软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(4):932−947 [doi: 10.13328/j.cnki.jos.005917]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于在线性能测试的概念漂移检测方法*

郭虎升[1,2], 张爱娟[1], 王文剑[1,2]

[1](山西大学 计算机与信息技术学院,山西 太原 030006)

[2](计算智能与中文信息处理教育部重点实验室(山西大学),山西 太原 030006)

通讯作者: 王文剑, E-mail: wjwang@sxu.edu.cn

摘 要: 概念漂移是动态流数据挖掘中一类常见的问题,但混杂噪声或训练样本规模过小而产生的伪概念漂移会引起与真实概念漂移相似的结果,即模型在线测试性能的不稳定波动,导致二者容易混淆,发生概念漂移的误报.针对流数据中真伪概念漂移的混淆问题,提出一种基于在线性能测试的概念漂移检测方法(concept drift detection method based on online performance test,简称 CDPT).该方法将最新获得的数据集进行均匀分组,在每组子数据集上分别进行在线学习,同时记录每组子数据集训练测试得到的分类精度向量,并计算相邻学习时间单元之间的精度落差,依据测试精度下降阈值得到有效波动位点.然后采用交叉检验的方式整合不同分组中的有效波动位点,以消除流数据在线学习过程中由于训练样本过小导致模型不稳定造成的检测干扰,根据精度波动一致性得到一致波动位点.最后,通过跟踪在线学习分类准确率,得到一致波动位点邻域参照点的测试精度变化,比较一致波动位点邻域参照点对应的模型测试精度下降幅度及收敛情况,以有效检测一致波动位点当中真实的概念漂移位点.实验结果表明,该方法能够有效辨识流数据在线学习过程中发生的真实概念漂移,并能有效避免训练样本过小或者流数据中噪声对检测结果的负面影响,同时提高模型的泛化性能.

关键词: 流数据;概念漂移;交叉检验;有效波动位点;一致波动位点;概念漂移位点

中图法分类号: TP181

中文引用格式: 郭虎升,张爱娟,王文剑.基于在线性能测试的概念漂移检测方法.软件学报,2020,31(4):932−947. http://www.jos.org.cn/1000-9825/5917.htm

英文引用格式: Guo HS, Zhang AJ, Wang WJ. Concept drift detection method based on online performance test. Ruan Jian Xue Bao/Journal of Software, 2020,31(4):932−947 (in Chinese). http://www.jos.org.cn/1000-9825/5917.htm

Concept Drift Detection Method Based on Online Performance Test

GUO Hu-Sheng[1,2], ZHANG Ai-Juan[1], WANG Wen-Jian[1,2]

[1](School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China)

[2](Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education (Shanxi University), Taiyuan 030006, China)

Abstract: Concept drift is a common problem in dynamic streaming data mining, but the false concept drift generated by the mixed noise data or too small scale size training data will cause similar results to the concept drift, that is, the instability fluctuation of model online testing performance, which leads to confusion between them, and the false alarm of concept drift. To address the problem which is easy to confuse the authenticity of concept drift, concept drift detection method based on online performance test, namely CDPT, is

* 基金项目: 国家自然科学基金(61503229, 61673249, U1805263); 山西省自然科学基金(201901D111033); 山西省重点研发计划(国际合作)(201903D421050)

Foundation item: National Natural Science Foundation of China (61503229, 61673249, U1805263); Natural Science Foundation of Shanxi Province of China (201901D111033); Key R&D Program of Shanxi Province (International Cooperation) (201903D421050)

本文由“非经典条件下的机器学习方法”专题特约编辑高新波教授、黎铭教授、李天瑞教授推荐.

收稿时间: 2019-03-08; 修改时间: 2019-07-11; 采用时间: 2019-09-20; jos 在线出版时间: 2020-01-10

CNKI 网络优先出版: 2020-01-14 09:53:12, http://kns.cnki.net/kcms/detail/11.2560.TP.20200114.0952.002.html

presented. With CDPT, the latest acquired data are evenly divided into groups, and online learning is performed on each group sub sets. At the same time, the classification accuracy vectors obtained by training and testing of each group sub sets are recorded, and the accuracy difference between adjacent learning time units is calculated. The effective fluctuation points are obtained according to the testing accuracy decline threshold. Then, the effective fluctuation points in different groups are integrated by cross checking to eliminate the detection interference caused by the instability of the model due to the small training samples in the online learning process of streaming data, and the consistent fluctuation points are obtained according to the consistency of accuracy fluctuation. Finally, by tracking the classification accuracy of online learning, the change of testing accuracy can be achieved of neighborhood reference points of consistent fluctuation points, and the decline and convergence of model testing accuracy can be compared of neighborhood reference points of consistent fluctuation points, so as to effectively detect the true concept drift points of the consistent fluctuation points. The experimental results demonstrate that the proposed CDPT method can effectively identify the true concept drift occurring in the online learning process of streaming data, effectively avoid the negative impact of too small training samples or noise on the detection results, and improve the generalization performance of the model.

Key words: streaming data; concept drift; cross checking; effective fluctuation point; consistent fluctuation point; concept drift point

新兴技术的迅猛发展使得海量数据不断涌现,如新闻媒体、物联网和云计算等不同行业产生的业务数据在一般场景中均可视为动态流数据[1].这些数据中包含大量有效信息亟待进行挖掘,然而,丰富的数据用于训练虽然能够降低风险,但在训练阶段将有大量时间损耗,且使计算复杂化.理想情况下,将每个样本都用于模型训练对分类性能会有很大提高,然而大量可用数据将会使训练集非常大以至于无法全部加载到内存中[2].针对传统数据挖掘方法在解决流数据挖掘问题时存在学习时间长、学习效率低等问题,一个有效的解决方式是使用在线学习技术,即每次训练都只将批量的流数据子集作为训练集[3].一般来说,按时间依次产生的时间序列数据,因数据量太大无法全部加载到内存中的大数据集都可用在线学习技术来进行处理.

就流数据的实时、多变性而言,并不能保证每次到达的新样本与前一次到达的样本完全保持独立同分布,因此不能每次都将由历史数据训练得到的分类模型作为后续样本的基准分类器[4,5].在训练模型阶段的每个新步骤中,当训练子集分布发生改变时,数据空间的底层结构也随之发生变化,从而导致概念漂移的发生[6].而当前由历史数据训练得到的分类器不符合新样本的分布规律,因此会引起分类精度下降[7].另外,流数据中存在噪声样本是不可避免的,尤其是当噪声规模较大时,造成分类器准确率大幅下降,其表现形式类似于发生真实的概念漂移,这种由噪声引起的伪概念漂移将严重干扰真实概念漂移的准确检测,即检测模型容易混淆真伪概念漂移而发生误判,从而导致模型对真实概念漂移或噪声引起的伪概念漂移做出错误的应对和处理.此外,在流数据学习过程中,若当前步骤中保留的训练样本信息有限,即当前所包含的训练样本信息无法代表全部样本的分布时,模型也会产生较大幅度的性能波动,此时,也会引起概念漂移的误检.

本文针对混杂噪声或训练样本规模过小而产生的伪概念漂移会引起与真实概念漂移混淆的问题,提出一种基于在线性能测试的概念漂移检测方法.该方法首先对当前得到的最新数据集进行分组学习,跟踪模型分类准确率并计算相邻时间单元之间的精度落差,通过衡量测试精度落差向量不同分量的大小,动态得到不同学习阶段每组数据流中的有效波动位点.然后根据不同分组中有效波动位点的同步情况,整合不同分组中的有效波动位点,即寻找一致波动位点.最后通过实时监测一致波动位点的邻域参考点精度变化及收敛趋势,比较一致波动位点的测试精度降幅及其邻域参考点的精度与收敛性,以有效检测一致波动位点中真实的概念漂移位点.本文主要提出了一种精确检测并辨识真伪概念漂移的方法,能够有效去除数据流中噪声数据以及训练集过小等问题对正常概念漂移检测的干扰,适用于常见的流数据在线学习及其概念漂移检测问题.

1 相关工作

在实际数据挖掘问题中,经常会遇到具有很长时间周期跨度的流数据,对于这类数据,其不同时刻的数据不能保证符合独立同分布的特性,特别是在复杂多变的环境中,样本实例随时间发生分布变化而引发概念漂移的概率很高.现实中,概念漂移的发生可能由多种原因引起,如在垃圾邮件过滤问题中,一方面,垃圾邮件的特征范畴可能随时间发生变化,另一方面,用户对垃圾邮件的定义也可能根据其喜好而有所改变,这两方面均可视为数

据流发生了概念漂移[8].再如天气或水温监测、传感器数据监测[9]、设备故障诊断等[10,11],这些现实问题产生的数据流何时发生漂移或是否发生漂移往往是提前未知的,因此准确检测概念漂移是否发生对处理许多现实问题具有重要的意义和价值[12].

