

# 基于主动-被动增量集成的概念漂移适应方法

祁晓博<sup>1,2</sup> 陈佳明<sup>1</sup> 史颖<sup>1,2,3</sup> 亓慧<sup>1,2</sup> 郭虎升<sup>3</sup> 王文剑<sup>4</sup>

**摘要** 数据流是一组随时间连续到来的数据序列,在数据流不断产生的过程中,由于各种因素的影响,数据分布随时间推移可能以不可预测的方式发生变化,这种现象称为概念漂移.在漂移发生后,当前模型需要及时响应数据流中的实时分布变化,并有效处理不同类型的概念漂移,从而避免模型泛化性能下降.针对这一问题,提出一种基于主动-被动增量集成的概念漂移适应方法(CDAM-APIE).该方法首先使用在线增量集成策略构建被动集成模型,对新样本进行实时预测以动态更新基模型权重,有利于快速响应数据分布的瞬时变化,并增强模型适应概念漂移的能力.在此基础上,利用增量学习和概念漂移检测技术构建主动基模型,提升模型在平稳数据流状态下的鲁棒性和漂移后的泛化性能.实验结果表明,CDAM-APIE能够对概念漂移做出及时响应,同时有效提高模型的泛化性能.

**关键词** 概念漂移,数据流分类,增量学习,在线集成

**引用格式** 祁晓博,陈佳明,史颖,亓慧,郭虎升,王文剑.基于主动-被动增量集成的概念漂移适应方法.自动化学报,2025,51(5):1131-1144

**DOI** 10.16383/j.aas.c240503 **CSTR** 32138.14.j.aas.c240503

## Concept Drift Adaptive Method Based on Active-Passive Incremental Ensemble

QI Xiao-Bo<sup>1,2</sup> CHEN Jia-Ming<sup>1</sup> SHI Ying<sup>1,2,3</sup> QI Hui<sup>1,2</sup> GUO Hu-Sheng<sup>3</sup> WANG Wen-Jian<sup>4</sup>

**Abstract** A data stream refers to a set of data sequences that arrive continuously over time. Due to various influencing factors, the data distribution may change in an unpredictable manner over time during the continuous generation of data streams, a phenomenon known as concept drift. After the drift occurs, the current model needs to respond promptly to the real-time distributional changes in the data stream and handle different types of concept drift efficiently, in order to avoid the degradation of the model generalization performance. Aiming at this problem, we propose a concept drift adaptation method based on the active-passive incremental ensemble (CDAM-APIE). Firstly, CDAM-APIE uses the online incremental ensemble strategy to construct a passive ensemble model, which makes real-time predictions on new samples to dynamically update the weights of the base model. It is beneficial for quickly responding to instantaneous changes in data distribution and enhancing the model's ability to adapt to concept drift. On this basis, an active basis model is constructed with incremental learning and concept drift detection techniques to improve the robustness of the model under steady data stream states and the generalization performance after drift. The experimental results show that CDAM-APIE can respond to concept drift promptly and effectively improve the generalization performance of the model.

**Key words** Concept drift, data stream classification, incremental learning, online ensemble

**Citation** Qi Xiao-Bo, Chen Jia-Ming, Shi Ying, Qi Hui, Guo Hu-Sheng, Wang Wen-Jian. Concept drift adaptive method based on active-passive incremental ensemble. *Acta Automatica Sinica*, 2025, 51(5): 1131-1144

收稿日期 2024-07-15 录用日期 2024-12-13

Manuscript received July 15, 2024; accepted December 13, 2024

国家自然科学基金(62476157, U21A20513, 62076154, 62276157),山西省专利转化专项计划项目(202302009, 202302012),山西省基础研究计划(自由探索类)项目(202103021223334),太原师范学院成果转化与技术转移基地(2023P003)资助

Supported by National Natural Science Foundation of China (62476157, U21A20513, 62076154, 62276157), the Shanxi Province Patent Transformation Special Programs (202302009, 202302012), the Basic Research Program (Free Exploration) of Shanxi Province (202103021223334), and Taiyuan Normal University Achievement Transformation and Technology Transfer Base (2023P003)

本文责任编辑 张敏灵

Recommended by Associate Editor ZHANG Min-Ling

1. 太原师范学院计算机科学与技术学院 晋中 030619 2. 智能优化计算与区块链技术山西省重点实验室 晋中 030619 3. 山西大学计算机与信息技术学院 太原 030006 4. 山西大学计算智能

大数据时代,数据流在医疗诊断、欺诈监测、气象预测等多个领域大量涌现<sup>[1-3]</sup>.相较于传统的静态数据,数据流通常以流的形式按时间顺序依次到达,具有时序性、动态性、无限性、不可重现性等特点<sup>[4-6]</sup>.数据流挖掘研究是为使在线学习模型更好地应对实时数据流中的动态变化,提高模型的适应性和泛化

与中文信息处理教育部重点实验室 太原 030006

1. School of Computer Science and Technology, Taiyuan Normal University, Jinzhong 030619 2. Shanxi Key Laboratory for Intelligent Optimization Computing and Blockchain Technology, Jinzhong 030619 3. School of Computer and Information Technology, Shanxi University, Taiyuan 030006 4. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006

性能<sup>[7]</sup>. 概念漂移是数据流挖掘中常见的一种现象, 其典型特征是样本的输入特征和输出标签之间的关系会随时间发生不可预见的变化<sup>[8-11]</sup>.

概念漂移会使基于历史数据训练的学习模型难以适应当前的实时数据变化<sup>[12-13]</sup>. 例如在信用卡欺诈检测中, 欺诈者更新一些伪装技术, 使欺诈特征随时间推移发生变化, 导致过去归为正常的交易记录可能在未来变成欺诈交易<sup>[14]</sup>; 在气象预测中, 相似的气温、压强、空气湿度等因素可能随季节变化造成不同的天气状况, 若模型更新不及时, 就会使当前的天气预测情况发生滞后<sup>[15]</sup>. 在工业生产及故障诊断领域, 工况或环境变化可能导致故障特征改变. 若故障诊断模型未能实时更新, 误诊和漏诊问题随时间累积, 可能引发生产中断、经济损失等问题. 因此, 在数据流挖掘中, 提高学习模型对概念漂移的适应能力, 维持其在数据变化下的准确性和有效性, 对于实际应用具有重要意义.

目前, 处理概念漂移数据流的方法大致分为主动检测方法和被动适应方法两类<sup>[16]</sup>. 主动检测方法通过监测模型的性能表现或数据变化判断概念漂移, 能够及时发现数据分布的变化, 迅速做出响应以保持模型的准确性, 但是可能会错误地将随机波动识别为概念漂移, 导致不必要的模型调整; 被动适应方法通过对模型的不不断调整来适应概念漂移, 即使在数据分布缓慢变化的情况下也能保持较好的性能, 但可能会忽略重要信息, 并且鲁棒性较差. 集成学习是常用的被动适应方法, 通过特定的结合策略, 将基于不同时序数据的多个基模型集成为一个泛化能力更强、性能更优的模型, 并通过灵活的指标调整以有效适应概念漂移<sup>[17-18]</sup>. 然而, 现有集成方法大多只能解决某一类型的概念漂移, 泛化能力不强, 对其他类型的概念漂移效果较差. 并且由于替换策略, 过拟合的基模型可能会替换掉泛化性能较好的基模型, 从而导致模型不稳定. 此外, 在概念漂移刚发生时, 集成模型中存在较多携带历史数据的基模型, 使模型难以迅速适应概念漂移.

为应对上述问题, 本文提出一种基于主动-被动增量集成的概念漂移适应方法 (Concept drift adaptation method based on active-passive incremental ensemble, CDAM-APIE). 该方法采用在线增量集成策略和漂移检测方法, 构建被动集成模型和主动基模型. 被动集成模型通过对数据块中的单一训练样本进行预测并动态调整基模型权重, 有利于对数据分布变化进行快速响应. 同时, 利用增量学习和概念漂移检测技术构建主动基模型, 提升模型在平稳数据流状态下的鲁棒性和漂移以后的泛化

性能. 本文的主要贡献如下:

- 1) 通过实时调整权重、周期性更新模型, 提高模型适应不同类型概念漂移的能力;
- 2) 主动方法与被动方法相结合, 提高模型对概念漂移的适应能力和泛化性能;
- 3) 使用增量学习方法, 缓解数据块大小对基模型性能的影响, 提高基模型的稳定性.

## 1 相关知识

针对数据流分类中的概念漂移问题, 目前已有很多研究成果, 常见的概念漂移方法可以分为两类, 即主动检测方法和被动适应方法. 在主动检测方法中, 通常使用概念漂移检测器监测分类器性能或窗口中数据分布的变化情况, 从而判断数据流是否处于稳定状态. 当检测指标超出设定阈值时, 则判断数据流不稳定, 模型做出相应调整以适应概念漂移. 例如 Gama 等<sup>[19]</sup>提出的概念漂移检测算法 (Drift detection method, DDM) 认为在稳定的数据流状态下, 随着训练数据增多, 学习模型的错误率会逐渐下降, 该方法通过检测当前总体错误率的增长情况判断概念漂移的发生. Hinder 等<sup>[20]</sup>提出动态适应窗口独立性漂移检测 (Dynamic adapting window independence drift detection, DAWIDD), DAWIDD 通过滑动窗口管理数据样本, 并结合独立性测试来识别数据分布的变化, 从而检测概念漂移. Wen 等<sup>[21]</sup>提出自适应树状神经网络 (Adaptive tree-like neural network, ATNN), ATNN 是一种多分支结构的自适应神经网络, 其通过概念漂移检测对网络进行调整, 确定何时添加新分支或激活旧分支, 以此适应概念漂移. 主动检测方法虽然在数据流平稳状态下效率较高, 但是在检测过程中可能会出现漏检、误检的情况, 导致模型错误更新.

