基于动态样本选择的概念漂移自适应预测方法

代 劲[1,3],李 昊[2,3],王国胤[3]
(1. 重庆邮电大学软件学院,重庆 400065;2. 重庆邮电大学计算机学院,重庆 400065;3. 计算智能重庆市重点实验室,重庆 400065)

摘 要: 概念漂移是影响流数据挖掘性能的重要因素,当前主要通过增量更新或重训练模型进行处理,但对已有知识并未充分利用. 从综合利用全体样本出发,本文构建了一种基于动态样本选择的概念漂移自适应分类方法. 该方法在新样本到来时进行基于局部一致性的漂移检测,在发现漂移发生时去除区域内的噪声样本,当检测到新概念出现时,对历史相似概念进行重用. 最后,对区域内不同类别样本进行多代表点归纳,并同步更新预测模型. 本文在含有不同漂移类型的合成数据集上进行去噪效果验证,并在真实数据集上进行预测任务. 实验结果表明,该方法可以有效去除因概念漂移而形成的漂移噪声,有效提升了预测模型性能,整体预测表现优于流行的概念漂移自适应模型.

关键词: 概念漂移;局部漂移检测;流数据;样本选择;样本去噪;自适应预测
基金项目: 国家自然科学基金(No.61936001,No.62002037);重庆市自然科学基金(No.cstc2021jcyj-msxmX0849,No.cstb2023nscq-LZX0006)
中图分类号: TP311 **文献标识码:** A **文章编号:** 0372-2112(2024)09-3228-12
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230124

Concept Drift Adaptive Prediction Method Based on Dynamic Sample Selection

DAI Jin[1,3], LI Hao[2,3], WANG Guo-yin[3]
(1. *School of Software*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*;
2. *School of Computer Science*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*;
3. *Chongqing Key Laboratory of Computational Intelligence*, *Chongqing* 400065, *China*)

Abstract: Concept drift is an important performance factor in stream data mining, mainly handled by incremental updating or retraining models, but not fully utilizing existing knowledge. This paper proposed an concept drift adaptive prediction method based on dynamic sample selection, starting from the comprehensive use of all samples. The method performs local consistency based drift detection when new samples arrive, removes noisy samples in the region when drift is detected, and reuses historically similar concepts when new concepts are detected. Finally, multi-representative point summarization is performed for different categories of samples in the region, and the prediction model is updated simultaneously. In this paper, the denoising effect is verified on synthetic datasets containing different drift types, and the prediction task is performed on the real dataset. The experimental results show that the method can effectively remove the drift noise due to conceptual drift, which effectively improves the performance of the prediction model. The prediction outperforms the popular concept drift adaptive model.

Key words: concept drift; local drift detection; stream data; sample selection; sample denoisy; adaptive forecast
Foundation Item(s): National Natural Science Foundation of China (No.61936001, No.62002037); Chongqing Natural Science Foundation (No.cstc2021jcyj-msxmX0849, No.cstb2023nscq-LZX0006)

1 引言

现代通信技术的飞速发展、物联网及传感网的广泛应用,流数据已经逐步成为生产生活信息记录的主要方式,并蕴藏着巨大的信息价值. 然而,在非稳态的流数据环境下,数据分布会不断发生变化,这种现象被称作概念漂移. 概念漂移的出现会导致待预测数据与已学习数据不满足独立同分布假设,从而降低模型预测效果. 概念漂移研究已经在许多领域中成为热点,如

收稿日期:2023-02-10;修回日期:2023-07-29;责任编辑:梅志强

社交网络[1]、金融分析[2]、垃圾邮件检测[3]、网络安全[4]以及工业应用[5]等.

概念漂移会使部分旧数据与新数据的映射关系不一致,使这些旧数据相对于新数据成为噪声,因此以这些数据训练的模型无法适用于新数据. 当前的研究主要集中在增量式学习方法上,通过学习新样本的同时减弱旧样本对预测的影响. 然而,概念漂移并非使全部旧数据都成为噪声,它通常只影响一定区域[6],例如网络热词的词义变化只影响部分社群. 因此,通过识别概念漂移影响的区域、只筛除掉失效噪声样本、保留有效样本用于后续训练,可以更高效地利用样本,有效提升概念漂移环境下预测准确率.

为解决上述问题,本文设计了基于动态样本选择的概念漂移自适应预测方法(Concept Drift Adaptive Prediction method based on Dynamic Sample Selection, CDAP-DSS). 该方法首先采用样本集动态维护策略,通过微簇聚类划分区域,并结合概念漂移检测来判断概念漂移发生的区域;其次,从样本池中删除掉受影响的噪声样本,并在新概念出现时重用旧样本以扩充样本池;最终,在动态样本集基础上进行模型学习,实现更准确的预测.

本文主要贡献如下:(1)对概念漂移噪声进行形式化定义,实现了基于局部一致性漂移检测的样本去噪方法,在保留有效数据时去除了无效噪声;(2)设计了基于样本选择的流式数据预测框架,并在多个概念漂移数据集下进行了验证,展示出较高的预测性能.

2　相关工作

在动态环境下预测的挑战在于概念漂移的影响,已有的研究工作主要聚焦于基于漂移检测的处理和基于自适应的漂移处理两个方面,这些方法对于提高预测效果具有重要意义.

2.1　基于漂移检测的处理

基于漂移检测的处理方法通过漂移检测技术来发现数据中发生的概念漂移. 发现数据平稳性的度量指标发生变化时,启动漂移处理策略,删除漂移前的数据并用新数据重新训练模型.

Drift Detection Method[7]是漂移检测中最经典的算法,通过检测总体错误率是否有显著增加,根据错误率变化的置信度激活相应级别的预警处理动作. Frias-Blanco发现大多数学者在进行漂移检测时都会假设一个基本的概率密度函数[8],而真实数据很少符合这些常见的分布函数,为此设计了基于霍夫丁不等式的非参数检测方法. Liu等人设计了基于直方图的空间划分算法EI-KMeans[9],检测数据密度信息的变化,从而实现有效的概念漂移检测. 近年来,一些学者将神经网络应用于概念漂移检测,如Zhe等提出了基于在线极限学习机的漂移检测方法[10],对新旧数据分别训练不同模型,通过对比模型间差异来实现概念漂移检测.

漂移检测通常被用来确定更新模型的时机,但它同样可用于样本选择[11]:即漂移点之前的数据全部为噪声,仅使用漂移点之后的数据作为模型训练样本. 但此做法存在较大问题:漂移检测通常是针对全局进行的,而概念漂移通常只发生在局部. 因此,将所有数据都当作失效数据并不合适,会导致过滤掉具有潜在信息价值的样本,历史样本也未得到充分利用.

2.2　基于自适应的处理

自适应的漂移处理方法通过被动的不断迭代更新模型来适应数据分布的变化,但它对旧概念的遗忘是缓慢的,为了应对突变型的概念漂移,一些方法会采用漂移适应技术来进行旧模型的替换.

Ryan等设计了基于在线集成学习的流式预测模型Learn++NSE[12],通过创建分类器学习新的知识,结合动态更新的投票权重,选择当前和历史分类器进行预测,其创新在于在适时的时候可以重新启用历史分类器. Heitor设计的自适应随机森林算法(Adaptive Random Forest, ARF)[13]为森林中的每棵霍夫丁树设计一个漂移监测器,在监测到预警时训练新树,在监测到警告时将树替换. Beringer等人设计了基于实例的流数据学习方法(SyncStream),动态地维护一组原型,设计由错误驱动的代表样本学习算法,并通过同步启发的约束聚类进一步总结成更小的原型集,同时引入概念漂移检测算法以更及时地适应突变型的概念漂移[14]. Raab等人将矢量量化学习和原型学习结合,解决数据中的缺失类,对频繁漂移的数据效果较好[15].

