 1159 – 1174

DOI: 10.7641/CTA.2021.00334

Overview of concept drift detection for industrial process soft sensor modeling

QIAO Jun-fei†, SUN Zi-jian, TANG Jian

(Faculty of Information Technology, Beijing University of Technology, Beijing, 100124, China;
Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing 100124, China)

Abstract: Data-driven soft sensor models are widely used for online measurement of difficult-to-measure parameters such as product quality and environmental protection indicators in industrial processes, and the concept drift in this process will lead to a decrease in model accuracy. Effective recognition of process concept changes and accurate detection of drift samples are the keys to improving model measure performance. This paper summarizes and analyzes the current research ideas and progress of drift detection, and provides design guidance for drift detection algorithms for industrial soft sensor modeling. First, the general definition of concept drift and its manifestation in the industrial process are introduced. Then, the current representative research methods are analyzed from the perspective of detection object and detection basis. Next, the technical characteristics of different algorithm strategies and the current research difficulties in the industrial field according to the literature are discussed. Finally, suggestions for future research directions are given.

Key words: industrial process; soft sensor; concept drift; process variable; sample distribution

Citation: QIAO Junfei, SUN Zijian, TANG Jian. Overview of concept drift detection for industrial process soft sensor modeling. *Control Theory & Applications*, 2021, 38(8): 1159 – 1174

1 引言

随着传感器技术与计算机水平的持续发展, 现代工业过程有望通过融入大量数据以期实现对运行状态的更精准有效控制. 为实现上述目标, 软测量建模方法被广泛用于具有连续化和复杂化等特点的工业系统, 其依据过程数据建立难测参数的测量模型[1–3]. 实际建模任务中, 过程数据因其随时间变化所具有的非平稳性引起了众多学者关注, 尤其是数据分布随时间发生变化导致旧模型无法适用于新样本的问题, 该现象被称为概念漂移[4], 其产生原因一般是工业中元器件老化或生产环境变化导致模型输入输出的关系改变, 其通常难以预知与量化. 为此, 建模过程通常引入在线学习方法(如非线性感知器[5]、正则化对偶平均[6]和LASSO[7]等)实现在线动态建模, 目的是使软

收稿日期: 2020−06−10; 录用日期: 2021−03−05.

†通信作者. E-mail: junfeiq@bjut.edu.cn; Tel.: +86 10-67391766.

本文责任编委: 阳春华.

国家自然科学基金项目(61703089, 61890930–5), 国家科技重大专项项目(2018YFC1900801)资助.

Supported by the National Natural Science Foundation of China (61703089, 61890930–5) and the National Science and Technology Major Project (2018YFC1900801).

测量模型能够根据新样本实时更新, 以在不断变化的数据环境中保持良好的测量精度, 同时有效缩减数据存储成本.

尽管在线动态建模使模型具有自主调整能力, 但在概念漂移环境中通常还需对模型更新方式进行引导, 否则模型将由于无法全面了解环境变化而长期处于频繁更新状态, 并因此消耗更多计算资源且易导致测量不及时或准确性下降, 此时有必要仅采用新概念样本对模型进行针对性更新, 以提高模型在环境变化时的适应速度[8].

为实现对新概念样本的精准筛选, 针对样本漂移检测的研究得到迅速发展. 图1展示了近20年内概念漂移相关文献的发表与引用数量变化情况[1].

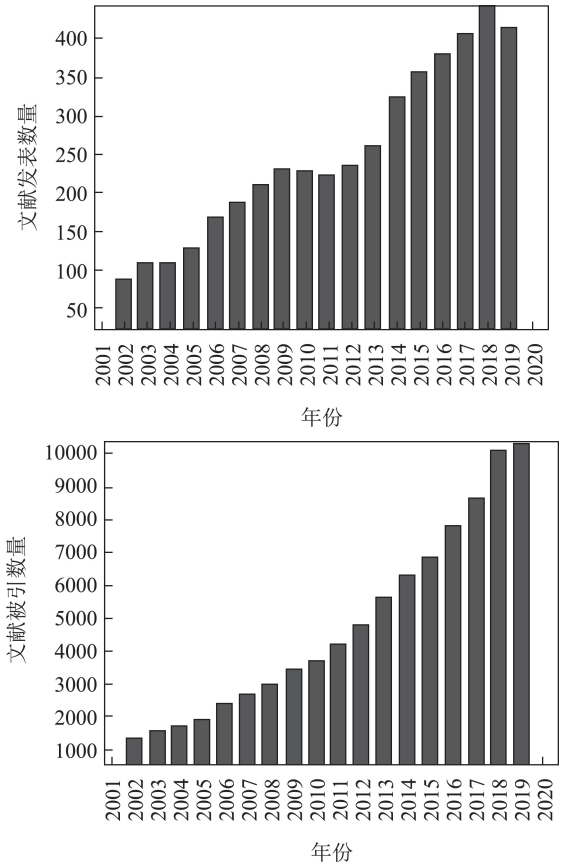


图 1 概念漂移研究现状

Fig. 1 Research status of concept drift

由图1可知, 该领域研究规模正逐渐扩大, 已有大量学者加入概念漂移相关问题讨论. 在这些研究中, 较早的漂移检测系统是FLORA系列算法[9], 该算法初步实现样本概念变化的判别与存储能力. 随后的工作中, 文献[4, 10–12]等进一步完善了概念漂移的产生原因、类型和定义; 文献[13–16]等研究了漂移检测算法的不同学习方式, 包括半监督学习、主动学习和重复概念学习; 文献[17–18]等结合现有测量模型与检测算法构造了特定的漂移适应性模型. 综上, 随着漂移检测技术的逐步完善与成熟, 为对实际工业过程中软测量建模任务提供有意义的应用帮助, 有必要对当前领域的研究动态与趋势进行有指导意义的总结与展望.

目前已存在的综述文献在不同方面介绍了漂移检测算法的研究进展, 如: 文献[19]归纳了面向分类任务的检测算法; 文献[20]围绕漂移的检测、理解和适应三个方面进行方法总结; 文献[15]中包含了较详细的无监督和半监督检测方式; 文献[21]重点介绍了基于测量误差、统计检验和模型结构的3种检测方式; 文献[22]对概念漂移检测在网络安全、金融市场和教育媒体等互联网领域内的应用情况做出详细分析. 但现有综述文献集中于对计算机等领域的应用描述, 且多数围绕分类任务特点开展, 仍缺少对工业过程的应用分析. 实际工业过程具有强耦合、大时滞和不确定等特性, 其概念变化情况相较有明确类别指示的任务而言更加复杂且不易区分, 因此需结合过程特点有针对性地对漂移检测方法进行综述.

本文以工业过程为背景, 围绕基于数据驱动的软测量模型对现有漂移检测算法进行综述, 主要贡献有: 1) 结合目前漂移检测领域内的研究成果与实际工业过程特点, 将现有算法的检测依据分为3类: 基于难测参数测量误差、基于过程变量和基于综合因素, 以此归纳现有方法的不同研究重点; 2) 新划分不同算法的检测对象, 即在不同检测依据的基础上进一步区分针对单样本和多样本的研究策略, 并说明不同检测对象对模型更新方式的影响; 3) 讨论并总结现有方法的技术特点与工业过程中常见的部分研究难点; 4) 提出面向工业过程检测算法的未来研究方向建议.

2 概念漂移现象描述

2.1 概念漂移的一般定义与类别

概念漂移指目标样本统计特性根据时间以随机方式变化[23], 其最早由文献[24]提出, 依据是噪声数据会在某些情况下得到与非噪声数据相同的特征从而被误认为正常数据, 且该变化通常难以直接测量[25]. 以数据驱动角度分析, 概念漂移的形式如图2所示.

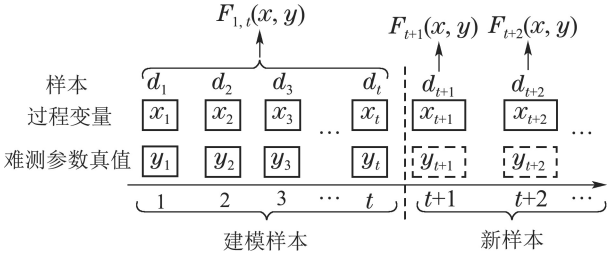


图 2 概念漂移的形式描述

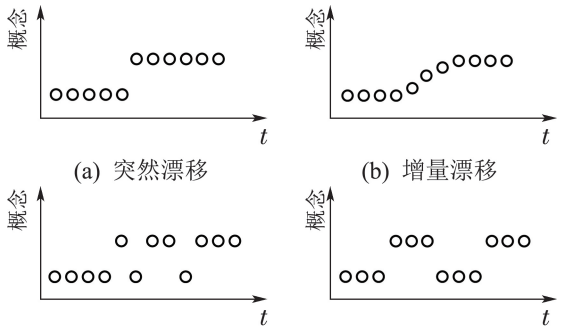
Fig. 2 Formal description of concept drift

结合图2, 可将其形式详细描述为: 给定$[1, t]$时刻

[1]Citation report of concept drift. Web of Science. www.isiknowledge.com.

内的建模样本集$S_{1,t}=\{\boldsymbol{d}_1,\cdots,\boldsymbol{d}_t\}$, 其中: $\boldsymbol{d}_i=(\boldsymbol{x}_i, y_i)(i\in[1,t])$是$S_{1,t}$中的一个样本, $\boldsymbol{x}_i$为样本过程变量(工业中对难测参数具有实际影响的温度、压力和流量等可实时测量参数), y_i为难测参数真值(约定真值[26], 即通过化验分析等方法确定的工业难测参数的最高基准值), $S_{1,t}$内样本均服从分布$F_{1,t}(\boldsymbol{x},y)$. 假定新时刻样本 $\boldsymbol{d}_k(k\in[t{+}1,\infty))$ 服从的分布为 $F_k(\boldsymbol{x},y)$, 当$F_{1,t}(\boldsymbol{x},y)\neq F_k(\boldsymbol{x},y)$时, 认为新样本$\boldsymbol{d}_k$相较建模样本$S_{1,t}$发生概念漂移.

依据不同视角, 现有研究将概念漂移划分为不同类别. 如: 文献[27]根据数据的产生环境差异提出虚、实概念漂移; 文献[28]根据漂移的产生原因将其描述为样本先验概率、类概率和后验概率的变化; 文献[29]依据时间序列分析思想将漂移分为随机噪声、随机趋势、随机替换和系统趋势; 文献[12]根据数据产生的多源性将概念漂移称为数据漂移. 上述研究均有助理解概念漂移本质. 目前, 多数漂移处理过程中, 最为常用的漂移类别为: 突然漂移、增量漂移、渐进漂移和重复漂移[4], 其示例如图3所示.



