

有限标签下的非平衡数据流分类方法

李艳红^{1,2}, 李志华^{1,2}, 郑建兴^{1,2}, 白鹤翔^{1,2}, 郭鑫^{1,2}

1. 山西大学计算机与信息技术学院, 山西 太原 030006;

2. 山西大学计算智能与中文信息处理教育部重点实验室, 山西 太原 030006

摘要

数据流分类是数据流挖掘的重要研究内容, 其核心任务是从实时到达的数据流中快速捕获概念漂移, 并及时调整分类模型。极限学习机具有训练速度快和泛化性能好的优点, 然而目前基于极限学习机的数据流分类方法很少可以同时处理数据流中常见的多类非平衡、概念漂移、标签成本昂贵的问题。为此, 提出了一种有限标签下的非平衡数据流分类方法。该方法定义了预测概率差值与信息熵相结合的样本预测确定性度量, 提出了不确定性标签请求策略; 定义了基于类不平衡比率和样本预测误差的样本重要性度量; 提出了基于概念漂移指数的分类器的更新与重构机制。在6个人工数据流和3个真实数据流上的对比实验表明, 本文提出方法的分类性能优于已有的6种数据流分类方法的分类性能。

关键词

数据流分类; 多类非平衡; 极限学习机; 概念漂移; 标签成本昂贵

中图分类号: TP181

文献标志码: A

doi:10.11959/j.issn.2096-0271.2025018

Imbalanced data stream classification method with limited labels

LI Yanhong^{1,2}, LI Zhihua^{1,2}, ZHENG Jianxing^{1,2}, BAI Hexiang^{1,2}, GUO Xin^{1,2}

1. School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China

2. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China

Abstract

Data stream classification is a crucial research area within data stream mining, with the core task of swiftly capturing concept drifts from real-time incoming data stream and promptly adjusting classification models. Extreme learning machine possesses advantages such as fast training speeds and excellent generalization performance. However, existing data stream classification methods based on extreme learning machine often struggle to simultaneously address common challenges in data stream, such as multi-class imbalance, concept drift, and the expensive labeling cost. For this reason, an imbalanced data stream classification with limited labels was proposed. We defined a sample prediction certainty measure that combined the difference in predicted probabilities and information entropy. An uncertainty label request strategy was introduced. Furthermore, we defined a sample importance measure based on class imbalance ratios and sample prediction errors. We also proposed an update and reconstruction mechanism for the classifier based on the

concept drift index. Comparative experiments on six synthetic data streams and three real data streams demonstrate that the proposed method outperforms six existing data stream classification methods in terms of classification performance.

Key words

data stream classification, multi-class imbalance, extreme learning machine, concept drift, expensive labeling cost

0 引言

随着计算机技术和通信技术的进步,信息技术逐渐渗透到人们生活的各个方面。来自社交网络、金融服务、交通运输、传感网络等领域的数据通常是以数据流的形式产生的。数据流具有实时、高速连续、快速变化、无限等特点^[1],数据流分类已成为数据挖掘领域的一项重要研究任务。

概念漂移是指在数据流分类任务中,随着时间的推移,数据流分布发生变化,导致分类模型出现性能下降的现象^[2]。根据数据分布随时间推移变化的速度和形式,概念漂移可分为4种类型:突变型、重复型、增量型和逐渐型^[3]。突变型概念漂移是指某一时刻旧数据分布突然变化为新数据分布;重复型概念漂移是指发生突然型概念漂移后,新数据分布维持一段时间又突然变为旧数据分布;增量型概念漂移是指旧数据分布在一段时间内缓慢变化为新数据分布,缓慢变化过程的数据分布为新旧数据分布的混合;逐渐型概念漂移是指在一段时间内新数据分布逐渐取代旧数据分布,在变化过程中新旧数据分布交替出现。

通常,处理数据流概念漂移的方法主要分为主动检测^[4]和被动适应^[5]方法两大类。主动检测方法通过监测数据分布或模型性能的变化来检测概念漂移,当检测到漂移时,更新分类模型^[6]来适应新的数据

流环境;被动适应方法不主动检测是否发生概念漂移^[7],而是定期更新分类器来适应新的数据分布。

被动适应方法在更新模型时速度恒定、缺乏针对性,导致时间开销较大^[8],因此本文采用主动检测方法处理概念漂移。现有的概念漂移检测方法通常假定可以不受限制地访问数据流中所有样本的标签(即有监督的学习方法),这在实际应用中是不现实的^[9]。本文采用主动学习方法,通过请求少量最有价值样本的标签检测概念漂移,并更新和重构分类模型,进而实现有限标签下的数据流分类。

非平衡数据流是指数据流中不同类别的样本数量不均衡的情况。在对非平衡数据流分类时,分类模型往往倾向于将样本标记为多数类,导致分类性能下降^[10]。数据流的非平衡问题可以通过采样和模型两种方法加以解决^[11-12]。采样方法是指通过减少多数类样本或增加少数类样本来平衡类分布,但这种方式会丢失有用信息或增加过拟合的风险。模型方法是指通过给少数类样本增加额外的损失代价或者基于Boosting方法重新训练分错的样本,使分类模型更加关注少数类样本。本文在模型初始化和更新时使用基于算法层面的类不平衡处理方式。

本文提出了一种有限标签下的非平衡数据流分类方法(imbalanced data stream classification method with limited labels, IDS-ELM),该方法分为初始化阶段和在线学习阶段。在初始化阶

段,训练初始极限学习机。在线学习阶段,首先使用初始极限学习机预测样本并计算预测确定性,结合阈值矩阵请求难分类和少数类样本标签;然后计算概念漂移指数并判断是否发生概念漂移,如果没有发生概念漂移,则计算请求到标签样本的重要性,更新极限学习机,如果发生漂移,则重构极限学习机。本文的主要贡献如下。

- 定义了预测概率差值与信息熵相结合的样本预测确定性度量,基于此提出了不确定性标签请求策略,兼顾了预测概率大的值之间的差异和预测概率的整体分布。
- 定义了基于类不平衡比率和样本预测误差的样本重要性度量,使极限学习机的更新兼顾了数据流的类不平衡比率和数据分布的变化。
- 提出了基于概念漂移指数的分类器更新与重构机制,使提出的数据流分类框架可以适用于4种不同类型的概念漂移。

1 相关工作

1.1 基于极限学习机的数据流分类方法

基于极限学习机的数据流分类方法具有学习速度快、泛化能力强和参数设置简单的优点。Huang等^[13]提出的极限学习机(extreme learning machine, ELM)是一种单隐藏层前馈神经网络,ELM的输入层权重和隐层偏差是随机生成的,在学习过程中保持不变,无须迭代调整,因此ELM具有学习速度快和泛化能力强^[13]的优点,但该方法并不具备数据流分类的能力。Mirza等^[14]提出了加权在线顺序极限学习机(weighted online sequential extreme learning machine, WOSELM),采用类不平衡比率对样本分配权重,以增强少数

类样本的作用,但该方法在评价样本重要性时只考虑了类不平衡比率,而没有考虑分类器对样本的预测误差。杜占龙等^[15]提出了基于正则化与遗忘因子的极限学习机(online sequential extreme learning machine based on regularization and forgetting factor, FROSELM),该方法采用遗忘机制和正则化机制,使分类器更关注最近到达的数据流。蔡桓等^[16]提出的自适应在线顺序极限学习机分类算法(adaptive online sequential extreme learning machine, AOSELM)通过引入自适应遗忘因子与概念漂移检测机制,能够根据动态变化的数据流进行自适应学习,但AOSELM不具备处理非平衡数据流的能力。陆克中等^[17]提出了一种具有自适应遗忘因子的加权在线顺序极限学习机集成算法(ensemble of weighted online sequential extreme learning machine with adaptive forgetting factor, EAF-WOSELM),该算法融合WOSELM的加权机制和FROSELM的遗忘机制,同时具备适应非平衡和概念漂移数据流的能力,但是该方法只适用于二分类数据流。

以上基于极限学习机的数据流分类方法需要获取全部样本的标签,然而获取代价非常高,限制了此类方法的应用。为此,本文提出了不确定性标签请求策略,通过请求部分难分类和少数类样本的标签对分类器进行训练和更新。

1.2 数据流主动学习方法

监督学习和半监督学习都是机器学习中常见的方法,它们都需要通过监督信息来训练模型。机器学习根据监督信息的获取方式分为被动学习和主动学习两种类型。被动学习的监督信息是预先给定的,此类

方法通常需要大量的标记数据, 监督信息的质量不能得到保证。与之不同, 主动学习允许在学习过程中灵活选择获取监督信息的样本, 以确保监督信息的质量, 降低获取代价^[18]。主动学习中标签请求策略至关重要^[19], 常见的标签请求策略包括不确定策略、随机策略和混合策略。不确定策略选择那些模型认为最不确定的样本; 随机策略随机地选择样本; 混合策略^[20]则综合考虑不确定性和随机性进行样本选择。

