计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.20210471
59(12):2844-2857,2022

面向概念漂移且不平衡数据流的 G-mean 加权分类方法

梁 斌 李光辉 代成龙
(江南大学人工智能与计算机学院 江苏无锡 214122)
(1634113866@qq.com)

G-mean Weighted Classification Method for Imbalanced Data Stream with Concept Drift

Liang Bin, Li Guanghui, and Dai Chenglong
(*School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, Jiangsu* 214122)

Abstract Concept drift and class imbalance in data stream seriously degrade the performance and stability of the traditional data stream classification algorithms. To solve this issue in binary classification of data stream, an online G-mean weighted ensemble classification method for imbalanced data stream with concept drift termed OGUEIL is proposed. It exploits the online update mechanism of component classifiers' weights to modify block-based ensemble algorithms, combining the hybrid resampling and adaptive sliding window algorithm. OGUEIL is based on the ensemble learning framework that once a new instance reaches, each component classifier in the ensemble and its weight are correspondingly updated online, and the minority class instance is randomly oversampled at the same time. Particularly, each component classifier determines its weight according to the G-mean performance on several recently incoming instances, where G-mean of each component classifier is calculated based on the time decay factor increment. At the same time, OGUEIL periodically constructs a balanced dataset according to the data in the current sliding window and trains a new candidate classifier, then adds it to the ensemble based on specific conditions. The experimental results on both real-world and synthesized datasets show that the comprehensive performance of the proposed method outperforms other baseline algorithms.

Key words data stream; concept drift; ensemble learning; class imbalance; classification

摘 要 数据流中的概念漂移和类别不平衡问题会严重影响数据流分类算法的性能和稳定性.针对二分类数据流中概念漂移和类别不平衡的问题,在基于数据块的集成分类方法上引入成员分类器权重的在线更新机制,结合重采样和自适应滑动窗口技术,提出了一种基于 G-mean 加权的不平衡数据流在线分类方法(online G-mean update ensemble for imbalance learning, OGUEIL).该方法基于集成学习框架,利用时间衰减因子增量计算成员分类器最近若干实例上的 G-mean 性能,并确定成员分类器权重,每到达一个新实例,在线更新所有成员分类器及其权重,并对少类实例进行随机过采样.同时,OGUEIL 会周期性地根据当前数据构造类别平衡数据集训练新的候选分类器,并选择性地添加至集成框架中.在真实和人工数据集上的结果表明,所提方法的综合性能优于其他同类方法.

收稿日期:2021-05-14;**修回日期**:2022-01-21
基金项目:国家自然科学基金项目(62072216)
This work was supported by the National Natural Science Foundation of China (62072216).
通信作者:李光辉(ghli@jiangnan.edu.cn)

关键词　数据流;概念漂移;集成学习;类别不平衡;分类

中图法分类号　TP391

信息的爆炸性增长导致数据流广泛出现在各个应用领域中,如无线传感器数据流、银行交易数据流等[1-3].数据流中的潜在分布或目标概念随着时间推移发生变化,这种现象被称为"概念漂移"[4-5].概念漂移会导致在过去数据上训练的分类模型性能显著下降,无法适应当前的新概念,这给传统的数据挖掘算法带来新的挑战.另一方面,当数据流中存在类别不平衡现象时,即某一类的实例数量显著多于其他类,数据流分类会变得更加困难,因为少类实例(minority class instance)[6-7]出现频率过低导致分类模型对它们学习不充分,而我们通常更关注少类的分类情况,因为误分类一个少类实例的代价通常远大于误分类一个多类实例(majority class instance)的代价,例如在癌症诊断中,将患癌人群诊断为健康会带来严重后果.

目前可以同时处理概念漂移和类别不平衡问题的数据流分类方法大多是基于集成学习的思想,主要包括在线集成和基于数据块的集成方法[8].在线集成方法以 Wang 等人[9]提出的 OOB(oversampling online bagging)和 UOB(undersampling online bagging)为代表,它们将过采样和欠采样技术与 Online Bagging[10]相结合,动态调整采样频率,有效解决了数据流中类别不平衡问题.在线集成方法通常还与针对不平衡数据流设计的漂移检测方法结合,例如 Wang 等人[11]提出的 DDM-OCI(drift detection method for online class imbalance learning)结合 Online Bagging,通过监测少类召回率的变化在不平衡数据流中检测漂移.但 DDM-OCI 假设数据流服从高斯分布,因而在实际应用中存在较高的误报率.为此,Wang 等人[12]又提出了 LFR(linear four rates)使用统计学检验分析中的 TPR(true positive rate),TNR(true negative rate),PPV(positive predicted value),NPV(negative predicted value)4 个指标的变化显著性来检测漂移,有效降低了 DDM-OCI 的误报率.而 Wang 等人[13]提出的 HLFR(hierarchical linear four rate)使用分层假设检测框架,在第 1 层使用 LFR 检测漂移,第 2 层使用排列检验(permutation test)验证漂移的真实性,进一步降低了 LFR 检测漂移的误报率.在所有基于数据块的集成方法中,Gao 等人[14]提出的 UB(uncorrelated bagging)是第一个解决数据流中类别不平衡的方法.UB 使用集成框架,不断累积数据流中的少类实例,然后添加到当前数据块中平衡数据分布.然而这种策略不仅需要大量的内存空间来存储累积的少类实例,而且没有考虑少类实例上可能发生概念漂移的问题,有较大的局限性.为此,Chen 等人[15]提出的 SERA(selectively recursive approach)改进了 UB,它使用马氏距离计算累积的少类实例和当前数据块中少类实例的相似度,只选择相似度较高的少类实例平衡当前数据块的类别分布.进一步,Chen 等人[16]又提出了 REA(recursive ensemble approach),该方法使用 KNN(k-nearest neighbors)计算相似度,替换 SERA 中的马氏距离度量,解决了少类实例中的子概念问题.而针对重采样过程存在的一些困难因素,例如异常数据,类别重叠等,Ren 等人[17]提出了 GRE(gradual recursive ensemble),它使用 DBSCAN 聚类技术将当前少类实例分为若干个簇,然后分别计算各个簇中实例和过去数据块中少类实例的相似度,选择部分少类实例填充至当前数据块,解决了重采样过程中数据异常和类别重叠问题.Wu 等人[18]提出的 DFGW-IS(dynamic feature group weighting with importance sampling)通过分析当前数据块和过去数据块的海林格距离差异来检测概念漂移,同时结合重要性采样处理类别不平衡问题.基于数据块的集成方法存在一个共性问题:它们都假设少类实例的概念不会发生变化,即过去数据块中的少类实例可以继续使用.然而在实际情况中,类的先验概率随时间也会发生变化,过去数据块中少类实例可能就是当前数据块中的多类实例.另外,重复访问历史数据也不符合数据流挖掘的要求.因此,以 Ditzler 等人[19]的 Lean++CDS 和 Lean++NIE 为代表,一些不需要保存历史数据的集成方法被提出.Lean++CDS 是 Learn++NSE 和 SMOTE(synthetic minority class oversampling technique)的简单结合,其中 Learn++NSE 用于处理概念漂移,而 SMOTE 产生新的少类实例以平衡当前数据块的类别分布,无需保存任何历史数据.Lean++NIE 也不需要访问历史数据,在每个数据块上对多类实例进行欠采样,结合 Bagging 技术生成一个由多个成员分类器组成的子集成模块,并根据成员分类器在过去和当前数据块上的 G-mean 性能分配权重,有效平衡每个类别的重要性.此外,Lu 等人[20]提出

的 DWMIL(dynamic weighted majority for imbalance learning)在集成模型中只保留有限数量的成员分类器,每个成员分类器的权重根据在当前数据块上的 G-mean 性能决定,并随着时间衰减,直至小于某个阈值被移除,兼顾了效率和性能.

基于上述分析,目前已有的方法主要存在 2 个问题:一是需要大量空间保存过去的少类实例进行重复使用,且没有考虑类先验概率变化的情况;二是集成方法中的成员分类器权重是基于数据块更新的,缺乏在线更新机制,面对突变型漂移或发生在数据块内的漂移时,难以快速应对.为此,针对二分类数据流,本文在基于数据块集成方法上引入了在线更新机制,提出了一种基于 G-mean 加权的在线不平衡数据流分类方法(online G-mean update ensemble for imbalance learning, OGUEIL),以集成框架为基础,每到达 1 个新实例,增量更新每个成员分类器及其权重,并对少类实例随机过采样,无须保存历史数据,同时周期性地训练多个具有差异性的候选分类器以提高集成模型的泛化能力.与同类方法相比,本文主要贡献有 3 个方面:

1) 提出了一种基于 G-mean 的在线加权策略,可以根据当前数据分布及时调整每个成员分类器的权重,有效解决不平衡数据流中的概念漂移问题.