与主动检测方法相比, 被动适应方法不需要进行概念漂移检测, 而是通过对模型的不不断调整来适应概念漂移<sup>[22]</sup>. 集成学习是常见的被动适应方法, 根据处理策略的不同分为两类: 一类是基于数据块的集成, 另一类是基于单样本的在线集成. 基于数据块的集成方式是每次对一整块数据进行批量学习和模型更新. 例如 Street 等<sup>[23]</sup>提出的一种基于数据块的数据流集成分类算法 (Streaming ensemble algorithm, SEA), SEA 使用最新数据块构造基模型, 并依据特定的启发性规则, 对集成模型中表现最差的基模型进行替换, 从而适应概念漂移. 该方法中新训练的基模型在整个集成中可能被旧的基模型压制, 导致无法及时适应概念漂移. 基于此, Wang 等<sup>[24]</sup>提出一种基于准确率加权集成算法 (Accuracy

weighted ensemble, AWE), AWE 对基模型赋予权重, 根据基模型在最新数据块上的误差率对其加权, 提升模型适应漂移的能力. Weinberg 等<sup>[25]</sup> 提出一种结合霍夫丁自适应树的集成方法 (Ensemble combined with Hoeffding adaptive tree, EnHAT), EnHAT 结合霍夫丁自适应树算法与基于数据块生成的决策树集成, 实现对概念漂移的快速适应. 通过周期性更新模型, 基于数据块的集成方式虽然能够有效应对渐变漂移, 但是不能及时应对突变漂移, 并且其性能受限于数据块的大小. 数据块较小虽能应对部分突变漂移, 但可能导致过拟合; 数据块较大虽有机会获得更好的基模型, 但可能出现一个数据块蕴涵多种概念的情况. 基于单样本的在线集成方式是每次只对一个数据进行学习和模型更新. 例如 Oza 等<sup>[26]</sup> 提出的 Oza Bagging, 该算法对最新的数据进行  $k$  次抽样, 并以此模拟数据流中的数据,  $k$  的取值服从参数为 1 的泊松分布. Kolter 等<sup>[27]</sup> 提出一种动态加权多数投票算法 (Dynamic weighted majority, DWM), DWM 依据基模型在最新样本上的分类结果动态调整其权值, 以提高应对突变漂移的能力. 郭虎升等<sup>[28]</sup> 提出一种基于在线集成的概念漂移自适应分类方法 (Adaptive classification method for concept drift based on online ensemble, AC\_OE), AC\_OE 将在线集成与增量学习结合, 提高模型整体泛化性能. 基于单样本的在线集成方式虽然能够及时应对突变漂移, 但是由于没有周期性更新, 对渐变漂移适应能力较差.

本文结合在线集成、周期性更新策略以及概念漂移检测方法, 提出 CDAM-APIE. 该方法既利用在线集成方式更新权重, 使模型能够有效应对突变漂移, 又利用周期性更新策略提升模型对渐变漂移的适应能力. 此外, CDAM-APIE 使用结合概念漂移检测方法的增量基模型, 避免在平稳数据流状态下持续更新基分类器造成的性能下降, 提升模型在平稳数据流状态下的鲁棒性和漂移以后的泛化性能.

## 2 本文方法

本文提出一种主动-被动增量集成的概念漂移适应方法, 整体框架如图 1 所示. 该方法首先通过连续捕获实例将数据流转换为一系列数据块, 然后分别采用被动集成模型和主动基模型进行训练更新, 被动集成模型通过在线集成策略进行构建, 主动基模型利用概念漂移检测方法进行增量训练, 最后将模型的两部分加权结合得到当前时刻的最终模型输出预测结果.

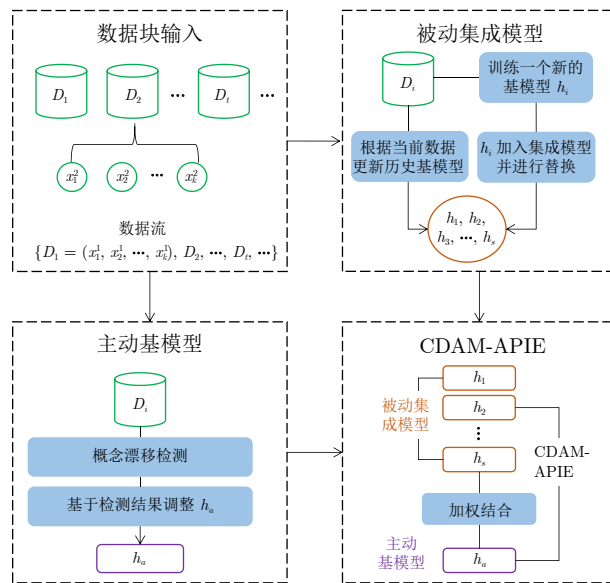


图 1 CDAM-APIE 整体框架

Fig. 1 The overall framework of CDAM-APIE

### 2.1 问题定义

数据流是指随时间不断变化、样本一个接一个到来, 具有实时性、连续性、不稳定性数据序列<sup>[29]</sup>, 可以表示为  $S = \{s_1, s_2, \dots, s_t, \dots\}$ , 其中,  $s_t = (x_t, y_t)$  表示  $t$  时刻的样本,  $x_t$  表示  $t$  时刻样本的特征向量,  $y_t$  表示  $t$  时刻样本的标签. 概念漂移是指底层数据分布发生变化, 假设数据流中的底层数据分布用联合概率分布, 表示为  $P(x, y)$ , 若在时刻  $t$ , 数据流发生概念漂移, 可以表示为

$$\exists x: P_{t-1}(x, y) \neq P_t(x, y) \quad (1)$$

目前, 研究人员根据变化率将概念漂移分为四种类型<sup>[29]</sup>: 突变漂移、渐变漂移、重复漂移和增量漂移. 四种类型的概念漂移如图 2 所示.

### 2.2 被动集成模型

数据流具有时序特性, 这与集成学习根据不同训练集构建基模型的机制高度契合, 且集成学习能够有效克服单一分类器在数据流挖掘中学习过多复杂数据分布的问题. 因此, 在不同时间节点构建基模型并将其加入集成, 是数据流挖掘的一种有效方法. 在线学习方法能够实时更新模型参数, 具备良好的适应性, 通过不断对单一样本进行处理, 可以迅速适应动态变化的数据分布, 因此, 在线集成策略可以实现对概念漂移的快速响应. 此外, 增量学习可以解决由固定数据单元训练的基模型泛化性能差的问题. 因此, 本文使用在线集成策略先对单一样本进行预测, 然后根据预测结果更新基模型权重并进行增量学习, 以快速适应数据流变化.

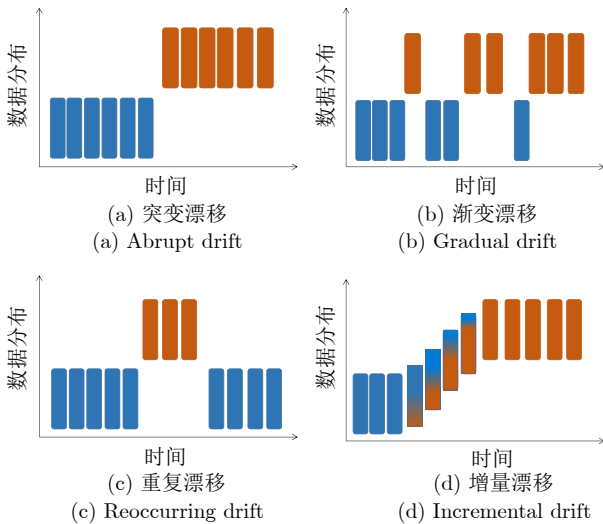


图 2 四种类型的概念漂移

Fig.2 Four types of concept drift

事实上, 在线集成策略中的权值更新能及时应对突变漂移, 但是对渐变漂移不太敏感, 因此, 本文采用实时更新权重和周期性更新模型的方式分别应对突变和渐变类型的概念漂移. 一方面, 通过实时更新基模型权重快速适应突变漂移; 另一方面, 在固定的学习单元后替换性能较差的基模型, 使集成模型对渐变漂移有较好的效果, 同时也能在一定程度上替换因增量学习而获取的驳杂概念的基模型, 保持模型的稳定性. 具体地, 假设被动集成模型为  $H_P = \{(h_1, w_1), (h_2, w_2), \dots, (h_s, w_s)\}$ , 其中,  $s$  表示集成模型中最大分类器数量,  $h_i$  表示第  $i$  个基模型,  $w_i$  表示其对应的权重, 初始情况下, 每个基模型对应权重  $w_i = 1/s$ . 假设数据流为  $SD = \{D_1, D_2, \dots, D_t, \dots\}$ , 固定数据单元为数据块  $D_t$ , 当时刻  $t$  的新样本  $x_j^t \in D_t$  输入后, 使用当前的被动集成模型中所有的基模型对其预测:

$$\tilde{y}_j^t = h_i(x_j^t) \quad (2)$$

若  $\tilde{y}_j^t \neq y_j^t$ , 则将该分类器的权重根据式 (3) 作更新, 反之基模型权重保持不变.

$$w_i = \beta \cdot w_i, \quad \beta \in (0, 1) \quad (3)$$

其中,  $\beta$  表示权重衰退率, 若某个基模型预测错误, 那么其权重按照一定比率下降.

