

DOI:10.3969/j.issn.1673-4785.201208012

网络出版地址: <http://www.cnki.net/kcms/detail/23.1538.TP.20121116.1701.004.html>

概念漂移数据流分类研究综述

文益民¹, 强保华¹, 范志刚²

(1. 桂林电子科技大学 计算机科学与工程学院, 广西 桂林 541004; 2. 中国科学院 上海高等研究院, 上海 201203)

摘要: 由于现有各种机器学习算法本质上都基于一个静态学习环境, 而以尽量保证学习系统泛化能力为目标的寻优过程, 概念漂移数据流分类给机器学习带来了巨大挑战。从数据流与概念漂移、概念漂移数据流分类研究的发展与趋势、概念漂移数据流分类的主要研究领域、概念漂移数据流分类研究的新动态4个方面展开了文献综述, 并分析了当前概念漂移数据流分类算法存在的问题。

关键词: 大数据; 概念漂移; 增量学习; 适应学习; 数据流; 机器学习

中图分类号: TP391.4 **文献标志码:** A **文章编号:** 1673-4785(2012)02-0095-10

A survey of the classification of data streams with concept drift

WEN Yimin¹, QIANG Baohua¹, FAN Zhigang²

(1. College of Computer Science and Engineering, Guilin University of Electronic Technology, Guilin 541004, China; 2. Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201203, China)

Abstract: Because the current machine learning algorithms all are essentially an optimization procedure that aims to ensure the generalization ability based on static learning environment, the classification data streams with concept drift has brought severe challenges to machine learning. In order to address these concerns, a survey was developed consisting of four aspects: the introduction to data streams and concept drift, the development process and future trends, the main research fields, and the new developments in the study field of the classification data streams with concept drift. The existing problems relating to classification data streams with concept drift were discussed at last.

Keywords: big data; concept drift; incremental learning; adaptive learning; data stream; machine learning

在社会生产和生活实践中, 有一类问题是数据所包含的概念^[1]可能随时间而变化。自动化生产线上, 相近原因的问题产品会连续出现, 然后问题产品的特征也随之发生变化; 商务活动中, 顾客的购买兴趣随时间而变化; 网络安全中, 网络的访问模式随用户不同而变化; 社交媒体上, 用户的实际行为随其注册位置而变化。这些问题的共同特点是: 不断产生数据形成流; 数据流没有终点; 数据流中数据包含的概念随时可能产生变化。数据流中这种概念的变化在文中被称为概念漂移^[2]。概念漂移要求学习系统能尽早

地检测到概念漂移, 并对自身进行适应概念漂移的调整, 以对不断到来的数据尽可能地正确判断。

概念漂移问题给机器学习带来了巨大的挑战。目前各种人工学习系统的构造算法在本质上都是基于一个静态的学习环境, 而以尽量保证学习系统泛化能力为目标的寻优过程, 所以现有的各种机器学习算法在本质上都不适应进行概念漂移数据流学习。这种不适应体现在: 计算模型或者缺乏获取新知识的能力, 或者不能保持原本学到的知识^[3]。

自“概念漂移”(concept drift)在1986年由Schlimmer和Granger^[2]首次提出后, 国内外众多研究人员对概念漂移数据流分类展开了深入研究。Kuncheva^[4]、Tsybmal^[5]、王涛^[6]、Zliobaite^[7]、Hoens^[8]和Gama^[9]等先后从各自的角度对概念漂移数据流分类研究进行了很好的文献综述。与以上

收稿日期: 2012-08-07. 网络出版日期: 2012-11-16.

基金项目: 湖南省自然科学基金资助项目(10JJ5067); 湖南省科技计划资助项目(2010GK3047); 广西省可信软件重点实验室(桂林电子科技大学)开放课题资助项目(KX201118).

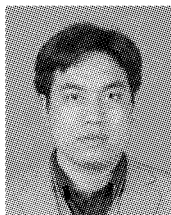
通信作者: 文益民. E-mail: ymwen2004@yahoo.com.cn.

[C]//Proceedings of the Eighth Australian Joint Conference on Artificial Intelligence. Canberra, Australia, 1995: 91-98.

[96]MOVELLAN J R, MINEIRO P. Robust sensor fusion: analysis and application to audio visual speech recognition [J]. Machine Learning, 1998, 32(2): 85-100.

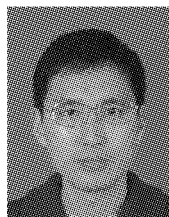
[97]BERNSTEIN A, VORBURGER P, EGGER P. A scenario-based approach for direct interruptability prediction on wearable devices [J]. International Journal of Pervasive Computing and Communications, 2008, 3(4): 426-438.

