计算机研究与发展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(4): 987−997, 2026

多类型概念漂移引导的在线分时适应性集成

郭虎升[1,2] 张 旭[1] 张羽桐[1] 王文剑[3]
[1](山西大学计算机与信息技术学院 太原 030006)
[2](计算智能与中文信息处理教育部重点实验室(山西大学) 太原 030006)
[3](数据智能与认知计算山西省重点实验室(山西大学) 太原 030006)
(guohusheng@sxu.edu.cn)

Online Time-Sharing Adaptive Ensemble Guided by Multi-Type Concept Drift

Guo Husheng[1,2], Zhang Xu[1], Zhang Yutong[1], and Wang Wenjian[3]
[1](*School of Computer and Information Technology*, *Shanxi University*, *Taiyuan* 030006)
[2](*Key Laboratory of Computational Intelligence and Chinese Information Processing* (*Shanxi University*), *Ministry of Education*, *Taiyuan* 030006)
[3](*Key Laboratory of Data Intelligence and Cognitive Computing of Shanxi Province* (*Shanxi University*), *Taiyuan* 030006)

Abstract Concept drift is an important feature of streaming data in the real world, and it is also an inevitable problem in data mining. In the multi-class concept drift adaptation problem, the training speed is often slow, so the overall adaptation performance is good but the drift recovery speed is slow. To address this issue, online time-sharing adaptive ensemble guided by multi-type concept drift (OTAE) method calculates the time offset distance between different data blocks, extracts the distance offset sequence, and identifies different drift types according to the sequence characteristics. According to different types of concept drift, combined with the regret boundary of exponential gradient descent model, the model weight is initialized asynchronously to achieve continuous asynchronous weighting of the model. Then, combined with the data classification features, the distance between inside and outside the sample class is calculated to extract the mixed density of samples, generate the density weight matrix, and realize the short-term weight control of the model. Finally, the long-term weight is combined with the short-term weight matrix to realize the two-stage weighted integration of the model. The experimental results show that the proposed method can significantly improve the adaptability of the model to different drift types and obtain good prediction performance.

Key words concept drift; integrated learning; drift type; two-stage weighted ensemble; range-offset sequence

摘 要 概念漂移是流数据在现实世界的一个重要特性,也是数据挖掘中不可避免的难题。在多类别概念漂移适应问题中,由于其训练速度较慢,故存在整体性能较好但收敛速度较慢的问题。为此,提出了一种多类型概念漂移引导的在线分时适应性集成(online time-sharing adaptive ensemble guided by multi-type concept drift, OTAE)方法,该方法通过计算不同数据块间的时间偏移距离提取距离偏移序列,根据序列特征识别不同漂移类型;针对不同类型概念漂移,结合指数梯度下降模型遗憾边界,动态异步初始化模型权重,实现模型的持续异步权重更新;之后,结合数据分类特征计算样本类内外距离,借此提取样本混合

收稿日期: 2025-06-06; **修回日期**: 2025-10-16
基金项目: 国家自然科学基金项目(62276157, U21A20513, 62476157, 61503229);山西省重点研发计划项目(202202020101003)
This work was supported by the National Natural Science Foundation of China (62276157, U21A20513, 62476157, 61503229) and the Key Research and Development Program of Shanxi Province (202202020101003).
通信作者: 王文剑(wjwang@sxu.edu.cn)

密度,生成密度权重矩阵,实现模型的短时权重调控;最后,将长期权重矩阵与短时权重矩阵结合,实现模型的双阶段加权集成。实验结果表明,该方法提高了模型对不同漂移类型的适应速度,取得了良好的预测性能。

关键词 概念漂移;集成学习;漂移类型;双阶段加权集成;距离偏移序列

中图法分类号 TP18

DOI: 10.7544/issn1000-1239.202550447 **CSTR:** 32373.14.issn1000-1239.202550447

大数据时代,动态数据不断涌现于各种领域,如:网购数据、气象数据、智能城市数据、舆情数据以及病毒传播数据等。相较于静态数据,这些数据源源不断产生,具有高速、实时、多变以及不可预知等特点,称之为流数据[1-2]。随着时间不断积累,数据分布往往会随着收集环境的动态变化而不断变化[3]。流数据挖掘的目的是使学习模型能够更准确地预测数据分布变化,但流数据的复杂特性给流数据挖掘任务带来了极大挑战[4]。

概念漂移(concept drift)是流数据在现实世界中的一个重要特征,其典型特点是随着时间推移或数据分布发生变化,样本的输入特征和输出标签之间的关系也发生改变的现象。然而,流数据分布变化方式的多样性易导致不同类型的概念漂移[5-6],如突变型、渐进型、瞬时型、增量型与重现型等,不同类型概念漂移的数据分布的变化时间、变化快慢、变化方式等都不相同。不同类型概念漂移的发生会导致模型性能发生不同程度的下降,因此处理流数据中的多类别概念漂移成为一个亟待解决的难题[7-9]。

目前,在线学习模型被广泛用于流数据的预测工作中,在整体预测性能上,此类模型均可取得不错的效果。然而,受其更新特性影响,当数据分布发生剧烈变化时,其模型适应速度会变得非常缓慢,这一现象被称为“慢开关现象”,该现象在多类别概念漂移中更为严重且频繁[10]。因此,针对含概念漂移数据流中的“慢开关现象”进行适应性建模十分必要。

为解决上述问题,本文提出了一种多类型概念漂移引导的在线分时适应性集成(online time-sharing adaptive ensemble guided by multi-type concept drift, OTAE)方法。该方法利用不同数据块间的时间偏移距离,提取距离偏移序列特征来识别概念漂移类型;其次,结合模型遗憾边界,动态异步初始化模型权重,实现持续异步模型加权;随后,提取样本混合密度,生成密度权重矩阵,实现短时权重调控;最后,结合长期权重与短时权重矩阵,实现模型的双阶段加权集成,提高模型对不同漂移类型的适应速度与预测性能。

本文主要贡献3个方面:

1)通过提取数据间的时间偏移距离生成距离偏移序列,可识别不同类型概念漂移;

2)提出持续异步权重更新与短时窗口密度驱动的权重调控,结合不同加权特征实现模型双阶段在线加权集成;

3)针对模型漂移恢复速度慢的问题,提供了稳定且高效的模型构建方法。

1 相关工作

目前,处理概念漂移的研究主要采用基于在线学习的集成和基于块的集成2种典型策略[11-12]。

在线集成方法以实时方式处理数据,按照时间顺序持续训练模型,使其能够动态适应概念漂移。OnlineBagging[13]采用滑动窗口进行数据采样,通过权重分配机制调整不同学习器的权重,以适应概念漂移;Promptboosting[14]在AdaBoost[15]的基础上,提出无梯度训练方法实时优化集成学习器;GOOWE-ML(geometrically optimum online weighted ensemble for multi-label classification)[16]利用空间建模为诸多增量多标签模型的分类器分配最佳权重,以提高模型预测性能;IWEB(incremental weighted ensemble broad learning system)[17]利用密度加权机制生成权重矩阵,并将该矩阵与增量集成方法结合,提高了模型的稳定性;CDD[18]实时跟踪分类器错误率变化的同时结合KL散度监测数据变化,当发生概念漂移时对模型进行再训练。DDG-DA(data distribution generation for predictable concept drift adaptation)[19]将历史数据重采样与漂移数据建模结合来更新学习器训练样本,提高了模型对概念漂移的有效性。在线集成方法能够实时处理新到达的数据,及时更新模型,但由于此类方法需要逐个处理数据,增加了计算资源,易导致学习效率较低。

