



计算机科学

COMPUTER SCIENCE

数据流概念漂移处理方法研究综述

陈志强, 韩萌, 李慕航, 武红鑫, 张喜龙

引用本文

陈志强, 韩萌, 李慕航, 武红鑫, 张喜龙. [数据流概念漂移处理方法研究综述](#)[J]. 计算机科学, 2022, 49(9): 14-32.

CHEN Zhi-qiang, HAN Meng, LI Mu-hang, WU Hong-xin, ZHANG Xi-long. [Survey of Concept Drift Handling Methods in Data Streams](#)[J]. Computer Science, 2022, 49(9): 14-32.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[基于对偶变分多模态注意力网络的不完备社会事件分类方法](#)

Dual Variational Multi-modal Attention Network for Incomplete Social Event Classification

计算机科学, 2022, 49(9): 132-138. <https://doi.org/10.11896/jsjcx.220600022>

[基于数据流特征的比较类函数识别方法](#)

Strcmp-like Function Identification Method Based on Data Flow Feature Matching

计算机科学, 2022, 49(9): 326-332. <https://doi.org/10.11896/jsjcx.220200163>

[监督和半监督学习下的多标签分类综述](#)

Survey of Multi-label Classification Based on Supervised and Semi-supervised Learning

计算机科学, 2022, 49(8): 12-25. <https://doi.org/10.11896/jsjcx.210700111>

[RIIM:基于独立模型的在线缺失值填补](#)

RIIM:Real-Time Imputation Based on Individual Models

计算机科学, 2022, 49(8): 56-63. <https://doi.org/10.11896/jsjcx.210600180>

[基于图卷积神经网络的文本分类方法研究综述](#)

Review of Text Classification Methods Based on Graph Convolutional Network

计算机科学, 2022, 49(8): 205-216. <https://doi.org/10.11896/jsjcx.210800064>

数据流概念漂移处理方法研究综述

陈志强 韩萌 李慕航 武红鑫 张喜龙

北方民族大学计算机科学与工程学院 银川 750021

(15720602388@163.com)

摘要 目前非稳态数据流中的概念漂移愈来愈呈现出不同速度、不同空间分布的趋势,给数据挖掘、机器学习等诸多领域带来了极大的挑战。近二十年来,许多致力于在非稳态数据流中处理概念漂移的技术方法被提出。从一种新颖的角度,分别针对主动检测的显式方法和被动自适应的隐式方法对目前的概念漂移处理技术方法进行了全面的阐述。首先,从处理某一特定类型和多种类型的概念漂移的角度对主动检测方法进行了分析,并从单学习器和集成学习的角度对被动自适应方法进行了分析;其次,对诸多概念漂移处理方法的对比算法、学习模型、适用漂移类型、算法的优缺点进行了全面总结;最后给出了未来的研究方向,包括类不平衡的数据流概念漂移处理方法、含新颖类的概念漂移数据流处理方法、含噪声的数据流概念漂移处理方法等方面。

关键词: 数据流;概念漂移;分类;主动方法;被动方法

中图法分类号 TP391

Survey of Concept Drift Handling Methods in Data Streams

CHEN Zhi-qiang, HAN Meng, LI Mu-hang, WU Hong-xin and ZHANG Xi-long

School of Computer Science and Engineering, North Minzu University, Yinchuan 750021, China

Abstract At present, concept drift in the nonstationary data stream presents a trend of different speeds and and different space distribution, which has brought great challenges to many fields such as data mining and machine learning. In the past two decades, many methods dedicated to handling concept drift in nonstationary data stream emerged. A novel perspective is proposed to classify these methods. The current concept drift handling methods are comprehensively explained from the explicit method of active detection and the implicit method of passive adaption. In particular, active detection methods are analyzed from the perspective of handling one specific type of concept drift and handling multiple types of concept drift, and passive adaptive methods are analyzed from the perspectives of single learner and ensemble learning. Many concept drift handling methods are analyzed and summarized in terms of the comparison algorithm, learning model, applicable drift type, advantages and disadvantages of algorithms. Finally, further research directions are given, including the concept drift handling methods in class-imbalanced data streams, the concept drift handling methods in data stream with the existence of novel classes, and the concept drift handling methods in the data stream with noise.

Keywords Data stream, Concept drift, Classification, Active methods, Passive method

1 引言

近年来,随着大数据、物联网技术以及人工智能技术的迅速发展,各行各业都在持续产生大量数据,而且一直以惊人的速度不断增长,这些数据因其自身特性被称为数据流,如网络数据、天气预报数据、无线传感数据、金融和电网数据等^[1]。由于数据流的时序性、高速性、多变性、潜在无限性等特性,传统的数据挖掘模型很难再适用于数据流挖掘,数据流挖掘

成为当前的热点研究方向。数据流挖掘在诈骗检测^[2]、垃圾邮件过滤^[3]、入侵检测^[4-5]、股票预测^[6]、用户兴趣预测^[7]等方面有着广泛的应用。在数据流挖掘的最初阶段,大部分数据流算法通常假定数据分布是稳定的,然而在真实世界的数据流应用中,数据分布在不断变化,模型需要不断更新以适应新的数据流环境。概念漂移通常表现为从样本属性到样本类别之间的映射关系,即目标概念(Target Concept)下随着数据流的不断变化而变化^[8]。概念漂移的存在成为数据流挖掘尤其

到稿日期:2021-07-12 返修日期:2021-12-10

基金项目:国家自然科学基金(62062004);宁夏自然科学基金(2020AAC03216)

This work was supported by the National Natural Science Foundation of China(62062004) and Ningxia Natural Science Foundation Project(2020AAC03216).

通信作者:韩萌(2003051@nmu.edu.cn)

是分类问题中的一个重要问题。在许多现实世界的应用场景中,数据的基本分布是不平稳的,随着时间的推移,以前有效的模型将不再有效。因此处理数据流中的概念漂移问题成为一个亟待解决的难题。

在处理非稳态数据流时,通常可以把处理概念漂移的方法分为主动检测方法和被动自适应方法。一方面,主动检测方法只有在出现漂移时才会更新学习器,因此更具反应性,从而可以节省时间和内存资源。此外,它们能够提供关于漂移的有用描述,如:速度、严重程度、发生的时间。另外,使用主动检测方法时,不仅要注意保持学习模型的性能,还要控制误检率和漏检率,以防止造成时间和内存的浪费。经典的算法主要有漂移检测法(Drift Detection Method, DDM)^[9]、早期漂移检测法(Early Drift Detection Method, EDDM)^[10]、自适应滑动窗口(Adaptive Windowing, ADWIN)^[11]等概念漂移检测器,以及自适应分类器集成(Adaptive Classifiers Ensemble, ACE)^[12]和自适应随机森林(Adaptive Random Forest, ARF)^[13]等集成方法。另一方面,被动自适应方法通常以固定的时间间隔隐式地使学习器来适应当前的概念,不使用任何漂移检测。它们以恒定的速度抛弃旧的概念,而不管是否发生了漂移。当连续数据源之间的差异与触发漂移不太相关时,这些方法对于处理渐变型概念漂移非常有用。经典的算法有自适应增量式决策树(Concept-adapting Very Fast Decision Tree Learner, CVFDT)^[14]、在线序列极限学习机(Online Sequential Extreme learning Machine, OS-ELM)^[15]等单学习器算法,以及流集成算法(Streaming Ensemble Algorithm, SEA)^[16]、精度加权集成算法(Accuracy Weighted Ensemble, AWE)^[17]和动态加权多数(Dynamic Weighted Majority, DWM)^[18]等集成学习算法。

本文的总体框架图如图1所示。本文首先阐述了概念漂移的定义及类型,然后从处理单一类型和多种类型概念漂移的角度分析主动检测方法,最后从单学习器和集成学习两个角度分析被动自适应方法。

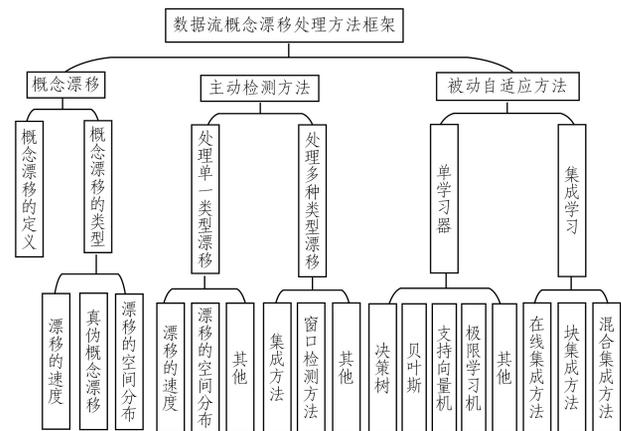


图1 总体框架图

Fig. 1 Overall framework

本文的主要贡献如下:

(1)对现有的概念漂移处理技术方法从基于主动检测方法和基于被动自适应方法的新颖角度进行了全面的阐述与分析。

(2)对算法的适用漂移类型、学习模型、算法的优缺点、对比算法等方面进行了阐述与总结。从处理概念漂移类型的角度对主动检测方法进行了分析,以及从单学习器和集成学习对被动自适应方法进行了分析。

(3)分析了现有算法中存在的问题,并提出了下一步的工作方向。

2 概念漂移

概念漂移是数据流挖掘中广泛存在的问题^[19],一般由数据流中的数据随着时间不断变化与发展而产生。数据流是一个潜在的、无限的、有序的数据项序列,并随着时间的推移按顺序到达^[20]。因此,数据流具有动态变化性,且形成数据项的分布会随着时间不断变化。在传统的数据挖掘中,一个数据集往往服从同一种数据分布,而数据流中的变化是不可预知的,数据项会过时,不再与当前的环境保持一致,过时数据会降低当前模型的训练能力。这就要求学习模型能够动态地调整自身以适应概念漂移。

2.1 概念漂移的定义

概念漂移是由流式数据随时间的变化或演变引起的。底层分布的改变会导致到达的实例的特征向量不再反映类标签。这对使用流数据分布进行预测的分类器的可靠性和准确性造成消极影响。

假设数据流是以连续的 (x_t, y_t) 实例的形式出现,其中 $t=1, 2, 3, \dots$,并且 x_t 是一个特征向量, y_t 则是属于一个具有 n 个类标签的集合,即 $y \in \{y_1, y_2, \dots, y_n\}$ 。预测器在特定时间基于特征向量 x_t 得到的一个预测结果可以用 \hat{y}_t 来表示。那么在 t_0 到 t_1 的时刻内的概念漂移可以被定义为式(1)所示^[21]。数据流分布的变化,即发生了概念漂移可在式(1)中的联合概率分布的变化中体现出来。

$$\exists x_t: p_t(x_t, y_t) \neq p_{t_0}(x_t, y_t) \quad (1)$$

其中, p_t 表示在 t 时刻特征向量 x_t 和目标类标签 y_t 之间的联合概率分布。

文献^[22]中对概念漂移进行了深一步的描述。在某一时刻,条件类概念分布 $p(x_t, y_t)$ 可由式(2)得到。

$$p(x_t, y_t) = p(y_t) p(x_t | y_t) \quad (2)$$

然后对输入的 x_t 进行预测,根据贝叶斯决策论可得后验概率分布,如式(3)所示:

$$p(y_t | x_t) = p(y_t) p(x_t | y_t) / p(x_t) \quad \text{where } y \in \{y_1, y_2, \dots, y_n\} \quad (3)$$

其中, $p(x_t) = \sum_{t=1}^n P(y_t) P(x_t | y_t)$ 。

以上是基于贝叶斯理论对数据流中的概念漂移问题的一般定义。同时,许多其他的文献中对概念漂移的定义有着不一样的解释^[19,23]。

2.2 概念漂移的类型

目前概念漂移的类型可根据不同的分类标准进行划分。本节将从漂移的速度、真伪概念漂移以及漂移的空间分布3个方面对其进行划分。

2.2.1 漂移的速度

随着时间的推移,数据分布的变化可能以不同的形式表现出来。通常把一个新目标概念取代旧目标概念的时间步称作概念漂移持续的时间,而完成漂移的持续时间越短,漂移的速度就越快。因此,根据概念漂移的速度,可以把概念漂移分为突变型、渐变型、增量型以及重复型^[19]。突变型概念漂移(Abrupt Concept Drift)表示在某一个时间步 t 或某一段较短的时间步 $[t, t+\Delta t]$ 内,旧的目标概念突然被新的目标概念所取代。而渐变型概念漂移(Gradual Concept Drift)表示旧的概念通常会在一段较长的时间步内才会被取代。与渐变型漂移类似,增量型概念漂移(Incremental Concept Drift)中的新旧概念的改变也比较缓慢,为了更好地解释它们之间的区别,文献^[21]中提出了中间概念(Intermediate Concept)。概念漂移不仅可能发生在一个精确的时间戳上,而且可能持续很长一段时间,在一个概念即起始概念,转变为另一个概念即结束概念时,可能就会出现中间概念^[23]。如图 2(b)所示,数据分布在经过一定的时间步之后,完成漂移并最终稳定下来。而介于初始数据分布与最终数据分布之间的即为中间概念。因此增量型漂移和渐变型漂移的主要区别在于概念漂移的初始阶段和最终阶段之间是否存在中间概念。

当概念随着时间变化,在某些情况下,先前的概念在一段时间后重新出现时,则会被视为重复型概念漂移(Recurring Concept Drift)。漂移的再次出现可能是周期性的,也可能是非周期性的^[24]。这种重复出现的概念可以通过学习算法来提升有限数据下的性能,因为学习器可以保留先前概念的知识。以上 4 种类型的概念漂移分布如图 2 所示。

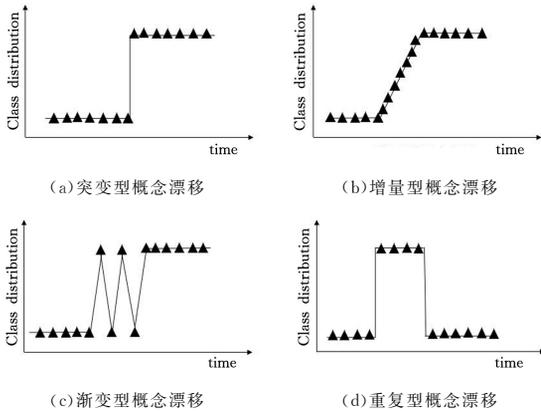


图 2 依据漂移的速度划分概念漂移

Fig. 2 Categorize concept drift according to drift speed

2.2.2 真伪概念漂移

目前一些研究^[8]还将概念漂移分为真实概念漂移(True Concept Drift)和伪概念漂移(Virtual Concept Drift)两种类型。将概念漂移视为数据分布的变化,当输出的条件分布发生变化而输入分布保持不变时,即是真实概念漂移。伪概念漂移的定义在目前的文献中有着不同的解释,包括数据分布的变化^[25]以及对目标概念的影响^[26]等。