除流数据本身可能发生的概念漂移外,其他因素也可能导致流数据在线学习过程中测试性能的波动.一方面,真实数据流中噪声是客观存在且不可避免的,噪声样本因偏离正常样本分布太多,会干扰流数据在线学习过程中的模型调整,导致对一些正常分布的数据分类错误,准确率降低.另一方面,在线学习过程中,当前的学习模型所包含的训练样本规模过小,得到的分类器不能稳定地表示数据的整体空间特征分布,因此,每次训练都可能会使模型发生较大改变,导致分类结果产生不稳定波动.这两者虽然不是概念漂移,但却与发生概念漂移产生了类似的测试精度波动的结果.与流数据中真实概念漂移相对应,这种非数据分布改变而导致的分类器波动称为伪概念漂移.流数据中由噪声或训练样本规模过小而导致的伪概念漂移将直接威胁到真实概念漂移的检测,降低在线学习模型的分类质量.尤其是当噪声样本含量较大或包含有效样本信息过少时,模型测试精度下降幅度增加,更容易误判为发生真实概念漂移,严重干扰了对真实概念漂移的检测结果.

目前,在处理流数据概念漂移问题中已经取得了很多研究成果.已有的处理方法大致可以分为两类:一是基于滑动窗口方法,二是自适应集成学习方法.滑动窗口方法通过时间窗口将当前训练数据限制在最新进入模型的样本中,并利用当前窗口中的新数据训练新的分类器,在预测时直接使用该分类器进行分类.在处理流数据概念漂移时,启发式地动态调整窗口大小,提高分类器适应能力.常见的方法有基于单窗口的概念漂移检测方法(SWCDS)[5].该算法通过周期性地检测滑动窗口中的分类错误率变化检测数据分布的变化.Du等人在2014年提出依据信息熵判定窗口新旧实例分布调整窗口方法(ADDM)[13],Ali 等人于 2016 年提出了加速 Hoeffding 漂移检测方法(FHDDM)[14]等.虽然以窗口为对象的检测方法在一定程度上提高了概念漂移检测的性能,但窗口大小直接决定了分类器的性能.较小的窗口敏感度较高,能够迅速对概念漂移做出反应,但在检测到漂移位点的同时也会对噪声等引起的不稳定波动做出反应.而较大窗口则难以及时检测到概念漂移,导致漂移检测滞后.

对于分类器建模过程,构造多个简单分类器比建立一个复杂单分类器更简单、易行.因此,许多研究者采用自适应集成学习方法处理流数据中的概念漂移问题.这种方法利用新到达的数据块训练一个基分类器,采用不断更替模型中性能较差的基分类器以保证系统中保存固定数目的较优基分类器,并使分类器能够适应概念漂移数据流.但这种方法只保证了分类器能够较好地适应最新的数据分布,却不能准确地检测到概念漂移何时发生.SEA 算法[15]通过分类器精度的阈值不断更新基分类器,以此保证分类器能够适应概念漂移.算法 AWE[16]在算法 SEA 的基础上进行改进,根据各个基分类器对当前数据块的分类精度为它们设置权值,更新分类器时直接替换权重最小的基分类器.CDOL 算法[17]由两个带权基分类器组成,检测到概念漂移后用新样本训练新的基分类器并移除原有基分类器中权重较小者.自适应集成学习方法仅根据各基分类器对当前数据块的分类精度来评估要删除哪些基分类器,忽略了基分类器的历史重要性.另外,一次小的概念波动可能导致历史上重要的基分类器被误删,造成保留的基分类器不是全局占优,难以获得好的预测结果[18].

检测概念漂移的一种主流思想是依据分类模型测试精度变化趋势判断漂移是否发生[19,20].一般来说,这种思想需要保证分类模型必须对新增样本做出决策之后才能继续进行下一次训练,以便检测器及时分析分类精度的变化趋势[21].通过监视分类模型的表现并结合某些决策方法以判断数据分布的变化情况,设置适当的置信水平控制预警信号,预测可能发生漂移的位点.然而,若预警信号着重关注于模型测试准确率的下降幅度,则一旦满足预先设置的临界值条件就会触发预警信号.如果没有进一步追溯下降原因,就会导致当训练集中噪声数据量足够多或者训练样本信息明显缺失时引起的伪概念漂移现象也被当作真实概念漂移来处理.而实际上,伪概念漂移与真实概念漂移存在本质上的差异,对应处理机制也不相同.真实概念漂移处理侧重于新样本分析及分类模型的实时更新,并通过摒弃历史数据来减少原始样本对模型更新的抑制作用,或者保存由历史数据训练所得到的分类器,以免当前数据分布形式再次出现[22].相反地,伪概念漂移处理方法需要利用历史数据并引进新的更多的样本,以降低训练集中噪声或样本规模过小对模型检测性能的负面效应,进而保证训练样本空间中数据分布及概念的纯洁性和有效性.将真实概念漂移处理机制作用于伪概念漂移会严重影响模型的收敛,降低模

型的性能,因此有必要在概念漂移检测的过程中对伪概念漂移进行辨识和区分.

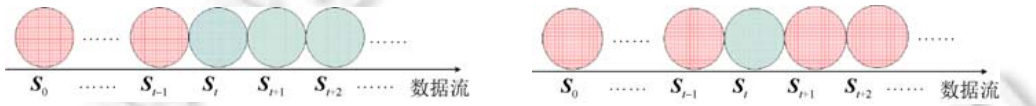
针对概念漂移检测机制中容易混淆真伪概念漂移的问题,本文提出一种有效检测概念漂移方法.主要工作如下.

1) 给出发生真伪概念漂移的特征描述,将引起分类模型准确率的不稳定波动原因分为数据发生概念漂移和数据中存在噪声或训练数据规模过小而引起的伪概念漂移,明确发生真实概念漂移与伪概念漂移之间的本质差异.

2) 以给出的真伪概念漂移特征描述为基础,提出基于在线性能测试的流数据概念漂移检测方法,包括有效波动位点检测、一致波动位点提取和概念漂移位点判定.该方法对当前获得的最新数据进行均匀的分组学习,然后根据相邻时刻模型精度差异来检测流数据中发生的有效波动位点,并通过交叉检验的方式提取其中的一致波动位点,最后跟踪一致波动位点的邻域参考点精度变化趋势,比较精度收敛偏差以准确检测概念漂移.

2 真伪概念漂移描述及分析

假设存在流数据序列 $\boldsymbol{S}=\{(x_t,y_t)\}_{t=0}^{T}$, 其中,$x_t\in\boldsymbol{R}^n$ 和 $y_t\in\{-1,+1\}$,流数据学习过程中样本可以逐个进入,也可分批到达,为简化问题描述,不妨将每次到达的样本集表示为 $\boldsymbol{S}_t=\{\boldsymbol{X}_t,\boldsymbol{Y}_t\}$.对于真实的概念漂移,漂移发生后新进入的样本与历史数据分布不一致,其在线学习样本序列为 $\boldsymbol{S}=\{\{\boldsymbol{X}_0,\boldsymbol{Y}_0\},\{\boldsymbol{X}_1,\boldsymbol{Y}_1\},\ldots,\{\boldsymbol{X}_{t-1},\boldsymbol{Y}_{t-1}\},\{\boldsymbol{\mathcal{X}}_t,\boldsymbol{\mathcal{Y}}_t\},\ldots,\{\boldsymbol{\mathcal{X}}_T,\boldsymbol{\mathcal{Y}}_T\}\}$, 其中,$\{\boldsymbol{X}_0,\boldsymbol{Y}_0\}$为初始离线样本, $\{\boldsymbol{X}_i,\boldsymbol{Y}_i\}_{i=0}^{t-1}\in P$,$\{\boldsymbol{\mathcal{X}}_j,\boldsymbol{\mathcal{Y}}_j\}_{j=t}^{T}\in\mathcal{P}$, 且 P 和 $\mathcal{P}$ 分别代表原始数据分布和发生概念漂移后的数据分布.伪概念漂移区别于真实概念漂移,并未发生数据分布改变,而是由于数据中存在噪声或训练样本包含有效信息过少等其他因素的存在导致分类器测试精度出现不稳定波动.以噪声引起的伪概念漂移为例,其对应的在线学习样本序列可以表示为 $\boldsymbol{S}=\{\{\boldsymbol{X}_0,\boldsymbol{Y}_0\},\{\boldsymbol{X}_1,\boldsymbol{Y}_1\},\ldots,\{\boldsymbol{X}_{t-1},\boldsymbol{Y}_{t-1}\},\{\boldsymbol{\mathcal{X}}_t,\boldsymbol{\mathcal{Y}}_t\},\{\boldsymbol{X}_{t+1},\boldsymbol{Y}_{t+1}\},\ldots,\{\boldsymbol{X}_T,\boldsymbol{Y}_T\}\}$, 其中, $\{\boldsymbol{X}_i,\boldsymbol{Y}_i\}_{i=0}^{t-1}\in P$ 且 $\{\boldsymbol{X}_j,\boldsymbol{Y}_j\}_{j=t+1}^{T}\in P$, 而 $\{\boldsymbol{\mathcal{X}}_t,\boldsymbol{\mathcal{Y}}_t\}$表示偏离原始数据分布的噪声样本.为方便描述,将噪声样本表示为数据分布为 $\mathcal{P}$ 的样本,即 $\{\boldsymbol{\mathcal{X}}_t,\boldsymbol{\mathcal{Y}}_t\}\in\mathcal{P}$, 两种流数据模式可用图 1 表示.