基于块的集成方法将新到达的数据划分为固定

大小或可变大小的数据块，并据此更新分类器，同时动态替换表现较差的学习器。EnHAT(ensemble combined with Hoeffding adaptive tree)[20]将霍夫丁自适应树算法与基于数据块生成的决策树集成相结合，以在更短时间内适应概念漂移；ACDWM(adaptive chunk-based dynamic weighted majority)[21]为每个数据块构建单独的分类器，根据它们的实时性能对其动态加权，并通过假设检验自适应选择数据块的大小；DiwE[22]利用区域分布变化设置实例权重，再结合不同的集成方案，构建多样的集成模型；SEOA(selective ensemble-based online adaptive deep neural network)[23]以神经网络不同层次的自适应密度单元作为基分类器，并根据其损失动态加权；Learn++.NSE[24]根据各分类器训练时间的一致性对基分类器进行加权，并采用暂时遗忘机制处理过时的分类器，有效适应了不同类型的概念漂移。基于不同数据块训练的学习器可以学习到数据的不同特征，从而增加学习器间的多样性，但可能会影响模型训练速度和性能。

目前，多数集成方法讨论如何提高概念漂移环境下模型的泛化性能，但缺乏对发生概念漂移后如何提高模型适应速度的研究，当数据流中包含多种类型的概念漂移时，传统方法可能存在不同类型漂移适应速度存在较大差异的问题，导致模型缺乏普适性。

现有多数集成方法普遍基于统一性能评价指标(如预测误差、分类损失)对基学习器进行加权更新或替换，忽略了概念漂移类型所隐含的差异化动态特征，在应对多种类型概念漂移时易出现响应滞后的问题。与此不同，本文提出了 OTAE 方法。该方法通过提取数据块间的距离偏移序列，实现不同概念漂移类型的有效识别；利用模型遗憾边界与数据分类特征，提出持续异步权重更新和短时窗口密度驱动的权重调控策略，设计了双阶段加权集成方法，有效增强集成模型在概念漂移数据流中的响应能力。该方法以漂移类型为驱动信号，通过双阶段加权集成实现对多类型漂移的差异化适应，提升了模型在非稳态环境下的响应速度与稳定性。

2 OTAE 方法介绍

本文提出了 OTAE 方法，模型的整体框架如图 1 所示。主要包括基于时序偏移的漂移类型识别和双阶段加权集成 2 部分。在基于时序偏移的漂移类型识别阶段，通过计算不同数据块间的时间偏移距离提取距离偏移序列，根据序列特征识别不同漂移类型。在双阶段加权集成阶段，利用指数梯度下降模型遗憾边界，计算权重初始化步长、动态异步初始化模型权重，实现权重持续异步更新；之后，利用样本类内与类外距离计算样本混合密度，生成密度权重矩阵，实现密度驱动权重调控；最后，通过结合长期权重与短时权重实现模型的双阶段加权集成。

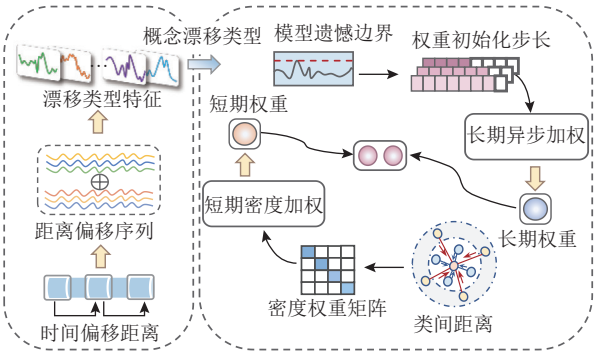


Fig. 1 Overall framework of OTAE method

图 1 OTAE 方法整体框架

2.1 基于时序偏移的漂移类型识别

为提高不同漂移类型下在线学习模型的适应速度，需先识别不同类型的概念漂移，不同漂移类型具有不同的漂移特性，这些特性会导致在线学习模型的适应速度存在较大差异，因此，针对不同漂移类型的模型更新具有必要性。

本文通过计算数据流中不同数据块间的时间偏移距离来实现漂移类型的检测。假设数据流中第 1 个数据块 D^{S_1} 不含概念漂移，在数据块上训练可得初始模型，利用该模型对数据块 D^{S_1} 与 D^{S_2} 进行预测得到的 2 个数据块先验精度为

$$\varepsilon_t^{S_1}=\left\{\varepsilon_{t,i}^{S_1}\right\}_{i=1}^{n},\varepsilon_t^{S_2}=\left\{\varepsilon_{t,i}^{S_2}\right\}_{i=1}^{n}, \tag{1}$$

其中 t 为当前时刻，n 为数据块大小。利用数据块间的先验精度，可计算出数据块 D^{S_1} 与 D^{S_2} 在时刻 t 的时间偏移距离并度量 2 个数据序列的相似度，该距离可以帮助计算序列间的变化和差异。数据块 D^{S_1} 与 D^{S_2} 在时刻 t 的时间偏移距离表示为：

$$d\left(\varepsilon_t^{S_1},\varepsilon_t^{S_2}\right)=\left|\varepsilon_t^{S_1}-\varepsilon_t^{S_2}\right| 。 \tag{2}$$

通过计算整个数据序列的时间偏移距离，可以得到 2 个数据块的距离偏移序列 $\{d_i\}_{i=1}^{n}$，利用匹配路径能够探测出 2 个数据序列间不匹配的部分。2 个数据序列的路径匹配过程表示为：

$$\varepsilon_{t,i}^{S_1}\leftarrow\varepsilon_{t,i}^{S_2} 。 \tag{3}$$

同样地，若提取 D^{S_1} 与后续的数据块间的时间偏移距离，便可得到一系列距离偏移序列，将该序列与

不同漂移类型特征结合，即可检测出概念漂移发生的区域并识别不同漂移类型，漂移类型识别过程如图 2 所示。

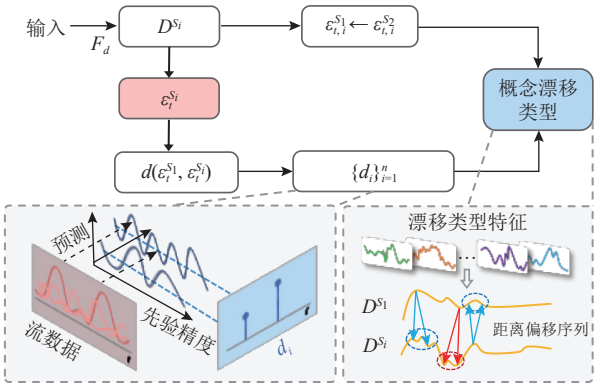


Fig. 2 Drift type identification process

图 2 漂移类型识别过程

2.2 持续异步模型权重更新

当发生不同类型概念漂移时，模型的漂移适应速度存在差异，特别地，若概念漂移较为剧烈时，模型漂移适应速度会明显下降，这一问题在在线学习模型中较为突出。如何在在线学习环境中更好地集成模型以适应概念漂移是一个重要的问题，指数梯度下降是一种常用的在线学习方法[25]。给定 n 个数据实例的数据集 $D=\{(x_i,y_i)\}_{i=1}^{n}$，以及 d 个具有不同参数的学习器 $F=\{\tilde{y}_i=f_i(x)\}_{i=1}^{d}$，则模型预测目标是将预测误差最小化为

$$\min_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w})=\left\|\sum_{i=1}^{d}w_i f_i(x)-y\right\|^2;\ \text{s.t. } \boldsymbol{w}\in\Delta, \tag{4}$$

