控制与决策
Control and Decision

基于工业视角的概念漂移检测与适应方法综述

周平, 张宇

引用本文:
周平, 张宇. 基于工业视角的概念漂移检测与适应方法综述[J]. 控制与决策, 2025, 40(6): 1774-1792.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2024.1431

基于工业视角的概念漂移检测与适应方法综述

周 平†, 张 宇

(东北大学 流程工业综合自动化全国重点实验室，沈阳 110819)

摘 要: 智能工业化的迅速发展推动了技术设备的持续创新, 随之而来产生大量实时数据流. 在这些数据流中, 数据的统计特性随时间可能发生变化, 这一现象称为概念漂移. 概念漂移对机器学习模型的性能产生显著影响, 未能及时识别和应对会导致模型性能的逐步下降, 进而引发错误决策, 从而在工业应用中造成不可忽视的损失. 鉴于此, 从工业应用的角度出发, 总结目前概念漂移检测与适应的研究进展. 首先, 聚焦于有监督环境下的工业概念漂移检测方法, 从基于性能、窗口技术和集成方法角度详细总结相关技术的发展现状; 其次, 针对工业场景中常见的标签稀缺问题, 系统介绍半监督学习和无监督学习在工业概念漂移检测中的应用方法, 此外讨论工业环境中普遍存在的不平衡类问题对概念漂移检测的影响, 并综述解决这一问题的相关策略; 最后, 针对工业环境下的概念漂移适应方法进行总结, 并提出未来研究的方向, 以进一步提升概念漂移检测方法在复杂动态环境中的表现.

关键词: 概念漂移；工业场景；标签稀缺；不平衡类；漂移适应；研究综述

中图分类号: TB118 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2024.1431

引用格式: 周平, 张宇. 基于工业视角的概念漂移检测与适应方法综述 [J]. 控制与决策, 2025, 40(6): 1774-1792.

A review of concept drift detection and adaptation methods from an industrial perspective

ZHOU Ping†, *ZHANG Yu*

(State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, China)

Abstract: The rapid development of intelligent industrialization has driven continuous innovation in technological equipment, resulting in the generation of large amounts of real-time data streams. Within these data streams, the statistical characteristics of the data may change over time, a phenomenon known as concept drift. Concept drift significantly impacts the performance of machine learning models. Failure to detect and address it in a timely manner can lead to a gradual decline in model performance, resulting in erroneous decisions and potentially causing substantial losses in industrial applications. This paper reviews the current research progress on concept drift detection and adaptation from the perspective of industrial applications. First, the paper focuses on supervised methods for industrial concept drift detection, providing a detailed overview of the development of relevant techniques, including performance-based methods, windowing techniques, and ensemble approaches. Second, to address the prevalent issue of label scarcity in industrial scenarios, the application of semi-supervised and unsupervised learning techniques in concept drift detection is systematically discussed. Furthermore, the paper discusses the impact of prevalent class imbalance challenge in industrial environments on concept drift detection and reviews strategies for addressing this issue. Finally, the paper summarizes concept drift adaptation methods in industrial settings and outlines potential directions for future research to enhance the performance of concept drift detection methods in complex and dynamic environments.

Keywords: concept drift; industrial scenarios; label scarcity; imbalanced classes; drift adaptation; review

0 引 言

现代工业物联网 (IIoT) 的快速发展和普及推动了智能工厂和工业 4.0 的不断进步, 也促使越来越多的先进检测技术与设备被应用于工厂环境, 助力工厂自动化与智能化视线[1-3]. 通过这些技术, 工业过程的大量日常运行数据得以连续、实时地采集并上传

收稿日期: 2024-12-10；**录用日期:** 2025-02-25.

基金项目: 国家重点研发计划项目 (2022YFB3304903)；国家自然科学基金项目 (U22A2049).

†通信作者. E-mail: zhouping@mail.neu.edu.cn.

至各个服务器或云平台, 为现场操作人员和管理者提供了极其重要的数据与信息支撑, 从而有助于其实现生产的高效操作和优化决策. 然而, 工业数据的快速积累不仅为模型学习、优化决策和性能提升提供了宝贵数据资源, 同时也带来了新的挑战. 特别是机器学习模型在实际工业场景的部署过程中, 现实世界不可避免的复杂动态性带来工业运行数据的概念漂移问题, 这对数据模型与智能模型的持续准确性和鲁棒性提出了更高的要求[4-6].

概念漂移是指目标变量的统计属性随时间不可预见地发生变化, 导致模型在实际环境中的性能下降[7]. 对于工业场景而言, 大数据场景下尤为显著, 主要原因是由于原材料与燃料等的不确定性和多变性, 在线工业环境本质上是复杂且不断变化的, 使得建立的模型即使在历史数据上表现优异, 也可能在实际部署中难以维持预期性能[8-9]. 影响概念漂移的因素包括原材料与燃料的波动引发工况动态时变、传感器等控制元器件故障、天气变化等不可抗拒因素以及数据采集方式差异等人为主观因素等, 这些因素使得工业数据驱动的系统在长期运行中面临极大的性能挑战[10-14]. 因此, 为了使模型能够在动态环境中快速且准确地适应各种动态变化, 针对概念漂移的精准检测与及时响应, 成为解决现实工业自动化问题的关键研究方向. 这不仅具有重要的学术意义, 也在实际应用中具有巨大的实用价值.

为系统梳理并总结近年来在工业应用场景下概念漂移检测与适应方法的相关研究成果, 本文从以下几个主要数据库中检索了近 10 年内发表在高水平期刊与重要会议上的研究工作: Science Direct, ACM Digital Library, IEEE Xplore 和 Springer Link. 检索的关键词主要包含“concept drift”“industrial”“drift detection”和“drift adaptation”等. 随后, 基于文献的题目、摘要以及与工业场景关联度等维度进行初步筛选; 对重复文献、无关文献进行剔除后, 进一步根据其影响力和研究主题的契合度进行二次筛选, 最终获得了约 100 篇高质量、紧密围绕概念漂移与工业应用的研究文献. 表 1 展示了本研究的文献检索与筛选流程.

近年来, 针对概念漂移问题的综述性研究逐渐增多, 许多学者从不同角度对概念漂移进行了系统性的调查和分析[15-22]. 例如, Lu 等[15] 对概念漂移的研究方法进行了详细地综述, 将相关研究划分为 3 大类别: 概念漂移检测、概念漂移理解以及概念漂移适应, 并分别进行了系统整理. Bayram 等[16] 针对基于模型性能的概念漂移检测方法进行专题性综述, 总结了统计过程控制、窗口技术和集成学习等主流方法. 此外, 一些研究聚焦于特定场景下的概念漂移问题. 例如: Lima 等[17] 总结了回归任务中不同机器学习模型应对概念漂移的适应策略; Li 等[18] 研究了机器学习任务中同时存在域漂移和概念漂移的情境, 提出三阶段分类方案以协同应对这两类挑战; Xiang 等[19] 从深度学习框架视角出发, 对基于判别学习、生成学习以及混合学习的概念漂移适应方法进行梳理, 揭示了深度学习框架下概念漂移适应的通用操作过程. 综上所述, 现有综述文献为概念漂移的研究奠定了重要基础, 但随着技术的发展和工业需求的提升, 针对特定场景和任务的概念漂移问题仍有深入探讨的空间. 本综述旨在进一步扩展已有研究视角, 为未来研究提供新的思路和方向.

本文聚焦于工业应用场景, 针对概念漂移问题及其在工业环境中与标签稀缺和数据不平衡共存的复杂情形展开研究, 总结了当前实际工业场景中应对概念漂移的各类解决方法. 本文的主要贡献包括以下几个方面:

1) 系统综述了当前工业应用背景下概念漂移的相关研究成果, 包括针对实际开源工业数据集的验证文献, 为工业生产中的决策支持提供了理论依据与技术参考.

2) 针对工业场景中普遍存在的标签稀缺性问题, 全面总结了适用于半监督和无监督环境的概念漂移检测方法. 这些方法在极少标签甚至完全无标签数据的情况下, 通过设计高效模型和算法, 实现了对概念漂移的精确检测, 为实际工业应用提供了解决方案.

表1　文献筛选过程

步骤	操作	结果
1. 关键词确定	(“concept drift” or “drift detection” or “drift adaptation”) and (“industrial” or “IIoT” or “manufacturing”)	—
2. 数据库选择	Science Direct, ACM DL, IEEE Xplore, Springer Link等	初步检索到约2 630篇文献
3. 初步筛选	根据题目和摘要判断是否与概念漂移及工业应用相关	剔除与工业场景或概念漂移无关的文献(约1 200篇)
4. 深度筛选	结合高引用率、期刊/会议等级、全文质量等标准进行再次排查	保留约500篇高质量文献供深入分析
5. 最终确定	对500篇文献进行全文审阅, 剔除工业背景不充分、实验不严谨等文献	最终确定约100篇核心文献

3) 聚焦工业中的不平衡类问题, 归纳现有基于块处理和在线处理的两类概念漂移检测技术, 为不同工业场景下的应用需求提供可行的技术路径与理论支持.

4) 对概念漂移适应方法进行分类总结, 分别从主动检测和被动适应两个角度整理工业环境中的现有解决方案, 全面呈现了应对概念漂移的多种策略与实践路径.

本文结构如图 1 所示. 第 1 节梳理概念漂移定义及其类型. 第 2 节聚焦于工业场景中专门针对概念漂移检测的方法. 第 3 节总结工业场景中同时存在标签稀缺和概念漂移问题的解决方法. 第 4 节归纳工业场景中同时面对数据不平衡与概念漂移问题的策略. 第 5 节梳理工业环境中针对概念漂移的适应方法. 第 6 节展望工业场景下概念漂移问题的未来研究方向.

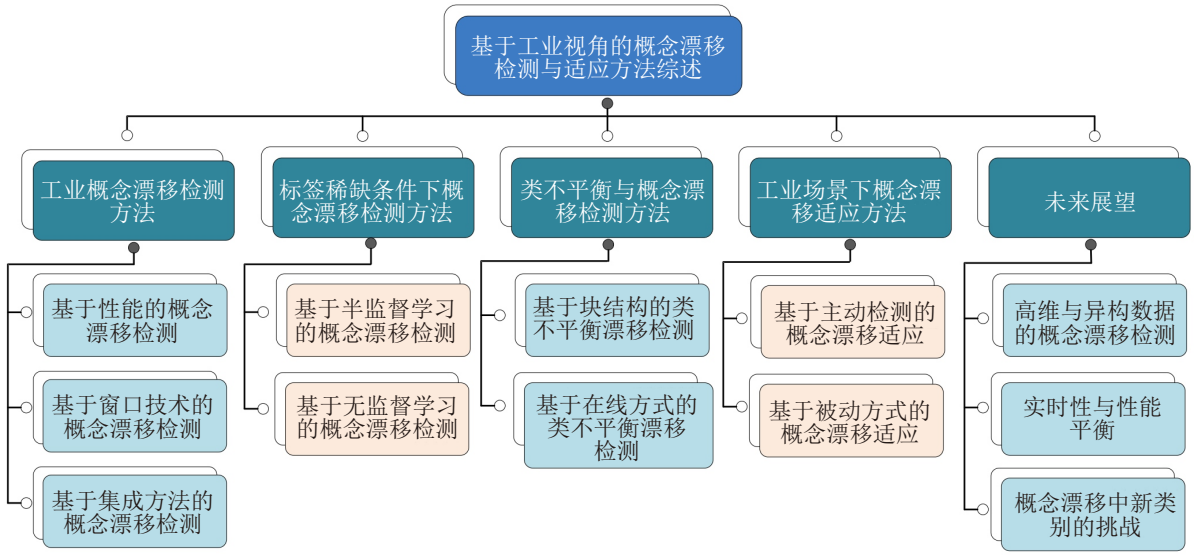


图1 本文总体结构

1 概念漂移定义及分类

概念漂移一词最早由 Schlimmer 等[20] 在 1986 年提出, 指底层数据分布中噪声信息随时间的变化. Baena-Garc 等[21] 指出概念漂移是一种目标变量的统计特性随时间变化的现象. 本文依据文献中普遍使用的概念漂移定义形式对其进行描述.