主动学习与极限学习机的结合可以生成有效的学习模型。一方面主动学习可以选择信息量最大的样本, 另一方面极限学习机计算成本低, 训练速度快, 请求样本标签后可以快速更新模型。Qin 等^[21]提出的增量加权在线主动极限学习机 (active learning for class incremental and weighted sequential extreme learning machine, AI-WSELM) 首先根据类不平衡比率对不同类的样本赋予不同的权重, 随后采用了不确定标签请求策略和相似性标签请求策略, 该方法不能主动检测数据流中的概念漂移。研究者还对主动学习与决策树相结合的数据流分类方法进行了研究。Liu 等^[22]提出了面向概念漂移和多类不平衡数据流的主动学习方法 (comprehensive active learning method for multiclass imbalanced streaming data with concept drift, CALMID), 该方法采用基于非对称边缘阈值矩阵的不确定标签请求策略和随机标签请求策略。李艳红等^[23]提出了基于 Adaboost 的非平衡数据流在线主动学习方法 (online active learning method for imbalanced data stream, OALM-IDS), 该方法采用不平衡比率和自适应遗忘因子定义了样本的重要性度量, 并使用样本的不确定程度自适应地调整边缘阈值矩阵。

以上数据流主动学习方法基于有限样本标签实现了分类器的训练和更新, 但在衡量样本重要性时, 只考虑了部分预测概率。因此本文提出了预测概率差值与信息熵相结合的样本预测确定性度量。

2 问题的形式化定义及相关算法

2.1 问题的形式化定义

设本文要处理的数据流 $DS = \{(x_1, y_1), (x_2, y_2), \dots, (x_t, y_t) \dots\}$, 其中样本 $x_t = [x_{t1}, x_{t2}, \dots, x_{td}]^T$, d 为样本的特征数, $y_t \in \{C_1, C_2, \dots, C_m\}$ 为样本 x_t 的类别标签, m 为样本的类别数。

为了对实时到达潜在无穷的数据流进行分类, 在初始化阶段, 利用最初的 S 个样本及其类别标签训练初始极限学习机 E (本文采用加权在线极限学习机作为分类器), 同时将这 S 个样本的类别标签存入标签窗口 $LW[S]$, 将预测结果窗口 $PW[S]$ 初始化为 Null。

在线学习阶段, 当新样本 x_t 到达后, 使用 E 对其进行类别预测, 并计算样本预测的确定性。使用阈值矩阵 $M[m \times m]$ 保存样本预测确定性的阈值, 用于筛选出难分类和少数类的样本, M 中元素的初值为 θ_0 。

本文采用基于样本预测确定性的标签请求策略与随机标签请求策略相结合的混合标签请求策略。 $LW[S]$ 用于保存最近 S 个连续样本的标签 (如果为基于随机策略请求到的样本, 则保存真实标签; 否则存入 Null), 用于计算不平衡比率 $imb_t = (imb_t^1, imb_t^2, \dots, imb_t^m)$ 。 $PW[S]$ 用于保存最近 S 个样本的预测结果 (如果为基于随机策略请求到的样本, 若预测正确则存入 True, 预测错误则存入 False; 否则存入 Null)。

每当请求到样本 x_i 的真实标签, 基于滑动窗口 PW 中保存的样本预测结果计算概念漂移指数 $CDI(x_i)$, 如果没有发生概念漂移, 则计算样本 x_i 的权重 $w(x_i)$, 并更新 E ; 如果发生概念漂移, 则使用漂移之后的 S 个样本重新构造 E 。

2.2 极限学习机算法

2.2.1 ELM 算法

ELM 算法^[13]是一种单隐藏层前馈神经网络算法, 适用于静态数据分类。假设训练集包括 S 个样本, ELM 算法有 L 个隐层节点, 通过求得使目标函数 $\|H_0\beta - Y_0\|$ 最小的输出层权重 β_0 来完成整个网络的训练过程。其中标签矩阵 $Y_0 = [y_1, \dots, y_S]^T$, 隐层输出矩阵 H_0 为:

$$H_0 = \begin{bmatrix} G(a_1, b_1, x_1) & \cdots & G(a_L, b_L, x_1) \\ \vdots & \ddots & \vdots \\ G(a_1, b_1, x_S) & \cdots & G(a_L, b_L, x_S) \end{bmatrix} \quad (1)$$

其中, a_i 和 b_i 分别表示第 i 个隐层节点的权重向量和偏置。根据广义逆的计算方法, 可以计算出 ELM 的输出层权重 β_0 为:

$$\beta_0 = R_0 H_0^T Y_0 \quad (2)$$

其中, $R_0 = (H_0^T H_0)^{-1}$ 。

2.2.2 WOSELM 算法

WOSELM 算法^[14]用于处理非平衡数据流, 该算法根据类不平衡比率为样本分配权重, 从而增强少数类的作用。WOSELM 算法分为两阶段, 分别为初始化阶段和在线学习阶段。WOSELM 在初始化阶段的输出层权重 β_0 为:

$$\beta_0 = R_0 H_0^T W Y_0 \quad (3)$$

$$R_0 = (H_0^T W_0 H_0)^{-1} \quad (4)$$

其中, $W = \text{diag}([w_1, w_2, \dots, w_S])$ 为初始化阶段样本的权重矩阵。在线学习阶段 WOSELM 的递推计算式为:

$$R_{k+1} = R_k - \frac{R_k h_{k+1}^T h_{k+1} R_k}{w_{k+1}^{-1} + h_{k+1}^T R_k h_{k+1}} \quad (5)$$

$$\beta_{k+1} = \beta_k + R_{k+1} h_{k+1}^T w_{k+1} (y_{k+1} - h_{k+1} \beta_k) \quad (6)$$

其中, w_{k+1} 为在线阶段样本权重。

3 本文方法

3.1 不确定性标签请求策略

不确定性标签请求策略旨在基于样本预测的不确定性筛选出难分类样本。为此, 首先需要定义样本预测的不确定性度量。

目前, 样本预测的不确定性度量通常采用基于边缘抽样或基于信息熵的方法。基于边缘抽样的方法度量样本的两个最大预测概率值之间的差值, 差值越小, 表明样本预测的不确定性越低。基于信息熵的方法度量样本的所有预测概率的混乱程度, 越混乱, 信息熵越大, 表明样本预测的不确定性越低。边缘抽样的方法只考虑了两个最大预测概率值之间的差值, 忽略了其他预测概率值的分布情况。基于信息熵的方法虽然考虑了所有的预测概率的分布, 但忽略了在评价样本预测的不确定性时, 预测概率大的值之间的差异要比预测概率小的值之间的差异更重要。

为了提高标签请求策略的有效性, 本文基于以下两个方面评价样本预测的不确定性。

- 分类器对样本的第一大预测概率和第二大预测概率的差值: 差值越小, 说明

对样本预测越不确定,反之则越确定。

- 分类器对样本预测概率的信息熵:信息熵越小表示预测概率的分布越混乱,说明对样本的预测越确定,反之则越不确定。

样本预测的确定性度量定义如式(7)所示。

$$D(x_t) = \alpha \times (P(\hat{y}_{c_1}|x_t) - P(\hat{y}_{c_2}|x_t)) + (1 - \alpha) \times \left(1 - \frac{\sum_{i=1}^m P(\hat{y}_{c_i}|x_t) \log_2 P(\hat{y}_{c_i}|x_t)}{\log_2 m} \right) \quad (7)$$

其中, $P(\hat{y}_{c_1}|x_t)$ 和 $P(\hat{y}_{c_2}|x_t)$ 分别为样本的第一大和第二大预测概率, $\alpha \in [0, 1]$ 为调节参数, $P(\hat{y}_{c_i}|x_t)$ 为样本 x_t 的第 i 大预测概率, m 为样本的类别数。

确定了样本预测的确定性度量后,还需要指定样本预测的确定性阈值,才能筛选出难分样本进行人工标注。本文使用阈值矩阵 M 保存样本预测确定性阈值。阈值矩阵的元素 M_{ij} 为样本的第一大预测概率对应的 c_i 类和第二大预测概率 c_j 类对应的样本预测确定性阈值。

基于样本预测的确定性度量 $D(x_t)$ 和阈值矩阵 M , 本文提出了一种不确定性标签请求策略, 具体的不确定标签请求策略如下。

- 如果 $D(x_t) > M[c_1][c_2]$, 表明样本 x_t 的最大预测概率与第二大预测概率的差值与所有预测概率的信息熵的加权和(乘积)在预定范围内, 即分类器对 x_t 的预测结果是可信的, 此时不需要请求 x_t 的真实标签, 也不需要调整阈值矩阵 M 。