其中 Δ 为一个 d 维单纯形，即 $\Delta=\{\boldsymbol{w}_t|w_{t,i}\geqslant 0 \text{ and 且} \|\boldsymbol{w}_t\|_1=1\}$，$t$ 为当前时刻，为方便起见，式(4)忽略了下标 t。

根据指数梯度下降方法，本文选取的单纯形的中心点为 $\boldsymbol{w}_1=(w_{1,i}=1/d)_{i=1}^{d}$，令 $\ell_{t,i}$ 表示学习器 f_i 在时刻 t 的损失，则每个学习器的权重 w_i 的更新规则为

$$w_{t+1,i}=\frac{w_{t,i}\exp(-\eta\ell_{t,i})}{Z_t}, \tag{5}$$

其中 $Z_t=\sum_{i=1}^{d}w_{t,i}\exp(-\eta\ell_{t,i})$，$\eta$ 为模型学习步长。

由模型权重更新可知，组合权重 $\boldsymbol{w}$ 依靠历史信息生成，这导致其不能快速适应数据分布的变化。若对模型学习步长 η 进行调整，当步长 η 很小时，权重变化不大，不会发生明显的自适应；当使用较大的步长 η 时，权重快速增加来适应环境变化，但是需要多次迭代才能完成。由此可知，无论如何调整步长，模型都无法在短时间内适应数据分布的变化。

根据上述分析，本文在指数梯度下降方法结合历史信息的基础上再结合漂移变化，提出了持续异步权重更新方法。具体地，通过丢弃部分历史信息，每隔 k 步重新初始化模型组合权重 $\boldsymbol{w}$，以加速模型的权重更新。此时，每隔 k 步模型组合权重可更新为

$$\boldsymbol{w}_{k+1}=\left(w_{k+1,i}=1/d\right)_{i=1}^{d}。 \tag{6}$$

通过对模型遗憾界限分析可知，当 $T>2\log d$ 时，设时间步长 $t=1,2,\cdots,T$ 的遗憾为 $R(T)$，取 $\eta=\sqrt{2\log d/T}$，此时指数梯度下降方法有遗憾界限为

$$R(T)=\sum_{t=1}^{T}\mathcal{L}(\boldsymbol{w}_t)-\inf_{\boldsymbol{u}}\sum_{t=1}^{T}\mathcal{L}(\boldsymbol{u})\leqslant\sum_{t=1}^{T}\sum_{i=1}^{d}w_{t,i}\|f_i(x)-y\|^2-\inf_{\boldsymbol{u}}\sum_{t=1}^{T}\mathcal{L}(\boldsymbol{u})\leqslant\sqrt{2T\log d}。 \tag{7}$$

证明.

$$\log\frac{Z_{t+1}}{Z_t}=\log\frac{\sum_{i=1}^{d}w_{t,i}\exp(-\eta\ell_{t,i})}{Z_t}=\log\sum_{i=1}^{d}p_{t,i}\exp(-\eta\ell_{t,i}), \tag{8}$$

其中 $p_{t,i}=w_{t,i}/Z_t\leqslant 1$ 是归一化权重。假设 $\eta\ell_{t,i}\in[0,1]$，利用泰勒展开 $\exp(-x)\leqslant 1-x+x^2/2$ 和对数不等式 $\log(1-x)\leqslant -x(x\in[0,1])$ 可以得到

$$\log\sum_{i=1}^{d}p_{t,i}\exp(-\eta\ell_{t,i})\leqslant\eta\sum_{i=1}^{d}p_{t,i}\ell_{t,i}+\eta^2/2, \tag{9}$$

$w_{t,i}$ 是非归一化权重，$w_{1,i}=1$ 且 $Z_1=d$，由此可以得到 $\log Z_{T+1}$ 的上下界：

$$\log Z_{T+1}\leqslant-\eta\sum_{t=1}^{T}\sum_{i=1}^{d}p_{t,i}\ell_{t,i}+\frac{T\eta^2}{2}+\log d, \tag{10}$$

$$\log Z_{T+1}\geqslant-\eta\sum_{i=1}^{d}\ell_{t,i}。 \tag{11}$$

令 $\zeta=\eta\left(\sum_{t=1}^{T}\sum_{i=1}^{d}p_{t,i}\ell_{t,i}-\inf_{\boldsymbol{u}}\sum_{t=1}^{T}\mathcal{L}(\boldsymbol{u})\right)$，结合上述推导 $\log Z_{t+1}$ 的上下界与步长 $\eta=\sqrt{2\log d/T}$，可以得到

$$\eta\left(\sum_{t=1}^{T}\sum_{i=1}^{d}p_{t,i}\ell_{t,i}-\sum_{i=1}^{d}\ell_{t,i}\right)\leqslant\zeta\leqslant\frac{T\eta^2}{2}+\log d。 \tag{12}$$

证毕.

对持续异步权重更新方法的遗憾界限分析可知，设 $\boldsymbol{I}=(l,l+1,\cdots,r)\in\{1,2,\cdots,T\}$ 为任意时间段，取 $|I|=T^n(0\leqslant n\leqslant 1)$。结合上式可以发现，若令 $k\in I$，可得 $R(k)\leqslant 2\sqrt{k\ln d(d-1)}$，此时存在 L/k 轮模型重新加权（L 为历史信息总数），并且模型遗憾界限将变为 $O(L/k\times 2\sqrt{k})$。考虑最坏情况，即 $L/k=k$，可得模型遗憾界限为 $O(k)$。综合以上所有情况可得模型遗憾界限为

$$R(I)\leqslant O(L/k\sqrt{k}+k)。\tag{13}$$

若令$L=T^n$，$k=T^m$，且$m\leqslant n$，当取$m=2n/3$时，模型遗憾界限将取得最小值$O(T^{2n/3})$。此时，在任意的$n<3/4$的区间内，都有$R(I)<O(T^{1/2})$，即当$0\leqslant k<T^{1/2}$时，持续异步权重更新方法的模型遗憾界限均比初始模型更小。此时，结合模型误差最小化过程，计算可得异步加权步长k为

$$\arg\min_{k}\mathcal{L}(\boldsymbol{w}_{k+1})=\left\|\sum_{i=1}^{d}\boldsymbol{w}_i f_i(x)-y\right\|^2。\tag{14}$$

通过持续异步权重更新方法，使模型取得了更低的遗憾界限，解决了在线学习模型漂移适应速度慢的问题，为模型快适应多类别概念漂移提供了基础。

2.3 短时窗口密度驱动的权重调控

由持续异步权重优化方法的遗憾界限分析可以发现，当异步加权步长较长时，模型性能提升并不明显，因此为提高模型适用性与预测性能，本文提出了短时窗口密度驱动的权重调控方法。该方法在持续异步加权的基础上，利用样本类内与类外距离计算样本混合密度，生成密度权重矩阵，实现持续异步权重优化与短时窗口密度驱动的权重调控的有效结合。