上述增量式自适应预测方法在遇到概念漂移发生时,会生成新模型替换旧模型,其缺点是旧模型中仍存在对当前以及未来有价值的部分. 虽然采用重用旧模型的机制可以利用一定历史知识,但无法区分出模型的有效部分,旧知识利用程度有限.

3　问题分析

3.1　问题定义

概念漂移是数据分布发生改变的一种现象(如图1所示),通常表现为特征分布的变化、标签分布的变化以及特征与标签之间映射关系的变化. 流数据处理过程中,由于各个时间片内的概念存在差异,在没有先验知识的情况下,很难判断哪一部分数据适合用来作为样本学习[16]. 为准确地描述概念漂移问题,本文首先对相关概念进行定义.

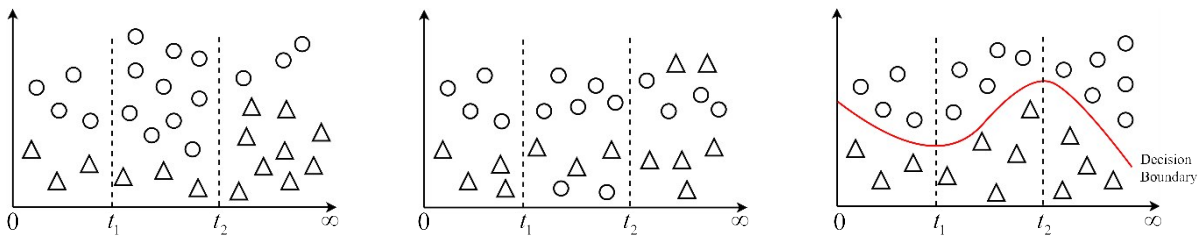
定义1　给定一段时期$[0,t]$,以及相应的数据集$S_{0,t}=\{d_0,\cdots,d_t\}$,其中$d_i=(x_i,y_i)$是一个数据样本,$x$表示

特征，y表示数据标签，数据集$S_{0,t}$服从分布$F_{0,t}(x,y)$. $t+1$时刻$F_{0,t}(x,y)\neq F_{t+1,\infty}(x,y)$，则定义在$t+1$时刻发生了概念漂移，记为$\exists t: F_t(x,y)\neq F_{t+1}(x,y)$[11].

定义 2 给定数据流$S=\{d_0,\cdots,d_t,\cdots\}$，其中$d_i=(x_i,y_i)$，若$S$在$t$时刻发生概念漂移，且概念漂移后的数据分布为$F_{t+1,\infty}(x)=y$. 对于符合$F_{t+1,\infty}(x_i)\neq y_i$的样本，即将与最新数据分布不一致的样本称为漂移噪声.

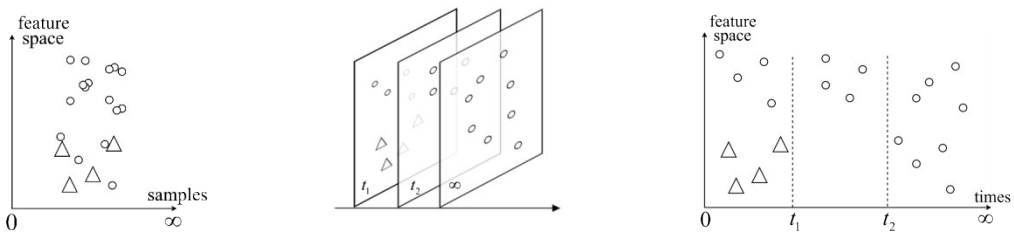
从本质上来说，漂移噪声是指不符合新概念的历史样本，是概念漂移的必然结果. 图2(a)所示区域存在多种类别的样本，其中一些通常会被视为噪声，若对样本的时间片加以区分，则如图2(b)所示，0~t_1时刻的数据由于概念漂移的发生与t_2时刻后的数据分布不一致，则0~t_1时刻的数据相对于t_2之后的数据可以被视为漂移噪声.

定义 3 给定数据流$S=\{d_0,\cdots,d_t,\cdots\}$，其中$d_i=(x_i,y_i)$，在$t$时刻的数据分布为$F_t(X,y)$，则在$t$时刻时，概念漂移下流数据样本选择任务为$S^*=\{(x_i,y_i)|i\leqslant t\wedge(x,y)\in S\wedge F_{t+1,\infty}(x)==y\}$，即选出与$t$时刻数据分布相一致的数据.



(a) 特征漂移 (b) 标签漂移 (c) 映射漂移

图1 概念漂移不同表现形式



(a) 全样本空间 (b) 样本按时间划分 (c) 不同时间下的样本

图2 不同时期下漂移噪声的表现

3.2 局部漂移检测可行性分析

概念漂移使部分旧数据失效成为噪声，需要概念漂移检测识别出这些噪声. 现有漂移检测技术主要针对全局进行检测，但概念漂移的发生通常是局部的[11]，需要能够有效识别漂移区域. 本文从理论层面进行了局部漂移检测的可行性证明.

首先，结合样本数据的映射关系变化，对概念漂移进行形式化定义：

$$\exists t: F_{0,t}(y|X)\neq F_{t,\infty}(y|X) \tag{1}$$

其中$F_{t_1,t_2}(y|X)$表示t_1到t_2时刻内数据的条件分布，概念差异度的定义如下：

$$D=\left|F_{0,t}(y|X)-F_{t,\infty}(y|X)\right| \tag{2}$$

为了便于分析，本文使用CK距离[17]来描述概念差异度：

$$I_i=I(x_i\leqslant x)\cdot I(y_i\leqslant y-F(y_i|x,\theta),$$
$$\mathrm{CK}_n=\max_{x,y}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}I_i\right| \tag{3}$$

其中，$\{(x_1,y_1),\cdots,(x_n,y_n)\}\in S$表示待检测的样本集，$F(y|x)$为给定的先验条件分布，$I(\cdot)$表示指示函数. CK距离是KS距离在多维条件分布上的拓展，其本质是统计待测样本经验累计条件分布与先验条件分布之间的最大差异距离，适合用于度量分布的差异性. 同时，CK距离的计算不依赖先验假设，适用于任何分布之间比较，可以广泛应用于各类场景中.

由于现实环境的复杂性，数据分布通常可能是多种分布的复合，同时概念漂移可能只发生在其中的一部分. 而整个样本空间的漂移检测依赖大量的数据才能实现可靠的结果，极大影响漂移检测效率. 调整合适的阈值可以改善这一问题[18]，但阈值选择又过度依赖对任务的理解. 因此，局部漂移检测是提升性能的有效措施. 从局部概念漂移检测出发，本文首先给出局部漂移检测的可行性证明.

定义 4 给定样本集S，已知目标累计条件分布$F(y|X)$，将样本空间划分为若干区域，则存在区域S'，$\mathrm{CK}_{S'}\geqslant\mathrm{CK}_S$，其中$\mathrm{CK}_S$表示样本集$S$与$F(y|X)$间的CK距

离,$CK_{S'}$同理.

证明　情况1,样本集分布与目标分布一致时,显然$\forall S_d \in S, CK_{S'}=CK_S=0$.

情况2,样本集分布与目标分布不一致时,由CK距离的定义可知,其是两个累计条件分布间的最大差值,为简化描述,本文将式(3)表述为式(4):

$$CK=\max_{x,y}\left|F(x,y)-G(x,y)\right| \tag{4}$$

其中,$F(x,y)$表示为先验累计条件分布,$G(x,y)$表示样本集S的经验累计条件分布. 相应的,任意连续子域S'内的数据分布如式(5)与式(6)所示:

$$F'(x,y)=\frac{F(x,y)-F(x_{\min},y_{\min})}{F(x_{\max},y_{\max})-F(x_{\min},y_{\min})} \tag{5}$$

$$G'(x,y)=\frac{G(x,y)-G(x_{\min},y_{\min})}{G(x_{\max},y_{\max})-G(x_{\min},y_{\min})} \tag{6}$$

其中,$x_{\min},y_{\min}$分别表示S'区域内的下界值,$x_{\max},y_{\max}$为上界.