- 如果 $D(x_t) \leq M[c_1][c_2]$, 表明分类器对样本 x_t 的预测结果不可信, 需要请求样本的真实类别标签 y_t 。若 $\hat{y}_{c_1} = y_t$, 表明阈值 $M[c_1][c_2]$ 对该样本过于严苛, 应减小。如果 $\text{imb}_t^{y_t} > 0.5$, 表明 x_t 属于多数类样本, 此时应进一步减小阈值 $M[c_1][c_2]$, 以实现

多数类样本比少数类样本具有更少的标签请求机会的目的。

阈值矩阵的调整方式如下:

$$M[c_1][c_2] = M[c_1][c_2] \times (1 - \gamma) \quad (8)$$

其中, $\gamma \in [0, 1]$ 。

3.2 更新样本的重要性度量

在线学习阶段, 每当请求到样本的标签, 均需要利用该样本更新极限学习机, 以适应数据流分布的变化。现有的方法通常根据类不平衡比率对样本赋权^[14], 以解决数据流中的类不平衡问题。在对样本赋权时, 除了类不平衡比率, 极限学习机对样本预测误差也是衡量样本重要性的重要指标, 样本的预测误差越大, 表明该样本与历史数据分布越不一致, 在更新极限学习机时应赋予更高的权重。为了提高极限学习机对数据流变化的适应性, 本文在衡量样本重要性时考虑以下两个方面的因素。

- 基于类不平衡比率为不同类的样本赋予不同的权重, 类不平衡比率小的样本权重较大, 反之则较小。

- 请求到样本真实标签后需要更新极限学习机, 此时需要考虑极限学习机对样本的预测误差, 预测误差大的样本权重较大, 反之则较小。

在线学习阶段样本的重要性(权重)的定义如下:

$$w(x_t) = e^{-\frac{\text{imb}_t^{y_t}}{f(x_t)}} \quad (9)$$

其中, $f(x_t)$ 为极限学习机对样本的预测误差, 定义如下:

$$f(x_t) = 1 - P(\hat{y}_t|x_t) \quad (10)$$

其中, $P(\hat{y}_t|x_t)$ 为极限学习机对样本 x_t 真实类别 y_t 的预测概率。 $\text{imb}_t^{y_t}$ 为 y_t 类在 t 时刻的

类不平衡比率, 计算方式如下:

$$\text{imb}_t^{y_i} = \frac{\text{num}_{y_i}}{(S - \text{num}_{\text{Null}})/m} \quad (11)$$

其中, num_{y_i} 为标签窗口 LW 中 y_i 类的标签个数, num_{Null} 为空标签 Null 的个数。

在初始化阶段, 根据类不平衡比率为样本加权, 样本的重要性, 即权重定义如下:

$$w(x_t) = e^{-\text{imb}_t^{y_i}} \quad (12)$$

其中, $\text{imb}_t^{y_i}$ 为 y_i 类在 t 时刻的类不平衡比率。

3.3 分类器的更新与重构机制

为了适应数据流的动态变化, 现有的工作往往采用具有遗忘机制的极限学习机^[15-17], 其核心是根据到达时间为样本分配不同的权重, 为最近到达的样本分配较高的权重, 并用于更新极限学习机。这种处理方法不考虑极限学习机的重构, 对漂移过程中旧的数据分布仍时有出现的情况是不适用的 (例如逐渐型概念漂移)。为了解决此类问题, 本文提出基于概念漂移指数的极限学习机更新与重构机制。

预测结果窗口 PW[S] 中保存了最近 S 个样本的预测结果, 如果 t 时刻 PW 中样本的预测准确率与最大预测概率 (数据流稳定时) 相差较大, 则认为 t 时刻发生概念漂移的可能性较大。可将概念漂移指数定义如下:

$$\text{CDI}(x_t) = \frac{P_{\max} - (P_t - S_t)}{S_{\max}} \quad (13)$$

其中, $P_{\max}$ 为数据流稳定时分类器预测准确率的最大值, $S_{\max} = \sqrt{P_{\max} \times (1 - P_{\max})}$ 为 $P_{\max}$ 的标准差, P_t 为 t 时刻 PW 中样本的预测准确率, $S_t = \sqrt{P_t \times (1 - P_t)}$ 为 P_t 的标准差。 P_t 的计算方法如下:

$$P_t = \frac{\text{num}_{\text{True}}}{S - \text{num}_{\text{Null}}} \quad (14)$$

其中, num_{True} 为预测结果窗口 PW 中 True 的个数, num_{Null} 为 PW 中空标签 Null 的个数。

当概念漂移指数 $\text{CDI}(x_t) < \varepsilon$ 时, 表明未检测到概念漂移, 使用请求到标签的样本 x_t 更新分类器 E ; 当概念漂移指数 $\text{CDI}(x_t) \geq \varepsilon$ 时, 表明发生了概念漂移, 使用新到达的样本重构分类器 E 。

3.4 算法框架

本文提出一种有限标签下的非平衡数据流分类方法 IDS-ELM, 包括初始化和在线学习两个阶段, IDS-ELM 算法的整体框架如图 1 所示。初始化阶段的主要任务是使用数据流最初到达的 S 个样本训练极限学习机 E 。在线学习阶段包括标签请求、概念漂移检测、分类器更新与重构 3 个任务。

(1) 标签请求

首先分类器预测新样本; 然后根据混合标签请求策略判断是否需要请求样本标签, 若需要则由专家标注该样本, 否则将预测结果作为样本标签, 分类器 E 继续预测下一个样本; 接下来更新标签窗口 LW 和预测结果窗口 PW; 最后根据样本预测结果和类不平衡比率调整阈值矩阵 M 。

(2) 概念漂移检测

在基于混合标签查询策略为样本请求到真实标签后, 基于预测结果窗口计算概念漂移指数, 并判断数据流中是否发生概念漂移。

(3) 分类器的更新与重构

若没有发生概念漂移, 则计算更新样本的重要性, 并用于更新分类器 E , 否则使用新到达的 S 个样本重新构造分类器 E 。

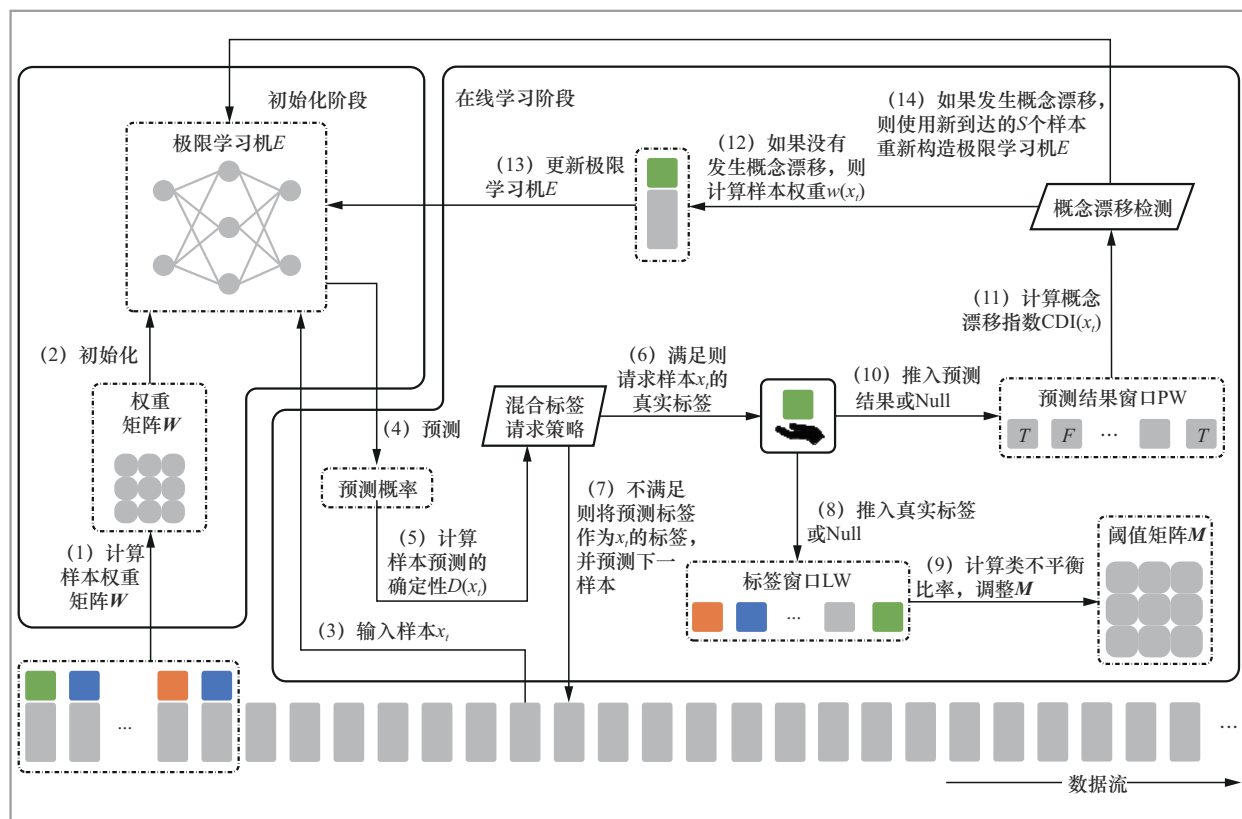


图1 IDS-ELM算法的整体框架

3.5 算法伪代码

本节给出有限标签下的非平衡数据流分类算法IDS-ELM以及不确定性标签请求算法 (uncertainty label request, ULR) 的伪代码。

3.6 复杂度分析

设初始化模型的样本数为 S ，在线学习的样本数为 N ， d 是数据流的特征数， L 是隐层的节点数， m 是类别数，初始化阶段训练极限学习机的时间复杂度为 $O(LS^2 + L^2S + L^3 + LSm)$ 。单个样本预测的时间复杂度是 $O(Ld + Lm)$ ，单个样本请求标签的时间复杂度是 $O(m)$ 。因此，在线学习阶段的时间复杂度为 $O(NLd + NLm + Nm)$ 。初始化阶