2) 在集成模型在线更新过程中引入了对少类实例的随机过采样策略,既提高了少类实例的召回率,又增加了集成的多样性.

3) 基于混合采样和自适应滑动窗口技术提出了一种候选分类器训练策略,周期性地对当前窗口上的数据同时使用边界人工少类实例合成技术[21]和随机欠采样技术生成多个具有差异性的候选分类器,并将它们选择性地添加至当前集成模型中,提高泛化能力.

1 相关知识

1.1 数据流概述

在数据流分类领域,数据流由大量按时间顺序到达的实例组成,表示为 $S=\{s_1,s_2,\cdots,s_t,\cdots\}$,其中 $s_t=(\boldsymbol{X}_t,y_t)$ 表示时刻 t 到达的实例,$\boldsymbol{X}_t=(d_1,d_2,\cdots,d_n)$ 代表 n 维向量,意味着数据流 S 是 n 维的,$y_t\in\{c_1,c_2,\cdots,c_k\}$ 表示实例 s_t 真实类别,k 为数据流 S 中所有类别数量.

1.2 概念漂移定义和分类

概念漂移是指数据流中的目标概念随时间发生改变,在数据流分类领域,目标概念一般指当前分类模型学习到的决策边界.具体而言,假设数据流 S 服从某分布 $F_t(X,y)$,$P(y|X)$ 表示 y 关于 X 的条件概率分布,代表决策边界,若在时刻 $t+1$ 有 $F_t(X,y)\neq F_{t+1}(X,y)$ 且 $P_t(y|X)\neq P_{t+1}(y|X)$,表明原有的决策边界发生变化,这种现象称为概念漂移[8,22].

概念漂移的分类普遍是基于概念变化的速度[22-23].当新旧概念过渡很快,旧的概念突然被另一个数据分布完全不同的新概念取代,这种漂移属于突变型概念漂移(abrupt concept drift);反之,新旧概念过渡较慢时,旧概念被新概念逐渐替换,且二者在漂移前后或多或少有些相似,则属于渐变型概念漂移(gradual concept drift).

1.3 在线过采样集成算法 OOB

针对数据流中的类别不平衡问题,Wang 等人在 OB(online bagging)[10]基础上提出了在线过采样集成算法 OOB(oversampling OB)[9].OB 将传统的集成学习算法 Bagging 从静态数据领域扩展到了数据流领域.Bagging 算法首先对所有样本放回随机采样,然后得到多个训练集,最后训练多个不同的成员分类器.因此每个样本会被重复选择 k 次,且 k 服从二项分布,如式(1)所示:

$$P(K=k)=\binom{N}{k}\left(\frac{1}{N}\right)^k\left(1-\frac{1}{N}\right)^{N-k}. \tag{1}$$

N 为所有样本数量,当 N 趋于无穷时,k 近似服从参数为 1 的泊松分布.在 OB 中,每到达一个新样本,实例中每个成员分类器使用该样本训练 k 次,其中 $k\sim$Poisson(1),最终可以近似达到 Bagging 的效果.但 OB 没有考虑类分布不平衡,OOB 通过动态调整泊松分布参数提高少类样本的学习频率,进而提升少类上的召回率.具体的,在时刻 t 每到达一个新样本,每个成员分类器使用该样本训练 k 次,若该样本属于少类,则 $k\sim$Poisson($s^t_{\text{maj}}/s^t_{\text{min}}$),$s^t_{\text{maj}}$,$s^t_{\text{min}}$ 分别表示时刻 t 多类和少类的实例数量.否则,$k\sim$Poisson(1).

2 基于 G-mean 加权的在线不平衡数据流分类方法

针对二分类数据流中的概念漂移和类别不平衡问题,本文提出了一种基于 G-mean 加权的数据流分类方法(OGUEIL).OGUEIL 属于在线集成方法,其主要思想是通过使用在线决策树 Hoeffding tree[24]

和基于 G-mean 的在线加权机制,在基于数据块的集成方法中引入在线更新机制,避免数据块大小难以选择的问题,可以有效处理各种类型的概念漂移,包括突变型、渐变型以及发生在数据块内部的漂移,提高分类性能.在线更新过程中,OGUEIL 结合 OOB[9] 对少类实例进行随机过采样,既提高了少类实例的召回率,又增加了集成的多样性,且不需要保存任何历史数据.此外,OGUEIL 会周期性地添加和淘汰集成中的成员分类器以维持集成模型的分类效率和性能.OGUEIL 包含更新、淘汰、候选分类器训练、加权和决策 5 个过程,下面分别详细介绍各过程的算法思路与伪代码.

2.1　在线更新和淘汰机制

在 OGUEIL 中,每获得一个新实例(x_t,y_t),所有成员分类器更新一次.为解决数据流中类别不平衡导致少类召回率过低的问题,OGUEIL 结合 OOB[9] 算法对少类实例随机过采样,即对每个少类实例学习 k 次,且 k 服从参数为 ξ 的泊松分布,ξ 为当前数据流中多类实例与少类实例的数量比,OOB 伪代码如算法 1 所示.由于数据流的不稳定性,类的先验分布可能发生变化,甚至少类和多类发生角色互换,因此 OGUEIL 需要实时监测数据流中多类实例和少类实例的分布情况.

算法 1. OOB[9].

输入:时刻 t 到达的实例(x_t,y_t),当前集成模型 Ω,当前多类实例数量$|Y_{\text{maj}}|$,当前少类实例数量$|Y_{\text{min}}|$;

输出:更新后的集成模型 Ω.

① while 到达一个新实例

②　对于当前集成模型 Ω 中的每一个分类器 C_i:

③　计算当前数据流的不平衡率 $\xi\leftarrow|Y_{\text{maj}}|/|Y_{\text{min}}|$;

④　if 当前实例属于少类

⑤　根据式(3)设置 $k\sim\text{Poisson}(\xi)$;

⑥　else

⑦　设置 $k\sim\text{Poisson}(1)$;

⑧　end if

⑨　更新 k 次分类器 C_i;

⑩ end while

OGUEIL 使用基于时间衰减因子的方法增量计算每个类的实例数量,二分类情况下(假设为正类和负类),在时刻 t 到达一个新实例 x_t,正类 c_{p} 和负类 c_{n} 的实例数量 $s_{\text{p}}^{(t)}$,$s_{\text{n}}^{(t)}$ 通过式(2)增量计算:

$$\begin{cases}s_{\text{p}}^{(t)}=\lambda s_{\text{p}}^{(t-1)}+(1-\lambda)[(x_t,c_{\text{p}})],\\ s_{\text{n}}^{(t)}=\lambda s_{\text{n}}^{(t-1)}+(1-\lambda)[(x_t,c_{\text{n}})],\end{cases}\tag{2}$$

如果 x_t 的真实类别是正类 c_{p},那么$[(x_t,c_{\text{p}})]=1$,否则$[(x_t,c_{\text{p}})]=0$,对于负类 c_{n} 也是同理,而 λ 为预设的时间衰减因子.区别于传统累加每个类别实例的方式,这种方式使用时间衰减因子进行指数平滑,强调当前数据的影响同时弱化旧数据的影响,更适合用在数据流中.然后根据式(3)确定少类和多类,其中 δ 为预设的阈值.若满足式(3),正类 c_{p} 被标记为多类,负类 c_{n} 为少类,反之亦然.

$$s_{\text{p}}^{(t)}-s_{\text{n}}^{(t)}>\delta,(0\leqslant\delta\leqslant1).\tag{3}$$

在本文中,参数 δ 是通过大量实验获得的经验值,δ 过大或过小均会影响到算法性能.在第 3 节实验中,本文将详细介绍各个参数的设置.为保证集成分类的效率和准确率,OGUEIL 使用淘汰机制优化集成结构:每当创建一个新候选分类器时,若集成模型的成员数量没有达到预设的最大值 m,直接添加成员,否则替换权重最小的成员,这样保证了集成模型的成员不会随时间无限增加,降低内存消耗.