(a) 突然漂移　(b) 增量漂移　(c) 渐进漂移　(d) 重复漂移

图 3 常见漂移类型图示

Fig. 3 Illustration of common drift types

图3中: 突然漂移与增量漂移分别表示样本概念在较短或较长的时间内改变; 渐进漂移表示在旧概念不完全消失的情况下新概念将其逐渐替代; 重复漂移表现为多种概念交替出现. 上述漂移类型的划分依据是样本概念变化的速度与幅度.

2.2 工业过程中的概念漂移

2.2.1 研究背景简述

当前工业过程主要存在两类软测量建模方式[27]: 机理驱动和数据驱动. 前者通常为特定工业过程开发并常用于推理控制, 该类模型缺点是: 1) 建模需大量经验知识; 2) 通常简化理论背景, 不符合真实过程状态; 3) 侧重描述工业过程的理想稳态, 不适合瞬态表达. 相反, 数据驱动模型基于对过程直接且详细的测量, 因此可从多方面描述实际工业过程. 现有漂移检测研究通常建立在基于数据驱动的软测量模型基础上, 其典型建模流程如图4所示.

根据图4, 可将该过程具体描述如下: 1) 第1阶段为数据初步检查阶段, 该阶段获得现有过程数据、识别建模时可能出现的问题并确定建模任务; 2) 第2阶段为建模数据选择阶段, 该阶段将选出处于平稳状态的、适合模型训练和评估的过程数据; 3) 第3阶段为数据预处理阶段, 该阶段通常将第2阶段选择后得到的过程数据进行标准化表示, 并进行特征处理和缺失数据标记等工作; 4) 第4阶段选择合适的模型进行训练与测试, 常用模型有决策树、支持向量机和神经网络等; 5) 第5阶段采用人工的或学习过程中得到的经验更新模型. 工业过程中漂移检测研究位于上述第3和第5阶段, 即首先对新样本进行漂移判别与处理, 然后将新概念样本用于更新模型, 以使模型在新概念环境下保持良好的鲁棒性与测量精度.

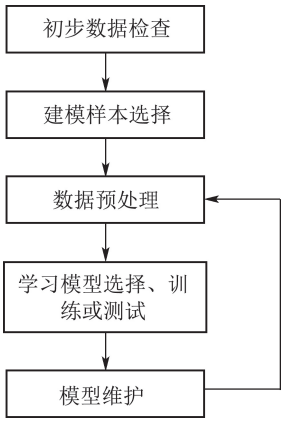


图 4 典型数据驱动软测量建模流程

Fig. 4 Typical data-driven soft sensor modeling process

此外, 相较其它应用领域, 工业过程中漂移检测研究通常还需考虑如下工业特点:

1) 回归任务多: 工业数据以连续型变量为主, 任务常集中于产品质量和环保指标等难测参数的软测量, 相较以分类任务为主的视觉识别等领域, 概念变化无法由类别改变直接表示, 通常需结合实际过程设定观测值阈值以确认漂移现象.

2) 工况变化复杂: 工业生产过程易受物料成分、生产环境变化等因素影响, 其工况变化形式与幅度较为复杂, 由此导致工业中概念漂移随机性较强, 且可能以多种类型共存的形式出现, 因此对检测算法的灵敏度和准确度均有较高要求.

3) 时效性要求高: 相较互联网中的消费心理、用户行为分析等领域, 工业概念漂移常预示潜在运行风险, 如无法及时检测与控制, 除造成经济损失外还可能引起人员伤亡及有毒污染物排放超标等严重运行事故.

2.2.2 漂移的实际影响与产生原因

概念漂移会使基于历史数据构建的软测量模型在面对漂移样本时测量性能下降, 进而影响工业系统的控制与决策[31]. 以现有研究为例:

文献[32]指出, 在流化床锅炉的燃烧质量与燃料流量测量过程中会出现概念漂移现象, 原因是燃料等级与成分改变使质量检测信号出现阶跃变化, 从而导致模型测量错误并使控制系统无法及时优化锅炉负载; 面向工业径向风扇自适应维护过程, 文献[33]提出变桨器机油中空气含量变化会影响旋转叶片仰角, 如无法及时检测并进行维护将降低风扇工作效率; 针对半导体蚀刻过程, 文献[34]指出不同材料的最佳蚀刻时间存在差异, 因此需要依据材料变化实时调整蚀刻时间, 否则将导致半导体结构宽度改变从而影响电路的电性能; 针对搅拌釜系统, 文献[35]指出换热器结垢参数值降低会使导体传热效率减小, 导致模型输出错误的测量值. 综上, 在软测量模型中引入概念漂移检测技术对提高工业过程控制效率具有重要意义.

根据漂移产生原因, 工业中将其分为过程漂移和传感器漂移[30]. 其中, 过程漂移一般有两种产生原因. 第一种是过程内部结构变化(机械元件磨损等), 如文献[36]提出图5所示的“可靠性浴盆曲线”, 表明一般情况下工业部件的可靠性会随时间变化并对过程本身产生影响; 第二种是过程外部条件变化(气候与工艺要求变化等), 以城市固废焚烧过程(municipal solid waste incineration, MSWI)为例, 固体废物含水率随季节与温度变化而改变, 炉膛温度依据实际燃烧状况进行实时调节, 这些变化均会影响出口烟气污染物的生成关系并进而对浓度测量产生干扰[37]. 以前文研究为例, 文献[32–34]属于工业过程外部条件变化引起的过程漂移, 这些漂移均由输入过程变量变化导致(燃料成分、机油质量和蚀刻材料), 文献[35]属于工业过程内部结构变化引起的漂移, 即由运行过程中参数变化导致(结垢参数). 传感器漂移也被称为测量漂移[38], 通常由传感器等硬件设施的测量精度改变导致, 因此该类漂移不反映运行过程的真实参数变化, 在漂移检测领域中研究较少.

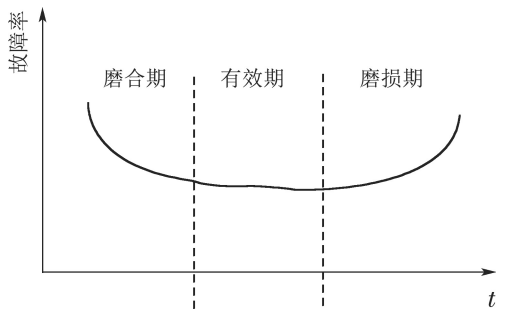


图 5 可靠性浴盆曲线

Fig. 5 Reliability bathtub curve

2.3 概念漂移的处理流程

针对漂移处理的理论研究包括: 文献[39]指出概念漂移检测可视为双重抽样问题, 即检查两个给定样本总体是否来自相同分布; 文献[40]从样本选择与加权角度对漂移样本检索方式进行讨论; 文献[28]基于贝叶斯理论将漂移归结为类概率、先验概率和后验概率的分布变化, 并以此指导漂移检测; 文献[10]给出了漂移的速度、持续时间和严重程度等定义; 文献[15]讨论了模型在漂移环境中的更新与适应方式.

基于上述研究, 文献[20]提出了如图6所示的概念漂移处理流程. 文中将概念漂移处理分为检测、理解和适应三个步骤. 其中, 漂移检测指通过识别变化点或变化间隔以表征和量化概念漂移的技术和机制; 漂移理解关注“何时”、“何地”和“如何”, 即识别漂移产生的时间、区域和程度等状态信息并将其作为漂移适应的输入; 漂移适应的目的是采用漂移状态信息更新模型, 其研究主要集中于简单再训练、集成再训练和模型调整3个方向.

基于上述通用概念漂移处理流程, 本文考虑实际工业过程中难测参数真值较难获得情况, 将工业过程概念漂移处理流程总结如图7所示. 图7中设置样本真值的查询与请求阶段的原因在于: 实际工业过程中部分难测参数真值通常无法及时获得, 如在MSWI过程中, 出口烟气污染物二噁英的浓度值需在专业检测中心经过多阶段核定, 其真值获得周期较长且费用高昂[37]. 此外, 现场人员通常根据工业过程的性能反馈有选择地标注样本, 以保证标注工作处于合理的经济范围内[41]. 上述情况常采用基于过程变量或综合因素的方法进行漂移检测, 具体细节将在第3章节介绍.

3 漂移检测方法综述

本节将分别从检测依据和检测对象两个视角对现有漂移检测算法进行归纳与讨论, 划分视角详情如表1所示.

表 1 综述视角划分

Table 1 Overview angle division

综述视角	划分详情
检测依据	基于难测参数测量误差
	基于过程变量
	基于综合因素
检测对象	针对单样本
	针对多样本

3.1 检测依据视角

本文将现有方法的检测依据分为3类: 基于难测参数测量误差、基于过程变量和基于综合因素. 其中: 基于难测参数测量误差的方法指通过模型测量误差的变化程度确认漂移; 基于过程变量的方法指通过分析样本过程变量间数值差异或分布变化进行检测; 基于综合因素的方法可视为前两种方法的结合.

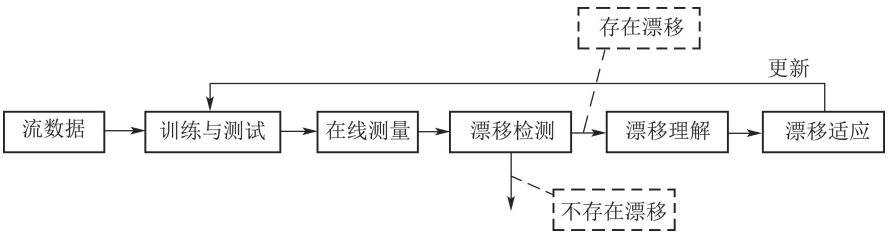


图 6 一般概念漂移处理流程

Fig. 6 General concept drift processing flow

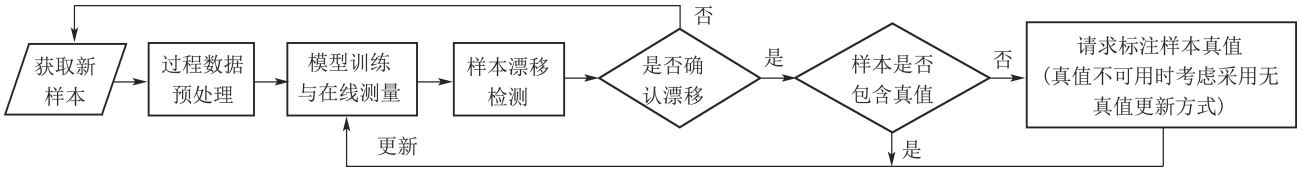


图 7 工业过程中概念漂移处理流程

Fig. 7 Concept drift processing flow in industrial process

3.1.1 基于难测参数测量误差的方法

在难测参数真值易获取的情况下, 测量误差是检测过程中最直观的判别标准之一, 因此仅基于难测参数测量误差的方法较为常见. 尽管测量误差变化通常无法说明样本分布的真实变化情况, 但仍可在一定程度反映变量输入输出关系的改变, 并能使该方法具有计算过程简便高效等特点.