目前区分真伪概念漂移的最普遍的定义如下^[21]。一个数据流是以连续的 (x_t, y_t) 实例的形式出现,其中 $t = 1, 2,$

$3, \dots$, 并且 x_t 是一个特征向量, y_t 则是属于一个具有 n 个类标签的集合,即 $y_t \in \{y_1, y_2, \dots, y_n\}$ 。预测器在特定时间基于特征向量 x_t 得到的一个预测结果可以用 \hat{y}_t 来表示。那么在 t_0 到 t_1 的时刻内的概念漂移可以被定义为:

$$\exists x_t : p_{t_0}(x_t, y_t) \neq p_{t_1}(x_t, y_t) \quad (4)$$

其中, p_t 表示在 t 时刻特征向量 x_t 和目标类标签 y_t 之间的联合概率分布。

那么根据类 y 的先验概率 $p(y_t)$ 或类条件概率分布 $p(x_t | y_t)$ 对类后验概率分布即 $p(y_t | x_t)$ 的影响,可以将概念漂移分为真实概念漂移和伪概念漂移。真实概念漂移表示,不管 $p(x_t)$ 是否发生变化,在 $p(y_t | x_t)$ 即类的后验概率分布上都会发生变化,这种变化会影响分类器的决策边界,进而降低学习器的性能。伪概念漂移则表示只有 $p(x_t)$ 即输入分布或者类分布即 $p(y_t)$ 发生变化,但未影响 $p(y_t | x_t)$ 的变化,在这种情况下,尽管输入分布发生了变化,但分类器的决策边界不受影响。真实概念漂移与伪概念漂移的数据分布如图 3 所示。

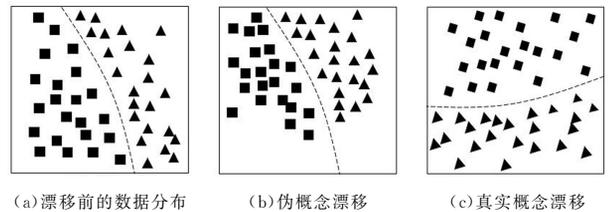


图 3 真实概念漂移与伪概念漂移的数据分布

Fig. 3 Data distribution of real drift and virtual drift

现有的处理概念漂移的研究大多集中在真实概念漂移上,概念漂移通过改变问题的真实决策边界来直接影响分类器的性能^[22]。然而,即使真实的类边界在虚拟漂移中没有改变,这种类型的漂移仍然可能导致所学的类边界变得不充分。如果后验概率不变,重新建立模型则没有意义,因为决策边界仍然是相同的。因此,处理真实漂移的技术可能仍然只适用于某些类型的虚拟漂移。伪概念漂移的检测也很重要,因为即使它不影响分类器的决策边界,但可能被检测并分类为真实概念漂移,所以也可能提供关于分类器再训练的错误决策^[27]。伪概念漂移会导致分类器的预测性能下降,因为即使伪概念漂移不会影响问题的真实决策边界,但在分类器不知道的搜索空间区域中观察到的情况也会影响所学习决策边界的适用性。因此,重新调整学习分类器的决策边界以避免错误分类变得十分重要^[27]。

2.2.3 漂移的空间分布

自概念漂移第一次被提出以来,已有大量文献从数据分布的角度来解决这个问题。然而,大多数基于数据分布的漂移检测方法都假设漂移发生在一个精确的时间点^[28]。因此,可以根据概念漂移完成后全局数据分布是否发生变化来划分,即分为局部概念漂移和全局概念漂移(Global Concept Drift)。

实例空间中某些区域在目标概念或数据分布上发生变化,同时变化的类型和严重程度又可能取决于变化在实例空间中的位置,因此可以将这种变化称为局部概念漂移^[29]。

在此类漂移中,局部区域发生了漂移,且总体变化不显著,因此难以解决此类变化带来的问题。例如,在垃圾邮件过滤系统中,如果用户改变了对某一主题的电子邮件的兴趣,但所有其他主题保持稳定,那么考虑到该主题在其所有电子邮件中所占的比例很小,这一小改变可能不会对用户的整体兴趣有重大影响。而全局概念漂移更容易检测,因为它会影响整个实例空间。在这种情况下,新旧概念之间的差异更为明显,并且可以更早地检测到漂移^[30]。局部概念漂移与全局概念漂移的数据分布如图4所示。

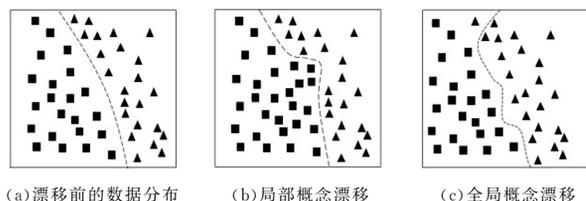


图4 局部概念漂移与全局概念漂移的数据分布

Fig. 4 Data distribution of local drift and global drift

3 主动检测方法

主动检测方法处理概念漂移旨在精确定义概念漂移的时间和严重程度,通过使用一种概念漂移触发机制,让学习器进行相应地调整以适应漂移^[31]。目前的一些研究方法旨在处理某一确定类型的概念漂移,即单一类型的概念漂移,这些方法根据当前类型漂移的具体特征有针对性地设计算法来更好地处理概念漂移。然而,现实世界中各种场景下的数据集中通常都是各种类型的概念漂移组合^[32],即多种类型的概念漂移。本节分别从处理单一类型概念漂移和多种类型概念漂移两个角度对主动检测方法处理概念漂移的算法进行阐述与分析。

3.1 处理单一类型概念漂移

在分析处理单一类型概念漂移算法时,从漂移的速度、漂移的空间分布以及其他方面分别进行阐述。

3.1.1 漂移的速度

根据漂移随时间的变化,可以把概念漂移分为突变型、渐变型、增量型和重复型。下文就从这4种漂移类型的角度进行阐述。

(1) 突变型漂移

当数据分布最终确定时,则表示概念漂移完成。时间步长可以是单个样本、一组样本或固定时间间隔的到达。概念完成所需的时间越短,漂移就越快。如果漂移在一个时间步内完成,则称为突变型概念漂移^[33]。

通常情况下,由于突变型概念漂移的特性,突变漂移很容易被检测出来。近年来,相当多的概念漂移检测器相继被提出。这些概念漂移检测器可分为基于统计的方法(DDM^[9], EDDM^[10], STEPDM^[34], DMDDM^[35]),基于窗口的方法(ADWIN^[11], SEED^[36], FHDDM^[37], MDDM^[38])和基于集成的方法(DDD^[39], ADOB^[40], BOLE^[41], AWOE^[42])。

其中,基于统计的方法中最具代表性的就是漂移检测方法(Drift Detection Method, DDM)^[9]。DDM估计分类器误差

及其标准差,如果分类器误差随着训练样本数的增加而增大,则表明可能发生了概念漂移。那么错误率达到一定级别时,DDM就会生成警告信号。如果达到警告级别,将在一个特殊窗口中记住新传入的样本;如果在此窗口内误差随时间增大并达到漂移水平,则重建当前的分类器。而EDDM(Early Drift Detection Method)^[10]则是通过比较两个连续错误率的距离来检测概念漂移。当数据流处于稳定状态时,连续错误之间的距离变大。因此EDDM更适合处理渐变型概念漂移。STEPDM(Detection Method Using Statistical Testing)^[34]则使用等比例的统计检验,并进行连续性校正,在两个处理过的数据窗口上计算。基于这样两个假设:在最近的 W 个实例上的分类精度将和学习算法初始的整体精度一致。因此,如果最近的分类精度降低幅度过大,则表明发生了漂移。此检验如式(5)所示:

$$T(r_0, r_r, n_0, n_r) = \frac{|r_0/n_0 - r_r/n_r| - 0.5 \left(\frac{1}{n_0} + \frac{1}{n_r} \right)}{\sqrt{\hat{p}(1-\hat{p}) \left(\frac{1}{n_0} + \frac{1}{n_r} \right)}} \quad (5)$$

其中, r_0 表示除最近的 W 个实例之外的总体 n_0 个实例中正确分类的数量, r_r 则表示 n_r 个实例中正确分类的数量。

Pesaranghader等^[37]提出使用Hoeffding不等式作为检测的统计检验(Fast Hoeffding Drift Detection Method, FHDDM)。与STEPDM类似,该方法比较了正确预测的最大总体概率和最近正确预测的概率。在HDDM(Drift Detection Method based on the Hoeffding's Inequality)^[43]中提出了两种方法,HDDM_A test和HDDM_W test,这两种方法都使用Hoeffding边界进行漂移检测,其中HDDM_A test比较移动平均值以检测漂移,HDDM_W test比较漂移检测的移动平均值的权重。值得注意的是,HDDM_A test更适用于突变型漂移,而HDDM_W test更适用于渐变型漂移。最近提出的漂移检测器DMDDM(Diversity Measure As a New Drift Detection Method)^[35]根据误差估计,并且根据不断变化的输入数据来计算分类器响应的多样性度量,并使用其计算和Page-Hinkley检验来检测概念漂移,该算法能以较少的时间和较少的内存消耗快速地对概念漂移做出反应。

自适应滑动窗口(Adaptive Windowing Algorithm, ADWIN)^[11]则是最具代表性的基于窗口的检测方法。ADWIN使用一个大小可变的 W 实例的滑动窗口,当检测到漂移时, W 减小。同时存储了两个能够动态调整的子窗口,分别表示较旧和最新的数据。当这些子窗口的平均值之差高于给定阈值时,则检测到漂移。基于ADWIN算法,SEED漂移检测方法^[36]比较了两个子窗口。当子窗口的平均值高于所选阈值时,旧的子窗口将被丢弃。SEED利用Hoeffding不等式和Bonferroni校正来计算检验统计量,并进行块压缩以消除割点和合并均匀块。

在基于集成的方法中,DDD(Diversity for Dealing with Drifts)^[39]分别在检测到概念漂移之前和之后的两个状态使用了具有高多样性和低多样性的4个集成。在检测到漂移之前,学习系统由具有较低分的集成和具有较高分的集成组成。

ADOB(Adaptable Diversity-based Online Boosting)^[40]则使用ADWIN 漂移检测器来监测分类的性能,ADOB 中的分类器和 Online Bagging and Boosting^[44]的分类器一样,返回结果如式(6)所示:

$$\begin{aligned} h(X) &= \arg \max_{y \in Y} \sum m: h_m(x) \\ &= y^{\log \frac{1}{\beta_m}} \text{ where } \beta_m = \frac{\epsilon_m}{1 - \epsilon_m}, \epsilon_m \\ &= \left(\frac{\lambda_m^{S_o}}{\lambda_m^{S_c} + \lambda_m^{S_o}} \right) \end{aligned} \quad (6)$$

其中, m 受基分类器的数量限制。

如果 ADWIN 检测器返回警告,则立即使用 ADOB 与现有的集成一起训练新的集成。确认发生漂移后,将使用新创建的集成,并删除代表最后一个分布的旧集成。基于对 ADOB 的修改,BOLE(Boosting-like Online Learning Ensemble)^[41]尝试了几种不同的策略来提高集成的精度,以应对尤其是存在突变漂移的场景。该算法首先减少控制允许哪些专家投票的前提条件并替换了几种在线学习方法中常用的概念漂移检测方法。Sun 等^[42]提出了一种新颖的基于自适应窗口的在线集成(Adaptive Window based Online Ensemble, AWOE)。该算法使用自适应窗口作为漂移检测器为每个集成成员分配不同大小的块。算法监测两个子窗口的平均值是否大于 Hoeffding 界限定义的阈值,并且该方法综合了两组集成来处理各种类型的概念漂移。

(2) 重复型漂移

在某些情况下,先前的概念在一段时间后重新出现时,会被视为重复型概念漂移。现有的处理重复型概念漂移的研究方法可分为两类。第一种方法是首先将过去的概念存储为模型,然后在触发概念漂移时使用元学习机制来寻找最佳匹配。第二种方法就是将过去学习到的概念存储在集成分类器中。

Gama^[45-46]等提出使用元学习技术来表征先前学习的模型的适用范围,元学习器可以检测到上下文的再次出现,并通过激活先前学习的模型来采取主动。该方法在重新利用以前学习的模型时,能够更快地检测出漂移。Angel 等^[47]为了处理重复概念,将一组代表漂移过程的实例发送给隐马尔可夫模型(Hidden Markov Model, HMM),从而从存储库中存储的模型中预测最可能重用的模型。Anderson 等^[48]提出一个名为 CPF(Concept Profiling Framework)的元学习器,该方法使用一个概念漂移检测器和一组分类模型,通过相似的分类行为来关联模型,可以对具有重复型概念漂移的数据流进行有效分类。在 CPF 的工作基础上,Anderson 等^[49]提出了 ECPF(Enhanced Concept Profiling Framework)。与 CPF 类似,ECPF 使用新分类器和现有分类器之间新数据分类的相似性来快速确定要重用的最佳分类器。

在重复型概念漂移中,集成模型中保留和重用旧模型的机制可以简化为重复概念重新训练新模型。

Gonçalves 等^[50]提出了一种检测重复概念漂移的方法(Recurring Concept Drifts, RCD),该方法为找到的每个上下文创建一个新的分类器,并存储起来用于构建它的数据样本。当出现新的概念漂移时,该算法使用非参数多元统计检验将

新的上下文与之前的上下文进行比较,以验证两个上下文是否来自同一分布。Sripirakas 等^[51]在文献[52]的基础上应用傅里叶编码光谱的集成来捕捉和挖掘数据流环境中的重复型漂移。同时采用两种机制来优化傅里叶谱的推导:一种用于能量阈值化,另一种用于加快傅里叶基函数的计算。Sidhu 等^[53]提出了一个维护两个集成的集成系统(Recurring Dynamic Weighted Majority, RDWM),用于处理重复型漂移。RDWM 有两个集成:一个主要的在线集成和一个次要集成。其中主要的在线集成代表当前概念并会像 DWM 中一样进行更新或删减,次要集成代表自学习开始以来的旧概念既不更新也不训练,只是从一次集成中复制最好的学习器。Gomes 等^[54]提出在特征空间中挖掘重复概念,用于解决在动态特征空间中学习循环概念的问题,同时降低与存储过去模型相关的内存成本,并能够利用学习过程的表现和上下文信息来检测和适应概念的变化。为了处理重复出现的概念,存储的模型被组合成一个动态加权的集成。HU 等^[55]则提出一种基于主要特征提取的概念聚类与预测算法(Recurring Concept Drift Base on Main Feature Extraction, MFCCP),通过计算不同批次样本的主要特征和影响因素的差异来识别重复出现的概念。