段需要使用标签窗口 $LW[S]$ 和预测结果窗口 $PW[S]$ ，因此，初始化阶段的空间复杂度为 $O(S)$ 。在线学习阶段需要使用阈值矩阵 $M[m][m]$ 、标签窗口 $LW[S]$ 和预测结果窗口 $PW[S]$ ，因此，在线学习阶段的空间复杂度为 $O(m^2 + S)$ 。

4 实验结果与分析

4.1 实验数据集和运行环境

本文实验环境为 Windows 10 操作系统，CPU 为 Intel Core i7-7700 3.60 GHz，内存为 8 GB，开发软件使用 PyCharm，所有分类算法均基于 Python 实现。

算法 1: IDS-ELM

Input: 数据流 DS, 已处理样本数 g , 带标签的样本数 l , 标签预算 B , 随机选择比率 η , 漂移检测阈值 ε , 窗口大小 S , 标签窗口 LW[S], 预测结果窗口 PW[S]

Output: 极限学习机 E 的预测结果

1. 使用最初的 S 个样本初始化极限学习机 E ;
2. **while**(输入一个样本 x_t) **do**
3. 输出 E 对样本 x_t 预测结果 $P(\hat{y}_C|x_t), 1 \leq j \leq m$;
4. $g = g + 1$;
5. labelling=False;
6. **if**($l/g < B$)**then**
7. 生成一个随机变量 $\zeta, \zeta \in [0, 1]$;
8. **if**($\zeta < \eta$)**then**
9. labelling=True;
10. 请求样本 x_t 的真实标签 y_t , 并将 y_t 存入 LW 中;
11. **if** ($y_t = \hat{y}_C$) **then**
12. 将 True 存入 PW 中;
13. **else**
14. 将 False 存入 PW 中;
15. **end if**;
16. **else if** (ULR(x_t)) **then**
17. labelling=True;
18. 请求样本 x_t 的真实标签 y_t ;
19. **end if**;
20. 将 Null 存入 LW 和 PW 中;
21. **end if**;
22. **if** ($y_t \in \{C_1, C_2, \dots, C_k\}$ AND labelling==True) **then**
23. $l = l + 1$;
24. 根据式 (13) 计算概念漂移指数 CDI(x_t);
25. **if** (CDI(x_t)< ε) **then**
26. 根据式 (9) 计算样本重要性 $w(x_t)$;
27. 根据式 (6) 用 x_t 对极限学习机 E 进行更新;
28. **else**
29. 根据式 (3) 用 x_t 之后 S 个样本重构极限学习机 E ;
30. **end if**;
31. **end if**;
32. **if** ($y_t \in \{C_1, C_2, \dots, C_k\}$ AND labelling== False) **then**
33. 将预测标签作为 x_t 的标签;
34. **end if**;
35. 继续预测下一样本;
36. **end while**.

算法 2：ULR

Input: 新到达样本 x_i , 阈值矩阵 $M[m \times m]$

Output: labelling $\in \{\text{True}, \text{False}\}$.

- 1. 根据式 (7) 计算样本预测的确定性 $D(x_i)$
- 2. **if** $(D(x_i) \leq M[c_1][c_2])$ **then**
- 3. labelling=True;
- 4. 请求样本的真实标签 y_i ;
- 5. **if** $(y_i = \hat{y}_{c_i})$ **then**
- 6. 根据式 (8) 调整矩阵阈值;
- 7. 根据式 (11) 计算 y_i 的类不平衡比率;
- 8. **if** $(imb_{y_i} > 0.5)$ **then**
- 9. 根据式 (8) 调整矩阵阈值;
- 10. **end if**;
- 11. **end if**;
- 12. **else**
- 13. labelling=False;
- 14. **end if**.

实验使用 6 个人工数据流和 3 个真实数据流，数据流的特征见表 1，其中 DS₁、DS₂、DS₃和 DS₄为类不平衡比率固定的非平衡数据流，DS₅和 DS₆为类不平衡比率变

化的非平衡数据流。在 6 个人工数据流中分别设置了 2 次概念漂移和 0.5% 的异常点。DS₁、DS₃和 DS₅中的 2 次概念漂移为突变型和增量型，分别位于为数据流的第 30 000 和 55 000 个样本处，增量型概念漂移的宽度为 5 000；DS₂、DS₄和 DS₆中的 2 次概念漂移为重复型和逐渐型，重复型概念漂移对应的 2 次漂移的位置分别为 25 000 和 50 000（重复型概念漂移相当于两次突变型概念漂移），逐渐型概念漂移的位置为 70 000，逐渐型概念漂移的宽度为 5 000。真实数据流 HAR（human activity recognition）、Letter 和 Avila 均来自 UCI 机器学习数据库，其中的概念漂移类型和异常点未知。

4.2 IDS-ELM 算法分类性能评价

用本文提出的 IDS-ELM 算法与 6 种分类算法（AOSELM^[16]、FROSELM^[15]、WOSELM^[14]、AI-WSELM^[21]、CALMID^[22]和 OALM-IDS^[23]）在 6 个人工数据流和 3 个真实数据流上进行分类性能比较，使用精确率（P）、召回率（R）、F1 值、Kappa 系数和 ROC 曲线作为评价指标。在

表1 数据流的特征

数据流	样本数	特征数	类别数	类分布	漂移类型	异常点
DS ₁	100 000	50	5	(0.4; 0.2; 0.2; 0.16; 0.04)	突变、增量	0.5%
DS ₂	100 000	50	5	(0.4; 0.2; 0.2; 0.16; 0.04)	重复、逐渐	0.5%
DS ₃	100 000	50	5	(0.6; 0.1; 0.1; 0.15; 0.05)	突变、增量	0.5%
DS ₄	100 000	50	5	(0.6; 0.1; 0.1; 0.15; 0.05)	重复、逐渐	0.5%
DS ₅	100 000	50	5	(0.4; 0.2; 0.2; 0.16; 0.04) (0.6; 0.1; 0.1; 0.15; 0.05)	突变、增量	0.5%
DS ₆	100 000	50	5	(0.4; 0.2; 0.2; 0.16; 0.04) (0.6; 0.1; 0.1; 0.15; 0.05)	重复、逐渐	0.5%
HAR	10 299	561	6	—	—	—
Letter	20 000	16	26	—	—	—
Avila	20 867	10	12	—	—	—

6 种对比分类算法中，AOSELM、FROSELM 和 WOSELM 为监督学习算法，AI-WSELM、CALMID 和 OALM-IDS 为主动学习算法。

IDS-ELM、AOSELM、FROSELM、WOSELM 和 AI-WSELM 均为基于极限学习机的分类算法，CALMID 和 OALM-IDS 为集成分类算法，基分类器为 Hoeffding 树。所有主动学习算法的标签预算均设为 20%，AOSELM、FROSELM、WOSELM、CALMID 和 OALM-IDS 的初始化实例数均为 500，CALMID 和 OALM-IDS 的滑动窗口大小为 500。IDS-ELM、FROSELM、WOSELM 和 AI-WSELM 的隐层节点数设

置为 500，所有基于极限学习机的分类算法的激活函数均采用 Sigmoid。算法 IDS-ELM 中标签窗口 **LW** 和预测结果窗口 **PW** 的大小 S 设置为 500，调节参数 α 设置为 0.8，阈值矩阵中元素的初始值 θ_0 设置为 0.2， ϵ 设置为 3。所有实验数据为 10 次实验结果的平均值。

由表 2~表 5 可知，本文提出的 IDS-ELM 算法在所有数据流上的 P 值和 Kappa 系数均优于其他 6 种对比分类算法，且在 6 个人工数据流和 2 个真实数据流（HAR 和 Letter）上的 R 值和 $F1$ 值上也均优于其他 6 种对比算法，但 IDS-ELM 算法在真实数据流 Avila 上 R 值和 $F1$ 值要略低于