该类研究中具代表性的算法是漂移检测法(drift detection method, DDM)[42], 其检测思路可描述为: 1) 首先依据二项式分布特点, 针对漂移程度定义漂移预警级别和漂移警告级别; 2) 然后使用窗口采集新样本(采集阶段), 计算窗口内样本的测量误差并存储其作为最新判别依据(在线测量阶段); 3) 最后通过计算模型当前的与历史的错误率差异判断(误差评估阶段): 当误差变化幅度达到漂移预警级别时, 存储当前窗口内样本, 并将这些样本用于构建新模型, 但此时仍然采用旧模型进行在线测量; 当错误率变化幅度达到漂移警告级别时, 采用此前构建的新模型代替当前模型进行在线测量(模型更新阶段). DDM的贡献是其初步提供了较为完整的检测框架(如图8所示), 即通过样本窗口、测量误差和级别定义完成新样本采集、在线测量、误差评估与模型更新.

后续较多研究均以图8所表示的检测框架为基础, 如:

1) 针对样本采集阶段的改进: 文献[43]采用衰落因子检索待测样本并结合Page-Hinkley方法对新样本检验, 结果表明该方式相较窗口式检索可有效降低检测延迟与存储成本; 文献[44]采用样本加权方式, 根据样本的采集顺序划分样本的概念变化权重, 以此筛选用于分析和比较的样本块.

2) 针对在线测量与误差评估阶段的改进: 文献[45]在算法中引入全局样本窗口以监视当前样本总体的测量误差, 并采用改进的等比例统计检验比较全局与新样本窗口内的在线测量误差差异以表征漂移; 文献[46]分别计算模型在总体样本和最近样本中可接受测量误差的出现概率, 采用Hoeffding不等式判断概率差异以确认漂移; 文献[44]采用指数加权移动平均(exponentially weighted moving average, EMWA)监视新样本真值与在线测量误差的平均值变化, 并同样通过Hoeffding不等式确认漂移.

3) 针对模型更新阶段的改进: 文献[47]采用集成方式设置多个并行样本窗口并在每个窗口内均建立在线测量模型, 当新样本到来时根据各模型测量误差分配窗口权重, 以权值最大窗口作为主模型以应对概念变化; 文献[48]提出双学习器概念, 即在算法中分别构造稳定和灵敏的在线测量模型, 根据两模型在不同概念环境中的测量精度交替使用.

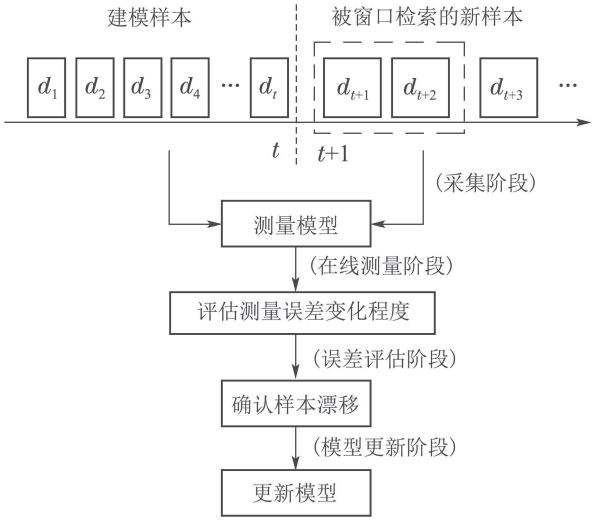


图 8 DDM检测框架

Fig. 8 DDM detection framework

此外, 文献[49]在DDM基础上将概念变化判别依据从测量误差的变化程度替换为两个错误测量之间的样本数量, 并因此表明算法的检测及时性得到改善; 文献[39] 提出基于支持向量机 (support vector machine, SVM)的检测方式, 即在两个样本中寻找最优线性间隔以使模型对两个样本的余量最大化, 通过观测两个线性间隔的相似度判别漂移; 文献[50]基于累积和控制图观察模型在线测量误差概率变化以反映样本分布差异; 文献[51]采用EMWA方法监控模型在线测量误差变化; 文献[52]基于在线随机权神经网络模型, 用新样本更新模型后, 量化并比较模型更新前后输出权重值的变化程度以表征漂移.

3.1.2 基于过程变量的方法

基于过程变量分布变化的常用漂移检测流程如图9所示.

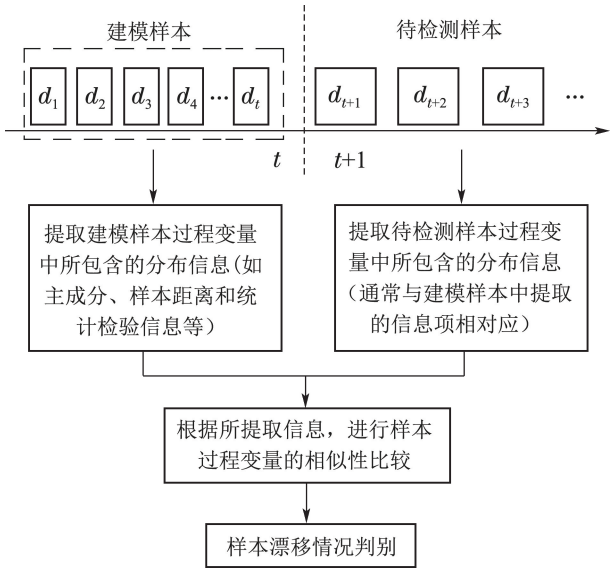


图 9 基于过程变量的检测流程

Fig. 9 Detection process based on process variable algorithm

根据图9可知, 该类算法首先提取过程变量中所包含的关键信息, 然后针对所提取信息进行相似性度量, 最后根据度量结果判断样本漂移情况. 本节将围绕上述过程中常见的3种检测策略展开描述, 分别是多元统计、距离度量和假设检验.

1) 多元统计策略.

该策略中较常见的方法是主成分分析(principal component analysis, PCA), 其被用于数据降维时表现出高效的数据分析能力, 因此也被广泛用于过程变量间相似性度量[53]. 该方法首先将新样本$\boldsymbol{d}_{t+1}$分为$\hat{\boldsymbol{d}}_{t+1}$和$\tilde{\boldsymbol{d}}_{t+1}$两部分

$$\boldsymbol{d}_{t+1} = \hat{\boldsymbol{d}}_{t+1} + \tilde{\boldsymbol{d}}_{t+1}, \tag{1}$$

$$\hat{\boldsymbol{d}}_{t+1} = \boldsymbol{d}_{t+1}\hat{P}_t\hat{P}_t^{\mathrm{T}}, \tag{2}$$

$$\tilde{\boldsymbol{d}}_{t+1} = \boldsymbol{d}_{t+1}(1 - \hat{P}_t\hat{P}_t^{\mathrm{T}}), \tag{3}$$

其中: $\hat{P}_t$为负荷矩阵, $\hat{\boldsymbol{d}}_{t+1}$和$\tilde{\boldsymbol{d}}_{t+1}$分别是$\boldsymbol{d}_{t+1}$在PCA模型的主元子空间和残差子空间上的投影.

然后, 计算新样本的平方测量误差(square prediction error, SPE)和Hotelling's T^2指标[51]:

$$\mathrm{SPE} \equiv \|\tilde{\boldsymbol{d}}_{t+1}\|^2 = \|\boldsymbol{d}_{t+1}(1 - \hat{P}_t\hat{P}_t^{\mathrm{T}})\|^2, \tag{4}$$

$$T^2 = \boldsymbol{d}_{t+1}\hat{P}_t\hat{\boldsymbol{\Lambda}}_t^{-1}\hat{P}_t^{\mathrm{T}}\boldsymbol{d}_{t+1}^{\mathrm{T}}, \tag{5}$$

$$\hat{\boldsymbol{\Lambda}}_t = \frac{\hat{T}_t^{\mathrm{T}}\hat{T}_t}{t-1} = \mathrm{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_h\}, \tag{6}$$

其中: $\hat{\boldsymbol{\Lambda}}_t$是由前$h$个特征值组成的特征向量; $\hat{T}_t$是得分矩阵.

最后, 判断当SPE和T^2满足如下条件时, 认为新样本的过程变量存在显著异常[52]:

$$\mathrm{SPE} > \mathrm{SPE}_{\alpha_{\mathrm{pro}}}, \tag{7}$$

$$T^2 > T^2_{\alpha_{\mathrm{pro}}}, \tag{8}$$

其中, $\mathrm{SPE}_{\alpha_{\mathrm{pro}}}$和$T^2_{\alpha_{\mathrm{pro}}}$表示SPE和$T^2$的控制限, 通常依据建模样本设定, 其定义详见文献[56].

除 PCA 外, 近似线性依靠 (approximate linear dependence, ALD)条件也被用于判别新样本与建模样本的过程变量间的相似程度[57]. 该方法首先在新样本$\boldsymbol{d}_{t+1}$的原始空间或核空间中计算其与建模样本间的ALD值δ_{t+1}如下:

$$\delta_{t+1} = \min \|\sum_{i=1}^{t}\alpha_i\boldsymbol{d}_i - \boldsymbol{d}_{t+1}\|^2. \tag{9}$$

然后, 根据δ_{t+1}和阈值v的关系判断是否异常: 当$\delta_{t+1} \leqslant v$时, 认为新样本的过程变量发生变化; 当$\delta_{t+1} > v$时, 认为新样本的过程变量未发生变化. 其中阈值v根据实际工业系统对建模精度与建模速度的要求确定, 其可表述为如下的单目标优化问题[58]:

$$\max J = \gamma_1 J_{\mathrm{pred}}(v_{j_v}) + \gamma_2 J_{\mathrm{time}}(v_{j_v}), \tag{10}$$

$$\text{s.t.} \begin{cases} J_{\mathrm{pred\_low}} < J_{\mathrm{pred}}(v_{j_v}) < J_{\mathrm{pred\_high}}, \\ J_{\mathrm{time\_low}} < J_{\mathrm{time}}(v_{j_v}) < J_{\mathrm{time\_high}}, \\ 0 < \gamma_1, \gamma_2 < 1, \\ \gamma_1 + \gamma_2 = 1, \end{cases} \tag{11}$$

其中: $J_{\mathrm{pred}}(v_{j_v})$和$J_{\mathrm{time}}(v_{j_v})$是采用阈值$v_{j_v}$时的建模精度和建模速度; $J_{\mathrm{pred\_low}}$ 和 $J_{\mathrm{pred\_high}}$, $J_{\mathrm{time\_low}}$ 和 $J_{\mathrm{time_high}}$是实际工业系统可接受建模精度和建模速度的上下限; γ_1和γ_2是在建模精度和建模速度之间均衡的加权系数.

此外, 偏最小二乘(partial least squares, PLS)[59]、独立成分分析(independent component analysis, ICA)[60]、费舍尔判别分析 (fisher discriminant analysis, FDA)[61] 和子空间辅助方法 (subspace aided approach, SAP)[62]等传统多元统计方法及它们的改进版本[63]均被证明可有效检测过程变量是否异常. 其中,

PLS常用于多输出过程分析, ICA在非高斯分布的异常检测中表现良好.