作者简介:

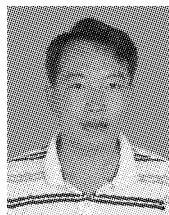


文益民,男,1969年生,教授,硕士生导师,博士,CCF高级会员,主要研究方向为机器学习与数据挖掘、极化 SAR 图像处理、社会计算。获得省部级教学、科研奖励5项,主持省部级科研项目8项,发表学术论文30余篇,

其中被SCI、EI检索18篇,翻译译著1部。



强保华,男,1972年生,教授,硕士生导师,博士,主要研究方向为Web信息处理、智能搜索、海量信息处理、网络信息集成。获西南大学教学成果奖三等奖1项,主持国家自然科学基金及省部级科研项目共8项。发表学术论文20余篇,其中被EI检索12篇,翻译译著1部。



范志刚,男,1978年生,副研究员,博士,主要研究方向为大规模数据挖掘和机器学习及其高性能分布式计算。曾主持中科院重点部署项目等科研项目。发表学术论文14篇,其中被SCI、EI检索10篇。

高级数据挖掘与应用国际学术会议

The 9th International Conference on

Advanced Data Mining and Applications (ADMA 2013)

ADMA 2013 will take place in Zhejiang University, Hangzhou, 14th-16th December 2013. It is our pleasure to invite you to contribute papers, register and participate at ADMA 2013. The conference aims at bringing together the experts on data mining from around the world, and providing a leading international forum for the dissemination of original research findings in data mining, spanning applications, algorithms, software and systems, as well as different applied disciplines with potential in data mining.

Topics of interest include but not limited to:

Advanced data ming

Mass data mining

Web mining

Social network mining

Document mining

High performance data mining

Economic data mining

Image mining

Data mining visualization

Parallel and distributed data mining

Multimedia mining

Security and privacy issues

Data stream mining

Biomedical data mining

Other issues on data mining

Graph mining

Complex data mining

Applications

Social network application

E-commerce and Web service

Data mining systems in finance, economic, e-commerce

Medical informatics

Emerging applications of mass data mining

Disaster prediction

Empirical study of data mining algorithm

Financial market analysis

Parallel data mining application

Data mining for education

DNA sequencing, Bioinformatics

Intelligent systems

Genomics and biometrics

Website: <http://www.adma2013.org/>

这些综述相比,本综述具有如下特色之处:1)剖析了概念漂移数据流分类研究产生和发展的脉络;2)包含了概念漂移数据流分类的最新研究动向——概念漂移数据流分类中的类别不平衡学习、重复概念学习及半监督学习和主动学习问题;3)深入分析了当前概念漂移数据流分类算法存在的问题。

1 数据流与概念漂移

数据流分类问题引起研究人员关注的原因主要有2个:1)因为自动数据获取技术的飞速发展使得人类获得了大量的数据,数据量太大时,数据不能被一次性装入内存;2)由于传感器技术的发展,使得人类获得了大量与时间和环境相关的数据^[10]。Gamma^[9]和Street^[11]讨论了数据流分类问题的顺序处理、单向通过、内存有限等特点。数据流分类通常被描述为在线分类模型^[12],也就是分类器每次只对一个样本分类,在完成对该样本的分类后,分类器将得到由专家给出的该样本的真实类别,该样本及其类别标志被用于分类器更新,当分类器完成更新后,将对下一个接收到的样本实施分类。在线分类模型通常又被扩展为分类器每次分类或学习一批样本。

数据流分为2种:1)数据源产生的数据独立同分布,研究人员称为稳定数据流^[13];2)数据源产生的数据不独立同分布,研究人员认为在数据产生过程中发生了“概念漂移”^[2],称其为动态数据流^[14]。研究人员对概念漂移的深入理解是通过分析概念漂移的种类及产生的原因逐步得到的。

Widmer等^[15]认为数据产生环境的变化导致了概念漂移,并将概念漂移区分为虚概念漂移和实概念漂移。Kelly等^[16]认为概念漂移是样本与其类别的联合概率随时间变化而产生,其产生原因分3种:1)某类的先验概率发生变化;2)某类的类概率发生变化;3)样本后验概率发生变化。Kuncheva^[4]引用时间序列分析方法将概念漂移分为4种:随机噪声、随机趋势、随机替换和系统趋势。其中,随机趋势中包含渐变性概念漂移,随机替换中包含突变性概念漂移,系统趋势中包含着重复性概念漂移。Narasimhamurthy等^[17]根据数据产生的多源性提出了概念漂移的产生模型。Zliobaite把概念漂移分为4种:突变性概念漂移、渐变性概念漂移、增量性概念漂移和重复性概念漂移^[7]。Minku等^[18]在总结他人工作的基础上,选择了纯度、速度、可预测度、频度、重复度5个维度,将概念漂移分成14种。对数据流中增加了类别的情形,学术界还没有达成一致认识——有

人认为是一种概念漂移^[19],有人则认为属于概念进化^[20]。

总之,学术界对概念漂移的认识日渐清晰,但目前还缺少对概念漂移的统一描述。Moreno-Torres等试图利用数据漂移(data shift)的概念来统一已有各种概念漂移的描述^[21]。