1.1 概念漂移定义

假设P_t表示在时间t处, 输入变量x和目标变量y的联合概念分布. P_{t+1}表示在时间$t+1$处, 输入变量x和目标变量y的联合概念分布. 如果下式成立, 则表明概念漂移的发生:

$$\exists t: P_t(x,y) \neq P_{t+1}(x,y). \tag{1}$$

概念漂移定义为在某一时刻联合概率的变化. 由联合概率分布的形式可以将其拆分成两部分$p(x,y)=p(x)p(y|x)$, 即式 (1) 可以进一步转变为如下形式:

$$\exists t: P_t(x)P_t(y|x) \neq P_{t+1}(x)P_{t+1}(y|x). \tag{2}$$

因此, $p(x)$和$p(y|x)$的单独变化或混合变化都可以导致概念漂移的发生.

1.2 概念漂移成因

根据概念漂移的定义以及联合概率的分解, 可以得出联合概率的变化源于以下 3 种情形[22]:

1) 虚拟概念漂移: 输入特征x的分布发生变化, 但在给定输入特征x的条件下目标变量y的条件概率保持不变, 即

$$P_{t_0}(x) \neq P_{t_1}(x) \text{且} P_{t_0}(y|x) = P_{t_1}(y|x). \tag{3}$$

这种漂移不影响决策边界, 仅改变输入特征空间, 因此称为虚拟概念漂移, 也称为特征漂移或数据漂移[23], 详见图 2(a).

2) 真实概念漂移: 在给定输入特征x的条件下, 目标变量y的条件概率发生变化, 而输入特征x的分布保持不变, 即

$$P_{t_0}(x) = P_{t_1}(x) \text{且} P_{t_0}(y|x) \neq P_{t_1}(y|x). \tag{4}$$

这种漂移直接影响模型的预测性能, 因此称为真实的概念漂移, 详见图 2(b).

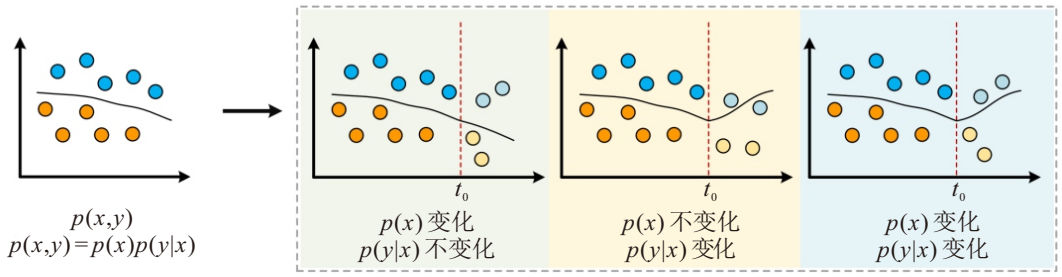
3) 混合概念漂移: 在实际场景中, 虚拟和真实概念漂移可能同时存在, 即

$$P_{t_0}(x) \neq P_{t_1}(x) \text{且} P_{t_0}(y|x) \neq P_{t_1}(y|x). \tag{5}$$

这种情况称为混合概念漂移, 详见图 2(c).

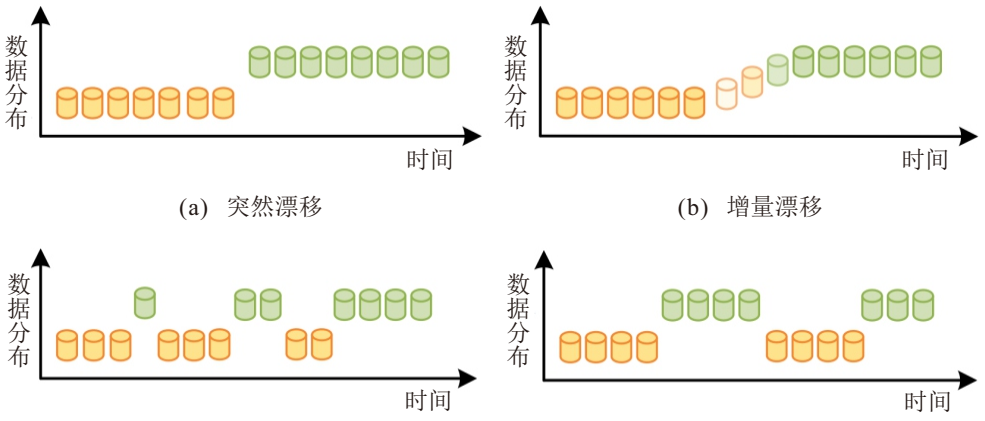
1.3 概念漂移类型

概念漂移不仅可能在某一个时刻突然发生, 也可能经历较长的时间逐渐演变. 大多数研究通常将



(a) 虚拟概念漂移　(b) 真实概念漂移　(c) 混合概念漂移

图2　概念漂移产生原因



(a) 突然漂移　(b) 增量漂移

(c) 逐渐漂移　(d) 重复漂移

图3　概念漂移类型

概念漂移分为 4 种类型: 突然漂移、增量漂移、逐渐漂移和重复漂移. 本文基于文献 [24] 中的数学定义, 对这 4 类漂移进行了总结:

1) 突然漂移 (abrupt drift): 指数据分布在极短时间内发生剧烈变化, 表现为分布突然从一种状态转变为另一种状态, 如图 3(a) 所示. 数学上可表示为

$$\exists t, P_t(X,y) \neq P_{t+\Delta t}(X,y),\ \Delta t < \delta. \tag{6}$$

其中δ为一个较小的阈值.

2) 增量漂移 (incremental drift): 指数据分布随着时间逐步变化, 经过长期累积, 最终转变为全新的分布, 如图 3(b) 所示. 其数学描述如下:

$$\begin{aligned}&\exists t, P_t(X,y) \neq P_{t+\Delta t}(X,y),\\ &\exists m, P_t(X,y) < P_m(X,y) < P_{t+\Delta t}(X,y),\\ &\quad t < m < t+\Delta t.\end{aligned} \tag{7}$$

3) 逐渐漂移 (gradual drift): 指旧分布逐渐过度到新分布, 数据分布在一定时间内逐步趋于新状态, 如图 3(c) 所示. 该过程的数学表达如下:

$$\begin{aligned}&\exists t, P_t(X,y) \neq P_{t+\Delta t}(X,y),\\ &\exists m, P_m(X,y) = \alpha(t)P_t(X,y)+\\ &(1-\alpha(t))P_{t+\Delta t}(X,y),\ t < m < t+\Delta t.\end{aligned} \tag{8}$$

其中$\alpha(t)$为一个介于 0 到 1 的系数.

4) 重复漂移 (recurring drift): 常见于季节性或周期性变化场景, 即历史数据分布在经过一段时间后重新出现, 如图 3(d) 所示. 数学上可表示为

$$\begin{aligned}&\exists t, P_t(X,y) \neq P_{t+\Delta t}(X,y),\\ &\exists \Delta m : P_{t+\Delta m}(X,y) = P_t(X,y),\ \Delta m > \Delta t.\end{aligned} \tag{9}$$

从工业视角理解这 4 类漂移, 以传感器异常为例. 突然的信号偏置是一种突然漂移, 振幅随时间增加的偏置是一种增量漂移, 在传感器临近故障之前频率增加的信号尖峰构成逐渐漂移, 在某些周期性工厂条件下发生的传感器读数构成重复漂移[15]. 针对不同的漂移类型, 已经开发了具体的概念漂移检测方法. 例如, DDM[25]、DetectA[26] 和 H-CDT[27] 等方法专门用于检测突然漂移, 而基于 Jensen-Shannon 散度的 ESCR 则用于识别反复出现的概念漂移[28].

1.4　概念漂移挑战

在实际工业应用中, 机器学习模型的核心目标之一是在未见数据上实现良好的性能, 这种能力称为模型的泛化能力. 理想状态下, 模型不仅应在训练数据上表现优异, 更需要在全新的数据环境中保持稳定. 然而, 正如图 4 中性能下降所示, 概念漂移现象往往会破坏这一假设. 随着数据环境随时间动态演变, 模型可能难以适应这些变化, 导致性能显著下降[29].

为应对概念漂移, 及时检测并适应这种变化至关重要. 这不仅是提升模型实际应用能力的必要措

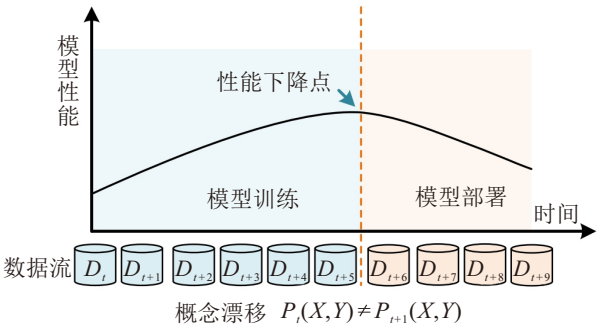


图4 概念漂移导致的模型性能下降

施, 也是增强其泛化能力的关键步骤. 通过监测数据分布的变化并相应地调整模型, 能够有效确保其在处理未知数据时, 仍然具备较高的准确性和鲁棒性, 从而在动态环境中实现长期稳定的性能.

2 工业场景下概念漂移检测方法

2.1 基于性能的概念漂移检测

2.1.1 直接基于错误率的概念漂移检测

基于性能的概念漂移检测方法主要通过监测在线学习模型的错误率直接评估模型的有效性[24, 30-32]. 由于该方法只有在模型性能真正受到影响时才触发漂移处理, 在工业应用中具有较高的实用性和较低的误报率. 其中漂移检测方法 (drift detection method, DDM) 是该类方法的典型代表, 最初由 Gama 等[25] 提出, 用于监测学习模型在数据流中的在线错误率, 这些变化往往会显著影响模型性能. DDM 通过设置预警和漂移两个阈值来实现检测: 当满足下式所示条件时, 触发预警:

$$p_i \geqslant p_{\min} + 2s. \tag{10}$$

其中: p_i为当前时刻的错误率, $p_{\min}$为观察到的最小误差率, s 为$p_{\min}$时刻的标准误差. 当满足如下条件时, 触发漂移:

$$p_i \geqslant p_{\min} + 3s. \tag{11}$$

在触发预警阶段, 系统通常会提前进行准备, 例如收集最新数据; 而在触发漂移阶段, 则会发出信号以调整或重新训练模型. DDM 方法实现简单, 计算开销较低, 能够有效检测对模型性能产生重大影响的数据流变化, 但其对阈值的选择较为敏感, 且对渐变或细微的漂移响应不足.

在 DDM 方法的基础上, 许多研究者提出了改进方案以适应不同的工业背景. 例如, Wang 等[33] 针对工业物联网中采集的温度、湿度、光照和速度数据, 提出了一种结合误报率的多标签漂移检测方法, 称为 DDM-FP-M. 该方法在借鉴传统 DDM 的基础上, 通过监测准确率和误报率的变化实现数据流中概念漂移的检测, 其检测过程分为预警和漂移两个阶段, 具体如下:

$$p_i + \mathrm{fpr}_{\overline{v}\text{-}i} \geqslant p_{\min} + 2 \times \mathrm{fpr}_{\overline{v}\text{-}\min}, \tag{12}$$

$$p_i + \mathrm{fpr}_{\overline{v}\text{-}i} \geqslant p_{\min} + 3 \times \mathrm{fpr}_{\overline{v}\text{-}\min}. \tag{13}$$

其中: p_i为第i个样本被错误分类的概率, $p_{\min}$为训练阶段的最小错误分类概率, $\mathrm{fpr}_{\overline{v}\text{-}i}$为第$i$个样本误报率的算术平均值. DDM-FP-M 通过将误报率纳入漂移检测机制, 特别适应多标签环境的需求. 在多标签的应用场景中, 误报对性能的负面影响尤为显著, 仅依赖总体错误率的监控可能不足以全面反映模型性能的衰退.