由式(4)可得,子域S'的CK距离如式(7):

$$CK_{S'}=\max_{x,y}\left|F'(x,y)-G'(x,y)\right| \tag{7}$$

令(x_k,y_k)表示取得最大分布差异时的样本点,对于任意样本点$(x_i,y_i)\in S$,满足式(8):

$$\left|F(x_k,y_k)-G(x_k,y_k)\right|\geqslant\left|F(x_i,y_i)-G(x_i,y_i)\right| \tag{8}$$

由上式可得:

$$\left|F'(x_k,y_k)-G'(x_k,y_k)\right|\geqslant\left|F(x_k,y_k)-G(x_k,y_k)\right| \tag{9}$$

结合式(7)与式(9)可知,任意包含(x_k,y_k)的样本空间子区域S'都满足$CK_{S'}\geqslant CK_S$.

证毕.

由定理1可知,当在样本空间任意区域上新旧样本间的CK距离超过阈值时,都是概念漂移发生的一种潜在信号. 因此,可以在局部区域进行一致性检验来检测概念漂移. 此外,概念分布具有内聚性特点,分区域检验通常可以更灵敏的检测出数据的变化[18],从而在更少的样本需求量下就可以判断出复杂的概念漂移.

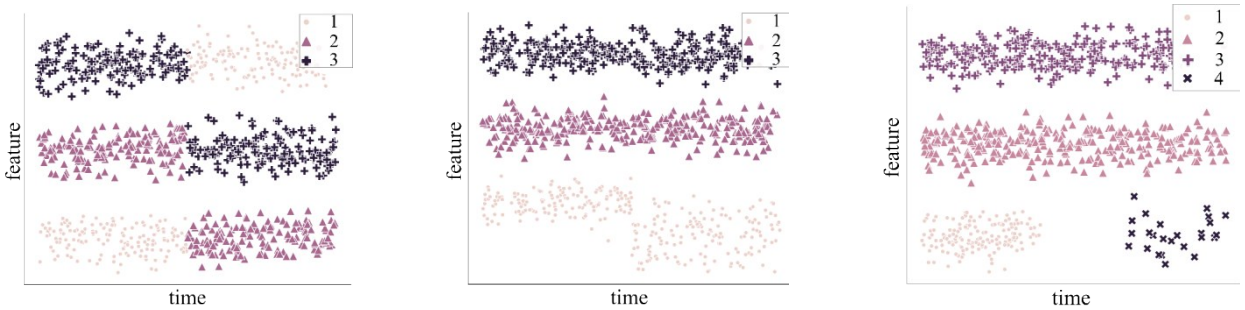
3.3　概念漂移分析

在局部视野观测数据时,概念漂移引起的分布变化可以呈现为标签分布变化,特征分布变化,时间分布变化[13],具体如图3所示.

从图3可以看出,特征分布变化与标签分布变化通常来说并不独立,它们共同构成了映射关系的变化. 而特征分布的变化又是在全局视野下最容易被观察到的(如图3(*b*)中所示),图中下方类别为1的样本在发生漂移后,其中心点与离散度等统计指标上发生了变化.

标签分布在整体空间上的变化通常并不显著,是因为不区分区域就无法正确感知各类标签的真实分布情况. 如图4(*a*)所示,全局下由于样本分散在整个空间上,虽然t_1时刻发生概念漂移,但难以发现标签分布的变化. 从在局部来观察数据分布,能更好判断局部的概念变化情况(图4(*b*)、图4(*c*)).

聚类研究中[19]对类簇有空间内聚性的假设,而在动态环境中内聚性同样存在于时间维度上. 如图3(*c*)所示,同一概念在时间维度上也会有聚集性. 不同的概念在时间上连续性也更差. 更一般情况是,当同一区域内的连续两次数据观测时间间隔较长时,其发生概念漂移的可能性也会变大. 如图5所示,原本在区域B的概念发生了漂移,而t_2时刻后区域B中再出现新样本时,已经不是原概念.



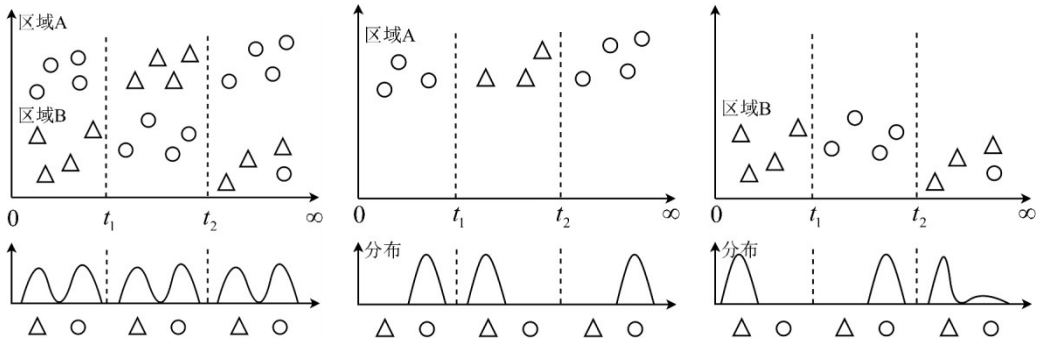
(*a*) 漂移在标签分布的表现　(*b*) 漂移在特征分布的表现　(*c*) 漂移在时间分布的表现

图3　漂移在标签分布、特征分布、时间上的表现

通过在概念域上监测特征、标签、时间维度上的变化,可有效判断概念漂移是否发生. 而基于概念域进行漂移检测可以灵敏识别概念漂移发生区域,从而保留有效区域同时过滤失效区域,实现在概念漂移环境下高效样本利用的动态样本选择.

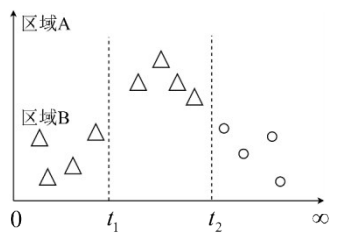
4　概念漂移自适应预测方法

本文构建的动态样本选择方法其基本思想是从局部动态地维护样本集,通过在局部进行样本去噪,并在适当的时机重用部分历史样本. 为此,设计了相应的处理策略:(1)基于局部漂移检测的样本去噪技术;(2)基

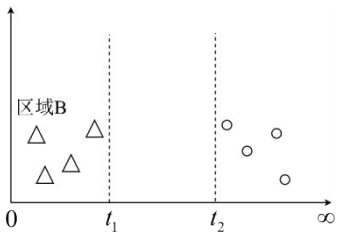


(*a*) 全样本分布 (*b*) 区域样本分布 (*c*) 区域样本分布

图4 标签分布变化在全局与局部的不同表现



(*a*) 全局视角下的样本



(*b*) 局部视角下的样本

图5 时间分布变化在全局与局部的不同表现

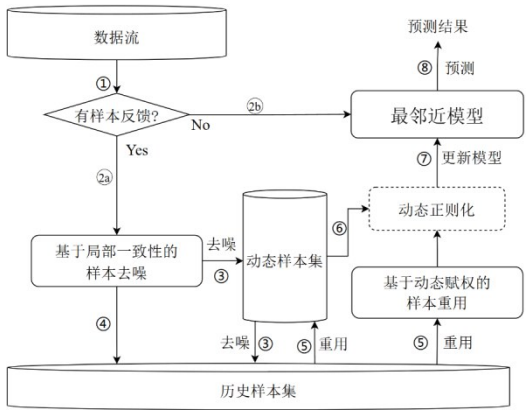


图6 基于动态样本选择的概念漂移自适应预测框架

于动态赋权的样本重用;(3)以及用于样本约简的多类代表点归纳;(4)基于动态样本集的预测方法.