(a) Schematic of true concept drift in streaming data (b) Schematic of false concept drift in streaming data (by noise)

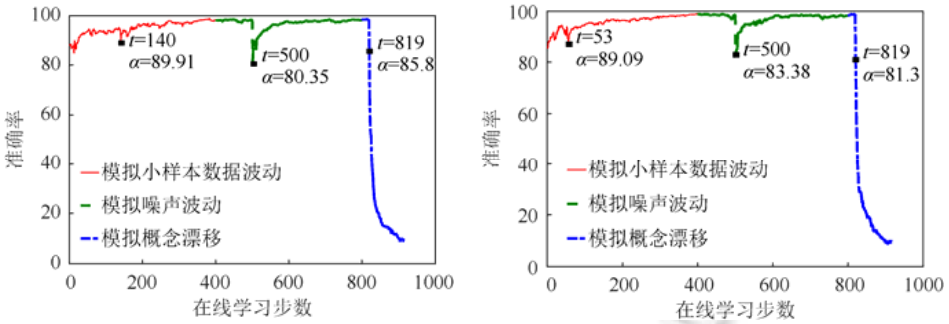
Fig.1 Schematic of concept drift in streaming data

图 1 流数据真伪概念漂移示意图

图 1 中样本数据 $\boldsymbol{S}=\{(\boldsymbol{X}_t,\boldsymbol{Y}_t)\}_{t=0}^{T}$ 沿时间轴按序到达并用于训练模型,其中,每个圆代表某一时间单元内到达的一批新样本 $\boldsymbol{S}_t=\{\boldsymbol{X}_t,\boldsymbol{Y}_t\}$,网格圆形和点阵圆形分别表示符合两种不同分布 P 和 $\mathcal{P}$ 的数据.假定在 t 时刻发生概念漂移,图 1(a)展示了流数据发生真实概念漂移的情况,从 t 时刻开始每个批次的数据均发生改变,即从分布 P 改变至分布 $\mathcal{P}$ 且短时间内不可逆,在分布变化发生后原始模型不再适用于新数据分布,此时必须利用新样本重新训练分类模型,以便尽快恢复模型分类性能[23].图 1(b)展示了含噪声的流数据产生伪概念漂移的形式,假定在 t 时刻到达的新样本 $\{\boldsymbol{\mathcal{X}}_t,\boldsymbol{\mathcal{Y}}_t\}\in\mathcal{P}$ 与原始数据分布 P 产生偏差,给学习器带来干扰性波动,但在 t+1 时刻数据 $\{\boldsymbol{X}_{t+1},\boldsymbol{Y}_{t+1}\}\in P$ 恢复原始分布状态,当在线学习的数据单元 $\boldsymbol{S}=\{\boldsymbol{X}_t,\boldsymbol{Y}_t\}$包含样本规模较小时,$t$ 时刻到达的样本在训练集中只占较小比例,因此学习器能够在后续时间单元内尽快恢复,但当噪声量较大时,这些异常数据将作为代表性样本大幅度更新模型,因此需要经过后续多步在线学习过程中新进入的样本对模型进行持续性校正,模型收敛速度较慢.

另一方面,在流数据在线学习初始阶段,训练样本规模较小,训练集未能覆盖完整的数据分布信息,因此,模型测试精度会产生不稳定波动.但这种情况与真实概念漂移及噪声引起的伪概念漂移存在明显不同,如在同一数据流中,假定该数据流在特定位置发生概念漂移,则由于训练数据规模过小而产生的波动位置会随着数据流特征随机产生,而真实概念漂移所引发的测试精度波动位点是固定的.例如,图 2 所示为将 UCI 上的数据集 *Image* 均匀划分为 5 组,两个子图分别对应随机取其中 4 组进行训练,剩余样本进行测试得到的结果(参数设置:

惩罚参数 C=1000,在线学习步长 O=10).图中横轴表示在线学习的每个学习步骤,其中,初始点 t=0 代表在初始的离线数据集上训练及测试得到的模型精度,在线学习的过程从 t=1 时刻开始,纵轴表示分类模型的测试精度.



(a) Simulation on the first group data (b) Simulation on the second group data

Fig.2 Comparison of model prediction accuracy

图 2 模型预测精度对比图

图 2 中,在(0,400)步为初始学习阶段,这个阶段内模型因训练样本规模较小,学习状态不稳定,不同分组数据得到的测试精度波动位点差别较大,如图 2(a)中 t=140,图 2(b)中 t=53 处分别存在有效波动位点且两者不一致.而在(400,800)的在线学习步内某个时刻(如 t=500)加入噪声,图 2(a)和图 2(b)中显示模型测试精度在 t=500 时刻均发生了突然下降.而在(800,915)内模拟真实概念漂移,并在某个时刻(如 t=819)突然改变新增样本分布,两次模拟结果均在 t=819 处检测到真实的概念漂移位点.因此可以看到,图 2 中由于训练样本规模小而引起的分类器测试精度波动在不同分组中会在不同位点被检测到.基于此,本文采用分组交叉检验的方法,消除由于训练样本规模过小而导致的分类器测试性能波动给概念漂移检测带来的负面影响,同时又不影响模型对于真伪概念漂移的检测过程.

3 基于在线性能测试的概念漂移检测

本文以流数据在线学习中由数据源分布变化引起的真实概念漂移以及由噪声样本或训练集规模过小而引起的伪概念漂移为研究对象,提出了基于在线性能测试的概念漂移检测方法(CDPT),包括有效波动位点检测、一致波动位点提取和概念漂移位点判定.有效波动位点检测环节通过监测在线学习分类器测试性能的变化,采用分段动态提取的方式获得测试精度波动较大的位点作为有效波动位点;一致波动位点提取环节通过对当前步骤中已获得的数据进行均匀分组,并检测每组数据集上的有效波动位点,然后取不同分组中的相同有效波动位点作为一致波动位点;概念漂移位点判定环节通过跟踪检测到的一致波动位点之后近邻参考点的精度变化及收敛趋势,比较漂移位点精度收敛偏差以判别概念漂移位点.本文中涉及到的有效波动位点、一致波动位点以及概念漂移位点之间的关系如图 3 所示.

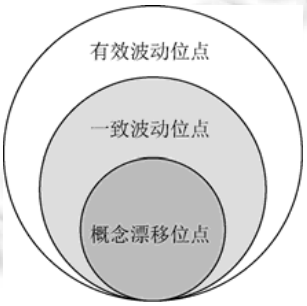


Fig.3 Relationship of different fluctuation points

图 3 不同波动位点的关系图

3.1 有效波动位点检测

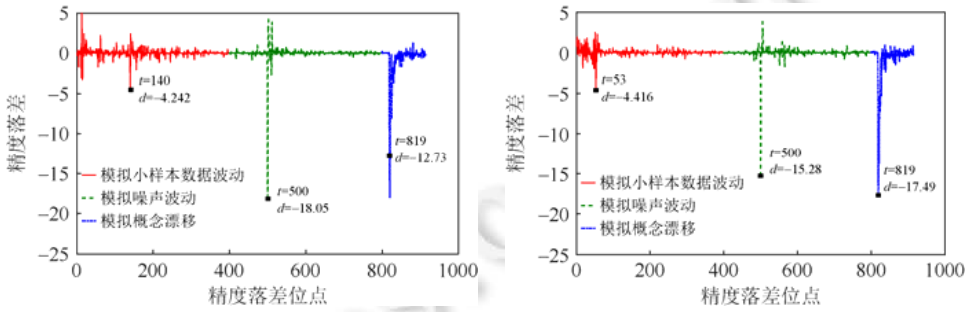
在流数据在线学习过程中,数据分布的改变可能导致分类器测试精度产生波动,即发生概念漂移;反之,当数据分布一致时,模型会稳步更新,即分类器测试精度会保持相对稳定.假设第 t 步训练测试得到的模型精度为 α_t,则第 t–1 步训练测试得到的模型精度为 α_{t-1},那么第 t 步与第 t–1 步的模型测试精度落差为

$$d_t = \alpha_t - \alpha_{t-1} \tag{1}$$

则一个完整的流数据在线学习过程中得到的一系列模型测试精度落差可表示为一个向量:

$$\boldsymbol{D} = \{d_t\}_{t=1}^{T} \tag{2}$$

图 4 展示了第 2 节中 *Image* 数据集上两次模拟流数据在线学习的测试精度落差.