此外, 在工业领域中还存在回归任务, 例如实际工业电价预测问题[34]. 针对此, Yan[35] 提出了精确概念漂移检测方法 (accurate concept drift detection method, ACDDM), 该方法利用 Hoeffding 不等式分析概念漂移检测误差率的不一致性. 其基本思路为: 首先, 计算时间t的错误p_t, 用到目前为止所有实例的平均误差表示为

$$p_t = \frac{1}{t}\sum_{i=1}^{t} L(y_i, \hat{y}_i). \tag{14}$$

其中: L为损失函数, y_i为真实标签, $\hat{y}_i$为预测标签. 随后使用 Hoeffding 不等式判断误差p_t是否显著偏离期望误差, 在数据平稳假设下满足下式:

$$P(|p_t - E[p]| \geqslant \varepsilon) \geqslant 2\mathrm{e}^{-2t\varepsilon^2}. \tag{15}$$

其中ε为偏离阈值, 如果超过设定的阈值, 则认为漂移存在. ACDDM 为漂移检测提供了坚实的理论基础, 能够在早期检测细微变化方面优于基于简单阈值设定的 DDM.

针对基于深度学习模型的工业预测任务, Kahraman 等[36] 提出了一种具有记忆机制的动态建模方法 (dynamic modeling with memory, DMWM), 该方法将概念漂移检测嵌入到工业机器实时能源消耗预测过程中. DMWM 首先将数据流动态划分为不同的数据块, 在第 1 个数据块上训练初始模型, 并利用该模型预测后续块中的能源消耗, 同时计算预测误差. 当实时计算的误差超过预设阈值时, 即判定为发生了概念漂移. 该方法依托深度学习技术, 更能捕捉数据流中非线性与多变量之间的复杂关系, 适用复杂工业场景. 尽管其对计算资源的需求较高, 但通过优化模型更新频率和降低重复训练次数, 长期来看能够提升整体效率.

2.1.2 间接反映错误率的概念漂移检测

除直接使用错误率或预测误差作为检测指标外, 还可以通过数据的一些统计特性间接反映错误率的

变化, 如均值、标准差、底层分布的参数以及特征频率等[37]. 其中, 累积和 (cumulative sum, CUSUM) 算法和指数加权移动平均 (exponentially weighted moving average, EWMA) 算法是两种较为经典的方法[38].

CUSUM 算法通过累积各输入数据与均值之间的偏差判断系统状态是否稳定. 假设X_i是第i个输入样本, μ_0是输入数据的均值, 则单侧上、单侧下 CUSUM 更新公式为

$$\begin{aligned} C_i^+ &= \max[0, X_i - (\mu_0 + K) + C_{i-1}^+], \\ C_i^- &= \max[0, (\mu_0 - K) - X_i + C_{i-1}^-]. \end{aligned} \tag{16}$$

其中: $C_0^+ = C_0^- = 0$; K是一个配置参数, 也称为参考值. C^+和C^-为每个新的输入数据进行更新, 如果其中一个大于H, 则会发出漂移信号. H是决策区间, 定义了均值周围的稳定边界. 需要注意的是, CUSUM 算法在检测漂移时可能存在一定延迟, 尤其在连续出现相反方向漂移时.

EWMA 算法则侧重于检测数据平均值的渐变变化, 通过对最新数据赋予更高权重, 使其对近期变化更为敏感. 其基本公式为

$$Z_i = \lambda X_i + (1 - \lambda) Z_{i-1}. \tag{17}$$

其中: Z_i为第i个样本的 EWMA 值, X_i为第i个实际观测值, λ为平滑系数 ($0 < \lambda < 1$). 上控制限 (UCL) 和下控制限 (LCL) 通常定义为

$$\mathrm{UCL} = \mu_0 + L\sigma\sqrt{\frac{\lambda}{2-\lambda}}, \tag{18}$$

$$\mathrm{LCL} = \mu_0 - L\sigma\sqrt{\frac{\lambda}{2-\lambda}}. \tag{19}$$

其中: μ_0为目标平均值, σ为数据的标准差, L为控制宽度因子. 需要指出的是, 传统方法中这两种算法无法动态更新目标值, 限制了其在某些工业应用中的适用性.

为解决这一问题, Estaji 等[39]对传统 EWMA 和 CUSUM 算法进行改进, 提出了动态更新目标值和控制限的方法, 使其能够适应工业过程数据中均值和标准差的逐渐变化, 同时不对数据的统计属性做过多假设, 从而更及时地捕捉到数据中的细微且渐近的变化.

此外, 随着深度学习技术在工业领域中不断发展, 越来越多的间接指标被用于概念漂移检测[40-41]. 例如, 在公开的网络入侵检测等数据集上, Zhang 等[42]提出了自适应在线增量学习算法 (adaptive online incremental learning, AOIL). 该方法利用自动编码器结合记忆模块, 通过比较输入数据的重构误差来判断概念漂移: 当输入数据的分布与自动编码器重构的分布不匹配时, 重构误差会显著增加, 从而触发漂移检测. 相较于传统的 EWMA 和 CUSUM 方法, AOIL 能够更动态地适应数据流的变化, 在新数据到来时迅速调整模型以应对新的数据分布. 然而, 该方法对计算资源的需求较高, 尤其在数据流变化较快的环境中, 需要额外维护和更新记忆模块.

2.2　基于窗口技术的概念漂移检测

基于窗口技术的概念漂移检测方法因其高效、实时的特点, 在大规模数据流和工业应用中备受关注. 总体而言, 这类方法利用滑动窗口将数据流按照数据量或时间间隔进行分段处理, 然后对各窗口内的数据特性进行监测和比较, 以判断是否存在概念漂移. 根据窗口划分和检测策略的不同, 主要可以分为单窗口检测和双窗口检测两大类.

2.2.1　基于单窗口的概念漂移检测

单窗口检测方法通常与基于性能指标 (如错误率、预测准确度) 的检测方法相结合, 其基本思想是对模型近期预测结果在固定或动态调整的滑动窗口内进行监控, 如图 5 所示. 例如, Pesaranghader 等[43]提出了一种基于 McDiarmid 不等式的漂移检测方法 (McDiarmid drift detection method, MDDM). 该方法在预测结果上滑动一个窗口, 其中正确预测记为 1, 错误预测记为 0. 随后, 根据样本的时序信息对窗口内的预测结果进行加权处理, 并将加权平均值与历史最大加权平均值进行比较, 从而判断是否发生概念漂移, 其判定条件为

$$\mu_w^m - \mu_w^t \geqslant \varepsilon_w. \tag{20}$$

其中: μ_w^m为迄今为止的最大加权平均值, μ_w^t为当前滑动窗口的加权平均值. 此外, Zhang 等[44]提出了一种基于动力学表征的奇异频谱分析进行工业概念漂移检测的方法, 并通过实际烧结过程的验证表明其有效性. 该方法利用奇异熵作为检测指标, 若满足如下条件, 则判断发生概念漂移:

$$E_{t+1} = \frac{e_{t+1} - e_t}{e_t} |\Delta^2(e_{t+1})| > z_q, \tag{21}$$

其中e_t为时间t的熵值. 单窗口方法计算量小、实时性好, 适用于对延迟要求较高的工业应用, 例如在线

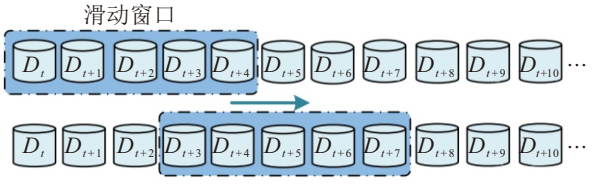


图5　基于单个滑动窗口的检测

监控、生产线质量检测. 然而, 固定窗口大小难以兼顾对快速突变和缓慢漂移的检测需求, 且当数据流规模不断增长时, 固定窗口的灵活性不足, 容易导致误判或漏判. 针对这一问题, Du 等[45] 提出了自适应滑动窗口检测方法 (adaptive sliding window-based detection method, ADDM), 通过监控窗口内数据熵的动态变化实时调整窗口大小, 提高了检测精度和响应时效, 但同时增加了算法的复杂性和调参难度.

在工业环境中, 如传感器监控、工艺流程控制等领域, 数据往往具有时变性和非平稳性. 单窗口方法在这些场景下能够快速捕捉异常, 但需要根据实际工业数据的特性 (例如数据量、噪声水平、采样频率等) 设计合适的窗口策略. 目前, 部分研究已在实际烧结过程、机械故障检测等数据集上进行验证[46], 但仍缺乏大规模工业数据集的系统评估.

2.2.2 基于双窗口的概念漂移检测

双窗口检测方法通常通过构建两个相互独立的时间窗口 (一个代表历史数据, 另一个代表最新数据) 对比数据分布或模型参数的差异, 从而实现概念漂移的检测. 该方法的主要思想是, 当新旧数据窗口之间的差异达到统计显著性水平时, 即认为概念漂移已经发生. 例如, Lu 等[15] 提出了统一的概念漂移检测框架, 如图 6 所示. 该框架通过对新旧数据窗口建模并进行差异度量以判断概念漂移.

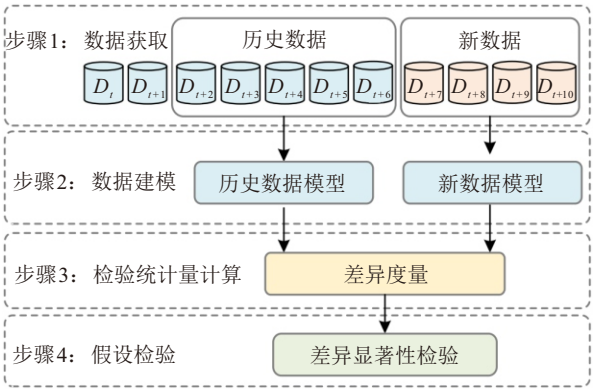


图6 基于双窗口的概念漂移检测一般框架

在工业电价预测、风力发电厂监控、机械设备健康管理等实际应用中, 数据分布可能由于设备老化、环境变化或工艺调整而发生漂移. 基于双窗口的检测方法通过对比历史与新数据, 不仅能够较为准确地捕捉概念漂移, 还可提供漂移发生的时间窗口和持续时长信息, 为工业系统及时调整模型提供依据. 在该类方法中, Yang 等[13] 在实际工业电价预测和风力发电厂预测任务中, 基于在线顺序极限学习机对两个阶段数据进行建模, 根据欧几里得距离量化模型输出权重的差异来检测概念漂移, 其检测准则为

$$D(M^*(t_1), M^*(t_1+\Delta t)) = \sqrt{\sum_{i=1}^{N_h}\sum_{j=1}^{S}\left(\beta_{t_{1_{i,j}}}, \beta_{t_1+\Delta t_{i,j}}\right)^2} > \mathrm{Th}. \tag{22}$$

其中: $M^*(t_1)$为该方法在时间t_1之前训练的模型, $M^*(t_1+\Delta t)$为该方法在时间t_1到$t_1+\Delta t$训练的模型, β为输出权重, Th为预设阈值. 当该不等式成立时, 即认为数据分布已发生显著变化.

此外, 为解决工业数据流中突变型与渐近型漂移检测的实际需求, 近年来基于窗口技术的概念漂移检测方法呈现出由简单到复杂、由单一指标到多维信息融合的递进趋势. 首先, Gözüaçık 等[47] 提出了一种单类概念漂移检测器 (one-class drift detector, OCDD), 其核心思想是利用两个滑动窗口分别存储新旧数据, 并通过计算窗口内分类器检测出的离群值百分比来判断概念漂移. 其主要优点在于对突变型和增量型漂移具有较高的敏感性和实时响应能力, 适合对异常数据反应迅速的工业应用 (如设备故障预警、生产线异常监控). 然而, 其不足之处在于当数据噪声较大或漂移模式较为复杂时, 单一的离群值比例可能不足以全面反映数据分布的细微变化, 从而增加了误判的风险.