预测方法的整体流程如图6所示,每当数据流提供样本反馈时,即提供一个样本(x,y),则通过样本去噪与样本重用维护动态样本集与历史样本集,并以动态样本集作为预测模型的基础训练集,同时通过重训练更新模型;当数据流仅提供特征x时,则使用预测模型对其进行预测.

4.1 基于局部漂移检测的样本去噪

针对概念漂移导致样本集中充斥漂移噪声的问题,本文设计了基于局部漂移检测的样本去噪方法. 该方法首先将样本空间划分为多个概念区域,然后根据各个概念域内概念漂移情况进行样本去噪处理,主要包括区域动态划分、局部漂移去噪两个流程.

区域动态划分的首要目标是适应数据流在特征空间上的变化,在这个过程中所有数据都将被逐个处理,用于动态地更新区域划分. 划分出的区域在特征空间上主要是以下两种形式变化:(1)区域移动,增量式的概念漂移[11]通常表现为概念的移动,其主要是表现在概念域的期望中心向着其他位置偏移;(2)区域边界变化,主要表现为区域内分布的离散程度或边界范围发生变化. 为处理上述两种变化,本文采用基于微簇的流式聚类方法[19]进行区域动态划分,通过微簇的增加或删除捕获区域的移动和边界的变化.

为了实现自适应地微簇半径选择,本文采用相互图聚类[20]来学习初始划分,将得到的初始概念域集合中半径频率最高的值作为默认半径. 其目的在于避免过小的簇出现,导致概念域过多而每个域中的样本过少,同样不会因为过大而导致覆盖区域不能有效区分样本,保证类簇划分结果具有较好的相关性. 此外,为了尽可能降低噪声对聚类效果的影响,本文加入标签约束使标签不同的样本不隶属于同一类簇,在生成样本相互关系的邻接矩阵时截断不同类别样本之间的连接. 具体如下所示:

$$m_{i,j}=\begin{cases}1, x_i\text{与}x_j\text{互为}k\text{最邻近样本}\&y_i=y_j\\0, \text{其他}\end{cases} \tag{10}$$

其中$m_{i,j}$表示邻接矩阵中样本x_i与x_j的连接关系.

在得到合适的区域划分后,将进行基于局部一致性检验的样本去噪. 在检测微簇内映射关系的变化时,

由于特征空间的限定,其映射关系由如下特性:

$$\lim_{x\to x_0} F(y|\{x|\text{dist}(x-x_0)\leqslant R\})=F(y|x_0) \tag{11}$$

其中,$\{x|\text{dist}(x-x_0)\leqslant R\}$中心点为$x_0$,半径距离为$R$的微簇区域内的特征空间. 从式(11)可看出,此情况下的CK距离又回退到KS距离,即映射关系的变化转换为标签分布的变化. 因此,两个样本集间的分布差异度可近似描述为:

$$D_{\text{label}}=\sup_y\left|Y_{\text{new}}(y)-Y_{\text{old}}(y)\right| \tag{12}$$

其中$Y_i(y)$表示标签的累计分布.

为度量漂移发生的显著程度,本文参考双样本KS检验,其非空假设为:

$$D_{\text{label}}>c(\alpha)\sqrt{\frac{n+m}{n\cdot m}} \tag{13}$$

其中n与m为两个样本集中样本的数量,$c(\alpha)$可通过下式计算:

$$c(\alpha)=\sqrt{-\ln(\frac{\alpha}{2})\cdot\frac{1}{2}} \tag{14}$$

其中α为预设的显著度值. 当式(14)成立时拒绝空假设,即没有发生概念漂移.

由于漂移检测是在局部进行,样本规模通常并不是特别庞大,节约的时间成本可以有效提升小样本检测的稳定性. 基于上文的分析,本文将标签分布与时间分布相结合,得到局部一致性分布差异度检验方法:

$$D_{\text{label}}>c(\alpha)\sqrt{T_{\text{d}}\cdot T_{\text{gap}}} \tag{15}$$

其中各项计算如下列公式所示:

$$T_{\text{gap}}=\max\{t_i-t_{i-1},t_i\in T_{\text{new}}\} \tag{16}$$

$$\bar{T}_{\text{new}}=\max(T_{\text{new}}),\ \underline{T}_{\text{new}}=\min(T_{\text{new}}) \tag{17}$$

$$T_{\text{d}}=\frac{\bar{T}_{\text{new}}-\underline{T}_{\text{new}}+\bar{T}_{\text{old}}-\underline{T}_{\text{old}}}{[\bar{T}_{\text{new}}-\underline{T}_{\text{new}}]\cdot[\bar{T}_{\text{old}}-\underline{T}_{\text{old}}]} \tag{18}$$

其中,T_{new}、T_{new}分别表示新旧样本集中样本的时间戳属性集合. 其本质上是利用时间因子替代了数量因子,这样做可以在新样本集中存在时间间隔较大的情况时,判断发生概念漂移的可能性也更大.

在一个微簇内检测概念漂移时,我们同样使用滑动窗口的方式,即最新的N个样本被认为是新样本集,而最久的N个样本被认为是旧样本集. 在发现概念漂移时,我们将该微簇内的最新的k个数据作为新概念保留,而此外的数据都将被视为失效的噪声数据进行删除. 虽然滑窗方法不能保样本是无噪声的,但在结合区域去噪处理后,在最大限度保留已有样本的同时尽可能减少噪声数据. 如算法1所示.

算法1　基于局部漂移检测的样本去噪算法

输入:样本(x_i,y_i,t_i),当前微簇集N_c,动态样本集S^*,历史样本集N_{history},置信度α,最小样本数N,潜在概念域最长保留时间T_{latent};

输出:去噪后的样本集S^*

```
IF C ← nearest(x_i, C.center), C ∈ N_c
  IF dist(x_i, C.center) ⩽ R:
    C = C + (x_i, y_i, t_i) /* 将样本加入到簇 C 中 */
    IF |C| < N:
      C.center = (C.center × |C| + x_i)/(|C| + 1); /* 更新簇的中心点 */
    ELSE:
      D_label ← 根据式(12)计算局部新旧概念间的分布差异度;
      IF 根据式(15)判断是否发生概念漂移:
        N_history = N_history + (C − newest(C)); /* 将旧样本保存至历史样本集 */
        C = newest(C); /* 微簇 C 保留 N 个最新的样本.
      END IF
    END IF
  END IF
ELSE:
  N_c = N_c + new(C ← {(x_i, y_i, t_i)});/* 用新样本构成新的微簇,并加入到微簇集中 */
END IF
FOR C IN N_c
  (x_new, y_new, t_new) ← 微簇 C 内最新的样本:
  IF t_new − i > T_latent:
    N_c = N_c − C;/* 长时间未使用的微簇被删除 */
    IF |N_c(x)| ⩾ N:
  N_history = N_history + old(N_c(x_c));/* 其中的样本放入到历史样本集中 */
    END IF
END IF
```

4.2　基于动态赋权的样本重用

基于局部漂移检测的样本去噪不可避免会提前过滤一些在未来仍然有价值的样本,其中可能存在着与当前概念相似的样本. 为了能够重用这些有价值的样本,本文将被判定为噪声删除的数据保存到历史样本集中,当出现新概念时,在历史样本集中寻找相似样本,将其中有效样本重新加入到当前样本集中,从而利用充分的历史样本丰富训练集.