在数据流中一旦发生概念漂移, 短时间内被动集成中大多数历史基模型的性能会显著下降, 导致其对应的权重急剧降低. 因此, 在所有基模型的权重更新完成后, 按照式 (4) 对权重进行标准化.

$$w_i = \frac{w_i}{\sum_{j=1}^s w_j} \quad (4)$$

当数据块  $D_t$  中的全部样本处理完毕后, 进行基模型的替换更新. 具体来说, 在固定数据单元  $k = 100$  的数据块  $D_t$  上训练构建新的基模型  $h_t$ , 同时依照式 (5) 的选择标准, 查找在数据块  $D_t$  上表现最差的基模型  $h_{\text{bad}}$ , 并用  $h_t$  进行替换.

$$h_{\text{bad}} = \arg \max_{h_i \in H_P} \frac{\sum_{j=1}^k h_i(x_j^t) \neq y_j^t}{k} \quad (5)$$

由于  $h_t$  基于最新数据块  $D_t$  训练构建, 代表目前数据流中的最新数据分布, 所以我们赋予其最高的权重, 计算方法为

$$w_t = \arg \max_{w_i \in H_P} (w_i) \quad (6)$$

被动集成模型的构建过程如图 3 所示.

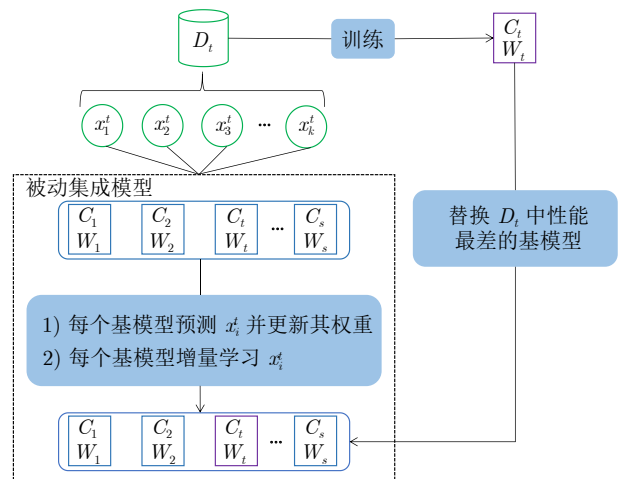


图 3 被动集成模型的过程

Fig.3 Process of passive ensemble model

### 2.3 主动基模型

在被动集成模型中, 增量学习虽然可以解决由固定数据单元训练基模型引起的泛化性能差的问题, 但也导致某些基模型在概念漂移期间学习到混合分布, 影响集成性能. 此外, 当数据流处于稳定状态时, 频繁替换基模型会导致部分代表性特征缺失, 使得模型的鲁棒性和泛化性能下降.

针对以上问题, 本文构建带有概念漂移检测方法的主动基模型. 当检测器发出警告信号时, 表示概念可能发生变化, 但还未达到漂移水平, 创建备用分类器并从当前数据位点开始学习, 同时让之前的历史模型也开始增量学习, 保证漂移发生时当前

模型可以更快适应概念漂移. 当检测器发出漂移信号时, 表示此时已经达到漂移水平, 备用模型替代历史模型. 本文借鉴 DDM<sup>[19]</sup> 的思想, 假设数据是符合独立同分布的, 随着数据的输入和学习, 模型错误率不断下降. 警告阈值和漂移阈值分别如式 (7) 和式 (8) 所示:

$$p_i + s_i \geq \min(p) + 2 \cdot \min(s) \quad (7)$$

$$p_i + s_i \geq \min(p) + 3 \cdot \min(s) \quad (8)$$

其中,  $p_i$  为第  $i$  个数据进入后的整体错误率,  $s_i$  为第  $i$  个数据进入后的整体标准差.  $\min(p)$  和  $\min(s)$  动态更新, 在  $t$  时刻时, 若  $p_t + s_t$  的值比当前的  $\min(p) + \min(s)$  小, 则使用  $p_t$  和  $s_t$  替换  $\min(p)$  和  $\min(s)$ ; 检测到概念漂移发生后,  $\min(p)$  和  $\min(s)$  重新取值. 主动基模型具体构建过程如图 4 所示.

## 2.4 CDAM-APIE

当新样本进入时, 将上述两个模型集成为当前时刻的测试模型. 对新样本进行加权投票得到预测结果:

$$H_t(x_j^t) = W_P \cdot H_P(x_j^t) + W_A \cdot H_A(x_j^t) \quad (9)$$

设置两个模型的权重为  $W_P$  和  $W_A$ , 初始权重设置为 0.5. 具体地, 当新样本进入时, 同时使用两种模型进行预测, 当使用被动集成模型  $H_P$  进行预测时:

$$H_P(x_j^t) = \sum_{i=1}^s w_i \cdot h_i(x_j^t) \quad (10)$$

若  $H_P(x_j^t) \neq y_j^t$ , 将当前的被动集成模型权重按式 (3) 进行更新, 否则不更新. 当使用主动基模型  $H_A$  预测时, 若预测错误, 也按式 (3) 更新权重, 否则不更新. CDAM-APIE 算法如算法 1 所示.

### 算法 1. CDAM-APIE

输入. 数据流  $SD = \{D_1, D_2, \dots, D_t, \dots\}$ , 主动基

模型  $H_A$ , 主动基模型权重  $W_A$ , 权重衰减系数  $\beta$ , 被动集成模型  $H_P = \{(h_1, w_1), (h_2, w_2), \dots, (h_s, w_s)\}$ , 被动集成模型权重  $W_P$

输出. 当前时刻的测试模型  $H_t = \{H_P, H_A\}$

- 1) **While** (数据流  $SD$  中  $t$  时刻对应的数据块  $D_t$  进入) **do**
- 2)   **While** ( $D_t$  中第  $j$  个样本  $x_j^t$  进入) **do**
- 3)      $H_t(x_j^t) = W_P \cdot H_P(x_j^t) + W_A \cdot H_A(x_j^t)$  // 使用当前时刻的集成模型预测
- 4)     **If** ( $H_A(x_j^t) \neq y_j^t$ ) **then**
- 5)        $W_A = \beta \cdot W_A$ ,  $\beta \in (0, 1)$  // 更新权重
- 6)     **End if**
- 7)     **If** ( $H_P(x_j^t) \neq y_j^t$ ) **then**
- 8)        $W_P = \beta \cdot W_P$ ,  $\beta \in (0, 1)$  // 更新权重
- 9)     **End if**
- 10)    **For** ( $h_i \subseteq H_P$ ) **do**
- 11)     **If** ( $H_i(x_j^t) \neq y_j^t$ ) **then**
- 12)        $w_i = \beta \cdot w_i$ ,  $\beta \in (0, 1)$  // 更新权重
- 13)     **End if**
- 14)      $h_i \leftarrow x_j^t$  // 基于  $x_j^t$  对  $h_i$  增量学习
- 15)     **End for**
- 16)      $w_i = w_i / \sum_{i=1}^s w_i$  // 权重标准化
- 17)      $H_A \leftarrow x_j^t$  // 基于  $x_j^t$  对  $H_A$  增量学习
- 18)    **If** (发出概念漂移警告信号且没有备用模型  $H_B$ ) **then**
- 19)     创建备用模型  $H_B$
- 20)      $H_B \leftarrow x_j^t$  // 基于  $x_j^t$  对  $H_B$  增量学习
- 21)     **End if**
- 22)    **If** (发出概念漂移信号) **then**
- 23)     置  $H_B$  为  $H_A$
- 24)     **End if**
- 25)    **End while**
- 26)    基于  $D_t$  训练最新的基模型  $h_t$
- 27)    **If** ( $H_P$  中的基模型数量  $< s$ ) **then**
- 28)     将最新的基模型  $h_t$  加入  $H_P$

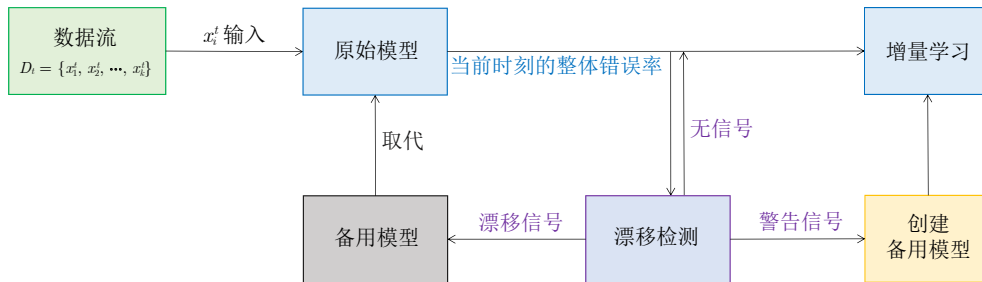


图 4 主动基模型的过程

Fig. 4 Process of active base model

29) **Else** 使用  $h_t$  替换  $h_{\text{bad}}$ , 权重设置

$$w_t = \arg \max_{w_i \in H_P} (w_i)$$

30) **End if**

31) **End while**

## 2.5 复杂度分析

CDAM-APIE 的计算成本主要集中在基模型训练更新、概念漂移检测和权重更新上。

假设在大小为  $m$  的数据集上训练 1 个基模型的时间复杂度为  $O(f(m))$ , 在被动集成模型中, 训练并更新  $s$  个基模型的时间复杂度为  $O(sf(m))$ . 对数据集进行分类以调整其权重的时间复杂度与数据集大小呈线性关系, 即  $O(ms)$ . 将数据集划分为  $m/k$  个大小为  $k$  的固定数据单元时, 替换基模型的时间复杂度为  $O(m/k)$ , 这在计算总的复杂度时可以忽略不计。

在主动基模型中, 概念漂移检测的时间复杂度为  $O(m)$ , 训练并更新模型的时间复杂度为  $O(f(m))$ , 由于替换基模型需要经过概念漂移检测, 其发生的频率远小于  $O(m)$ , 因此可以忽略不计. 对两个模块进行加权集成的时间复杂度为  $O(m)$ .