为进一步提升检测的细粒度和准确性, Guo 等[48] 提出了一种基于多滑动窗口的概念漂移类型识别方法, 简称为 CDT_MSW. 该方法在保留了双窗口对比的基础上, 通过整合多个滑动窗口的信息, 不仅实现了概念漂移的识别, 而且能够量化漂移持续的时长, 从而进一步对漂移类型进行区分. 这一方法的优势在于它能更全面地捕捉数据流中的渐近性变化, 特别适用于那些漂移具有阶段性和多样性特征的工业场景. 但同时, 复杂的窗口管理和多参数调节增加了算法实现的难度, 对计算资源和实时性提出了更高要求, 这在高频数据流或资源受限的工业环境中可能成为限制因素. 此外, 文献 [49] 提出了一种基于 Kolmogorov-Smirnov(KS) 检验的双窗口概念漂移检测方法, 称为 KSWIN. 该方法通过比较两个滑动窗口中数据分布的差异, 利用 KS检验的统计显著性结果来判断是否发生概念漂移. 相对于前述方法, KSWIN 的优势在于它借助严谨的统计检验理论, 能够在一定程度上降低因窗口参数选择不当而带来的误判风险, 适合于那些对检测结果可靠性要求较高的工业应用. 但其不足之处在于, 检测效果对窗口大小和数据分布特性较为敏感, 同时在处理数据量庞大或分布动态变化较快的场景中, 统计检验的计算复杂度可能限制其实时性.

综上所述, 双窗口方法能够有效捕捉新旧数据

表2　概念漂移检测集成方法对比

方法	集成方式	集成内容	漂移检测策略
drift detection ensemble(DDE)[51]	概念漂移检测器集成	DDM、ADWIN、ECDD、HDDM	阈值比较
Toumi等[52]	概念漂移检测器集成	ADWIN、DDM、STEPD	阈值比较
statistical tests ensemble detector(STED)[53]	统计测试集成	Brown-Forsythe、O'Brien、ANOVA	多数投票 early-find-early-report
e-Detector[54]	概念漂移检测器集成	DDM、EDDM、ADWIN、STEPD、EnDDM	early-find-early-report

间的细微差异, 适合检测渐变型漂移, 并且通过对比建模可以从定量上评估漂移程度. 然而, 双窗口方法通常依赖于预先设定的窗口大小和边界, 其性能对窗口参数高度敏感; 同时, 如何在工业环境中合理划分"历史"和"新数据"窗口仍是一个开放性问题. 此外, 双窗口方法的计算复杂度相对较高, 在数据流速率较快或数据量庞大的工业场景中可能面临实时性挑战[50].

2.3　基于集成方法的概念漂移检测

受在线分类器集成技术在性能和通用性方面表现的启发, 集成思想逐渐被引入概念漂移检测领域. 现有研究主要通过集成多种漂移检测器或统计测试提升检测精度、鲁棒性和适用性, 特别适用于工业环境中数据流噪声大、非平稳性强的情况. 总体上, 基于集成方法的概念漂移检测可分为基于检测器集成和基于统计测试集成两大类. 表 2 总结了部分方法的集成方式、所使用的检测器或统计测试, 以及对应的漂移检测策略.

2.3.1　基于检测器集成的方法

该类方法通过聚合多个独立的漂移检测器的检测信号或警告, 实现综合判定. 设有N个基础检测器, 其中第i个检测器在时刻t的输出为$d_i(t) \in \{0, 1\}$(1 表示检测到漂移, 0 表示无漂移), 则集成检测器的决策可表示为

$$D(t) = I\left(\sum_{i=1}^{N} d_i(t) \geqslant \theta\right). \tag{23}$$

其中: $I(\cdot)$为示性函数, θ为预设的投票阈值. 例如, Maciel 等[51] 提出一种小型集成概念漂移检测器 (drift detection ensemble, DDE), 通过聚合多个概念漂移检测器的警告信号和检测信号, 显著提升了整体检测精度. 该方法引入灵敏度参数, 用于指定触发警告或漂移信号所需的检测器数量, 并基于不同的灵敏度阈值划分为 3 种集成配置. 在实际电力系统监控中, DDE 在处理高噪声、复杂电力数据流时展现了较高的鲁棒性. 然而, 其检测性能高度依赖于基础检测器的选择和灵敏度参数的调优; 同时, 多检测器并行计算可能导致较高的计算开销, 给实时监控带来挑战. 基于此思路, Toumi 等[52] 进一步利用 3 个不同的检测器对云计算环境下的资源使用率数据进行监控, 其决策同样可采用上述多数投票机制, 并在实际云平台资源管理中取得了良好效果, 表明了该方法在工业应用中的适用性.

2.3.2　基于统计测试集成的方法

另一类方法将多种统计测试 (如 Kolmogorov-Smirnov 检验、CUSUM 等) 整合到同一检测框架中, 通过多数投票机制和"早发现–早报告" (early-find-early-report) 策略实现漂移检测. Pérez 等[53] 提出的基于多个统计测试的集成检测器 (statistical tests ensemble detector, STED) 便是典型代表, 其通过在一个检测器中融合多种统计测试, 实现了对数据分布变化的多角度捕捉. 为了将各测试的结果进行集成, STED 使用 Fisher 联合检验方法. 具体而言, 计算联合统计量

$$T(t) = -2\sum_{i=1}^{m} \ln(p_i(t)). \tag{24}$$

在零假设 (即数据无漂移) 的条件下, 统计量$T(t)$服从自由度为$2m$的卡方分布. 因此, 可以设定一个临界值$\chi^2_{\alpha,2m}$, 其中α为显著性水平. 如果联合统计量超过预设的临界值, 则可以拒绝零假设, 认为在该窗口内发生了概念漂移. 该方法在传感器监控和生产质量控制等工业数据集中能够较全面地反映数据特性变化, 但由于大多数统计测试均依赖于数据分布的假设, 其在处理非平稳或高噪声数据时可能受到限制, 同时多重测试的并行计算也增加了实时性挑战.

在上述两类方法的基础上, Du 等[54] 提出的 e-Detector 采用了一种选择性集成策略, 旨在同时兼顾检测敏感性与实时响应性. 该方法也采用 early-find-early-report 策略, 即一旦任何基本检测器发现概念漂移, e-Detector 便立即标记漂移, 从而实现快速响应. 此外, e-Detector 基于选择性标准将候选检测器进行排序, 方式如下:

$$\mathrm{CoF} = \frac{\sum_{i=1}^{n_a} \mathrm{Perr}_i \times \mathrm{Rdr}_i}{n_a} + \frac{\sum_{j=1}^{n_g} \mathrm{Rerr}_j}{n_g}. \tag{25}$$

其中: Rdr_i和Perr_i分别为检测器在第 i 个参考数据集上的相对检测率和先验错误率; n_a和n_g分别为包

含突然漂移和逐渐漂移的总数. 在设备故障预警和生产线实时监控等工业应用中, e-Detector 能够迅速响应数据流中的漂移信号, 但同时也面临计算资源分配和参数调节的挑战, 尤其是在高频数据流的场景下.

综上所述, 基于集成方法的概念漂移检测通过检测器和统计测试的多样化组合, 不仅提高了整体检测精度和鲁棒性, 而且为工业环境中应对数据流噪声大、非平稳性强等问题提供了灵活的解决方案. 在电力系统监控、云计算资源管理、设备故障预警以及生产质量控制等实际应用中, 集成方法的优势已得到充分验证; 但在检测器组合、参数调优和计算资源要求等方面仍存在不足.

3 工业场景下标签稀缺与概念漂移检测方法

在工业场景中, 由于标注数据往往稀缺且获取成本高昂, 标签稀缺在概念漂移检测中的问题日益受到关注. 为解决这一问题, 近年来研究者在半监督学习和无监督学习两个方向上开展了大量工作. 前文主要回顾了基于监督学习的工业概念漂移检测方法, 而本节重点探讨半监督和无监督学习在概念漂移检测中的应用, 并对典型方法的优势与不足进行归纳, 同时结合电力、传感器监控等工业数据集的实际应用背景进行分析.

3.1 半监督学习中的概念漂移检测

半监督学习通过充分利用有限的标记数据和大量无标签数据, 有效缓解了标签稀缺对模型性能的影响. 在概念漂移检测中, 半监督方法主要集中在自训练和主动学习两大类.

3.1.1 基于自训练的半监督检测方法

自训练方法是最为常见的技术, 其基本思想是: 首先利用已有的标记数据对模型进行初始训练; 随后对无标签数据进行预测, 并选择置信度较高的预测结果作为伪标签加入训练集中, 迭代更新模型, 直至充分利用所有数据, 具体如图 7 所示. 例如, Khezri 等[55] 提出的 STDS 方法在自训练过程中结合聚类算法与分类器预测的集成策略, 从而选出高置信度的伪标签样本. 该方法利用选定样本的 Kullback-Leibler (KL) 散度来衡量数据分布的变化, 当 KL 散度超过预设阈值时, 即认为发生了概念漂移. 此方法在工业应用 (如大规模制造数据流或电力系统监控) 中, 能够在仅有少量标记数据的情况下, 有效捕捉环境或设备状态的变化, 从而保证系统的鲁棒性. 然而, 由于该方法对聚类算法参数和置信度阈值较为敏感, 加之计算复杂度较高, 在大规模工业数据流的实时处理上可能存在一定挑战. 针对上述实时性不足的问题, 文献 [56] 提出的 CPSSDS 框架则在自训练过程中引入了共形预测. 该方法为每个无标签样本计算一个p值, 用以衡量标签信息的可靠性. 当样本的p值大于或等于预设阈值时, 将其纳入训练集. 随后, 通过对新旧数据块的p值分布采用 Kolmogorov-Smirnov(KS) 检验, 以判断是否存在概念漂移. 此框架在电力系统的实时监控与异常检测中, 不仅兼顾了模型性能的提升, 也实现了对数据分布变化的敏感捕捉. 然而, 由于阈值设置和参数调优较为复杂, 该方法需要针对具体工业场景进行精细调整, 以满足实时性和精确度的双重要求.

为了解决静态阈值难以快速适应数据变化的问题, 文献 [57] 提出了 DyDaSL 算法, 使用基于改进自训练方法 FlexCon-C 阈值的半监督学习方法, 进一步增强其在概念漂移检测中的性能. 该方法在伪标签生成时引入动态更新的置信阈值, 以优化样本选择, 更新公式如下:

$$\mathrm{conf}(t_{e+1}) = \begin{cases} \mathrm{conf}(t_e), & \mathrm{mp}-\varepsilon < \mathrm{acc} < \mathrm{mp}+\varepsilon; \\ \mathrm{conf}(t_e)-\mathrm{cr}, & \mathrm{acc} \geqslant \mathrm{mp}+\varepsilon; \\ \mathrm{conf}(t_e)+\mathrm{cr}, & \mathrm{acc} \leqslant \mathrm{mp}-\varepsilon. \end{cases} \tag{26}$$

其中新的置信阈值$\mathrm{conf}(t_{e+1})$由当前置信阈值$\mathrm{conf}(t_e)$和变化率cr决定. 置信阈值的更新依赖于分类器的精度acc、最小可接受精度mp和可接受变化ε. 上述方法均在 Elec 等数据集上进行验证, 表现出了显著的有效性. 其中 Elec 数据集是一个广泛应用于概念漂移检测研究的电力系统数据集, 该数据集记录了电力市场中电力负荷、价格等关键指标, 具有明显的时序特征和动态变化性. 在工业应用中, Elec 数据集常用于电力系统监控、异常检测等场景, 由于其数据量大且变化多样, 成为验证概念漂移检

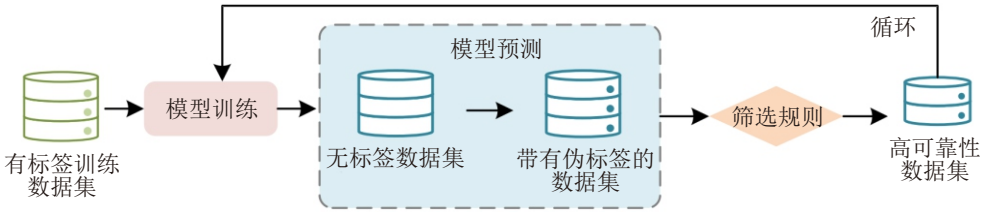


图7 基于自训练的半监督方法

测方法性能的重要基准[58].

3.1.2 基于主动学习的半监督检测方法

主动学习作为一种选择性标注技术, 通过智能选择最有价值的样本进行人工标注, 可以进一步缓解标签稀缺问题. 例如, Tan 等[59] 提出了一种基于主动学习与积极无标签学习的半监督概念漂移检测方法. 该方法基于后验概率分布来识别概念漂移, 并在 Elec 数据集上取得了较好的实验效果. 但需要注意的是, 主动学习的样本选择策略对数据分布的变化较为敏感, 可能存在关键样本漏标或误标的风险; 同时, 依赖专家标注也使得实际应用成本增加.