(a) Accuracy drop on the first experiment (b) Accuracy drop on the second experiment

Fig.4 Comparison of model accuracy drop

图 4 模型精度差异度对比图

图 4 中所标注的模型出现明显测试精度落差现象的位置,均已在图 2 中被标注.另外,流数据在线学习过程中不同时间段模型波动情况不同,在初始学习阶段,样本信息积累较少,模型测试精度波动较大,但是随着在线学习进程的推进,在样本信息规模较大时,模型变得相对稳定.因此,为了能够更有效地挖掘不同学习阶段内的漂移现象,本文将测试精度落差向量根据学习时间步骤平均分为 3 个子向量,即 $\boldsymbol{D}=[\boldsymbol{D}_f;\boldsymbol{D}_m;\boldsymbol{D}_l]$(即分别对应在线学习过程的初期、中期和后期).在每个子向量中,采用动态替换的方式选择 n 个有效波动位点,即随着流数据挖掘的推进,首先依次提取 n 个测试精度下降位点构成初始波动位点集,并记录初始波动位点集中对应的测试精度下降幅度的最小值 $d_{\min}$,然后继续选择后续的测试精度下降位点,若后续位点的测试精度下降幅度大于当前波动位点集合中的最小下降幅度 $d_{\min}$,则采用对应的后续位点代替当前波动位点集中的最小下降幅度对应的波动位点,并更新波动位点集中对应的最小下降幅度 $d_{\min}$,即更新波动位点集,如此循环往复,即可提取得到每个子向量上对应的 n 个下降幅度最大的波动位点,构成相应的有效波动位点集.当数据集规模较大时,由于每个下降位点都需要进行比较,因此这种动态比较次数会较大,位点的替换次数可能较多.此时,可以设置一个精度落差临界阈值 $\varepsilon>0$,只有当精度落差超过 ε 时,才认为概念处于非稳定状态,即当前位点作为波动位点参与比较和动态替换的过程,反之,则认为概念处于稳定状态,不作为波动位点处理(当然,虽然这种增加约束的动态调整替换方式提高了有效波动位点的提取效率,但有可能导致某些区间上提取到的有效波动位点个数小于 n,此时,一方面可以通过减小精度落差临界阈值 ε 来增加提取的波动位点个数,也可以直接取实际提取的波动位点).记有效波动位点集合为 $\boldsymbol{E}$,可知,若满足式(3)的条件,则认为概念处于非稳定状态,且模型测试精度落差位点为波动位点.

$$\begin{cases} d_t < 0 \\ |d_t| > \varepsilon \end{cases}, \quad t = 1,...,T \tag{3}$$

在实际应用中,每个区间的有效波动位点个数 n 应该设置成一个恰当值,n 值过大,则有可能导致提取到的一致波动位点过多,后续概念漂移检测的过程变得复杂且效率较低.反之,若 n 值太小,则可能检测不到实际存在的概念漂移位点.在处理实际的流数据挖掘问题时,可以结合问题背景设置合适的有效波动位点提取个数参数,

比如类似于天气预报的流数据、股票检测的流数据,数据发生概念漂移的模式比较频繁,则可以设置较大的有效波动位点提取个数参数 n 值.反之,在类似于系统故障诊断等一般情况下都比较稳定的流数据挖掘任务中,则更应设置较小的有效波动位点提取个数参数.假设 3 个阶段内提取的有效波动位点集合分别表示为 $\boldsymbol{E}_f$、$\boldsymbol{E}_m$ 和 $\boldsymbol{E}_l$,则整个数据集上的有效波动位点集合为

$$\boldsymbol{E}=\boldsymbol{E}_f\cup\boldsymbol{E}_m\cup\boldsymbol{E}_l \tag{4}$$

3.2 一致波动位点提取

由第 2 节分析可知,由于训练样本规模过小而引起的模型测试精度波动是随机的,为消除其对概念漂移检测的影响,本文采用交叉检验的方式提取一致波动位点,以减少由训练样本规模过小而导致的模型不稳定波动带来的问题.假设在线的训练样本集为 $\boldsymbol{S}=\{(\boldsymbol{X}_t,\boldsymbol{Y}_t)\}_{t=0}^{T}$,将该训练集平均划分为 k 个互不重叠的子序列,且每组数据集中包含相同规模的不同类别样本,即:

$$\boldsymbol{S}=\bigcup_{i=1}^{k}\boldsymbol{S}^i \tag{5}$$

采用第 3.1 节中的有效波动位点提取方法分别提取每个分组中的 $3n$ 个有效波动位点,不妨假设子数据流 $\boldsymbol{S}^i$ 的有效波动位点集合为 $\boldsymbol{E}^i$,然后整合不同分组中的有效波动位点集合,以消除训练样本规模过小而带来的精度波动问题,整合过程为

$$\boldsymbol{W}=\bigcap_{i=1}^{k}\boldsymbol{E}^i \tag{6}$$

其中,$\boldsymbol{W}$ 为一致波动位点集.本文采用 5 折交叉检验的方式进行一致波动位点的提取.另外,本文的方法尽管在第 3.1 节提取到的有效波动位点较多,但经过一致波动位点提取后,一方面,避免了训练样本过小导致模型测试性能随机波动带来的概念漂移误检问题,另一方面,也大幅度减少了实际处理的波动位点个数,提高了后续概念漂移检测的效率.

3.3 概念漂移位点判定

流数据发生真实概念漂移表明,当前由历史数据训练得到的分类模型不再适用于新数据,假定发生概念漂移后原始数据分布不再出现,可知旧模型在新样本中得到的测试精度将持续降低,最终趋于稳定,如图 2 所示中(800,1000)步内所展示的模型测试精度趋势,即使在某个时刻精度有所波动,也很难恢复到最初状态.而噪声数据的表现形式则不同,其所得到的模型测试精度的下降位点只存在于某一时刻或某一时间段内,当样本恢复原始分布后,模型测试精度也会相应提升.

基于上述分析,概念漂移位点的判定可通过一致波动位点的后续近邻参考点对应的测试精度下降及收敛情况进行判定.这里定义一个模型测试精度的收敛偏差 P,以记录特定的一致波动位点后不同近邻参考点对应的模型测试精度的恢复情况,以实时地区分当前一致波动位点处概念漂移的真伪性.假设当前检测到的一致波动位点为 w,则在其后续在线学习过程中,第 t 步近邻参考点在线学习的测试精度收敛偏差 $P_t(w)$定义为

$$P_t(w)=\frac{\alpha_t-\alpha_w}{\alpha_{w-1}-\alpha_w} \tag{7}$$

其中,α_t 为一致波动位点 w 的近邻参考点 t 处模型得到的测试精度,α_w 为一致波动位点处模型的测试精度,α_{w-1} 为一致波动位点前一步所得到的测试精度.依据真实概念漂移与噪声引起的伪概念漂移在模型收敛速度上的差异性,设置一致波动位点的近邻参考点精度收敛偏差临界阈值参数 τ,比较一致波动位点的近邻参考点对应的收敛偏差与临界阈值之间的关系,若当前近邻参考点的测试精度收敛偏差大于给定阈值,则从当前近邻参考点来看,所对应的一致波动位点更趋向于发生了伪概念漂移,反之,则认为所对应的一致波动位点更趋向于发生了真实概念漂移.

对于收敛偏差的临界阈值参数设定,本文采用动态放缩的方法,且只考虑一致波动位点后续的 5 个近邻参考点.一致波动位点 w 的近邻参考点 t 的测试精度收敛偏差 $P_t(w)$的临界阈值参数 τ_t 的设定方法可以表示为

$$\tau_t=\begin{cases}0, & t=w+1\\ \dfrac{\alpha_{t-2}-\alpha_{t-1}}{\alpha_{w-1}-\alpha_w}, & t=w+2,...,w+5\end{cases} \tag{8}$$

采用这种动态调整的方法进行收敛偏差临界阈值的设定,有助于进一步提高检测的效能.在概念漂移位点判定过程中,对于某个一致波动位点 w,若前驱近邻参考点 t–1 判定其为真实概念漂移,说明其具有明显的精度下降现象.为了后续近邻参考点 t 判定 w 为真实概念漂移的概率有所增加,则应进一步扩大判定下降幅度,当精度下降幅度足够大时,说明真实概念漂移发生.反之,若前驱近邻参考点 t–1 判定一致波动位点 w 为伪概念漂移时,说明当前位点精度具有明显的回升.为了使得后继近邻参考点 t 判定 w 为伪概念漂移的概率有所增加,则应进一步缩小判定恢复程度,也即增大临界阈值.当精度仍具有回升趋势时,可判定为发生伪概念漂移.

对于提取到的一致波动位点 w,依次计算每组子数据流 $\boldsymbol{S}^i$ 中其近邻参考点的判定结果,经过 5 个近邻参考点的判定,若对于当前的一致波动位点 w,k 组子数据流得到的概念漂移位点平均判定值 $Cd(w)$大于 50%,则认为当前一致波动位点为真实的概念漂移位点,否则认为,当前一致波动位点为由噪声引起的伪概念漂移位点.一致波动位点的平均判定值计算可表示如下:

$$Cd(w)=\frac{1}{k}\sum_{i=1}^{k}\frac{\sum_{t=w+1}^{w+5}f(P_t^i(w)\leqslant\tau_t^i)}{5} \tag{9}$$

其中,$P_t^i(w)$ 代表一致波动位点 w 在子数据流 $\boldsymbol{S}^i$ 中的近邻参考点 t 处的测试精度收敛偏差,τ_t^i 代表子数据流 $\boldsymbol{S}^i$ 中近邻参考点 t 对应的测试精度收敛偏差临界阈值参数,函数 $f(X)$为条件判别函数,即当条件 X 成立时,该函数取值为 1,否则,取值为 0.这里需要注意区分标号 t 和 i,其中,标号 t 的取值从 w+1 到 w+5 代表一致波动位点的近邻参考点,它们用来判定一致波动位点是否为真实的概念漂移位点;而标号 i 的取值从 1 到 k 代表对整个流数据集均匀划分为 k 个子数据流,即通过采用交叉检验的方法消除由于训练样本规模过小而带来的随机波动.