综上所述, CDM-APIE 的时间复杂度为  $O(sf(m) + ms + m + f(m) + m) = O(sf(m) + m)$ .

## 3 实验结果及分析

本文在包含不同类型的概念漂移数据集上进行实验, 实验方法采用 python3.9 编写和运行. 对比方法选取经典的基于数据块分类算法、基于单样本的在线集成算法以及深度学习算法, 包括准确率加权集成算法 (AWE)<sup>[24]</sup>、Oza Bagging 算法<sup>[26]</sup>、动态加权多数投票算法 (DWM)<sup>[27]</sup>、在线欠装袋算法

(Online under over bagging, OOB)<sup>[30]</sup>、基于在线集成的概念漂移自适应分类方法 (AC\_OE)<sup>[28]</sup> 以及自适应树状神经网络 (ATNN)<sup>[21]</sup>.

### 3.1 实验数据集

为评估算法适应各种类型概念漂移的能力, 本文共使用 12 个数据集, 其中 10 个数据集来源于文献 [28], 包含 6 个带有渐变、突变和增量类型概念漂移的合成数据集<sup>[28]</sup> 和 4 个真实数据集<sup>[28]</sup>, 另外, 本文还使用 python3.9 生成 2 个具有重复漂移的合成数据集, 生成方式如下。

1) Sea-re<sup>[23]</sup>: 每个样本的基本结构为  $\{f_1, f_2, f_3, C\}$ , 其中  $\{f_1, f_2, f_3\}$  为特征, 类别  $C$  仅与  $\{f_1, f_2\}$  两个特征相关, 当满足分类函数  $f_1 + f_2 < \theta$  时, 类别  $C$  为正类, 反之为负类. 在生成过程中, 我们首先置  $\theta = 8$ , 之后变化为  $\theta = 9$ , 周期变化 2 次生成重复漂移。

2) Sine<sup>[19]</sup>: 每个样本的基本结构为  $\{f_1, f_2, f_3, f_4, C\}$ , 其中  $\{f_1, f_2, f_3, f_4\}$  为特征, 每个特征的值均匀分布在  $[0, 1]$  中, 类别  $C$  仅与  $\{f_1, f_2\}$  两个特征相关. 当满足分类函数  $f_1 < \sin(f_2)$  时, 类别  $C$  为正类, 反之为负类. 在生成过程中, 我们首先选择分类函数  $f_1 < \sin(f_2)$ , 之后将分类函数反转为  $f_1 > \sin(f_2)$ , 周期变化 2 次生成重复漂移。

数据集的具体信息如表 1 所示。

### 3.2 实验参数设置

1) 固定数据单元  $k$ . 选择一个合适的合适的数据大小替换基模型是非常重要的. 数据单元过小可能会导致基模型过拟合, 对集成性能造成影响; 数据单元过大可能会导致适应渐变漂移较慢, 对适应速度有

表 1 实验所用数据集  
Table 1 Datasets used in experiment

数据集	特征个数	类别个数	样本个数 (k)	漂移类型	漂移次数	漂移位点 (k)
Hyperplane	10	2	100	增量	-	-
Sea	3	2	100	渐变	3	25, 50, 75
Sea-re	3	2	100	重复	3	25, 50, 75
LED-gradual	24	10	100	渐变	3	25, 50, 75
LED-abrupt	24	10	100	突变	1	50
RBFblips	20	4	100	突变	3	25, 50, 75
Tree	30	10	100	突变	3	25, 50, 75
Sine	4	2	100	重复	3	25, 50, 75
KDDcup99	41	23	494	-	-	-
Electricity	6	2	45	-	-	-
Coverttype	54	7	581	-	-	-
Weather	9	3	95	-	-	-

所影响. 本文实验采取数据单元  $k = 100$ .

2) 权重衰退率  $\beta$ . 权重衰退率  $\beta$  是影响算法集成性能的关键. 衰退率大时, 虽然在漂移发生后能够更快适应新的数据分布, 但是模型性能极其不稳定; 反之, 虽然模型相对稳定, 但是在漂移发生后无法快速适应新的数据分布. 本文实验采用的衰退率  $\beta = 0.95$ .

3) 基模型. AC\_OE 方法沿用文献 [28] 中的参数和 LIBSVM, 其余所有集成方法均使用 Hoeffding 树, 基模型数量  $s$  设置为 10.

### 3.3 评价指标

为验证所提 CDAM-APIE 的合理性, 本文从精度、恢复速率和鲁棒性等方面对算法进行分析, 具体指标如下.

1) 平均实时精度 (Average real-time accuracy)<sup>[28]</sup> 表示模型在每个时间步数下实时精度的均值, 定义如下:

$$A_{\text{acc}} = \frac{1}{T} \sum_{t=1}^T \text{racc}_t \quad (11)$$

其中,  $A_{\text{acc}}$  表示平均实时精度,  $T$  表示总的时间步数,  $\text{racc}_t$  表示模型在时间步数  $t$  下的实时精度, 计算式为

$$\text{racc}_t = \frac{n_t}{n} \quad (12)$$

其中,  $n_t$  表示在时间步数  $t$  下分类正确的样本数量,  $n$  表示在一个时间步数内处理的样本总数.  $A_{\text{acc}}$  值越大表明模型的实时性能越好.

2) 累积精度 (Cumulative accuracy)<sup>[28]</sup> 反映模型从开始时刻到当前时刻的性能, 定义如下:

$$C_{\text{acc}} = \frac{1}{T_t \cdot n} \sum_{t=1}^T n_t \quad (13)$$

其中,  $C_{\text{acc}}$  表示累积精度,  $T_t$  表示当前的累积步数.  $C_{\text{acc}}$  值越大表明模型的整体性能越好.

3) 恢复速率 (Recovery speed under accuracy)<sup>[28]</sup> 是评价模型在数据流发生概念漂移后实时精度稳定到新概念所需的步数, 定义如下:

$$RSA = \text{Step} \cdot (1 - \text{racc}_t) \quad (14)$$

其中,  $RSA$  表示恢复速率,  $\text{Step}$  表示模型从漂移点到收敛点所需要的时间步数,  $\text{racc}_t$  表示收敛点后一个时间节点的实时精度. 本文根据数据集漂移位点后的实时精度变化确定当前时间节点是否为收敛点, 若当前时间节点以及后一个时间节点的精度变化均小于阈值, 则判断其为收敛点, 定义如式 (15)

所示:

$$\begin{aligned} |\text{racc}_t - \text{racc}_{t-1}| &\leq \text{Thr} \text{ 且} \\ |\text{racc}_{t+1} - \text{racc}_t| &\leq \text{Thr} \end{aligned} \quad (15)$$

其中,  $\text{Thr}$  表示阈值. 由于模型在不同数据集上的波动不同, 将阈值设置为各数据集在收敛期间实时精度变化的均值加上其三倍标准差, 如式 (16) 所示:

$$\text{Thr} = \frac{1}{T} \sum_{t=1}^T (\text{racc}_{t+1} - \text{racc}_t) + 3\sigma \quad (16)$$

其中,  $\sigma$  表示标准差.  $RSA$  值越小表明模型的恢复速率越快.

4) 鲁棒性 (Robustness)<sup>[31]</sup> 是评价模型稳定性的重要指标, 能够在一定程度上评价模型的泛化性能, 通过平均实时精度分析不同算法的鲁棒性. 算法  $A$  在数据集  $D$  上的鲁棒性定义为

$$ROB_A(D) = \frac{A_{\text{acc-A}}(D)}{\min A_{\text{acc-all}}(D)} \quad (17)$$

其中,  $A_{\text{acc-A}}(D)$  表示算法  $A$  在数据集  $D$  上的平均实时精度,  $\min A_{\text{acc-all}}(D)$  表示在数据集  $D$  上所有算法中的最小平均实时精度. 算法  $A$  的整体鲁棒性为算法  $A$  在所有数据集上的鲁棒性之和, 假设有  $N$  个数据集, 具体定义如下:

$$ROB_A = \sum_{i=1}^N ROB_A(D_i) \quad (18)$$

$ROB_A$  数值越大表明算法  $A$  的整体鲁棒性越好.

### 3.4 实验结果及分析

为有效评估 CDAM-APIE 的分类效果、概念漂移发生后适应漂移的能力以及模型的稳定性, 从模型精度、恢复速率、鲁棒性、消融效果以及参数敏感性等方面进行实验分析.