随着机器学习技术的逐渐发展, 除了生成伪标签的方法外, 研究人员还探索了在检测阶段无需依赖真实标签的概念漂移检测方案. 例如, Ali 等[60] 提出一种基于自编码器的半监督概念漂移检测方法, 其目的是在不需要真实标签的情况下识别概念漂移. 该方法在离线训练阶段利用标注数据构建模型, 而在在线检测阶段通过监测重构误差来判断是否发生概念漂移, 从而降低对实时标签的依赖, 提高检测的实时性和鲁棒性.

3.2 无监督学习中的概念漂移检测

在工业场景中, 标签数据完全缺失的情况下, 无监督学习在概念漂移检测中的应用越来越受到关注. 由于无需依赖标注数据, 无监督方法通常通过以下信息实现概念漂移检测: 基于两个分布之间的统计检验或散度度量[61-64]、跟踪模型参数的变化[65] 以及预测结果[66-67].

3.2.1 基于分布间统计检验或散度度量的方法

这类方法通过比较参考窗口与新数据窗口间的统计分布判断是否发生概念漂移. 例如, DRIFTLENS 方法[61] 使用 Frechét 漂移距离来衡量两个多元正态分布之间的差异, 其计算公式为

$$\mathrm{FDD}(b,w)=\\ \|\mu_b-\mu_w\|_2^2+\mathrm{Tr}(\Sigma_b+\Sigma_w-2\sqrt{\Sigma_b\Sigma_w}). \tag{27}$$

其中μ_b, μ_w和Σ_b, Σ_w分别为参考窗口和新窗口的多元正态分布的均值向量和协方差矩阵. 当距离值超过设定阈值时, 即判定发生了概念漂移. 该方法提供了一种数学上明确且直观的方式来量化分布变化. 然而, 在许多工业应用中, 数据往往呈现出复杂、非正态且噪声较多的特性, 这使得基于多元正态分布假设的 DRIFTLENS 在实际应用时可能出现较高的误判率或漏判风险. 这一局限性促使研究者探索更为灵活的分布比较方法.

为了解决正态分布假设带来的限制, Bu 等[62] 提出了基于最小二乘密度差的概念漂移检测方法, 其使用的散度定义如下:

$$D^2(p,q)=\int(p(x)-q(x))^2\mathrm{d}x. \tag{28}$$

其中$p(x)$和$q(x)$分别为连续概率密度函数. 该方法采用三级阈值机制, 包括安全阈值、警告阈值和变化阈值, 以提高对漂移的敏感性. 然而, 由于计算散度时需要对概率密度函数进行较为复杂的估计, 该方法在面对工业现场中动态且高噪声数据时, 计算开销较大且容易受到异常值的干扰, 导致阈值设置成为一大挑战. 正因如此, 研究者进一步引入了非参数检验方法, 以规避对具体分布形式的依赖. 增量 KS 检验[63] 正是在这种背景下被提出, 其通过计算两个数据窗口经验累积分布函数之间的最大差异来检测漂移, 如下所示:

$$D=\frac{1}{|A|}\max\{(\max_{x\in A\cup B}G(x)),-(\min_{x\in A\cup B}G(x))\}. \tag{29}$$

其中$G(x)=F'_A(x)-F'_B(x)$, $F'_A(x)$和$F'_B(x)$分别为两个分布的经验累积分布函数. 当统计量超过阈值时, 拒绝原假设, 表明检测到概念漂移. 该方法的一个显著优势是无需对数据分布做出任何假设, 因此在面对工业数据时具有较好的通用性. 然而, 工业过程中常伴随自然波动和过程噪声, 增量 KS 检验对微小波动极为敏感, 容易在正常波动中误判漂移, 从而影响系统的稳定性和可靠性. 为了解决这一问题, 进一步减少对分布细节的依赖, Gözüaçık 等[64] 提出了判别式漂移检测器 (discriminative drift detector, D3), 该方法摒弃了直接计算 KL 散度等复杂过程, 通过构建带滑动窗口的判别分类器, 利用 AUC 值的变化衡量特征空间的连续差异, 若满足$\mathrm{AUC}(C,S)\geqslant\tau$则表示概念漂移发生. 在工业环境中, 这种方法由于避免了对全局分布的精细估计, 在一定程度上降低了计算成本, 并对复杂数据具有更好的适应性, 但同时也对滑动窗口大小和分类器参数较为敏感, 需要针对具体生产过程行精细调校.

3.2.2 基于模型参数变化的方法

考虑到直接对数据分布进行检测在某些情况下可能无法捕捉到内部系统变化的全部信息, 另一类方法转而关注模型自身参数的变化. Wang 等[65] 提出的基于深度学习模型参数跟踪的概念漂移检测方法 (model-centric concept drift detection, MCDD) 是此类方法的代表. 该方法通过预训练和迁移学习获取深度模型的权重, 并实时监控这些参数的变化, 以判断是否发生了概念漂移. 在工业场景中, 尤其是当

系统内部特征较为隐蔽时, 这种方法可以提前捕捉到微妙的变化. 然而, 其检测效果高度依赖于预训练模型的质量, 同时模型参数的变化也可能受到多种非漂移因素的影响, 为结果的解释带来一定困难.

3.2.3 基于预测结果的方法

鉴于上述方法在直接监控数据分布或模型参数时均可能存在局限性, 近年来研究者又转向关注模型预测结果的变化来进行漂移检测. Cerqueira 等[66]提出了一种基于师生学习范式的无监督概念漂移检测方法 STUDD, 该方法流程如图 8 所示, 图中T为教师模型, S为学生模型. 通过创建一个辅助模型(学生模型) 来模仿主要模型 (教师模型) 的行为, 通过使用教师模型来预测新的实例, 并监控学生模型在运行期间的模仿损失, 最后采用 PageHinkley 方法跟踪模仿损失进行漂移检测, 同时在多个实际工业数据集上进行了验证. 该方法直接反映了模型预测性能的异常, 对于工业生产中对产品质量和过程效率的严格要求具有较强的针对性; 但与此同时, 维护两个模型的复杂性和噪声对模仿损失的干扰也增加了系统实施的难度和调校工作量. 类似地, Nunes 等[67]提出了一种新的基于典型化和离心数据分析的无监督概念漂移检测器 (concept drift detector based on typicality and eccentricity data analytics, TEDA-CDD), 该方法使用两个模型监测数据流特征, 并使用 Jaccard Index(JI) 度量两个模型的相似性. 根据该方法, 当相似性显著降低时, 表明发生了概念漂移. 这种方法在工业应用中能够更直观地反映数据流特征的变化, 但其参数的调整和模型维护同样需要在实际生产环境中反复验证和优化.

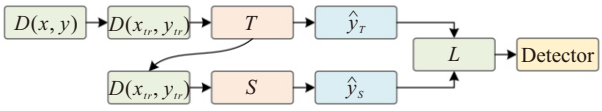


图8 基于师生学生范式的无监督概念漂移检测器

此外, 还有部分方法通过直接衡量模型在数据集上的泛化误差来进行漂移检测. 例如, Oikarinen 等[30]提出了基于泛化误差的漂移检测方法, 通过计算数据集上的泛化误差评估漂移风险. 泛化误差定义如下:

$$\mathrm{RMSE}(f,D)=\left(\sum_{i=1}^{n}[f(x_i')-y_i']^2/n\right)^{1/2}. \quad (30)$$

其中: $f(x_i')$为段模型在样本x_i'上的预测值, y_i'为整体回归模型在样本x_i'上的预测值. 这种方法在工业场景中具有直接关联模型性能的优势, 使得检测结果与实际生产中的风险紧密对应; 然而, 其缺点在于往往只能在漂移较为明显、模型性能明显下降后才能检测出来, 从而可能延迟对问题的响应. 类似地, Mayaki 等[31]提出的基于自回归模型的概念漂移检测方法 (autogressive drift detection method, ADDM) 以及 Zenisek 等[32]在工业径向风扇应用中采用的预测性能监控方法, 通过分析机器学习模型预测值与时间序列模型预测值之间误差的变化, 直接反映了工业系统中预测精度的波动, 但同样面临检测响应滞后的风险, 这在对实时性要求较高的生产环境中可能带来不利影响.

总体而言, 工业现场的数据往往呈现出高维、噪声多且分布复杂的特征, 各种无监督概念漂移检测方法各具优势与局限. 采用统计检验和散度度量的方法尽管在理论上具有严谨的数学基础, 但在面对复杂非理想数据分布时, 其适用性和鲁棒性可能受到制约; 基于模型参数变化的方法能够敏锐捕捉系统内部细微变动, 然而其效果高度依赖于预训练模型的质量, 同时对参数变化与概念漂移之间因果关系的解释能力相对较弱; 而基于预测结果的方法则更贴合工业实际, 能够直接反映系统性能的波动, 但在某些情况下可能存在响应滞后的问题. 因此, 工业实际应用中往往需要综合利用多种方法的优势, 通过混合或多层次的检测策略构建出既实时又稳健的概念漂移检测系统, 以切实保障生产过程的安全、稳定和高效运行.

4 工业场景下类不平衡与概念漂移检测方法

随着智能制造技术的持续发展与完善, 工业设备运行的可靠性显著提升, 故障或异常事件的发生频率显著降低. 这一趋势导致异常数据在整体数据中占比极低, 与正常数据相比呈现出严重的不平衡特性. 在此背景下, 类不平衡与概念漂移问题成为工业场景中的关键挑战. 类不平衡通常会导致分类器在少数类上的泛化能力较差, 而概念漂移的存在进一步加剧了这一问题. 为同时解决类不平衡与概念漂移的挑战, 许多研究致力于开发更高效的学习算法. 本节将系统性回顾已在实际数据集上验证的基于块结构与基于在线方式的类不平衡概念漂移检测方法, 旨在为工业应用提供实践参考.

4.1 基于块结构的类不平衡概念漂移检测

在基于块结构的方法中, 通常采用欠采样和过采样技术来缓解类别不平衡问题, 从而构建出平衡的数据集以训练分类器. 例如, 基于欠采样的方法在工业应用中常用于通过保留过去数据块中关键的少数类 (例如设备异常或故障) 样本, 同时对当前数据

块中的多数类 (正常运行状态) 样本进行随机抽样, 从而达到样本平衡[68]. 设当前数据块中多数类样本数为N_M和少数类样本数为N_m, 常见的策略是将多数类样本数调整为$\widetilde{N}_M = N_m$, 以构建平衡训练集. 这种方法具有实现简单、计算高效的优点, 适用于实时生产数据的离线批处理. 但在实际工业环境中, 由于多数类中也可能包含部分边缘信息, 过度欠采样可能会丢失这些有效信息, 从而影响整体分类性能.

另一类方法则利用过采样技术生成新的少数类样本形成平衡数据集, 其生成样本的质量对分类精度和学习效率具有直接影响[69]. 例如, Ditzler 等[69]在 Learn++.NSE 集成模型中引入合成少数类过采样技术 (synthetic minority class oversampling technique, SMOTE). 其基本原理是在少数类样本x_i及其最近邻x_{nn}间通过线性插值生成新样本, 即

$$x_{new} = x_i + \lambda(x_{nn} - x_i),\ \lambda \sim U(0,1). \tag{31}$$

扩展了少数类样本空间. 该方法在工业数据中能够有效提高模型对稀有故障或异常的检测率, 但传统 SMOTE 对所有少数类样本采用固定邻居数, 未能适应概念漂移后数据分布的动态变化, 降低了模型对新概念的适应性. 为此, 文献 [70] 提出 AnnSMOTE (SMOTE with adaptive nearest neighbors) 方法, 根据信息统计结果动态调整邻居数量, 使得生成的新样本更贴合最新数据分布, 从而提高模型在工业现场面对概念漂移时的适应性和鲁棒性.

此外, 针对块结构方法在数据块选择上的局限性, Lu 等[71]提出了一种基于块的增量学习方法, 即基于自适应块的动态加权多数方法 (adaptive chunk-based dynamic weighted majority, ACDWM). 该方法利用统计假设检验自适应确定数据块的大小, 从而使模型在应对概念漂移时能够充分利用历史数据的统计特性, 实现动态加权更新. 对于工业应用而言, ACDWM 方法能够更好地捕捉长期运行中设备状态的微妙变化, 并及时调整决策模型, 尽可能减少因突变漂移带来的误判风险.