表2 7种算法的P值

数据流	AOSELM	FROSELM	WOSELM	AI-WSELM	CALMID	OALM-IDS	IDS-ELM
DS ₁	0.7294	0.9105	0.8581	0.8839	0.8940	0.9314	0.9453
DS ₂	0.7137	0.9042	0.8362	0.8646	0.8750	0.9248	0.9393
DS ₃	0.7262	0.9031	0.8431	0.8704	0.8926	0.9224	0.9390
DS ₄	0.6936	0.8823	0.8196	0.8599	0.8688	0.9162	0.9281
DS ₅	0.7045	0.8918	0.8280	0.8593	0.8847	0.9172	0.9301
DS ₆	0.6806	0.8682	0.8087	0.8422	0.8619	0.8920	0.9125
HAR	0.8106	0.8705	0.8461	0.8674	0.6946	0.7098	0.8867
Letter	0.6102	0.6957	0.7941	0.7768	0.6347	0.6547	0.7989
Avila	0.5471	0.5814	0.6373	0.5737	0.5306	0.5619	0.6401

表3 7种算法的R值

数据流	AOSELM	FROSELM	WOSELM	AI-WSELM	CALMID	OALM-IDS	IDS-ELM
DS ₁	0.7062	0.8896	0.8740	0.8866	0.8541	0.9005	0.9429
DS ₂	0.6952	0.8816	0.8557	0.8207	0.8278	0.8891	0.9322
DS ₃	0.7023	0.8870	0.8719	0.8844	0.8420	0.8982	0.9373
DS ₄	0.6767	0.8762	0.8496	0.8203	0.8072	0.8813	0.9262
DS ₅	0.6975	0.8814	0.8671	0.8741	0.8248	0.8832	0.9228
DS ₆	0.6674	0.8634	0.8319	0.8154	0.7953	0.8649	0.9168
HAR	0.7967	0.8638	0.8435	0.8506	0.6735	0.6635	0.8769
Letter	0.5947	0.6888	0.7871	0.7878	0.6448	0.6264	0.7884
Avila	0.4609	0.5758	0.5963	0.5361	0.4807	0.5144	0.5870

表4 7种算法的F1值

数据流	AOSELM	FROSELM	WOSELM	AI-WSELM	CALMID	OALM-IDS	IDS-ELM
DS ₁	0.7176	0.8999	0.8660	0.8852	0.8736	0.9157	0.9441
DS ₂	0.7043	0.8928	0.8458	0.8421	0.8507	0.9066	0.9357
DS ₃	0.7141	0.8950	0.8573	0.8773	0.8666	0.9101	0.9381
DS ₄	0.6850	0.8792	0.8343	0.8396	0.8369	0.8984	0.9271
DS ₅	0.7010	0.8866	0.8471	0.8666	0.8537	0.8999	0.9264
DS ₆	0.6739	0.8665	0.8201	0.8286	0.8273	0.8775	0.9146
HAR	0.8036	0.8672	0.8448	0.8589	0.6839	0.6859	0.8817
Letter	0.6024	0.6922	0.7906	0.8077	0.6497	0.6305	0.7931
Avila	0.4829	0.5785	0.6161	0.5543	0.4904	0.5371	0.6124

表5 7种算法的Kappa值

数据流	AOSELM	FROSELM	WOSELM	AI-WSELM	CALMID	OALM-IDS	IDS-ELM
DS ₁	0.7030	0.9124	0.8692	0.8884	0.8697	0.9093	0.9495
DS ₂	0.6973	0.9080	0.8380	0.7629	0.8389	0.8982	0.9359
DS ₃	0.6956	0.9042	0.8683	0.8415	0.8628	0.8902	0.9418
DS ₄	0.6823	0.8941	0.7913	0.7412	0.8165	0.8901	0.9333
DS ₅	0.6826	0.8882	0.8021	0.8382	0.8335	0.8854	0.9314
DS ₆	0.6711	0.8733	0.7701	0.7225	0.8061	0.8762	0.9273
HAR	0.7781	0.8357	0.8309	0.8262	0.6946	0.7041	0.8591
Letter	0.5823	0.6821	0.7796	0.7716	0.6320	0.6133	0.7815
Avila	0.4313	0.5434	0.5352	0.4761	0.4459	0.4687	0.5455

WOSELM算法，其原因在于WOSELM采用了基于类不平衡比率的样本加权机制，并且需要使用所有样本的标签信息。

由表2~表5的实验结果可知，除了AOSELM算法，其他4种基于极限学习机的算法在3个真实数据流上的分类性能均优于两种基于决策树的算法，AOSELM算法隐层节点数较少，导致其在所有数据流上的分类性能都低于其他6种算法；本文提出的IDS-ELM算法在所有数据流上的分类性能均优于其他3种主动学习算法；所有算法在不平衡比率较小（DS₁和DS₂）、不平衡比率较大（DS₃和DS₄）和不平衡比

率变化（DS₅和DS₆）3类数据流上的分类性能依次下降；所有算法在含有突变型和增量型概念漂移数据流上的分类性能都优于含有重复型和逐渐型概念漂移数据流上的分类性能，这是由于含有重复型和逐渐型概念漂移的数据流相当于有3次漂移（一次重复型概念漂移相当于两次突变型概念漂移），并且逐渐型概念漂移过程中新旧数据分布交替出现，算法需要更长的适应时间。

