面向非平衡数据流的重采样集成分类方法研究

章涂义,刘三民,陈燕菲,余文韬,朱 健
安徽工程大学 计算机与信息学院,安徽 芜湖 241000

摘 要:目的 类不平衡和概念漂移是数据流分类任务中的两个主要挑战,当它们同时发生时,将显著影响数据流分类算法的性能,因此,针对传统数据流分类算法难以应对类别不平衡和概念漂移同时存在的问题,提出一种专注于非平衡数据流的重采样集成模型。**方法** 首先,设计一种适用于数据流的边界过采样方法,利用三角形重心的特点,在边界样本内侧合成新样本,使得块中的少数类得到增强的同时,尽可能保持数据原有分布并且避免引入新的概念,有效改善数据块中类别不平衡情况;在此基础上,融合时间衰减策略和加权集成策略,设计基于马修斯相关系数作为权重的动态加权集成模型,解决概念漂移问题,同时增强分类挖掘模型的自适应性和健壮性。**结果** 在 3 个真实数据流和 6 个模拟数据流上的仿真实验结果表明:所提方法在非平衡数据流场景中,展现出对多数类和少数类均有高效的识别能力,并且对突变和增量概念漂移都具有更好的漂移感知和适应能力,分类模型整体性能优于对比算法。**结论** 实验验证:所提方法构建出一种鲁棒的非平衡数据流分类模型,在处理非平衡数据流和适应两种类型的概念漂移方面具有更好的优势。

关键词:非平衡数据流;概念漂移;集成学习;马修斯相关系数
中图分类号:TP311.13 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2025.0003.005

Research on Resampling Ensemble Classification Method for Imbalanced Data Streams

ZHANG Tuyi, LIU Sanmin, CHEN Yanfei, YU Wentao, ZHU Jian
School of Computer and Information, Anhui University of Technology, Anhui Wuhu 241000, China

Abstract: Objective Class imbalance and concept drift are two main challenges in data stream classification tasks. When they occur simultaneously, they significantly affect the performance of data stream classification algorithms. Therefore, to address the difficulty of traditional data stream classification algorithms in handling the simultaneous occurrence of class imbalance and concept drift, a resampling ensemble model focused on imbalanced data streams was proposed. **Methods** Firstly, a boundary oversampling method tailored for data streams was designed. By leveraging the characteristics of the triangular center of gravity, new samples were synthesized inside boundary samples to enhance the minority class within the block, while striving to maintain the original data distribution and avoid introducing new concepts. This effectively improved the class imbalance in the data block. On this basis, a dynamic weighted ensemble model based on Matthews correlation coefficient as weights was designed by integrating the time decay strategy and weighted ensemble strategy. This model solved the problem of concept drift and enhanced the adaptability and robustness of the classification mining model. **Results** Simulation experiments on three real data streams and six simulated data streams demonstrated that the proposed

收稿日期:2023-11-07 **修回日期:**2024-01-15 **文章编号:**1672-058X(2025)03-0034-10
基金项目:安徽省自然科学基金项目(2308085MF220);安徽省高校自然科学研究重点项目(2022AH050972,KJ2021A0516).
作者简介:章涂义(1998—),男,安徽宣城人,硕士研究生,从事数据流分类与概念漂移研究.
通信作者:刘三民(1978—),男,安徽岳西人,教授,博士,从事数据挖掘与机器学习等研究. Email:aqlsm@163.com.
引用格式:章涂义,刘三民,陈燕菲,等. 面向非平衡数据流的重采样集成分类方法研究[J]. 重庆工商大学学报(自然科学版),2025,42(3):34-43.
ZHANG Tuyi, LIU Sanmin, CHEN Yanfei, et al. Research on resampling ensemble classification method for imbalanced data streams[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2025, 42(3): 34-43.

method exhibited efficient identification capabilities for both majority and minority classes in imbalanced data stream scenarios, as well as better drift perception and adaptation capabilities for sudden and incremental concept drifts. The overall performance of the classification model was superior to the comparison algorithms. **Conclusion** The experiments verify that the proposed method constructs a robust classification model for imbalanced data streams, which has better advantages in dealing with imbalanced data streams and adapting to two types of concept drift.

Keywords: imbalanced data stream; concept drift; ensemble learning; Matthews correlation coefficient

1 引　言

随着互联网技术的发展,数据从社交媒体、网络购物、视频监控等场景源源不断产生。为从这些实时动态数据流中提取出有价值的潜在信息,数据流挖掘已经成为机器学习领域的热点问题之一[1]。在这些数据流中,数据分布会随时间发生变化,该现象称为概念漂移[2],这导致模型无法适应新的数据变化,模型性能将退化;与此同时,现实情况中数据流往往存在类别分布极不平衡的情况,如网络入侵检测、金融欺诈、故障诊断等[3-4],严重影响分类器在少数实例上的分类准确性。由此可见,类不平衡和概念漂移都会影响数据流分类模型的性能。当它们同时发生时,往往会互相影响[5],使得数据流的分类任务更为复杂。因此如何有效解决数据流中概念漂移和类不平衡问题成为数据流挖掘中的关键问题。

当前针对具有概念漂移的非平衡数据流分类问题的研究,已经取得显著进展。Sun 等[6]通过引入两阶段代价敏感学习框架,将特征选择和分类阶段的代价信息以及漂移检测机制相结合,实现对类别不平衡和概念漂移的综合处理,从而有效提升数据流分类性能;高源等[7]提出面向动态不平衡数据流的集成极限学习机算法,通过设计检测方法来跟踪数据流中不平衡比率变化,并在超限学习机基础上将增量学习与集成学习相结合,应对数据流发生数据分布的变化;陆克中等[8]开发了一种结合自适应遗忘因子的加权在线顺序极限学习机集成模型,该模型根据分类性能自适应地更新遗忘因子、投票权重以及类别权重,使模型更灵活地适应数据流变化;李艳红等[5]通过综合考虑概念漂移指数、自适应遗忘因子与不平衡比率来计算训练样本的重要性,成功将概念漂移程度纳入模型重构过程,使传统集成学习方法 AdaBoost M2 适用于非平衡数据流,为非平衡概念漂移数据流场景下的集成学习提供了模型重构的新思路;此外,董明刚等[9]探索通过计算存储小类样本的相似性来解决不平衡数据流问题,并根据熵值大小来调整基分类器,以解决概念漂移问题;Grzyb 等[10]针对概念漂移的非平衡数据流提出一种基于 Hellinger 距离的加权集成方法,该方法结合数据流分类的精度加权集成方法和 Hellinger 距离度量来确定分类器的权重,以更好地适应不平衡数据,然而该方法并没有考虑数据层的处理。综上所述,面向存在概念漂移且类不平衡的数据流分类环境时,在数据层进行预处理可以较好地解决类不平衡问题,集成学习方法在应对数据流概念漂移时也表现出优秀的适应性。