现有研究中, 文献[64]采用PCA检测水泥回转窑运行过程状态, 并引入EWMA方法自适应调整PCA模型控制限阈值; 文献[65]针对乙烯裂解过程, 在PCA基础上结合基于知识的符号有向图(signed directed graph, SDG)推理方法, 实现检测变量变化的同时确定变化原因; 文献[66]采用ALD条件逐个分析待测样本的概念变化情况, 并将新概念样本用于PCA模型更新以使其获得自适应调整能力; 文献[67]面向传感器网络概念漂移现象, 根据子空间学习思想将PCA和基于角度优化的全局降维算法(angle optimized global embedding, AOGE)相结合, 以从多角度分析待测样本的主成分变化情况; 文献[68]采用统计矩与功率谱分别度量样本过程变量的均值、方差、偏度、峰度、幅度和频率变化等因素以表征漂移.

2) 距离度量策略.

该策略采用距离(欧式距离、马氏距离和余弦距离等)对样本过程变量间的相似关系进行量化, 特点是无需过程变量服从特定分布(高斯或非高斯分布等), 且漂移判别标准设置相对灵活, 因此已成为目前基于过程变量的漂移检测算法中最常见的一类方法[18].

现有研究中, 文献[69]较早为样本差异分析中距离函数的设计提供了指导, 其采用L1范数度量样本距离关系, 并结合Chernoff界和Vapnik-Chervonenkis维数确定距离簇变化程度; 文献[70]采用Hellinger距离检测渐进或突然的概念变化, 计算新旧样本中每个变量间的Hellinger距离, 并将所有变量距离的均值作为最终距离后计算其与预设基准距离的差异; 文献[71]结合距离度量与最近邻思想, 首先计算相邻样本块中各样本间的异构欧式距离, 然后根据最近邻样本的标签一致程度计算样本漂移度; 文献[72]将历史样本拆分为多个样本块, 并将每个历史样本块映射为不同的概念向量后进行聚类, 当新样本块到达时计算概念向量与历史样本聚类中心的距离差异以检测漂移; 文献[73]提出基于Kullback-Leibler距离的决策树分布检测模型; 文献[74]采用马氏距离和欧式距离互补的方式对样本过程变量的不同子空间进行度量, 根据预设差异指标指示概念变化; 文献[75]对样本聚类后, 通过比较相邻样本块的领域熵差值以检测漂移.

3) 假设检验策略.

常见的假设检验策略可分为参数检验与非参数检验, 前者需在样本总体分布信息已知情况下进行, 而后者不依赖样本总体分布. 常用的参数检验包括t检验和F检验, 分别观测样本总体均值和方差的相似性; 常用的非参数检验包括 Wilcoxon 检验、置换检验和Kolmogorov-Smirnov检验, 相应地分别观测样本秩和、均数和频数的相似性.

现有研究中, 文献[23]根据案例推理分类思想提出基于能力模型的检测法, 其构造样本间基于能力的经验距离并对该距离进行置换检验以检测漂移; 文献[76]提出基于重采样和t检验的多尺度检测法, 首先在训练集中提取具有典型概念特征的样本, 然后将这些样本组成规模较小的且具有多样概念的子集, 最后通过t检验比较该子集与待检测样本的总体均值差异以检测漂移; 文献[77]在多集理论基础上提出基于累计区域密度差异的检验方法, 该方法计算样本块中不同过程变量值的所占比例, 并通过Monte-Carlo置换检验判断相邻样本块中过程变量值的比例分布差异指示漂移.

3.1.3 基于综合因素的方法

综合因素法结合了基于难测参数测量误差与基于过程变量的方法, 相较单一检测方法可提供更全面检测信息, 因此该类方法被用于解决实际问题.

文献[78]在基于专家知识构建的模糊推理模型的基础上, 结合样本相对ALD值和相对测量误差值(relative prediction error, RPE)有效识别新概念样本, 文中表明该算法相较仅基于ALD和仅基于RPE的样本识别方法可详细反映样本漂移程度, 且能提高模型可解释性与测量精度. 相对ALD值的计算方式如下:

$$a_{t+1}=\frac{a_{t+1}^{\mathrm{abs}}}{\dfrac{\sum\limits_{i=1}^{t}a_i^{\mathrm{abs}}}{t}}, \tag{12}$$

其中: a_{t+1}^{abs}是新样本$\boldsymbol{d}_{t+1}$相对于训练集$S_{1,t}$的ALD绝对值[61]; a_i^{abs}表示训练集中第i个样本相对于其他所有$t-1$个样本的ALD值.

RPE值计算方式如下:

$$e_{t+1}=\frac{\left|\dfrac{\hat{y}_{t+1}-y_{t+1}}{y_{t+1}}\right|}{\dfrac{1}{t}\cdot\sum\limits_{i=1}^{t}\left|\dfrac{\hat{y}_i-y_i}{y_i}\right|}, \tag{13}$$

其中: $\hat{y}_{t+1}$和$\hat{y}_i$分别表示模型对新样本$\boldsymbol{d}_{t+1}$和训练样本$\boldsymbol{d}_i$的在线测量值.

算法中结合相对ALD值和RPE值后提出的模糊规则$F_{\mathrm{com}}(\cdot)$与判别准则JC_{t+1}的表达式分别为

$$u_{s(t+1)}=F_{\mathrm{com}}(a_{t+1},e_{t+1}), \tag{14}$$

$$JC_{t+1}=\begin{cases}1,\ u_{st+1}\geqslant\theta_{\mathrm{com}},\\ 0,\ u_{st+1}<\theta_{\mathrm{com}},\end{cases} \tag{15}$$

其中: θ_{com}为样本选择阈值. 当$JC_{t+1}=1$时, 表示新样本$\boldsymbol{d}_{t+1}$发生漂移, 否则认为样本正常.

文献[79]面向在线数据维护提出了名为P树的模型结构, 在监视模型测量性能变化的基础上结合PCA和Wilcoxon检验对样本的类分布与后验分布变化进行检测. 文中所提算法框架如图10所示.

根据图10, 可将综合因素法检测思路描述为: 1) 依据模型性能变化(测量误差与误差率等)检索异常样本; 2) 采用基于过程变量的方法分析异常样本分布变化情况; 3) 根据分析结果定义漂移变化指标实现漂移检测.

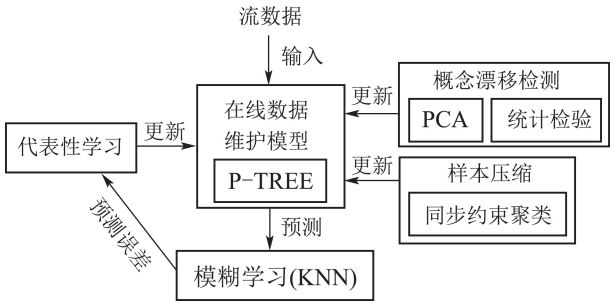


图 10 基于P树的在线数据维护框架

Fig. 10 P-tree-based online data maintenance framework

3.1.4 基于其他检测依据的方法

其它研究中, 文献[31]针对质量在线测量过程, 对新样本进行窗口检索后采用3种方式检测样本概念变化, 即模型均方测量误差、非参数U检验和观测均值分析; 文献[80]提出具有滑动窗口的符号回归集成模型, 首先根据模型测量误差触发样本变化检验, 然后计算新旧样本的平方皮尔逊相关系数, 最后判断当相关系数大于预设阈值时认为概念变化; 文献[81]提出双准则主动学习算法, 首先建立逻辑回归模型监测模型性能变化, 然后对样本聚类并结合贝叶斯思想判别样本块间概率密度差异, 最后综合上述变化确认样本漂移情况; 文献[82]提出层次假设检验框架: 第1层监视分类器的在线错误率, 第2层采用置换检验分析样本过程变量的相似性; 文献[83]针对三聚氰胺树脂生产过程的漂移现象, 在具有滑动窗口的集成PLS模型中引入Page-Hinkly检测以检索漂移样本; 文献[84]基于模型解释思想, 首先计算样本块中各过程变量的贡献度, 然后采用欧式距离度量不同样本块中变量贡献度差异, 最后通过Page-Hinkly检测判断差异是否显著; 文献[85]基于DDM思想对异常样本进行检索, 并通过监视异常样本集中马尔可夫链随时间的转变概率变化表征漂移.

除上述有监督方法外, 现有研究还针对实际问题中难测参数真值难以获取的情况提出了半监督综合检测方法, 如: 文献[86]提出基于边际密度的半监督检测方法, 采用分类器边际密度作为无监督漂移指标检索待标注样本, 在样本获取标注后再基于模型性能变化进行第二次漂移确认; 文献[16]在Page-Hinkly检测基础上加入下降指示器和衰减因子并采用Hoeffding界定义检测阈值, 依据单次主动学习思想定义半监督性能指标, 实验表明该方法具有接近有监督方法的检测效率, 其样本真值需求量仅为后者的20%.

3.2 检测对象视角

现有漂移检测研究工作中暂未有明确的针对单样本与多样本的算法描述, 但在部分文献中存在与该工作类似的研究与讨论, 主要集中在样本窗口大小的选择问题.

样本窗口的目的是依据样本数量或时间步长将部分流数据组织为样本块后进行漂移分析, 采用该策略的原因是部分学者认为单个样本难以携带足够信息推断总体分布[87], 因此有必要将数据组织为有意义的模式或知识[88]. 目前, 样本窗口设置方式已成为漂移检测研究的重点之一, 较为典型的是基于固定窗口[42]、滑动窗口[23]和多窗口[45]的检测策略. 此外, 文献[89]指出, 大尺寸窗口虽可覆盖更多新概念样本但会导致检测不及时, 小尺寸窗口虽可保证检测及时性但易增大计算消耗, 因此该文提出自适应窗口, 即窗口大小可依据概念变化速度与幅度实时调节.

实际上无论以何种方式划分样本窗口, 均无法避免的问题是: 在样本块组织过程中, 可能丢失关键漂移时刻信息或由于无法及时更新模型导致测量精度持续恶化. 因此, 有学者认为逐样本检测方式可显著提升检测的时效性, 即单个样本可在一定程度上表征漂移现象[89].

基于上述工作, 本文提出针对单样本和多样本的算法检测框架, 如图11所示.

图11所示漂移检测框架的依据为: 实际工业过程中, 部分检测任务侧重对过程反应变化规律进行探索, 如烟气污染物的排放浓度变化趋势观察[37]和生成物质量监测实现锅炉优化[31]等, 因此需采用样本窗口方式获得更加精确的变化关系, 此时由检索过程造成的检测延时通常可被接受. 而在另一些检测任务中, 概念变化通常预示生产过程意外改变, 此时若无法及时检测与处理漂移可能引起更大工程事故, 因此需进行逐样本分析以及时杜绝潜在运行风险. 综上, 以单样本与多样本视角对现有研究进行讨论可有效区分各检测方式在工业应用中的及时性与准确性, 有助于为不同建模任务选择合适的漂移检测算法.