(3) 渐变型与增量型漂移

在渐变型概念漂移和增量型概念漂移中,概念之间的转变是缓慢发生的。这使得检测这两种漂移的难度提升^[56]。

Alippi 等^[57]引入一种新的 JIT 自适应分类器集成了一个估计过程动态的指标。这使得该算法能够利用监督和非监督样本来提高当观测跟随这种漂移时分类器的精度,能够跟踪和调整其知识库以适应渐变型概念漂移,同时又保持其处理突变型漂移的有效性。Liu 等^[58]提出了一种新的模糊窗口概念漂移自适应方法(Fuzzy Windowing Drift Adaptation, FWDA)。该算法允许滑动窗口保持一个重叠周期,从而可以更精确地确定属于不同概念的数据实例,以获得更高的精度。Abdualrhman 等^[59]提出一种确定性概念漂移检测方法(Deterministic Concept Drift Detection, DCDD)。为了处理渐变漂移,该模型通过比较用于训练分类器的记录和缓冲在当前窗口中的记录的属性投影值的分布和动量相似性来检测概念漂移。

3.1.2 漂移的空间分布

在现实环境中,概念漂移的发生往往是局部的。例如,只有特定的细菌可能对某些抗生素产生耐药性,而对其他抗生素的耐药性可能保持不变^[60]。现有的概念漂移检测方法很少关注局部概念漂移(Local Concept Drift)。局部漂移在某些情况下会被视为噪声,使得模型不稳定,因此模型必须有效区分局部变化和噪声以及处理代表漂移的实例的稀缺性,以便有效地更新学习器^[22]。因此,目前的研究主要关注处理局部概念漂移。全局概念漂移则更容易被检测,因为它会影响整个实例空间。

Gama 等^[61]提出利用决策树的内部节点来检测实例空间的局部区域型概念漂移。该算法连续监测学习算法的序贯

误差,搜索与先前稳定状态的显著偏差。在决策树和规则学习器等对实例空间区域进行不同功能匹配的决策模型中,该方法可用来监测实例空间区域的误差,具有模型自适应速度快的优点。针对集成方法无法处理局部概念漂移的现状,Tsymbol等^[29]提出用动态集成处理局部概念漂移。该方法根据每个实例的局部精度为其分配相应的权重。在动态集成中,每个基分类器接收与其在当前测试实例附近的局部精度成比例的权重,而不是全局分类精度。Ikonomovska等^[62]使用全局模型并结合局部漂移处理。通过一个回归树,以一种简洁的方式将特征空间划分为局部片段。因此,在漂移阶段记录的样本会影响到树的某个分支,然后对该分支进行调整并以更大的灵活性进行遍历。Shaker等^[63]提出对不同遗忘强度的特征空间局部区域进行处理,引入局部遗忘因子。这是通过在流学习中使用模糊模型结构来实现的,其结构组件提供了特征空间的局部划分。在发现区域型漂移与相邻数据实例变化之间关系的基础上,Liu等^[64]提出一种新的基于区域密度估计的漂移检测方法,即基于最近邻的密度变化识别(Nearest Neighbor-based Density Variation Identification, NN-DVI)。NN-DVI基于 K 近邻的空间划分模式,将不可测量的离散数据实例转换为一组共享子空间以进行密度估计,并通过距离函数累积这些子空间中的密度差异,量化了总体差异。Liu等^[65]提出了一种能够连续监测区域密度变化的局部漂移度(Local Drift Degree, LDD)检测方法。该算法同时考虑了时间相关和空间相关的漂移信息,来度量每个可疑区域发生区域漂移的可能性,并通过分析局部区域的密度增减情况来采取相应的行动。Liu等^[66]提出了一种基于局部漂移不一致的概念漂移自适应多实例加权集成算法(Diverse Instance-weighting Ensemble, DiWE),该算法通过定义不同的构造区域集,可以根据新出现的概念动态地改变实例的权重,选择多样性最高的集成。同时根据估计的局部漂移风险增量地调整实例权重,并在漂移变得具有统计意义之前将这些信息传递给学习模型。

3.1.3 其他

目前一些研究^[8]将概念漂移分为真实概念漂移和伪概念漂移两种类型。由于伪概念漂移不影响分类器的决策边界并可能被检测和分类为真实概念漂移,因此也可能提供关于分类器再训练的错误决策^[27]。因此,准确区分真伪概念漂移尤为重要。Oliveira等^[67]基于GMM增量方法创建了一个新的高斯函数,可用来快速适应系统的真实概念漂移。同样地,可通过对观测值最近点高斯分布的更新,使系统适应伪概念漂移。Oliveira等^[28]提出一种高斯混合模型以处理伪概念漂移和真实概念漂移,该模型通过更新和生成高斯函数来处理伪概念漂移,并通过重置系统来处理真实概念漂移。Almeida等^[68]提出一种动态分类器选择方法。他们提出可以用一个非常大的窗口来处理伪漂移,因为其中将包含许多与当前概念相对应的观测值。另一方面,可以用一个小窗口来处理真实漂移,因为通过加速转换到新的数据可以更快地排除其观测值。

3.1.4 小结

已有很多研究针对特定类型的概念漂移的人工数据集与真实数据集所提出的算法进行评估。

目前针对突变型、渐变型、增量型、重复型漂移的评估数据集主要包括垃圾邮件数据集(Spam)、电力数据集(Electricity)等真实数据集以及SINE1, STAGGER, MIXED, Hyperplane等工数据集生成器生成的人工数据集。最近,Pe-saranghader等提出的快速霍夫丁漂移检测方法(FHD-DM)^[37]、McDiarmid漂移检测方法(McDiarmid Drift Detection Method, MDDM)^[38]、Gözüaçık等提出的使用单类分类器漂移检测方法(One-Class Drift Detector, OCDD)^[69]以及Myuu提出的演变数据流中概念漂移的精确检测方法(Accurate Concept Drift Detection Method, ACDDM)^[70]利用机器学习中的不等式(如Hoeffding不等式和McDiarmid不等式),并与滑动窗口机制相结合,通过监督或无监督的方式检测平稳状态下和非平稳状态下的数据流的差异性来检测漂移,这些方法在延迟性、真阳性、假阳性和假阴性比率等评估指标上都优于其他方法。特别地,Micevska等提出的统计概念漂移检测方法(Statistical Drift Detection Method, SD-DM)^[71]在概念漂移探测器中引入可解释性,直接从数据中量化漂移以发现漂移的原因和漂移模式,因此在实际应用中比仅指示是否发生漂移的其他方法更具优势。

另外,利用Museba等提出的重复自适应分类器集成(RACE)^[72]将知识转移和漂移检测相结合,最大限度地提高多样性并创建动态决策边界来分隔训练实例、漂移检测和基于多样性的策略,以最少的计算开销及时学习重复概念并最终在精确度上优于其他算法。以Anderson等提出的增强概念分析框架(ECPF)^[49]为代表的元学习机制方法同样也是目前处理重复型概念漂移性能较好的方法。ECPF在新分类器和现有分类器之间快速确定最佳分类器以供重用,能够比最先进的集成技术即ARF^[13]更准确地对具有重复漂移的人工数据集进行分类,对真实数据集的分类速度是自适应随机森林(ARF)的6倍,而且准确度并未明显降低。目前,以主动方式处理渐变型概念漂移性能最佳的算法是Abdualrhman等提出的确定性概念漂移检测方法(DCDD)^[59]和Blanco等提出的霍夫丁漂移检测方法(HDDM)^[43]。DCDD通过比较用于训练分类器的记录和缓冲在当前窗口中的记录的属性投影值的分布和动量相似性来检测渐变漂移,其渐变漂移检测率达到了 0.957 ± 0.011 。

同样地,目前处理局部概念漂移性能最佳的算法是Liu等提出的基于局部漂移不一致的概念漂移自适应多实例加权集成算法(DiWE)^[66]以及Liu等^[64]提出的一种新的基于区域密度估计的漂移检测方法等算法。

3.2 处理多种类型概念漂移

用于处理数据流的现有分类模型大多专门处理某一特定类型的概念漂移,但是在现实世界中的数据总是包含多种类型的概念漂移,因此要求算法能够处理多种类型的概念漂移成为必然。将处理多种概念漂移的主动方法分为集成分类

方法、基于窗口的方法以及基于检测器集合等其他方法。

3.2.1 集成分类方法

由于数据流的先前数据项总是被丢弃以限制存储使用,单学习器需要使用最新的数据来不断地调整自己,因此单分类器不能利用先前流数据上的信息。然而,集成模型可以利用集成成员保存先前的信息。因此,为了处理非稳态数据流,漂移检测器可以与集成学习相结合以更好地处理概念漂移。

Nishida 等^[12]提出的自适应分类器集成算法由一个漂移检测器和基于块的集成组成。该方法通过跟踪每个输入样本的分类器错误率,对突变漂移做出反应,同时用大量样本缓慢重建分类器集成。Minku 等^[39]提出一种处理漂移的多样性在线集成学习方法(Diversity for Dealing with Drifts, DDD)。DDD 利用多样性的优势来处理漂移,并且能够获得比其他方法更好的精度。GOEL 等^[73]提出一种基于集成的在线多种漂移检测方法(Ensemblebased Online Diversified Drift detection, En-ODDD),通过使用一个动态更新的集成,来加快对不断变化的分布的适应性。该方法利用多种自适应环境下的在线漂移检测和剪枝技术,同时嵌入一个漂移检测器的多专家集成,保证了对概念漂移的及时反应。Ren 等^[32]提出知识最大化集成(Knowledge-Maximized Ensemble KME)。KME 利用有监督和无监督知识识别重复概念,并评估集成成员的权重,同时隐式地重用过去块中保留的标记实例,提高候选成员的识别能力以适应渐变型概念漂移。KME 通过每个集成成员的加权投票来实现最终决策,如式(7)所示。

$$(x_t) = \arg \max_{y_t \in Y} \left(\sum_{j=1}^{i-1} w_{i,j} * f(k_j(x_t), y_t) + (w_i * f(k_i(x_t), y_t)) \right) \quad (7)$$

其中, $B(x_t)$ 是使用集成 B 对标签 x_t 进行预测的结果, $f(k_j(x_t), y_t)$ 和 $f(k_i(x_t), y_t)$ 为指示函数。

在文献[74]中,两个窗口由一个随机搜索树维护,该树保存最近的数据和概念。该方法使用增量 K-S 检验算法来检测没有真实标签的概念漂移,同时允许使用随时间变化的两个样本立即执行 K-S 假设检验。Sun 等^[42]提出一种新的集成模式 AWOE,其结合了基于块的加权和在线处理,同时使用自适应窗口作为变化检测器,为每个集成成员分配不同大小的块,一旦检测到一个变化,就会构建一个新的分类器来替换集成中最差的分类器。Cabrera 等^[75]提出一种新的基于集成的在线自适应分类器集成算法(Online Adaptive Classifier Ensemble, OACE)。该算法在控制、警告和失控 3 个漂移阶段之间交替处理每个基分类器中的概念漂移。Dong 等^[76]提出一种模糊加权集成(Active Fuzzy Weighting Ensemble, AFWE)算法。其采用模糊实例加权和动态投票策略,将现有的基分类器组织起来,构建集成学习模型。该算法集成了漂移检测机制、模糊实例加权和动态投票策略,用于构建自适应集成学习模型。基于漂移度的判断指标,Zhang 等^[77]应用一种基于模糊积分的新算法对模型进行自适应更新,得到相应的分类结果。为了检测概念漂移,该算法提出了一种漂移程度的判断指标,并引入欧氏距离来实现这一目标。其计算式如式(8)所示。

$$d_E(x_1, x_2) = \sqrt{\sum_{a=1}^m d_a^2 \left(\frac{X_{1a} - X_{2a}}{\text{range}_a} \right)} \quad (8)$$

其中, x_1 表示当前的数据块中的实例, x_2 表示在之前数据块中最近的所有实例, m 代表实例的属性。

Montiel 等^[78]提出一种自适应梯度增强模型,其使用 ADWIN^[11]来跟踪性能的变化并以此来更新集成。这些变化是由分类准确度等指标来衡量的。

随机森林同样也是一种能够较好地处理概念漂移的集成方案。最初的针对概念漂移数据流分类问题的随机森林方法是 Abdulsalam 等^[79]提出的一种在线增量流分类集成算法(Streaming Random Forests, SRF),是标准随机森林分类算法的扩展。它能够处理具有进化性质和训练/测试数据记录的随机到达率的数据流。在不损失分类精度的前提下,SRF 算法能够处理底层数据发生概念漂移的问题,并且能够处理训练记录块不足以建立或更新分类模型的情况。Abdulsalam 等^[80]提出的动态流随机森林(Dynamic Streaming Random Forests, DSRF)对 SRF 进行了扩展,使用基于熵的漂移检测技术来处理不断变化的数据流。Gomes 等^[81]提出了一种自适应随机森林算法(Adaptive Random Forest, ARF)来对不断演变的数据流进行分类。与之前随机森林对数据流设置的适应性相比,ARF 使用了一种基于在线 Bagging 的理论上合理的双重采样方法和一种自适应策略来应对漂移。这种自适应策略基于对每棵树使用一个漂移检测器来跟踪警告和漂移,并在替换它们之前就训练新的树。受随机子空间方法^[82]和在线装袋^[44]的启发,Gomes 等^[81]提出了流随机补丁算法(Streaming Random Patches, SRP),SRP 采用了主动漂移检测策略,每个基模型都是在随机数据块(即特征和实例的随机子集)上进行训练。Luong 等^[83]提出流深度森林算法(Streaming Deep Forest, SDF),SDF 调整 gcForest^[84]模型,通过重用 gcForest 的级联结构,保留了 gcForest 的表示学习能力。为了动态更新模型,该算法将每层的经典随机森林替换为自适应随机森林(ARF)^[81]。该算法还提出了增强变量不确定性(AVU)的主动学习策略,以降低流媒体环境中的标记成本。

3.2.2 窗口检测方法

基于窗口的方法大致可分为基于单个窗口和基于两个窗口这两类方法^[22]。

EDDM^[10]使用单个窗口,通过监测两个误差 μ_i 和 σ_i 之间的平均距离来对其进行连续调整。然后,将两个误差 μ_i 和 σ_i 与误差间距离分布最大时达到的 μ_{\max} 和 σ_{\max} 进行比较,以检测漂移。Haque 等^[85]提出一种有效的基于半监督滑动窗口框架的 SAND,在分类每个测试样本时估计分类器置信度,存储在滑动窗口中,并使用漂移检测技术跟踪置信度估计中的任何显著变化,分类器置信度的变化表明概念漂移的发生。基于 SAND^[85]中使用基于分类器的置信度的漂移检测来检测概念漂移的思想,Haque 等^[86]提出了包含一个集成分类器并基于 KNN 型聚类的模型框架 ECHO。ECHO 维护了一个可变大小的窗口 W 来监视分类器对最近数据实例的置信度,提出了