2.2　候选分类器训练

如何训练泛化能力强的候选分类器是克服多类别不平衡、提高少类分类准确率的关键.普遍的解决方案是对多类实例欠采样或对少类实例过采样,这 2 种方法都有各自的优点和缺陷.本文结合过采样和欠采样,提出了一种基于混合采样的候选分类器训练方法(candidate classifier training, CCT),如算法 2 所示.OGUEIL 每隔固定周期检测当前窗口中各类实例的数量是否均超过预设值 β,若满足则开始训练 $T(T>1)$个新候选分类器.首先确定当前窗口中所有类实例数量的最大值(max)和最小值(min),然后在 min 和 max 之间随机取值 N 作为之后每类实例的重采样数量.对于实例数量少于 N 的类,OGUEIL 使用边界人工合成少类样本方法(BorderlineSMOTE)[21] 将其数量过采样至 N,值得注意的是,它属于过采样方法的一种,通过在决策边界附近人工合成少类样本来平衡数据分布,既增强了决策边界,又降低了过拟合的风险;而对于实例数量大于 N 的类,通过随机欠采样(RUS)将其数量削减,最终使用类分布相对平衡的数据集训练候选分类器.由于 OGUEIL 每次生成不止一个候选分类器,且每次的采样数量都是随机选取,因此可以最大限度减少有价值的信息的丢失.同时,由于训练每个候选分类器的数据集都不同,OGUEIL 会得到一组具有足够多样性的候选分类器,可以增强整体集成

分类器的泛化能力.生成 T 个候选分类器后,此时如果当前集成规模 $|\Omega|$ 与 T 之和小于预设的集成最大成员数 m,直接添加成员分类器,否则移除集成中权重最小的成员分类器,直至满足 $|\Omega|+T<m$,新的候选分类器的权重均初始化为当前所有成员分类器权重的最大值.

算法 2. CCT.

输入:当前窗口 W 中的数据 D;

输出:新的候选分类器.

① 确定 D 中所有类实例数量的最大值 max 和最小值 min;

② 在[min,max]内随机取值 N 作为之后每个类实例的重采样数量;

③ 对实例数量少于 N 的类使用 Borderline-SMOTE 过采样至 N;

④ 对实例数量大于 N 的类使用 RUS 欠采样至 N;

⑤ 使用处理后的数据集 D 训练一个新的候选分类器.

2.3 加权和决策机制

数据流集成分类方法的加权机制大都是基于数据块的,即每到达一个数据块,集成中每个成员分类器的权重由在当前数据块上的分类精度决定.当面对突变型漂移或发生在数据块内的漂移时,基于数据块的加权机制难以快速调整成员分类器的权重.此外,基于分类精度的加权机制容易受到类分布的影响,导致成员分类器偏向多类,忽略少类.为此本文提出了一种基于 G-mean 的在线加权机制,它的特点是每到达一个新实例而不是一个完整的数据块,所有成员分类器的权重更新一次且不受类分布的影响.更新成员分类器时既考虑该分类器创建的时间,又考虑它在最近 d 个数据上的 G-mean 性能.二分类数据流中,G-mean 就是正类 c_{p} 上的准确率 PR 和负类 c_{n} 上的准确率 NR 的几何平均值,如式(4)所示:

$$Gmean=\sqrt{PR\times NR}. \tag{4}$$

G-mean 对数据的类分布不敏感,可以平衡多类和少类的重要性,在平衡和不平衡数据流中都可以很好地反映一个分类器的性能.根据文献[25],PR 和 NR 可以通过时间衰减因子增量计算,增量计算的时间复杂度为 $O(1)$,在时刻 t 到达一个新实例(x_t,y_t),第 j 个成员分类器 T 的 PR_j^t 计算式如式(5)所示:

$$PR_j^t=\lambda PR_j^{t-1}+(1-\lambda)\times\{y_t'=y_t\},\quad j=1,2,\cdots,m. \tag{5}$$

其中,λ 是时间衰减因子,y_t 为时刻 t 的真实标签,y_t' 为时刻 t 的预测标签,使用伯努利随机变量 I_t 表示$\{y_t'=y_t\}$的结果,若 y_t' 与 y_t 相同,$I_t=1$,否则 $I_t=0$.根据文献[12],式(5)可以表示为伯努利随机变量的几何加权和,如式(6)所示:

$$PR_j^t=(1-\lambda)\sum_{i=1}^{t}\lambda^{(t-i)}I_i,j=1,2,\cdots,m. \tag{6}$$

同理,第 j 个成员分类器 NR_j^t 的计算式如式(7)所示:

$$NR_j^t=(1-\lambda)\sum_{i=1}^{t}\lambda^{(t-i)}I_i,j=1,2,\cdots,m. \tag{7}$$

在时刻 t,每个成员分类器的权重通过式(8)~(11)更新:

$$PR_j^t=\begin{cases}\lambda PR_j^{t-1}+(1-\lambda)I_t, & 1<t-\tau\leqslant d,\\ \lambda PR_j^{t-1}+(1-\lambda)I_t-(1-\lambda)\lambda^{(t-1)}I_{t-d}, & t-\tau>d,\end{cases}\quad j=1,2,\cdots,m. \tag{8}$$

$$NR_j^t=\begin{cases}\lambda NR_j^{t-1}+(1-\lambda)I_t, & 1<t-\tau\leqslant d,\\ \lambda NR_j^{t-1}+(1-\lambda)I_t-(1-\lambda)\lambda^{(t-1)}I_{t-d}, & t-\tau>d,\end{cases}\quad j=1,2,\cdots,m. \tag{9}$$

$$Gmean_j^t=\sqrt{PR_j^t\times NR_j^t},j=1,2,\cdots,m. \tag{10}$$

$$w_j^t=\frac{1}{1-Gmean_j^t+\varepsilon},j=1,2,\cdots,m. \tag{11}$$

其中,τ 代表成员分类器的创建时刻,w_j^t 代表时刻 t 的成员分类器权重,ε 为一个极小的正实数,防止式(11)的分母为 0.一个新的成员分类器在时刻 τ 被创建,权重使用当前所有分类器权重的最大值初始化.当 $1<t-\tau\leqslant d$ 时,成员分类器的权重根据在当前 $t-\tau$ 个数据上的 G-mean 值增量计算;当 $t-\tau>d$ 时,成员分类器的权重计算只考虑在最近 d 个数据上的 G-mean,其中 d 代表预设的检测周期.OGUEIL 根据加权多数投票原则对每个输入实例预测,在时刻 t,集成模型 Ω 根据每个成员分类器 C_i 的权重 w_i 和预测结果 $C_i(x_t)$ 对实例 x_t 预测,结果为 y_t',如式(12)所示:

$$y_t'=\mathrm{sgn}\Big(\sum_{i=1}^{m}w_iC_i(x_t)\Big). \tag{12}$$

其中,sgn(·)为符号函数,若括号中结果大于 0,返回 1,代表正类 c_{p};否则返回 -1,代表负类 c_{n}.OGUEIL 的伪代码如算法 3 所示:

算法 3. OGUEIL.

输入:数据流 S、检测周期 d、集成模型容量 m、成员分类器 C_i、少类实例数量最小值 β、滑动窗口 W、候选分类器个数 T;

输出:加权集成模型 Ω.

① while 每到达一个新实例(x_t,y_t)

② 　根据式(12)得到 x_t 的预测结果 y_t';

③ 　根据式(2)增量计算每个类的实例大小;

④ 　根据式(3)确定当前数据流中的少类和多类;

⑤ 　把新实例(x_t,y_t)添加到窗口 W 中;

⑥ 　根据式(8)～(11),使用(x_t,y_t)更新集成中每个成员分类器 C_i 的权重;

⑦ 　每隔 d 个实例:

⑧ 　　if 窗口 W 中的少类实例数量大于 β:

⑨ 　　　调用 CCT 算法 T 次,训练 T 个新的候选分类器;

⑩ 　　end if

⑪ 　　if 当前集成中的分类器数量小于 $m-T$

⑫ 　　　直接将这 T 个分类器添加至集成中;

⑬ 　　else

⑭ 　　　使用替换集成中权重最小的分类器;

⑮ 　　end if

⑯ 　　清空窗口 W 中数据;

⑰ 　根据 OOB 算法,使用新实例(x_t,y_t)对每个成员分类器训练 k 次;

⑱ end while

2.4　计算复杂度分析

OGUEIL 集成模型使用 Hoeffding tree 做基分类器,Hoeffding tree 学习每个实例的时间复杂度为 $O(1)$,故含有 m 个 Hoeffding tree 的集成模型学习时间复杂度为 $O(m)$.OOB 使每个 Hoeffding tree 训练 k 次,k 服从泊松分布,OGUEIL 的时间复杂度变为 $O(km)$.每个类的数量计算均通过增量计算,所以时间复杂度为 $O(1)$.CCT 算法创建 T 个候选分类器的时间复杂度为 $O(2TN)$,N 代表采样数量,而每个基分类器通过式(8)～(11)加权需要 $O(1)$ 时间,对 m 个基分类器加权的时间复杂度为 $O(m)$.综上,OGUEIL 的时间复杂度为 $O(km+2TN+m)$,由于 k,m,N,T 均与输入数据流的规模无关,故 OGUEIL 关于数据流规模的时间复杂度可解析为 $O(1)$.