在给定数据集$D=\{(x_i,y_i)\}_{i=1}^n$的基础上，设x_i所属类定义为C_p，其余类定义为C_n。利用这2类数据，可计算出x_i在类C_p中的k个近邻数据的类内距离$\tilde{d}_i^p=\{d_{i1}^p,d_{i2}^p,\cdots,d_{ik}^p\}$及$x_i$在类$C_n$中的$h$个近邻数据的类外距离$\tilde{d}_i^n=\{d_{i1}^n,d_{i2}^n,\cdots,d_{ih}^n\}$。利用2种距离可计算出$x_i$在类$C_p$与$C_n$中的类内密度$\delta_i^p$与类外密度$\delta_i^n$分别为

$$\delta_i^p=\frac{1}{\mathbb{E}[\tilde{d}_i^p]}=\frac{1}{\dfrac{1}{k}\displaystyle\sum_{j=1}^{k}d_{ij}^p},\tag{15}$$

$$\delta_i^n=\frac{1}{\mathbb{E}[\tilde{d}_i^n]}=\frac{1}{\dfrac{1}{h}\displaystyle\sum_{j=1}^{h}d_{ij}^n}。\tag{16}$$

由式(15)(16)可以发现，类内密度越大，类外密度越小，表示x_i越靠近类C_p中心；类内密度越小，类外密度越大，表示x_i越靠近类C_p边界。结合上述特征，本文利用混合密度V_i进一步度量数据的类别分布，混合密度可表示为：

$$V_i=\beta\delta_i^p+(1-\beta)\delta_i^n,\tag{17}$$

其中β为优化密度的调节参数，由式(17)可以发现，混合密度越大时样本越靠近类别中心。

由持续异步权重更新方法可知，当异步加权步长$k\geqslant T^{1/2}$时，模型效果较差，这可能是因为模型训练数据漂移样本较多导致。因此本文结合样本混合密度，针对数据中较难适应的少数类样本，生成x_i的密度权重C_{ii}，使模型的训练更加关注少数类样本，减小噪声对模型的影响。其可表示为

$$C_{ii}=\gamma_i\times(1/N_{ik}),\tag{18}$$

其中γ_i用于样本混合密度的归一化，其可表示为$\gamma_i=V_i/\sum_{i=1}^{n}V_i$，$N_{ik}$表示第$i$个样本所属$k$类样本的总数。经多次迭代可得一个$n\times n$的密度权重矩阵$\boldsymbol{C}_w$，该矩阵为对角矩阵。

通过持续异步权重更新方法可得模型长期权重$\boldsymbol{w}_t=(w_{t,1},w_{t,2},\cdots,w_{t,d})$，将其与密度权重矩阵结合，更新模型权重矩阵可得$\tilde{\boldsymbol{w}}_t=\boldsymbol{w}_t\cdot\boldsymbol{C}_w$。利用更新后的模型权重，可在提高模型适应速度的同时，提高模型预测性能。此时，模型的预测误差可表示为

$$\mathcal{L}(\tilde{\boldsymbol{w}})=\left\|\sum_{i=1}^{d}\tilde{\boldsymbol{w}}_i f_i(x)-y\right\|^2。\tag{19}$$

2.4 总体结构

在流数据环境中，当发生多类别概念漂移时，模型的漂移适应速度存在差异，需针对不同漂移类型特征对模型继续调整，使模型保持良好泛化性能的同时兼顾较快的适应速度。如图3所示，OTAE首先利用不同数据块间的时间偏移距离提取距离偏移序列特征来识别概念漂移类型。其次，结合模型遗憾边界，动态异步初始化模型模型权重，实现模型持续异步权重更新。之后，利用样本类内与类外距离生成密度权重矩阵，实现模型短时窗口密度驱动的权重调控。最后，结合长期权重与短时权重矩阵，实现模型双阶段加权集成。OTAE方法的实现过程如算法1所示。

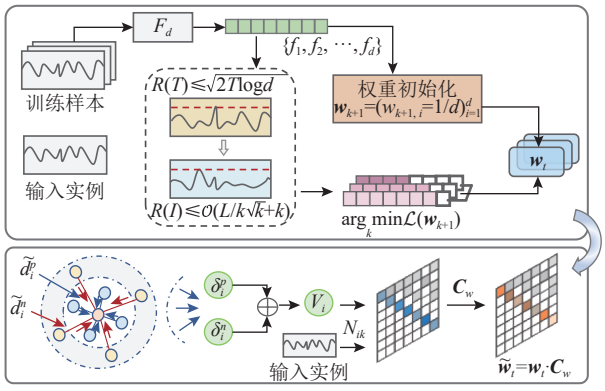


Fig. 3 Two-stage weighted ensemble process

图3 双阶段加权集成过程

算法 1. OTAE 方法。

输入: 数据流 $SD=D^{S_1},D^{S_2},\cdots,D^{S_T}$, 模型加权维度 d, 模型学习步长 η, 优化密度调节参数 β;

输出: 双阶段加权更新的模型权重矩阵 $\tilde{\boldsymbol{w}}_t$。

① 初始化指数梯度下降模型 $F_d=\{f_1,f_2,\cdots,f_d\}$;

② for $D^{S_i}\in SD$ do

③ 对 D^{S_i} 与 $D^{S_{i+1}}$ 预测的先验精度 $\varepsilon_t^{S_{i+1}}$ 与 $\varepsilon_t^{S_{i+2}}$;

④ 根据式(2)计算时间偏移距离 $d(\varepsilon_t^{S_{i+1}},\varepsilon_t^{S_{i+2}})$;

⑤ 生成距离偏移序列 $\{d_i\}_{i=1}^n$ 识别漂移类型;

⑥ 根据式(14)计算异步加权步长 k;

⑦ if $i/k=0$ then

⑧ 根据式(6)初始化模型组合权重 $\boldsymbol{w}$;

⑨ end if

⑩ 根据式(15)(16)计算类内外密度 δ_i^p, δ_i^n;

⑪ 根据式(17)计算样本混合密度 V_i;

⑫ for $i=1:n$ do

⑬ 根据式(18)生成密度权重矩阵 $\boldsymbol{C}_w$;

⑭ end for

⑮ 结合长期权重与密度权重矩阵更新模型权重矩阵 $\tilde{\boldsymbol{w}}_t=\boldsymbol{w}_t\cdot\boldsymbol{C}_w$;

⑯ end for

3 实验分析

为了验证 OTAE 方法的有效性, 本文在 8 个具有不同类型概念漂移的数据集上进行实验。分别从实时分类精度、累积分类精度、漂移发生后模型的收敛性以及不同重构方法对模型性能影响这 4 个方面将本文方法与 11 种不同的方法进行性能对比。实验平台是搭载酷睿 i5-8265U 的 Window 10, 内存是 8 GB。实验是在 Python 3.7 环境中编写和运行的。

3.1 实验数据集

为了检验本文方法对不同类型概念漂移的处理能力, 本文使用海量在线分析(massive online analysis, MOA)平台中的流数据生成器[26]生成了 5 个具有突变式、增量式、渐进式、瞬时式以及重现式的概念漂移数据集。除合成数据集外, 实验中还采用了 3 个真实数据集。实验中所用数据集的具体信息如表 1 所示。