关联度和纯度两个估计量,来估计分类器的置信度。

基于两个窗口的方法通常用一个窗口存储过去有用的信息,另一个窗口则用于收集数据流中最新到达的数据信息。若两个窗口的分布发生了显著差异,则表示发生了概念漂移。

自适应滑动窗口(ADWIN)是一种经典的使用两个可变大小窗口的方法,算法将监测分类器预测精度^[11],如果检测到性能的变化,则概念已经改变。ADWIN使用一个窗口 W 来识别分布的变化,并将 W 分成两个自适应子窗口,并比较底层的统计数据。ADWIN2作为ADWIN的一个改进版本,拥有一个可变大小的窗口。DDM^[9]也是一种统计比较两个窗口并控制预测过程中学习模型所产生误差的方法。一个窗口包含所有数据,另一个窗口仅包含从数据流到预测模型的错误率开始上升时之间的数据。STEPD^[34]在两个处理过的数据窗口上计算,使用等比例的统计检验,并进行连续性校正。STEPD基于这样一个假设:在最近的 W 个实例上的分类精度与学习算法初始的整体精度一致。因此,如果最近的分类精度降幅过大,则表明发生了漂移。Zliobaite^[87]主要解决了在延迟标签下的检测概念漂移的问题。为了检测 N 时间步内的概念漂移,该方法比较了来自原始数据的两种样本,分别是一个固定长度为 w 的检测窗口 X^T 和一个相同长度的参考窗口 X^R 。与文献[87]中的增量Kolmogorov-Smirnov算法不同,Reis^[74]选择固定其中一个窗口,而不是同时维护两个连续的滑动窗口。该方法能够在没有真实标签信息的情况下识别概念漂移,只需要一部分真实标签就可以使模型适应新的概念。Raab等^[88]提出了一种基于原型的自适应策略。该方法在滑动窗口 ψ 中存储 n 个样本,在每一个时间步内均匀地选取 r 个样本,并与窗口中最新的 r 个样本进行比较。第一个窗口为 $R = \{x_i \in \psi\}_{i=n-r+1}^n$,拥有 ψ 滑动窗口中的所有最新的样本 r 。第二个窗口 W 是由 ψ 滑动窗口中非最近部分均匀地采样来创建,其计算式如式(9)所示:

$$W = \{x_i \in \psi \mid <n-r+1 \wedge p(x) = u(x_i \mid 1, n-r)\} \quad (9)$$

该方法是对原有RSLVQ算法的一大改进。Yuan等^[89]提出一种基于多尺度滑动窗口的无监督概念漂移检测算法,通过 K 均值聚类和多尺度窗口得到总平均距离,作为概念漂移的检测指标,并利用统计过程控制系统对概念漂移进行检测,确定索引阈值的范围。Khamassi等^[30]提出了一种新的漂移检测机制EDIST2。EDIST2通过一个自适应窗口来监控学习器的性能,该窗口通过统计假设检验自动调整来有效地检测漂移。

3.2.3 其他

处理多概念漂移的另一种方法则是对漂移检测器进行集成。漂移检测器的集合可以灵活地对集合中的检测器进行添加或删除,以应对出现的新型概念漂移。该方法可以处理现阶段各种类型的概念漂移。

Du等^[90]提出一种基于选择性的检测器集合(Ensemble of Detectors, e-Detector)。e-Detector集成了不同的检测器来寻找不同速度的概念漂移,可以同时检测突变漂移和渐变漂移,并采用选择性集成策略以发现单个基于变化指标的

方法无法检测到的概念漂移。Hu等^[91]提出的漂移检测器集合框架(Ensemble Framework for Drift Detection, EFDD)则使用基于类型的投票,对于覆盖同一类型概念漂移的单个算法,多数表决将决定检测结果。在对每种类型的概念漂移检测进行判断后,将产生的来自所有不同类型漂移的检测判决的联合作为最终的检测判断。Korycki等^[92]首次提出使用无监督的检测器集合进行局部检测的方法,并使用增量K-S检验来识别特征子空间中的局部变化,而无需额外的监督。

聚类分析同样也是一种重要的数据流挖掘方法。Li等^[93]提出一种增量熵聚类算法,通过增量熵确定数据点与聚类的相异度,该算法能够自主确定相异距离的阈值,能同时检测出概念漂移和离群点。Bai等^[94]提出了分类数据流聚类的优化模型,通过不断更新聚类参数,使当前聚类模型与前一个聚类模型之间的误差尽可能小,参数可以通过EM算法求解。为了检测概念漂移,定义了一种新的测度。如果新度量的值大于预定义的阈值,则可以检测概念漂移。Katakis等^[95]提出一个利用流聚类对数据流进行分类的框架,动态地创建和维护一个集成的分类器。当新的一批实例到达时,该算法识别该批实例所属的簇,并应用相应的分类器来预测其实例的真实标签。Xu等^[96]提出的混合数据流的自适应邻域密度聚类方法利用一个显著性准则对分类属性值进行评价,使分类属性值变为数值,并提出一种基于邻域相似度的非线性降维方法对数据进行降维。在聚类方法中,通过邻域密度对每个点进行评价。根据粗糙集理论确定 k 点后,从互距最大的数据集中选取 k 点。

3.2.4 小结

在包含某种特定类型的概念漂移和包含混合类型概念漂移的人工数据集以及Poker hand、Coverttype、Spam、Electricity等经典分类评估数据集上进行实验评估。以集成方法主动处理多种混合类型概念漂移的方法中,Goel等提出的基于动态更新的多样化集成的概念漂移处理方法(En-ODDD)^[73]中的多数加权机制和在线漂移检测器的结合使得其在所有可能的漂移场景(突变、渐变、重复和混合)中都表现最佳。同时,通过使用在线装袋和多样化的更新机制修改传入的训练实例来引入多样性是En-ODDD在准确性方面优于大多数算法的主要原因。Museba等提出的基于多样性的处理概念漂移集成方法(Adaptive Diversified Ensemble Selection Classifier, ADES)^[97]为不同类型的概念漂移创建高多样性和低多样性的集成,通过多样性显著优化了对不同类型概念漂移的适应,使其在处理多种类型概念漂移时优于其他算法。在以窗口机制处理多种混合类型概念漂移的方法中,Yuan等^[89]提出的基于多尺度窗口的无监督漂移检测器通过 k 均值聚类和多尺度窗口得到总平均距离,将其作为概念漂移的检测指标,然后利用统计过程控制系统确定指标阈值的范围。5种不同维数据集的渐变和突变概念漂移的检测实验证明,该算法具有良好的概念漂移检测效果。表1列出了本节所述的基于主动检测方法的算法。

表 1 主动方法总结

Table 1 Summary of active methods

算法	适用漂移	学习模型	优点	缺点
AWOE ^[42]	突变漂移	集成方法	很好地解决了难以确定数据块的大小的问题,且相比其他集成方法在精度和内存使用上更好	获取所有真实值标签不现实
ADOB ^[40]	突变漂移	集成方法	在基学习器之间更有效地分配实例,快速从频繁的漂移中恢复	未考虑多样性与准确性对运行时间和内存使用情况的影响
BOLE ^[41]	突变漂移	集成方法	尝试对检测器进行更积极的参数化,使其对概念漂移更敏感	没有彻底分析改变界限参数和改变分类器权重对基分类器多样性和精确定度的影响
RCD ^[50]	重复漂移	集成方法	与使用同一基学习器的集成方法单分类器相比,在具有突变和渐变漂移中表现更好	未能根据特定的问题选择最佳的配置
MM-PRec ^[47]	重复漂移	元学习模型	比其他算法更加平缓地适应概念漂移,同时保持较好的预测性能	训练元学习模型需要额外的计算资源
CPF ^[48]	重复漂移	元学习模型	比 RCD 更精确,且运行速度比 RCD 快得多	重用的分类器将对传入的实例分类不准确
ECPF ^[49]	重复漂移	元学习模型	在保持分类精度的情况下,运行时间远少于其他算法少	对于何时复制与更改分类器有待改进
MFCCP ^[55]	重复漂移	元学习模型	比 CCP 的精度降低的幅度小得多	没有考虑真实数据集下的表现
RDWM ^[53]	重复漂移	集成方法	即使在资源受限的环境中也能保持较高的精确度	未能解决数据流中的新颖类问题
Adaptive JIT ^[57]	渐变漂移	JIT 模型	比 JIT 能够更好地处理渐变漂移,同时在突变漂移下也能保持高精度	在非线性漂移中,漂移期间无法恢复与在平稳环境下相同的性能
FW-DA ^[58]	渐变漂移	窗口模型	允许滑动窗口保持一个重叠周期,能够更精确地确定属于不同概念的数据实例,从而获得更高的精度	运行时间略长于其他算法
NN-DVI ^[64]	局部漂移	最近邻	能准确地检测出人工数据集集中的概念漂移,有利于解决现实世界中的概念漂移问题	NN-DVI 的性能会受窗口大小的限制
LDD ^[65]	局部漂移	最近邻	从一个新的角度解决了数据流挖掘中的一个关键问题,与现有的算法相比取得了很好的效果,并且在测试数据集上具有最佳性能	在高维数据上的执行时间比其他算法稍长
DiwE ^[66]	局部漂移	集成方法	整体性能与其他最先进的算法具有较好的可比性	数据集集中的漂移数量可能会增加算法的执行时间
ACE ^[12]	突变、渐变	集成方法	能够很好地处理噪声以及各种类型的概念漂移	不能有效地解决数据块大小难以确定的问题和内存消耗的限制
DDD ^[39]	突变、渐变	集成方法	比 EDDM 具有更高的精度,对虚警也有相当好的鲁棒性	对重复型、周期性漂移的处理效果有待改进
En-ODDD ^[73]	突变、渐变、重复	集成方法	在所有漂移下都具有稳定的精度性能,优于对比较算法;统计检验表明,En-ODDD 在不同的漂移流条件下具有较高的精度	对多种漂移共存的漂移场景的组合仍有待研究
KME ^[32]	渐变、重复	集成方法	避免半监督方法的标注成本,及时补充监督信息;分量评估和加权机制使得 KME 对随机斑点具有鲁棒性	不能处理具有标称属性的数据集中发生的突变漂移
AWOE ^[42]	突变、渐变	集成方法	很好地解决了设置适当块大小的问题,更适合不同类型漂移的场景,同时算法在精度和效率方面比其他集成方法更有效	不能很好地处理未标记数据流
OACE ^[75]	突变、渐变	集成方法	在突变和渐变漂移、不同噪声水平、无关属性和缺失属性值下都有优秀的性能	替换其他学习算法作为基分类器有待研究
ARF ^[81]	突变、渐变、增量	集成方法	在标签延迟和即时情况下都有很好的分类性能,特别是在真实数据集上,可用于处理具有大量特征的数据流	在所有特征都需要构建合理模型的数据集中表现不好
SDF ^[83]	突变、渐变、增量、重复	集成方法	在标记开销仅为 70% 的情况下,采用 AVU 主动学习策略的 SDF 算法显著优于其他用所有实例来训练的方法	由于使用多层结构,SDF 的运行时间较长
EDIST2 ^[30]	突变、渐变	窗口机制	能够较早地检测到漂移并保持较小的虚警率	未能解决延迟标签的问题
SAND ^[85]	突变、渐变	窗口机制	无论所使用标记数据的数量大小,SAND 都是有效的	资源消耗效率有待提升
e-Detector ^[90]	突变、渐变	检测器的集合	具有很强的泛化能力和自寻优能力	时空资源消耗过大
EFDD ^[91]	突变、渐变、非平稳突变、渐变	检测器的集合	可以检测多种复杂类型的漂移	未考虑半监督或无监督
SNDC ^[96]	突变、渐变	聚类分析	在混合数据集中效果较好	参数值对漂移的敏感性影响较大

4 被动自适应方法

被动自适应方法处理概念漂移将假设数据环境可能随时变化。随着数据的不断输入,该方法不断调整自身模型,不断地从环境中学习,而不使用漂移检测机制^[22]。根据所使用学习器的数量,被动处理概念漂移的技术方法通常可分类为单学习器方法和集成学习方法。

4.1 单学习器

在存在概念漂移的非稳态环境中,使用单学习器的自适应

学习算法被广泛应用。使用单学习器的方法可以很好地控制系统的复杂性,因为它们被设计成实时适应,并且计算工作量最小。

4.1.1 决策树

为适应数据流环境,学者们探索了基于传统分类模型(包括决策树模型、贝叶斯模型、贝叶斯网络、神经网络、支持向量机等)的改进算法,其中基于决策树模型的数据流分类方法成为主流方法之一。

增量式决策树模型 VFDT (Very Fast Decision Tree

learner)^[98]最早于2000年被提出。该算法采用增量式的决策树,通过不断地将叶节点替换为决策节点而生成。当某一节点的统计量达到一定阈值,则在节点进行分割测试,但是该算法未能解决概念漂移的问题。CVFDT(Concept adapting Very Fast Decision Tree learner)算法^[14]周期性地扫描整棵决策树,依据分类错误率检测内部节点是否发生了概念漂移,从而在漂移节点生成相应的替换子树。一旦替换子树的预测能力超过当前节点所在子树的预测能力,则用新子树替换掉旧子树。为了处理概念漂移,Blake等^[99]提出了一种重置操作,允许对树的过时部分进行局部重新学习。这种方法可以更好地适应概念漂移和特征空间的变化,同时仍然可以生成一个短而高精度的决策树。Jankowski等^[100]提出一种基于数据流的增量决策树归纳算法,基于进化算法,其可以同时优化不同的目标。使用增量学习策略,能够实时工作,同时可以积累或修改新知识,而不会忘记旧知识。Jankowski等^[101]提出结合树学习器和进化算法的概念漂移数据流增量学习方法,该算法逐步学习决策树,并将所有信息存储在种群的内部结构中。同时,Costa等^[102]提出了基于VFDT的SVFDT算法(Strict Very Fast Decision Tree, SVFDT),SVFDT比VFDT生成的树浅得多,适合具有内存和时间限制的数据流挖掘方法,同时在基于集成的解决方案中可被作为基学习器。Bifet等^[103]提出STREAM-C++系统,它是一种利用决策树和C++集成的挖掘流的新系统。它包含能够适应数据流的变化自适应决策树和强大的集成方法,如装袋、提升和随机森林。