ROC曲线下面积能够直观地反映出分类性能的好坏，所有分类算法在6个人工数据流和3个真实数据流的ROC曲线如

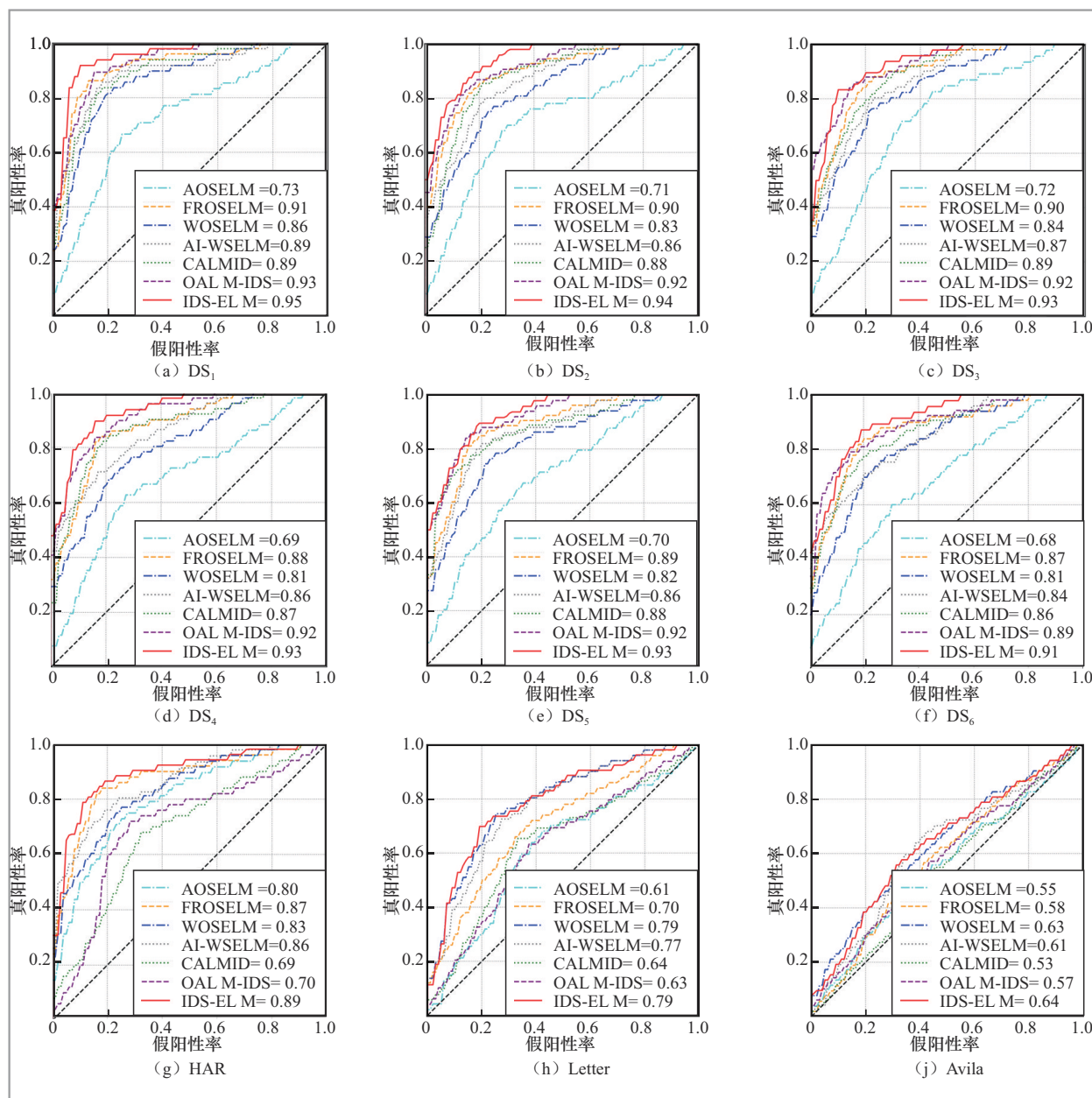


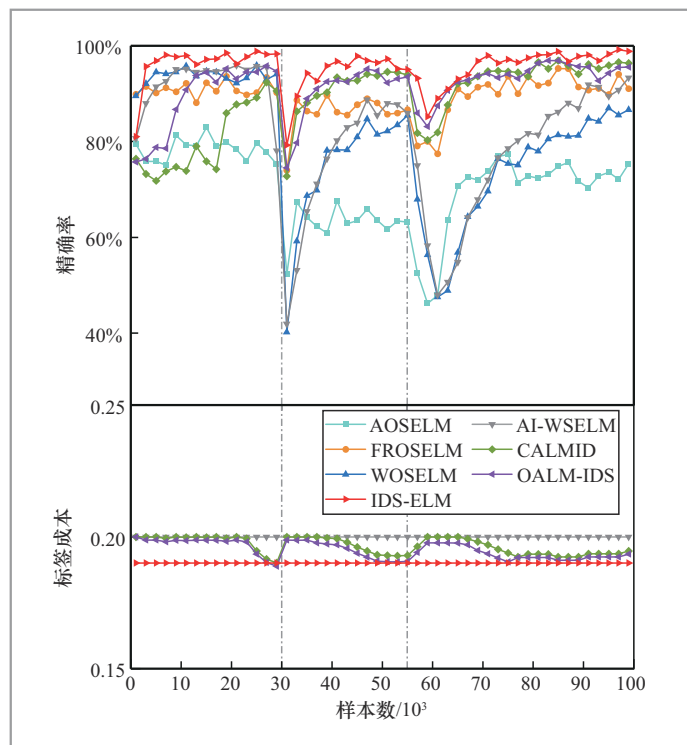
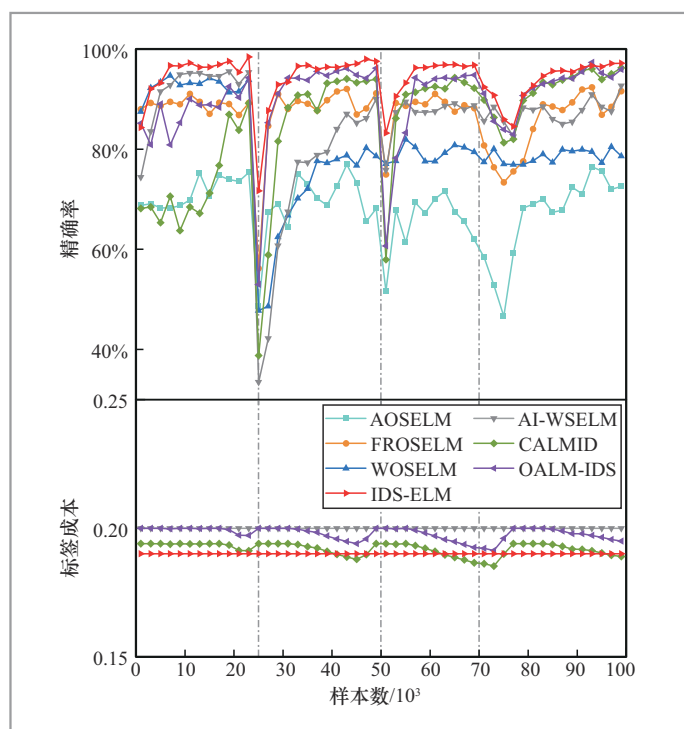
图2 7种算法的ROC曲线

图2所示。由图2可知，除AOSELM算法以外，所有基于极限学习机的算法（IDS-ELM、FROSELM和AI-WSELM）在所有真实数据流上的ROC曲线下面积均优于两种基于决策树的算法（CALMID和OALM-IDS），IDS-ELM除了在Letter数据流上与WOSELM算法的ROC曲线下

面积相同外，在其他数据流上的ROC曲线下面积均优于其他分类算法。

DS₅和DS₆和Letter上的分类精确率曲线如图3~图5所示，图中虚线表示概念漂移发生的位置。

由图3~图5可知，IDS-ELM算法在这3个数据流上的分类精确率均优于其他6种

图3 DS₅上的分类精确率曲线图4 DS₆上的分类精确率曲线

算法。由图3、图4可知,在发生突变型、重复型、增量型和逐渐型概念漂移时,采用重构机制的算法(IDS-ELM、CALMID和OALM-IDS)的分类精确率可快速提升,这是由于分类器在发生概念漂移之后重新训练,可以更快地适应概念漂移,由于WOSELM和AI-WSELM没有应对概念漂移的机制,在发生概念漂移之后,分类精确率明显下降。FROSELM和AOSELM采用了遗忘机制,为最近到来的样本分配较高的权重并用于更新分类器,因此在发生逐渐型概念漂移时,相比增量型概念漂移分类性能较差;在逐渐型概念漂移期间,FROSELM和AOSELM的平均分类精确率分别为76.82%和52.62%;在增量型概念漂移期间,FROSELM和AOSELM的平均分类精确率分别为81.92%和54.01%。相比增量型和逐渐型概念漂移,发生突变型和重复型概念漂移时,所有算法的分类精确率均下降更快。IDS-ELM算法在这3个数据流上均可以用最少的标签成本获得最高的精确率。

基于上述实验结果可知,本文所提的IDS-ELM算法在各项评价指标上整体优于其他对比分类算法。

4.3 消融实验

为了验证IDS-ELM中引入的样本预测确定性度量、样本重要性度量和极限学习机更新与重构机制的有效性,在人工数据流DS₆上进行了4种消融实验。

(1) 将IDS-ELM的样本预测确定性度量替换为只考虑两个最大预测概率值的差值,得到IDS-ELM-rfs算法。

(2) 将IDS-ELM的混合标签请求策略替换为随机标签请求策略,得到IDS-ELM-r算法。

(3) 将IDS-ELM-r的样本的重要性度量替换为只考虑类不平衡比率, 得到IDS-ELM-ri算法。

(4) 将IDS-ELM-ri分类器的更新与重构机制替换为只对分类器进行更新而不重构, 得到IDS-ELM-riu算法。

消融实验的结果如图6所示。由图6可知, 只考虑两个最大概率值差值的IDS-ELM-rfs算法比IDS-ELM算法的精确率低; 只考虑随机标签请求策略的IDS-ELM-r算法比IDS-ELM算法的精确率更低; 只考虑类不平衡比率的样本重要性度量的IDS-ELM-ri算法比IDS-ELM-r算法的精确率进一步下降; 只对分类器更新的IDS-ELM-riu算法的精确率在漂移之后相比IDS-ELM-ri算法的精确率明显下降。因此, 本文在样本预测确定性度量、样本重要性度量和分类器的更新与重构机制方面的改进能够有效提升数据流分类的性能。

4.4 参数敏感性分析

样本预测确定性度量中的调节参数 α , 用于调节样本的第一大和第二大预测概率差值与样本预测概率信息熵的作用, 参数 θ_0 为阈值矩阵的元素初值。接下来测试参数 α 和 θ_0 对IDS-ELM算法分类性能的影响。实验使用含有异常点和概念漂移的3个合成数据流和3个真实数据流, 结果如图7、图8所示。

图7所示为 α 对IDS-ELM算法分类性能的影响, 由图7可知, 当 α 为0.8时, 在6个数据流上的 P 值、 R 值、 $F1$ 值、 $Kappa$ 系数都达到了最大值, 表明在度量样本预测确定性时, 第一大预测概率和第二大预测概率的差值非常重要; 当 α 大于0.8时, 算法分类性能下降, 可知分类器预