Table 1 Datasets Information

表 1 数据集信息

数据集	实例数	维度	类别数量	漂移类型	漂移数量	漂移发生位点
RBFblips	100 000	10	2	突变式	3	25 000, 50 000, 75 000
Hyperplane	100 000	10	2	增量式	3	25 000, 50 000, 75 000
SEA_gradual	100 000	3	2	渐进式	3	25 000, 50 000, 75 000
SEA_blip	100 000	3	2	瞬时式	3	25 000, 50 000, 75 000
SEA_recurrent	100 000	3	2	重现式	3	25 000, 50 000, 75 000
Electricity	45 300	8	2	未知	未知	未知
Spam	9 300	500	2	未知	未知	未知
Weather	25 600	8	2	未知	未知	未知

3.2 参数设置

为了充分验证本文方法应对概念漂移的有效性, 本节在不同参数下进行了实验研究。

1)模型加权维度 d。此参数作为影响模型加权及预测的关键参数, 直接作用于模型的训练及预测过程, 可能影响集成模型的整体表现。为了避免 d 值选取不当, 本节分别在 $d\in\{10,20,40,80,100\}$ 的情况下展开讨论, 通过对比不同 d 值下的模型性能, 筛选出最优的模型加权维度。

2)数据块大小 w。由于过小的数据块难以提供足够的样本特征, 导致基学习器稳定性较差; 而过大的数据块可能包含概念漂移, 影响模型分类效果。因此, 本文统一将 w 设为 100。

3)优化密度调节参数 β。该参数的大小直接影响短时窗口密度驱动的权重调控过程中模型对类内与类外密度信息的考虑, 当 β 较大时, 导致模型忽视不同样本类间的关联, 当 β 较小时则相反。因此, 本节将 β 设置为 0.5。

4)模型学习步长 η。该参数控制模型权重的更新幅度。使用较大的步长 η 时, 模型权重快速更新来适应环境变化, 需要多次迭代才能完成; 步长很小时, 权重变化不大, 不会发生明显的自适应, 本文采用动态调整策略, 统一设置 $\eta=\sqrt{2\log d/T}$, 其中 d 为模型加权维度, T 为时间步长, 参数 η 值会随 T 的增大而逐渐减小, 既保证了初期快速学习的能力, 又确保了后期训练的稳定性。

3.3 评价指标

为了评估 OTAE 方法的性能, 本文主要从 3 个指标进行分析。

1)平均实时精度(average real accuracy, *Avgracc*)表示模型在每个时间戳的实时精度的平均值, 用于反映模型的实时性能, 是分类结果的有效衡量标准。其定义为:

$$Avgracc=\frac{1}{T\times n}\sum_{t=1}^{T}n_t\,, \tag{20}$$

其中 T 表示经过的时间步长, n_t 表示第 t 个时间戳正确分类的样本数, n 表示每个时间戳的样本数。平均

实时精度越大表明模型的分类性能越好。

2）累积精度（cumulative accuracy, *Cumacc*）。表示过去T_t个时间戳内累积预测正确样本数与累积获取样本数的比值。它反映了模型在过去所有时间步上的性能，其定义为：

$$Cumacc=\frac{1}{T_t\times n}\sum_{t=1}^{T_t}n_t\ 。\tag{21}$$

3）收敛速度（convergence speed, *CS*）。表示在发生概念漂移后模型收敛到新分布数据的能力。其定义为：

$$CS=step\times avge\ ,\tag{22}$$

其中*step*表示模型从概念漂移位点到收敛位点所用的步数，*avge*表示收敛过程中分类错误率的均值。本文统一设定收敛判定阈值$\rho=0.02$，如果一个位点的性能与其后m个参考位点的平均性能之差小于收敛判定阈值ρ，并且参考位点前半部分与后半部分性能差异小于收敛判定阈值ρ，则认为该位点为收敛位点，即

$$\left|\frac{\sum_{i=1}^{\lfloor m/2\rfloor}acc_{t+i}}{\lfloor m/2\rfloor}-\frac{\sum_{i=\lceil \frac{m}{2}\rceil}^{m}acc_{t+i}}{m-\lceil m/2\rceil+1}\right|<\rho\ 且\ \left|acc_t-\frac{\sum_{j=1}^{m}acc_{t+j}}{m}\right|<\rho,\tag{23}$$

其中$acc_t=n_t/n$，*CS*越小表明模型能越快收敛到漂移前的性能。

3.4 实验结果与分析

为了评估OTAE的性能，本文选择ARF[27]，IMB-WAE[28]，LEV[29]，WAEM[30]，SRP[31]，RUS[32]，DNN，HBP[33]，Highway[34]，Resnet[35]，Densenet[36]进行对比实验和结果分析。

3.4.1 精度分析

本节首先分析了OTAE在不同模型加权维度d下的性能表现。表2展示了不同d值下的平均实时精度。其中，当$d=80$时，模型的平均实时精度相对较高。当d值过低或过高时，可能会受模型权重设置不合理影响，出现欠拟合或过拟合问题，导致模型性能下降。因此本文选择在$d=80$的情况下进行对比分析。

Table 2 Average Real-Time Accuracy with Different Model Weighted Dimension *d*

表2 不同模型加权维度 *d* 下的平均实时精度

数据集	*d*=10	*d*=20	*d*=40	*d*=80	*d*=100
Hyperplane	0.857 2	0.879 3	0.887 4	**0.888 6**	0.854 5
RBFblips	0.881 6	0.876 6	0.884 0	0.883 6	**0.884 6**
SEA_blip	0.940 6	0.938 7	**0.942 8**	0.939 7	0.936 5
SEA_gradual	0.870 5	0.872 7	0.874 2	**0.874 7**	0.867 4
SEA_recurrent	0.899 4	0.897 2	0.897 7	**0.904 5**	0.898 3
Electricity	0.935 9	0.938 2	0.935 6	0.936 2	**0.939 4**
Spam	**0.973 9**	0.971 4	0.968 5	0.971 6	0.961 7
Weather	0.858 6	**0.859 5**	0.858 2	0.856 9	0.856 0
平均排名	3.38	3.25	2.88	**2.25**	3.75

注：黑体数值表示最高平均实时精度。

表3展示了不同方法在各个数据集上的平均实时精度、本文方法与最优对比方法的综合排名情况。从表3中可以看出，在合成数据集与真实数据集上，OTAE方法的平均实时精度均高于其他方法，在多数数据集上，性能差异均较为明显。这是因为OTAE方法在持续异步权重更新保证模型预测性能的基础上，引入短时窗口密度驱动的权重调控方法，提高了模型的漂移适应速度，进一步提高了模型的预测性能。