#### 3.4.1 模型精度评估

不同方法在每个数据集上的平均实时精度和平均排名如表 2 所示. 从表中可以看到, CDAM-APIE 的精度最高, 平均排名第 1, ATNN、Oza Bagging 和 DWM 紧随其后, OOB、AWE 和 AC\_OE 比较接近. CDAM-APIE 充分发挥被动集成模型和主动基模型的作用, 在平稳的数据流环境下保持较高精度, 在数据流发生概念漂移后能够有效学习最新数据分布, 提高模型的整体性能. CDAM-APIE 在大多数数据集上的表现都优于其他方法, 与平均排名次优的方法相比, 平均实时精度在 12 个数据集上平均高出 1.3%. 在 RBFblips 数据集上, CDAM-APIE 与 ATNN 和 AC\_OE 方法相差较大, 因为该数据集由随机径向基函数生成, 在 LIBSVM 上适配

表 2 不同方法在各数据集上的平均实时精度  
Table 2 Average real-time accuracy of different methods on each dataset

数据集	AWE	Oza Bagging	DWM	OOB	AC_OE	ATNN	CDAM-APIE (本文)
Hyperplane	0.8882 (4)	0.8758 (5)	0.9029 (2)	0.8223 (6)	0.8966 (3)	0.8195 (7)	<b>0.9088 (1)</b>
Sea	0.8335 (3)	0.8159 (4)	0.8410 (2)	0.7754 (7)	0.8027 (5)	0.7871 (6)	<b>0.8432 (1)</b>
Sea-re	0.8564 (4)	0.8596 (2)	0.8581 (3)	0.8030 (7)	0.8055 (6)	0.8166 (5)	<b>0.8605 (1)</b>
LED-gradual	0.6055 (4)	0.5979 (5)	0.5022 (7)	0.6163 (3)	0.5054 (6)	0.6282 (2)	<b>0.6330 (1)</b>
LED-abrupt	0.5944 (5)	0.6075 (3)	0.4918 (7)	0.5948 (4)	0.5178 (6)	0.6147 (2)	<b>0.6240 (1)</b>
RBFblips	0.8208 (5)	0.8852 (3)	0.7861 (6)	0.7811 (7)	0.9316 (2)	<b>0.9855 (1)</b>	0.8309 (4)
Tree	0.3630 (7)	0.4982 (6)	0.6449 (3)	0.6938 (2)	0.5480 (5)	0.6300 (4)	<b>0.8072 (1)</b>
Sine	0.9331 (4)	0.7489 (7)	0.9363 (3)	0.8595 (6)	0.9155 (5)	<b>0.9381 (1)</b>	0.9374 (2)
KDDcup99	0.9796 (4)	0.9920 (2)	0.9793 (5)	0.9913 (3)	0.9446 (7)	0.9589 (6)	<b>0.9926 (1)</b>
Electricity	0.7678 (7)	0.7928 (5)	0.8153 (3)	0.8110 (4)	0.7919 (6)	<b>0.8912 (1)</b>	0.8300 (2)
Covertime	0.2288 (7)	0.8735 (2)	0.8135 (4)	0.8052 (5)	0.7813 (6)	<b>0.9362 (1)</b>	0.8400 (3)
Weather	0.8893 (7)	0.9952 (2)	0.9941 (3)	0.9862 (4)	0.9069 (6)	0.9616 (5)	<b>0.9956 (1)</b>
平均排名	5.1	3.8	4.0	4.8	5.3	3.4	<b>1.6</b>

注: 括号内数值表示各方法的排名名次, 加粗字体表示在同一数据集上各方法中的最优结果。

性较高. 由于 Covertime 数据集数据分布倾斜且波动较小, 在该数据集上, CDAM-APIE 略差于 Oza Bagging. 此外, ATNN 以神经网络强大的表示能力在 RBFblips、Electricity、Covertime、Sine 数据集上强于传统集成学习方法. 实验结果表明, CDAM-APIE 可以有效适应不同类型的概念漂移, 但是对于数据分布倾斜的数据集, 分类性能仍需提升, 这主要是由于被动方法比较容易替换泛化性能较好的基模型, 从而影响模型性能.

不同方法在各数据集上的累积精度如图 5 所示. 从图中可以看到, CDAM-APIE 在 Hyperplane、Sea、Weather 数据集上略优于其他方法, 在 LED-gradual、LED-abrupt、Tree 数据集上优势显著, 在 RBFblips、Covertime、Electricity 数据集上较 Oza Bagging、AC\_OE、ATNN 方法略差. 总体来说, 本文提出的 CDAM-APIE 在所有数据集上表现较好.

不同方法在各数据集上的实时精度如图 6 所示. 在大部分数据集上, CDAM-APIE 的实时精度比其他方法的精度更高且更稳定. 在 RBFblips、Electricity、Covertime、Sine 数据集上, ATNN 在大多数时间都高于 CDAM-APIE 和其他集成方法, 这得益于神经网络的强大表示能力和多分支结构的动态调整. 发生概念漂移后, 所有方法的实时精度均明显下降, 但 CDAM-APIE 在大多数时候比其他方法更稳定、精度下降幅度更小、适应新概念更快.

实验结果表明, CDAM-APIE 在数据流平稳状态下能够保持较高的精度, 在概念漂移发生之后能及时捕捉新数据分布, 精度下降幅度较小, 其采用的增量学习和实时模型权重更新能够快速适应概念

漂移, 并保证方法的稳定性.

在性能分析中, 本文使用非参数检验方法 Friedman-test<sup>[32]</sup> 和 Bonferroni-Dunn 测试<sup>[33]</sup> 对所提方法与对比方法进行统计分析以验证其差异.

Friedman-test 可以对上述方法的性能优劣进行统计检验. 针对特定的  $K$  (7) 种方法和  $N$  (12) 个数据集, 令  $r_i^j$  为第  $j$  个方法在第  $i$  个数据集上的秩, 则第  $j$  个方法的秩和平均为

$$R_j = \frac{1}{N} \sum_{i=1}^N r_i^j \quad (19)$$

零假设  $H_0$  假定所有方法的性能相同. 在此前提下, 当  $K$  和  $N$  足够大时, Friedman-test 统计值  $\tau_F$  服从第一自由度  $K-1$  和第二自由度  $(K-1)(N-1)$  的 F 分布为

$$\begin{cases} \tau_F = \frac{(N-1)\tau_{X^2}}{N(K-1) - \tau_{X^2}} \\ \tau_{X^2} = \frac{12N}{K(K+1)} \left[ \sum_{j=1}^K R_j^2 - \frac{K(K+1)^2}{4} \right] \end{cases} \quad (20)$$

若计算得到的统计值超过特定显著水平 ( $\alpha$ ) 下 F 分布的临界值, 则拒绝零假设  $H_0$ , 表示各方法的秩和存在明显差异, 即不同方法性能具有显著差异. 反之, 接受零假设  $H_0$ , 所有方法的性能无明显差异.

对上述不同方法的平均实时精度进行统计检验, 计算得到 Friedman-test 统计值  $\tau_F = 5.7746$ , 在显著水平  $\alpha = 0.05$  的情况下, F 分布临界值为  $\tau_F^{0.05} = 2.239$ , 因此拒绝零假设  $H_0$ , 表明所有方法性能存

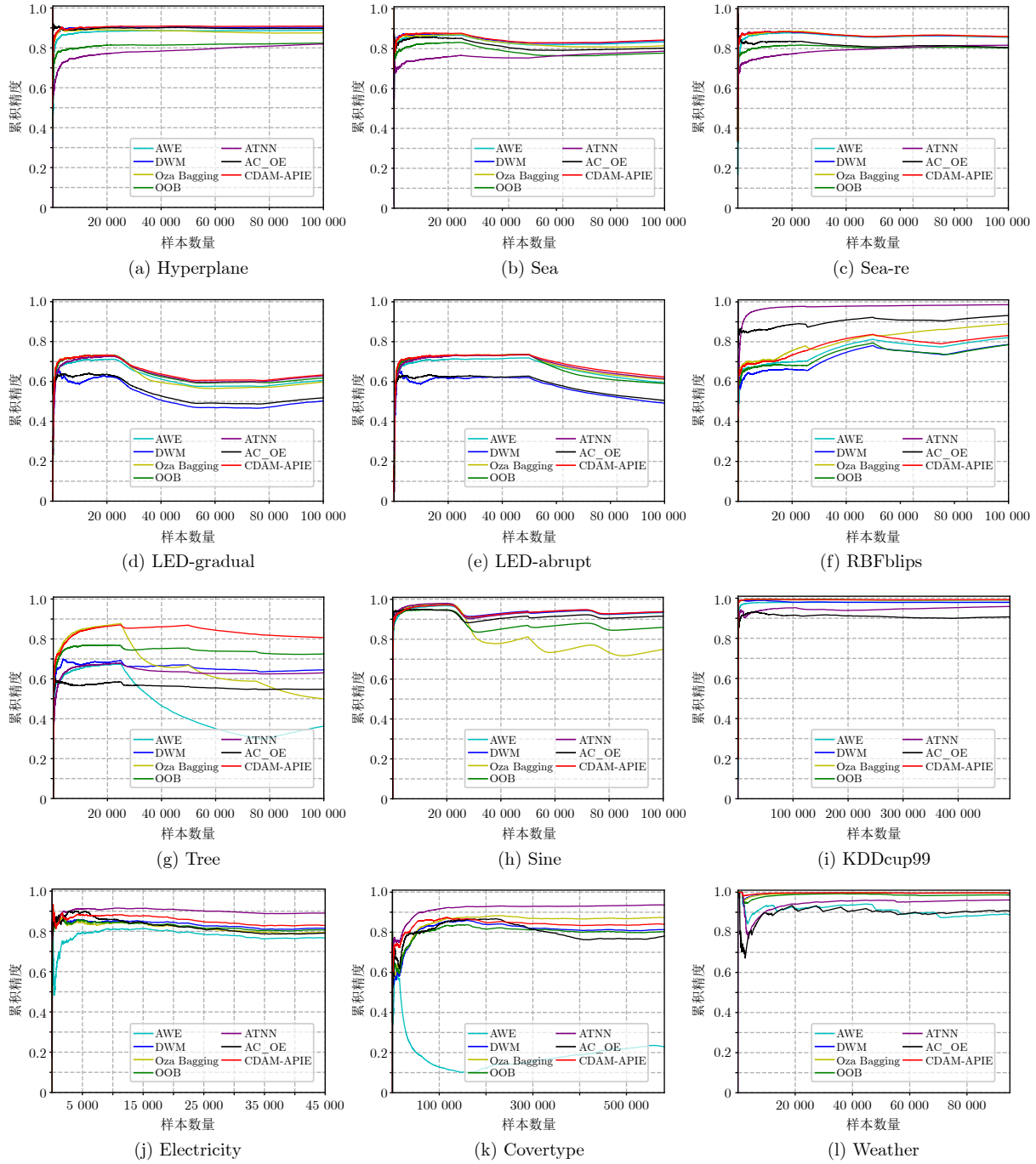


图 5 不同方法的累积精度  
Fig. 5 Cumulative accuracy of different methods

在显著差异.