4.1.2 贝叶斯

贝叶斯分类器因其在线性、低复杂度和低内存消耗等优点而成为数据流挖掘的一种常用模型。现已出现了许多以贝叶斯模型为基础的改进算法。

Krawczyk等^[104]提出一种简单而有效的用于流挖掘的朴素贝叶斯分类器(Weighted Naive Bayes for Concept Drift, WNB-CD)。为了使分类器能够快速调整其属性以适应不断变化的数据,该方法将遗忘机制作为权重衰减来实现。随着每次迭代的进行,先前对象的重要性级别都会降低,直到它们从数据收集中被丢弃为止。Zhao等^[105]提出一种新的动态朴素贝叶斯方法,将朴素贝叶斯方法推广到预测模型因概念漂移而改变的时间。该方法可以有效地预测何时更新预测模型以反映数据分布的实质性变化。Liu等^[106]提出一种自适应的快速切换朴素贝叶斯算法。该方法实现了基于残差的双样本测试,并根据评估窗口大小动态选择的漂移阈值来操纵一个滑动贝叶斯分类器和一个增量朴素贝叶斯分类器,进而检测漂移。Kishore等^[107]将粗糙集理论与朴素贝叶斯分类器相结合,提出了一种新的分类器模型,称为粗糙高斯朴素贝叶斯分类器(Rough Gaussian Naive Bayes Classifier, RGN-BC)。RGNBC利用粗糙集理论来检测概念漂移。Kishore等^[108]还提出了Pearson Gaussian Naive Bayes分类器(PGN-BC)。该方法是对现有的Gaussian Naive Bayes分类器(GN-BC)的改进,额外增加了属性之间的相关性。对于数据流的分类,PGNBC是基于概念漂移的频繁更新。

4.1.3 支持向量机

Syed等^[109]首次提出增量支持向量机学习来解决概念漂移问题,实验证明支持向量机在学习过程中能够很好地处理概念漂移。Ruping^[110]提出一种基于支持向量机的增量学习方法,该方法将数据分批提交给算法,使算法在每一个训练步骤后都产生一个初步的结果,不允许直接使用最后一个训练步骤中的所有数据,以减少时间和空间的消耗。Yalcin等^[111]提出了一种基于集成的支持向量机(SVM)分类器增量学习方法。该算法提出加入遗忘机制来消除环境变化中冗余数据的影响,这使得增量学习算法能够很容易地适应不断变化的环境。Ayad^[112]提出根据环境条件的变化定期更新支持向量机。该方法使用一个监视度量来观察类的条件概率分布在增长时间窗内的变化。如果检测到一个严重的变化,支持向量机参数将被更新。在数据流中存在概念漂移的情况下,在学习到的标签空间之外进行分类。ZareMoodi等^[113]提出利用基于支持向量机的方法(Support Vector Based Method, SVDD)来维持流中观察到的类的边界,通过构造邻域图来分析位于这些边界之外的新到达的实例,以检测所学习的标签空间之外的新类的出现。在核空间中通过缩小、扩大和合并球体来动态地保持边界,该方法能够适应数据中的渐变型和突变型漂移。

4.1.4 极限学习机

近年来,随着机器学习的迅速发展,涌现出了越来越多的学习算法,极限学习机(Extreme Learning Machine, ELM)就是其中之一。

ELM是Zhu等^[114]提出的一种单隐层前馈神经网络学习算法,具有学习速度快、泛化能力强等优点。最近提出了一些基于ELM的概念漂移处理算法,这类算法训练速度快,能够快速适应数据流的变化,并且能够检测和防止概念漂移。然而,ELM只能解决批量学习问题。为了将ELM应用于数据流,Liang等^[115]提出了一种名为OS-ELM的增量学习算法。与ELM相比,OS-ELM能够以固定或可变的块大小逐个或逐块学习数据,并且因其快速的学习速度而成为顺序学习的标准算法之一。基于OS-ELM的算法是解决概念漂移问题最常用的方法。例如,IDS-ELM利用搜索方法在隐藏层中找到适当数量的神经元。当隐藏在数据流中的概念发生变化和演化时,不符合数据流分类要求的分类器将被剔除并重新训练^[116]。与ELM相似,由于OS-ELM中加法节点的输入权值和偏差是随机选取的,输出权值是解析确定的,因此在不同的仿真实验中仍然存在不稳定性。对于大数据集,当存在未学习的训练样本时,OS-ELM不会停止学习,导致学习时间较长。为了解决这些问题,Zhai等^[117]提出了一种集成OS-ELM对大数据集进行分类的算法E-OS-ELM(Ensemble Online Sequential ELM)。为了提高OS-ELM的稳定性并考虑漂移问题,Lan等引入了EOS-ELM(Ensemble of Online Sequential ELM)^[118],EOS-ELM由多个OS-ELM网络组成,集成中每个OS-ELM输出的平均值被作为网络性能的最终度量。同时,Xu等^[119]提出了一种用于数据流分类的动态极限学习机(Dynamic ELM, DELM)。DELM采用在线学习机制训练ELM并将其作为基分类器,采用双隐层结构来提升

ELM 的性能。当设置了概念漂移警报时,在 ELM 中加入更多的隐层节点,以提高分类器的泛化能力。如果测量概念漂移值达到上限或 ELM 的精度较低,则删除当前分类器。Liu 等^[120]提出了一种遗忘参数极值学习机(Forgetting Parameters ELM, FP-ELM),其主要机制是,在学习过程中,当新数据到达时,为旧数据分配一个遗忘参数。与其他基于遗忘机制或遗忘因子的 ELM 算法相比,FP-ELM 算法中的遗忘参数根据学习器的实时性进行简单计算,并采用正则化优化方法避免了收敛问题。Silva 等^[121]提出一种基于 ELM 的半监督在线算法(SSOE-ELM);针对概念漂移问题,提出了一种改进的 SSOE-ELM 算法 SSOE-FP-ELM。SSOE-FP-ELM 采用半监督概念漂移检测器和遗忘参数。

4.1.5 其他

Pratama 等^[122]针对现实数据流中的概念漂移和不确定性问题,提出了一种演变的 Type-2 模糊神经网络(Evolving Type-2 Recurrent Fuzzy Neural Network, eT2RFNN)。eT2RFNN 提出了一种完全灵活且计算效率高的工作原理,可以对模糊规则进行自适应、增长、剪枝和合并。它具有在线特征选择技术,能够很好地解决高维问题。eT2RFNN 采用一种新的递归网络结构,具有双递归层。作为第三代人工神经网络的主要代表之一,尖峰神经网络(SNNs)天生就适合轻松快速地适应不断变化的环境。Lobo 等^[123]提出一种新的尖峰神经网络的在线学习情景下的概念漂移。通过限制神经元库的大小,将传统的 eSNN 技术应用于在线数据流,得到了所谓的 OeSNN,并采用选择性和生成性数据简化技术来优化神经元存储库的内容,从而使模型更好地适应处理数据流中不断变化的概念。Andrade 等^[124]提出一种用于聚类数据流的快速进化算法(Fast Evolutionary Algorithm for Clustering data streams, FEAC),该算法使用 K-均值聚类,从数据流中自动估计 k 。FEAC 使用 Page-Hinkley 检验来识别聚类质量的下降,并用进化算法重新估计 k 。它认为部分未知数据可以提供有效的流知识。

K-最近邻(KNN)是另一种适合处理概念漂移的学习算法,它不需要在训练期间做任何工作,而是使用整个数据集来预测测试示例的类标签^[125]。Losing 等^[126]提出了自适应内存模型的 K 最近邻算法,为 KNN 构成了一个经验证的分类器内的流设置。SAM-KNN 利用生物启发记忆模型及其协调来处理异构概念漂移,即不同的漂移类型和速率。其基本思想是为当前和以前的概念建立专门的模型,并根据特定情况的需要加以应用。Tennant 等^[127]提出了微聚类最近邻(Micro-Cluster Nearest Neighbour, MC-NN)数据流分类器。MC-NN 基于最近邻分类的数据和近邻的统计摘要,统计摘要由一组基于方差的微簇(MCs)构成,微聚类通过吸收新的数据实例不断适应概念漂移。

4.2 集成学习

集成学习方法被认为是目前处理数据流中的概念漂移非常有效的方法之一。依据对数据流中数据的处理方式,被动的自适应集成方法可以大致分为基于块的集成方法和在线集成方法。

4.2.1 块集成方法

处理具有概念漂移的非稳态环境的基于块的集成方法通常是在训练样本数据中以固定大小或者可变大小的数据块的形式来进行处理,该方法将在新的数据块中调整或创建新的基分类器来适应概念漂移。

Street 等^[16]提出的 SEA 算法(Streaming Ensemble Algorithm)是最早的基于块的集成方法之一。SEA 使用新生成的基分类器替换分类中性能最弱的基分类器,同时所有数据的基础上构建一个决策树,可快速调整以适应突变和渐变漂移。Wang 等^[17]提出了一种精度加权集成(AWE)的算法。AWE 根据在最新训练块上的预测误差,将权重分配给集成的每个分类器。AWE 可以去掉会阻碍预测的分类器,包括可以学习新概念的新分类器。其分类器的权重如式(10)所示:

$$\omega_i = MSE_r - MSE_i \quad (10)$$

其中, MSE_i 与 MSE_r 分别表示分类器 C_i 的均方误差与随机预测的均方误差,其计算式分别如式(11)、式(12)所示:

$$MSE_i = \frac{1}{|S_n|} \sum_{(x,c) \in S_n} (1 - f_i^c(x))^2 \quad (11)$$

$$MSE_r = \sum_c p(c)(1 - p(c))^2 \quad (12)$$

在 AWE 的基础上,Brzezinski 等^[128]提出的精度更新集成(Accuracy Updated Ensemble, AUE)则是另一种基于块的集成方法。在这个集成中,所有的组件分类器都用新块中的一部分示例进行增量更新,从而有助于减少块较小导致的与创建较差的基分类器相关的问题。在 AUE 的基础上,Brzezinski 等^[129]提出了一种新的加权和更新机制(Accuracy Updated Ensemble, AUE2)。与 AUE 相比,AUE2 引入了一个新的权重函数,不需要对候选分类器进行交叉验证,不保留分类器缓冲区,删除其基学习器。Polikar 等^[130]提出另一种经典的被动自适应集成的概念漂移增量学习方法(Learn++-NSE),该方法以非平稳环境(NSEs)为特征,其数据分布随时间变化。它不需要对漂移的性质或速率作任何假设就可以从连续的数据批中学习,并且可以从经历恒定或可变漂移速率、概念类的添加或删除以及周期性漂移的环境中学习。

最近,Li 等^[131]提出一种动态更新集成的增量集成算法(Dynamic Updated Ensemble, DUE)。DUE 为每个块创建几个候选分类器,并使用分段加权方法对其加权。每个候选组件的训练集通过类似 Bagging 的方法进行平衡,之前的组件会定期更新,以使集成对不同种类的概念漂移做出反应。Klikowski 等^[132]提出了一种多采样随机子空间集成方法(Multi Sampling Random Subspace Ensemble, MSRS),该算法采用随机子空间方法并使用各种采样方法来平衡数据,以确保分类器集成的适当多样性,并通过使用各种过采样技术来确保多样性。Wozniak 等^[133]提出一种新的基于滑动窗口的方法,该方法允许实现遗忘机制,即在分类器更新过程中不考虑来自过时模型的旧对象;另一方面假设只有部分到达的样本可以被标记,因为假设标签预算有限;同时采用主动学习范式来选择一个“有趣”的对象进行标记。Jackowski 等^[134]引入了两种新的误差趋势多样性度量方法来比较分类器在处理后续样本时的误差趋势,目的是检测对数据流波动反应的

差异,这使得它们特别适用于具有概念漂移的问题。此外,算法能够自动控制集成大小,使其保持相对较小,并根据分类问题和数据波动的特点进行调整。Bertini 等^[135]提出了一种迭代 Boosting 流集成的方法 (Iterative Boosting Streaming Ensemble, IBS)。IBS 依赖于将 Boosting 应用到新的数据块中,其更新机制增强了模型的灵活性,从而快速降低概念漂移导致的高错误率。Ren 等^[136]提出渐进重采样集成 (Gradual Resampling Ensemble, GRE) 来处理概念漂移和类不平衡的数据流。GRE 采用一种可避免漂移数据的选择性重采样方法,从以前的少数样本中选取一部分样本,对当前的少数样本集进行放大;同时使用最新实例更新以前的组件分类器,无论概念漂移的类型如何都能快速适应。

在出现概念漂移时,仅基于分类精度的集成分量选择并不是最优解。因此引入一个额外的度量来确保组件的多样性,对集成算法在非平稳环境中的性能有积极的影响。Duda^[137]研究了数据流分类集成中的组件选择问题,提出了 ASE-GD (Automatically Sized Ensemble for Gradual Drift), 该算法在处理渐变漂移时,当海林格距离的预测结果与集成的其他部分有显著差异,则可以应用海林格距离来添加新的组件。基于集成的处理时间和内存的瓶颈, Bertini 等^[138]提出一种快速的基学习算法,算法将每个属性范围离散为不相交的区间,同时维持一个静态的集成 DISCO 在不替换基学习器的情况下处理概念漂移。

4.2.2 在线集成方法

在线集成方法会分别学习传入的每个训练样本,而不是分块学习。在线集成方法能够一次学习数据流,相比基于块的方法,其需要的内存与时间更少。该方法还避免了需要在基于块的集成中选择合适的的数据块大小。

Kolter 等提出的动态加权多数 (DWM)^[18]方法是最著名的在线集成方法之一。在 DWM 中,每个分类器做出错误的预测时,其权重都会减小,并乘以常数 $\beta(0 \leq \beta < 1)$ 。为了加快对概念漂移的反应,DWM 添加新的分类器或删除现有的分类器。当集成对给定的训练示例进行错误分类时,将添加新的分类器。最近, Bach 等^[139]利用这两个学习器之间的相互作用及其在准确性上的差异来应对概念漂移,将反应型学习器作为漂移的指标,使用稳定型学习器进行预测。在线精度更新集成 (Online Accuracy Updated Ensemble, OAUE)^[140]继承了 AUE 的一些解决方案,比如基分类器的增量更新和在某些时间步学习新的分类器。然而,为了更有效地处理传入的单个示例和权重分量分类器,引入了一种新的具有成本效益的函数。其在基分类器上的权重 ω_j 如式 (13) 所示:

$$\omega_j = \frac{1}{MSE_j + MSE_c + \epsilon} \quad (13)$$

其中, MSE_j 是基分类器, C_i 是对最后 d 个样本的预测误差, MSE_c 是随机预测基分类器的均方误差, ϵ 则表示当前集成内的基分类器数量。

Jaber 等^[141]提出了一种新的二阶学习机制 (Anticipative Dynamic Adaptation to Concept Change, ADACC), 该机制能够检测环境的相关状态,以识别重复出现的上下文并主动应对概念漂移,但是并未采用任何漂移检测器。该方法通过使用

一个存储最佳候选分类器的列表来识别重复出现的概念并实现预测。Krawczyk 等^[142]提出使用一个确定性阈值来确定哪些集成成员可以参与投票,同时引入一个动态自适应的阈值来监视集成的输出,并允许利用基础多样性来有效地预测漂移。Zhang 等^[143]提出一种基于类不平衡数据流的在线主动学习集成对 (Online Active Learning paired ensemble for Drifting streams with Class Imbalance, OAL-DI)。该方法提出了一种新的混合标记策略,能有效地处理不同类别不平衡率下的概念漂移。Sidhu 等^[144]提出的多样化动态加权多数的在线集成方法 (DDWM) 则保持了两组在多样性水平上不同的加权集成。