关于方法的空间复杂度,由于 OGUEIL 使用滑动窗口处理数据,创建分类器时需存储 N 个样本数据.因此方法的空间复杂度为 $O(TN)$,这里 T 为候选分类器个数.显然,滑动窗口大小、采样数量和分类器个数均与输入数据流的规模无关,故关于输入数据流规模的空间复杂度仍可视为 $O(1)$.

3　实验结果及其分析

为验证 OGUEIL 方法的性能,本节将 OGUEIL 和其他 5 种同类方法在人工和真实数据集上进行实验比较.对比方法可分为 2 类:一类是基于数据块的集成方法:DWMIL[20],Learn++NIE[19](后面简称为 LPN)和 REA[16];另一类是在线集成方法:OAUE[26],OOB[9].实验环境:1 台处理器为 Intel Core i7-7700HQ,内存为 16 GB 的笔记本电脑,运行 Windows 10 系统和 python3.7.在该环境下,分别实现了本文方法和对比方法,对比方法的参数设置均各参照对应文献.OGUEIL 的参数设置为:成员分类器使用 python 的 scikit-multiflow 包[27]的 Hoeffding tree 使用默认设置;时间衰减因子 λ 和类别不平衡检测阈值 δ 设置参照文献[9],分别设为 0.9 和 0;p 根据大量实验确定,设为 500;集成最大成员数量 $m=15$;创建候选分类器所需的最小少类实例数量 $\beta=15$;$\varepsilon=0.000\,000\,1$.

3.1　性能评价指标

本文利用以下指标对方法进行评价,包括二分类数据流中的分类准确率 ACC(accuracy)、几何均值 $Gmean$(geometry mean)、少类召回率 MCR(minority class recall),其具体定义如式(13)(14)所示.

$$ACC=\frac{tp+tn}{tp+tn+fp+fn}. \tag{13}$$

$$PR=\frac{tp}{tp+fn},$$

$$NR=\frac{tn}{tn+fp},$$

$$Gmean=\sqrt{PR\times NR}. \tag{14}$$

如果正类是少类,即正类实例数量小于负类数量,少类召回率 $MCR=\sqrt{\dfrac{tp}{tp+fn}}$;否则,$MCR=\sqrt{\dfrac{tn}{tn+fp}}$,其中 tp,tn,fp,fn 的定义如表 1 所示:

Table 1 Confusion Matrix

表 1 混淆矩阵

预测类别	真实类别	
	正类	负类
正类	tp	fp
负类	fn	tn

3.2 数据集介绍

实验共用到 6 个人工数据集和 2 个真实数据集,详情如下:

Sine 数据集[4].该数据集生成器有 2 个属性 x 和 y.分类函数是 $y=\sin(x)$,在第 1 次漂移之前,函数曲线下方的实例被标记为正类,曲线上方的实例被标记为负类,共有 2 个类别.在漂移点,通过反转分类规则来产生漂移.Sine 共包含 100 000 个实例,每隔 20 000 个实例产生 1 次漂移,类分布平衡,含 10%噪声.

Sea 数据集[28].该数据集生成器有 3 个属性,其中第 3 个属性与类别无关,如果 $x_1+x_2<\alpha$,实例分类为正,否则为负,x_1,x_2 表示前 2 个属性.通过欠采样生成 2 个新的数据集:1)Sea_{ac}通过欠采样产生类别不平衡,不平衡率(指少类实例所占百分比)初始化为 0.05,在数据流中某处会突然上升至 0.95,即多类实例变为少类实例;2)Sea_{nc}通过欠采样产生类别不平衡,不平衡率固定为 0.05.

Circle 数据集[4].该数据集生成器有 2 个属性 x 和 y.4 个不同圆方程表示 4 个不同概念.圆内的实例被分类为正,圆外为负,共 2 个类别.在漂移点通过更换圆的方程来产生漂移.Circle 数据集共包含 50 000 个实例,每隔 12 500 个实例产生 1 次漂移,类分布平衡,含 10%噪声.

Hyper Plane 数据集[28].该数据集生成器有 10 个属性,通过连续旋转决策超平面产生漂移.Hyper Plane_{nc}包含 50 000 个实例,不平衡率固定为 0.05.

Gaussian 数据集[28].该数据集生成器有 2 个属性,通过改变高斯成分的均值和方差产生漂移.本实验中通过欠采样产生类别不平衡数据集 Gaussian_{gc},不平衡率初始化为 0.05,然后逐渐上升至 0.95.

Electricity 数据集[4].该数据集为真实数据集,收集了澳大利亚新南威尔士州电力市场的 45 312 个电价数据,包含 8 个属性和 2 个类别.

Weather 数据集[20].该数据集为真实数据集,包含贝尔维尤和内布拉斯加州 50 多年来的天气信息.任务是预测一天是否下雨.本实验中通过欠采样实现类别不平衡[20],不平衡固定为 0.05,包含 18 159 个实例,有 8 个属性和 2 个类别.

表 2 总结了所有数据集的信息.实验用到的 8 个数据集进一步可分为四大类,模拟 4 种不同场景:1)概念漂移的类平衡数据集,包括 Sine,Circle,Electricity;2)有概念漂移的类别不平衡数据集且包含不平衡率突然变化的情况,包括 Sea_{ac};3)有概念漂移的类别不平衡数据集且包含不平衡率逐渐变化的情况,包括 Gaussian_{gc};4)有概念漂移的类别不平衡数据集且不平衡率固定不变,包括 Sea_{nc},Hyper Plane_{nc}.

Table 2 Description of Datasets

表 2 数据集的描述

数据集	数据量	属性数	噪声率	漂移类型	不平衡率
Sine	100 000	2	0.1	Abrupt	0.49
Circle	50 000	2	0	Gradual	0.50
Electricity	45 312	8			0.41
Sea_{ac}	48 780	3	0	Gradual	0.05 变化至 0.95
Gaussian_{gc}	51 893	2	0	Gradual	0.05 变化至 0.95
Sea_{nc}	100 000	3	0	Abrupt	0.05
Hyper Plane_{nc}	50 000	10	0	Gradual	0.05
Weather	18 159	7	0		0.05

3.3 参数实验

本节用 OGUEIL 的参数 p(基分类器更新周期)的不同值对算法 G-mean 性能进行了实验,结果如表 3 所示.

由表 3 中数据可知,参数 p 的不同取值对 OGUEIL 的 G-mean 性能影响较小,同时 $p=500$ 时在 8 个数据集上的平均排名最高,所以最终 OGUEIL 的参数 p 设置为 500.

Table 3　G-Mean Results of OGUEIL Under Different p Values

表 3　不同 p 值下的 OGUEIL 的 G-mean 结果

数据集	p=500	p=750	p=1000	p=1250	p=1500
Sine	**0.862 0(1)**	0.859 2(2)	0.848 9(3)	0.842 6(4)	0.840 1(5)
Circle	**0.973 6(1)**	0.972 9(2)	0.970 6(3)	0.969 9(4)	0.968 6(5)
Electricity	**0.906 0(1)**	0.902 7(3)	0.905 0(2)	0.895 5(4)	0.892 3(5)
Weather	**0.664 2(1)**	0.643 9(4)	0.658 2(2)	0.658 8(3)	0.642 2(5)
Sea_{ac}	0.832 0(2)	0.815 0(5)	**0.848 7(1)**	0.821 0(4)	0.827 3(3)
$Gaussian_{gc}$	0.929 2(3)	0.931 3(2)	0.923 9(4)	**0.948 5(1)**	0.924 3(5)
Sea_{nc}	0.960 9(2)	0.959 3(5)	**0.961 5(1)**	0.960 3(4)	0.960 8(3)
$HyperPlane_{nc}$	**0.865 2(1)**	0.856 4(3)	0.858 1(2)	0.855 1(4)	0.855 4(5)
平均排名	**1.5**	3.2	2.2	3.5	4.5

注:加粗项表示最优值,括号中数字代表横向排名,数字越小越好.