以上方法在数据处理层对非平衡数据流进行处理时,常使用静态数据预处理方法,并未充分考虑流式数据动态变化的特点。当过采样合成的新样本落在分类器决策边界时,容易产生噪声甚至新的概念。随着新数据不断到达,这种负面影响在数据流分类任务中更为突出,严重影响整体分类器的决策精度。此外,如何使用集成学习方法在非平衡数据流环境下对突变和渐变两种概念漂移进行有效感知,并迅速做出更高效的反应,仍值得深入研究。

因此,本文在已有文献的基础上,设计了一种面向非平衡数据流的重采样集成模型(Resampling Ensemble Model for Imbalanced Data Streams,REM-IDS)解决上述问题。本文工作主要包括两个部分:

(1) 设计一种适用于非平衡数据流的过采样方法。通过此方法,改善了边界样本合成方法对数据分布变化的影响,使得数据流中过采样合成的新样本尽可能不影响决策边界的变化,提升分类模型的健壮性。

(2) 提出一种新的数据流动态加权集成学习方法。该方法在基分类器训练阶段结合了数据流时间衰减的特性,使基分类器更加关注新样本的信息。通过计算基分类器在数据块中的马修斯相关系数值对其进行在线评估和权重更新,使集成模型能及时更新,提高模型对概念漂移的感知与处理能力。

2 模型设计

本文设计的重采样集成分类挖掘模型包括两个阶段:数据过采样阶段和动态加权集成阶段。整体框架如图 1 所示。

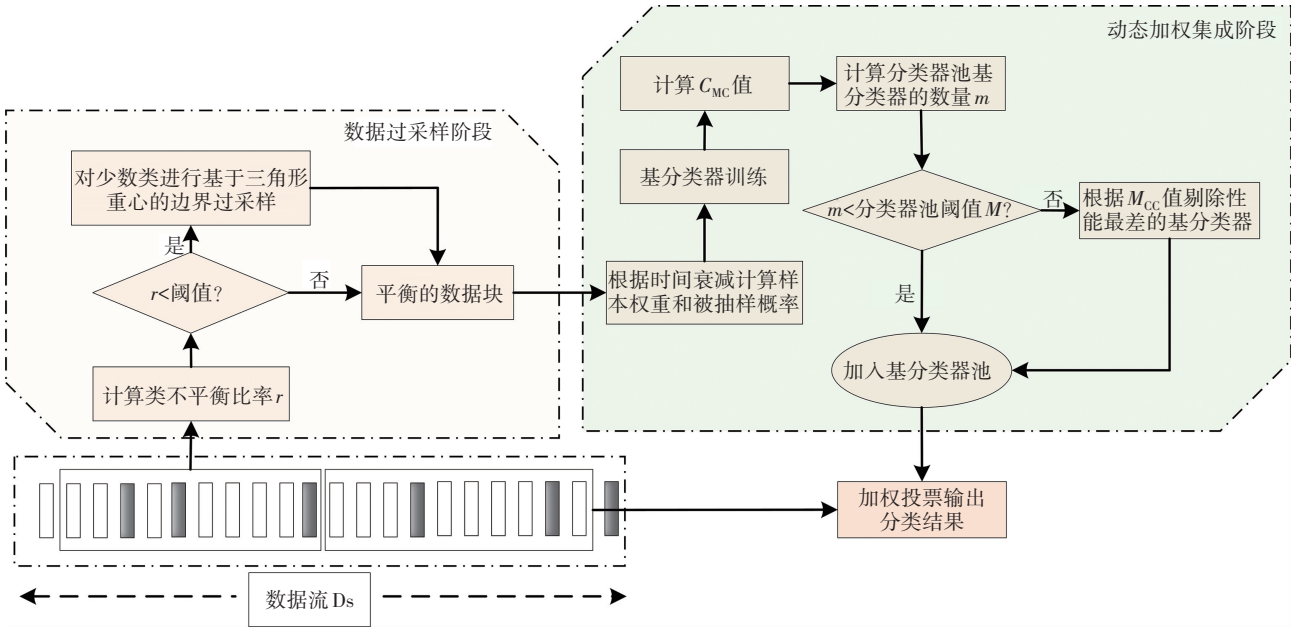


图 1 REM-IDS 框架流程图

Fig. 1 Flowchart of the REM-IDS framework

在数据过采样阶段,首先按批处理方式将数据流分成连续的数据块,每块根据式(1)计算不平衡比率 r。当不平衡比率小于设定阈值时,采用基于三角形重心的边界过采样方法对少数类进行处理,使数据块达到较为平衡的状态。

$$r=\frac{S_{\text{minority}}}{S_{\text{majority}}} \tag{1}$$

其中,S_{minority} 为少数类样本集,S_{majority} 为多数类样本集。为及时跟踪概念变化,提升分类模型的健壮性,本文设计一种动态加权集成分类策略。在每个基分类器的训练过程中,首先利用数据流样本的时间衰减特性,赋予新样本更高的抽样概率,可以使分类器更加关注新样本,更好地适应数据流变化。此外,在增量学习过程中,本文使用基于马修斯相关系数对基分类器进行加权和更新,筛选剔除性能差的分类器,激励当下性能更优秀的分类器更积极地参与投票决策,以此来提高整个集成模型在动态数据流中的自适应能力。下面对数据过采样阶段和动态加权集成阶段进行详细介绍。

2.1 数据过采样阶段

2.1.1 传统过采样方法

在处理非平衡数据流时,许多技术利用经典静态数据过采样方法进行插值过采样,如 SMOTE、Borderline-SMOTE[11] 等。Borderline-SMOTE 方法是通过考虑样本 k 近邻的类别将少数类样本分类为安全、危险和噪声 3 类,只对那些被分类为“危险”的少数类别样本进行过采样。这使得过采样过程更加关注边界样本,有助于保持数据的原始结构。然而,使用 Borderline-SMOTE 方法在边界合成新样本时,合成的样本可能会落入多数类区域,如图 2 所示。在数据流环境下,这一现象的影响更为突出,随着样本不断到达,愈发导致数据分布发生变化,继而引起概念的漂移甚至引入新概念,从而严重影响分类器的决策准确性。因此,一些经典的过采样方法在数据流分类任务中并不适用。