考虑投资回报率 (ROI) 问题, Olorunnimbe 等^[145]提出了自适应集成大小算法 (Adaptive Ensemble Size, AES), AES 根据 ROI 使用模式动态调整集成的大小。结果表明,在不影响预测精度的前提下, AES 算法具有较高的投资回报率 (ROI) 和快速适应概念漂移的能力。Ghomeshi 等^[146]提出了一种新的集成技术以适应非平稳数据流中不同的概念漂移。基于 3 层体系结构来产生不同大小的分类类型,一种进化算法即复制子动力学 (RD) 被用来无缝地适应不同的概念漂移。此外,所有分类类型中所选择的特征组合分别使用粒子群优化 (PSO) 技术的非规范版本对每个层进行优化。Shan^[147]提出了一种新的基于混合标记策略的在线主动学习集成框架,该框架将改进的不确定性策略和随机策略这两种不同的主动学习策略相结合,选择具有代表性的实例进行标注。

4.2.3 混合集成方法

Lu 等^[148]提出了基于块的自适应动态加权多数集成方法 (Adaptive Chunk-Based Dynamic Weighted Majority, ACDWM), ACDWM 根据分类器在当前数据块上的分类性能对各个分类器进行动态加权。ACDWM 为每个块创建一个单独的分类器,并根据它们在当前块上的性能对其进行加权。因此,最近训练的分类器或基于与当前块相似的概念训练的分类器将在集成中接收高权重以帮助预测。Ren 等^[132]提出了一种混合模型 KME, 其结合了在线集成和基于块的集成机制来应对各种类型的概念漂移。在平稳期, KME 是以逐块的方式工作的,而概念漂移检测系统是在窗口级对一系列数据进行检测。Cano 等^[149]提出一种 Kappa 更新集成算法 (Kappa Updated Ensemble, KUE)。KUE 结合了在线集成和基于块的集成方法,使用 Kappa 统计量对基分类器进行动态加权和选择。每个基学习器使用不同的特征子集进行训练,并按照泊松分布用给定概率的新实例进行更新。KUE 中的每个基分类器都可在投票中弃权,从而提高 KUE 的整体鲁棒性。

4.3 小结

在以被动方式处理概念漂移的方法中,集成学习方法因表现出优于单学习器的泛化能力而受到广泛关注。集成学习方法依赖于两个基本点即多样性和自适应性来处理概念漂移,保证了基学习器的良好组合,成为处理概念漂移数据流最有前途的方法之一。其中, Lu 等提出的基于自适应块的动态加权多数集成 (ACDWM)^[148]在 HyperPlane, Spiral 和 Checkerboard 等数据集上的实验证明了其权重调整机制可以很好地应对混合类型概念漂移。Cano 等提出的 Kappa 更新集成

算法(KUE)^[149]结合了在线集成和基于块的集成方法,使用Kappa统计量对基分类器进行动态加权和选择,对不同的特征子集进行训练以获得更高的多样性。实验表明,KUE在标准和不平衡漂移数据流上的性能优于最先进的集成,

同时计算复杂度较低。

集成学习方法凭借其动态组合机制、训练数据更新机制、集成成员更新机制和集成结构变化机制等成为了适应概念漂移的最好的方法之一。表2列出了基于被动自适应方法的算法。

表2 被动方法总结

Table 2 Summary of passive methods

算法	对比算法	学习模型	优点	缺点
CVFDT ^[14]	VFDT	决策树	在处理大数据流以及频繁的概念变化时保持很好的效果	丢弃了已经过时但是之后可能会重用的子树
CEVOT ^[101]	HAT, HT, NB, AUE	决策树	将树学习算法和演化算法相结合,逐步学习决策树	由于演化算法非常耗时,CEVOT运行时间较长
WNB-CD ^[104]	NB, HT, DHT	贝叶斯	在保持非常低的时间和内存复杂性的同时,在统计分方面超越了所有对比算法	没有遗忘机制,不能自适应任意类型概念漂移
RGNBC ^[107]	MReC-DFS	贝叶斯	在大数据集上对RGNBC与现有的MReC-DFS算法进行了灵敏度,特异性和准确性的比较,RGNBC的最大精度达到74.5%	考虑除了贝叶斯模型外的学习算法
PGNBC ^[108]	MReC-DFS, RGNBC	贝叶斯	在各个数据集上,PNGBC的灵敏度,精度等都优于RGNBC	较依赖于数据块的大小
DELM ^[119]	SAELMs, BagELMs, EOS-ELM, WELM等	极限学习机	可以在精度和时间开销之间取得更好的平衡,DELM更适合于数据流分类	当概念漂移速度过快时,DELM算法的精度并不令人满意
SSOE-ELM ^[121]	FP-ELM	极限学习机	在两种不同概念漂移类型下的精度优于其他算法,且训练时间没有明显增加	SSOE-FP-ELM的训练时间稍慢
eT2RFNN ^[122]	FAOSPFNN, eT2Class, PANFIS, GENEFIS等	神经网络	能够很好地解决高维问题,在人工和真实数据流上有更可靠的预测精度,同时保持了较低的复杂性	下一步工作采用元认知框架学习算法。
OeSNN ^[123]	OeSNN-PS, OeSNN-PG	神经网络	所需计算资源相对较少,处理时间和内存使用在处理高流数据效率较低	没有考虑漂移的速度,严重性等先验信息
AWE ^[17]	ensemble naive Bayesian, ensemble RIPPER等	基于块的集成	提出了分类器集成方法新的解决路径,相比单分类器提供了更准确的性能	只能适应潜在的概念漂移和少量数据
SEA ^[16]	RTS, RTC, SEA1, SEA2, SEA3, AGRAWAL-1等	基于块的集成	一种新颖的集成方法,比单分类器更有效	旧的分类器可能超越新的分类器,从而有可能减慢对新概念的适应
Learn++ NSE ^[130]	DWM, SEA, AdaBoost	基于块的集成	Learn++ NSE可以快速地适应变化的环境,而不考虑概念漂移的类型。	计算效率有待提升
AUE2 ^[128]	ACE, AUE, DDM, OZA, DWM, NSE等	基于块的集成	适用于涉及多种漂移和静态环境的场景;提供了最佳的平均分类精度,实验证明比其他整体方法消耗的内存更少	没有全部采用增量学习方式
DUE ^[131]	AWE, UB, SERA, CDS, OOB, UOB	基于块的集成	一次只学习一个块而不需要访问以前的数据;能够及时地对多种概念漂移做出反应;在学习非平稳不平衡数据流方面的有效性	不能处理具有倾斜类分布和复杂数据分布的多标签数据流
GRE ^[136]	AWE, SERA, MuSeRA, SMOTE, UB, OOB	基于块的集成	相比其他算法,GRE算法在不牺牲大多数类性能的前提下,能够在少数类上保持良好的性能	考虑解决多标签问题
DWM ^[18]	DWM-ITI, DWM-NB	在线集成	维持了相当数量的基学习器,但实现了更高的预测准确度	从概念漂移中恢复的时间较长
OAUE ^[140]	ACE, DWM, Lev Bag, OAUE	在线集成	无论漂移存在与否或类型如何,都能在线环境下提供较高的分类精度;在所有对比算法中,提供了最好的平均分类精度,并且是时间和内存消耗最少的算法之一	较大的窗口会导致较高的时间和内存开销
OAL-DI ^[143]	OB, OOB	在线集成	能以更低的标记成本获得更高的精度值	某些情况下需要更高的标签成本
DDWM ^[144]	EDDM, DWM DDWM, NB, WM	在线集成	能较好地处理超大数据集中的突变漂移	处理突变漂移的周期值和基分类器的最大数量的值可能影响运行时间和内存
RED-PSO ^[146]	ARF, DWM, Lev-Bag, OAUE等	在线集成	在即时和延迟预测评估环境下性能较好	当目标数据流具有大量特征时,其总体评估时间较长
ACDWM ^[148]	UB, REA, Learn++ NSE, OOB, DDM-OCI等	块集成与在线集成结合	能快速适应新的概念;不需要存储以前的数据;存储的分类器数量有限,以确保高效率;自适应地选择概念漂移环境中的块大小	计算代价较高
KME ^[32]	KME, ADACC, NB, VFDT, AWE等	块集成与在线集成结合	可以避免半监督方法的标注成本,及时补充监督信息;对随机斑点具有鲁棒性	KME中的漂移检测系统不能处理具有标称属性的数据集中发生的突变漂移
KUE ^[149]	LNSE, DWM, ADACC, SAE2	块集成与在线集成结合	在标准和不平衡漂移数据流上的性能优于最先进的集成,同时计算复杂度较低	有待研究多标签问题

5 未来研究方向

基于主动检测方法和被动自适应方法来处理概念漂移的

研究有很多。但是目前仍然存在很多未能解决的概念漂移的相关挑战,如类不平衡的数据流概念漂移处理方法、概念演化数据流处理方法、含噪声的数据流概念漂移处理方法。

(1)类不平衡的数据流概念漂移处理方法

除概念漂移的问题外,非稳态数据流还可能受到其他数据复杂性因素的影响,特别是涉及类不平衡的问题,即少数类代表性不足的挑战。传统分类器的决策边界往往倾向于少数样本,这必然会导致对低代表性数据的预测精度很低^[133]。类不平衡存在于许多实际应用中,如网络入侵检测和信用卡交易。以往大多数的数据流分类模型并未关注不平衡数据流中的概念漂移问题,大多数概念漂移处理方法是基于数据流平衡的假设而设计的^[142]。虽然已经有了很多研究概念漂移和类不平衡问题的算法,但这两者共存的问题大多没有得到充分的探讨。解决概念漂移和类不平衡的联合问题要求模型能够及时修改自身,并保存有价值的信息,即存储少数实例,以便对当前表示不足的数据进行过采样,最后在不牺牲多数样本的预测精度的条件下,平衡少数和多数情况下模型的预测精度。同时在重采样过程中,要考虑离群点、小的间断点等因素以避免类重叠^[130]。更复杂的情况是,不平衡比率和少数类的概念可能会随着时间的推移而改变。因此下一步的工作旨在处理同时存在类不平衡和概念漂移的非稳态数据流。

(2)含新颖类的数据流概念漂移处理方法

在现实中,随着时间的推移,数据流中会出现离群点和新颖类的现象即概念演化,这将导致传统数据流分类算法的精度逐渐下降,因此学习算法必须准备好处理新的、看不见的数据,而这些数据不遵循先前的分布^[84]。在一些实际的真实世界应用中,可以合理地假设类的数量不是一直不变,这种情况发生在新样本被视为新颖类时。因此,为了处理出现新类的分类情况,模型必须具有识别变化和检测数据流中新颖类的能力以适应概念演变。这样的情况可能是由数据流中的噪声引起的,或者实际上可能源于开始出现的新概念,因此必须同时处理多个类之间的识别问题和可能出现的新类的检测。目前关于概念演化(即新颖类检测)的研究大致可分为两类,即基于模型的学习方法和基于聚类的方法^[150]。基于模型的方法致力于寻找合适的模型来识别新类;而基于聚类的方法则依赖于聚类,将传入的实例与相关的不相交聚类相关联以识别新颖类。下一步工作旨在在数据流环境中对新颖类进行更高效的检测。

(3)含噪声的数据流概念漂移处理方法

数据流中包含的噪声数据同样是数据流挖掘中的挑战与难题。数据流中的噪声通常指罕见和不常见的数据项。数据流中的噪声会影响学习算法的准确性并且容易与概念漂移混淆。噪声不应被视为概念漂移,因为它代表不显著的随机波动。因此处理概念漂移时需要将噪声保持鲁棒性,能够将噪声与概念漂移区分开来的同时很好地适应概念漂移。下一步工作旨在在存在噪声的概念漂移数据流时保持鲁棒性。

结束语 本文对以往处理数据流中概念漂移问题的算法从基本概念、使用技术、算法流程、算法优缺点这几个方面进行了综述,并从主动检测方法和被动自适应方法两个方面对处理概念漂移的技术方法进行了详细的讨论,其中分别从处理单一类型概念漂移和处理多种类型概念漂移的角度分析了主动检测方法,并从单学习器和集成学习的角度对被动自适应方法进行了详细的阐述与分析,并对算法的对比算法、学习

模型、适用漂移类型、优缺点等方面进行了总结。最后,基于现有处理概念漂移技术方法所面临的挑战,提出了类不平衡的数据流概念漂移处理方法、含新颖类的数据流概念漂移处理方法和含噪声的数据流概念漂移处理方法的下一步研究工作。

参 考 文 献

- [1] DONGRE P B, MALIK L G. A review on real time data stream classification and adapting to various concept drift scenarios [C]//Proceedings of 2014 IEEE International Advance Computing Conference(IACC). Gurgaon; IEEE, 2014: 533-537.
- [2] MASSI MC, IEVA F, LETTIERI E. Data mining application to healthcare fraud detection: a two-step unsupervised clustering method for outlier detection with administrative databases[J]. BMC Medical Informatics and Decision Making, 2020, 20(1): 1-11.
- [3] FDEZ-RIVEROLA F, IGLESIAS E, DIAZ F, et al. Applying lazy learning algorithms to tackle concept drift in spam filtering [J]. Expert Systems with Applications, 2007, 33(1): 36-48.
- [4] AHMIM A, GHOULMI-ZINE N. A new adaptive intrusion detection system based on the intersection of two different classifiers [J]. International Journal of Security and Networks, 2014, 9(3): 125-132.
- [5] DEMERTZIS K, ILIADIS L, ANEZAKIS V. A Dynamic Ensemble Learning Framework for Data Stream Analysis and Real-Time Threat Detection [C]//Proceedings of Artificial Neural Networks and Machine Learning (ICANN). Greece: Springer, 2018: 669-681.
- [6] DASH R, SAMAL S, DASH R, et al. An integrated TOPSIS crow search based classifier ensemble: In application to stock index price movement prediction [J]. Applied Soft Computing, 2019, 85: 105784.
- [7] DU S Y, HAN M, SHEN M Y, et al. Survey of Ensemble Classification Algorithms for Data Streams with Concept Drift [J]. Computer Engineering, 2020, 46(1): 15-24.
- [8] IWASHITA A S, PAPA J P. An Overview on Concept Drift Learning [J]. IEEE Access, 2018, 7: 1532-1547.
- [9] GAMA J, MEDAS P. Learning with Drift Detection [J]. Advances in Artificial Intelligence - SBIA, 2004, 3171: 286-295.
- [10] BAENA-GARC M, CAMPO-ÁVILA J D, FIDALGO-MERINO R, et al. Early drift detection method [J]. International Workshop on Knowledge Discovery from Data Streams, 2006, 6: 77-86.
- [11] BIFET A, GAVALDÁ R. Learning from Time-Changing Data with Adaptive Windowing [C]//Proceedings of the Seventh SIAM International Conference on Data Mining. Minneapolis: SIAM, 2007: 443-448.
- [12] NISHIDA K, YAMAUCHI K, OMORI T. ACE: Adaptive Classifiers-Ensemble System for Concept-Drifting Environments [C]//Proceedings of International Conference on Multiple Classifier Systems. Verlag: Springer, 2005, 176-185.
- [13] GOMES H M, BIFET A, READ J, et al. Adaptive random forests for evolving data stream classification [J]. Machine Lear-