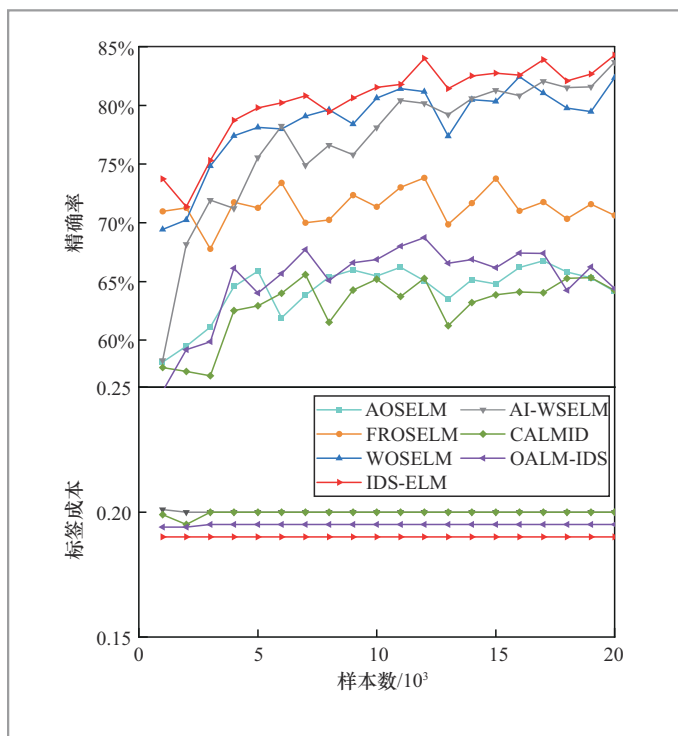


图5 Letter上的分类精确率曲线

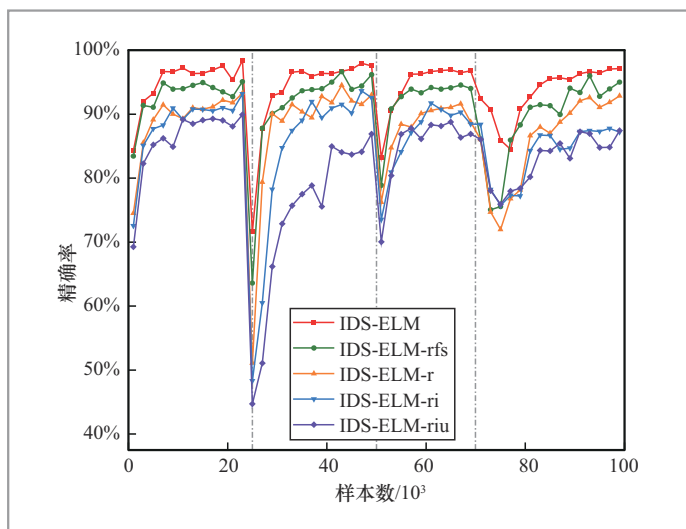
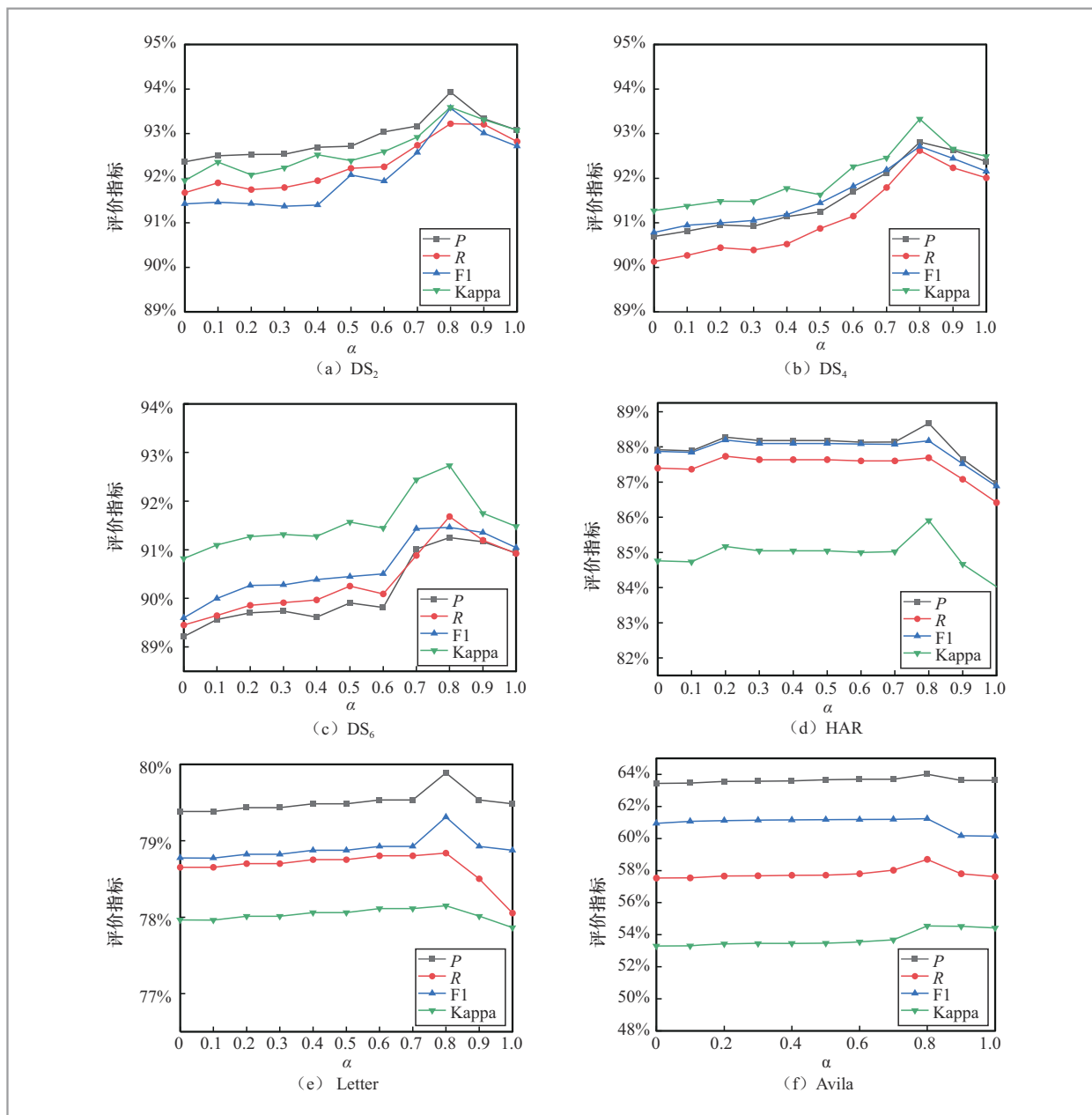


图6 消融实验的结果

测概率的信息熵的作用也至关重要。

图8所示为 θ_0 对IDS-ELM算法分类性能的影响, 由图8可知, 当 θ_0 为0.2时, 在6个数据流上的 P 值、 R 值、 $F1$ 值、 $Kappa$