3.4 CDPT算法

本文提出的基于在线性能测试的概念漂移检测方法通过分析真实概念漂移、由噪声引起的伪概念漂移及由训练样本规模过小导致的伪概念漂移之间的异同,并充分利用分类模型在相邻在线学习时间单元中预测精度表现出的不同趋势,对概念漂移进行检测.由于支持向量机(support vector machine,简称 SVM)建立在传统学习理论和结构风险最小化基础之上,具有良好的泛化性,较强的鲁棒性等优势[24],因此本文在模拟流数据在线学习实验的过程中,以 SVM 作为基准分类器,以 SVM 增量学习方式进行在线学习.本文所提出的方法构造了模型测试精度波动的落差向量,通过动态调整替换的方式提取有效波动位点;通过交叉检验,比较不同分组波动位点的一致性,提取一致波动位点;最后,通过一致波动位点近邻参考点的模型测试精度恢复情况,来判别最终的概念漂移位点.具体算法如下.

算法. 基于在线性能测试的概念漂移检测.

初始化. 流数据序列 $\boldsymbol{S}=\{(\boldsymbol{X}_t,\boldsymbol{Y}_t)\}_{t=0}^{T}$, 其中,$\{\boldsymbol{X}_0,\boldsymbol{Y}_0\}$为初始离线样本,$\boldsymbol{X}_t\in\boldsymbol{R}^n$ 和 $\boldsymbol{Y}_t\in\{-1,+1\}$,测试集 $\boldsymbol{TS}$,其与初始训练集符合独立同分布;

Step 1. 将流数据 $\boldsymbol{S}$ 按照式(5)划分为 k 个子序列,即: $\boldsymbol{S}=\bigcup_{i=1}^{k}\boldsymbol{S}^i$;

Step 2. 分别对每个子序列并行进行在线学习,记录每次学习得到的准确率α;

Step 3. 根据式(1)计算得到每个子序列 $\boldsymbol{S}^i$ 对应的测试精度落差向量 $\boldsymbol{D}^i$,并将其根据学习时间步骤平均分为 3 个子向量,即 $\boldsymbol{D}^i=[\boldsymbol{D}_f^i;\boldsymbol{D}_m^i;\boldsymbol{D}_l^i]$,然后采用第 3.1 节中动态替换的方式在线提取每个子向量中满足式(3)的 n 个最大值(不足 n 个的取实际满足式(3)的所有位点),然后按照式(4)得到每个子序列 $\boldsymbol{S}^i$ 上对应的有效波动位点集合 $\boldsymbol{E}^i$;

Step 4. 根据式(6)得到整个数据集上的一致波动位点集合 $\boldsymbol{W}$;

Step 5. 根据式(7)计算每个一致波动位点 w 后续 5 个近邻参考点 t 的收敛精度偏差 $P_t(w)$,并采用式(8)动态设置精度收敛偏差阈值τ,对于每个一致波动位点 w,根据每个子数据流中收敛精度偏差和临界值τ的大小关系,

根据式(9)计算一致波动位点对应的平均判定值 $Cd(w)$,若其大于 50%,则该一致波动位点为预测的概念漂移位点,否则该一致波动位点为噪声引起的伪概念漂移位点;

Step 6. 分析数据流中概念漂移的检测情况,即将检测到的概念漂移位点与数据集中存在的真实概念漂移位点及噪声位点进行对比分析,这里,所有在真实概念漂移位点近邻范围内捕获的预测概念漂移位点均被认为准确检测到了概念漂移;

Step 7. 算法结束.

4 实验与性能分析

为验证本文所提出的真伪概念漂移检测算法中近邻参考点对于概念漂移位点检测的平均判定值及概念漂移的检测情况(误检、漏检、延时等),本文分别在 7 个标准数据集和 3 个真实数据集上对算法进行了测试.实验平台为 Windows 7,CPU 为酷睿 i5-3210,内存为 8GB,开发环境为 Matlab 2016b.实验中,将本文提出的基于在线性能测试的概念漂移检测方法与目前主要的两类概念漂移检测方法,即基于滑动窗口的概念漂移检测方法 SWCDS[5]和基于自适应集成学习的漂移检测算法 CDOL[17]进行了对比.

4.1 数据集

1) UCI 标准数据集:选择 UCI 数据集的目的在于其可以方便地模拟各种存在概念漂移及噪声的环境,设置真实的概念漂移发生位点及由噪声引起的伪概念漂移位点,有利于发现算法在不同噪声环境下的概念漂移检测性能.为确保数据环境的稳定,在训练集中利用正反例互换形式使得每个数据集中存在两种不同概念.具体所用到的 7 个数据集的详细信息见表 1.

Table 1 UCI datasets used in experiment

表 1 实验中使用的 UCI 数据集

Datasets	Features	Categories	Size of data
Banana	2	2	9 800
Breast_cancer	9	2	2 770
Diabetis	8	2	7 680
German	20	2	5 000
Image	18	2	11 550
Spambase	57	2	4 000
Thyroid	2	2	4 300

2) 真实数据流:实验选择了真实数据库 MITface 数据集、Usps 数据集[17]以及 KDDCup99[25]数据集.其中,MITface 来自 MIT 的人脸数据库,Usps 来自美国邮政手写数据集,KDDCup99 数据集是网络入侵检测竞赛的测试数据,该数据集常用于模拟漂移数据验证检测漂移算法的有效性,实验中对该数据集进行抽样压缩,抽样后各类别数量比例仍与原数据集相同.表 2 给出了 3 个真实数据集的具体信息描述.

Table 2 Realistic datasets used in experiment

表 2 实验中使用的真实数据集

Datasets	Features	Categories	Size of data
KDDCup99	41	5	24 000
MITface	361	2	6 977
USPS	256	2	2 930

4.2 评估指标

本文主要从如下 3 个方面评价方法的性能.

首先,对于概念漂移位点判定的步骤中,采用平均判定值(见式(9))来衡量采用近邻参数法进行真伪概念漂移判定时,近邻参考点对于概念漂移位点的平均决策值.

其次,对于最终的概念漂移检测结果,本文提出漏检率(MDR)和误检率(FDR)两个指标进行度量,漏检率用于衡量实际的概念漂移位点中没有被检测到的概率;误检率代表检测到的概念漂移位点中没有发生概念漂移的概率.在概念漂移检测问题中,大多数模型携带部分旧样本信息,因此检测到的正确概念漂移位点多数情况下

会比实际插入概念漂移的位点有一定延时,本文把实际插入概念漂移的位点后续 5 个近邻检测到的概念漂移位点均视为正常检测.假设实验中得到的一致波动位点集合为 $\boldsymbol{W}$,其中,根据近邻参考点判定为概念漂移的一致波动位点集合为 $\boldsymbol{W}^T$.假设真实的概念漂移插入位点为 $\boldsymbol{C}=\{c_i\}_{i=1}^{3}$ (其中,c_1、c_2、c_3 分别为在流数据前期/中期/后期不同阶段插入的真实概念漂移位点),则漏检率为

$$MDR=\frac{\left|\{c_i\left|w^t\notin\boldsymbol{C}_i[5]\ \&\ w^t\in\boldsymbol{W}^T\}\right.\right|}{|\boldsymbol{C}|} \tag{10}$$

其中,$\boldsymbol{C}_i[5]$代表由实际插入概念漂移位点 c_i 及其连续后继 5 个位点构成的集合,即 $\boldsymbol{C}_i[5]=\{c_i+j\}_{j=0}^{j=5}$,因此漏检率即反映了实际插入概念漂移的位点中在允许的延时范围内均未检测到概念漂移的概率.而误检率则表示为

$$FDR=\frac{\left|\{w^t\left|w^t\notin\boldsymbol{C}_i[5]\ \&\ w^t\in\boldsymbol{W}^T\}\right.\right|}{|\boldsymbol{W}^T|} \tag{11}$$

误检率反映了判定为概念漂移的位点中不在实际插入概念漂移位点允许延时范围内的概率.显然,检测结果评测中,得到的漏检率和误检率越低,说明模型对于概念漂移检测和识别能力越好.

此外,由于概念漂移检测问题中,判定为概念漂移的位点多数情况下会比实际插入概念漂移的位点有一定延时,因此,实验中还衡量了实际插入概念漂移位点的检测延时,其定义如下:

$$Delay_{(c_i)}=\begin{cases}\min\limits_{w^t\in\boldsymbol{C}_i[5]}(w^t-c_i), & \exists w^t\in\boldsymbol{W}^T, w^t\in\boldsymbol{C}_i[5]\\ +\infty, & \forall w^t\in\boldsymbol{W}^T, w^t\notin\boldsymbol{C}_i[5]\end{cases} \tag{12}$$

概念漂移检测延时有效评估了在线检测概念漂移发生位点的实时性和精准性,延时越小,表明算法对于概念漂移的检测越敏感.由于每个真实的概念漂移位点及其连续后继 5 个位点中检测到概念漂移,均视为正确检测到了概念漂移,因此检测延时的取值范围为[0,5].