3.4　实验结果分析

本节比较了 OGUEIL 和其他 5 种方法在上述 8 个数据集上的分类准确率,G-mean 和少类召回率,结果如表 4～6 所示.表 4 给出了所有方法的 8 个数据集上的准确率结果.根据表 4 可以看出:其一,Sine,Circle,Electricity 3 个数据集的类分布相对平衡,准确率可以较好地反映每种方法的性能,OGUEIL 在这 3 个数据集上准确率均排在第 1,表明 OGUEIL 可以很好地处理各种类型概念漂移,紧接着是 OAUE 和 DWMIL,二者结果相近;其二,在其余类分布不平衡数据集上,OAUE 均排名第 1,但这不能表明 OAUE 处理类别不平衡数据流中概念漂移的能力强于其他方法,因为数据流的类分布严重不平衡时,准确率会偏向于多类,意味着一个方法只有把所有实例预测为多类就可以获得很高的准确率,严重忽略少类实例,不能合理地反映方法性能.表 5 给出了各方法 G-mean 的实验结果,G-mean 对类分布不敏感,在平衡或不平衡数据流中都可很好地反映一个方法的性能.结果显示:OGUEIL 在 7 个数据集上平均排名最高,DWMIL 次之,而 OAUE 的 G-mean 性能很差,在 Weather 上甚至为 0,但它的准确率很高,这表明它的多类性能很好而少类性能很差,主要因为它没有处理类别不平衡的机制,容易将少类实例误分类为多类实例.REA 是针对不平衡数据流的方法,但它的 G-mean 性能很差,甚至弱于 OAUE,主要因为它保存过去所有数据块中的少类实例,然后通过 KNN(k-nearest neighbors)选择部分少类实例平衡当前数据块的类分布,这种机制很容易受到概念漂移的影响,当少类上的概念发生漂移时,少类实例会和多类实例大量重叠,严重影响方法 G-mean 性能.少类召回率的结果如表 6 所示,OGUEIL 和 DWMIL 的平均排名并列第 1,特别地,在 Sine,Circle,Electricity 这 3 个类分布相对平衡的数据集上,OGUEIL 的少类召回率高于 DWMIL,

Table 4　Accuracy Results of All Datasets

表 4　所有数据集上的准确率结果

数据集	OGUEIL	OAUE	DWMIL	OOB	LPN	REA
Sine	**0.862 0(1)**	0.834 1(3)	0.845 6(2)	0.592 5(5)	0.718 9(4)	0.569 4(6)
Circle	**0.973 6(1)**	0.958 8(2)	0.926 2(3)	0.900 8(5)	0.906 8(4)	0.693 1(6)
Electricity	**0.906 0(1)**	0.835 8(2)	0.760 9(5)	0.768 6(3)	0.764 9(4)	0.734 7(6)
Weather	0.680 7(6)	**0.952 2(1)**	0.686 5(5)	0.935 9(2)	0.863 5(4)	0.923 1(3)
Sea_{ac}	0.914 2(3)	**0.984 4(1)**	0.772 6(5)	0.847 1(4)	0.916 8(2)	0.663 9(6)
$Gaussian_{gc}$	0.975 5(2)	**0.997 1(1)**	0.889 6(4)	0.928 5(3)	0.566 7(5)	0.057 0(6)
Sea_{nc}	0.980 9(3)	**0.986 2(1)**	0.946 5(6)	0.968 4(4)	0.961 9(5)	0.984 3(2)
Hyper $Plane_{nc}$	0.921 7(2)	**0.971 9(1)**	0.831 5(5)	0.898 4(3)	0.883 4(4)	0.802 0(6)
平均排名	2.3	**1.5**	4.3	3.6	4	5.1

注:加粗项表示最优值,括号中数字代表排名,数字越小越好.

Table 5 G-Mean Results of All Datasets

表 5 所有数据集上 G-mean 结果

数据集	OGUEIL	OAUE	DWMIL	OOB	LPN	REA
Sine	**0.862 0(1)**	0.834 1(3)	0.845 5(2)	0.596 5(6)	0.717 4(4)	0.569 3(5)
Circle	**0.973 6(1)**	0.958 6(2)	0.925 2(3)	0.900 7(5)	0.905 4(4)	0.667 0(6)
Electricity	**0.906 2(1)**	0.823 1(2)	0.750 3(4)	0.759 1(3)	0.744 7(5)	0.649 3(6)
Weather	0.664 2(2)	0.000 0(6)	**0.711 3(1)**	0.311 5(5)	0.574 0(3)	0.415 1(4)
Sea_{ac}	**0.832 0(1)**	0.400 8(6)	0.807 6(3)	0.811 4(2)	0.764 8(4)	0.757 5(5)
$Gaussian_{gc}$	**0.929 2(1)**	0.834 6(4)	0.909 7(2)	0.842 8(3)	0.727 6(5)	0.210 5(6)
Sea_{nc}	**0.960 9(1)**	0.846 1(6)	0.952 4(2)	0.928 0(4)	0.936 2(3)	0.898 3(5)
Hyper $Plane_{nc}$	**0.865 2(1)**	0.713 0(4)	0.823 2(2)	0.670 1(5)	0.796 5(3)	0.611 2(6)
平均排名	**1.1**	4.1	2.3	4.1	3.8	5.3

注:加粗项表示最优值,括号中数字代表排名,数字越小越好.

Table 6 Minority Class Recall Results of All Datasets

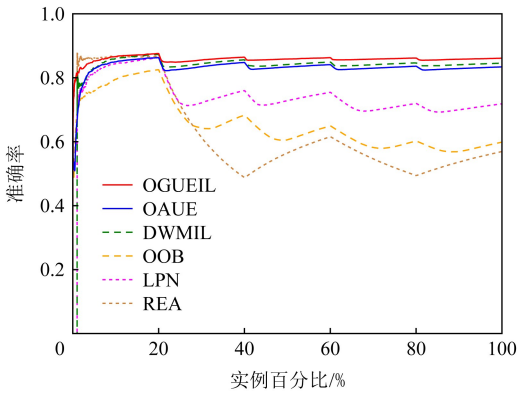
表 6 所有数据集上少类召回率结果

数据集	OGUEIL	OAUE	DWMIL	OOB	LPN	REA
Sine	**0.861 4(1)**	0.833 7(3)	0.844 0(2)	0.610 8(5)	0.732 1(4)	0.580 5(6)
Circle	**0.984 0(1)**	0.979 0(2)	0.885 4(4)	0.919 8(3)	0.858 0(5)	0.504 9(6)
Electricity	**0.907 2(1)**	0.759 3(2)	0.696 7(4)	0.706 0(3)	0.654 9(5)	0.445 3(6)
Weather	0.646 4(2)	0.000 0(6)	**0.740 0(1)**	0.099 2(5)	0.371 1(3)	0.179 4(4)
Sea_{ac}	0.756 7(4)	0.161 3(6)	0.845 5(2)	0.773 0(3)	0.635 4(5)	**0.867 7(1)**
$Gaussian_{gc}$	0.884 8(3)	0.696 9(6)	0.934 6(2)	0.778 8(5)	**0.963 6(1)**	0.878 8(4)
Sea_{nc}	0.949 4(2)	0.716 1(6)	**0.958 8(1)**	0.889 5(4)	0.910 5(3)	0.813 2(5)
Hyper $Plane_{nc}$	0.806 9(2)	0.510 5(4)	**0.834 3(1)**	0.488 4(5)	0.711 0(3)	0.455 5(6)
平均排名	2	4.3	**2.1**	4.1	3.6	4.7

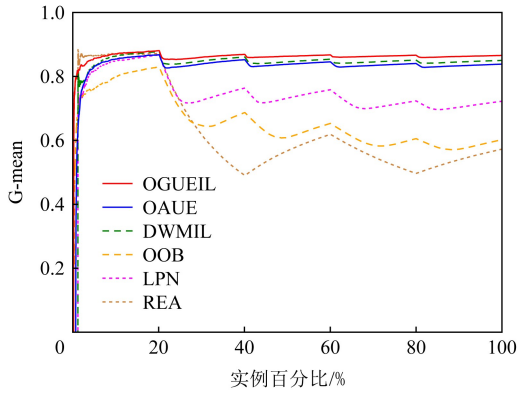
注:加粗项表示最优值,括号中数字代表排名,数字越小越好.