3.2.1 单样本漂移检测

文献[91] 提出基于测量误差限 (prediction error band, PEB)的单样本检测算法, 其误差e_k采用下式计算:

$$e_k = y_k - f(\boldsymbol{d}_k),\ k = t+1, t+2, \cdots, \tag{16}$$

其中: y_k是样本$\boldsymbol{d}_k$对应的难测参数真值; $f(\boldsymbol{d}_k)$为模型测量函数. 当PEB满足如下条件时, 认为当前样本发生漂移

$$e_k > \text{Rule}(\{\delta_k^1, \delta_k^2, \cdots\}), \tag{17}$$

其中: $\delta_k^1, \delta_k^2, \cdots$表示依据先验知识设定的不同阈值; $\text{Rule}(\cdot)$表示依据经验设定的判断规则.

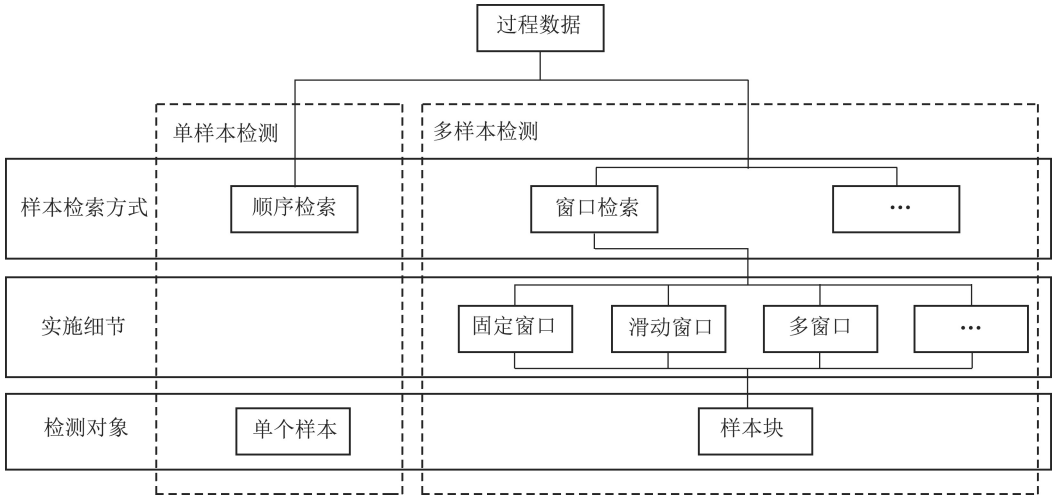


图 11　针对单样本与多样本的漂移检测框架

Fig. 11　Detection framework for single and multiple samples

其它研究中, 文献[49]针对两个相邻漂移样本之间的样本数量进行分析; 文献[39]通过观察两个样本所对应的模型最优线性间隔进行检测; 文献[50–51]分别依据模型对每个样本的测量错误可能性和测量错误率变化; 文献[67]按采集顺序对待测样本的主成分变化情况进行多角度分析; 文献[74]针对单个样本中过程变量的不同空间进行距离度量; 文献[66, 79]分别在PCA和RPE基础上结合ALD条件实现逐样本分析.

3.2.2　多样本漂移检测

前文所述研究中, 文献[43–44]采用样本加权方式将待测样本组织为样本块进行分析; 文献[45–48]通过监测样本窗口内的模型性能变化确认漂移; 文献[64–65]采用控制图方式监控样本块概念变化; 文献[69–71]分析了不同样本块间的距离变化关系; 文献[72]采用样本聚类方式分析; 文献[73]对两个样本块之间的相对熵差异进行检测; 文献[23, 76–77]均针对样本块所携带的分布信息进行假设检验分析; 文献[80–84]采用综合型方法对样本窗口内分布变化进行检测; 文献[16, 87]所提的半监督检测方法对异常样本集合请求标注后进行二次检验以确认漂移.

3.3　检测方法汇总

根据上述讨论, 本节结合检测依据、检测对象和具体检测方式对具有代表性的检测方法进行归纳, 结果如表2所示.

4　讨论与分析

4.1　现有检测方法特点

现有各类检测方法特点总结如表3所示. 根据表3, 可将各方法特点详细描述为:

1) 基于难测参数测量误差的检测方法: 该类方法观测概念漂移产生的最直接变化, 即模型输入输出关系变化导致的模型测量误差显著升高, 因此其检测速度相对较快, 能及时反映漂移可能发生的时间与位置, 且该过程实现较为简便, 易于理解. 但该类方法检测效率较依赖模型性能与其构建方式, 且由于难以详细反映样本分布变化信息, 可能导致模型长期处于频繁的更新过程从而使测量精度不稳定, 同时该类方法无法在难测参数真值难以获得的情况下使用.

2) 基于过程变量的检测方法: 该类方法检验样本过程变量的显著变化, 可较全面反映变量变化情况, 且该过程不依赖特定模型与难测参数真值. 但有时过程变量变化无法充分说明样本概念分布发生变化, 以同时包含x_1, x_2和x_3的三维过程变量集$\boldsymbol{x}_t = [1, 1, 1]$与$\boldsymbol{x}_{t+1} = [1, 3, 9]$为例, 可观察到两个变量集中的变量数值与其变化幅度有明显差异, 但当$\boldsymbol{x}_t$和$\boldsymbol{x}_{t+1}$的对应样本均满足简单线性映射关系$f(x) = x_1 + \alpha x_2 + \beta x_3 (\alpha, \beta \to 0)$时, 变量间的数值差异难以准确反映样本的概念变化, 该情况下对模型的更新可能是不必要的.

3) 基于综合因素的检测方法: 该类方法可通过多视角分析概念变化情况以得到较为准确的漂移检测结果, 在一定程度弥补了上述方法的缺点, 但也因此要求不同检测策略之间具有合理的触发机制与科学的资源分配机制, 否则任一策略偏差均可能导致算法检测效率低下甚至失效, 需在方法构建时充分考虑实际应用环境以及各策略适用性.

4) 针对单样本的检测方法: 目前针对单样本的研究较少, 原因是单个样本携带的分布变化信息相较多样本更难评估, 但现有研究方法证明针对单样本的漂移检测是可行的, 且该类方法所具有的时效性对于分析工业中过程环境变化及预估漂移程度与规模有重要意义.

5) 针对多样本的检测方法: 样本块通常携有丰富变化信息, 现有研究表明该类方法具有更高检测精度, 但其需要更长的检索与检测时间且在此期间内难以维持模型性能, 同时现有研究中多数方法未能对样本漂移程度进行区分.

表 2 多视角下的算法特点总结

Table 2 Summary of algorithm characteristics under multiple angles

检测依据	检测对象	文献编号	发表年份	检测方式
基于难测参数测量误差	单样本	[49]	2006	相邻漂移样本间的样本数量+定义漂移级别
		[39]	2009	测量误差+线性间隔相似度
		[91]	2010	测量误差+经验规则
		[51]	2012	测量错误率+预设阈值
		[50]	2015	测量错误的概率+ 累积和控制图
	多样本	[42]	2004	测量错误率+定义漂移级别
		[45]	2007	测量误差+等比例统计检验
		[48]	2008	测量误差+预设阈值
		[44]	2014	真值与误差均值+Hoeffding不等式
		[46]	2016	测量错误的概率+Hoeffding不等式
		[47]	2019	测量误差+预设阈值
		[52]	2019	模型输出权重
		[43]	2020	测量错误率+Page-Hinkley检验
基于过程变量	单样本	[66]	2012	PCA+ALD
		[74]	2015	欧式距离+马氏距离
		[67]	2017	PCA+AOGE
	多样本	[69]	2004	L1范数
		[73]	2006	Kullback-Leibler 距离+决策树
		[72]	2010	聚类距离
		[70]	2011	Hellinger距离
		[71]	2011	异构欧式距离+最近邻
		[23]	2014	能力经验距离+置换检验
		[64]	2017	PCA+EWMA
		[76]	2018	重采样+t检验
		[77]	2018	Monte-Carlo置换检验
		[65]	2018	PCA+SDG
		[68]	2019	统计矩+功率谱
		[75]	2020	领域熵
基于综合因素	单样本	[79]	2016	ALD+RPE
	多样本	[32]	2009	模型均方测量误差+ U检验+原始样本统计分析
		[81]	2016	模型性能+样本概率密度差异
		[16]	2016	Page-Hinkley检验+ 单次主动学习
		[80]	2017	测量误差+平方皮尔逊相关系数
		[78]	2017	测量误差+PCA+Wilcoxon检验
		[86]	2017	边际密度+模型性能
		[83]	2018	集成PLS模型+Page-Hinkley检验
		[84]	2018	过程变量贡献度+Page-Hinkley检验
		[82]	2019	在线错误率+置换检验
		[85]	2019	测量误差+马尔可夫链转变概率

4.2 相似研究

目前工业领域中的部分研究虽未指明概念漂移问题, 但其研究思路与技术路线均与漂移检测具有相似之处. 为对后续漂移检测工作提供不同借鉴方案, 此处对部分相似研究进行整理, 如下所示:

与基于难测参数测量误差视角相似的方法: 文献[92]采用自回归滑动平均模型应对动态研磨过程中由环境变化或传感器故障引起的软测量模型性能下降; 文献[93]采用有限冲激响应和SVM分析过程变量的动态与静态关系, 并以此构建动态软测量测量模型; 文献[94]针对动态工业过程, 采用时间差分模型减弱由机械元件老化导致的模型测量精度下降. 在最近的研究中: 文献[95]提出基于长短期记忆神经网络的动

态测量维护框架, 通过比较设备在新时刻与历史时刻的性能差异估算当前设备故障概率; 文献[96]采用自适应标准化的局部窗口对新样本检索后, 基于包含双向自编码器的深度神经网络模型分析窗口内样本的分布差异; 文献[97]结合样本时滞、动态时间和测量误差提出基于最小二乘SVM的氮氧化物浓度实时动态测量模型.

与基于过程变量视角相似的方法: 文献[98]基于趋势分析思想, 采用动态特征同步算法对过程变量的变化趋势量化, 并通过与历史趋势进行相似性比较以确认连续生产过程中的工况切换状态; 文献[99]基于子空间辨识思想, 采用滑动窗口检索新样本后计算窗口内样本子空间的马尔可夫参数向量, 通过比较不同窗口内样本参数向量的均值与方差差异判断模型是否失配; 文献[100]采用基于概率的慢特征分析方法提取过程变量的潜在变化, 并以此提高软测量模型在动态工业环境的测量精度; 文献[101]指出时变工业过程中具有影响力的过程变量通常变化缓慢, 因此提出慢特征分析方法对时间序列数据中不同过程变量变化情况分析. 在最近的研究中: 文献[102]通过聚类获取时序数据变化特点, 并根据数据状态趋势检测过程异常; 文献[103]面向多模态化工过程, 通过结合迁移学习与深度神经网络, 使工业测量模型能快速检测并适应源域与目标域间的数据分布差异; 文献[104]将几何字典学习思想用于工业过程监控, 通过K近邻模型对历史样本中过程变量的几何特征进行编码, 进而在字典学习框架下分析新旧样本间的信息差异; 文献[105]采用欧式距离和时间加权距离度量样本在空间与时间尺度中的相似性, 并结合支持向量数据描述(support vector data description, SVDD)建立过程监控模型; 文献[106]结合PCA与SVDD处理动态、非线性和非高斯分布的故障检测问题.