为实现以上目标,本文提出基于动态赋权的样本重用方法. 首先,对于一个样本实例(x,y),定义其在t时刻的样本有效性权重为:

$$w(x,y,t)=\frac{1}{|C(x)|}\sum_{y_i\in C(x)}I(y_i=y) \tag{19}$$

其中$C(x)$表示当前样本集中位于以x为中心点的微簇内的样本集合,样本有效性权重会受到新样本的影响,因此它的更新公式为:

$$w(x,y,t+1)=\frac{|N_c(x)|\cdot w(x,t)+\Delta(y_i=y)}{|N_c(x)|+1} \tag{20}$$

从实际来看,对全部历史样本进行权重更新是极

其耗时的,同时也包含大量无意义操作. 为了简化这个过程,可在每次新样本到来时仅针对微簇进行更新,也就是查找与该样本相近的历史微簇,然后修改这个微簇的动态权重.

需要特别注意的是,旧样本的重用应持保守策略:旧概念和新概念虽然相似,但在新概念被完全认识之前,两者一致程度难以保证. 而当新概念被完全认识时,旧样本重用的必要性又将不复存在. 因此,旧样本加入到新概念后,新样本会不断地取代掉这些旧样本直到一个稳定的微簇形成. 算法2阐述了样本重用的实现流程.

算法2 基于动态赋权的样本重用

输入:样本(x_i,y_i,t_i);当前样本集S^*;历史样本集N_{history};样本权重阈值th_w

输出:重用后的样本集S^*

$C\leftarrow\{(x,y,t)|\text{dist}(x_i,x)\leqslant R\},(x,y,t)\in N_{\text{history}}$/* 在历史样本集中与新样本最邻近的微簇 */

FOR (x,y,t) IN C:

  根据式(20)更新C内样本的权重

  IF $w(x,y,t)\geqslant\text{th}_w$:

    $S^*=S^*+(x,y,t)$/* 高权重样本重新被放入当前样本集中 */

  END IF

END FOR

4.3 局部多类别的代表点归纳

对于数据规模特别大的情况下,过度膨胀的样本集会严重减缓模型学习以及样本查询的效率. 为避免这一问题,本文采用数据代表点归纳来构建更高效可用的动态样本集.

如算法3所示,通过局部概念区域划分过程生成的微簇,其中心点对于区域内样本已具有高度概括性. 为避免分类边界模糊的问题,本文对含有多个类别的样本区域进行多语义归纳,在这些区域内将生成多个代表点,对每个占比超过阈值的样本类别分别归纳一个代表点,并以相应的标签作为代表点的预测目标属性. 该方法的优势还包括当区域划分算法效果较差时,会尽可能的降低对原始数据的信息损失.

算法3 局部多类别的代表点归纳

输入:样本集$S=\{(x_1,y_1,t_1),\cdots,(x_n,y_n,t_n)\}$;微簇集$N_c$;

输出:归纳后的样本集S^*

$S^*_{\text{reduce}}=\{\}$

FOR C IN N_c:

  FOR y IN set($C.y$): /*set($C.y$)表示保留微簇C中的所有无重复标签值y集合 */

$$x=\frac{1}{|\{d.y==y|d\in C\}|}\sum_{d\in C}^{C}I(d.y==y)\cdot d.x$$

/* 计算各类别样本的中心点,其中$d.x$表示样本d的x属性 */

    $S^*_{\text{reduce}}=S^*_{\text{reduce}}+(x,y)$/* 将样本插入到集合中 */

  END FOR

END FOR

4.4 基于样本选择的学习

本文的最终目标是在概念漂移数据集中可以取得稳定的预测效果,核心问题在于适应概念漂移的动态样本集构建及相应的学习模型. 当前,采用以下两种策略进行模型更新:(1)采用周期性重训练或检测到漂移后重训练策略. 该方式可达到大多数任务的最好预测水平,但对于许多学习模型来说,重训练时间消耗巨大;(2)采用懒学习策略[15],即采用最邻近预测模型. 通过搜索最邻近样本进行预测,可以节省昂贵的训练时间,但预测时间受到样本集规模的影响,同时预测准确率会受到距离度量方式的影响. 最终构建的基于动态样本选择的概念漂移自适应预测算法如下所示.

由于上述过程中存在着大量的空间搜索操作,为了提高算法时间效率,本文使用K维树(K-Dimensional Tree, KDTree)[21]来维护历史微簇集. 具体来说,当每一个样本实例(x,y)可获取时,根据特征x在KDTree中查找相应的历史微簇,随后将进行概念漂移检测. 当发现新概念时,对权重超过期望阈值的历史微簇重新激活加入当前样本集中. 对于失效的噪声微簇,则将中心点x_N增量式地插入到KDTree中. 算法4详尽阐述了整个自适应预测方法的实现流程.

算法4 基于动态样本选择的自适应预测方法

输入:样本集$S=\{(x_1,y_1,t_1),\cdots,(x_n,y_n,t_n)\}$;相互邻近阶数$k$;稳定概念所需的最小样本数阈值$N$;潜在概念域最长保留时间$T_{\text{latent}}$;

输出:持续的预测结果y^*

/ * 初始化阶段 */

利用初始样本依据式(10)计算k阶相互邻接矩阵$\boldsymbol{M}$;

对$\boldsymbol{M}$中各个极大连通图构建概念域,形成概念集合N_c;

删除样本个数低于N的概念域,形成动态样本集S^*;

$R\leftarrow$概念域集合中半径频率最高的值作为默认半径使用;

Learner$\leftarrow$根据S^*初始化预测模型;

/ * 流式处理阶段 */

FOR $i\leftarrow 1,2,\cdots,n$ LOOP

  $y_i^*\leftarrow$Learner(x_i) // 预测

  / * 增量自适应阶段 */

  $\boldsymbol{S}^*\leftarrow$算法1, 样本去噪

  $\boldsymbol{S}^*\leftarrow$算法2, 样本重用

  $\boldsymbol{S}^*_{\text{reduce}}\leftarrow$代表点归纳, 返回全部当前微簇集的$C$

  Learner$\leftarrow$根据S^*_{reduce}更新

END FOR

假设数据量为n,选用m个样本作为初始化使用,相互图聚类(算法1第4行)时间复杂度为$O(m^2)$,本文仅探讨基于懒策略的模型时间复杂度,样本去噪对当

前概念域的搜索需要 $O(|N_c|^2)$ 时间,基于局部一致性的漂移检测需要 $O(|C|)$,样本重用时耗时主要为通过KDTree对历史样本的搜索以及KDTree的维护更新,而这一阶段耗时平均不超过 $O(n\log(n))$,代表点归纳只是使用概念域更新的结果此阶段耗时为 $O(|N_c|)$,总耗时为 $O(n\cdot m^2+n\cdot|N_c|^2+n\cdot|C|+n\cdot n\log(n)+n)$,而由于在实际应用中 n 远大于其他值,则平均时间复杂度可视为 $O(n^2\log(n))$,当概念域划分完全错误的最差情况下,此时所有数据都保存在动态样本集中,则最大时间复杂为 $O(n^3)$.

5　实验与分析

针对本文所提的基于动态样本选择自适应预测算法,本文从两个角度进行验证:(1)算法在概念漂移数据上样本去噪效果. 过在合成数据集上主动引入概念漂移,使数据发生多种分布变化,观察本算法在多种概念漂移下的表现;(2)算法概念漂移数据的预测性能. 用现实环境获取的数据中进行漂移预测,并与其他算法对比结果.

实验软硬件平台如下:Intel(R) Core(TM) i5-9500 CPU,16 GB 内存,INTEL 512 GB 硬盘,操作系统为Win10,使用语言为Python3.8版本.