而在剩下的类分布不平衡数据集上 OGUEIL 的少类召回率低于 DWMIL.结合表 4,5,OGUEIL 在准确率和 G-mean 上的表现均优于 DWMIL,表明 OGUEIL 在维持少类性能的同时没有过多牺牲多类的性能,在 2 个类上的性能达到了最佳平衡.

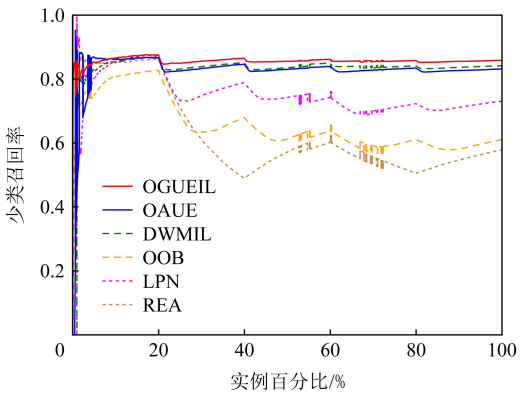
图 1 为 Sine 数据集上的结果,该数据集为类分布平衡数据集,可以发现各方法在准确率、G-mean 和少类召回率上的性能变化曲线基本一致,以图 1(a)的准确率结果为例,可以得到以下观测结果:1)OGUEIL 的准确率最高,OAUE 和 DWMIL 次之,REA 的准确率最低,表明 OGUEIL,OAUE 和 DWMIL 的抵抗概念漂移能力较强.2)Sine 数据集每隔全部数据的 20%,通过反转分类规则产生一次突变型概念漂移,OGUEIL,OAUE,DWMIL 受影响较小,准确率轻微下降后迅速恢复,其中 OGUEIL 得益于它的在线更新和在线加权机制,发生漂移后迅速更新所有成员分类器及其权重值,最先完成新概念的学习,准确率曲线率先上升.3)LPN,OOB,REA 受概念漂移影响严重,尤其是 REA,准确率甚至下降至 0.5 左右,这主要因为 REA 所有成员分类器无法增量更新,集成模型缺少成员分类器的淘汰机制,遭遇概念漂移时,在旧概念上训练的大量成员分类器既不能增量更新,也不被淘汰,从而严重影响性能.LPN 和 REA 类似,所有成员分类器也无法增量更新,集成模型也没有淘汰机制,但它有独特的加权机制,LPN 中每个成员分类根据分类性能调整权重时,会使用 sigmoid 函数对它在当前数据块上的性能和过去所有数据块上的性能加权,可以快速地消除旧概念对当前集成模型的影响,同时若发现某个成员分类器的性能弱于随机分类器,该成员分类器的权重置则被设置为 0,消除它对最终决策的负面影响,故它处理概念漂移的能力强于 REA.OOB 没有加权机制和成员分类器淘汰机制,但它的成员分类器是在线分类器,遭遇概念漂移时通过在线更新缓慢适应新的



(a) 准确率



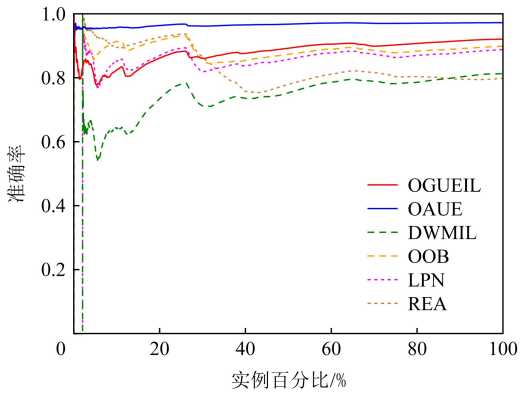
(b) G-mean



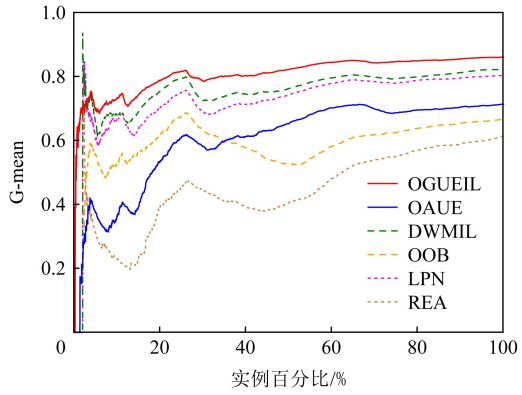
(c) 少类召回率

Fig. 1　Experimental results on the Sine dataset

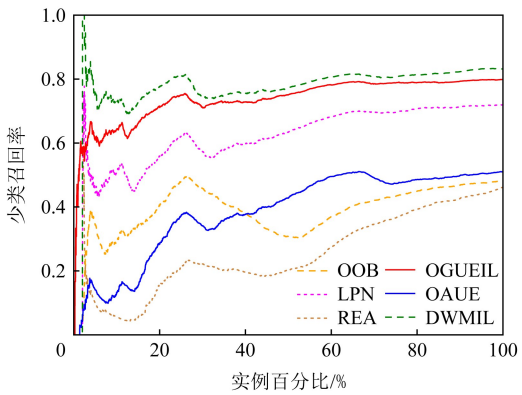
图 1　Sine 数据集上的实验结果



(a) 准确率



(b) G-mean



(c) 少类召回率

Fig. 2　Experimental results on the Hyper Plane_{nc} dataset

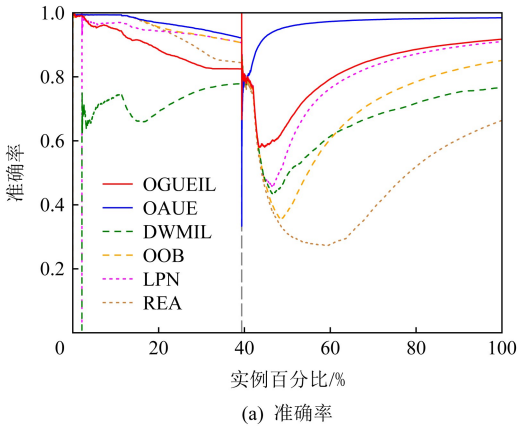
图 2　Hyper Plane_{nc} 数据集上的实验结果

概念,整体效果略好于 REA.

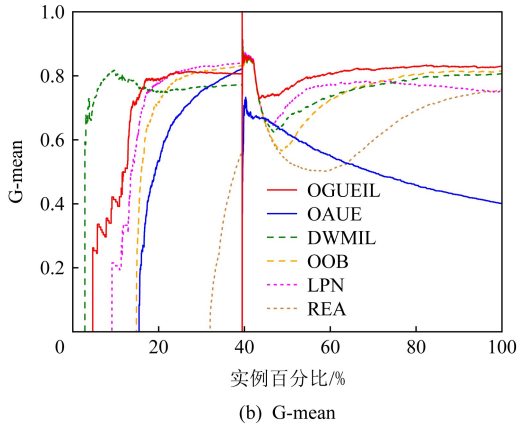
图 2 为 Hyper Plane_{nc} 上的结果,该数据集是类分布不平衡的且不平衡率固定为 5%,包含渐变型概念漂移.OAUE 的准确率始终保持较高水平,但这以严重牺牲少类上的性能为前提,它的少类召回率远低于其他方法.OGUEIL 在 3 个评价指标上的性能曲线都没有较大波动,始终保持着较高的水平,表现出较强的抗概念漂移能力.并且它在准确率和少类召回率上排名第 2,在 G-mean 上排名第 1,这表明 OGUEIL 很好地平衡了在每个类上的性能.DWMIL 在少类召回率上性能很好,排名第 1,但它准确率排在第 5,这表明 DWMIL 以大幅牺牲多类上的性能为代价提高它在少类上的性能,处理类分布不平衡的策略有些激进.LPN 的 G-mean 曲线和 DWMIL 的 G-mean 曲线十分接近,但它的少类召回率低于 DWMIL 少类召回率而准确率高于 LPN 的准确率,表明 LPN 处理类分布不平衡的策略较

DWMIL 保守一些,没有为了提高少类上的性能而过多牺牲多类上的性能.OOB 和 REA 在少类召回率上保持稳定,都高于 OAUE 的少类召回率,但在准确率和 G-mean 上低于 OAUE 的且存在较大的波动,主要因为它们对概念漂移的响应较慢,影响了在多类上的性能.