表 3 各类漂移检测方法特点

Table 3 Characteristics of various drift detection methods

方法类型	方法优点	方法缺点
基于难测参数测量误差	检测速度快, 计算复杂度低	难以详细反映分布变化信息
基于过程变量	可全面检测变量变化信息	有时无法准确反映漂移
基于综合因素	检测效率高, 结果可信度强	计算资源消耗较大, 各检验方法的触发阈值设置较为复杂
针对单样本	模型快速适应过程环境变化	难以对漂移形式与程度进行推断
针对多样本	详细分析过程变化情况, 有针对性更新模型	较长的检测周期可能错过关键漂移信息, 且模型更新不及时

与综合因素视角相似的方法: 文献[107]面向时变化工过程提出具有定时功能的模糊Petri网算法, 在获取过程动态特性的同时监测工况异常变化及其发生时间; 文献[108]面向非线性系统, 采用包含摄动信号与模型残差的互信息矩阵量化多变量系统中的模型失配程度. 在最近的研究中: 文献[109]针对工业过程中老化与时变特性提出基于动态多属性决策的控制性能评价方法, 通过计算超调量、非线性、输出方差和控制阀黏滞指标权重变化获得过程动态评价基准; 文献[110]从设备历史故障中提取受故障影响最大的过程变量, 并在运行过程中观测上述变量的综合变化幅度判断设备故障状态; 文献[111]提出基于随机森林的实时控制图, 在监视模型性能变化基础上结合过程变量重要性实现故障检测与故障原因识别; 文献[112]结合深度信念网络和SVDD提出分层表示学习方法, 在分析模型测量误差变化的同时融入贝叶斯诊断框架表征过程变量中的故障信息.

4.3 工业漂移检测研究难点

结合以上分析, 本文对工业过程中概念漂移检测的部分研究难点总结如下:

1) 难测参数的真值获取难: 工业过程中由于技术局限与经济性考虑, 通常无法为难测参数提供足够的真值, 因此要求检测方法能在样本少量标记的情况下对样本分布变化做出有效分析. 为此, 基于无监督或半监督的检测研究是有必要的[16], 但无监督方法在变量变化情况较为复杂时可能无法保证检测结果准确性, 而目前针对半监督方法的研究相对缺乏.

针对真值无法及时标注问题, 面向分类任务, 文献[113]采用神经网络测量无标注样本的最大类别概率以生成样本伪标签; 文献[114]采用SVM分析同一样本在不同类别下对模型决策边界的影响程度从而推断无标注样本标签. 针对半监督学习, 文献[115]提出基于协同学习的半监督回归策略, 文中建立不同的K近邻测量模型并基于测量一致性输出置信度最高的

样本测量值; 文献[116]面向多媒体信息处理领域提出基于分歧的半监督主动学习方法. 上述工作均为半监督漂移检测方法设计提供了支撑, 但如何将其应用于连续型变量伪真值生成及具有概念变化的工业回归任务中仍需深入研究.

2) 样本的期望分布获取难: 现有工作多围绕分类任务进行, 因此样本概念通常可根据标签或类别等具有明显区分性质的信息划分. 但实际工业过程多为回归任务, 此时二项分布、Hoffeding不等式和分类器决策边界等常用的阈值界定方法难以直接应用.

在基于分布的虚拟样本生成研究中, 文献[117]基于信息扩散准测提出整体趋势扩散技术, 通过监视数据变化趋势估计其合理分布范围; 文献[118]基于模糊理论提出扩散神经网络, 以观测样本视为模糊正态分布中心并采用对称的扩散函数获取其理论分布范围. 在基于特征的迁移学习研究中, 文献[119]基于降维思想, 采用再生核希尔伯特空间度量样本分布差异; 文献[120]采用协同聚类获取源域数据的特征表示. 上述工作均有助于提取工业过程变量的潜在概念, 但如何将其与漂移检测技术结合并定义过程变量的概念变化阈值仍需结合实际工业过程的特点进行讨论.

3) 噪声等异常数据区分难: 实际工业系统结构较为复杂, 各监测环节扰动均会为样本采集过程混入噪声等异常数据, 这些数据同样会导致模型性能改变从而易与漂移现象相混淆, 显然, 采用噪声样本对模型进行更新是无意义的.

现有研究中, 文献[121]在集成软测量模型中采用基于分区、层次和密度的聚类方法去除噪声建模样本; 文献[122]面向分类任务, 提出基于k近邻感知的标签噪声过滤算法; 文献[123]通过集成投票策略评估噪声得分以确认噪声样本. 上述工作均为工业过程中异常数据辨识提供了思路, 但如何将其与漂移样本合理区分仍需进一步分析.

5 总结与展望

本文介绍了当前工业中的概念漂移现象, 总结了概念漂移的定义、形式以及现有的部分研究工作, 分析了各类检测方法的特点与针对工业领域的部分难点, 旨在为工业过程中概念漂移检测算法的设计与应用提供指导.

结合文中分析结果, 在此提出对未来工作的研究方向与建议:

1) 加强半监督检测算法研究: 目前半监督检测方法相对较少, 该类方法在难测参数的真值难以获得时具有较强的研究意义. 因此, 在实际算法设计时可进一步结合虚拟样本生成和小样本分析等技术以充分利用已有真值样本的分布信息, 同时建立可靠的无监督检测策略进行异常样本筛选.

2) 加强单样本检测算法研究: 现有工作中针对单样本的算法较为缺乏, 由于单个样本所携带分布信息有限, 未来应结合基于综合因素的方法从样本输出空间、变量空间和变量子空间等方面进行多角度并行分析, 同时引入多步测量与变化率分析等技术思想, 实现对未来发生漂移的可能性、时间和程度等信息进行预判, 以充分发挥单样本检测的时效性特点.

3) 加强多样本检测算法研究: 现有多样本检测算法可初步保证检测准确性, 在此基础上未来应加强对漂移现象的理论研究. 如, 建立漂移变化指标以量化历史样本的漂移速度与新样本漂移幅度, 从而衡量不同形式漂移对软测量模型的影响程度并以此指示模型在当前环境中的更新方式与必要性, 实现在加强模型适应性的同时避免模型因频繁更新导致的计算资源消耗与短期性能下降.

4) 加强与实际工业过程联系: 算法设计时除对检测功能进行完善外, 仍需考虑在工业运行过程中的适用性. 如, 在算法中引入噪声识别等数据预处理技术以应对过程数据的复杂性, 同时结合专家知识与工艺机理充分了解运行过程中的易变工况, 并建立多模式集成或自适应调整的漂移检测模型, 提高工业环境中的漂移检测效率.

此外, 本文仅针对概念漂移的检测方式进行综述介绍, 其它研究内容如漂移理解、漂移适应性模型的构建与更新策略等仍需进一步讨论.

参考文献:

[1] TANG J, QIAO J F, LIU Z, et al. Mechanism characteristic analysis and soft measuring method review for ball mill load based on mechanical vibration and acoustic signals in the grinding process. *Minerals Engineering*, 2018, 128: 294 – 311.

[2] TANG Jian, XIA Heng, QIAO Junfei, et al. Deep ensemble forest regression modeling method with its application research. *Journal of Beijing University of Technology*, 2020, 1 – 13[2020-07-30]. http://kns.cnki.net/kcms/detail/11.2286.T.20200723.1048.002.html.
(汤健, 夏恒, 乔俊飞, 等. 深度集成森林回归建模方法及应用研究. 北京工业大学学报, 2020, 1 – 13[2020-07-30]. http://kns.cnki.net/kcms/detail/11.2286.T.20200723.1048.002.html.)

[3] YIN Erxin, DONG Ze, CAO Xiaolin. Dynamic data-driven modeling of industrial systems based on state optimization. *Computer Integrated Manufacturing Systems*, 2018, (5): 133 – 136, 141.
(尹二新, 董泽, 曹晓玲. 基于状态寻优的工业系统动态数据驱动建模. 计算机仿真, 2018, (5): 133 – 136, 141.)

[4] ZLIOBAITE I. Learning under concept drift: an overview. *Computer Science*, 2010, 4(2): 107 – 194.

[5] SHALEV-SHWARTZ S. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 2011, 4(2): 107 – 194.

[6] XIAO L. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 2010, 11: 2543 – 2596.

[7] ZHI J L, YUAN X L, FENG W, et al. Efficient and accelerated online learning for sparse group LASSO. *Proceedings of IEEE International Conference on Data Mining Workshop*. Shenzhen: IEEE, 2014: 1171 – 1177.

[8] LI Zhijie, LI Yuanxiang, WANG Feng, et al. Overview of online learning algorithms for big data analysis. *Journal of Computer Research and Development*, 2015, 52(8): 1707 – 1721.
(李志杰, 李元香, 王峰, 等. 面向大数据分析的在线学习算法综述. 计算机研究与发展, 2015, 52(8): 1707 – 1721.)

[9] WIDMER G, KUBAT M. Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 1996, 23(1): 69 – 101.

[10] MINKU L L, WHITE A P, YAO X. The impact of diversity on online ensemble learning in the presence of concept drift. *IEEE Transactions on Knowledge and Data Engineering*, 2009, 22(5): 730 – 742.

[11] ELWELL R, POLIKAR R. Incremental learning of concept drift in nonstationary environments. *IEEE Transactions on Neural Networks*, 2011, 22(10): 1517 – 1531.

[12] MORENO-TORRES J G, RAEDER T, ALAIZ-RODRíGUEZ R O, et al. A unifying view on dataset shift in classification. *Pattern Recognition*, 2012, 45(1): 521 – 530.

[13] WU X, LI P, HU X. Learning from concept drifting data streams with unlabeled data. *Neurocomputing*, 2012, 92: 145 – 155.

[14] ZHU X, ZHANG P, LIN X, et al. Active learning from stream data using optimal weight classifier ensemble. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2010, 40(6): 1607 – 1621.

[15] KHAMASSI I, SAYED-MOUCHAWEH M, HAMMAMI M, et al. Discussion and review on evolving data streams and concept drift adapting. *Evolving Systems*, 2018, 9(1): 1 – 23.

[16] LUGHOFER E, WEIGL E, HEIDL W, et al. Recognizing input space and target concept drifts in data streams with scarcely labeled and unlabelled instances. *Information Sciences*, 2016, 355(C): 127 – 151.