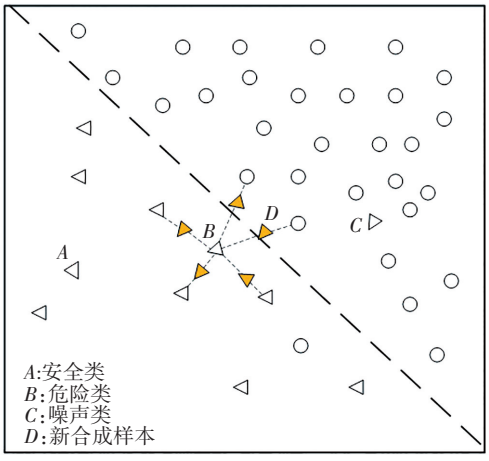


图 2 近邻个数为 5 的 Borderline-SMOTE 方法

Fig. 2 Borderline-SMOTE method with 5 near neighbors

2.1.2 基于三角形重心的边界插值过采样方法

为解决此问题,本文在 Borderline-SMOTE 方法的基础上进行了优化和改进,设计出一种基于三角形重心的边界插值过采样方法。该方法使过采样过程中合成的新样本分布更加聚集,尽可能保持原有样本分布特性的同时,对决策边界不产生影响。具体示意图如

图 3 所示。

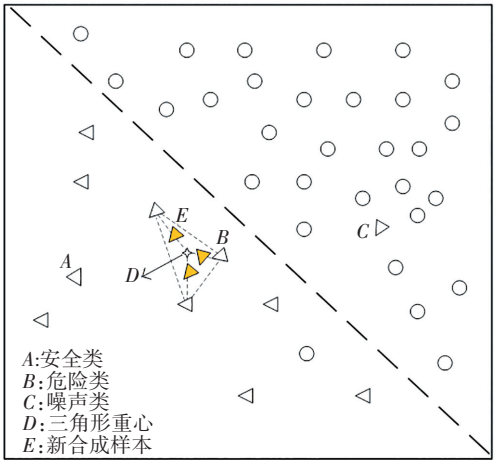


图 3　基于三角重心的边界过采样方法

Fig. 3　Boundary oversampling method based on triangular center of gravity

基于三角形重心的边界过采样方法的描述如下:

算法 1　基于三角形重心的边界过采样策略

输入:不平衡数据块 D_{Si},邻居数量 k。

输出:过采样后的平衡数据块 B_{Di};

将不平衡数据块 D_{Si} 分为少数类样本集 S_{minority} 和多数类样本集 S_{majority};

初始化合成样本集合为空;

For each 少数类样本 in 少数类样本集 do:

根据欧氏距离找到少数类样本 x_i 的 k 个最近邻样本集合;

根据少数类样本 x_i 的 k 个最近邻样本中与其不同类别的样本数量找到危险样本;

For each 危险样本 in 危险样本集合 do:

找到危险样本的 2 个最近邻样本集合 neighbors;

If 危险样本和其最近邻的 2 个样本 neighbor1, neighbor2 共线:

根据式(3)插值合成新样本;

Else:

将该危险样本和其最近邻的 2 个样本 neighbor1, neighbor2 组成三角形;

根据式(2)计算三角形的重心坐标 x_o;

连接重心与三角形顶角,根据式(3)进行插值合成新样本;

End If;

End for;

将少数类样本和合成样本集合并为过采样样本集合;

End for;

将过采样样本集合和多数类样本并为最终的数据块;

Return 平衡后的数据块 B_{Di}。

首先将少数类样本根据欧氏距离找到其 k 个最近邻样本集合,再根据少数类样本 x_i 的 k 个近邻样本中与其类别不同,将样本数量划分为 3 类,分别为安全、危险和噪声。其中,在 k 近邻样本中,若有数量超过一半的样本属于多数类样本,则将该少数类样本称为“危险类”样本,即边界样本,本方法仅对边界样本进行操作。在第 1 步—第 5 步,首先找到少数类样本中的危险类样本;第 6 步—第 15 步,对危险类样本进行具体操作:在危险样本集合中任选一个危险类样本,令为 A,找到其最近邻的 2 个少数类样本令为 B 和 C,若 A、B、C 共线,则选择在 AB 或 AC 连线上使用式(3)进行插值合成新样本;若 A、B、C 不同线,则连接 A、B、C 形成三角形,再根据式(2)找到三角形 ABC 的重心 O;再在三角形重心 O 和任意顶点的连线间根据式(3)进行插值生成新的少数类样本 x_{new};最后,将合成样本与原来的样本集合并成为最终平衡后的数据块。

三角形重心计算公式:

$$\boldsymbol{x}_O = \frac{1}{3}\left(\sum_{i=1}^{3} x_{i1}, \sum_{i=1}^{3} x_{i2}, \cdots, \sum_{i=1}^{3} x_{i3}\right) \tag{2}$$

式(2)中,$\boldsymbol{x}_O$ 表示由点 A、B、C 所构成的三角形的 n 维重心向量,x_i 中,i 取值为 A,B,C。

插值合成少数类样本 x_{new} 的计算公式如式(3):

$$x_{\text{new}} = x_i + \text{rand}(0,1) * (x_o - x_i) \tag{3}$$

2.2　动态加权集成阶段

2.2.1　基分类器训练

在数据流挖掘任务中,新近的数据往往包含丰富的信息并带来较高的价值。因此,与历史样本相比,新样本应被赋予更高的权重。为体现流式数据时间衰减特性,在本文中,假设样本的重要性随时间衰减遵循高斯衰减模型。据此设计样本采样的权重计算如下:

$$\varphi(x_i) = \frac{1}{2\pi} e^{-\frac{T(x_i)\xi}{2}} \tag{4}$$

其中,$T(x_i)$ 是一个迭代器,用于表示已存储样本集合的数量,它的初始值是已存储样本集合。通过在每次迭代时递减 $T(x_i)$,可以在计算权重时考虑已存储样本集合的索引位置。这样设计的目的是让早期添加的样本具有较大的索引值,从而在计算权重时,较早期的样本具有较低的权重。ξ 是一个可调参数,用于控制权重随时间衰减的强度,本文设置为 0.01。数据块中的样本越老,权重越小,这直接关系块中数据被抽取的机会。

$$P(x_i) = \frac{\varphi(x_i)}{\sum_{n=1}^{N} \varphi(x_i)} \tag{5}$$

块中样本被抽取的概率由 $P(x_i)$ 表示,$P(x_i)$ 计算如式(5)所示。其中,$\varphi(x_i)$ 是样本被抽取机会的权重,N 是块中包含的样本总数。因此,样本越新,样本的权重越大,被抽取的机会 $P(x_i)$ 越大。