Bonferroni-Dunn 测试用于比较多个方法之间的显著性差异. 如果两种方法的秩和平均差值超过临界差值 (Critical difference, CD), 就认为它们的性能存在显著差异:

$$CD = q_{\alpha} \sqrt{\frac{K(K+1)}{6N}} \quad (21)$$

其中,  $q_{\alpha}$  是在显著性水平  $\alpha$  下的学生化极差分布临

界值.

通过计算, 在显著性水平  $\alpha = 0.05$  的情况下, 临界差值  $CD = 2.3265$ . 不同方法在平均实时精度上的统计分析结果如图 7 所示. 结果表明, CDAM-APIE 明显优于 ATNN、Oza Bagging、DWM、OOB、AC\_OE 和 AWE.

### 3.4.2 模型恢复速率评估

当数据流发生概念漂移之后, 模型能否迅速调

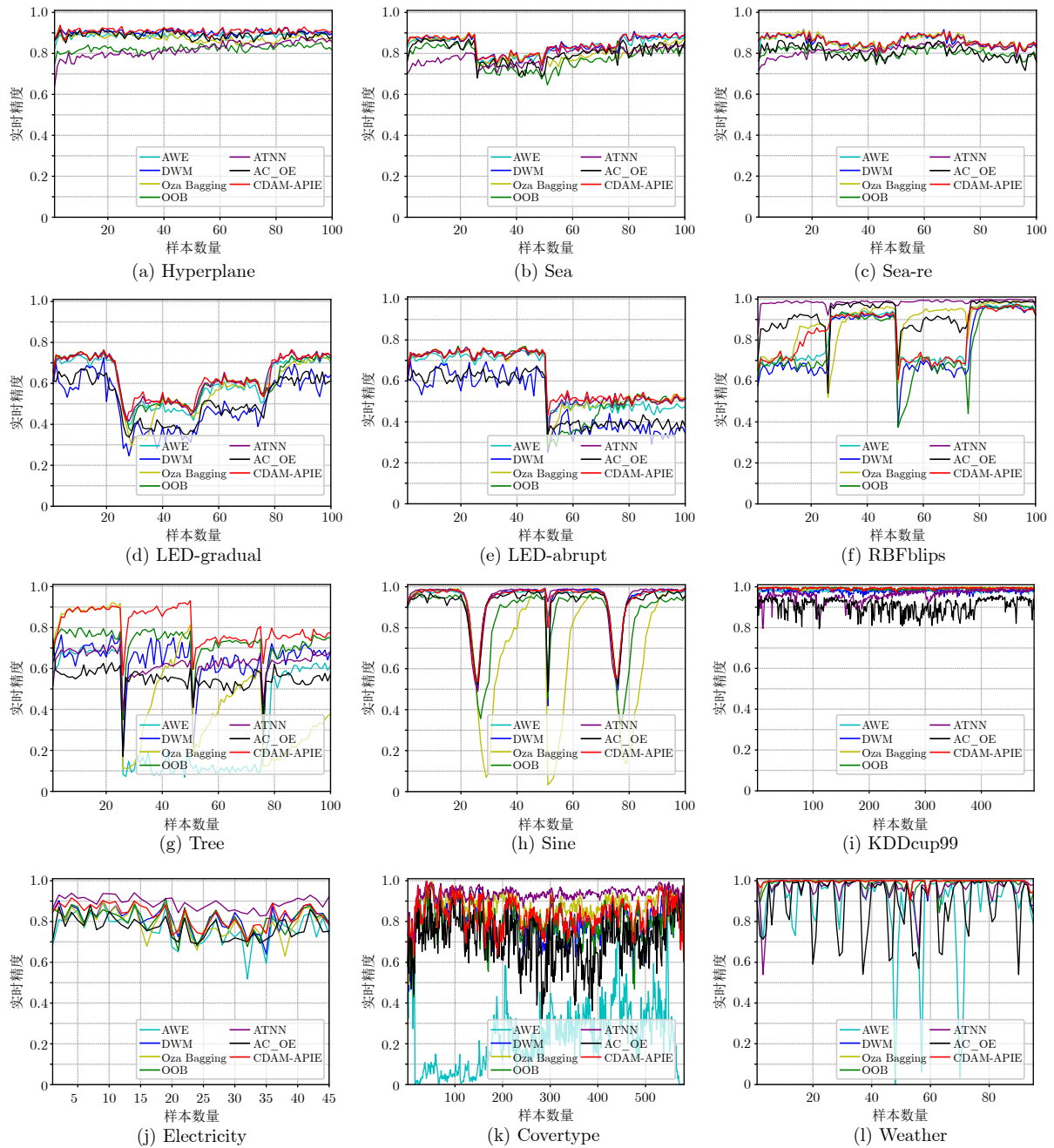


图 6 不同方法的实时精度

Fig.6 Real-time accuracy of different methods

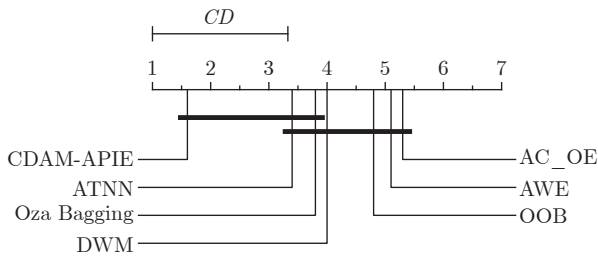


图 7 不同方法平均实时精度的 Bonferroni-Dunn 检验结果

Fig.7 Bonferroni-Dunn test for average real-time accuracy of different methods

整以适应新的概念是衡量算法的一个重要指标. 表 3 为不同方法在目前已知概念漂移位点的 5 个合成数据集上进行的模型恢复速率分析, 其中, LED-gradual, RBFblips, Sea, Tree 数据集有 25 k, 50 k, 75 k 三个漂移位点, LED-abrupt 数据集只有 50 k 一个漂移位点. 从表中可以看出, CDAM-APIE 总体表现最好, 在多数情况下恢复速率较好, 除了在 LED-gradual 数据集 (25 k, 75 k)、RBFblips 数据集 (25 k, 50 k, 75 k)、LED-abrupt 数据集 (50 k) 的恢复速

表 3 不同方法的恢复速率  
Table 3 Recovery speed of different methods

漂移位点	数据集	AWE	Oza Bagging	DWM	OOB	AC_OE	ATNN	CDAM-APIE (本文)
25 k	Sea	0.48	0.52	<b>0.46</b>	0.53	0.76	0.56	<b>0.46</b>
	LED-gradual	1.15	1.20	<b>0.72</b>	1.09	1.33	1.12	1.16
	RBFblips	0.29	0.38	0.28	0.37	0.17	<b>0.07</b>	0.15
	Tree	3.60	1.77	1.39	0.89	1.29	1.33	<b>0.58</b>
	平均排名	4.8	5.8	2.8	3.8	5.0	3.5	<b>2.3</b>
50 k	Sea	0.20	0.53	0.19	1.17	0.45	0.36	<b>0.17</b>
	LED-gradual	0.58	0.57	0.63	0.58	0.61	0.54	<b>0.53</b>
	LED-abrupt	1.63	1.36	1.35	1.36	<b>1.28</b>	1.66	1.47
	RBFblips	0.86	0.36	1.26	1.17	0.52	<b>0.04</b>	0.87
	Tree	<b>0.89</b>	1.55	1.60	1.16	1.44	1.56	<b>0.89</b>
平均排名	3.6	3.8	5.0	4.6	3.8	4.0	<b>2.6</b>	
75 k	Sea	0.52	0.44	0.33	0.61	0.47	<b>0.20</b>	0.33
	LED-gradual	1.09	<b>0.44</b>	0.51	1.06	1.23	0.99	0.47
	RBFblips	0.11	0.13	0.24	0.19	0.07	<b>0.02</b>	0.08
	Tree	2.20	1.74	1.04	2.62	1.44	1.18	<b>0.76</b>
	平均排名	5.5	3.8	3.5	6.3	4.5	2.3	<b>2.0</b>

率略低于其他方法之外, 其他位点上的恢复速率均优于其他方法. 由于 CDAM-APIE 通过实时处理最新样本更新基模型, 使模型能够快速适应新的数据分布, 但是在处理一些波动较大或较小的数据集时, 概念漂移检测技术可能会遇到误检或漏检的问题, 影响主动基模型的替换, 从而导致适应较慢.