[17] NUNEZ M, FIDALGO R, MORALES R. Learning in environments with unknown dynamics: Towards more robust concept learners. *Journal of Machine Learning Research*, 2007, 8: 2595 – 2628.

[18] KOTLER J, MALOOF M. Dynamic weighted majority: A new ensemble method for tracking concept drift. *Proceedings of IEEE International Conference on Data Mining*. Melbourne: IEEE, 2003: 123 – 130.

[19] WEN Yimin, QIANG Baohua, FAN Zhigang. A survey of the classification of data streams with concept drift. *CAAI Transactions on Intelligent Systems*, 2013, 8(2): 95 – 104.
(文益民, 强保华, 范志刚. 概念漂移数据流分类研究综述. 智能系统学报, 2013, 8(2): 95 – 104.)

[20] LU J, LIU A, DONG F, et al. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 2018, 31(12): 2346 – 2363.

[21] RAMAKRISHNA B, RAO S K M. Concept drift detection in data stream mining: The review of contemporary literature. *Global Journal of Computer Science and Technology*, 2017, 17(2): 1 – 9.

[22] ZLIOBAITE I, PECHENIZKIY M, GAMA J. An overview of concept drift applications. *Big Data Analysis: New Algorithms for a New Society*, Berlin: Springer, 2016: 91 – 114.

[23] LU N, ZHANG G, LU J. Concept drift detection via competence models. *Artificial Intelligence*, 2014, 209: 11 – 28.

[24] SCHLIMMER J C, GRANGER R H. Incremental learning from noisy data. *Machine Learning*, 1986, 1(3): 317 – 354.

[25] LIU A, SONG Y, ZHANG G, et al. Regional concept drift detection and density synchronized drift adaptation. *Proceedings of International Joint Conference on Artificial Intelligence*. Melbourne: IGCAI, 2017: 2280 – 2286.

[26] YANG Junzhi. Full analysis on accuracy and related terms. *Science of Surveying and Mapping*, 2011, 36(1): 75 – 76.
(杨俊志. 测量准确度及相关术语辨析. 测绘科学, 2011, 36(1): 75 – 76.)

[27] WIDMER G, KUBAT M. Effective learning in dynamic environments by explicit context tracking. *Proceedings of European Conference on Machine Learning*. Berlin: Springer, 1993: 227 – 243.

[28] KELLY M G, HAND D J, ADAMS N M. The impact of changing populations on classifier performance. *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Diego: ACM, 1999: 367 – 371.

[29] KUNCHEVA L I. Classifier ensembles for changing environments. *Proceedings of the Fifth Workshop on Multiple Classifier Systems*. Cagliari: Springer, 2004: 1 – 15.

[30] KADLEC P, GABRYS B, STRANDT S. Data-driven soft sensors in the process industry. *Computers & Chemical Engineering*, 2009, 33(4): 795 – 814.

[31] YUAN Xiaofeng, GE Zhiqiang, SONG Zhihuan. Adaptive soft sensor based on time difference model and locally weighted partial least squares regression. *CIESC Journal*, 2016, 67(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, 67(3): 724 – 728.)

[32] BAKKER J, PECHENIZKIY M, ZLIOBAITE I, et al. Handling outliers and concept drift in online mass flow prediction in CFB boilers. *Proceedings of the Third International Workshop on Knowledge Discovery from Sensor Data*. New York: ACM, 2009: 13 – 22.

[33] ZENISEK J, HOLZINGER F, AFFENZELLER M. Machine learning based concept drift detection for predictive maintenance. *Computers & Industrial Engineering*, 2019, 137: 106031.

[34] ROSENTHAL F, VOLK P B, HAHMANN M, et al. Drift-aware ensemble regression. *Proceedings of International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Leipzig: Springer, 2009: 221 – 235.

[35] SHANG L, LIU J, ZHANG Y, et al. Efficient recursive canonical variate analysis approach for monitoring time-varying processes. *Journal of Chemometrics*, 2017, 31(1): e2858.

[36] ANDREWS J D, MOSS T R. *Reliability and Risk Assessment*. 2nd Edition , USA: Wiley-Blackwell, 2002: 1 – 540.

[37] QIAO Junfei, GUO Zihao, 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.)

[38] ROWCLIFFE W. *Learning in the Presence of Sudden Concept Drift and Measurement Drift*. Ames, Iowa: Iowa State University, 2013.

[39] DRIES A, RUCKERT U. Adaptive concept drift detection. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 2009, 2(5/6): 311 – 327.

[40] KLINKENBERG R. Learning drifting concepts: Example selection vs example weighting. *Intelligent Data Analysis*, 2004, 8(3): 281 – 300.

[41] LUGHOFER E. Hybrid active learning (HAL) for reducing the annotation efforts of operators in classification systems. *Pattern Recognit*, 2012, 45(2): 884 – 896.

[42] GAMA J, MEDAS P, CASTILLO G, et al. Learning with drift detection. *Proceedings of Brazilian Symposium on Artificial Intelligence*. Sao Luis: Springer, 2004: 286 – 295.

[43] MAHDI O A, PARDEDE E, ALI N, et al. Diversity measure as a new drift detection method in data streaming. *Knowledge-Based Systems*, 2020, 191: 105227.

[44] FRIAS-BLANCO I, DEL CAMPO-AVILA J, RAMOS-JIMENEZ G, et al. Online and non-parametric drift detection methods based on Hoeffding's bounds. *IEEE Transactions on Knowledge and Data Engineering*, 2014, 27(3): 810 – 823.

[45] NISHIDA K, YAMAUCHI K. Detecting concept drift using statistical testing. *Proceedings of International Conference on Discovery Science*. Sendai: Springer, 2007: 264 – 269.

[46] PESARANGHADER A, VIKTOR H L. Fast hoeffding drift detection method for evolving data streams. *Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Riva del Garda: Springer, 2016: 96 – 111.

[47] LU Y, CHEUNG Y M, TANG Y Y. Adaptive chunk-based dynamic weighted majority for imbalanced data streams with concept drift. *IEEE Transactions on Neural Networks and Learning Systems (Early Access)*, 2019: 1 – 15.

[48] BACH S H, MALOOF M A. Paired learners for concept drift. *Proceedings of the Eighth IEEE International Conference on Data Mining*. Pisa: IEEE, 2008: 23 – 32.

[49] BAENA-GARCIA M, DEL CAMPO-áVILA J, FIDALGO R, et al. Early drift detection method. *Proceedings of the Fourth International Workshop on Knowledge Discovery from Data Streams*. Berlin, Germany: IEEE, 2006, 6: 77 – 86.

[50] MARTINEZ-REGO D, FERNáNDEZ-FRANCOS D, FONTENLA-ROMERO O, et al. Stream change detection via passive-aggressive classification and Bernoulli CUSUM. *Information Sciences*, 2015, 305: 130 – 145.

[51] ROSS G J, ADAMS N M, TASOULIS D K, et al. Exponentially weighted moving average charts for detecting concept drift. *Pattern Recognition Letters*, 2012, 33(2): 191 – 198.

[52] YANG Z, AL-DAHIDI S, BARALDI P, et al. A novel concept drift detection method for incremental learning in nonstationary environments. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, 31(1): 309 – 320.

[53] DING S, ZHANG P, DING E, et al. On the application of PCA technique to fault diagnosis. *Tsinghua Science and Technology*, 2010, 15(2): 138 – 144.

[54] YUE H H, QIN S J. Reconstruction-based fault identification using a combined index. *Industrial & Engineering Chemistry Research*, 2001, 40(20): 4403 – 4414.

[55] LIU J. On-line soft sensor for polyethylene process with multiple production grades. *Control Engineering Practice*, 2007, 15(7): 769 – 778.

[56] TANG J, YU W, CHAI T Y, et al. 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.

[57] ENGEL Y, MANNOR S, MEIR R. The kernel recursive least-squares algorithm. *IEEE Transactions on Signal Processing*, 2004, 52(8): 2275 – 2285.

[58] TANG J, CHAI T Y, ZHAO L J. On-line KPLS algorithm with application to ensemble modeling parameters of mill load. *Acta Automatica Sinica*, 2013, 39(5): 471 – 486.

[59] KHAN A A, MOYNE J R, DAWN M. Virtual metrology and feedback control for semiconductor manufacturing processes using recursive partial least squares. *Journal of Process Control*, 2008, 18(10): 961 – 974.

[60] GIROLAMI M. *Self-organising Neural Networks*. London: Springer, 1999: 47 – 75.

[61] DUDA R O, HART P E, STORK D G. *Pattern Classification*. USA: John Wiley & Sons, 2012: 1 – 654.

[62] FAVOREEL W, DE M B, VAN O P. Subspace state space system identification for industrial processes. *Journal of Process Control*, 2000, 10(2-3): 149 – 155.

[63] YIN S, DING S X, HAGHANI A, et al. A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. *Journal of Process Control*, 2012, 22(9): 1567 – 1581.

[64] BAKDI A, KOUADRI A, BENSMAIL A. Fault detection and diagnosis in a cement rotary kiln using PCA with EWMA-based adaptive threshold monitoring scheme. *Control Engineering Practice*, 2017, 66: 64 – 75.

[65] HAN X, TIAN S, ROMAGNOLI J A, et al. PCA-SDG based process monitoring and fault diagnosis: Application to an industrial pyrolysis furnace. *IFAC-PapersOnLine*, 2018, 51(18): 482 – 487.

[66] TANG J, YU W, CHAI T Y, et al. On-line principal component analysis with application to process modeling. *Neurocomputing*, 2012, 82: 167 – 178.

[67] LIU S, FENG L, WU J, et al. 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.

[68] MELLO R F, VAZ Y, GROSSI C H, et al. On learning guarantees to unsupervised concept drift detection on data streams. *Expert Systems with Applications*, 2019, 117: 90 – 102.

[69] KIFER D, BEN-DAVID S, GEHRKE J. Detecting change in data streams. *Proceedings of the 30th International Conference on Very Large Data Bases*. Toronto: ACM, 2004, 4: 180 – 191.

[70] DITZLER G, POLIKAR R. Hellinger distance based drift detection for nonstationary environments. *Proceedings of IEEE Symposium on Computational Intelligence in Dynamic and Uncertain Environments*. Paris: IEEE, 2011: 41 – 48.

[71] SOBHANI P, BEIGY H. New drift detection method for data streams. *Proceedings of International Conference on Adaptive and Intelligent Systems*. Klagenfurt: IEEE, 2011: 88 – 97.

[72] KATAKIS I, TSOUMAKAS G, VLAHAVAS I. Tracking recurring contexts using ensemble classifiers: an application to email filtering. *Knowledge and Information Systems*, 2010, 22(3): 371 – 391.

[73] DASU T, KRISHNAN S, VENKATASUBRAMANIAN S, et al. An information-theoretic approach to detecting changes in multidimensional data streams. *Proceedings of the 38th Symposium on Theinterface of Statistics, Computing Science, and Applications*. Pasadena, CA: IEEE, 2006: 1 – 24.