- ning, 2017, 106(9/10):1469-1495.
- [14] HULTEN G, SPENCER L, DOMINGOS P. Mining Time-Changing Data Streams[C]//Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2001:97-106.
- [15] LIANG N, HUANG G, SARATCHANDRAN P, et al. A Fast and Accurate Online Sequential Learning Algorithm for Feed-forward Networks[J]. IEEE transactions on neural networks, 2006, 17(6):1411-1423.
- [16] STREET W N, KIM Y S. A streaming ensemble algorithm (SEA) for large-scale classification[C]//Proceedings of the Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2001:377-382.
- [17] WANG H, YU P S, HAN J. Mining Concept-Drifting Data Streams[M]//Data Mining & Knowledge Discovery Handbook. 2003:789-802.
- [18] KOLTER J Z, MALOOF M A. Dynamic weighted majority: a new ensemble method for tracking concept drift[C]//Proceedings of the Third International IEEE Conference on Data Mining. Los Alamitos: IEEE Computer Society, 2003:123-130.
- [19] WARES S, ISAACS J, ELYAN E. Data stream mining: methods and challenges for handling concept drift [J]. SN Applied Sciences, 2019, 1(11):1412.
- [20] RAMIREZ-GALLEGO S, KRAWCZYK B, GARCIA S, et al. A survey on data preprocessing for data stream mining: Current status and future directions [J]. Neurocomputing, 2017, 239:39-57.
- [21] GAMA J, LIOBAIT I, BIFET A, et al. A Survey on Concept Drift Adaptation[J]. ACM Computing Surveys, 2014, 46(4):4410-4437.
- [22] KHAMASSI I, SAYED-MOUCHAWEH M, HAMMAMI M, et al. Discussion and review on evolving data streams and concept drift adapting[J]. Evolving Systems, 2016, 9(1):1-23.
- [23] LU J, LIU A, DONG F, et al. Learning under Concept Drift: A Review[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 31(12):2346-2363.
- [24] MINKU L, WHITE A, XIN Y. The Impact of Diversity on Online Ensemble Learning in the Presence of Concept Drift[J]. IEEE, 2010, 22(5):730-742.
- [25] TSYMBAL A. The Problem of Concept Drift: Definitions and Related Work[J/OL]. Journal of Information Security. <https://www.researchgate.net/publication/228723141>.
- [26] DELANY S J, CUNNINGHAM P. A case-based technique for tracking concept drift in spam filtering [J]. Knowledge-Based Systems, 2004, 18(4/5):187-195.
- [27] KRAWCZYK B, MINKU L L, GAMA J, et al. Ensemble learning for data stream analysis: A survey[J]. Information Fusion, 2017, 37:132-156.
- [28] OLIVEIRA G H F M, MINKU L L, OLIVEIRA A L I. GMMVRD: A Gaussian Mixture Model for Dealing With Virtual and Real Concept Drifts [C]//Proceedings of International Joint Conference on Neural Networks (IJCNN). Budapest: IEEE, 2019:1-8.
- [29] TSYMBAL A, PECHENIZKIY M, CUNNINGHAM P, et al. Handling Local Concept Drift with Dynamic Integration of Classifiers: Domain of Antibiotic Resistance in Nosocomial Infections [C]//Proceedings of 19th IEEE International Symposium on Computer-Based Medical Systems (CBMS). Salt Lake City: IEEE, 2006:679-684.
- [30] KHAMASSI I, SAYED-MOUCHAWEH M, HAMMAMI M, et al. Self-Adaptive Windowing Approach for Handling Complex Concept Drift[J]. Cognitive Computation, 2015, 7(6):772-790.
- [31] ŽLIOBAITE I, BUDKA M, STAHL F. Towards cost-sensitive adaptation: When is it worth updating your predictive model [J]. Neurocomputing, 2015, 150:240-249.
- [32] REN S, LIAO B, ZHU W, et al. Knowledge-maximized ensemble algorithm for different types of concept drift [J]. Information Sciences, 2018, 430/431:261-281.
- [33] HU H, KANTARDZIC M, SETHI T S. No Free Lunch Theorem for concept drift detection in streaming data classification: A review [J/OL]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, <https://doi.org/10.1002/widm.1327>.
- [34] NISHIDA K, YAMAUCHI K. Detecting Concept Drift Using Statistical Testing[M]. Berlin: Springer, 264-269.
- [35] MAHDI O A, PARDEDE E, ALI N, et al. Diversity measure as a new drift detection method in data streaming [J]. Knowledge-Based Systems, 2020, 191:105227.
- [36] HUANG D T J, KOH Y S, DOBBIE G, et al. Detecting Volatility Shift in Data Streams [C]//Proceedings of IEEE International Conference on Data Mining (ICDM). Shenzhen: IEEE, 2014:863-868.
- [37] PESARANGHADER A, VIKTOR H L. Fast Hoeffding Drift Detection Method for Evolving Data Streams [C]//Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2016:96-111.
- [38] PESARANGHADER A, VIKTOR H L. McDiarmid Drift Detection Methods for Evolving Data Streams [C]//Proceedings of 2018 International Joint Conference on Neural Networks. 2018:1-9.
- [39] MINKU L L, YAO X. DDD: A New Ensemble Approach for Dealing with Concept Drift [J]. IEEE transactions on Knowledge and Data Engineering, 2012, 24(4):619-633.
- [40] SANTOS S G T D, GONÇALVES P M, SILVA G D D S, et al. Speeding Up Recovery from Concept Drifts [C]//Proceedings of Machine Learning and Knowledge Discovery in Databases-European Conference. Nancy: Springer, 2014:179-194.
- [41] BARROS R S M D, GARRIDO T, DE CARVALHO SANTOS S, et al. A Boosting-like Online Learning Ensemble [C]//International Joint Conference on Neural Networks (IJCNN). Vancouver: IEEE, 2016:1871-1878.
- [42] SUN Y, WANG Z, LIU H, et al. Online Ensemble Using Adaptive Windowing for Data Streams with Concept Drift [J]. International Journal of Distributed Sensor Networks, 2016, 12(5):4218973.
- [43] FRIAS-BLANCO I, CAMPO-AVILA J D, RAMOS-JIMENEZ G, et al. Online and Non-Parametric Drift Detection Methods Based on Hoeffding Bounds [J]. IEEE Transactions on Know-

- ledge and Data Engineering, 2015, 27(3): 810-823.
- [44] OZA N C. Online Bagging and Boosting[C]// Proceedings of 2005 IEEE International Conference on Systems, Man and Cybernetics. Waikoloa, IEEE, 2005: 2340-2345.
- [45] GAMA J, KOSINA P. Tracking Recurring Concepts with Meta-learners[C]// Proceedings of 14th Portuguese Conference on Artificial Intelligence. Aveiro: Springer, 2009: 423-434.
- [46] GAMA J, KOSINA P. Recurrent concepts in data streams classification[J]. Knowledge and Information Systems, 2014, 40(3): 489-507.
- [47] ANGEL A, JOÃO BARTOLO G, ERNESTINA M. Predicting recurring concepts on data-streams by means of a meta-model and a fuzzy similarity function[J]. Expert Systems with Applications, 2016, 46: 87-105.
- [48] ANDERSON R, KOH Y S, DOBBIE G. CPF: Concept Profiling Framework for Recurring Drifts in Data Streams[M]. Cham: Springer International Publishing, 2016: 203-214.
- [49] ANDERSON R, KOH Y S, DOBBIE G, et al. Recurring Concept Meta-learning for Evolving Data Streams[J]. Expert Systems with Applications, 2019, 138: 112832.
- [50] GONÇALVES J P M, BARROS R S M D. RCD: A recurring concept drift framework[J]. Pattern Recognition Letters, 2013, 34(9): 1018-1025.
- [51] SAKTHITHASAN S, PEARS R, BIFET A, et al. Use of ensembles of Fourier spectra in capturing recurrent concepts in data streams[C]// Proceedings of International Joint Conference on Neural Networks. Killarney: IEEE, 2015: 1-8.
- [52] SRIPRAKAS S, PEARS R. Mining Recurrent Concepts in Data Streams Using the Discrete Fourier Transform[J]. Data Warehousing and Knowledge Discovery, 2014, 8646: 439-451.
- [53] SIDHU P, BHATIA M P S. A two ensemble system to handle concept drifting data streams; recurring dynamic weighted majority[J]. International Journal of Machine Learning and Cybernetics, 2019, 10(3): 563-578.
- [54] GOMES J B, GABER M M, SOUSA P A C, et al. Mining Recurring Concepts in a Dynamic Feature Space[J]. IEEE Transaction on Neural Networks and Learning Systems, 2014, 25(1): 95-110.
- [55] HU J, CHEN J, QIN X. Algorithm of Recurring Concept Drift Base on Main Feature Extraction, 2019[C]// Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence (ICCAD). Bali: ACM, 2019: 59-65.
- [56] ZHENG X, LI P, HU X, et al. Semi-supervised classification on data streams with recurring concept drift and concept evolution [J]. Knowledge-Based Systems, 2021, 215: 106749.
- [57] ALIPPI C, BORACCHI G, ROVERI M. An effective just - in - time adaptive classifier for gradual concept drifts[C]// Proceedings of the 2011 International Joint Conference on Neural Networks (IJCNN). IEEE, 2011: 1675-1682.
- [58] LIU A, ZHANG G, JIE L. Fuzzy Time Windowing for Gradual Concept Drift Adaptation[C]// Proceedings of 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). Naples: IEEE, 2017: 1-6.
- [59] ABDUALRHMAN M A A, PADMA M C. Deterministic Concept Drift Detection in Ensemble Classifier Based Data Stream Classification Process [J]. International Journal of Grid and High Performance Computing, 2019, 11(1): 29-48.
- [60] TSYMBAL A, PECHENIZKIY M, CUNNINGHAM P, et al. Dynamic integration of classifiers for handling concept drift[J]. Information Fusion, 2008, 9(1): 56-68.
- [61] GAMA J, CASTILLO G. Learning with Local Drift Detection [C]// Proceedings of Advanced Data Mining and Applications, Second International Conference (ADMA). Xi'an: Springer, 2006: 42-55.
- [62] IKONOMOVSKA E, GAMA J, SEBASTIAO R, et al. Regression trees from data streams with drift detection[C]// Proceedings of Discovery Science, 12th International Conference. Porto: Springer, 2009: 121-135.
- [63] SHAKER A, LUGHOFFER E. Self-adaptive and local strategies for a smooth treatment of drifts in data streams[J]. Evolving-Systems, 2014, 5(4): 239-257.
- [64] LIU A, SONG F, ZHANG G, et al. Regional Concept Drift Detection and Density Synchronized Drift Adaptation[C]// Proceedings of Twenty-Sixth International Joint Conference on Artificial Intelligence. 2017: 2280-2286.
- [65] LIU A J, LU J. Accumulating regional density dissimilarity for concept drift detection in data streams[J]. Pattern Recognition, 2018(76): 256-272.
- [66] LIU A, LU J, ZHANG G. Diverse Instance-Weighting Ensemble Based on Region Drift Disagreement for Concept Drift Adaptation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(1): 293-307.
- [67] OLIVEIRA L S, BATISTA G E. Igmm-cd: a gaussian mixture classification algorithm for data streams with concept drifts [C]// Proceedings of BRACIS, 2015 Brazilian Conference on Intelligent Systems (BRACIS). IEEE, 2015: 55-61.
- [68] ALMEIDA P R, OLIVEIRA L S, BRITTO A S, et al. Adapting dynamic classifier selection for concept drift[J]. Expert Systems with Applications, 2018, 104: 67-85.
- [69] GZÜAK M, CAN F. Concept learning using one-class classifiers for implicit drift detection in evolving data streams[J]. Artificial Intelligence Review, 2020(3): 1-23.
- [70] YAN M M W. Accurate detecting concept drift in evolving data streams[J]. ICT Express, 2020, 6(4): 332-338.
- [71] MICEVSKA S, AWAD A, SAKR S. SDDM: an interpretable statistical concept drift detection method for data streams[J]. Journal of Intelligent Information Systems, 2021, 56: 459-484.
- [72] MUSEBA T, NELWAMONDO F, OUAHADA K, et al. Recurrent Adaptive Classifier Ensemble for Handling Recurring Concept Drifts [J]. Applied Computational Intelligence and Soft Computing, 2021, 13: 5533777.
- [73] GOEL K, BATRA S. Dynamically updated diversified ensemble-based approach for handling concept drift[J]. Turkish Journal of Electrical Engineering & Computer Sciences, 2020, 28(1): 556-574.
- [74] REIS D, FLACH P, MATWIN S, et al. Fast Unsupervised Online Drift Detection Using Incremental Kolmogorov-Smirnov Test[C]// Proceedings of the 22nd ACM SIGKDD International