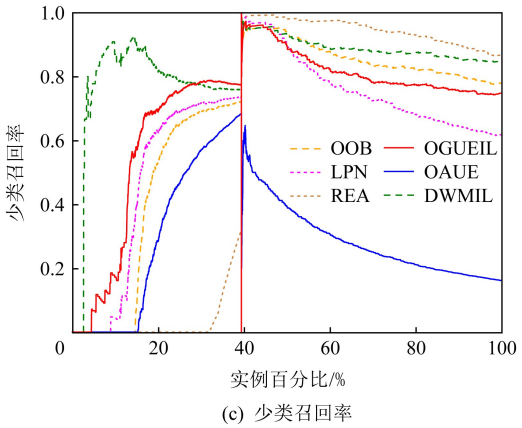
图 3 给出了各方法在 Sea_{ac} 上的实验结果,该数据集是包含渐变漂移的类分布不平衡数据集,而且不平衡率会在数据集中发生突变,导致多类实例和少类实例的角色互换,因此在不平衡率变化处重置所有评价指标,如图 3 中虚线所示(数据流 40%的位置,虚线和实线部分重合).OGUEIL 在各项性能指标上一直比较稳定,不平衡率突变后,它会根据当前数据流中类分布快速识别出多类实例和少类实例,然后调整集成中的所有成员分类器过采样的目标,性能恢复最快,最终在准确率上排名第 2,G-mean 上排名第 1,少类召回率上排名第 4.而 LPN,DWMIL,OOB,REA 的恢复速度依次递减.至于 OAUE,它在准确率上受不平衡率突变影响最小,始终保持较高水平,原因与 Hyper $Plane_{nc}$ 数据集上的



(a) 准确率



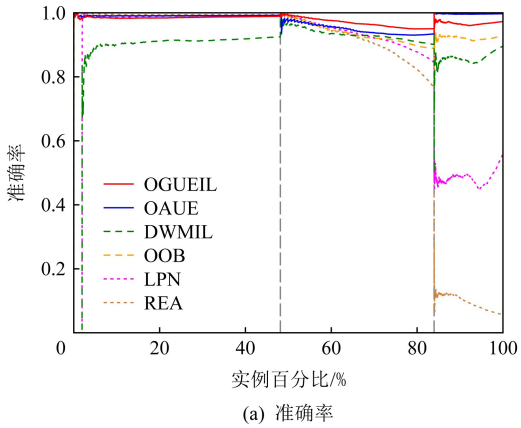
(b) G-mean



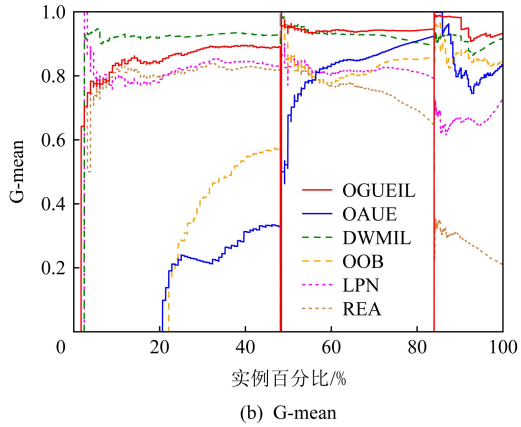
(c) 少类召回率

Fig. 3 Experimental results on the Sea_{ac} dataset

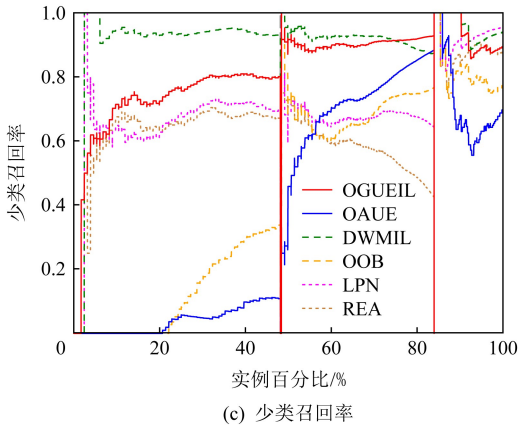
图 3 Sea_{ac} 数据集上的实验结果



(a) 准确率



(b) G-mean



(c) 少类召回率

Fig. 4 Experimental results on the $Gaussian_{gc}$ dataset

图 4 $Gaussian_{gc}$ 数据集上的实验结果

相同,但它在 G-mean 和少类召回率上波动很大,在不平衡率突变前,随着少类实例的增加,各项性能逐渐上升,突变后,由于缺乏处理类分布不平衡机制,一直处在下降状态.值得注意的是,REA 的少类召回率在不平衡率突变前很低,突变后,少类召回率大幅上升,甚至最后排名第 1,这是因为突变前 REA 将大量多类实例预测为少类实例,突变后,多类实例变为少类实例,从而获得了较高的少类召回率.

图 4 为 Gaussian_{gc} 上的结果,该数据集是包含渐变漂移的类分布不平衡数据集,而且不平衡率会在数据集中逐渐发生变化,数据流的状态由不平衡逐渐变到平衡然后又到不平衡,因此在 2 次不平衡率变化处重置所有评价指标,如图 4 中虚线所示.在第 1 次不平衡率变化后,数据流状态由类分布严重不平衡逐步过渡到平衡状态,除 REA 外所有方法的性能都保持稳定或上升状态.第 2 次变化后,数据流又从平衡状态转变至不平衡状态,多类实例变为少类实例,而少类实例变为多类实例,由于类分布的变化,所有方法的性能都有所下跌,然后随数据流增加逐渐上升.OGUEIL 的准确率基本保持稳定,最终排名第 2,在 G-mean 和少类召回率上在短暂下降后迅速恢复,最终排名分别为第 1 和第 3,整体上在 3 个性能指标上没有出现较大波动,始终保持较高水平,表明 OGUEIL 有效地降低了概念漂移和类分布变化对集成性能的影响.除了少类召回率,REA 和 LPN 在准确率和 G-mean 上均显著低于没有处理类分布不平衡机制的 OAUE 的准确率和 G-mean,可能的原因是该数据集上的概念漂移严重影响了 REA 和 LPN 在多类实例上的准确率.

3.5　运行时间比较

8 个数据集上,所有方法的运行时间如表 7 所示.平均运行时间最短的是 OOB,主要因为 OOB 的方法结构简单,它没有加权机制,没有成员分类器的添加和淘汰机制,也无需保存任何历史数据,只需维护集成模型在线更新和对少类实例的过采样.OGUEIL 在所有数据集上的运行时间都慢于 OAUE,二者的加权和集成模型成员的创建、添加和淘汰操作的耗时相近,主要差别在于 OGUEIL 整合了 OOB,集成模型的在线更新比 OAUE 增加了少类实例的过采样.REA 集成模型的成员分类器为静态批处理方法,无法在线更新,减少了时间消耗,但是它的集成模型没有淘汰机制,会保留所有成员分类器,同时它需要从历史数据块中寻找 k 个最近邻用以平衡当前数据块的类分布,这些机制导致了 REA 在小规模数据集上的运行效率较高,例如 Electricity,Gaussian_{ac} 等,而大规模的数据集上效率较低,例如 Sine,Circle 等,因为数据量越大,REA 创建的成员分类器越多,搜索 k 个最近邻的耗时也越高.DWMIL 和 LPN 的运行时间明显高于其他方法,主要因为二者都使用由若干静态批处理分类器组成的集成分类器作为集成模型的成员分类器,不过 DWMIL 的集成模型有剪枝策略,LPN 没有,这就意味着 LPN 的规模会随数据流无限扩大,导致决策时间消耗越来越大.此外,LPN 在权重计算阶段,不仅要考虑每个成员分类器在当前数据块上的性能,还要考虑它在之前每个数据块上的性能,这也会严重增加时间消耗.