2.2.2 加权集成更新策略

为提高集成模型在非平稳数据流中的自适应能力,本文提出通过计算马修斯相关系数作为基分类器权重的方法来更新分类器池。这种方法旨在利用集成模型的多样性,提高集成模型应对概念漂移的适应能力,保证当前时刻性能更好的基分类器在投票中占据更重要的地位,从而提升整个集成模型的性能。

马修斯相关系数 C_{MC} 是一种用于评估二分类模型性能的指标,使用表 1 混淆矩阵计算。它结合了真正例 T_P、真负例 T_N、假正例 F_P 和假负例 F_N 所有 4 个信息,因此它对二元分类问题最具信息量。C_{MC} 的取值范围在-1 到+1 之间,其中+1 表示完全正确预测,0 表示随机预测,-1 表示完全错误预测。因此 C_{MC} 值越大,则基分类器性能越好;相反地,C_{MC} 值越少,则代表基分类器越差。

表 1 混淆矩阵

Table 1 Confusion matrix

	预测结果	
真实结果	正例	反例
正例	T_P	F_N
反例	F_P	T_N

马修斯相关系数的计算如式(6):

$$C_{\mathrm{MC}}=\frac{T_P\times T_N-F_P\times F_N}{\sqrt{(T_P+F_N)\times(T_P+F_P)\times(T_N+F_P)\times(T_N+F_N)}} \tag{6}$$

第 i 个基分类器的权重 ω_i 由马修斯相关系数赋值,如式(7)所示:

$$C_{\mathrm{MC}\cdot i}\leftarrow w_i, i=1,2,\cdots,m \tag{7}$$

此外,如果计算出基分类器的 C_{MC} 值小于等于 0,则失去投票资格;否则,通过计算基分类器在实例 x_j 和类预测 y'_j 的 C_{MC} 值对基分类器进行加权多数投票。集成分类器的最终预测结果由所有基分类器的预测结果进行加权投票,加权投票的计算如式(8)所示:

$$y'_j=\operatorname{argmax}\sum_{i=1}^{m}\begin{cases}\omega_i I(C_i(x_j)=y'_j), \omega_i>0\\ 0, \text{other}\end{cases} \tag{8}$$

其中,y'_j 表示整个集成分类器的预测结果,$I(C_i(x_j)=y'_j)$ 是由基分类器 C_i 预测 x_j 为 y'_j 的标签,argmax 表示从多个基分类器的预测结果中选择具有最高概率值的类别标签。

基于马修斯相关系数动态集成学习方法的具体步骤如算法 2 所示。

首先初始化基分类器池,对于每个基分类器,初始化平均分配基分类器权重;接着通过式(6)计算每个候选基分类器在当前数据块上的马修斯相关系数,并将计算的马修斯相关系数作为基分类器的权重;再计算基分类器池中基分类器的数量 m,如果基分类器池大小小于预设最大值,则将基分类器添加入分类器池中,步骤 10—步骤 12 是将计算的马修斯相关系数进行归一化处理,并赋值给基分类器,当基分类器池大小超过预设最大值时,则移除最差的基分类器;最后根据式(8)使用加权投票方法进行投票预测,得到集成器的分类结果。通过这个过程,基于数据流的集成学习可以逐步训练和选择最佳的基分类器,获得集成器的最终分类结果,并且还可以及时对基分类器进行更新与替换,应对概念漂移的发生。具体算法如下:

算法 2 动态加权更新集成方法

输入:数据流中经过算法 1 处理后的第 i 个数据块 DS_i,集成器最大基分类器数量 M。

输出:集成器投票分类结果;

基分类器池 $\Pi\leftarrow\varnothing$;

For each D_{Si} in D_S do:

从数据块 D_{Si} 中抽取样本训练基分类 C_k;

初始化基分类器权重 $\omega(C'_k)=1/m$;

通过式(6)计算块中的 C_{MC} 值;

将 C_{MC} 值赋值给基分类器权重;

计算基分类器池中基分类器的数量 m;

If 分类器池大小 $m<M$;

将 C_k 添加到基分类器池;

For each 基分类器 C_k in Π do:

将权重值归一化后值赋予 $\omega(C'_k)$;

End for;

Else

移除最差的基分类器;

End If;

根据式(8)使用加权投票方法输出分类结果;

End for。

2.3 整体流程

由图 1 的整体框架图可知,REM-IDS 方法包括 2.1 节数据过采样阶段和 2.2 节动态加权集成方法两个部分,其完整步骤如下:

算法 3 REM-IDS

输入:数据流 D_S,不平衡阈值 θ;集成模型最大基分类器个数 M。

输出:分类结果;

基分类器池 $\prod\leftarrow\varnothing$;

For each 数据块 D_{Si} in D_S do:

将数据块 D_{Si} 分为 S_{minority} 和 S_{majority};

根据式(1)计算 DS_i 中的不平衡比率 r;

If $r<$阈值 θ

对少数类进行算法 1;

续表(算法3)

算法 3　REM-IDS
Else:
直接进入动态加权集成阶段;
End If;
使用算法 2 动态加权集成方法进行分类决策;
输出分类结果;
End for。

3　仿真实验和结果分析

为验证提出的 REM-IDS 模型的有效性,以及其对概念漂移和类不平衡数据流的适应性,本文在 3 个真实数据集和 6 个模拟数据集上对比了 6 种有代表性的数据流集成学习算法,分别为 HDWE[9]、AWE[11]、OUSE[12]、REA[13]、Learn++CDS[14] 和 SEA[15]。本文方法以及对比实验算法均在 Pycharm 环境下使用 Stream-Learn[16] 和 Scikit-Multiflow[17] 平台进行实验。实验中,不平衡阈值 θ 设置 0.45,基分类器数量最大值为 10 个,基分类器使用的是 Hellinger 距离决策树[18]。

3.1　数据集

实验数据集包括真实数据集和模拟数据集,其中真实数据集描述如下:

(1) covtypeNorm-1-2vsAll。数据集 covtype 包含有关森林覆盖类型信息,初始版本有 7 个类,本文实验合并第一类和第二类,并将它们与其他类进行比较,此时分类问题变成二元分类。数据集包含 267 000 个样本,54 个特征,不平衡比率 25%,实验设置块大小为 2 000。