### 3.4.3 模型鲁棒性评估

鲁棒性是评价算法稳定性的关键指标. 图 8 展示了 7 种方法在 12 个数据集上的鲁棒性, 每一列

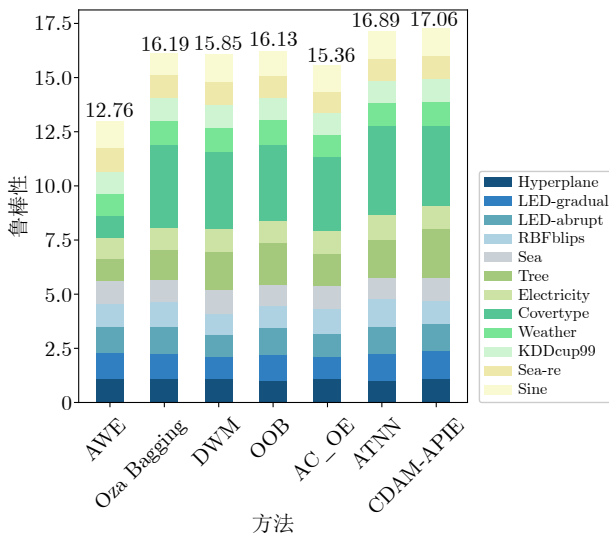


图 8 不同方法的鲁棒性比较  
Fig.8 Comparison of robustness of different methods

表示一种方法的整体鲁棒性, 其中不同颜色的堆叠块代表该方法在不同数据集上的鲁棒性. 由图 8 可以看到, CDAM-APIE 在大多数数据集上的鲁棒性优于其他 6 种方法, 且整体鲁棒性最高, 达到 17.06.

为进一步说明方法的稳定性和适用性, 图 9 显示了不同方法的平均排名与标准差. 从图中可以看到, CDAM-APIE 不仅平均排名第一, 而且标准差最小、稳定性较高. 实验结果表明, CDAM-APIE 有较好的鲁棒性.

### 3.4.4 消融效果评估

本文在被动集成模型的基础上加入带有概念漂移检测的主动基模型. 为证明其有效性, 分别采用基模型、被动集成模型、主动基模型和 CDAM-

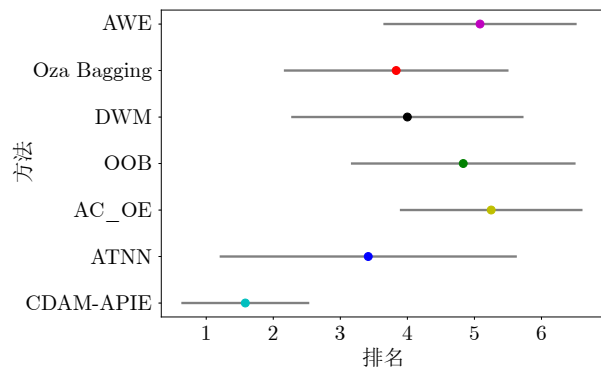


图 9 不同方法的平均排名 (平均值 ± 标准差)  
Fig.9 Average ranking of different methods (mean ± standard deviation)

表 4 消融效果分析  
Table 4 Analysis of ablation effect

数据集	基模型	被动集成模型	主动基模型	CDAM-APIE
Hyperplane	0.8603	0.8904	0.8716	<b>0.9088</b>
Sea	0.8118	0.8315	0.8333	<b>0.8432</b>
Sea-re	0.8554	0.8505	0.8517	<b>0.8605</b>
LED-gradual	0.5859	0.6266	0.6217	<b>0.6330</b>
LED-abrupt	0.6057	0.6154	0.6197	<b>0.6240</b>
RBFblips	<b>0.8430</b>	0.7825	0.8097	0.8309
Tree	0.4902	0.7873	0.7917	<b>0.8072</b>
Sine	0.6562	0.8941	0.9372	<b>0.9374</b>
KDDcup99	0.9904	0.9786	0.9922	<b>0.9926</b>
Electricity	0.7828	0.8117	0.8092	<b>0.8300</b>
Covertime	0.8247	0.8178	0.8246	<b>0.8400</b>
Weather	0.9953	0.9771	0.9929	<b>0.9956</b>

APIE 进行分析. 表 4 展示了不同数据集下各方法的平均实时精度. 从表中可以看出 CDAM-APIE 在除 RBFblips 以外的 11 个数据集上性能表现最佳. 在 RBFblips 数据集上基模型效果最好可能是因为其漂移前存在关键性样本无法在漂移后被保留. 实验结果充分表明 CDAM-APIE 将被动方法与主动方法结合的有效性以及合理性.

### 3.4.5 参数敏感性评估

本节对 CDAM-APIE 在不同固定数据单元  $k$  和权重衰退率  $\beta$  下的平均实时精度进行分析. 表 5 展示了 CDAM-APIE 在不同参数下的平均实时精

度. 结果显示, 在不同的固定数据单元  $k$  下, CDAM-APIE 在大多数数据集上的性能差异并不显著. 这是由于 CDAM-APIE 采用增量学习方法, 有效减轻了不同固定数据单元对模型性能的影响.

通过对  $k$  的总体标准差分析发现, 随着  $k$  值的增加, 平均实时精度的总体标准差呈现出先升后降的趋势. 当  $k$  取值较小时, 模型可能在某些数据集上表现较好, 但未能很好地泛化到所有数据集上. 当  $k$  取值较大时, 模型在大多数数据集上表现最好, 但在某些数据集上表现欠佳. 当  $k = 100$  时, 尽管精度略有下降, 但总体标准差较小, 表明模型更加稳定可靠. 另外, 现有方法常采用的固定数据单元一般为  $k = 100$ , 为了模型稳定性和便于比较, 故本文选择固定数据单元  $k = 100$ .

从表 5 可以看出, 随着  $\beta$  的增加出现两种情况, 精度先升后降或持续上升. 这是由于  $\beta$  决定模型权重的变化速度. 较小的  $\beta$  使得权重变化过快, 导致暂时效果不佳的基模型难以在后续过程中发挥作用. 而  $\beta$  较高时, 权重变化速度减慢, 使得模型无法及时响应概念漂移. 在相同  $k$  下的大多数数据集中,  $\beta = 0.95$  时的平均实时精度最好, 因此, 本文选择权重衰退率  $\beta = 0.95$ .

## 4 结束语

现有方法通常只能对某种特定类型的概念漂移做出有效应对, 且无法及时响应数据流, 为此本文提出 CDAM-APIE 方法. 该方法首先结合基于块

表 5 CDAM-APIE 在不同参数下的平均实时精度  
Table 5 Average real-time accuracy of CDAM-APIE under different parameters

	固定数据单元 $k$											
	50				100				150			
	权重衰退率 $\beta$				权重衰退率 $\beta$				权重衰退率 $\beta$			
	0.80	0.85	0.90	0.95	0.80	0.85	0.90	0.95	0.80	0.85	0.90	0.95
Hyperplane	0.9005	0.9020	0.9046	0.9076	0.9050	0.9062	0.9076	0.9088	0.9087	0.9099	0.9106	<b>0.9113</b>
Sea	0.8388	0.8400	0.8413	0.8420	0.8417	0.8423	0.8430	<b>0.8432</b>	0.8431	<b>0.8432</b>	0.8431	0.8427
Sea-re	0.8573	0.8584	0.8591	0.8600	0.8596	0.8599	0.8603	0.8605	0.8604	0.8607	<b>0.8609</b>	0.8606
LED-gradual	0.6309	0.6313	0.6317	0.6320	0.6314	0.6319	0.6328	0.6330	0.6348	0.6353	0.6358	<b>0.6362</b>
LED-abrupt	0.6220	0.6223	0.6230	0.6234	0.6229	0.6233	0.6237	0.6240	0.6228	0.6235	0.6241	<b>0.6243</b>
RBFblips	0.8366	0.8369	0.8378	0.8368	0.8308	0.8314	0.8312	0.8309	0.8503	0.8503	0.8505	<b>0.8500</b>
Tree	0.8073	0.8079	0.8086	<b>0.8089</b>	0.8066	0.8068	0.8071	0.8072	0.7960	0.7961	0.7963	0.7970
Sine	0.9372	0.9372	0.9372	0.9372	0.9371	0.9373	0.9371	0.9374	0.9378	0.9379	<b>0.9381</b>	<b>0.9381</b>
KDDcup99	0.9922	0.9922	0.9922	0.9922	0.9922	0.9922	0.9924	0.9926	0.9931	0.9931	0.9931	<b>0.9932</b>
Electricity	<b>0.8548</b>	0.8515	0.8482	0.8416	0.8401	0.8385	0.8358	0.8300	0.8205	0.8205	0.8207	0.8204
Covertime	<b>0.8569</b>	0.8543	0.8510	0.8460	0.8455	0.8444	0.8426	0.8400	0.8451	0.8448	0.8441	0.8415
Weather	0.9946	0.9945	0.9947	0.9956	0.9956	<b>0.9957</b>	0.9956	0.9956	0.9951	0.9952	0.9952	0.9953
总体标准差	0.121 0				<b>0.120 8</b>				0.121 0			

和单样本集成方法的优势, 能够更好地适应不同类型的概念漂移; 其次将被动和主动方法进行动态加权结合, 提高模型的泛化性能和适应能力. 此外, 增量学习用于缓解数据块大小对基模型性能的影响, 提高模型的鲁棒性. 实验结果表明, CDAM-APIE 通过动态加权的方式调节两个模块, 充分利用两个模块在不同数据集上的优势, 使模型在平稳状态下保持较高性能并在数据流发生概念漂移后也能快速适应新的数据分布, 对多种类型的概念漂移都具有较好的效果. 然而, 数据流中通常还伴随着数据不平衡等问题, 我们计划未来对含有概念漂移的非平衡数据流继续进行更进一步的尝试.