2 概念漂移数据流分类研究的发展与趋势

自“概念漂移”首次提出后,得到了学术界的日益重视。从1986—2000年左右这段时间的研究主要围绕单分类器展开——使用单个分类器实现概念漂移数据流分类。Kilander等提出了COBBIT^[22];Widmer和Kubat提出了FLORA^[15];Hulten等提出了CVFDT^[14];Black等提出了CD3^[23];同时研究人员开始关注概念漂移数据流分类的理论问题^[24-26]。为深入探讨数据产生的情境(context)与概念漂移的关系,1998年由Dietterich、Widmer和Kubat发起,Machine Learning出版了研究概念漂移数据流的专刊^[27]。

由于使用单分类器处理概念漂移数据流时需要不断更新分类模型且分类器泛化能力不高,Street等^[11]首次将集成学习引入概念漂移数据流分类,提出了SEA算法。因此,从2000年左右开始,研究人员对概念漂移数据流分类的研究开始转移到分类器集成上来。通过多分类器集成,实现对历史样本的选择,提高分类器泛化能力。Wang等提出的AWE^[28]、Kolter等提出的DWM和AddExp^[29-30]和Bifet等提出的ADWIN和ASHT^[31]是这个领域里非常有影响的成果。

2000年左右,概念漂移数据流分类研究进入了快速发展期,研究人员开始考虑更加接近实际状况的概念漂移数据流。Klinkenberg和Lanquillon比较早地研究了在检测概念漂移时只有部分样本获得用户反馈或者没有反馈的情形^[32-34]。2004年由Intelligent Data Analysis期刊出版的概念漂移数据流专刊^[35]主要探讨了如何利用增量学习方法以较小的代价使已有分类器适应概念漂移;之后概念漂移数据流分类中的类别不平衡学习^[36-42]、概念重复学习^[43-45]、半监督学习^[46-48]、主动学习^[49-54]等问题开始得到较多关注。2010年IOS还出版了《Adaptive Stream Mining: Pattern Learning and Mining from Evolving Data Streams》。从近年机器学习与数据挖掘领域的一些国际权威期刊和国际顶级会议上发表的论文来看,概念漂移数据流分类的研究正日益成为学术界关注的焦点,对概念漂移数据流的研究已经

开始与转移学习^[55-56]、进化计算^[57-58]、特征选择^[59]、聚类^[60]、时间复杂度分析^[61-62]、社会计算^[63]等结合起来. 由于运动是物质的本质,概念漂移也是数据的本质. 因此,从趋势上来讲,已有各种模式分类的理论和算法都可与概念漂移相结合而引出更多新的研究问题.

3 概念漂移数据流分类的主要研究领域

3.1 概念漂移数据流学习器的构建

对已有各种学习器进行调整使之适应概念漂移数据流学习是目前主要的研究方向. 这些算法可分为2类:1)通过单分类器实现;2)通过多分类器集成实现.

利用单分类器进行概念漂移数据流学习的方法有4种:1)选择训练样本. 该类方法的主要思路是:从采集的训练样本中选择一部分最合适对未来数据实施准确分类的样本训练分类器,其主要做法有滑动窗口法、自适应滑动窗口调整法^[64-66]以及动态样本选择法^[67-68]. 2)给训练样本赋以权值. 该类方法的主要思路是对最新的训练样本赋以最大的权值,以提高对新概念的反应速度^[69-71]. 3)调整学习器的结构. 该类方法的特点是动态调整分类器的内部结构,以适应概念漂移检测的要求. Hulten 等提出的 CVFT^[14]用适合新概念的子树替换旧子树;Nunez 等提出了 OlineTree2,该算法中决策树的每个叶子节点都维护一个时间窗口和局部性能测度,当测度下降时,时间窗口将减小^[72]. 4)是各种方法的组合^[6,14].

多分类器集成是机器学习的研究热点之一. 国内外学者在利用集成学习策略实施概念漂移数据流学习方面已经做了许多探索,具体的研究内容主要分2个方面:1)利用集成学习策略对数据流实施分块学习;2)基于在线学习模型对整个数据流实施集成学习,即所有基分类器采用相同的学习算法,它们各自的训练样本来自同一数据流.

利用集成学习策略对数据流实施分块学习,使用了滑动窗口技术. 这类算法通常假定最近获得的训练样本与即将要采集的样本同分布. Street 等^[11]提出了 SEA (streaming ensemble algorithm) 算法,该算法根据一个预设的质量标准使用新分类器替代不必要的旧分类器,而保持分类器总数不变,而实现对新概念的学习. 然而当概念漂移突然发生时,体现新概念的分类器不足以跟旧分类器相抗衡,因此 SEA 在一段时间内不能识别新概念. Wang 等^[28]提出了