5.1　实验数据

为验证算法样本去噪性能,本文选择4个概念漂移合成数据集进行实验,分别由不同的概念漂移类型生成. 为验证样本选择的有效性,选择了5个来自于真实生产环境的流数据进行实验集.

数据集 Drift Moon、SEA[22]、Moving Squares[23]、Mixed Drift[23]为人工合成的数据集,包含有突变性、渐变型、增量型、反复型概念漂移,一些概念漂移是局部的,会更考验算法的局部适应能力;Elec2,Weather[8],Bike[23],Poker[23],Colliery Recall 数据集均收集于不同的现实场景中,包括了气象、工业、销售等场景,能够代表概念漂移的典型场景. 各数据集规模、特征数、标签数如表1所示.

表1　实验数据

分组	项目	样本数	特征数	标签数
合成数据	Drift Moon	6 000	2	2
	SEA	50 000	3	2
	Moving Squares	200 000	2	3
	Mixed Drift	200 000	2	15
真实数据	Elec2	45 312	8	2
	Weather	18 159	8	2
	Bike	17 379	4	2
	Poker	829 201	21	40
	Colliery Recall	329 201	71	5

5.2　实验设计与结果分析

本文设计算法对数据的处理是流式的而非批式,意味着数据处理过程是逐个进行. 标签延迟设定为1,这表示在 t 时刻见到样本的特征 x_t 并对它进行预测,在 t+1时刻会得到样本的真实标签 y_t. 因此,整个流数据预测在每个新样本到来后,都可以立即进行增量式学习. 为了让部分学习算法可以稳定的启动,本文选择少量样本(500条)作为初始训练样本,其后数据作为流数据测试样本逐个进行预测,最终统计在测试样本上的平均性能指标.

5.2.1　样本选择实验

合成数据集是手动生成的数据集,可以清晰的获得每一个阶段数据的分布情况. 实验采用4个合成数据进行样本去噪对比实验,验证CDAP-DSS算法去噪性能. 为了量化去噪效果,采用噪声率作为评价指标,具体如下:

$$\mathrm{NR}=\frac{N_{\mathrm{noisy}}}{N} \tag{21}$$

其中 N_{noisy} 为噪声样本的数量,N 为总样本数.

此外,实验会与其它概念漂移样本选择方法进行对比:(1)基础方法(Base Line, BL)[24],只使用初始样本作为训练集,之后不更新训练集;(2)滑动窗口类方法(Top Line, TL)[24],此类方法为业界最常用方法,其在众多数据集上通常可取得非常优异的效果,本文将具体滑窗策略定为两种,TL1方法将滑动窗口内的全部样本作为训练集,TL2方法将全部历史样本作为训练集;(3)基于漂移检测的方法,本文选择基于霍夫丁边界的概念漂移检测算法(Hoeffding`s bounds based Drift Detection Method, HDDM)[8],它被广泛用于漂移检测工作,本文将其的样本选择策略设定为使用概念漂移后的数据作为训练集;(4)基于原型的方法:反应式软原型学习(Robust Soft Learning Vector Quantization, RSLVQ)[15],基于预测性能和聚类划分进行原型选择,再根据原型的预测结果进行概念漂移检测,进而更新样本选择结果,在代表点选择上具有较优能力;(5)对于本文所提出的CDAP-DSS方法,在进行样本选择时使用的 k 参数设为3,最小样本阈值数为5,潜在概念域最长保留时间取2 000.

表2展示了在4个合成数据集上的实验结果. 使用CDAP-DSS在4个数据集上进行样本选择时均取得了较低的噪声率,由于MovingSquares数据集的增量漂移形式,从而TL1方法样本选择结果在该数据集上几乎是无

噪声的;Drift Moon以及Mixed Drift数据集上一部分概念漂移仅发生在局部,而CDAP-DSS是从局部角度进行的漂移检测,可以更快速地发现概念漂移发生,从而使得样本中所含的噪声量更小.

表2 实验数据

方法	SEA	Drift Moon	Moving Squares	Mixed Drift
BL	0.167 7	0.491 9	0.671 9	0.580 6
TL1	0.132 0	0.090 6	**0.000 0**	0.211 6
TL2	0.165 4	0.372 1	0.696 2	0.796 3
HDDM	0.197 5	0.225 5	0.061 9	0.243 1
RSLVQ	0.298 4	0.580 8	0.247 2	0.259 6
CDAP-DSS	**0.130 9**	**0.063 0**	0.050 9	**0.098 6**

为了更好呈现漂移发生时各算法性能的变化,图7展示了各方法在DriftMoon数据集上各个时刻的噪声率,可以看出在发生漂移时,所有算法都会出现剧烈抖动,虽然CDAP-DSS同样会发生抖动,但可以看出在3个漂移点上,CDAP-DSS方法噪声率下降速度会更快.这是因为CDAP-DSS在面对漂移时,从局部视野发现漂移,其对漂移的反映更加灵敏,样本更新更快.

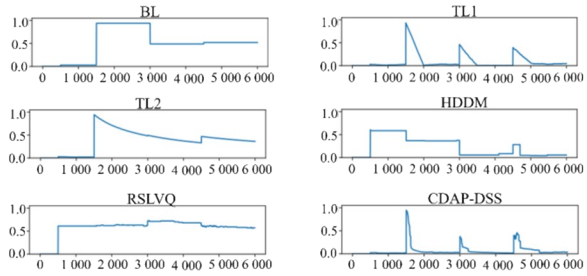


图7 各方法在Drift Moon数据集上实时噪声率

图8展示了CDAP-DSS在Drift Moon数据集4个阶段的样本选择结果图.在第一个概念出现阶段,算法保留了部分随机噪声,用于避免不确定性导致的概念漂移误判;在随后的各个概念变化阶段中,CDAP-DSS也都得到了相对优秀的样本选择结果.在最后一个阶段时,由于该阶段发生的是增量式漂移变化,CDAP-DSS对于一些历史数据并没有进行全部的删除,而是保留了部分未与新概念重叠的区域样本.其目的在于该样本虽然不代表当前概念,但对反复型漂移来说仍存在