4.3 参数设置

本节对实验模型检测概念漂移影响参数进行讨论,以研究模型参数对方法性能的影响,进而实现模型的有效推广.概念漂移检测过程中对于有效波动位点的提取个数 n 决定着后续一致波动位点的提取数量以及得到的一致波动位点能否覆盖到真实概念漂移位点,它的设定对模型的检测性能起着重要作用.为了获得最为合适的 n 值,针对由不同 n 值实验后得到的 FDR 和 MDR 决策值进行分析讨论.为简洁起见,本文仅介绍 UCI 标准数据集 German 的具体分析过程.图 5 给出了不同 n 值情况下得到的 FDR 和 MDR 决策值(参数设置:惩罚参数 C=1000,在线学习步长 O=5).图中横轴表示提取有效波动位点个数 n,本文在测试过程中取值为 n={5,10,15,20},纵轴表示在对应取值情况下得到的 FDR 和 MDR.

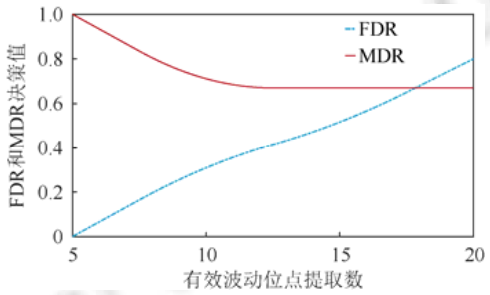


Fig.5 Trend charts of FDR and MDR under different n values

图 5 不同 n 值下 FDR 与 MDR 趋势图

从图 5 可以看出,随着有效波动位点提取个数的增加,FDR 呈持续上升趋势,相反地,MDR 呈现下降趋势.由上一节分析可知,FDR 和 MDR 结果越小,则模型对概念漂移检测性能越好.因此,为了让有效波动位点提取数 n

对评估结果起到正面作用,通过权衡 n 的取值与评估标准之间的变化趋势关系,将有效波动位点提取个数 n 值取[10,20]区间内的整数是较好的选择.考虑到 n 的取值越大,检测到的有效波动位点越多,增加计算时间,取值太小又不能完全覆盖真实概念漂移位点,在本文实验中,将每个学习阶段的有效波动位点提取个数取值为 15,即整个流数据上最大提取到的有效波动位点不超过 45 个.

因为本文实验采用的是动态替换的方式选择 n 个有效波动位点,因此根据第 3.1 节所给出的描述,实验中将得到的测试精度下降幅度的最小值 $d_{\min}$ 赋值给精度落差临界阈值 ε,即 $\varepsilon=|d_{\min}|$,然后继续进行后续的测试.当 $d_{\min}$ 更新之后,相应地也将 ε 进行更新.

为检验本文提出的概念漂移检测方法的性能,本文对不同大小的在线学习单元、不同概念漂移发生位置以及噪声对概念漂移检测的影响等多个方面进行了测试.首先,由于在线学习过程中,每个时间单元内到达的样本数量对于模型有较为明显的影响,因此,本文分别对每个时间单元内到达的样本数目为 $O=\{1,5,10,20\}$ 的各个子数据流进行测试.其次,由于数据流挖掘过程中,不同位置分类器所能采集到的样本数量和有效信息有一定差异,概念漂移位点所反映出来的模型测试性能也会有所不同,因此,实验中分别设置 $G=\{10,\lceil T/2\rceil,T-50\}$ 这 3 个位点作为概念漂移及噪声的插入位点(其中,T 为流数据中存在的学习单元个数),以反映本文方法在流数据分类前期/中期/后期不同阶段概念漂移检测的性能,后续的实验结果均为 3 个不同阶段位点测试结果的均值.此外,为分析噪声引起的伪概念漂移对结果的影响,以反映本文提出的基于在线性能测试的概念漂移检测方法对于噪声的鲁棒性,实验中在与概念漂移相同的位置插入噪声的同时进行了检测.实验中 SVM 采用高斯核,核参数 p 取默认值,即 $p=1/n$(其中,n 为样本维度),SVM 惩罚参数 C 取 1 000,SVM 本身的模型选择不是本文重点研究的问题,相关方法可参考文献[26].实验中,采用的概念漂移检测对比方法 SWCDS 和 CDOL 的参数选择分别参考文献[5,17].

4.4 实验结果与分析

由于流数据在线学习过程中噪声是不可避免的,因此有效区分噪声引起的伪概念漂移和真实概念漂移对在线学习算法的抗噪性具有重要作用.图 6 展示了本文提出的 CDPT 方法中近邻参考点对于概念漂移位点的平均判定值,用于反映近邻参考点对于真实概念漂移的贡献和对于伪概念漂移的识别情况.图中横轴表示不同的数据集,纵轴表示近邻参考点对于概念漂移的平均判定值,“*”表示判定值的平均值.平均判定值越大,则采用近邻参考点的方法对于概念漂移检测的贡献率越大,这里需要注意的是,实验中统计的是真实概念漂移位点及其后续 5 个近邻位点中的一致波动位点,因为这些点均被认为是真实概念漂移附近的一致波动位点,其平均判定值大小才能准确反映近邻参考点对于真实位点的判定情况.

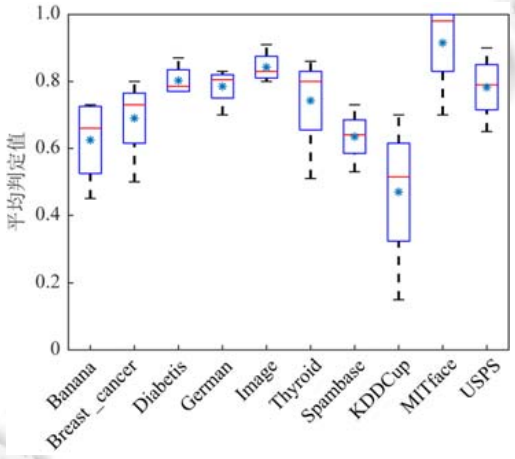


Fig.6 Average decision value of nearest neighbor reference points for concept drift

图 6 近邻参考位点对概念漂移的平均判定值

从图 6 可以看出,除了数据集 KDDCup 之外,所用到的数据集对应的概念漂移平均判定值的均值都在 60%以上,且最大值与最小值之间的差异不大,这充分说明在多数数据集上,近邻参考点对于概念漂移位点具有较好的判定准确性,同时能够有效区分噪声样本引起的伪概念漂移.而在数据集 KDDCup 检测结果中,存在部分真实概念漂移位点邻域内的一致波动位点多数被误判为噪声的情况,造成这种结果的原因是,在线处理的数据单元过小,概念漂移检测不及时所致.另外,数据集 KDDCup 为严重非平衡分布的数据集,个别类别不足其他类别样本规模的百分之一,因此在概念漂移过程中无法检测到极少数类样本存在的漂移问题.

漏检率用于检验概念漂移检测算法对真实概念漂移位点检测的完备性(类似召回率),即能否对每一次真实概念漂移的发生都精确地检测到.较低的漏检率能够说明算法可以识别尽可能多的概念漂移位点.图 7 记录了本文提出的 CDPT 概念漂移检测算法在不同规模的数据单元下的概念漂移漏检率,其中,“CDPT+*O*”代表本文提出的 CDPT 方法,且其对应的在线数据单元规模为 *x*.由于模拟实验中只在前期、中期、后期分别插入了一个真实的概念漂移位点,因此,这里概念漂移的漏检率取值只有 0、1/3、2/3 和 1 这 4 个值.

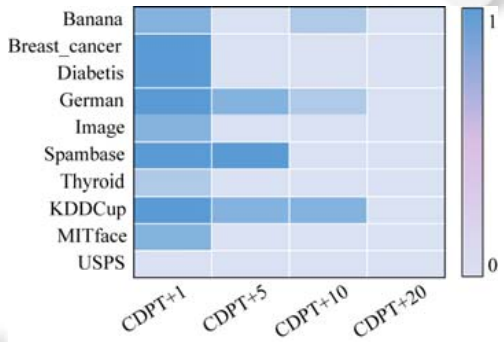


Fig.7 Miss detection rate (MDR) of CDPT algorithm

图 7 CDPT 算法漏检率(MDR)

从图 7 可以看出,当在线学习的数据单元规模扩大时,CDPT 算法得到的漏检率减小,即在线学习数据单元规模越大,概念漂移被检测到的概率就越大.另一方面,当数据块较小时(如 *O*=1),每个在线时间单元内到达的数据单元包含的样本数较少,以至于单次训练不足以引起模型预测结果产生明显波动,因此,相应地插入概念漂移的位点也很难被准确检测.实验发现,其他两种概念漂移检测方法 SWCDS 与 CDOL 随着在线数据单元规模的扩大,漏检率也相应减小(这里没有专门列出),表 3 列出了 CDPT 算法与其他两种传统概念漂移检测算法在同一数据单元规模参数下(*O*=20)的漏检率取值结果.

Table 3 Comparison of MDR of different algorithms

表 3 不同算法漏检率比较

Dataset \ Algorithm	CDPT	SWCDS	CDOL
Banana	0	0.67	0
Breast_cancer	0	0.67	0
Diabetis	0	0.67	0
German	0	0.67	1
Image	0	0.67	0
Spambase	0	0.67	1
Thyroid	0	0.67	0
KDDCup	0	1	1
MITface	0	0.67	1
USPS	0	0.67	1

从表 3 可以看出,与传统的 SWCDS 和 CDOL 两种概念漂移检测算法相比,相同条件下本文提出的基于在线性能测试的 CDPT 算法能够更准确地检测到漂移的发生位点,且发生漏检的几率更小.