在多类不平衡问题中, 其复杂性和挑战性更为突出. 图 9 直观展示了多类不平衡场景下概念漂移的现象, 例如在一条生产线上, 不同类型的异常事件各自发生频率低且存在相互干扰的情况. 虽然已有部分研究关注多类不平衡问题, 例如 Wang 等[72]探讨了具有可变不平衡率的虚拟概念漂移, Mirza 等[73]研究了固定不平衡比率下的概念漂移, 但针对可变不平衡比率的系统解决方案依然相对欠缺. 针对这一挑战, 文献 [74] 首次提出了元认知在线顺序极限学习机 (meta-cognitive online sequential extreme learning machine, MOS-ELM), 该方法通过元认知框架将重采样与代价敏感学习相结合, 有效解决了逐渐漂移与突变漂移情况下的多类不平衡问题. 在工业场景中, MOS-ELM 能够根据不断变化的生产数据自动调整采样策略与学习参数, 从而更好地应对不同故障模式和异常事件的识别任务.

总体而言, 基于块结构的类不平衡概念漂移检测方法在工业应用中具有显著优势, 能够充分利用历史数据的统计特性实现数据平衡和模型更新. 然而, 这类方法对数据块大小的选择、样本生成过程以及采样策略的动态调整较为敏感, 在面对突变漂移时响应可能不够及时. 因此, 在实际应用中, 需结合具体工业场景对方法参数进行微调, 以确保生产过程的安全、稳定和高效运行.

4.2　基于在线方式的类不平衡概念漂移检测

虽然基于块结构的方法在某些工业场景中能够取得不错的效果, 但由于其对数据块大小的自适应性有限, 往往难以捕捉单个数据块内发生的细微漂移. 在许多实际工业应用中, 如实时监控生产线异常、设备故障预警和网络流量异常检测, 数据通常以高速流式传输, 此时采用基于在线方式的检测方法显得尤为重要. 这类方法能够逐样处理数据流, 迅速响应概念漂移变化, 从而更及时地维护系统稳定性和生产安全.

早期在线方法中, 在线 Bagging(online bagging, OB) 方法通过泊松分布实现在线重采样, 从而近似生成平衡的训练数据以训练分类器[75]. 在工业现场, 比如实时监控系统中, OB 方法凭借其实现简单和计

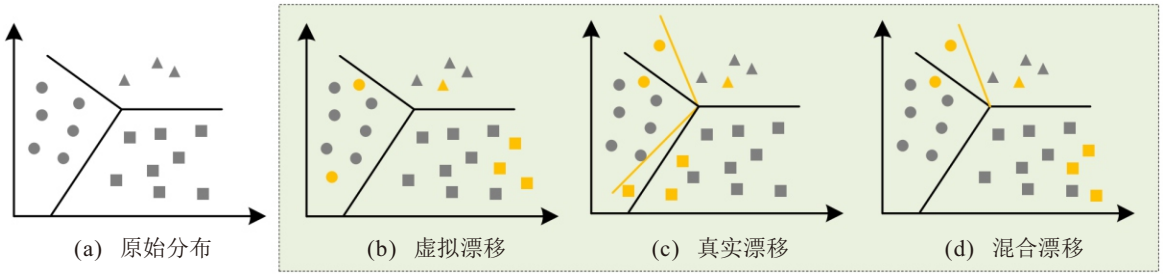


图9　多类不平衡概念漂移

算高效的特点, 便于快速部署和更新. 然而, 工业数据往往伴有高噪声, 因此 OB 方法在面对数据采样波动时可能导致检测结果不稳定, 进而影响异常事件的及时识别. 为了提高检测精度, Wang等[76] 提出了在线类不平衡漂移检测方法 (drift detection method for online class imbalance, DDM-OCI), 通过监测少数类召回率来检测类不平衡学习中概念漂移. 此方法在工业应用中能够较为敏感地捕捉少数类异常事件 (如罕见设备故障或质量异常), 但其检测效果高度依赖于召回率的稳定性, 容易受到数据波动的影响而产生误报, 特别是在噪声较大的生产环境中. 在此基础上, Vafaie 等[77] 提出了一种改进的在线集成方法, 其核心思路是基于实时监测的召回率对数据实例进行动态在线重采样. 当少数类的召回率低于平均水平时, 该方法采用过采样策略生成更多少数类样本; 反之, 则对多数类进行欠采样. 这样不仅可以平衡数据分布, 还能在实时性与准确性之间达到较好的平衡, 但需要针对不同工业数据的具体特性调优采样策略和阈值, 以确保在复杂环境下的鲁棒性.

此外, 为解决多类不平衡数据流中概念漂移问题, Han 等[78] 设计了一种自适应主动学习方法 (AdaAL-MID), 结合动态阈值矩阵与不确定性策略, 优化少数类样本的标签请求与分类性能. 在实际工业场景中, 如多故障类型设备监控, AdaAL-MID 能够根据数据流中各类别的变化动态调整采样和学习策略, 从而提高对各类异常事件的识别率. 与此同时, Liu 等[79] 提出了一种基于在线主动学习的框架, 专门用于应对网络流量中的多类不平衡与概念漂移问题. 该框架通过不断更新模型参数和采样策略, 并结合实时反馈机制, 显著提升了整体检测性能, 为动态工业数据处理提供了一种新颖而有效的解决方案, 尤其适用于工业网络安全监控和流量异常检测等场景.

总体而言, 基于在线方式的检测方法因其能够实时处理数据流、迅速响应概念漂移而在工业应用中具有较大优势. 然而, 这类方法对采样策略和阈值设定依赖较强, 容易受到噪声和数据波动的影响. 因此, 如何在保证实时性的同时兼顾检测的鲁棒性, 仍然是未来研究的重要课题.

5 工业场景下概念漂移适应方法

在工业实际应用中, 概念漂移对监控系统、预测模型和设备状态评估等环节具有重要影响. 为保障生产安全与高效运行, 概念漂移检测往往与相应的适应机制相结合, 以便在漂移发生时及时更新先验知识和学习模型. 根据适应方式的不同, 概念漂移适应策略主要分为基于主动检测的策略和基于被动适应的策略.

5.1 基于主动检测的概念漂移适应

基于主动检测的适应策略要求在模型学习过程中嵌入漂移检测算法, 当检测到概念漂移时立即触发适应机制, 更新或调整模型. 这类方法可进一步划分为 3 大类: 简单再训练、保留旧模型再训练 (基于集成学习) 以及模型调整.

5.1.1 简单再训练

简单再训练方法在检测到漂移后, 利用最新数据重新训练整个模型, 以取代旧模型, 如图 10 所示. 一般情况下, 重新训练操作发生在检测到概念漂移的具体位置[80-82]. 例如, 设在概念漂移检测到达时刻 t_d 后, 构建新的训练集 D_{new} 并求解损失函数

$$\theta^* = \arg\min_{\theta} L(\theta; D_{\text{new}}), \tag{32}$$

从而获得新的模型参数 θ^*. 在实际应用中, 简单再训练方法可以结合特定的机器学习算法进行优化. 例如, 文献 [83] 结合自适应调整隐藏层节点数的方式处理概念漂移. 当分类错误率增加时, 可能表明概念漂移的存在, 此时通过增加网络节点数提升分类能力. 这种方法的优点在于响应迅速、实现简单, 特别适用于工业生产线中出现明显故障或状态突变的紧急情况; 但其缺点在于每次漂移均需要完全重训模型, 计算成本较高, 并可能在重新训练期间影响系统的实时服务能力.

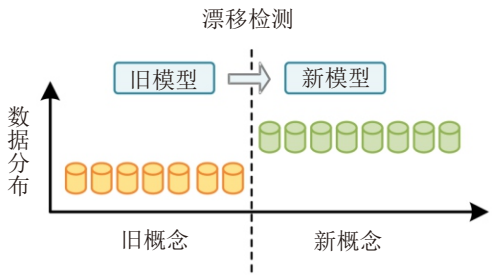


图10 简单再训练适应策略

5.1.2 保留旧模型再训练

为了降低计算成本并缓解频繁重训练带来的风险, 一种策略是保留部分历史模型信息, 通过集成学习实现平滑过渡. 此方法通常包括一组分类器, 通过一定的投票规则组合分类器结果[84-85]. 如图 11 所示, 这种策略并非直接用新模型替换旧模型, 而是将新模型加入旧模型集. 例如, Kolter 等[86] 提出的动态加权多数集成方法 (dynamic weighted majority, DWM) 将多个基本分类器进行加权投票, 其预测结果为

$$\hat{y} = \text{argmax}_y \sum_i w_i \cdot I(h_i(x) = y). \quad (33)$$

其中: $I(\cdot)$为示性函数, 权重w_i根据分类器在新数据上的表现动态调整. 当检测到漂移时, 若某基分类器的错误率过高, 则其权重降低甚至被移除, 同时引入新的基本分类器以增强系统适应性. 该方法在工业现场 (如设备状态监控和生产线异常诊断中) 具有较高的鲁棒性和容错性, 但需要维护一个分类器集成, 系统结构相对复杂, 且权重更新策略的设计需要精细调校.

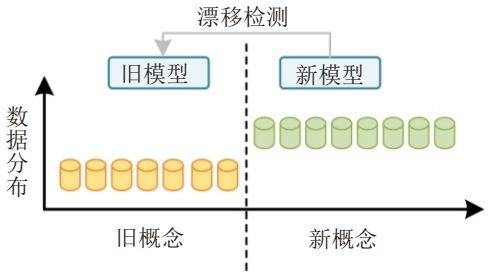


图11　保留旧模型再训练适应策略

5.1.3　模型调整

在深度学习领域, 模型调整方法通过微调现有模型的参数或结构, 以适应概念漂移带来的数据分布变化, 如图 12 所示. 当检测到概念漂移后, 采用参数更新和模型结构调整的方法适应漂移[87]. 例如, Diez-Olivan 等[88] 提出了一种基于树突状细胞算法的概念漂移检测与动态自适应学习方法. 在检测到概念漂移后, 该方法利用核密度估计器生成合成数据$D_{\text{synthetic}}$, 然后通过梯度下降对部分网络参数θ_{partial}进行微调, 有

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \nabla L(\theta_{\text{old}}; D_{\text{synthetic}}). \quad (34)$$

其中η为学习率. 这种方法能够在不完全重训的情况下快速恢复模型性能, 计算开销较低, 特别适用于工业实时监控的场景; 但若漂移幅度过大, 则微调可能不足以使模型完全适应新分布, 从而影响预测准确性.

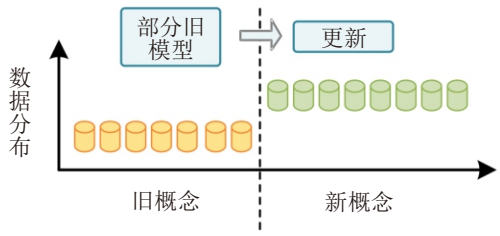


图12　部分模型调整适应策略

综上所述, 基于主动检测的适应策略能够在检测到漂移时快速响应, 适用于对漂移变化敏感且要求即时调整的工业场景, 如实时故障预警和生产质量监控. 然而, 其依赖于准确的漂移检测和高效的适应机制, 一旦检测失误或适应策略不当, 可能会导致系统过度反应或适应不足, 进而影响整体性能.

5.2　基于被动方式的概念漂移适应

与主动检测不同, 被动适应策略无需显式检测漂移, 而是在新数据不断到达时, 持续更新模型参数以逐步适应数据分布的变化. 根据更新方式不同, 基于被动适应的方法主要分为单一分类器方法和基于集成的适应方法.

5.2.1　单一分类器方法

在单一分类器方法中, Hulten 等[89] 提出了概念漂移快速决策树 (concept drift very fast decision tree, CVFDT), 该方法通过自适应窗口以增量方式处理非平稳环境中的数据变化. 这种方法计算成本低, 适合大规模数据流场景, 如工业传感器数据实时监控或设备状态评估. 但在面对突变型漂移时, 其响应可能滞后, 影响异常事件的及时检测. 为此, Liu 等[90] 提出了一种改进的 CVFDT, 即高效 CVFDT(efficient CVFDT, E-CVFDT). E-CVFDT 引入缓存机制, 有效处理了突然型和渐变型等多种漂移类型. 基于单一分类器的被动适应方法在计算成本上具有优势, 因此适合于大多数数据流场景.