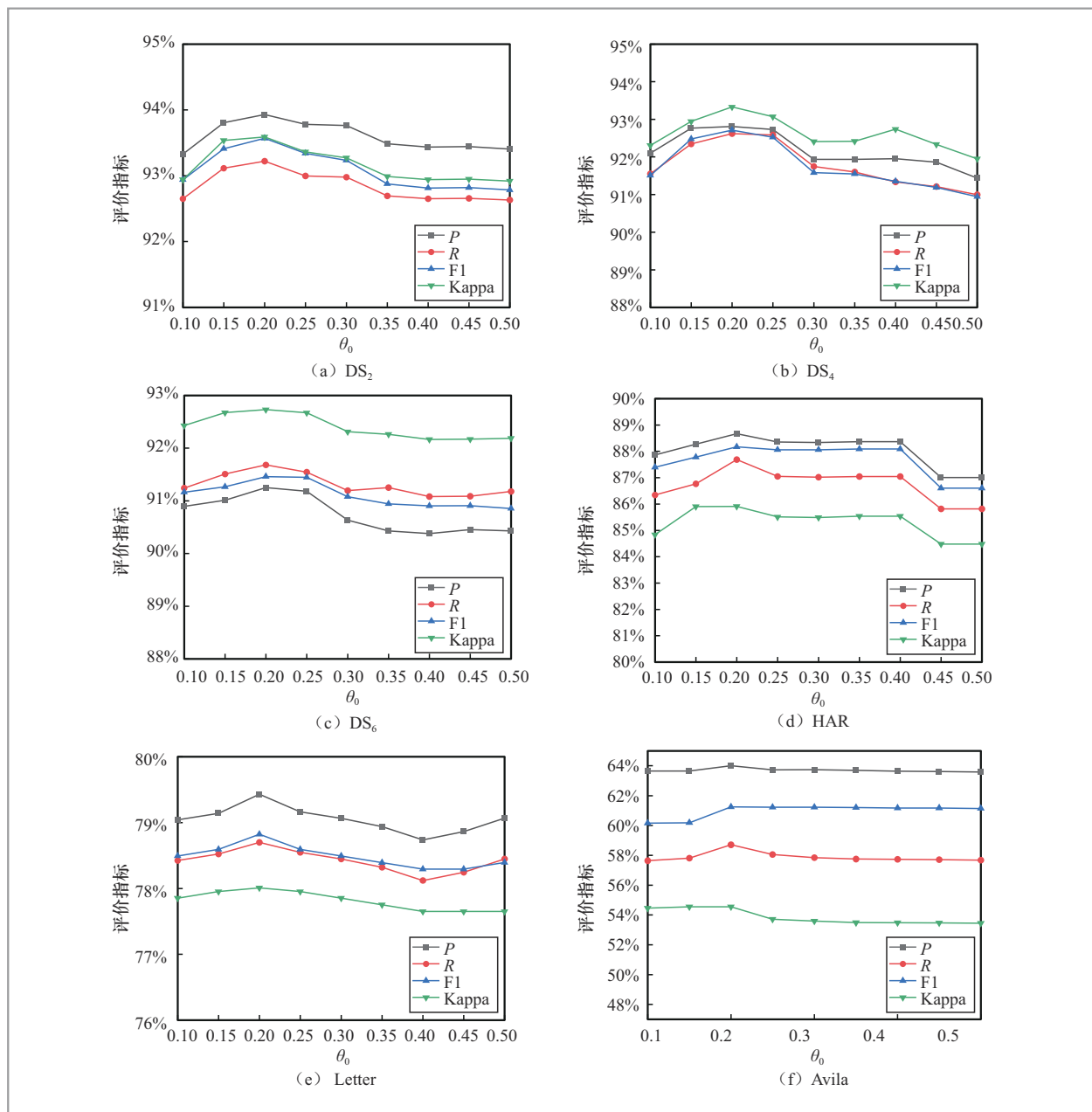
图7 参数 α 对算法的影响

系数达到最大值；当 θ_0 小于0.2时，请求不到难分类和少数类样本的标签，从而使IDS-ELM算法的分类性能下降；当 θ_0 大于0.2时，请求到的样本标签数量增加，但由于样本标签预算是有限的，也会导致不能请求到足够的难分类和少数类样本标签，

从而使IDS-ELM算法的分类性能下降。

5 结束语

本文研究了有限标签下的非平衡数据

图8 参数 θ_0 对算法的影响

流分类方法, 定义了预测概率差值与信息熵相结合的样本预测确定性度量, 并提出了不确定性标签请求策略, 兼顾了预测概率大的值之间的差异和预测概率的整体分布; 定义了基于类不平衡比率和样本预测误差的样本重要性度量, 用于极限学习机

的更新; 提出了基于概念漂移指数的分类器的更新与重构机制, 使本文提出的数据流分类框架可以适用于4种不同类型的概念漂移。

为了增强有限标签下的非平衡数据流分类方法的鲁棒性, 在未来工作中, 笔者

将关注以下问题。首先,已有的基于极限学习机的数据流分类方法通常需要提前指定隐层节点数,且保持隐层节点数不变,不能适应数据流的变化,需要研究不同数据流上最优隐层节点数的确定问题。其次,基于单分类器的数据流分类方法存在稳定性差的问题,难以适用于不同类型的数据流,因此需要研究有限标签下的数据流集成学习方法。

参考文献:

- [1] LIAO G B, ZHANG P, YIN H P, et al. A novel semi-supervised classification approach for evolving data streams[J]. Expert Systems with Applications, 2023, 215: 119273.
- [2] WANKHADE K K, DONGRE S S, JONDALE K C. Data stream classification: a review[J]. Iran Journal of Computer Science, 2020, 3(4): 239-260.
- [3] 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.
- [4] GULCAN E B, CAN F. Unsupervised concept drift detection for multi-label data streams[J]. Artificial Intelligence Review, 2023, 56(3): 2401-2434.
- [5] YU H L, WEBB G I. Adaptive online extreme learning machine by regulating forgetting factor by concept drift map[J]. Neurocomputing, 2019, 343: 141-153.
- [6] YANG L, SHAMI A. A lightweight concept drift detection and adaptation framework for IoT data streams[J]. IEEE Internet of Things Magazine, 2021, 4(2): 96-101.
- [7] 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.
- [8] KARIMIAN M, BEIGY H. Concept drift handling: a domain adaptation perspective[J]. Expert Systems with Applications, 2023, 224: 119946.
- [9] SHAHRAKI A, ABBASI M, TAHERKORDI A, et al. Active learning for network traffic classification: a technical study[J]. IEEE Transactions on Cognitive Communications and Networking, 2022, 8(1): 422-439.
- [10] WANG S, LUO H, HUANG S S, et al. Counterfactual-based minority oversampling for imbalanced classification[J]. Engineering Applications of Artificial Intelligence, 2023, 122: 106024.
- [11] DU H L, ZHANG Y, GANG K, et al. Online ensemble learning algorithm for imbalanced data stream[J]. Applied Soft Computing, 2021, 107: 107378.
- [12] WANG W Y, SUN D C. The improved AdaBoost algorithms for imbalanced data classification[J]. Information Sciences, 2021, 563: 358-374.
- [13] HUANG G B, ZHU Q Y, SIEW C K. Extreme learning machine: theory and applications[J]. Neurocomputing, 2006, 70(1/2/3): 489-501.
- [14] MIRZA B, LIN Z P, TOH K A. Weighted online sequential extreme learning machine for class imbalance learning[J]. Neural Processing Letters, 2013, 38(3): 465-486.
- [15] 杜占龙, 李小民, 郑宗贵, 等. 基于正则化与遗忘因子的极限学习机及其在故障预测中的应用[J]. 仪器仪表学报, 2015, 36(7): 1546-1553.
- DU Z L, LI X M, ZHENG Z G, et al. Extreme learning machine based on regularization and forgetting factor and its application in fault prediction[J]. Chinese

- Journal of Scientific Instrument, 2015, 36(7): 1546-1553.
- [16] 蔡桓, 陆克中, 伍启荣, 等. 面向概念漂移数据流的自适应分类算法[J]. 计算机研究与发展, 2022, 59(3): 633-646.
CAI H, LU K Z, WU Q R, et al. Adaptive classification algorithm for concept drift data stream[J]. Journal of Computer Research and Development, 2022, 59(3): 633-646.
- [17] 陆克中, 陈超凡, 蔡桓, 等. 面向概念漂移和类不平衡数据流的在线分类算法[J]. 电子学报, 2022, 50(3): 585-597.
LU K Z, CHEN C F, CAI H, et al. On-line classification algorithm for concept drift and class imbalance data stream[J]. Acta Electronica Sinica, 2022, 50(3): 585-597.
- [18] 孙必成. 基于主动学习的聚类算法研究[D]. 合肥: 安徽大学, 2022.
SUN B C. Research on clustering algorithms based on active learning[D]. Hefei: Anhui University, 2022.
- [19] YIN C Y, CHEN S S, YIN Z C. Clustering-based active learning classification towards data stream[J]. ACM Transactions on Intelligent Systems and Technology, 2023, 14(2): 38.
- [20] XU W H, ZHAO F F, LU Z C. Active learning over evolving data streams using paired ensemble framework[C]// Proceedings of the 2016 Eighth International Conference on Advanced Computational Intelligence (ICACI). Piscataway: IEEE Press, 2016: 180-185.
- [21] QIN J M, WANG C, ZOU Q H, et al. Active learning with extreme learning machine for online imbalanced multiclass classification[J]. Knowledge-Based Systems, 2021, 231: 107385.
- [22] LIU W K, ZHANG H, DING Z Y, et al. A comprehensive active learning method for multiclass imbalanced data streams with concept drift[J]. Knowledge-Based Systems, 2021, 215: 106778.
- [23] 李艳红, 王甜甜, 王素格, 等. 非平衡概念漂移数据流主动学习方法[J]. 自动化学报, 2024, 50(3): 589-606.
LI Y H, WANG T T, WANG S G, et al. Active learning method for imbalanced concept drift data stream[J]. Acta Automatica Sinica, 2024, 50(3): 589-606.

作者简介



李艳红 (1977-), 女, 博士, 山西大学计算机与信息技术学院副教授, 主要研究方向为数据挖掘和机器学习。



李志华 (1998-), 男, 山西大学计算机与信息技术学院硕士生, 主要研究方向为数据挖掘和机器学习。



郑建兴（1985-），男，博士，山西大学计算机与信息技术学院教授，主要研究方向为中文信息处理和数据挖掘。



白鹤翔（1980-），男，博士，山西大学计算机与信息技术学院教授，主要研究方向为空间数据挖掘。



郭鑫（1982-），女，博士，山西大学计算机与信息技术学院副教授，主要研究方向为自然语言处理。

收稿日期: 2024-05-20
通信作者: 李艳红, liyh@sxu.edu.cn
基金项目: 国家自然科学基金项目 (No.62272286, No.41871286); 山西省基础研究计划项目 (No.202203021221001, No.202203021221021)
Foundation Items: The National Natural Science Foundation of China (No.62272286, No.41871286), The Fundamental Research Program of Shanxi Province (No.202203021221001, No.202203021221021)