误检率是检验概念漂移检测算法对漂移检测的精准性,即用于衡量检测到的概念漂移位点中并未真实发

生概念漂移的比例.较低的误检率能够说明算法对于由于训练样本规模过小或在线流数据中存在噪声等引起的伪概念漂移有足够的鲁棒性,即尽可能地不将伪概念漂移位点错误地识别为概念漂移位点.图 8 分别给出了不同算法在多个数据集中的误检率实验结果,图中横轴表示不同在线学习数据单元规模的 CDPT 算法与其他对比算法(对比算法对应的在线学习数据单元为固定值 O=20),图中灰色虚线条形表示当前方法未检测到概念漂移位点,因此误检率分母为 0.

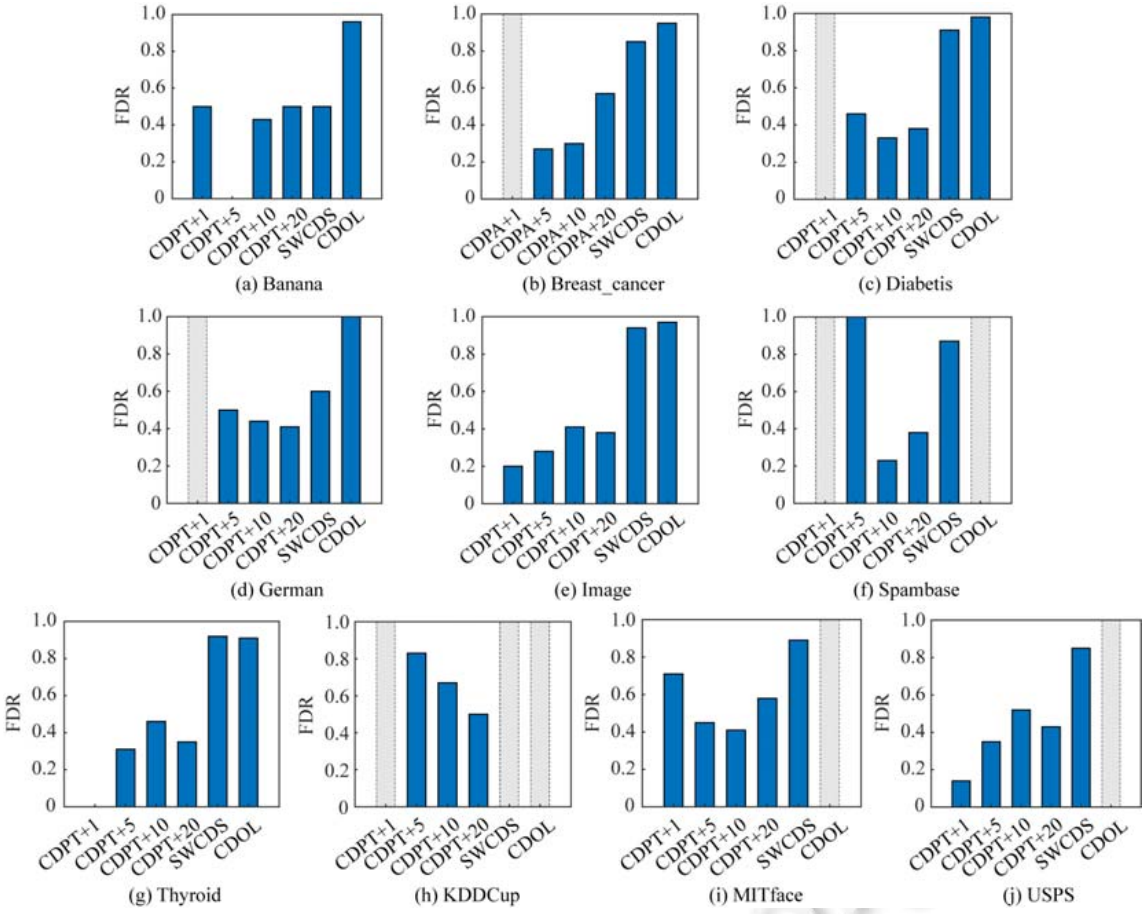


Fig.8 Comparison of FDR of different algorithms

图 8 不同算法误检率比较

从图 8 可以看出,对于本文提出的 CDPT 算法,当在线学习数据单元为 1 时,检测效果并不好,其中,在数据集 Breast_cancer、Diabetis、German、Spambase 以及 KDDCup 上未检测到概念漂移位点,而其在数据集 Image、Thyroid 和 USPS 上得到的误检率却最小,这充分说明,当在线学习数据单元取 1 时,每次只有一个样本新加入参与训练,由于在线学习模型中可能包含较多的历史数据信息,因此概念漂移发生后过小的在线数据单元有可能不足以引起模型的有效波动,因此模型表现不稳定.其次,从实验结果可以看出,除了在线学习单元过小的情形之外,本文提出的 CDPT 方法的概念漂移误检率基本在 20%~50%左右浮动,而对比的两个概念漂移检测算法 SWCDS 和 CDOL,其概念漂移的误检率一般都达到 80%以上,明显大于本文提出的 CDPT 算法,特别是 CDOL 方法在许多数据集上会存在检测不到概念漂移的现象.因此,本文提出的 CDPT 算法能够以较小的代价(检测到的概念漂移位点数少)检测到真实存在的概念漂移位点,可有效提高概念漂移检测的性能.

表 4 给出了 CDPT 算法在不同的在线学习数据单元下,在概念漂移检测过程中,检测到的概念漂移位点与

实际插入的概念漂移位点相比的平均延时,以分析算法在检测概念漂移时的实时性.根据概念漂移检测延时的定义(见式(12)),若某个概念漂移位点没有检测到,则计其延时为+∞.实验共分前期、中期、后期 3 个阶段,对应插入了 3 个概念漂移位点,表中列出了 3 个概念漂移位点检测延时的平均值,如果其中某个实际插入概念漂移的位点存在未检测到概念漂移的情况,依然列出了剩余检测到的概念漂移位点的平均延时值,但做了相应标注,如“2(∞)”表示存在一个未检测到的概念漂移点,剩余两个检测到的概念漂移点的平均延时为 2,“2(∞/∞)”表示存在两个未检测到的概念漂移位点,剩余一个检测到概念漂移的位点的检测延时为 2,而单纯的一个数值“2”则表示 3 个概念位点均被检测到,且其平均检测延时为 2.若 3 个概念漂移位点均未检测到,则不存在平均延时,其标记为“–”,此时,3 个真实概念漂移位点的延时均为+∞.

Table 4 Concept drift average detection delay of CDPT

表 4 CDPT 算法概念漂移平均检测延时

Dataset \ Algorithm	CDPT+1	CDPT+5	CDPT+10	CDPT+20
Banana	0(∞/∞)	0.33	1(∞)	1
Breast_cancer	–	0	0	1.33
Diabetis	–	2	0.33	0
German	–	2(∞/∞)	1(∞)	0
Image	1(∞/∞)	0	0.67	1.67
Spambase	–	–	0.67	0
Thyroid	1(∞)	0	2.33	1.33
KDDCup	–	1(∞/∞)	0(∞/∞)	1.67
MITface	1(∞/∞)	0.33	0.67	2.67
USPS	0.33	1	1.33	1

从表 4 可以看出,本文提出的 CDPT 算法在学习单元为 1 时,尽管概念漂移平均检测延时小,但其存在很多检测不到真实概念漂移位点的情况,即说明很小的学习单元的分布变化不足以使模型分类准确率发生显著改变,分类器准确率下降缓慢,无法检测到概念漂移;但是随着在线学习单元的增大,尽管概念漂移检测的平均延时略有增加,但检测不到概念漂移的情况能够被有效避免,特别地,当在线学习数据单元规模达到 20 时,在所有数据集上,在线学习的前期、中期、后期的概念漂移位点都能被 CDPT 方法有效地检测到.表 5 列出了在线数据单元为 20 时,CDPT 算法与其他两种对比算法的平均检测时延.

Table 5 Comparison of MDD in different algorithms

表 5 不同算法中漂移平均检测延时比较

Dataset \ Algorithm	CDPT	SWCDS	CDOL
Banana	1	3(∞/∞)	1
Breast_cancer	1.33	2(∞/∞)	0.67
Diabetis	0	1(∞/∞)	1
German	0	2(∞/∞)	–
Image	1.67	1(∞)	1.33
Spambase	0	2(∞/∞)	–
Thyroid	1.33	1.5(∞)	1
KDDCup	1.67	–	–
MITface	2.67	3(∞/∞)	–
USPS	1	1	–

从表 5 统计的概念漂移检测方法的漂移平均检测延时对比结果可以看出,与 SWCDS 算法相比,本文提出的 CDPT 算法平均检测延时要明显更小,且对比的 SWCDS 算法在多数情况下都未检测到概念漂移,这是由于本文把实际插入概念漂移的位点后续 5 个近邻检测到的概念漂移位点均视为检测正常检测,而 SWCDS 算法在流数据学习的中期和后期的平均检测时延都达到 20~30 个位点,因此,在中期和后期插入概念漂移的多数情况下都没有检测到概念漂移.尽管 CDOL 在部分数据集上的平均检测延时很短,但其在很多数据集上所有的概念漂移都未检测到,这说明该方法对数据集分布高度敏感,缺乏通用性.由此可见,与其他两种对比方法相比,本文提出

的基于在线性能测试的概念漂移检测方法能够更及时地检测到流数据中不同位置发生概念漂移的位点,可以广泛地应用于不同分布的数据集.