Table 3 Average Real-Time Accuracy of Different Methods on Each Dataset

表3 不同方法在各数据集上的平均实时精度

数据集	ARF	IMB-WAE	LEV	WAEM	SRP	RUS	DNN	HBP	Highway	Resnet	Densenet	OTAE(本文)
Hyperplane	0.779 7	0.782 0	0.734 8	0.766 0	0.666 6	0.713 7	0.827 8	0.822 2	0.841 5	0.839 7	0.846 5	**0.863 4**
RBFblips	0.817 8	0.554 0	0.671 6	0.629 0	0.629 4	0.905 8	0.801 9	0.823 0	0.848 2	0.850 0	0.845 9	**0.907 2**
Sea_blip	0.797 3	0.790 7	0.747 8	0.616 9	0.631 2	0.580 0	0.723 9	0.711 1	0.735 6	0.705 5	0.756 2	**0.931 4**
Sea_gradual	0.837 3	0.833 7	0.776 4	0.776 3	0.835 9	0.827 8	0.831 1	0.807 3	0.823 8	0.867 7	0.853 3	**0.869 9**
Sea_recurrent	0.863 8	0.864 0	0.803 2	0.805 3	0.863 0	0.842 1	0.801 6	0.855 0	0.838 8	0.857 3	0.826 6	**0.864 6**
Electricity	0.753 6	0.765 7	0.627 1	0.640 2	0.628 7	0.633 0	0.612 2	0.591 3	0.648 6	0.635 6	0.682 5	**0.907 6**
Spam	0.899 9	0.849 9	0.903 3	0.897 9	0.895 9	0.924 2	0.930 7	0.873 2	0.876 3	0.879 2	0.949 9	**0.965 9**
Weather	0.791 3	0.693 8	0.755 1	0.796 4	0.824 4	0.813 1	0.799 9	0.803 4	0.823 3	0.655 0	0.824 7	**0.861 5**
平均排名	5.41	7.02	8.90	9.37	7.45	7.19	7.58	8.16	6.15	6.47	3.97	**1.00**

注：黑体数值表示最高平均实时精度。

图 4 展示了 OTAE 与对比方法在不同数据集上的累积精度。从图 4 中可以看出，在合成数据集与真实数据集上，OTAE 的累积精度均高于其他方法，且在漂移位点处该方法的漂移适应速度均优于其他方法。这是由于该方法利用混合密度，针对数据中较难适应的少数类样本生成了密度权重矩阵，使模型更加关注少数类样本，减小了漂移数据对模型的影响。当发生不同类型概念漂移时，模型均能快速适应。

表 4 展示了不同方法在已知漂移位点的 5 个数据集上对应漂移位点处的收敛速度。统一设置参考位点个数 $m=20$，表中每个方法对应数据集的 3 个值分别代表前、中、后 3 个位点的收敛值。从表 4 中可以看出，在 5 个不同漂移类型的数据集上，OTAE 均具有较好的收敛性能，再次证明该方法的有效性。这主要是由于 OTAE 将持续异步加权的 k 步初始化模型权重的策略与短时窗口密度矩阵有效结合，提高了模型的收敛速度。

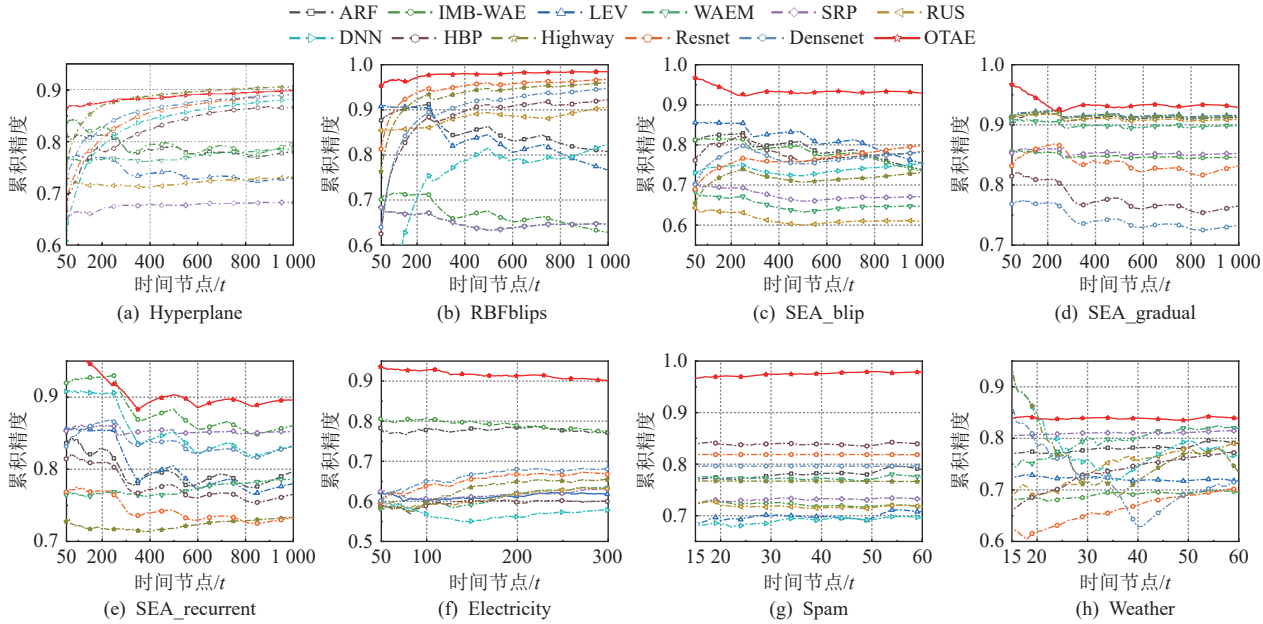


Fig. 4 *Cumacc* of different methods on each dataset

图 4 不同方法在各数据集上的累积精度

Table 4 Convergence Speed of Different Methods on Each Dataset

表 4 不同方法在各数据集上的收敛速度

数据集	ARF	IMB-WAE	LEV	WAEM	SRP	RUS
Hyperplane	3.26/10.33/2.31	16.33/15.37/6.59	**2.11**/5.33/3.81	3.82/6.11/5.32	3.56/5.37/4.69	3.41/4.68/4.50
RBFblips	14.70/9.96/10.11	11.20/9.37/10.76	1.89/5.38/2.64	**1.43**/5.33/3.26	2.37/5.31/3.45	8.33/23.98/9.61
Sea_blip	6.39/5.31/3.54	4.60/3.91/5.34	7.28/6.31/7.59	10.67/10.39/5.87	11.21/11.08/5.21	4.67/3.94/3.91
Sea_gradual	23.59/20.77/8.28	18.46/32.11/23.86	13.77/14.21/23.69	7.31/7.28/7.33	**6.27**/7.39/8.99	21.79/6.81/15.46
Sea_recurrent	15.31/7.93/2.70	7.51/5.33/4.31	12.57/11.24/4.21	6.14/10.65/7.51	6.76/10.32/8.47	15.99/26.97/33.75
平均排名	7.22	8.27	7.33	6.27	7.05	7.16

数据集	DNN	HBP	Highway	Resnet	Densenet	OTAE（本文）
Hyperplane	3.77/4.45/3.99	4.15/2.33/5.19	3.96/4.58/3.33	3.87/5.32/9.31	3.33/4.79/3.24	2.79/**1.63**/**1.10**
RBFblips	11.24/11.27/21.12	6.01/12.38/13.56	7.51/13.58/10.67	11.70/8.67/10.21	1.74/3.77/3.60	5.11/**1.75**/**0.94**
Sea_blip	6.21/7.37/3.51	4.30/4.21/4.97	3.31/5.27/4.88	6.33/5.24/**3.44**	6.46/7.81/7.81	**3.06**/**3.50**/5.76
Sea_gradual	10.67/9.31/10.08	10.75/13.28/15.08	13.33/6.72/17.55	12.70/11.77/7.52	12.90/14.76/11.23	7.65/**6.37**/**5.25**
Sea_recurrent	18.31/14.22/7.07	**4.36**/3.74/4.90	7.30/12.57/8.74	12.11/9.79/13.57	5.37/7.48/7.37	5.50/**1.42**/**0.77**
平均排名	7.00	6.22	6.05	6.88	5.93	**1.95**