5.2.2　基于集成的被动适应方法

基于集成的适应方法通过构建多个候选分类器, 并定期更新其权重或结构来应对数据流中不断变化的模式. 例如, Li 等[91] 提出动态更新集成 (dynamic updating ensemble, DUE) 方法. 所提出 DUE 为每个数据块D_k上创建若干候选分类器$\{h_1, h_2, \cdots, h_n\}$, 并根据各分类器在$D_k$上的准确率动态调整权重, 有

$$w_i = f(\text{acc}(h_i; D_k)), \quad (35)$$

最终通过加权投票得到预测结果. 集成方法在处理数据流和概念漂移方面具有较高的鲁棒性和适应性, 特别适用于数据分布复杂且变化频繁的工业场景, 如网络流量异常检测和多传感器数据融合. 但其缺点在于系统设计较为复杂, 计算资源消耗较大, 对实时性要求较高的场合可能需要优化实现策略.

综上所述, 基于被动方式的适应策略适合数据分布变化较为平缓的环境, 其更新机制无需依赖显式的漂移检测, 系统设计较为简单, 适合大规模数据流场景. 然而, 在工业应用中, 数据往往存在高噪声和突变情况, 如何在保持实时性更新的同时保证检测的鲁棒性, 仍然是未来研究的重点.

6　未来展望

近年来, 概念漂移检测与适应领域取得了显著

进展, 但随着数据复杂性和应用场景的不断拓展, 仍然存在许多挑战性的研究方向和未解难题. 未来研究可以从以下几个方向深入探索.

6.1 高维数据与异构数据中的概念漂移检测

随着数据复杂性和维度的不断增加, 高维数据和异构数据场景日益普遍. 传统的概念漂移检测方法在处理高维特征时, 可能面临维度灾难问题带来的挑战. 此外, 高维数据中的特征交互关系更加复杂, 如何有效捕获和建模特征间的关系成为研究重点. 对于异构数据, 概念漂移检测的复杂性来源于数据模态的差异. 例如, 在多模态数据 (如文本、图像和时间序列) 中, 不同模态数据可能具有各自独特的漂移特征, 这些特征需要被统一建模[92-94]. 如何开发适用于高维和异构数据的高效检测算法, 是未来研究的一个重要方向, 这可能涉及深度学习模型的自适应学习、图结构学习、以及特征降维与特征选择技术的结合.

6.2 实时性与性能的平衡

在资源受限的工业环境中, 例如边缘计算设备或实时监控系统, 如何平衡概念漂移检测的速度与准确性是一个关键问题[95]. 实时性需求通常要求算法具有高效的计算性能, 但这往往会以一定的检测精度为代价. 未来的研究需要重点探索轻量级、高效的在线检测算法, 可能结合增量学习、稀疏建模等技术, 以在资源受限环境中实现检测速度与准确性的最优平衡.

6.3 概念漂移中新类别的挑战

在实际应用场景中, 数据流的变化不仅可以表现为现有类别分布的改变, 还可以涉及新类别的出现[96-97]. 例如, 在实时监控系统中, 可能会检测到从未见过的异常模型. 在新类别刚刚出现时, 往往只有极少量样本, 这种情况对模型的学习能力提出了严格要求. 如何在极少样本或零样本的情况下实现对新类别的快速适应, 是未来研究的重要方向之一, 需要结合零样本学习、小样本学习以及基于先验知识或迁移学习的方法.

7 结 语

本文系统性地探讨和综述了适用于工业场景的概念漂移检测与适应方法. 针对有监督的概念漂移检测方法, 主要总结了基于性能、窗口以及集成的检测策略. 然而, 在工业实际应用中, 标签稀缺和类别不平衡问题常对概念漂移检测产生显著影响. 对此, 本文分别从半监督和无监督的角度探讨了解决标签稀缺问题的各类方法, 并针对类别不平衡问题提出了基于块处理和在线更新的两类应对方案. 此外, 本文详细归纳了概念漂移适应方法, 涵盖了主动检测触发的概念漂移策略与被动适应策略的特点及应用场景. 通过对比各方法的优缺点和适用条件, 本文为工业环境中应对复杂动态的数据分布提供了方法论指导和实践参考. 总之, 针对工业生产中不断变化的数据分布, 研究人员可根据具体应用场景, 进一步探索和融合多种概念漂移检测与适应技术, 以构建高效的智能监控系统, 保障生产过程的稳定性和持续优化.

参考文献 (References)

[1] Lin C C, Deng D J, Kuo C H, et al. Concept drift detection and adaption in big imbalance industrial IoT data using an ensemble learning method of offline classifiers[J]. IEEE Access, 2019, 7: 56198-56207.

[2] Oztemel E, Gursev S. Literature review of industry 4.0 and related technologies[J]. Journal of Intelligent Manufacturing, 2020, 31(1): 127-182.

[3] Hu Y J, Jia Q M, Yao Y, et al. Industrial internet of things intelligence empowering smart manufacturing: A literature review[J]. IEEE Internet of Things Journal, 2024, 11(11): 19143-19167.

[4] Chi H R, Wu C K, Huang N F, et al. A survey of network automation for industrial internet-of-things toward industry 5.0[J]. IEEE Transactions on Industrial Informatics, 2023, 19(2): 2065-2077.

[5] de Barros R S M, de Carvalho Santos S G T. An overview and comprehensive comparison of ensembles for concept drift[J]. Information Fusion, 2019, 52: 213-244.

[6] 霍海丹, 阎高伟, 程兰, 等. 基于低秩重构表示的动态回归迁移模型[J]. 控制与决策, 2024, 39(8): 2511-2520. (Huo H D, Yan G W, Cheng L, et al. Dynamic transfer regression model based on low-rank reconstruction representation[J]. Control and Decision, 2024, 39(8): 2511-2520.)

[7] Gama J, Žliobaitė I, Bifet A, et al. A survey on concept drift adaptation[J]. ACM Computing Surveys, 2014, 46(4): 1-37.

[8] Yang Z Y, Zheng J H, Ge Z Q. Lifelong Bayesian learning machines for streaming industrial big data[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2023, 53(3): 1554-1565.

[9] Hovakimyan G, Bravo J M. Evolving strategies in machine learning: A systematic review of concept drift detection[J]. Information, 2024, 15(12): 786.

[10] Sun L J, Ji Y J, Zhu M R, et al. A new predictive method supporting streaming data with hybrid recurring concept drifts in process industry[J]. Computers & Industrial

Engineering, 2021, 161: 107625.

[11] Grote-Ramm W, Lanuschny D, Lorenzen F, et al. Continual learning for neural regression networks to cope with concept drift in industrial processes using convex optimisation[J]. Engineering Applications of Artificial Intelligence, 2023, 120: 105927.

[12] Zhang T M, Yan G W, Ren M F, et al. Dynamic transfer soft sensor for concept drift adaptation[J]. Journal of Process Control, 2023, 123: 50-63.

[13] Yang Z, Al-Dahidi S, Baraldi P, et al. A novel concept drift detection method for incremental learning in nonstationary environments[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(1): 309-320.

[14] Sun Z J, Tang J, Qiao J F, et al. Review of concept drift detection method for industrial process modeling[C]. 2020 39th Chinese Control Conference. Shenyang, 2020: 5754-5759.

[15] Lu J, Liu A J, Dong F, et al. Learning under concept drift: A review[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(12): 2346-2363.

[16] Bayram F, Ahmed B S, Kassler A. From concept drift to model degradation: An overview on performance-aware drift detectors[J]. Knowledge-Based Systems, 2022, 245: 108632.

[17] Lima M, Neto M, Filho T S, et al. Learning under concept drift for regression — A systematic literature review[J]. IEEE Access, 2022, 10: 45410-45429.

[18] Li J L, Hsu C F, Chang M C, et al. A comprehensive review of machine learning advances on data change: A cross-field perspective[J/OL]. 2024, arXiv: 2402.12627.

[19] Xiang Q Y, Zi L L, Cong X, et al. Concept drift adaptation methods under the deep learning framework: A literature review[J]. Applied Sciences, 2023, 13(11): 6515.

[20] Schlimmer J C, Granger R H. Incremental learning from noisy data[J]. Machine Learning, 1986, 1(3): 317-354.

[21] Baena-Garc M, Jose D C A, Fidalgo R, et al. Early drift detection method[J]. International Workshop on Knowledge Discovery from Data Streams, DOI: oai:eprints.pascal-network.org:2509.

[22] Jameel S M, Ahmed M, Alhussain H, et al. A critical review on adverse effects of concept drift over machine learning classification models[J]. International Journal of Advanced Computer Science and Applications, 2020, 11(1): 0110127.

[23] Ramírez-Gallego S, Krawczyk B, García S, et al. A survey on data preprocessing for data stream mining: Current status and future directions[J]. Neurocomputing, 2017, 239: 39-57.

[24] Yuan L, Li H, Xia B, et al. Recent advances in concept drift adaptation methods for deep learning[C]. IJCAI. Vienna, 2022: 5654-5661.

[25] Gama J, Medas P, Castillo G, et al. Learning with drift detection[C]. The 17th Brazilian Symposium on Artificial Intelligence. Sao Luis, 2004: 286-295.

[26] Escovedo T, Koshiyama A, da Cruz A A, et al. DetectA: Abrupt concept drift detection in non-stationary environments[J]. Applied Soft Computing, 2018, 62: 119-133.

[27] Alippi C, Liu D R, Zhao D B, et al. Detecting and reacting to changes in sensing units: The active classifier case[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2014, 44(3): 353-362.

[28] Gama J, Sebastião R, Rodrigues P P. On evaluating stream learning algorithms[J]. Machine Learning, 2013, 90(3): 317-346.

[29] Cai S H, Zhao Y W, Hu Y K, et al. CD-BTMSE: A concept drift detection model based on bidirectional temporal convolutional network and multi-stacking ensemble learning[J]. Knowledge-Based Systems, 2024, 294: 111681.

[30] Oikarinen E, Tiittanen H, Henelius A, et al. Detecting virtual concept drift of regressors without ground truth values[J]. Data Mining and Knowledge Discovery, 2021, 35(3): 726-747.

[31] Mayaki M Z A, Riveill M. Autoregressive based drift detection method[C]. International Joint Conference on Neural Networks. Padua, 2022: 1-8.

[32] Zenisek J, Holzinger F, Affenzeller M. Machine learning based concept drift detection for predictive maintenance[J]. Computers & Industrial Engineering, 2019, 137: 106031.

[33] Wang P F, Jin N L, Fehringer G. Concept drift detection with False Positive rate for multi-label classification in IoT data stream[C]. International Conference on UK-China Emerging Technologies. Glasgow, 2020: 1-4.

[34] Harries M, Wales N S. Splice-2 comparative evaluation: Electricity pricing[D]. Sydney: University of New South Wales, 1999.

[35] Yan M M W. Accurate detecting concept drift in evolving data streams[J]. ICT Express, 2020, 6(4): 332-338.

[36] Kahraman A, Kantardzic M, Kotan M. Dynamic modeling with integrated concept drift detection for predicting real-time energy consumption of industrial machines[J]. IEEE Access, 2022, 10: 104622-104635.

[37] Aguiar G J, Cano A. A comprehensive analysis of concept drift locality in data streams[J]. Knowledge-Based Systems, 2024, 289: 111535.

[38] Han M, Mu D L, Li A, et al. Concept drift detection methods based on different weighting strategies[J]. International Journal of Machine Learning and Cybernetics, 2024, 15(10): 4709-4732.

[39] Estaji A, Götzinger M, Tutzer B, et al. Evaluation of drift detection algorithms in the condition monitoring domain[J]. IEEE Transactions on Industrial Informatics, 2025, 21(1): 317-326.

[40] Chikushi R T M, de Barros R S M, da Silva M G N M, et al. Using spectral entropy and bernoulli map to handle concept drift[J]. Expert Systems with Applications, 2021, 167: 114114.