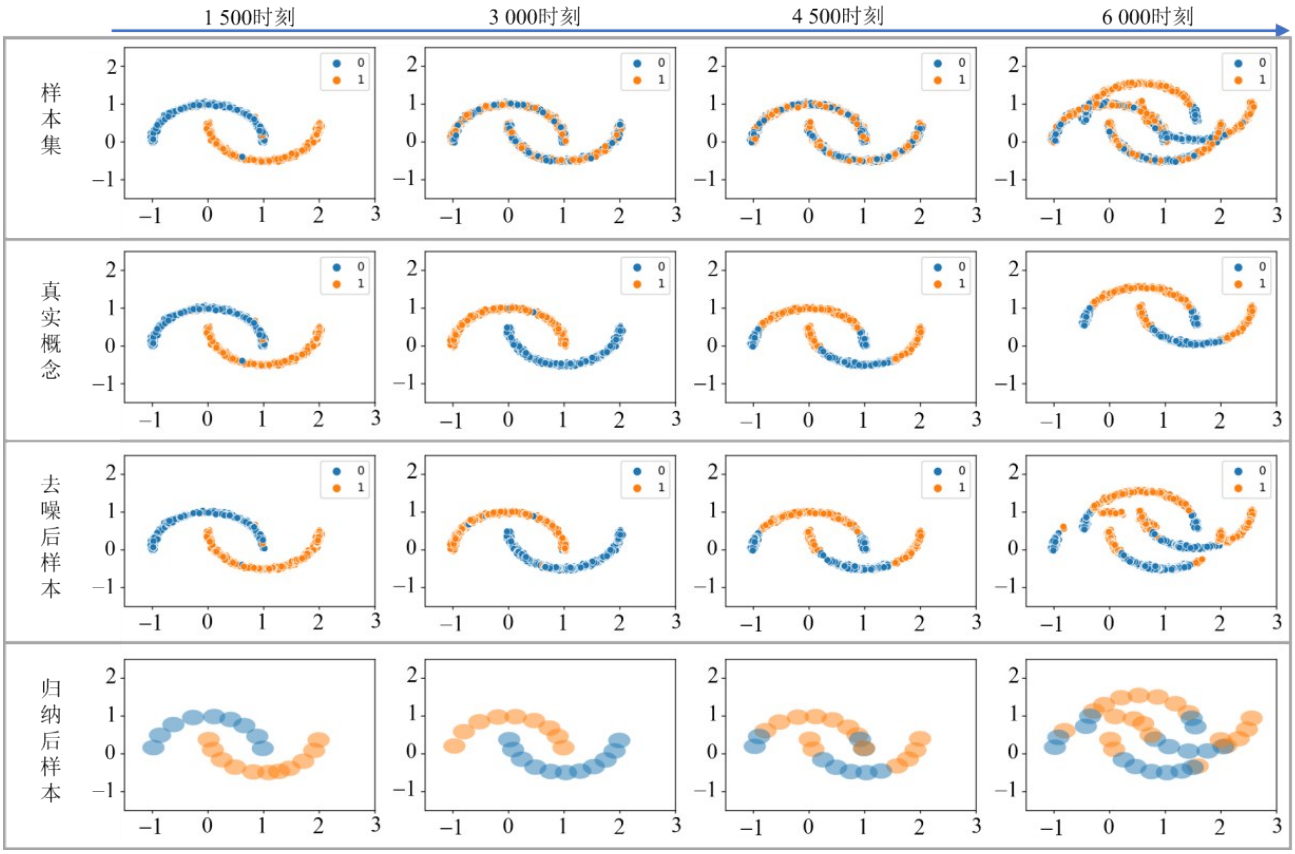


图8 Drift Moon数据集各阶段概念及样本选择结果

潜在价值. 需要注意的是,这种保留也可能导致空间分布的不均匀,在一定程度上会影响模型学习的效果.

5.2.2　预测实验

由于样本选择的本质目标是提高预测效果,所以对CDAP-DSS方法在5个真实数据集上进行流式预测实验,同时对部分方法设置使用动态采样与不使用动态采样的情况对比. 实验时,以一定量的数据作为初始数据进行训练学习,对随后的数据用流式逐个的进行预测,并在预测后给模型及时提供真实结果,以便其对变化做出快速的适应.

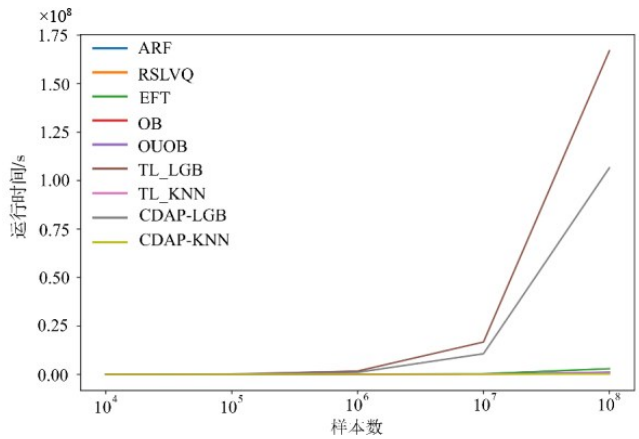
预测实验对比方法包括:(1)基础方法(BL):只使用固定数量的数据作为训练集,在后续过程中不做任何更新操作;(2)滑窗方法(TL):与样本选择实验中的设计一致,但这里会在每个新样本加入进来后,就对模型进行重训练;(3)流数据分类算法:除了上文介绍过的自适应随机森林算法ARF[13]与反应式软原型学习算法RSLVVQ[15],对比方法还包括基于在线集成技术的在线提升算法(Online Boosting, OB)和在线欠采样过采样袋装算法(Online Under Over Bagging, OUOB)[12],以及基于树模型自适应迭代的极速决策树模型(Effective Fast Tree, EFT)[25]

实验中参数的相关设置如下:(1)对于概念漂移检测算法,漂移阈值设定大小选取范围为(0.05,0.005,0.000 5,0.000 05);(2)对于滑窗类方法,窗口大小的选择在(200,400,600,800,1 000,1 500,2 000)范围,预测结果选择最优结果展示;(3)数据集初始样本量为1 000. 各对比方法根据范围内的值进行多次交叉实验,取最好结果展示.

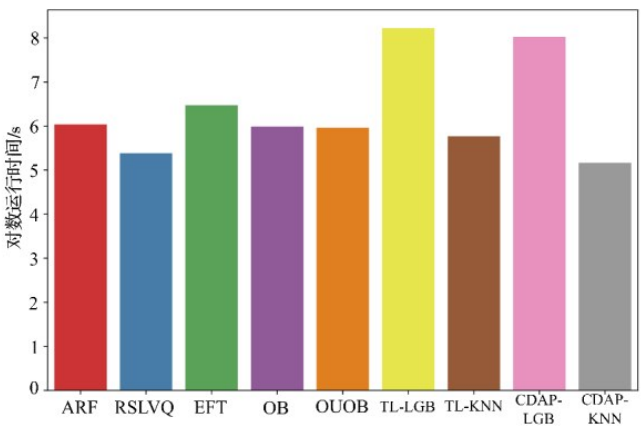
本文所提出的CDAP-DSS方法,将采用轻量级梯度提升机(light gradient boosting machine, LGBM)[26]与k近邻算法(Nearest Neighbor, KNN)[27]作为基方法分别记为CDAP-LGBM和CDAP-KNN. 大多数参数使用算法包提供的默认值. 特别的是CDAP-KNN方法的k值选择为3,CDAP-DSS在进行样本选择时使用的k参数设计为3,最小样本的阈值数N为5,潜在概念域最长保留时间T_{latent}取2 000.

表3展示了在真实数据集上进行预测实验的结果. 可以发现在大多数数据集上CDAP-DSS都取得了不错效果,而由于基模型学习能力不同,在各数据集上表现略有差异. 通过与基线方法、滑窗方法的对比,可以发现在使用动态样本选择后确实有较大提升,对于部分数据集虽然不能达到最高水平,其相差也不至于过大. CDAP-LGBM模型在许多数据集上基本可以达到最高水平,对比滑窗重训练方法也可以实现不低的预测水平,这是因为滑窗方法丢失了历史有效样本,同时也不能保证滑窗内样本无噪声,而CDAP-LGBM学习的是经过去噪的样本,可以有效提高数据质量,从而提高预测准确率. 但这种方法的问题在于其高昂的重训练时间. 而在Colliery Recall实验中由于样本重用的原因,导致CDAP-DSS可以在更大的范围内间接地实现数据相对平衡,尤其在召回类别实现了更准确的预测. 而基于懒学习策略的CDAP-KNN在预测表现上可以接近OB算法,充分验证了本文所提方法在流数据分类上选取的代表点对样本具有较高的代表性,适合用于流数据分类预测任务.