综上可以看出,本文提出的基于在线性能测试的概念漂移检测方法通过采用分组交叉检验提取一致波动点的方式去除由于训练样本过小而引起的伪概念漂移,通过分析一致波动点的近邻参考点测试性能的大小及收敛情况去除了由于噪声引起的伪概念漂移,检测结果具有较低的误检率和漏检率,即能够准确、充分地识别流数据中发生的概念漂移,同时其检测延时较小,可以更加实时地检测概念漂移的发生,符合流数据学习中高时效性要求的特点.

5 结束语

针对动态流数据中存在的伪概念漂移混淆引起真实概念漂移检测困难的问题,本文提出一种基于在线性能测试的概念漂移检测算法,通过对流数据学习中获得的当前数据分组,并对每组进行在线学习测试,根据每个分组相邻测试精度之间的精度落差判断有效波动位点.然后,采用交叉检验方式,精度波动的一致性提取一致波动位点.最后通过跟踪一致波动位点的近邻参考点测试精度的下降幅度和收敛速率来进一步减小噪声引起的伪概念漂移对真实概念漂移检测的负面影响,准确区分概念漂移的真伪性.本文所提出的方法能够在流数据在线学习过程中准确、充分地检测到概念漂移的发生,具有较低的漏检率和误检率,且平均检测延时短,检测结果具有较好的时间效能.同时,基于在线性能测试的方法对于流数据中存在的伪概念漂移有较高的识别能力,在真伪概念漂移同时存在且未知的环境下,能够有效地检测概念漂移并区分其真伪性,减小噪声及训练样本过小而导致的流数据在线学习及概念漂移检测困难的问题.

References:

[1] Yi Y, Wu JS, Xu W. Incremental SVM based on reserved set for network intrusion detection. Expert Systems with Applications, 2011,38(6):7698–7707. [doi: 10.1016/j.eswa.2010.12.141]

[2] García-García D, Parrado-Hernández E, Diaz-de-Maria F. State-space dynamics distance for clustering sequential data. Pattern Recognition, 2011,44(5):1014–1022. [doi: 10.1016/j.patcog.2010.11.018]

[3] Lu J, Liu AJ, Dong F, Gu F, Gama J, Zhang GQ. Learning under concept drift: A review. IEEE Trans. on Knowledge and Data Engineering, 2018. 1. [doi: 10.1109/TKDE.2018.2876857]

[4] Hulten G, Spencer L, Domingos P. Mining time-changing data streams. In: Proc. of the 7th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2001. 97–106.

[5] Zhu Q, Zhang YH, Hu XG, *et al*. A double-window-based classification algorithm for concept drifting data streams. Journal of Automatica Sinica, 2011,37(9):1077–1084 (in Chinese with English abstract). [doi: 10.3724/SP.J.1004.2011.01077]

[6] Liu B, Xiao YS, Yu PS, Cao LB, Zhang Y, Hao ZF. Uncertain one-class learning and concept summarization learning on uncertain data streams. IEEE Trans. on Knowledge & Data Engineering, 2014,26(2):468–484. [doi: 10.1109/TKDE.2012.235]

[7] Liu AJ, Lu J, Liu F, Zhang GQ. Accumulating regional density dissimilarity for concept drift detection in data streams. Pattern Recognition, 2018,76:256–272. [doi: 10.1016/j.patcog.2017.11.009]

[8] Méndez JR, Glez-Peña D, Fdez-Riverola F, Díaz F, Corchado JM. Managing irrelevant knowledge in CBR models for unsolicited e-mail classification. Expert Systems with Applications, 2009,36(2):1601–1614. [doi: 10.1016/j.eswa.2007.11.037]

[9] Yang KL, Wang L, Ryu KH. A system architecture for monitoring sensor data stream. In: Proc. of the Int'l Conf. on Computer and Information Technology. DBLP, 2007. 1026–1031. [doi: 10.1109/CIT.2007.178]

[10] Havens TC, Bezdek JC, Leckie C, Hall LO, Palaniswami M. Fuzzy *c*-means algorithms for very large data. IEEE Trans. on Fuzzy Systems, 2012,20(6):1130–1146. [doi: 10.1109/TFUZZ.2012.2201485]

[11] Žliobaitė I, Pechenizkiy M, Gama J. An overview of concept drift applications. In: Big Data Analysis: New Algorithms for a New Society, Studies in Big Data. Springer-Verlag, 2016,16:91–114. [doi: 10.1007/978-3-319-26989-4_4]

[12] Minku LL, Yao X. DDD: A new ensemble approach for dealing with concept drift. IEEE Trans. on Knowledge & Data Engineering, 2012,24(4):619–633. [doi: 10.1109/TKDE.2011.58]

[13] Du L, Song QB, Jia XL. Detecting concept drift: An information entropy based method using an adaptive sliding window. Intelligent Data Analysis, 2014,18(3):337–364. [doi: 10.3233/IDA-140645]

[14] Pesaranghader A, Viktor HL. Fast Hoeffding drift detection method for evolving data streams. In: Proc. of the Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. 2016. 96–111. [doi:10.1007/978-3-319-46227-17]

[15] Street WN, Kim YS. A streaming ensemble algorithm (SEA) for large-scale classification. In: Proc. of the 7th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2001. 377–382. [doi: 10.1145/502512.502568]

[16] Wang HX, Fan W, Yu P. S, Han JW. Mining concept-drifting data streams using ensemble classifiers. In: Proc. of the 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2003. 226–235. [doi: 10.1145/ 956750.956778]

[17] Zhao PL, Hoi SCH, Wang JL, Li B. Online transfer learning. Artificial Intelligence, 2014,216(16):76–102. [doi: 10.1016/j.artint.2014.06.003]

[18] Zhao QL, Jiang YH, Lu YT. Ensemble model and algorithm with recalling and forgetting mechanisms for data stream mining. Ruan Jian Xue Bao/Journal of Software, 2015,26(10):2567–2580 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4747.htm [doi: 10.13328/j.cnki.jos.004747]

[19] Ross GJ, Adams NM, Tasoulis DK, Hand DJ. Exponentially weighted moving average charts for detecting concept drift. Pattern Recognition Letters, 2012,33(2):191–198. [doi: 10.1016/j.patrec.2011.08.019]

[20] Gonçalves Jr PM, Santos SGTC, Barros RSM, Vieira DCL. A comparative study on concept drift detectors. Expert Systems with Applications, 2014,41(18):8144–8156. [doi: 10.1016/j.eswa.2014.07.019]

[21] Ditzler G, Royveri M, Alippi C, Polikar R. Learning in nonstationary environments: A survey. IEEE Computational Intelligence Magazine, 2015,10(4):12–25. [doi: 10.1109/MCI.2015.2471196]

[22] Tang SQ, Wen YM, Qin YX. Online transfer learning from multiple sources based on local classification accuracy. Ruan Jian Xue Bao/Journal of Software, 2017,28(11):2940–2960 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5352.htm [doi: 10.13328/j.cnki.jos.005352]

[23] Gonçalves Jr PM, Barros RSM. RCD: A recurring concept drift framework. Pattern Recognition Letters, 2013,34(9):1018–1025. [doi: 10.1016/j.patrec.2013.02.005]

[24] Gomes HM, Bifet A, Read J, Barddal JP, Enembreck F, Pfharinger B, Holmes G, Abdessalem T. Adaptive random forests for evolving data stream classification. Machine Learning, 2017,106(9):1469–1495. [doi: 10.1007/s10994-017-5642-8]

[25] KDD cup 99 data. 2010. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[26] Guo HS, Wang WJ. An active learning-based SVM multi-class classification model. Pattern Recognition, 2015,48(5):1577–1597. [doi: 10.1016/j.patcog.2014.12.009]

附中文参考文献:

[5] 朱群,张玉红,胡学钢,李培培.一种基于双层窗口的概念漂移数据流分类算法.自动化学报,2011,37(9):1077–1084.

[18] 赵强利,蒋艳凰,卢宇彤.具有回忆和遗忘机制的数据流挖掘模型与算法.软件学报,2015,26(10):2567–2580. http://www.jos.org.cn/1000-9825/4747.htm [doi: 10.13328/j.cnki.jos.004747]

[22] 唐诗淇,文益民,秦一休.一种基于局部分类精度的多源在线迁移学习算法.软件学报,2017,28(11):2940–2960. http://www.jos.org.cn/1000-9825/5352.htm [doi: 10.13328/j.cnki.jos.005352]

郭虎升(1986－),男,博士,副教授,CCF 专业会员,主要研究领域为数据挖掘,机器学习,计算智能.

王文剑(1968－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为机器学习,数据挖掘,计算智能.

张爱娟(1993－),女,硕士生,主要研究领域为流数据挖掘,机器学习.