一个类似 SEA 的 AWE (accuracy-weighted ensembles) 算法,该算法根据最新采集的训练样本集计算各基分类器的权值. Kyosuke^[73]提出了 ACE (adaptive classifier-ensemble) 算法,该算法在集成分类器的基础上增加了一个能进行在线学习的分类器,以提高算法对概念漂移的反应速度. Muhlbaier 等^[19,74]提出了 Learn + NSE,该算法根据已训练的基分类器和集成分类器在最新采集训练样本集上的分类性能,来调整各样本对应的权值和各基分类器的权值. 文益民等^[75]提出了基于分类置信度的概念漂移检测方法,根据分类置信度实现分类器的选择集成,使得集成分类器能快速适应新概念. 关菁华等^[76]提出了一种选择性集成方法,该算法根据各基分类器在验证数据集上的输出结果与参考向量之间的角度来选择参与集成的分类器. 朱群等^[77]提出了一种基于双层窗口方法,该算法将滑动窗口分解成多个基本窗口,以基本窗口为单位进行概念漂移检测. 以上这些算法还有一个共同点,即假定一个数据块中没有概念漂移,因此需要事先了解数据流的结构.

基于在线学习模型对整个数据流实施集成学习的思路与 Adaboosting 算法的思路类似. Kolter 等提出了基于动态带权多数投票 (DWM) 的学习算法^[29]和 AddExp^[30],该算法能根据集成分类器的分类性能动态地增加或者删除基分类器;孙岳等^[78]提出了一种基于多分类器的概念漂移挖掘算法,该算法的思路与 AddExp 类似,只是基分类器的权值设置不同;辛轶等^[79]提出了 IKnnM-DHecoc 算法,该算法根据新到样本的概念漂移度调整编码矩阵,然后根据调整后的编码矩阵更新训练基分类器.

3.2 概念漂移数据流学习理论的研究

概念漂移数据流学习理论主要是基于在线学习模型展开研究. Kun^[24]给出了在前一个概念是可能近似正确学习 (PAC) 的条件下,实现对下一个概念的 PAC 学习所需的最多样本数的上界;Helmbold 等^[25]在稳定渐变概念漂移和确保未来样本错误率下界的条件下,给出了概念漂移程度的上界;Barve 等^[26]研究了在确保分类器泛化能力的条件下,数据流中相邻2个样本概念漂移的最大幅度;Wang 等^[28]给出了使用带权多数投票进行概念漂移学习的理论证明;Kolter 等^[30]证明了 AddExp 的错误率的界;Kuncheva 等^[80]试图建立确定滑动窗口大小的理论根据,给出了分类错误率与滑动窗口大小的关系;Minku 等^[18,81]分析了分类器之间多样性 (Diversity) 对概念漂移检测的影响.

3.3 概念漂移的检测

检测概念漂移大致有3种方法:性能法、距离法和性质法。

1) 性能法. 跟踪当前分类器对最新采集训练集的分类性能,如分类性能出现较大下降,这说明最新采集训练集中包含有概念漂移. Widmer 和 Kubat^[15]提出的 FLORA 系列算法,根据分类器对正类样本覆盖率以及分类准确率调整滑动窗口; Klinkenberg 等^[82]使用对训练集的分类准确率、训练集中某类的分类准确率和召回率(recall)来实施概念漂移检测,以调整窗口大小;他们还通过估算各个滑动窗口上得到的支持向量机的泛化能力来确定滑动窗口的大小^[64]; Last 等^[83]提出了 OLIN,该算法通过比较分类器在训练集与验证集上的错误率来判断是否产生了概念漂移; Gama^[84]通过计算一个训练样本被错误分类的概率和其变化的范围来检测训练集中概念漂移的起点和终点; Nishida 等^[85]使用分类器对最新采集训练样本的分类准确率和对全部训练样本的分类准确率来检测概念漂移; 罗秀等^[86]提出了基于误差率的检测方法. 性能法是最常用的概念漂移检测方法,但当数据流中存在类别不平衡时或只有部分训练样本具有类别标志时,性能法将不适合用于概念漂移检测。

2) 距离法. Katakis 等^[87]将一个数据块映射成一个“概念向量”,然后对多个概念向量实施聚类,由一个聚类代表一个概念. 当采集到一个数据集时,计算该数据集对应的概念向量与各个聚类中心的距离,以检测是否产生概念漂移. Katakis 用该方法实现了对重复概念的检测,但该方法有一个前提:一个数据块中的各数据属于同一个概念。

3) 性质法. 分析最近获得训练集的一些统计特性:各类的分布、各特征值的分布等来实现对概念漂移的检测. Alippi 等^[88-90]利用中心极限定理,设计了一种不依赖数据分布模型的,不需要任何先验信息概念漂移检测算法; Peter 等^[91]提出了基于熵的概念漂移检测方法,通过一种熵的计算来评测训练集之间样本分布的区别; Kuncheva^[92]在 KL 距离和 T 平方测试的基础上通过假定数据服从组合正态分布导出了 SPLL 概念漂移检测方法。