图9展示了各个算法的运行时间效率. BL算法由于仅需初次适应,后续并没有任何更新过程,所以在本次实验并不参与统计;TL类算法与CDAP-LGBM虽然能够取到较高的预测准确率,但大量时间用于重训练上,导致运行时长过长,尽管如此,大部分对实时性要求不高的应用场景仍然可以使用该类方法;而其他增量式方法由于在千万规模的数据集上其运行时间略有差异,KNN方法由于没有任何主动的模型重构建



(a) 各方法运行时间



(b) 各方法对数运行时间

图9　各方法在合成数据上的运行时间

表3 真实数据集上预测准确率

方法	Elec2	Weather	Bike	Poker	Colliery Recall
TL1-LGBM	0.906 7	0.773 4	0.712 1	0.851 3	0.321 1
TL2-LGBM	0.885 2	0.791 5	0.711 1	0.810 2	0.383 1
BL-LGBM	0.610 1	0.701 1	0.485 5	0.353 2	0.171 5
TL1-KNN	0.731 7	0.773 0	0.716 1	0.681 5	0.113 1
TL2-KNN	0.747 4	0.785 0	0.693 9	0.681 1	0.153 1
BL-KNN	0.578 8	0.751 1	0.481 3	0.507 3	0.101 1
CDAP-LGBM	**0.919 0**	**0.792 2**	0.741 7	**0.907 1**	**0.432 3**
CDAP-KNN	0.744 5	0.766 4	0.723 1	0.701 1	0.398 2
OB	0.788 0	0.717 5	0.725 5	0.732 9	0.291 4
OUOB	0.773 4	0.769 8	0.722 6	0.724 5	0.313 5
ARF	0.884 7	0.754 1	0.759 9	0.856 4	0.374 3
RSLVQ	0.844 8	0.637 8	**0.785 0**	0.694 1	0.251 2
EFT	0.836 0	0.735 4	0.739 2	0.715 1	0.269 7

使其在各类方法中速度最快,这使得基于懒策略的方法CDAP-KNN、RSLVQ取得较为优异的预测速度.

6 结论

针对流数据挖掘中的概念漂移问题,本文从动态样本选择的角度出发,提出了一种基于动态样本选择的概念漂移自适应预测方法,通过充分利用历史样本有效解决了概念漂移问题. 特别的是,本文深入分析了从局部进行适应概念漂移的可行性,并设计了基于概念域的动态样本选择方法,在真实数据与合成数据上充分验证了其良好的去噪能力以及高效的预测性能.

需要注意的是,该方法在时间效率上还具有较大的优化空间. 此外,当前模型在预测准确率与时间效率方面还存在无法同时优化问题,可以尝试通过动态性度量方式或者响应式更新策略进行改善,这也是未来需要努力的方向.

参考文献

[1] MOHAWESH R, TRAN S, OLLINGTON R, et al. Analysis of concept drift in fake reviews detection[J]. Expert Systems with Applications, 2021, 169: 114318.

[2] WANG L, WU C. Dynamic imbalanced business credit evaluation based on Learn++ with sliding time window and weight sampling and FCM with multiple kernels[J]. Information Sciences, 2020, 520: 305-323.

[3] HENKE M, SANTOS E, SOUTO E, et al. Spam detection based on feature evolution to deal withConcept drift[J]. Journal of Universal Computer Science, 2021, 27(4): 364-386.

[4] YANG L M, GUO W B, HAO Q Y, et al. CADE: Detecting and explaining concept drift samples for security applications[C]//30th USENIX Security Symposium. Berkeley: USENIX, 2021: 2327-2344.

[5] 韩光洁, 赵腾飞, 刘立, 等. 基于多元区域集划分的工业数据流概念漂移检测[J]. 电子学报, 2023, 51(7): 1906-1916.
HAN G J, ZHAO T F, LIU L, et al. Concept drift detection of industrial data flow based on multivariate region set partition[J]. Acta Electronica Sinica, 2023, 51(7): 1906-1916. (in Chinese)

[6] 陆克中, 陈超凡, 蔡桓, 等. 面向概念漂移和类不平衡数据流的在线分类算法[J]. 电子学报, 2022, 50(3): 585-597.
LU K Z, CHEN C F, CAI H, et al. Online classification algorithm for concept drift and class imbalance data stream [J]. Acta Electronica Sinica, 2022, 50(3): 585-597. (in Chinese)

[7] GAMA J, MEDAS P, CASTILLO G, et al. Learning with drift detection[C]//Advances in Artificial Intelligence-SBIA 2004. Berlin: Springer Berlin Heidelberg, 2004: 286-295

[8] FRIAS-BLANCO I, DEL CAMPO-AVILA J, RAMOS-JIMENEZ G, et al. Online and non-parametric drift detection methods based on hoeffding´s bounds[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(3): 810-823.

[9] LIU A J, LU J, ZHANG G Q. Concept drift detection via equal intensity k-means space partitioning[J]. IEEE Transactions on Cybernetics, 2021, 51(6): 3198-3211.

[10] YANG Z, AL-DAHIDI S, BARALDI P, et al. A novel concept drift detection method for incremental learning in nonstationary environments[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(1): 309-320.

[11] LU J, LIU A J, DONG F, et al. Learning under concept drift: A review[J]. IEEE transactions on Knowledge and Data Engineering, 2019, 31(12): 2346-2363.

[12] ELWELL R, POLIKAR R. Incremental learning of concept drift in nonstationary environments[J]. IEEE Transactions on Neural Networks, 2011, 22(10): 1517-1531.

[13] GOMES H M, BIFET A, READ J, et al. Adaptive random forests for evolving data stream classification[J]. Machine Learning, 2017, 106(9): 1469-1495.

[14] SHAO J M, AHMADI Z, KRAMER S. Prototype-based learning on concept-drifting data streams[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:

ACM, 2014: 412-421.
[15] RAAB C, HEUSINGER M, SCHLEIF F M. Reactive soft prototype computing for concept drift streams[J]. Neurocomputing, 2020, 416: 340-351.
[16] RAMÍREZ-GALLEGO S, KRAWCZYK B, GARCÍA S, et al. A survey on data preprocessing for data stream mining: Current status and future directions[J]. Neurocomputing, 2017, 239: 39-57.
[17] ANDREWS D W K. A conditional Kolmogorov test[J]. Econometrica, 1997, 65(5): 1097-1128.
[18] LIU A J, LU J, SONG Y L, et al. Concept drift detection delay index[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(5): 4585-4597.
[19] ZUBAROĞLU A, ATALAY V. Data stream clustering: A review[J]. Artificial Intelligence Review, 2021, 54(2): 1201-1236.
[20] GAO Y, CHANDRA S, LI Y F, et al. SACCOS: A semi-supervised framework for emerging class detection and concept drift adaption over data streams[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(3): 1416-1426.
[21] XU W, CAI Y X, HE D J, et al. FAST-LIO2: Fast direct LiDAR-inertial odometry[J]. IEEE Transactions on Robotics, 2022, 38(4): 2053-2073.
[22] BRODER A Z. On the resemblance and containment of documents[C]//Proceedings. Compression and Complexity of SEQUENCES 1997. Piscataway: IEEE, 1997: 21-29.
[23] LOSING V, HAMMER B, WERSING H. KNN classifier with self adjusting memory for heterogeneous concept drift[C]//2016 IEEE 16th International Conference on Data Mining (ICDM). Piscataway: IEEE, 2016: 291-300.
[24] DOS REIS D M, 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. New York: ACM, 2016: 1545-1554.
[25] MANAPRAGADA C, WEBB G I, SALEHI M. Extremely fast decision tree[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2018: 1953-1962.
[26] KE G L, MENG Q, FINLEY T, et al. Lightgbm: A highly efficient gradient boosting decision tree[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc, 2017: 3149-3157.
[27] COVER T, HART P. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27.

作者简介

代　劲　男,1978年出生. 教授,博士,从事大数据知识工程、智能信息处理等研究.
E-mail: daijin@cqupt.edu.cn

李　昊　男,1996年出生,硕士研究生,主要研究领域为机器学习,流数据挖掘.
E-mail: S200201023@stu.cqupt.edu.cn

王国胤　男,1970出生. 教授,博士生导师,从事粒计算主要研究领域为多粒度认知计算、认知计算、智能信息处理等研究.
E-mail: wanggy@cqupt.edu.cn