Table 7　Comparison of Running Time

表 7　运行时间对比　s

数据集	OGUEIL	OAUE	DWMIL	OOB	LPN	REA
数据集	OGUEIL	OAUE	DWMIL	OOB	LPN	REA
Sine	433.8(3)	221.9(2)	779.0(5)	**107.1(1)**	1 456(6)	604.9(4)
Circle	179.3(3)	103.9(2)	433.3(5)	**60.4(1)**	435.5(6)	184.7(4)
Electricity	360.8(5)	185.6(3)	243.5(4)	**128.1(1)**	421.3(6)	155.8(2)
Weather	33.8(4)	23.9(3)	35.2(5)	**19.9(2)**	35.3(6)	5.5(1)
Sea_{ac}	174.7(4)	78.7(3)	293.6(5)	**47.94(1)**	325.2(6)	66.5(2)
Gaussian_{gc}	108.7(4)	78.3(2)	703.9(6)	**43.62(1)**	392.2(5)	80.7(3)
Sea_{nc}	377.7(3)	308.7(2)	1 891(5)	**157.6(1)**	1 927(6)	398.5(4)
Hyper Plane_{nc}	523.1(6)	209.3(3)	313.8(4)	**128.6(2)**	427.9(5)	92.7(1)
平均排名	4	2.5	4.8	**1.2**	5.7	2.6

注:加粗项表示最优值,括号中数字代表排名,数字越小越好.

4 结束语

本文针对数据流中存在概念漂移和类别不平衡的问题,提出了一种新的不平衡数据流分类方法OGUEIL,它基于集成学习框架,综合基于数据块的方法和在线方法的优点,可以有效处理不平衡数据流中的概念漂移.OGUEIL 是基于完全增量的方法,无需保存任何历史数据,使用在线分类器作为成员分类器,每到达一个实例,对集成模型中的所有成员在线更新的同时根据每个成员在最近若干数据上的 G-mean 性能加权,性能越好的成员获得权重值也越大.每隔固定周期,OGUEIL 检查当前是否满足创建新候选分类器条件,若满足就通过混合采样创建多个具有差异性的候选分类器,然后选择性地添加至集成中,并使用 2 种淘汰机制控制集成模型的规模,保持决策的高效性和准确性.

本文利用 6 个人工数据集和 2 个真实数据集模拟了 4 种不同场景,对 OGUEIL 与 5 种主流的同类方法进行了全面的对比实验.结果表明,OGUEIL 在少类数据上保持良好性能的同时没有牺牲在多类数据上的性能,在平衡与不平衡数据流下都可以有效处理概念漂移,综合性能优于其它方法,具有较强的鲁棒性.

作者贡献声明:梁斌提出了算法思路和实验方案,完成实验并撰写论文;李光辉和代成龙提出了指导意见并修改论文.

参 考 文 献

[1] Sun Yange. Research on concept drift data stream classification algorithm [D]. Beijing: Beijing Jiaotong University, 2019 (in Chinese)
(孙艳歌. 概念漂移数据流分类算法研究[D]. 北京: 北京交通大学, 2019)

[2] Guo Husheng, Ren Qiaoyan, Wang Wenjian. Concept drift category detection based on time series window [J]. Journal of Computer Research and Development, 2022, 59(1): 127-143
(郭虎升, 任巧燕, 王文剑. 基于时序窗口的概念漂移类别检测[J]. 计算机研究与发展, 2022, 59(1): 127-143)

[3] Guo Husheng, Zhang Aijuan, Wang Wenjian. Concept drift detection method based on online performance test [J]. Journal of Software, 2020, 31(4): 932-947 (in Chinese)
(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[J]. 软件学报, 2020, 31(4): 932-947)

[4] Pesaranghader A, Viktor H L. Fast hoeffding drift detection method for evolving data streams [C] //Proc of the 21st Joint European Conf on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2016: 96-111

[5] Lu Jie, Liu Anjin, Fan Dong, et al. Learning under concept drift: A review [J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(12): 2346-2363

[6] Ren Siqi. Research on data stream ensemble classification algorithm based on concept drift [D]. Changsha: Hunan University, 2018 (in Chinese)
(任思琪. 基于概念漂移的数据流集成分类算法研究[D]. 长沙: 湖南大学, 2018)

[7] Li Zeng, Huang Wenchao, Yan Xiong, et al. Incremental learning imbalanced data streams with concept drift: The dynamic updated ensemble algorithm [J]. Knowledge-Based Systems, 2020, 195(4): 105-120

[8] Wang Shuo, Minku L L, Yao Xin. A systematic study of online class imbalance learning with concept drift [J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(10): 4802-4821

[9] Wang Shuo, Minku L L, Yao Xin. Resampling-based ensemble methods for online class imbalance learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(5): 1356-1368

[10] Oza N C. Online bagging and boosting [C] //Proc of the 8th IEEE Int Conf on Systems, Man and Cybernetics. Piscataway, NJ: IEEE, 2005: 2340-2345

[11] Wang Shuo, Minku L L, Yao Xin. A learning framework for online class imbalance learning [C] //Proc of the 13th IEEE Symp on Computational Intelligence and Ensemble Learning. Piscataway, NJ: IEEE, 2013: 36-45

[12] Wang Hua, Abraham Z. Concept drift detection for streaming data [C] //Proc of the 21st Int Joint Conf on Neural Network. Piscataway, NJ: IEEE, 2015: 1-9

[13] Wang Shuo, Minku L L, Yao Xin. Dealing with multiple classes in online class imbalance learning [C] //Proc of the 25th Int Joint Conf on Artificial Intelligence. Palo Alto, CA, AAAI, 2016: 2118-2124

[14] Gao Jing, Fan Wei, Han Jing, et al. A general framework for mining concept-drifting data streams with skewed distributions [C] //Proc of the 17th Int Conf on Data Mining. Philadelphia: SIAM, 2007: 3-14

[15] Chen Sheng, He Haibo. Sera: Selectively recursive approach towards nonstationary imbalanced stream data mining [C] //Proc of the 18th Int Joint Conf on Neural Network. Piscataway, NJ: IEEE, 2009: 2053-2060

[16] Chen Sheng, He Haibo. Towards incremental learning of nonstationary imbalanced data stream: A multiple selectively recursive approach [J]. Evolving Systems, 2010, 2(1): 35-50

[17] Ren Siqi, Liao Bo, Zhu Wen, et al. The gradual resampling ensemble for mining imbalanced data streams with concept drift [J]. Neurocomputing, 2018, 286(12): 150-166

[18] Wu Ke, Edwards A, Fan Wei, et al. Classifying imbalanced data streams via dynamic feature group weighting with importance sampling [C] //Proc of the 24th Int Conf on Data Mining. Philadelphia, PA: SIAM, 2014: 722-730

[19] 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

[20] Lu Yang, Cheung Yiuming, Tang Yuanyan. Dynamic weighted majority for incremental learning of imbalanced data streams with concept drift [C] //Proc of the 26th Int Joint Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2017: 2393-2399

[21] Han Hui, Wang Wenyuan, Mao Binghuan. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning [C] //Proc of the 9th Int Conf on Intelligent Computing. Berlin: Springer, 2005: 878-887

[22] Lu Jie, Liu Anjin, Fan Dong, et al. Learning under concept drift: A review [J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(12): 2346-236

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

[24] Domingos P, Hulten G. Mining high-speed data streams [C] //Proc of the 6th Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2000: 71-80

[25] Wang Shuo, Minku L L, Ghezzi D, et al. Concept drift detection for online class imbalance learning [C] //Proc of the 22nd Int Joint Conf on Neural Network. New York: ACM, 2013: 1-10

[26] Brzezinski D, Stefanowski J. Combining block-based and online methods in learning ensembles from concept drifting data streams [J]. Information Sciences, 2014, 265(7): 50-67

[27] 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): 2915-2914

[28] Lu Yang, Cheung Yiuming, Tang Yuan. 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

Liang Bin, born in 1996. Master. His main research interests include concept drift detection, data stream mining, and ensemble learning.

梁　斌,1996 年生.硕士.主要研究方向为概念漂移检测、数据流挖掘和集成学习.

Li Guanghui, born in 1970. PhD, professor and PhD supervisor. Senior member of CCF. His main research interests include wireless sensor network, data stream mining and intelligent nondestructive detection technology.

李光辉,1970 年生.博士,教授,博士生导师,CCF 高级会员.主要研究方向为无线传感网、数据流挖掘和智能无损检测技术.

Dai Chenglong, born in 1992. Lecturer. His main research interests include electroencephalogram processing, electroencephalogram analyzing, and data mining.

代成龙,1992 年生.讲师.主要研究方向为脑电图处理、脑电图分析和数据挖掘.