3.4 概念漂移数据流分类研究使用的数据集

到目前为止,概念漂移数据流分类技术被用于以下实际问题的解决: Web 数据^[13-14]、英国银行数据^[16]、天气预报^[19]、Reuters 语料^[34,52]、KDDCup 数据^[20,40,45,48,59-62,66,77,79]、信用卡欺诈数据^[28,66]、日程

提醒^[29]、电价预测^[29,31,39,44,52,68,72,80]、TREC^[32,33,69,82]、垃圾邮件过滤^[53,62,86]、Netflix 电影等级数据集^[63]、Yahoo 购物数据^[48,77]、邮件链表^[87]、集群计算机负载均衡^[93]、传感网数据^[94]、交通数据^[94]、金融时间序列^[95]、视听说话识别^[96]、可穿戴设备^[97]、航班延误^[52]和电影标注数据集^[52]。

使用频率比较高的人工概念漂移数据集包括: STAGGER 数据集^[2,15,29,30,31,37,40,57,65,70,84-85]、SEA 数据集^[11,19,29,31,37,57,72-73,75-78]、高斯数据^[19,41,61,84-85]、旋转平面数据集^[14,28,31,42,44,49,61,66,76-79,81,85-86]。

另外,UCI 中的一些数据集也常用于概念漂移数据流分类^[40,42-43,45-46,52,61-62,68,72,77,92]。

4 概念漂移数据流分类研究的新动态

4.1 概念漂移数据流中的类别不平衡学习问题

类别不平衡使得概念漂移数据流问题更加复杂,因此直到最近学术界才开始关注这方面的研究,总的思路是将已有处理类别不平衡的算法加以改进,以适应概念漂移数据流中的类别不平衡问题. Gao 等^[36]研究了概念漂移数据流学习中的类别不平衡问题,对最新采集训练集中的多类(样本数量多的类)样本采取多轮“下采样”,将所有已学习过的少类(样本数量少的类)样本和最新采集训练集中的少类样本合并成一个子集,然后将该子集分别与属于多类每轮“下采样”得到的子集合并训练分类器,以实现对最新采集样本集的集成学习; Chen 等^[37-39]利用 Mahalanobie 距离从已学习过的所有少类样本中选择一部分样本与最新采集的不平衡样本集合并,以减轻类别不平衡; Lichtenwalter 等^[40]对属于多类的样本进行下采样,将多类中被错误分类的样本与全部少类样本构成训练集; Gregory 等^[41]采用 Gao 采取的方法,实现对最新采集样本集的集成学习,然后通过修改 Learn++NSE 中的权值使其偏向多类和少类的查全率,实现其与已训练好分类器的集成; Zhang 等^[42]将分类错误的少类样本加入训练集并使用 F_1 值控制分类器更新频率. 由于存在概念漂移,这些方法都不能取得在数据分布不改变情形下类别不平衡学习算法所能达到的性能。

4.2 概念漂移数据流中的概念重复学习问题

Widmer 比较早地注意到了概念会重复出现^[15],但是概念漂移数据流中的概念重复学习问题直到最近才得到学术界的广泛关注. Widmer 等将已经学习过的概念描述保存起来,当已学习过的概念重新出现时,保存的分类器被重新激活; Ramamur-

thy^[43]提出将得到的分类器放到全局集中,当采集到新数据块时,通过评测全局集中的分类器在新数据块上的分类准确率来判断新到数据块是否属于新概念,如果不是新概念则从全局集中挑选部分分类器组成集成分类器去检测新到数据块是否属于原来出现过的概念;Katakis^[87]将一个数据块转化成一个概念向量,然后通过数据流聚类方法将不断到来的数据块聚类成多个概念向量的集合,每个概念向量的集合通过增量学习得到一个分类器,最新采集数据块用于最近获得的训练集所属概念相对应的分类器分类;Li等^[44]提出了通过概念聚类,形成概念列表的方式检测新到样本是否是属于已有概念;Masud等^[45]使用主集成和辅集成来检测新到样本是否属于已学习概念.以上这些算法主要的不足在于数据块的大小不容易确定(太大可能包含多个概念,太小则数据的分布不够体现概念本身).

4.3 概念漂移数据流中的半监督学习与主动学习问题

Klinkenberg^[32]较早地关注到了概念漂移数据流分类中的半监督学习问题.在假定当前最新训练集与最新测试集同分布的条件下,他提出使用没有类别标志的样本估计滑动窗口的大小;Xue等^[46]针对某类的先验概率发生变化以及最新采集的训练样本无类别标志的情形,探讨了如何利用无类别标志样本估计各类别样本的数量及实施半监督学习;Zhang等^[47]提出了RK-TS3VM算法,该算法根据最新采集训练集中的样本是否具有类别标志和是否产生了概念漂移而分为4个子集,然后根据子集的特点选择TS3VM或RK算法实现学习和分类;Li等^[48]提出了SUN算法,该算法使用基于k-Modes的聚类方法来实现无类别标志样本的标注和重用,进而使用概念聚类方法实现概念漂移检测.