### References

- Din S, Yang Q, Shao J, Mawuli C, Ullah A, Ali W. Synchronization-based semi-supervised data streams classification with label evolution and extreme verification delay. *Information Sciences*, 2024, **678**: Article No. 120933
- Liao G, Zhang P, Yin H, Deng X, Li Y, Zhou H, et al. A novel semi-supervised classification approach for evolving data streams. *Expert Systems With Applications*, 2023, **215**: Article No. 119273
- Zheng X, Li P, Wu X. Data stream classification based on extreme learning machine: A review. *Big Data Research*, 2022, **30**: Article No. 100356
- Agrahari S, Singh A. Concept drift detection in data stream mining: A literature review. *Journal of King Saud University-Computer and Information Sciences*, 2021, **34**(10): 9523–9540
- Krempel G, Zliobaite I, Brzezinski D, Hullermeier E, Last M, Lemaire V, et al. Open challenges for data stream mining research. *ACM SIGKDD Explorations Newsletter*, 2014, **16**(1): 1–10
- Lughofer E, Pratama M. Online active learning in data stream regression using uncertainty sampling based on evolving generalized fuzzy models. *IEEE Transactions on Fuzzy Systems*, 2018, **26**(1): 292–309
- Zhai Ting-Ting, Gao Yang, Zhu Jun-Wu. Survey of online learning algorithms for streaming data classification. *Journal of Software*, 2020, **31**(4): 912–931 (翟婷婷, 高阳, 朱俊武. 面向流数据分类的在线学习综述. 软件学报, 2020, **31**(4): 912–931)
- Li H, Zhao T. A dynamic similarity weighted evolving fuzzy system for concept drift of data streams. *Information Sciences*, 2024, **659**: Article No. 120062
- Du Hang-Yuan, Wang Wen-Jian, Bai Liang. A novel evolving data stream clustering method based on optimization model. *Scientia Sinica: Informationis*, 2017, **47**(11): 1464–1482 (杜航原, 王文剑, 白亮. 一种基于优化模型的演化数据流聚类方法. 中国科学: 信息科学, 2017, **47**(11): 1464–1482)
- Wang P, Jin N, Davies D, Woo W. Model-centric transfer learning framework for concept drift detection. *Knowledge-Based Systems*, 2023, **275**: Article No. 110705
- Guo Hu-Sheng, Zhang Ai-Juan, Wang Wen-Jian. Concept drift detection method based on online performance test. *Journal of Software*, 2020, **31**(4): 932–947 (郭虎升, 张爱娟, 王文剑. 基于在线性能测试的概念漂移检测方法. 软件学报, 2020, **31**(4): 932–947)
- Karimian M, Beigy H. Concept drift handling: A domain adaptation perspective. *Expert Systems With Applications*, 2023, **224**: Article No. 119946
- Wozniak M, Zyblewski P, Ksieniewicz P. Active weighted aging ensemble for drifted data stream classification. *Information Sciences*, 2023, **630**: 286–304
- Cherif A, Badhib A, Ammar H, Alshehri S, Kalkatawi M, Imine A. Credit card fraud detection in the era of disruptive technologies: A systematic review. *Journal of King Saud University-Computer and Information Sciences*, 2023, **35**(1): 145–174
- Halstead B, Koh Y, Riddle P, Pears P, Pechenizkiy M, Bifet A, et al. Analyzing and repairing concept drift adaptation in data stream classification. *Machine Learning*, 2022, **111**(10): 3489–3523
- Jiao B, Guo Y, Gong D, Chen Q. Dynamic ensemble selection for imbalanced data streams with concept drift. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(1): 1278–1291
- Liu N, Zhao J. Streaming data classification based on hierarchical concept drift and online ensemble. *IEEE Access*, 2023, **11**: 126040–126051
- Wilson J, Chaudhury S, Lall B. Homogeneous-heterogeneous hybrid ensemble for concept-drift adaptation. *Neurocomputing*, 2023, **557**: Article No. 126741
- Gama J, Medas P, Castillo G, Rodrigues P. Learning with drift detection. In: Proceedings of the 17th Brazilian Symposium on Artificial Intelligence. Maranhao, Brazil: Springer, 2004. 286–295
- Hinder F, Artelt A, Hammer B. Towards non-parametric drift detection via dynamic adapting window independence drift detection (DAWIDD). In: Proceedings of the 37th International Conference on Machine Learning. New York, USA: PMLR, 2020. 4249–4259
- Wen Y, Liu X, Yu H. Adaptive tree-like neural network: Overcoming catastrophic forgetting to classify streaming data with concept drifts. *Knowledge-Based Systems*, 2024, **293**: Article No. 111636
- Pratama M, Pedrycz W, Lughofer E. Evolving ensemble fuzzy classifier. *IEEE Transactions on Fuzzy Systems*, 2018, **26**(5): 2552–2567
- Street W, Kim Y. A streaming ensemble algorithm (SEA) for large-scale classification. In: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2001. 377–382
- Wang H, Fan W, Yu P, Han J. Mining concept-drifting and noisy data streams using ensemble classifiers. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2003. 226–235
- Weinberg A, Last M. EnHAT-Synergy of a tree-based ensemble with Hoeffding adaptive tree for dynamic data streams mining. *Information Fusion*, 2023, **89**: 397–404
- Oza N, Russell S. Experimental comparisons of online and batch versions of bagging and boosting. In: Proceedings of the 7 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2001. 359–364
- Kolter J, Maloof M. Dynamic weighted majority: An ensemble method for drifting concepts. *Journal of Machine Learning Research*, 2007, **8**(12): 2755–2790
- Guo Hu-Sheng, Cong Lu, Gao Shu-Hua, Wang Wen-Jian. Adaptive classification method for concept drift based on online ensemble. *Journal of Computer Research and Development*, 2023, **60**(7): 1592–1602 (郭虎升, 丛璐, 高淑花, 王文剑. 基于在线集成的概念漂移自适应分类方法. 计算机研究与发展, 2023, **60**(7): 1592–1602)
- Gama J, Zliobaite I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. *ACM Computing Surveys*, 2014, **46**(4): 1–37
- Wang B, Pineau J. Online bagging and boosting for imbalanced data streams. *IEEE Transactions on Knowledge and Data Engineering*, 2016, **28**(12): 3353–3366
- Zhao Peng, Zhou Zhi-Hua. Learning from distribution-changing

data streams via decision tree model reuse. *Scientia Sinica: Informationis*, 2021, **51**(1): 1–12  
(赵鹏, 周志华. 基于决策树模型重用的分布变化流数据学习. 中国科学: 信息科学, 2021, **51**(1): 1–12)

- 32 Pereira D, Afonso A, Medeiros F. Overview of Friedman's test and post-hoc analysis. *Communications in Statistics-Simulation and Computation*, 2015, **44**(10): 2636–2653
- 33 Demsar J. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 2006, **7**: 1–30



**祁晓博** 太原师范学院计算机科学与技术学院副教授. 主要研究方向为数据挖掘与机器学习.

E-mail: [xbqi@tynu.edu.cn](mailto:xbqi@tynu.edu.cn)

(**QI Xiao-Bo** Associate professor at the School of Computer Science and Technology, Taiyuan Normal University. Her research interest covers data mining and machine learning.)

Her research interest covers data mining and machine learning.)



**陈佳明** 太原师范学院计算机科学与技术学院硕士研究生. 主要研究方向为数据挖掘与机器学习.

E-mail: [20222551035@stu.tynu.edu.cn](mailto:20222551035@stu.tynu.edu.cn)

(**CHEN Jia-Ming** Master student at the School of Computer Science and Technology, Taiyuan Normal University. His research interest covers data mining and machine learning.)

His research interest covers data mining and machine learning.)



**史颖** 山西大学计算机与信息技术学院博士研究生. 主要研究方向为图像处理与机器学习.

E-mail: [sy@tynu.edu.cn](mailto:sy@tynu.edu.cn)

(**SHI Ying** Ph.D. candidate at the School of Computer and Information Technology, Shanxi University. Her research interest covers image processing and machine learning.)

Her research interest covers image processing and machine learning.)



**齐慧** 太原师范学院计算机科学与技术学院教授. 主要研究方向为数据挖掘与机器学习.

E-mail: [qihui@tynu.edu.cn](mailto:qihui@tynu.edu.cn)

(**QI Hui** Professor at the School of Computer Science and Technology, Taiyuan Normal University. Her research interest covers data mining and machine learning.)

Her research interest covers data mining and machine learning.)



**郭虎升** 山西大学计算机与信息技术学院教授. 主要研究方向为数据挖掘与计算智能.

E-mail: [guohusheng@sxu.edu.cn](mailto:guohusheng@sxu.edu.cn)

(**GUO Hu-Sheng** Professor at the School of Computer and Information Technology, Shanxi University. His research interest covers data mining and computational intelligence.)

His research interest covers data mining and computational intelligence.)



**王文剑** 山西大学计算智能与中文信息处理教育部重点实验室教授. 主要研究方向为数据挖掘与机器学习. 本文通信作者.

E-mail: [wjwang@sxu.edu.cn](mailto:wjwang@sxu.edu.cn)

(**WANG Wen-Jian** Professor at the Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University. Her research interest covers data mining and machine learning. Corresponding author of this paper.)

Her research interest covers data mining and machine learning. Corresponding author of this paper.)