[74] 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.

[75] XU S, FENG L, LIU S, et al. Self-adaption neighborhood density clustering method for mixed data stream with concept drift. *Engineering Applications of Artificial Intelligence*, 2020, 89: 103451.

[76] WANG X S, KANG Q, ZHOU M C, et al. A multiscale concept drift detection method for learning from data streams. *Proceedings of the 14th International Conference on Automation Science and Engineering*. Munich: IEEE, 2018: 786 – 790.

[77] LIU A, LU J, LIU F, et al. Accumulating regional density dissimilarity for concept drift detection in data streams. *Pattern Recognition*, 2018, 76: 256 – 272.

[78] SHAO J, HUANG F, YANG Q, et al. Robust prototype-based learning on data streams. *IEEE Transactions on Knowledge and Data Engineering*, 2017, 30(5): 978 – 991.

[79] TANG J, CHAI T Y, LIU Z, et al. Adaptive ensemble modelling approach based on updating sample intelligent identification. *Acta Automatica Sinica*, 2016, 42(7): 1040 – 1052.

[80] ZENISEK J, AFFENZELLER M, WOLFARTSBERGER J, et al. Sliding window symbolic regression for predictive maintenance using model ensembles. *Proceedings of International Conference on Computer Aided Systems Theory*. Las Palmas de Gran Canaria: Springer, 2017: 481 – 488.

[81] MOHAMAD S, BOUCHACHIA A, SAYED-MOUCHAWEH M. A bi-criteria active learning algorithm for dynamic data streams. *IEEE Transactions on Neural Networks and Learning Systems*, 2016, 29(1): 74 – 86.

[82] YU S, ABRAHAM Z, WANG H, et al. Concept drift detection and adaptation with hierarchical hypothesis testing. *Journal of the Franklin Institute*, 2019, 356(5): 3187 – 3215.

[83] NIKZAD-LANGERODI R, LUGHOFER E, CERNUDA C, et al. Calibration model maintenance in melamine resin production: Integrating drift detection, smart sample selection and model adaptation. *Analytica Chimica Acta*, 2018, 1013: 1 – 12.

[84] DEMSAR J, BOSNIC Z. Detecting concept drift in data streams using model explanation. *Expert Systems with Applications*, 2018, 92: 546 – 559.

[85] ROVERI M. Learning discrete-time markov chains under concept drift. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, 30(9): 2570 – 2582.

[86] SETHI T S, KANTARDZIC M. On the reliable detection of concept drift from streaming unlabeled data. *Expert Systems with Applications*, 2017, 82: 77 – 99.

[87] LU N, LU J, ZHANG G, et al. A concept drift-tolerant case-base editing technique. *Artificial Intelligence*, 2016, 230: 108 – 133.

[88] RAMIREZ-GALLEGO S, KRAWCZYK B, GARCIA S, et al. A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing*, 2017, 239: 39 – 57.

[89] BIFET A, GAVALDA R. Learning from time-changing data with adaptive windowing. *Proceedings of the 2007 SIAM International Conference on Data Mining*. Minneapolis: SIAM, 2007: 443 – 448.

[90] CHANNOI K, MANEEWONGVATANA S. Concept drift for CRD prediction in broiler farms. *Proceedings of the 12th International Joint Conference on Computer Science and Software Engineering*. Songkhla: JCSSE, 2015: 287 – 290.

[91] LIU Y, WANG H, YU J, et al. Selective recursive kernel learning for online identification of nonlinear systems with NARX form. *Journal of Process Control*, 2010, 20(2): 181 – 194.

[92] CASALI A, GONZALEZ G, TORRES F, et al. Particle size distribution soft-sensor for a grinding circuit. *Powder Technology*, 1998, 99(1): 15 – 21.

[93] SHANG C, GAO X, YANG F, et al. Novel Bayesian framework for dynamic soft sensor based on support vector machine with finite impulse response. *IEEE Transactions on Control Systems Technology*, 2013, 22(4): 1550 – 1557.

[94] KANEKO H, FUNATSU K. Maintenance-free soft sensor models with time difference of process variables. *Chemometrics & Intelligent Laboratory Systems*, 2011, 107(2): 312 – 317.

[95] NGUYEN K T P, MEDJAHER K. A new dynamic predictive maintenance framework using deep learning for failure prognostics. *Reliability Engineering & System Safety*, 2019, 188: 251 – 262.

[96] WU H, ZHAO J. Self-adaptive deep learning for multimode process monitoring. *Computers & Chemical Engineering*, 2020, 141: 107024.

[97] YANG T, NA K, LV Y, et al. Real-time dynamic prediction model of NOx emission of coal-fired boilers under variable load conditions. *Fuel*, 2020, 274: 117811.

[98] SUNDARRAMAN A, SRINIVASAN R. Monitoring transitions in chemical plants using enhanced trend analysis. *Computers & Chemical Engineering*, 2003, 27(10): 1455 – 1472.

[99] YIN F, WANG H, XIE L, et al. Data driven model mismatch detection based on statistical band of Markov parameters. *Computers & Electrical Engineering*, 2014, 40(7): 2178 – 2192.

[100] SHANG C, HUANG B, YANG F, et al. Probabilistic slow feature analysis-based representation learning from massive process data for soft sensor modeling. *AIChE Journal*, 2015, 61(12): 4126 – 4139.

[101] HUANG D X, SUYKENS J A K, HUANG X L, et al, et al. Concurrent monitoring of operating condition deviations and process dynamics anomalies with slow feature analysis. *AIChE Journal*, 2015, 61(11): 3666 – 3682.

[102] YANG Hui, MAO Haohao, HUO Weigang. The application of clustering HMM model in QAR data analysis. *Computer Applications and Software*, 2018, 35(1): 85 – 91.
(杨慧, 毛好好, 霍纬纲. 聚类HMM模型在QAR数据分析中的应用研究. 计算机应用与软件, 2018, 35(1): 85 – 91)

[103] VENKATASUBRAMANIAN V, RENGASWAMY R, YIN K, et al. A review of process fault detection and diagnosis: Part I: Quantitative model-based methods. *Computers & Chemical Engineering*, 2003, 27(3): 293 – 311.

[104] HUANG K, WEN H, LIU H, et al. A geometry constrained dictionary learning method for industrial process monitoring. *Information Sciences*, 2020, 546: 265 – 282.

[105] DONG J, ZHANG C, PENG K. A novel industrial process monitoring method based on improved local tangent space alignment algorithm. *Neurocomputing*, 2020, 405: 114 – 125.

[106] ZHANG Y, LI X. Two-step support vector data description for dynamic, non-linear, and non-Gaussian processes monitoring. *The Canadian Journal of Chemical Engineering*, 2020, 98(10): 2109 – 2124.

[107] LIU Z, LI H, ZHOU P. Towards timed fuzzy Petri net algorithms for chemical abnormality monitoring. *Expert Systems with Applications*, 2011, 38(8): 9724 – 9728.

[108] CHEN G, XIE L, ZENG J, et al. Detecting model-plant mismatch of nonlinear multivariate systems using mutual information. *Industrial & Engineering Chemistry Research*, 2013, 52(5): 1927 – 1938.

[109] LUO Lin, YANG Bo, LI Hongguang. A dynamic multi-attribute decision making approach to industrial process control performance evaluations. *CIESC Journal*, 2018, 69(S1): 94 – 101.
(罗琳, 杨博, 李宏光. 基于动态多属性决策方法的工业过程控制性能评价. 化工学报, 2018, 69(S1): 94 – 101.)

[110] WANG J, LIANG Y, ZHENG Y, et al. An integrated fault diagnosis and prognosis approach for predictive maintenance of wind turbine bearing with limited samples. *Renewable Energy*, 2020, 145: 642 – 650.

[111] LEE I, PARK S H, BAEK J G. Random-forest-based real-time contrasts control chart using adaptive breakpoints with symbolic aggregate approximation. *Expert Systems with Applications*, 2020, 158: 113407.

[112] WANG Y, JIANG Q. Data-driven nonlinear chemical process fault diagnosis based on hierarchical representation learning. *The Canadian Journal of Chemical Engineering*, 2020, 98(10): 2150 – 2165.

[113] LEE D H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *Proceedings of the 30th International Conference on Machine Learning*. Atlanta: ACM, 2013, 3(2): 1 – 6.

[114] BENNETT K P, DEMIRIZ A. Semi-supervised support vector machines. *Proceedings of the Conference and Workshop on Neural Information Processing Systems*. Denver: MIT Press, 1999: 368 – 374.

[115] ZHOU Z H, LI M. Semi-supervised regression with co-training. *Proceedings of the International Joint Conference on Artificial Intelligence*. Edinburgh: Morgan Kaufmann, 2005, 5: 908 – 913.

[116] ZHOU Z H, CHEN K J, DAI H B. Enhancing relevance feedback in image retrieval using unlabeled data. *ACM Transactions on Information Systems*, 2006, 24(2): 219 – 244.

[117] LI D C, WU C S, TSAI T I, et al. Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge. *Computers & Operations Research*, 2007, 34(4): 966 – 982.

[118] HUANG C, MORAGA C. A diffusion-neural-network for learning from small samples. *International Journal of Approximate Reasoning*, 2004, 35(2): 137 – 161.

[119] PAN S J, KWOK J T, YANG Q. Transfer learning via dimensionality reduction. *Proceedings of the National Conference on Artificial Intelligence*. California: AAAI, 2008, 8: 677 – 682.

[120] DAI W, XUE G R, YANG Q, et al. Co-clustering based classification for out-of-domain documents. *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Jose: ACM, 2007: 210 – 219.

[121] WANG L, MAO S, WILAMOWSKI B M, et al. Ensemble learning for load forecasting. *IEEE Transactions on Green Communications and Networking*, 2020, 4(2): 616 – 628.

[122] SJIANG Gaoxia, FAN Ruixuan, WANG Wenjian. Label noise filtering via perception of nearest neighbors. *Pattern Recognition and Artificial Intelligence*, 2020, 33(6): 518 – 558.
(姜高霞, 樊瑞宣, 王文剑. 近邻感知的标签噪声过滤算法. 模式识别与人工智能, 2020, 33(6): 518 – 558)

[123] YUAN W, GUAN D, MA T, et al. Classification with class noises through probabilistic sampling. *Information Fusion*, 2018, 41: 57 – 67.

作者简介:

乔俊飞 教授, 目前研究方向为智能控制、神经网络分析与设计等, E-mail: junfeiq@bjut.edu.cn;

孙子健 硕士研究生, 目前研究方向为数据驱动软测量建模、概念漂移检测等, E-mail: sunzj@emails.bjut.edu.cn;

汤 健 教授, 目前研究方向为小样本数据建模、固废焚烧过程智能建模与控制等, E-mail: freeflytang@bjut.edu.cn