(2) Weather 数据集。包含 1949—1999 年在内布拉斯加州收集的天气信息,包含 18 159 个实例。数据集一共包含 8 个相关属性,目的是预测给定日期是否下雨,不平衡比率 45%,实验设置块大小为 200。

(3) IoT_2020_b_0.01 数据集。此数据集使用正常和攻击物联网设备的物联网网络流量数据生成。包含 83 种不同的网络特征,用于网络攻击检测。在使用 k-means 聚类抽样方法后,抽取具有代表性的 IoT_2020 子集,共 6 253 个实例,不平衡比率 6%,实验设置块大小为 150。

模拟数据流由 Stream-learn 生成,生成的数据流分为两种类型:突变漂移和增量漂移。突变概念漂移意味一种状态发生迅速变化,并且分布不适合这种状态。在增量概念漂移中,数据分布的变化是逐渐而缓慢的。本次实验使用的模拟数据流样本为 100 000 个,实验选择每个数据块大小为 250 个样本。每个样本有 20 个特征,包含 15 个有效信息特征和 5 个冗余特征。DS1-DS3 是发生 3 次突变漂移情况下不同平衡比的数据流,DS4-DS6 是发生 3 次增量漂移情况下不同平衡比的数据流。

3.2　评价指标

对于不平衡数据集分类问题,传统性能评价指标如准确率等并不能很好地反映出分类模型的性能。因此本文选取不平衡分类任务中常用的 G_{mean} 值和平衡准确率 B_{AS} 作为评价指标。

几何均值 G_{mean} 指标可以较全面反映分类模型的总体性能,是衡量类不平衡数据分类性能的重要指标。只有当正、负两类样本的分类精度均处于一个较高水准时,G_{mean} 才会较大。因此该指标的数值越大,模型的总体分类效果越好。

$$G_{\text{mean}}=\sqrt{\frac{T_P}{T_P+F_N}+\frac{T_N}{T_N+F_P}}$$

平衡准确率 B_{AS} 是一个用于评估二分类模型性能的指标,适用于不平衡数据集的情况。它由对分类器在每一个类别上的准确率计算均值得到:

$$B_{\text{AS}}=\frac{R_{\text{ecall}}+S_{\text{pecificity}}}{2}$$

其中,召回率 R_{ecall} 表示正例样本被正确分类为正例的比例,特异度 $S_{\text{pecificity}}$ 表示负例样本被正确分类为负例的比例。

3.3　实验结果与分析

3.3.1　真实数据流上的对比实验

表 2、表 3 分别呈现了对比算法在真实数据集下的实验结果。图 4 至图 5 分别给出 3 个真实数据集在不同对比算法中的平衡准确率和 G_{mean} 表现曲线图。

根据表 2、表 3 的数据,可以明显看出 REM-IDS 方法在 covtypeNorm-1-2vsAll 数据集上表现出最佳性能。其平衡准确率相比第 2 名的 Learn++CDS 高出了 2.72%,而 G_{mean} 值高出 2.87%。这充分显示了本文方法的有效性。对于 Weather 数据集,尽管 HDWE 取得了最高的平衡准确率,但 REM-IDS 紧随其后,两者之间的差距微小。并且,在 G_{mean} 指标上,REM-IDS 取得了最优的性能。这表明在 Weather 数据集下,本文设计的动态权重更新方法在保持模型整体性能稳定方面发挥了重要作用。在 IoT_2020_b_0.01 数据集中,本文方法在平衡准确率和 G_{mean} 指标下均排名第 1,表明本文设计的基于三角形重心的边界过采样方法在面对数据类别极不平衡挑战时,仍能保持稳定的高性能。综合表 2、表 3 的数据,可以看出本方法整体表现出色,说明本文方法能够较好应对复杂的真实数据流,在实际应用中能够较好应对非平衡数据流的挑战。

从图 4 可以清晰观察出对比算法在 CovtypeNorm-1-2vsAll 数据集下的分类性能。在 20—40 个数据块时,即发生第一次漂移后,REA 和 OUSE 的平衡准确率下降明显且无法恢复到发生漂移前的水平。在 60—100 个数据块时,所有分类器的平衡准确率和 G_{mean} 都

开始快速降低,相较于对比的 6 种算法,REM-IDS 的平衡准确率和 G_{mean} 受影响最小。尽管 G_{mean} 轻微下降,但迅速恢复,且平衡准确率最先在第 80 个块的时候开始上升。这得益于 REM-IDS 方法的在线更新和加权机制,在发生漂移后能迅速应对,并及时更新基分类器。在发生漂移后能迅速适应新的数据分布,并及时更新基分类器,从而保持更稳定的性能。

图 5 反映对比算法在 Weather 数据集下的性能曲线图。可以清晰地观察到:与 REA、AWE 等对比算法相比,本方法在 G_{mean} 指标上表现较为平稳,且一直保持在较高水平。在这 7 个算法中,HDWE、REM-IDS、SEA 和 Learn++CDS 的 G_{mean} 表现明显优于 AWE、REA 和 OUSE。相比之下,对比算法 REA 和 AWE 在 Weather 数据集上的性能呈现较大波动。这是由于面对非平衡数据流,本方法使用基于三角重心改进的边界插值过采样方法,这种方法能够早合成新样本的同时有效控制数据分布的漂移,从而减少非平衡数据流引起的性能波动,使模型能更稳定地适应不同类别的数据。

表 2 对比算法在真实数据流上的平衡准确率

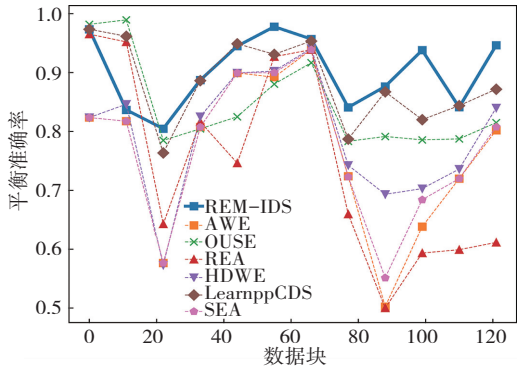
Table 2 Balance accuracy of the compared algorithms on real data streams