- Conference on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016: 1545-1554.
- [75] VERDECIA-CABRERA A, BLANCO I F, CARVALHO A C P L. An online adaptive classifier ensemble for mining non-stationary data streams[J]. *Intelligent Data Analysis*, 2018, 22(4): 787-806.
- [76] DONG F, LU J, ZHANG G, et al. Active Fuzzy Weighting Ensemble for Dealing with Concept Drift[J]. *International Journal of Computational Intelligence Systems*, 2018, 11(1): 438.
- [77] ZHANG B, CHEN Y, XUE L. Research on Concept-Drifting Data Stream Based on Fuzzy Integral Ensemble Classifier System[C]// *Proceedings of Communications, Signal Processing, and Systems*. Singapore: Springer, 2019: 225-232.
- [78] MONTIEL J, MITCHELL R, FRANK E, et al. Adaptive XG-Boost for Evolving Data Streams[C]// *Proceedings of 2020 International Joint Conference on Neural Networks (IJCNN)*. Glasgow: IEEE, 2020: 1-8.
- [79] ABDULSALAM H, SKILLICORN D B, MARTIN P. Streaming Random Forests[C]// *Proceedings of Eleventh International Database Engineering and Applications Symposium (IDEAS 2007)*. Banff: IEEE Computer Society, 2007: 22-36.
- [80] ABDULSALAM H, SKILLICORN D B, MARTIN P. Classifying Evolving Data Streams Using Dynamic Streaming Random Forests[C]// *Proceedings of Database and Expert Systems Applications*, 19th International Conference. Turin: Springer, 2008: 643-651.
- [81] GOMES H M, BIFET A, READ J, et al. Adaptive random forests for evolving data stream classification[J]. *Machine Learning*, 2017, 106(9/10): 1469-1495.
- [82] HO T K. The random subspace method for constructing decision forests[J]. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 1998, 20(8): 832-844.
- [83] LUONG A V, NGUYEN T T, LIEW A W. Streaming Active Deep Forest for Evolving Data Stream Classification[J]. *arXiv*: 2002.11816, 2020.
- [84] ZHOU Z H, FENG J. Deep Forest: Towards An Alternative to Deep Neural Networks[C]// *Proceedings of Twenty-Sixth International Joint Conference on Artificial Intelligence*. Melbourne: ijcai, 2017: 3553-3559.
- [85] HAQUE A, KHAN L, BARON M, et al. SAND: semi-supervised adaptive novel class detection and classification over data stream[C]// *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. Phoenix: AAAI Press, 2016: 1652-1658.
- [86] HAQUE A, KHAN L, BARON M, et al. Efficient handling of concept drift and concept evolution over Stream Data[C]// *Proceedings of Efficient Handling of Concept Drift and Concept Evolution Over Stream Data*. Helsinki: IEEE, 2016: 481-492.
- [87] ZLIOBAITE I. Change with Delayed Labeling: When is it Detectable? [C]// *Proceedings of The 10th IEEE International Conference on Data Mining Workshops*. Sydney: IEEE, 2010: 843-850.
- [88] RAAB C, HEUSINGER M, SCHLEIF F. Reactive Soft Prototype Computing for Concept Drift Streams[J]. *Neurocomputing*, 2020, 416: 340-351.
- [89] YUAN Y, WANG Z, WANG W. Unsupervised concept drift detection based on multi-scale slide windows[J]. *Ad Hoc Networks*, 2021, 111: 102325.
- [90] DU L, SONG Q, ZHU L, et al. A Selective Detector Ensemble for Concept Drift Detection[J]. *The Computer Journal*, 2015, 58(3): 457-471.
- [91] HU H, KANTARDZIC M, LYU L. Detecting Different Types of Concept Drifts with Ensemble Framework[C]// *Proceedings of 17th IEEE International Conference on Machine Learning and Applications*. Orlando: IEEE, 2018: 344-350.
- [92] KORYCKI L, KRAWCZYK B. Unsupervised Drift Detector Ensembles for Data Stream Mining[C]// *Proceedings of 2019 IEEE International Conference on Data Science and Advanced Analytics*. Washington: IEEE, 2019: 317-325.
- [93] LI Y, LI D, WANG S, et al. Incremental entropy-based clustering on categorical data streams with concept drift[J]. *Knowledge-Based Systems*, 2014, 59(1): 33-47.
- [94] BAI L, CHENG X, LIANG J, et al. An optimization model for clustering categorical data streams with drifting concepts[J]. *IEEE Trans. Knowl. Data Eng.* 2016, 28(11): 2871-2883.
- [95] KATAKIS I, TSOUMAKAS G, VLAHAVAS I. Tracking recurring contexts using ensemble classifiers; an application to email filtering[J]. *Knowledge and Information Systems*, 2010, 22(3): 371-391.
- [96] XU S, FENG L, LIU S, et al. Self-adaption neighborhood density clustering method for mixed data stream with concept drift[J]. *Engineering Applications of Artificial Intelligence*, 2020, 89: 103451.
- [97] MUSEBA T, NELWAMONDO F, OUAHADA K. ADES: A New Ensemble Diversity-Based Approach for Handling Concept Drift[J]. *Mobile Information Systems*, 2021, 2021(4): 1-17.
- [98] DOMINGOS P, HULTEN G. Mining high-speed data streams, 2000[C]// *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Boston: ACM, 2000: 71-80.
- [99] BLAKE C, NTOUTSI E. Reinforcement Learning Based Decision Tree Induction Over Data Streams with Concept Drifts, 2018[C]// *Proceedings of Reinforcement Learning Based Decision Tree Induction Over Data Streams with Concept Drifts*. Singapore: IEEE, 2018: 328-335.
- [100] JANKOWSKI D, JACKOWSKI K. An Increment Decision Tree Algorithm for Streamed Data[C]// *Proceedings of 2015 IEEE TrustCom/BigDataSE/ISPA*. Helsinki: IEEE, 2015: 199-204.
- [101] JANKOWSKI D, JACKOWSKI K, CYGANEK B. Learning Decision Trees from Data Streams with Concept Drift[J]. *Procedia Computer Science*, 2016, 80: 1682-1691.
- [102] COSTA T, GUILHERME V. Strict Very Fast Decision Tree: a memory conservative algorithm for data stream mining[J]. *Pattern Recognition Letters*, 2018, 116(1): 22-28.
- [103] BIFET A, ZHANG J, FAN W, et al. Extremely Fast Decision Tree Mining for Evolving Data Streams[C]// *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax: ACM, 2017: 1733-1742.
- [104] KRAWCZYK B, WOZNIAK M. Weighted Naive Bayes Classi-

- fier with Forgetting for Drifting Data Streams[C]//Proceedings of 2015 IEEE International Conference on Systems, Man, and Cybernetics(SMC). Hong Kong:IEEE,2015:2147-2152.
- [105]ZHAO Q,KLAUE C,LAI C.Predicting concept drift via dynamic Naive Bayes [C] // Proceedings of 2017 IEEE International Conference on Big Data. Boston:IEEE,2017:2420-2425.
- [106]LIU A,ZHANG G. Fast Switch Naive Bayes to Avoid Redundant Update for Concept Drift Learning[C] // Proceedings of 2020 International Joint Conference on Neural Networks(IJCNN). Glasgow:IEEE,2020:1-7.
- [107]KISHORE B D,RAMADEVI Y,RAMANA K V.RGNBC: Rough Gaussian Naive Bayes Classifier for Data Stream Classification with Recurring Concept Drift[J]. Arabian Journal for Science and Engineering,2016,42(2):705-714.
- [108]BABU D K,RAMADEVI Y,RAMANA K V.PGNBC:Pearson Gaussian Naive Bayes classifier for data stream classification with recurring concept drift[J]. Intelligent Data Analysis,2017,21(5):1173-1191.
- [109]SYED N,LIU H,SUNG K. Handling concept drifts in incremental learning with support vector machines[C]//Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego:ACM,1999:317-321.
- [110]RUPING S.Incremental Learning with Support Vector Machines[C]//Proceedings of the 2001 IEEE International Conference on Data Mining. San Jose:IEEE,2001:641-642.
- [111]YALCIN A,ERDEM Z,GURGEN F. Ensemble based incremental SVM classifiers for changing environments[C] // Proceedings of 2007 22nd International Symposium on Computer and Information Sciences. Ankara:IEEE,2007.
- [112]AYAD O. Learning under Concept Drift with Support Vector Machines[C]//Proceedings of 24th International Conference on Artificial Neural Networks. Hamburg:Springer,2014:587-594.
- [113]ZAREMOODIP,SIAHROUDI S K,BEIGY H. A support vector based approach for classification beyond the learned label space in data streams [C] // Proceedings of the 31st Annual ACM Symposium on Applied Computing. Pisa:ACM,2016:910-915.
- [114]ZHU Q Y,SIEW C K. Extreme Learning Machine:A New Learning Scheme of Feedforward Neural Networks[J]. Neurocomputing,2006,70:489-501.
- [115]LIANG N,HUANG G,SARATCHANDRAN P,et al. A Fast and Accurate Online Sequential Learning Algorithm for Feedforward Networks[J]. IEEE Transactionson Neural Networks,2006,17(6):1411-1423.
- [116]XU S L,B J W A. A fast incremental extreme learning machine algorithm for data streams classification [J]. Expert Systems with Applications,2016,65:332-344.
- [117]ZHAI J,WANG J G. Ensemble online sequential extreme learning machine for large data set classification[C]// Proceedings of 2014 IEEE International Conference on Systems, Man, and Cybernetics(SMC). San Diego:IEEE,2014:2250-2255.
- [118]LAN Y,SOH Y C,HUANG G. Ensemble of online sequential extreme learning machine[J]. Neurocomputing,2009,72(13/14/15):3391-3395.
- [119]XU S,WANG J. Dynamic extreme learning machine for data stream classification[J]. Neurocomputing,2017,238:433-449.
- [120]LIU D,WU Y X,JIANG H. FP-ELM: An online sequential learning algorithm for dealing with concept drift[J]. Neurocomputing,2016,207:322-334.
- [121]DA SILVA C A S,KROHLING R A. Semi-Supervised Online Elastic Extreme Learning Machine with Forgetting Parameter to deal with concept drift in data streams[C]// Proceedings of International Joint Conference on Neural Networks. Budapest:IEEE,2019:1-8.
- [122]PRATAMA M,LU J,LUGHOFER E,et al. An Incremental Learning of Concept Drifts Using Evolving Type-2 Recurrent Fuzzy Neural Networks[J]. IEEE Transactions on Fuzzy Systems,2017,25(5):1175-1192.
- [123]LOBO J L,BAI L,JAVIER D S,et al. Evolving Spiking Neural Networks for online learning over drifting data streams[J]. Neural Networks,2018,108:1-19.
- [124]ANDRADE S J D,HRUSCHKA E R,GAMA J . An evolutionary algorithm for clustering data streams with a variable number of clusters[J]. Expert Systems with Applications,2017,67:228-238.
- [125]BAHRI M,MANIU S,BIFET A. A Sketch-Based Naive Bayes Algorithms for Evolving Data Streams[C]//Proceedings of 2018 IEEE International Conference on Big Data. Seattle:IEEE,2018:604-613.
- [126]LOSING V,HAMMER B,WERSING H. KNN Classifier with Self Adjusting Memory for Heterogeneous Concept Drift,2016 [C] // Proceedings of IEEE 16th International Conference on Data Mining. Barcelona:IEEE,2016:291-300.
- [127]TENNANT M,STAHL F,RANA O,et al. Scalable real-time classification of data streams with concept drift[J]. Future Generation Computer Systems,2017,75:187-199.
- [128]BRZEZINSKI D,STEFANOWSKI J. Accuracy Updated Ensemble for Data Streams with Concept Drift[C]//Proceedings of International Conference on Hybrid Artificial Intelligent Systems. Wroclaw:Springer,2011:153-160.
- [129]BRZEZINSKI D,STEFANOWSKI J. Reacting to Different Types of Concept Drift:The Accuracy Updated Ensemble Algorithm[J]. IEEE Transactions on Neural Networks and Learning Systems,2014,25(1):81-94.
- [130]ELWELL R,POLIKAR R. Incremental Learning of Concept Drift in Nonstationary Environments[J]. IEEE Transactions on Neural Networks,2011,22(10):1517-1531.
- [131]LI Z,HUANG W,XIONG Y,et al. Incremental learning imbalanceddata streams with concept drift:The dynamic updated ensemble algorithm [J]. Knowledge-Based Systems,2020,195:105694.
- [132]KLIKOWSKI J,WOŹNIAK M. Multi Sampling Random Subspace Ensemble for Imbalanced Data Stream Classification[C]// Proceedings of International Conference on Computer Recognition Systems. Polanica-Zdrój:Springer,2019:360-369.
- [133]WOZNIAC M,KSINIOWICZ P,CYGANEK B,et al. Active Learning Classification of Drifted Streaming Data[J]. Procedia

Computer Science, 2016, 80(C):1724-1733.

- [134] JACKOWSKI K. New diversity measure for data stream classification ensembles[J]. Engineering Applications of Artificial Intelligence, 2018, 74:23-34.
- [135] BERTINI J J R, NICOLETTI M D C. An iterative boosting-based ensemble for streaming data classification[J]. Information Fusion, 2019, 45:66-78.
- [136] REN S, LIAO B, ZHU W, et al. The Gradual Resampling Ensemble for mining imbalanced data streams with concept drift [J]. Neurocomputing, 2018, 286:150-166.
- [137] DUDA P. On Ensemble Components Selection in Data Streams Scenario with Gradual Concept-Drift[C]// Proceedings of 17th Artificial Intelligence and Soft Computing International Conference, Zakopane; Springer, 2018:311-320.
- [138] BERTINI J J R. A Discretization-based Ensemble Learning Method for Classification in High-Speed Data Streams [C]// Proceedings of International Joint Conference on Neural Networks. Budapest; IEEE, 2019:1-8.
- [139] BACH S H, MALOOF M A. Paired Learners for Concept Drift [C]// Proceedings of the 8th IEEE International Conference on Data Mining, Pisa; IEEE, 2008:23-32.
- [140] BRZEZINSKI D, STEFANOWSKI J. Combining block-based and online methods in learning ensembles from concept drifting data streams[J]. Information Sciences, 2014, 265:50-67.
- [141] JABER G, CORNUÉJOLS A, TARROUX P. A New On-Line Learning Method for Coping with Recurring Concepts: The ADACC System[C]// Proceedings of Neural Information Processing-20th International Conference, Daegu; Springer, 2013: 595-604.
- [142] KRAWCZYK B, CANO A. Online ensemble learning with abstaining classifiers for drifting and noisy data streams[J]. Applied Soft Computing, 2018, 68:677-692.
- [143] ZHANG H, LIU W, SHAN J, et al. Online Active Learning Paired Ensemble for Concept Drift and Class Imbalance [J]. IEEE Access, 2018, 6:73815-73828.
- [144] SIDHU P, SIDHU P, BHATIA M P S, et al. A novel online ensemble approach to handle concept drifting data streams; diversified dynamic weighted majority[J]. International Journal of Machine Learning and Cybernetics, 2018, 9(1):37-61.
- [145] OLORUNNIMBE M K, VIKTOR H L, PAQUET E. Dynamic adaptation of online ensembles for drifting data streams [J]. Journal of Intelligent Information Systems, 2018, 50 (2): 291-313.
- [146] GHOMESHI H, GABER M M, KOVALCHUK Y. A non-canonical hybrid metaheuristic approach to adaptive data stream classification[J]. Future Generation Computer Systems, 2020, 102:127-139.
- [147] SHAN J, ZHANG H, LIU W, et al. Online Active Learning Ensemble Framework for Drifted Data Streams [J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(2): 486-498.
- [148] YANG L, CHEUNG Y M, YUAN Y T. Adaptive Chunk-Based Dynamic Weighted Majority for Imbalanced Data Streams With Concept Drift [J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(8):2764-2778.
- [149] CANO A, KRAWCZYK B. Kappa Updated Ensemble for Drifting Data Stream Mining [J]. Machine Learning, 2020, 109(1): 175-218.
- [150] SCHOLZ M, KLINKENBERG R. An ensemble classifier for drifting concepts [C]// Proceedings of the Second International Workshop on Knowledge Discovery from Data Streams (IWK-DDS'05), 2005:53-64.



CHEN Zhi-qiang, born in 1998, post-graduate. His main research interests include data stream classification and so on.



HAN Meng, born in 1982, Ph.D, associate professor, graduate supervisor, is a member of China Computer Federation. Her main research interests include data mining and so on.

(责任编辑:杨雪敏)