注：*/*/*分别表示 3 个漂移位点（25 000，50 000，75 000）处的收敛速度，且黑体数值表示最高平均实时精度。

本文采用 Bonferroni-Dunn 检验方法分析了 OTAE 方法与其他对比方法之间的显著性差异。该检验方法将各方法平均序值差与临界值 CD 对比，若其大于临界值则说明平均序值高的方法优于平均序值低的方法，反之则没有差异。

$$CD = q_a \sqrt{\frac{K(K+1)}{6N}}, \tag{24}$$

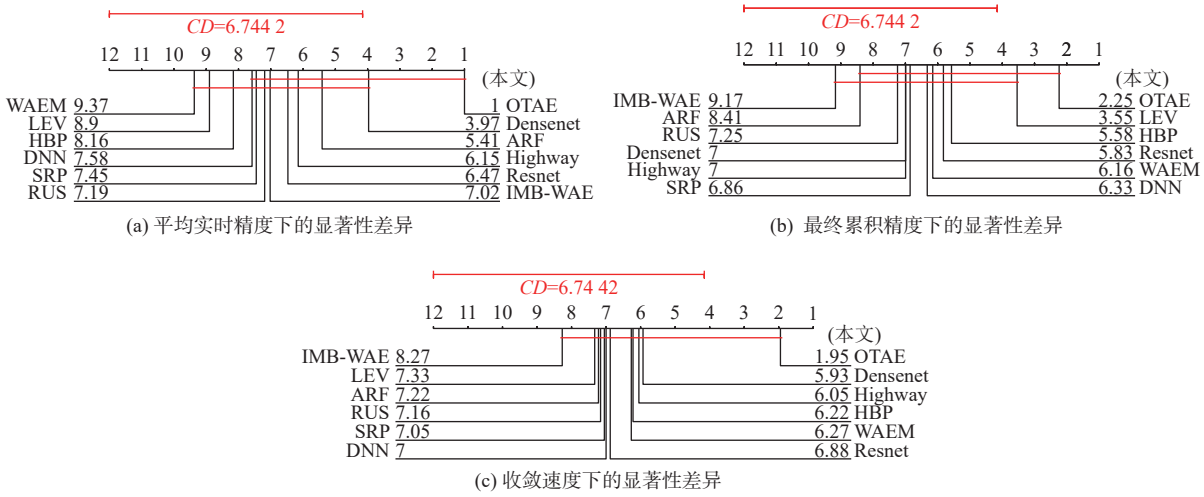
其中 K 为方法的个数，N 为数据集的个数，q_a 为显著水平 α 下的临界值，经计算可得，当 $\alpha = 0.01$ 时，$CD = 6.744\,2$。统计分析结果如图 5 所示，结果表明，OTAE 方法具有明显的优势。

3.4.2 消融实验

为测试双阶段加权集成(TSWE)的有效性，实验比较了模型在持续异步权重优化(CAWO)与短时窗口密度驱动的权重调控(SDWC)下的性能。由图 6 可以看出，若单独采用 SDWC，模型在大部分数据集上表现较差，这是由于此时模型过度关注漂移数据而忽视正常数据，导致模型原有权重受到破坏，降低了模型原有的适应性；若单独采用 CAWO，模型不会根据当前样本进行调整，在大部分数据集上表现也较差。结果表明，将 2 种方法结合后，模型取得了良好的泛化性能，有效验证了 OTAE 的合理性。

4 结束语

针对流数据中的多类别概念漂移存在的模型适应速度较慢且差异较大的问题，本文提出一种多类型概念漂移引导的在线分时适应性集成方法(OTAE)。该方法通过提取距离偏移序列识别概念漂移类型，其次根据模型遗憾边界分析权重初始化

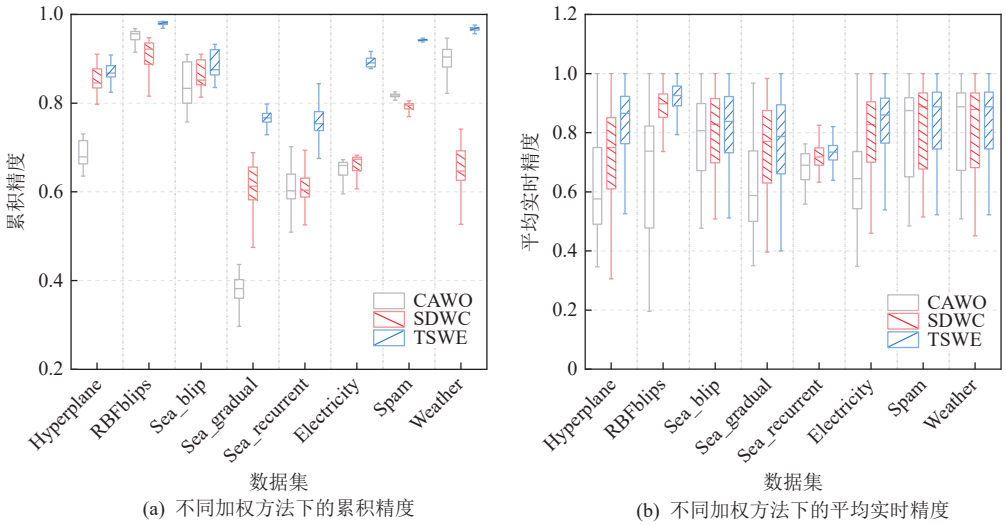


(a) 平均实时精度下的显著性差异　(b) 最终累积精度下的显著性差异

(c) 收敛速度下的显著性差异

Fig. 5 Bonferroni-Dunn test results of different methods

图 5 不同方法的 Bonferroni-Dunn 检验结果



(a) 不同加权方法下的累积精度　(b) 不同加权方法下的平均实时精度

Fig. 6 Influence of three weighted methods on the model stability

图 6 3 种加权方法对模型稳定性的影响

步长,以动态异步初始化模型权重。随后,利用样本混合密度生成密度权重矩阵。最后,将长期权重与短时权重矩阵结合,实现模型的2阶段加权集成。本文在8个具有不同漂移类型的数据集上进行了实验。实验结果表明,与其他方法相比,该方法在不同类型概念漂移的适应方面具有显著优势。然而,在多类别概念漂移的适应过程中,漂移检测与模型更新的持续性对漂移适应也有较大影响,因此在下一步研究中,考虑将OTAE方法与其他持续集成技术结合,以进一步提高模型性能。

作者贡献声明:郭虎升提出思想、设计方法,撰写及修改论文;张旭撰写论文、实现代码、测试数据及修改论文;张羽桐撰写论文及实现代码;王文剑指导写作、修改和审定论文。

参 考 文 献

[1] Blázquez-García A, Conde A, Mori U, et al. A review on outlier/anomaly detection in time series data[J]. ACM Computing Surveys, 2021, 54(3): 1−33

[2] Habeeb R A A, Nasaruddin F, Gani A, et al. Real-time big data processing for anomaly detection: A survey[J]. International Journal of Information Management, 2019, 45(C): 289−307