数据集	AWE	CDS	OUSE	SEA	REA	HDWE	**REM-IDS**
Covtype	0.769 1	0.860 1	0.832 3	0.773 6	0.717 5	0.799 5	**0.887 3**
Weather	0.633 2	0.682 0	0.580 6	0.654 6	0.609 0	**0.693 3**	0.685 8
IoT	0.966 8	0.946 0	0.937 0	0.963 2	0.989 7	0.970 1	**0.990 7**

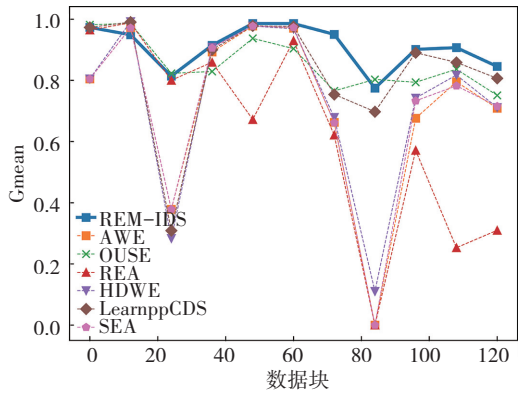
表 3 对比算法在真实数据流上的 G_{mean}

Table 3 G_{mean} of the compared algorithms on real data streams

数据集	AWE	CDS	OUSE	SEA	REA	HDWE	**REM-IDS**
Covtype	0.683 4	0.840 9	0.817 1	0.692 9	0.578 6	0.749 4	**0.869 6**
Weather	0.536 1	0.661 7	0.430 0	0.582 5	0.384 3	0.661 4	**0.662 7**
IoT	0.964 9	0.941 7	0.935 6	0.961 0	0.989 3	0.968 3	**0.990 3**



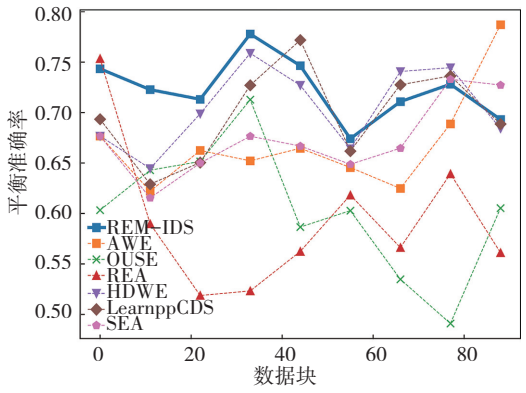
(a) 平衡准确率



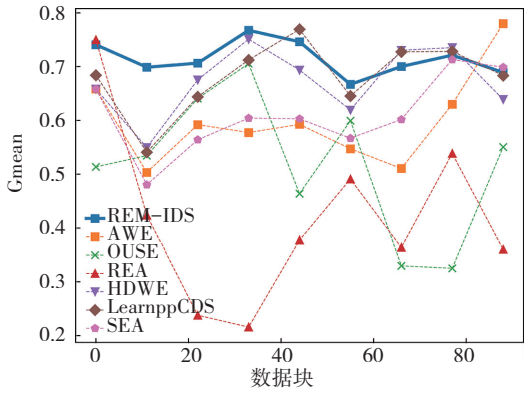
(b) G_{mean}

图 4 在 covtypeNorm-1-2vsAll 上的曲线图

Fig. 4 Curves on covtypeNorm-1-2vsAll



(a) 平衡准确率



(b) G_{mean}

图 5 在 Weather 上的曲线图

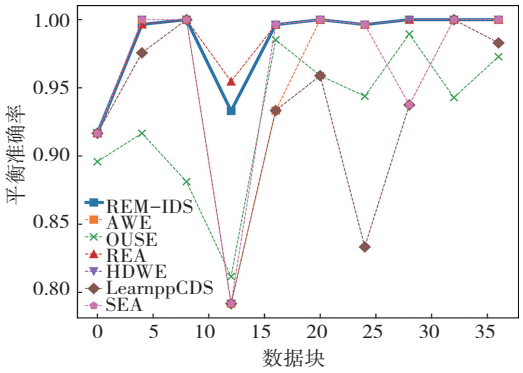
Fig. 5 Curves on Weather

图 6 呈现出对比算法在真实数据流 IoT_2020_b_0.01 下的分类性能。在物联网真实数据流下,REM-IDS 和 REA 平衡准确率的差别很小,且均高于其他分类器。再次印证本文设计的基于三角形重心的边界过采样方法,在应对极度不平衡的类别分布挑战时,表现出稳定的高性能。

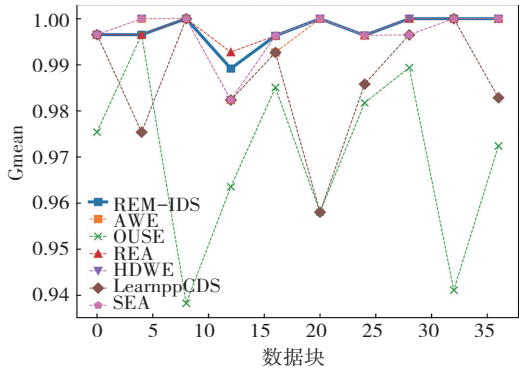
3.3.2　模拟数据流上的对比实验

表 4、表 5 呈现出 REM-IDS 与 6 种对比算法在 3 种不平衡比率(1:9,2:8,3:7)情况下,发生突变漂移和增量漂移的平衡准确率和 G_{mean} 表现。

表 4 显示,在发生突变漂移时,仅在 1:9 的不平衡比情况下,除 OUSE 算法的 G_{mean} 略优于本方法外,在其他不平衡比率下,本文提出的 REM-IDS 方法的 G_{mean} 和平衡准确率均排名第一。此外,表 5 也能反映出,在发生增量漂移时,REM-IDS 能保持最高或次高的性能表现。表 4、表 5 的模拟数据流实验结果表明:本文所提出的 REM-IDS 方法在适应增量和突变概念漂移的非平衡数据流时,都能够表现出更优秀稳定的性能;不仅能够在不同的不平衡比率下保持高水平的分类性能,还能够在面对不同类型的漂移情况时保持其优越性能,显示出 REM-IDS 方法对于不同类型数据流的鲁棒性。



(a) 平衡准确率



(b) G_{mean}

图 6　在 IoT_2020_b_0.01 上的曲线图

Fig. 6　Curves on IoT_2020_b_0.01

表 4　对比算法在突变漂移数据流上的表现

Table 4　Experimental results of the compared algorithms on sudden drift data streams