[41] Xu L, Ding X, Peng H, et al. ADTCD: An adaptive anomaly detection approach toward concept drift in iot[J]. IEEE Internet of Things Journal, 2023, 10(18): 15931-15942.

[42] Zhang S S, Liu J W, Zuo X. Adaptive online incremental learning for evolving data streams[J]. Applied Soft Computing, 2021, 105: 107255.

[43] Pesaranghader A, Viktor H, Paquet E. McDiarmid drift detection methods for evolving data streams[J/OL]. 2017, arXiv: 1710.02030.

[44] Zhang Y Y, Liu Z, Yang C J, et al. Unveiling dynamics changes: Singular spectrum analysis-based method for detecting concept drift in industrial data streams[J]. Knowledge-Based Systems, 2024, 293: 111640.

[45] Du L, Song Q B, Jia X L. Detecting concept drift: An information entropy based method using an adaptive sliding window[J]. Intelligent Data Analysis, 2014, 18(3): 337-364.

[46] Raeiszadeh M, Ebrahimzadeh A, Glitho R H, et al. Real-time adaptive anomaly detection in industrial IoT environments[J]. IEEE Transactions on Network and Service Management, 2024, 21(6): 6839-6856.

[47] Gözüaçık Ö, Can F. Concept learning using one-class classifiers for implicit drift detection in evolving data streams[J]. Artificial Intelligence Review, 2021, 54(5): 3725-3747.

[48] Guo H S, Li H, Ren Q Y, et al. Concept drift type identification based on multi-sliding windows[J]. Information Sciences, 2022, 585: 1-23.

[49] Raab C, Heusinger M, Schleif F M. Reactive soft prototype computing for concept drift streams[J]. Neurocomputing, 2020, 416: 340-351.

[50] Xu L J, Han Z Y, Zhao D W, et al. Addressing concept drift in IoT anomaly detection: Drift detection, interpretation, and adaptation[J]. IEEE Transactions on Sustainable Computing, 2024, 9(6): 913-924.

[51] Maciel B I F, Santos S G T C, Barros R S M. A lightweight concept drift detection ensemble[C]. IEEE 27th International Conference on Tools with Artificial Intelligence. Vietri sul Mare, 2015: 1061-1068.

[52] Toumi H, Brahmi Z, Gammoudi M M. Extended hoeffding adaptive tree based-server load prediction in cloud computing environment[C]. Proceedings of the International Conference on High Performance Computing in Asia-Pacific Region. Fukuoka, 2020: 161-168.

[53] Pérez J L M, Barros R S M, Santos S G T C. Statistical tests ensemble drift detector[C]. IEEE Symposium Series on Computational Intelligence. Canberra, 2020: 1021-1028.

[54] Du L, Song Q B, Zhu L, et al. A selective detector ensemble for concept drift detection[J]. The Computer Journal, 2015, 58(3): 457-471.

[55] Khezri S, Tanha J, Ahmadi A, et al. STDS: Self-training data streams for mining limited labeled data in non-stationary environment[J]. Applied Intelligence, 2020, 50(5): 1448-1467.

[56] Tanha J, Samadi N, Abdi Y, et al. CPSSDS: Conformal prediction for semi-supervised classification on data streams[J]. Information Sciences, 2022, 584: 212-234.

[57] Gorgonio A C, de P Canuto A M, Vale K M O, et al. A semi-supervised based framework for data stream classification in non-stationary environments[C]. International Joint Conference on Neural Networks. Glasgow, 2020: 1-8.

[58] Wang P F, Yu H, Jin N L, et al. QuadCDD: A quadruple-based approach for understanding concept drift in data streams[J]. Expert Systems with Applications, 2024, 238: 122114.

[59] Tan C H, Lee V C, Salehi M. Online semi-supervised concept drift detection with density estimation[J/OL]. 2019, arXiv: 1909.11251.

[60] Ali U, Mahmood T. A novel framework for concept drift detection using autoencoders for classification problems in data streams[J]. International Journal of Machine Learning and Cybernetics, 2025, 16(1): 397-418.

[61] Greco S, Vacchetti B, Apiletti D, et al. Unsupervised concept drift detection from deep learning representations in real-time[J/OL]. 2024, arXiv: 2406.17813.

[62] Bu L, Alippi C, Zhao D B. A pdf-free change detection test based on density difference estimation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(2): 324-334.

[63] dos Reis D M, Flach P, Matwin S, et al. Fast unsupervised online drift detection using incremental Kolmogorov-Smirnov test[C]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, 2016: 1545-1554.

[64] Gözüaçık Ö, Büyükçakır A, Bonab H, et al. Unsupervised concept drift detection with a discriminative classifier[C]. Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing, 2019: 2365-2368.

[65] Wang P F, Jin N L, Davies D, et al. Model-centric

transfer learning framework for concept drift detection[J]. Knowledge-Based Systems, 2023, 275: 110705.

[66] Cerqueira V, Gomes H M, Bifet A, et al. STUDD: A student–teacher method for unsupervised concept drift detection[J]. Machine Learning, 2023, 112(11): 4351-4378.

[67] Nunes Y T P, Guedes L A. Concept drift detection based on typicality and eccentricity[J]. IEEE Access, 2024, 12: 13795-13808.

[68] Gao J, Fan W, Han J W, et al. A general framework for mining concept-drifting data streams with skewed distributions[C]. Proceedings of the 2007 SIAM International Conference on Data Mining. Piscataway: IEEE, 2007: 3-14.

[69] Ditzler G, Polikar R. Incremental learning of concept drift from streaming imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(10): 2283-2301.

[70] Jiao B T, Guo Y N, Gong D W, et al. Dynamic ensemble selection for imbalanced data streams with concept drift[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(1): 1278-1291.

[71] Lu Y, Cheung Y M, Tang Y Y. Adaptive chunk-based dynamic weighted majority for imbalanced data streams with concept drift[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 31(8): 2764-2778.

[72] Wang S, Minku L L, Yao X, et al. Dealing with multiple classes in online class imbalance learning[C]. Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, 2016: 2118-2124.

[73] Mirza B, Lin Z P, Liu N. Ensemble of subset online sequential extreme learning machine for class imbalance and concept drift[J]. Neurocomputing, 2015, 149: 316-329.

[74] Mirza B, Lin Z P. Meta-cognitive online sequential extreme learning machine for imbalanced and concept-drifting data classification[J]. Neural Networks, 2016, 80: 79-94.

[75] Oza N C. Online bagging and boosting[C]. IEEE International Conference on Systems, Man and Cybernetics. Waikoloa, 2005: 2340-2345.

[76] Wang S, Minku L L, Ghezzi D, et al. Concept drift detection for online class imbalance learning[C]. The 2013 International Joint Conference on Neural Networks. Dallas, 2013: 1-10.

[77] Vafaie P, Viktor H, Michalowski W. Multi-class imbalanced semi-supervised learning from streams through online ensembles[C]. International Conference on Data Mining Workshops. Sorrento, 2020: 867-874.

[78] Han M, Li C P, Meng F X, et al. An adaptive active learning method for multiclass imbalanced data streams with concept drift[J]. Applied Sciences, 2024, 14(16): 7176.

[79] Liu W K, Zhu C, Ding Z Y, et al. Multiclass imbalanced and concept drift network traffic classification framework based on online active learning[J]. Engineering Applications of Artificial Intelligence, 2023, 117: 105607.

[80] Kifer D, Ben-David S, Gehrke J, et al. Detecting change in data streams[C]. Proceedings of the 20th International Conference on Very Large Data Bases. New York, 2004: 180-191.

[81] Soares S G, Araújo R. An adaptive ensemble of on-line extreme learning machines with variable forgetting factor for dynamic system prediction[J]. Neurocomputing, 2016, 171: 693-707.

[82] Manly B F J, MacKenzie D. A cumulative sum type of method for environmental monitoring[J]. Environmetrics, 2000, 11(2): 151-166.

[83] Mariano-Hernández D, Hernández-Callejo L, Solís M, et al. Analysis of the integration of drift detection methods in learning algorithms for electrical consumption forecasting in smart buildings[J]. Sustainability, 2022, 14(10): 5857.

[84] Cano A, Krawczyk B. ROSE: Robust online self-adjusting ensemble for continual learning on imbalanced drifting data streams[J]. Machine Learning, 2022, 111(7): 2561-2599.

[85] Yang L, Shami A. IoT data analytics in dynamic environments: From an automated machine learning perspective[J]. Engineering Applications of Artificial Intelligence, 2022, 116: 105366.

[86] Kolter J Z, Maloof M A. Dynamic weighted majority: An ensemble method for drifting concepts[J]. Journal of Machine Learning Research, 2007, 8: 2755-2790.

[87] Guo L T, Lu J, An J P, et al. DSIL: An effective spectrum prediction framework against spectrum concept drift[J]. IEEE Transactions on Cognitive Communications and Networking, 2024, 10(3): 794-806.

[88] Diez-Olivan A, Ortego P, Del Ser J, et al. Adaptive dendritic cell-deep learning approach for industrial prognosis under changing conditions[J]. IEEE Transactions on Industrial Informatics, 2021, 17(11): 7760-7770.

[89] Hulten G, Spencer L, Domingos P, et al. Mining time-changing data streams[C]. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, 2001: 97-106.

[90] Liu G, Cheng H R, Qin Z G, et al. E-CVFDT: An improving CVFDT method for concept drift data stream[C]. International Conference on

Communications, Circuits and Systems. Chengdu, 2013: 315-318.

[91] Li Z, Huang W C, Xiong Y, et al. Incremental learning imbalanced data streams with concept drift: The dynamic updated ensemble algorithm[J]. Knowledge-Based Systems, 2020, 195: 105694.

[92] Li W F, Li B, Wang Z R, et al. A drift detection method for industrial images based on a defect segmentation model[J]. Knowledge-Based Systems, 2024, 301: 112320.

[93] Tran Q T, Kuppa A, Bertolotto M, et al. A new approach for concept drift detection in visual data[C]. Conference on Information Technology and its Applications. Cham: Springer Nature Switzerland, 2024: 172-183.

[94] Yang X Y, Lu J, Yu E. Adapting multi-modal large language model to concept drift from pre-training onwards[J/OL]. 2024, arXiv: 2405.13459.

[95] Souza V M A, Parmezan A R S, Chowdhury F A, et al. Efficient unsupervised drift detector for fast and high-dimensional data streams[J]. Knowledge and Information Systems, 2021, 63(6): 1497-1527.

[96] Chambers L, Gaber M M, Abdallah Z S. DeepStreamCE: A streaming approach to concept evolution detection in deep neural networks[J/OL]. 2020, arXiv: 2004.04116.

[97] Kuppa A, Le-Khac N A. Learn to adapt: Robust drift detection in security domain[J]. Computers and Electrical Engineering, 2022, 102: 108239.

作者简介

周平 (1980−), 男, 教授, 博士, 博士生导师, 主要研究方向为人工智能与机器学习算法及其在建模、控制与优化中的应用和工程实践, E-mail: zhouping@mail.neu.edu.cn;

张宇 (1998−), 女, 博士生, 主要研究方向为工业图像异常检测、工业概念漂移检测, E-mail: 2290097@stu.neu.edu.cn.

科研团队简介

周平教授科研团队立足于柴天佑院士领导的东北大学流程工业综合自动化全国重点实验室, 长期专注于复杂工业过程运行优化控制方法、技术及工程应用的研究. 目前, 课题组有在读博士生 6 人、在读硕士生 10 人. 近年来, 课题组在 Automatica、CEP、JPC 等 IFAC 会刊, IEEE T CST、IEEE T ASE 等 IEEE 汇刊以及《中国科学: 信息科学》《自动化学报》等国内外权威刊物发表期刊论文 140 余篇, 出版学术专著 4 部, 授权中美发明专利和计算机软件著作权 60 余件. 先后主持国家自然科学基金重大课题项目、国家自然科学基金区域创新联合基金重点项目等国家级科研项目多项. 课题负责人周平教授为国家级青年人才入选者, 曾获得中国自动化学会自然科学一等奖、教育部自然科学一等奖、浙江省科技进步一等奖等 6 项省部级科技奖, 以及中国自动化学会青年科技奖、首届中国自动化学会优秀博士学位论文奖等荣誉.