[3] Zhao Peng, Zhou Zhihua. Learning from distribution-changing data streams via decision tree model reuse[J]. SCIENTIA SINICA Informationis, 2021, 51(1): 1−12 (in Chinese)
(赵鹏, 周志华. 基于决策树模型重用的分布变化流数据学习[J]. 中国科学: 信息科学, 2021, 51(1): 1−12)

[4] Zhai Tingting, Gao Yang, Zhu Junwu. Survey of online learning algorithms for streaming data classification[J]. Journal of Software, 2020, 31(4): 912−931 (in Chinese)
(翟婷婷, 高阳, 朱俊武. 面向流数据分类的在线学习综述[J]. 软件学报, 2020, 31(4): 912−931)

[5] Liu Sanmin, Xue Shan, Wu Jia, et al. Online active learning for drifting data streams[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 34(1): 186−200

[6] Guo Husheng, Ren Qiaoyan, Wang Wenjian. Concept drift class detection based on time window[J]. Journal of Computer Research and Development, 2022, 59(1): 127−143 (in Chinese)
(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)

[7] Sato D M V, De Freitas S C, Barddal J P, et al. A survey on concept drift in process mining[J]. ACM Computing Survey, 2021, 54(9): 1−38

[8] Suárez-Cetrulo A L, Quintana D, Cervantes A. A survey on machine learning for recurring concept drifting data streams[J]. Expert Systems with Applications, 2023, 213: 118934

[9] 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)

[10] Cesa-Bianchi N, Lugosi G. Prediction, Learning, and Games[M]. New York: Cambridge University Press, 2006

[11] Guo Husheng, Zhang Yang, Wang Wenjian. Dynamical targeted ensemble learning for streaming data with concept drift[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(12): 8023−8036

[12] Chen Wuxing, Yang Kaixiang, Yu Zhiwen, et al. A survey on imbalanced learning: Latest research, applications and future directions[J]. Artificial Intelligence Review, 2024, 57(6): 137

[13] Oza N C, Russell S J. Online bagging and boosting[C]//Proc of the 8th Int Workshop on Artificial Intelligence and Statistics. New York: PMLR, 2001: 229−236

[14] Hou B, O'connor J, Andreas J, et al. Promptboosting: Black-box text classification with ten forward passes[C]//Proc of the 40th Int Conf on Machine Learning. New York: PMLR, 2023: 13309−13324

[15] Li Xuchun, Wang Lei, Sung E. AdaBoost with SVM-based component classifiers[J]. Engineering Applications of Artificial Intelligence, 2008, 21(5): 785−795

[16] Büyükçakir A, Bonab H, Can F. A novel online stacked ensemble for multi-label stream classification[C]//Proc of the 27th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2018: 1063−1072

[17] Yang Kaixiang, Yu Zhiwen, Chen C L P, et al. Incremental weighted ensemble broad learning system for imbalanced data[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 34(12): 5809−5824

[18] Jain M, Kaur G, Saxena V. A K-Means clustering and SVM based hybrid concept drift detection technique for network anomaly detection[J]. Expert Systems with Applications, 2022, 193: 116510

[19] Li Wendi, Yang Xiao, Liu Weiqing, et al. DDG-DA: Data distribution generation for predictable concept drift adaptation[C]//Proc of the 36th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2022, 36(4): 4092−4100

[20] Weinberg A I, Last M. EnHAT—Synergy of a tree-based ensemble with hoeffding adaptive tree for dynamic data streams mining[J]. Information Fusion, 2023, 89: 397−404

[21] Bifet A, Gavalda R. Learning from time-changing data with adaptive windowing[C]//Proc of the 7th SIAM Int Conf on Data Mining. Philadelphia, PA: SIAM, 2007: 443−448

[22] Liu Anjin, Lu Jie, Zhang Guangquan. Diverse instance-weighting ensemble based on region drift disagreement for concept drift adaptation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(1): 293−307

[23] Guo Husheng, Zhang Shuai, Wang Weijian. Selective ensemble-based online adaptive deep neural networks for streaming data with concept drift[J]. Neural Networks, 2021, 142: 437−456

[24] Elwell R, Polikar R. Incremental learning of concept drift in

nonstationary environments[J]. IEEE Transactions on Neural Networks, 2011, 22(10): 1517−1531

[25] Hill S I, Williamson R C. Convergence of exponentiated gradient algorithms[J]. IEEE Transactions on Signal Processing, 2001, 49(6): 1208−1215

[26] Bifet A, Holmes G, Pfahringer B, et al. MOA: Massive online analysis, a framework for stream classification and clustering[C]// Proc of the 1st Workshop on Applications of Pattern Analysis. New York: PMLR, 2010: 44−50

[27] Ridder F, Chen K H, Alachiotis N. Accelerated real-time classification of evolving data streams using adaptive random forests[C]//Proc of the 23rd Int Conf on Field Programmable Technology. Piscataway, NJ: IEEE, 2023: 232−237

[28] Wegier W, Maczyński M, Woźniak M. Changing lineup classifier ensemble for drifting imbalanced data streams[C]//Proc of the 24th IEEE Int Conf on Data Mining Workshops. Piscataway, NJ: IEEE, 2024: 238−245

[29] Bifet A, Holmes G, Pfahringer B. Leveraging bagging for evolving data streams[C]//Proc of the European Conf on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer, 2010: 135−150

[30] Mulimani D, Kanakaraddi S G, Totad S G, et al. Weighted averaging ensemble model for concept drift adaptation in streaming data[C]//Proc of the 2nd Int Conf on Intelligent Technologies. Piscataway, NJ: IEEE, 2022: 1−8

[31] Gomes H M, Read J, Bifet A. Streaming random patches for evolving data stream classification[C]//Proc of the 19th IEEE Int Conf on Data Minin. Piscataway, NJ: IEEE, 2019: 240−249

[32] Wang Boyu, Pineau J. Online bagging and boosting for imbalanced data streams[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(12): 3353−3366

[33] Sahoo D, Pham H Q, Lu J, et al. Online deep learning: Learning deep neural networks on the fly[C]//Proc of the 27th Int Joint Conf on Artificial Intelligence. Morgan Kaufmann: IJCAI, 2018: 2660−2666

[34] Srivastava R K, Greff K, Schmidhuber J. Training very deep networks[C]//Proc of the 29th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2015: 2377−2385

[35] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770−778

[36] Gao Huang, Zhuang Liu, Pleiss G, et al. Convolutional networks with dense connectivity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 44(12): 8704−8716



Guo Husheng, born in 1986. PhD, professor, PhD supervisor. Distinguished member of CCF. His main research interests include machine learning, data mining, and computational intelligence.

郭虎升, 1986 年生。博士，教授，博士生导师。CCF 杰出会员。主要研究方向为机器学习、数据挖掘、计算智能。



Zhang Xu, born in 2001. Master candidate. Her main research interests include stream data mining and online machine learning.

张　旭, 2001 年生。硕士研究生。主要研究方向为流数据挖掘、在线机器学习。



Zhang Yutong, born in 1999. Master. His main research interests include stream data mining and online machine learning。

张羽桐, 1999 年生。硕士。主要研究方向为流数据挖掘、在线机器学习。



Wang Wenjian, born in 1968. PhD, professor, PhD supervisor. Distinguished member of CCF. Her main research interests include machine learning, data mining, and computational intelligence.

王文剑, 1968 年生。博士，教授，博士生导师。CCF 杰出会员。主要研究方向为机器学习、数据挖掘、计算智能。