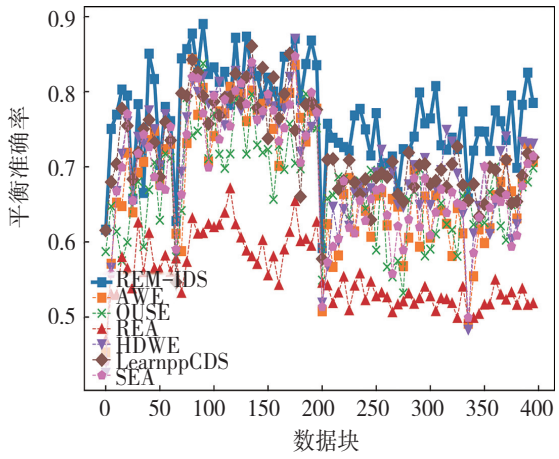
			AWE	CDS	OUSE	SEA	REA	HDWE	REM-IDS
平衡准确率	r=1:9	DS1	0.617 7	0.693 4	0.697 3	0.622 8	0.616 4	0.658 3	**0.711 7**
	r=2:8	DS2	0.747 2	0.767 8	0.705 6	0.745 1	0.609 0	0.767 7	**0.806 4**
	r=3:7	DS3	0.809 7	0.796 7	0.698 8	0.811 1	0.608 6	0.814 1	**0.849 2**
G_{mean}	r=1:9	DS1	0.453 1	0.631 8	**0.691 0**	0.468 4	0.462 2	0.549 4	0.644 8
	r=2:8	DS2	0.703 6	0.749 8	0.700 2	0.699 6	0.454 5	0.735 6	**0.787 6**
	r=3:7	DS3	0.797 9	0.791 0	0.692 9	0.799 2	0.452 2	0.804 6	**0.844 0**

表 5　对比算法在增量漂移数据流上的表现

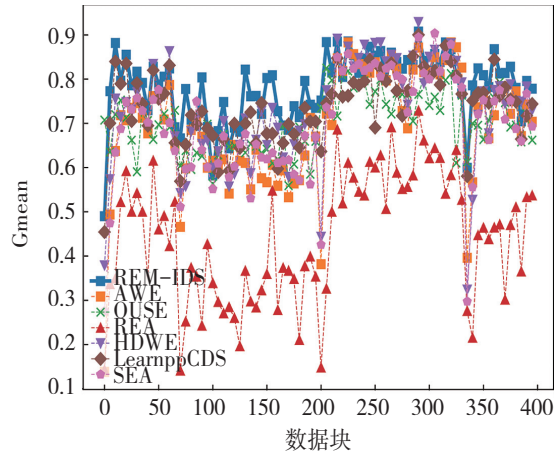
Table 5　Experimental results of the compared algorithms on incremental drift data streams

			AWE	CDS	OUSE	SEA	REA	HDWE	REM-IDS
平衡准确率	r=1:9	DS4	0.599 1	0.690 4	0.607 4	0.607 9	0.681	0.647 5	**0.691 2**
	r=2:8	DS5	0.719 2	0.701	0.737 6	0.597 2	0.754 1	0.755 4	**0.786 7**
	r=3:7	DS6	0.796 7	0.697 6	0.799 2	0.603 9	0.784 2	0.804 2	**0.830 3**
G_{mean}	r=1:9	DS4	0.406 7	**0.684 3**	0.438 5	0.442	0.614 4	0.533 9	0.615 8
	r=2:8	DS5	0.664 7	0.694 9	0.692 7	0.429 8	0.735	0.722 2	**0.765 2**
	r=3:7	DS6	0.784 1	0.691 7	0.786	0.440 8	0.778 1	0.794 7	**0.824 2**

为更为直观地表现对比算法在不同平衡比率下模拟数据流上性能表现的动态变化,本文选取在平衡比为 2∶8 的情况下,对比算法发生 3 次突变和 3 次增量漂移的性能曲线图,如图 7、图 8 所示。



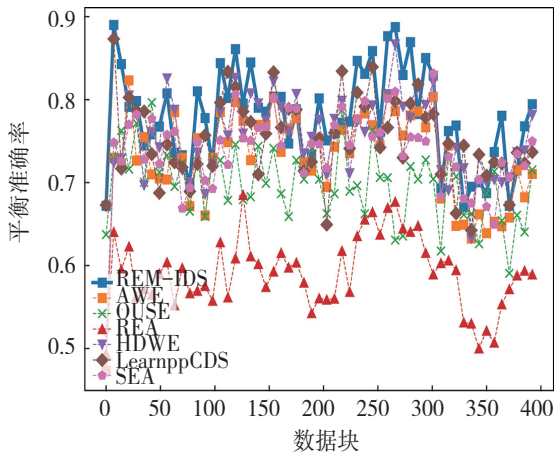
(a) 平衡准确率



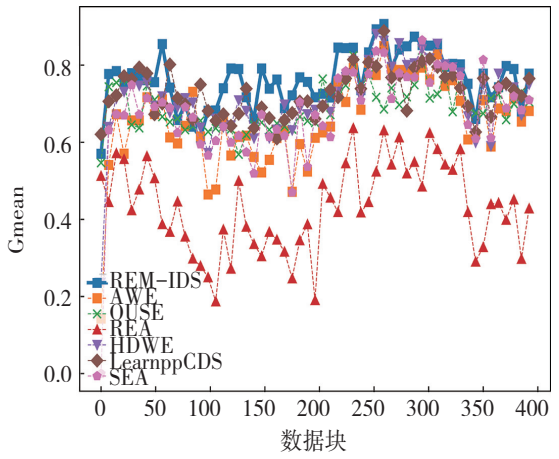
(b) G_{mean}

图 7 对比算法在突变漂移数据流上的曲线图

Fig. 7 Curves of comparison algorithms on sudden drift data streams



(a) 平衡准确率



(b) G_{mean}

图 8 对比算法在增量漂移数据流上的曲线图

Fig. 8 Curves of comparison algorithms on incremental drift data streams

图 7 表示的是在不平衡比 2∶8 的情况下,发生 3 次突变漂移时,本方法与 6 种对比算法的平衡准确率和 G_{mean} 折线图。可以反映出,在数据类别不平衡的情况下,本文方法的整体表现均为最好,尤其是在数据流发生突变漂移后,本文方法反应灵敏,调整迅速;并且,本文方法在平衡准确率曲线图上,在 200 个块后,即发生第二次突变漂移后,平衡准确率指标大幅上升,优势明显。这进一步突显了 REM-IDS 方法在应对非平衡数据流中正负样本的识别能力都很出色。