Fan等^[49]提出了不需要事先获得样本类别标志就可以计算的2个检测概念漂移的指标:基于叶子节点的分布统计指标以及验证错误率与期望错误率的差,当检测到概念漂移后,根据样本类别标注代价,选择性地给出一部分样本的类别标志;Zhu等^[50]选择使得集成分类器能产生最大方差的不确定样本,给出类别标志,并通过分类器的加权集成,使得集成分类器能适应概念漂移;Masud等^[51]使用主动学习策略,去分辨数据流中是否产生了新的类别以及挑选需要标注的样本;Zliobaite等^[52]针对概念漂移发生的位置有的在分类面附近而有的远离分类面的特点,提出了可变不确定性策略和随机不确

定性策略以主动选择无类别标志样本,尽量使得主动标注的样本能体现概念漂移;Chu等^[53]提出了基于贝叶斯概率的在线主动学习算法,该算法利用重要性采样方法实现对无类别标志样本的主动无偏选择,使得主动选择的样本符合当前概念的概率分布.

以上这些算法或使用半监督学习策略尽量发挥没有类别标志样本的作用,或使用主动学习策略选择能体现概念漂移的样本进行标注,但半监督学习和主动学习并没有很好地结合起来.Kholghi^[54]提出了一个将半监督和主动学习相结合进行概念漂移数据流分类的算法框架.

5 存在的问题

通过以上分析发现,已有的各种概念漂移数据流学习算法在处理概念漂移数据流分类时存在以下5方面的问题:

1)冷启动.由于在一段时间里概念漂移的次数无法预知,学习系统只有在发生分类错误后才能得到调整,这导致了属于新概念的样本在刚开始出现时被错误分类.而且在只有部分样本具有类别标志的情形下,学习系统将很难知道是否产生了错误分类.如果将这样的分类器应用于工业流水线生产,将带来巨大的损失.

2)只能实施单概念学习.目前主要使用滑动窗口、训练样本赋权、分类器自适应调整等方法实现对新概念的学习.这些方法只能跟踪学习一个概念,这导致了已学习知识难以保持.由于某些概念不定期重复出现,其产生的时刻和延续时间未知,导致现有的单概念学习模式无法适应.对概念漂移中重复概念的关注说明多概念学习问题刚得开始到学术界的关注.

3)概念漂移检测难以准确实施.现有的各种概念漂移算法主要依靠最近获得的训练样本集.且当最近获得的训练样本集中存在类别不平衡时会导致对少类的忽视;当其中只有部分样本具有类别标志时已有概念漂移检测方法将无法实施.

4)最新采集样本与最近获得训练样本同分布的假设不正确.从某一时刻开始最新采集样本中可能包含有跟最近获得训练样本不同分布的新概念,甚至既包含属于新概念的样本又包含属于旧概念的样本.如果最新采集样本与最近获得训练样本总是不同分布,现有概念漂移学习算法将无法实施.

5)理论基础研究缺乏.尽管概念漂移数据流研究已经取得丰硕的研究成果,然而到目前为止,概念

漂移的定义、概念漂移的速度与幅度、概念漂移数据流分类器的错误率的界、概念的容量、概念漂移检测等都没有得到严格的数学描述,已取得的一些理论上的结果也有待验证和深化。

6 结束语

概念漂移数据流分类给机器学习带来了巨大的挑战,本论文从数据流与概念漂移、概念漂移数据流分类研究的发展与趋势、概念漂移数据流分类的主要研究领域、概念漂移数据流分类研究的新动态4个方面分析总结了学者们在概念漂移数据流分类问题上杰出的研究工作,并分析了当前概念漂移数据流分类存在的问题。尽管概念漂移数据流分类已经取得了丰硕的研究成果,且已经成为机器学习研究的热点之一,然而其研究远还没有成熟,还存在着众多挑战,期待着学术界能有更大的突破。

参考文献:

- [1] MITCHELL T M. Machine learning [M]. New York: McGraw-Hill, 1997: 15-37.
- [2] SCHLIMMER J, GRANGER R. Incremental learning from noisy data[J]. Machine Learning, 1986, 1(3): 317-354.
- [3] GROSSBERG S. Nonlinear neural networks: principles, mechanisms and architecture[J]. Neural Network, 1988, 1(1): 17-61.
- [4] KUNCHEVA L I. Classifier ensembles for changing environments[C]//Proceedings of the Fifth Workshop on Multiple Classifier Systems. Cagliari, Italy, 2004: 1-15.
- [5] TSYMBAL A. The problem of concept drift: definitions and related work. TCD-CS-2004-15[R]. Dublin: Department of Computer Science Trinity College, 2004.
- [6] 王涛, 李舟军, 颜跃进, 等. 数据流挖掘分类技术综述[J]. 计算机研究与发展, 2007, 44(11): 1809-1815.
WANG Tao, LI Zhoujun, YAN Yuejin, et al. A survey of classification of data streams[J]. Journal of Computer Research and Development, 2007, 44(11): 1809-1815.
- [7] ŽLIobaite I. Learning under concept drift: an overview [EB/OL]. [2012-01-12]. http://zliobaite.googlepages.com/Zliobaite_CDoverview.pdf.
- [8] HOENS T R, POLIKAR R, CHAWLA N V. Learning from streaming data with concept drift and imbalance: an overview[J]. Progress in Artificial Intelligence, 2012, 1(1): 89-101.
- [9] GAMA J. A survey on learning from data streams: current and future trends[J]. Progress in Artificial Intelligence, 2012, 1(1): 45-55.
- [10] OVERPECK J T, MEEHL G A, BONY S, et al. Climate data challenges in the 21st century[J]. Science, 2011, 331(6018): 700-702.
- [11] STREET W N, KIM Y S. A streaming ensemble algorithm (SEA) for large-scale classification[C]//Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2001: 377-382.
- [12] ANGLUIN D. Queries and concept learning[J]. Machine Learning, 1988, 2(4): 319-342.
- [13] DOMINGOS P, HULTEN G. Mining high-speed data stream[C]//Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, USA, 2000: 71-80.
- [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. San Francisco, USA, 2001: 97-106.
- [15] WIDMER G, KUBAT M. Effective learning in dynamic environments by explicit context tracking[C]//Proceedings of the Sixth European Conference on Machine Learning. Vienna, Austria, 1993: 69-101.
- [16] KELLY M, HAND D, ADAMS N. The impact of changing populations on classifier performance[C]//Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 1999: 367-371.
- [17] NARASIMHAMURTHY A, KUNCHEVA L I. A framework for generating data to simulate changing environments[C]//Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications. Innsbruck, Austria, 2007: 384-389.
- [18] MINKU L L, WHITE A P, YAO X. The impact of diversity on online ensemble learning in the presence of concept drift[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(5): 730-742.
- [19] ELWELL R, POLIKAR R. Incremental learning of concept drift in nonstationary environments[J]. IEEE Transactions on Neural Networks, 2011, 22(10): 1517-1531.
- [20] MASUD M, GAO J, KHAN L, et al. Integrating novel class detection with classification for concept-drifting data streams[C]//Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. Bled, Slovenia, 2009: 79-94.
- [21] MORENO-TORRES J G, RAEDER T, ALAIZ-RODRIGUEZ R. A unifying view on dataset shift in classification[J]. Pattern Recognition, 2012, 45: 521-530.
- [22] KILANDER F, JANSSON C. COBBIT—a control procedure for COBWEB in the presence of concept drift[C]//Proceedings of the Sixth European Conference on Machine