图 8 表示在不平衡比 2∶8 的情况下,发生 3 次增量漂移时,本文方法与 6 种对比算法的平衡准确率和 G_{mean} 折线图。可以清晰看出,在绝大多数情况下,本文方法都能取得最高或次高的表现,尤其是在图 8(b)中,对比算法的 G_{mean} 表现波动不定,而本文方法波动不大并且保持最高或次高的性能表现,展现出本方法的健壮性和自适应能力。

以上实验结果表明:在不同平衡比的非平衡数据流场景下,基于三角形重心的边界过采样机制都行之有效,并且在面向增量漂移和突变漂移的数据流时,本文设计的方法都能有更好更稳健的表现。

4 结 论

本文聚焦于解决类不平衡和概念漂移共存情况下的数据流在线分类问题,提出一种面向非平衡漂移数据流的集成学习模型 REM-IDS。该模型包括基于三角重心改进的边界插值过采样方法和一种新的动态加权集成方法。首先,通过基于三角形重心的边界过采样方法强化数据流中的少数类,通过此方法更好地保留

少数类的关键特征,并较好控制因过采样合成新样本而引起的数据分布变化;然后,考虑数据流样本的衰减特性,使得分类器的训练样本更加关注新出现的概念;继而,设计了基于马修斯相关系数的加权集成更新策略,在发生概念漂移时实时做出动态调整;最后,在 3 个真实数据流和 6 个模拟数据流上验证了该模型在应对各种非平衡数据流场景均表现出色。在未来工作中,还将考虑在数据流中存在多类不平衡时,如何更好地提高分类模型的性能。

参考文献(References):

[1] 崔瑞华, 綦小龙, 刘艳芳, 等. 面向概念漂移数据流的在线集成自适应算法[J]. 南京大学学报(自然科学版), 2023, 59(1): 134-144.
CUI Rui-hua, QI Xiao-long, LIU Yan-fang, et al. Online ensemble adaptive algorithm for concept drift of streaming data[J]. Journal of Nanjing University (Natural Science Edition), 2023, 59(1): 134-144.

[2] 文益民, 刘帅, 缪裕青, 等. 概念漂移数据流半监督分类综述[J]. 软件学报, 2022, 33(4): 1287-1314.
WEN Yi-min, LIU Shuai, MIAO Yu-qing, et al. Survey on semi-supervised classification of data streams with concept drifts[J]. Journal of Software, 2022, 33(4): 1287-1314.

[3] 张喜龙, 韩萌, 陈志强, 等. 动态集成选择的不平衡漂移数据流 Boosting 分类算法[J]. 山东大学学报(工学版), 2023, 53(4): 83-92.
ZHANG Xi-long, HAN Meng, CHEN Zhi-qiang, et al. Boosting classification algorithm for imbalanced drift data stream based on dynamic ensemble selection[J]. Journal of Shandong University (Engineering Science Edition), 2023, 53(4): 83-92.

[4] 张金润, 胡森荣, 洪炎. 基于 LoRa 技术的便携式健康监测系统设计[J]. 重庆工商大学学报(自然科学版), 2022, 39(1): 56-61.
ZHANG Jin-run, HU Sen-rong, HONG Yan. Design of portable health monitoring system based on LoRa technology[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2022, 39(1): 56-61.

[5] 李艳红,任霖,王素格,等. 非平衡数据流在线主动学习方法[J]. 自动化学报, 2024, 50(7): 1-13.
LI Yan-hong, REN Lin, WANG Su-ge, et al. Online active learning method for imbalanced data stream [J]. Acta Automatica Sinica, 2024, 50(7): 1-13.

[6] SUN Y, SUN Y, DAI H. Two-stage cost-sensitive learning for data streams with concept drift and class imbalance[J]. IEEE Access, 2020(8): 191942-191955.

[7] 高源, 施伟谊, 周亦华, 等. 一种面向动态不平衡数据流的集成超限学习机分类算法[J]. 复旦学报(自然科学版), 2023, 62(3): 352-361.
GAO Yuan, SHI Wei-yi, ZHOU Yi-hua, et al. An ensemble classification method of extreme learning machine for dynamic imbalanced data streams [J]. Journal of Fudan University (Natural Science Edition), 2023, 62(3): 352-361.

[8] 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 Ke-zhong, CHEN Chao-fan, CAI Huan, et al. Online classification algorithm for concept drift and class imbalance data stream[J]. Acta Electronica Sinica, 2022, 50(3): 585-597.

[9] 董明刚, 张伟, 敬超. 面向不平衡数据流的动态权重集成分类算法[J]. 小型微型计算机系统, 2020, 41(8): 1649-1655.
DONG Ming-gang, ZHANG Wei, JING Chao. Dynamic weight ensemble integration classification algorithm for imbalanced data stream based on sampling [J]. Journal of Chinese Computer Systems, 2020, 41(8): 1649-1655.

[10] GRZYB J, KLIKOWSKI J, WOŹNIAK M. Hellinger Distance Weighted Ensemble for imbalanced data stream classification[J]. Journal of Computational Science, 2021, 51: 101314.

[11] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning [C]// Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005: 878-887.

[12] WANG H, FAN W, YU P S, et al. Mining concept-drifting data streams using ensemble classifiers [C]//Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2003: 226-235.

[13] GAO J, DING B, FAN W, et al. Classifying data streams with skewed class distributions and concept drifts [J]. IEEE Internet Computing, 2008, 12(6): 37-49.

[14] CHEN S, HE H. Towards incremental learning of nonstationary imbalanced data stream: A multiple selectively recursive approach[J]. Evolving Systems, 2011, 2(1): 35-50.

[15] 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.

[16] STREET W N, KIM Y, STREET W N, et al. A streaming ensemble algorithm (SEA) for large-scale classification[C]// Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2001: 377-382.

[17] KSIENIEWICZ P, ZYBLEWSKI P. Stream-learn-open-source Python library for difficult data stream batch analysis [J]. Neurocomputing, 2022, 478: 11-21.

[18] MONTIEL J, READ J, BIFET A, et al. Scikit-multiflow: A multi-output streaming framework[J]. The Journal of Machine Learning Research, 2018, 19(1): 2905-2914.

[19] CIESLAK D A, HOENS T R, CHAWLA N V, et al. Hellinger distance decision trees are robust and skew-insensitive[J]. Data Mining and Knowledge Discovery, 2012, 24(1): 136-158.

责任编辑:李翠薇