- Learning. Helsinki, Finland, 1993: 244-261.
- [23] BLACK M, HICKEY R J. Maintaining the performance of a learned classifier under concept drift [J]. *Intelligent Data Analysis*, 1999, 3(6): 453-474.
- [24] KUN A, PETSCH T, RIVEST R L. Learning time-varying concepts[C]//*Proceedings of the Conference Neural Information Processing Systems*. Denver, USA, 1990: 183-189.
- [25] HELMBOLD D P, LONG P M. Tracking drifting concepts by minimizing disagreements [J]. *Machine Learning*, 1994, 14(1): 27-45.
- [26] BARVE R D, LONG P M. On the complexity of learning from drifting distributions [C]//*Proceedings of the Ninth Annual Conference on Computational Learning Theory*. Desenzano Sul Garda, Italy, 1996: 122-130.
- [27] DIETTERICH T, WIDMER G, KUBAT M. Special issue on context sensitivity and concept drift [J]. *Machine Learning*, 1998, 32(2): 83-201.
- [28] WANG H, FAN W, YU P S, et al. Mining concept-drifting data streams using ensemble classifiers [C]//*Proceedings of the Ninth International Conference on Knowledge Discovery and Data Mining*. Washington, DC, USA, 2003: 226-235.
- [29] KOLTER J Z, MALOOF M A. Dynamic weighted majority: a new ensemble method for tracking concept drift [C]//*Proceedings of the Third IEEE International Conference on Data Mining*. Washington, DC, USA, 2003: 123-130.
- [30] KOLTER J Z, MALOOF M A. Using additive expert ensembles to cope with concept drift [C]//*Proceedings of the 22nd International Conference on Machine Learning*. Bonn, Germany, 2005: 449-456.
- [31] BIFET A, HOLMES G, PFAHRINGER B, et al. New ensemble methods for evolving data streams [C]//*Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Paris, France, 2009: 139-148.
- [32] KLINKENBERG R. Using labeled and unlabeled data to learning drifting concepts [C]//*Proceedings of the Workshop Notes of the IJCAI-01 Workshop on Learning from Temporal and Spatial Data*. Menlo Park, USA, 2001: 16-24.
- [33] KLINKENBERG R. Detection concept drift with support vector machines [C]//*Proceedings of the Seventeenth International Conference on Machine Learning*. Stanford, California, USA, 2000: 487-494.
- [34] LANQUILLON C. Information filtering in changing domains [C]//*Proceedings of the 16th International Joint Conference on Artificial Intelligence*. Stockholm, Sweden, 1999: 41-48.
- [35] KUBAT G, GAMA J, UTGOFF P. Special issue on incremental learning systems capable of dealing with concept drift [Z]. Amsterdam: *Intelligent Data Analysis*, 2004.
- [36] GAO J, DING B, FAN W, et al. Classifying data stream with skewed class distributions and concept drifts [J]. *IEEE Transactions on Internet Computing*, 2008, 12(6): 37-49.
- [37] CHEN S, HE H. SERA: selectively recursive approach towards nonstationary imbalanced stream data mining [C]//*Proceedings of the 2009 International Joint Conference on Neural Networks*. Atlanta, USA, 2009: 522-529.
- [38] CHEN S, HE H, LI K, et al. MuSeRA: multiple selectively recursive approach towards nonstationary imbalanced stream data mining [C]//*Proceedings of the 2010 International Joint Conference on Neural Networks*. Barcelona, Spain, 2010: 1-8.
- [39] CHEN S, HE H. Towards incremental learning of nonstationary imbalanced data stream: a multiple selectively recursive approach [J]. *Evolving Systems*, 2010, 2(1): 35-50.
- [40] LICHTENWALTER R N, CHAWLA N V. Adaptive methods for classification in arbitrarily imbalanced and drifting data streams [C]//*Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Bangkok, Thailand, 2009: 53-75.
- [41] GREGORY D, ROBI P. An ensemble based incremental learning framework for concept drift and class imbalance [C]//*Proceedings of the 2010 International Joint Conference on Neural Networks*. Barcelona, Spain, 2010: 9-17.
- [42] ZHANG J, HU X, ZHANG Y, et al. An efficient ensemble method for classifying skewed data streams [C]//*Proceedings of the 7th International Conference on Intelligent Computing*. Zhengzhou, China, 2011: 144-151.
- [43] RAMAMURTHY S, BHATNAGAR R. Tracking recurrent concept drift in streaming data using ensemble classifiers [C]//*Proceedings of the 6th International Conference on Machine Learning and Applications*. Cincinnati, USA, 2007: 404-409.
- [44] LI P P, WU X D, HU X G. Mining recurring concept drifts with limited labeled streaming data [C]//*Proceedings of the 2th Asian Conference on Machine Learning*. Tokyo, Japan, 2010: 241-252.
- [45] MASUD M M, AL-KHATEEB T M, KHAN L, et al. Detecting recurring and novel classes in concept-drifting data streams [C]//*Proceedings of the 2011 International Conference on Data Mining*. Vancouver, Canada, 2011: 1176-1181.
- [46] XUE J C, WEISS G M. Quantification and semi-supervised classification methods for handling changes in class distribution [C]//*Proceedings of the 15th ACM SIGKDD*

- International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 897-906.
- [47] ZHANG P, ZHU X, GUO L. Mining data streams with labeled and unlabeled training examples [C]//Proceedings of the Ninth IEEE International Conference on Data Mining. Miami, USA, 2009: 627-636.
- [48] LI P, WU X D, HU X G. Learning from concept drift data streams with unlabeled data [C]//Proceedings of the 24th AAAI Conference on Artificial Intelligence. Atlanta, USA, 2010: 1945-1946.
- [49] FAN W, HUANG Y, WANG H, et al. Active mining of data streams [C]//Proceedings of the 4th SIAM International Conference on Data Mining. Lake Buena Vista, USA, 2004: 457-461.
- [50] ZHU X, ZHANG P, LIN X. Active learning from stream data using optimal weight classifier ensemble [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2010, 40(6): 1607-1621.
- [51] MASUD M, GAO J, KHAN L, et al. Classification and novel class detection in data streams with active mining [C]//Proceedings of the 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Hyderabad, India, 2010: 311-324.
- [52] ZLIOBAITE I, BIFET A, PFAHRINGER B, et al. Active learning with evolving streaming data [C]//Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Athens, Greece, 2011, 11: 597-612.
- [53] CHU W, ZINKEVICH M, LI L, et al. Unbiased online active learning in data streams [C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 195-203.
- [54] KHOLGHI M, KEYVANPOUR M R. Active learning framework combining semi-supervised approach for data stream mining [J]. Intelligent Computing and Information Science, 2011, 135: 238-243.
- [55] FORMAN G. Tackling concept drift by temporal inductive transfer [C]//Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto, Canada, 2006: 252-259.
- [56] BEN D S, BORBELY R S. A notion of task relatedness yielding provable multiple-task learning guarantees [J]. Machine Learning, 2008, 73(3): 273-287.
- [57] FOLINO G, PAPUZZO G. Handling different categories of concept drifts in data streams using distributed GP [C]//Berlin: Springer-Verlag, 2010: 74-85.
- [58] YANG S, YAO X. Population-based incremental learning with associative memory for dynamic environments [J]. IEEE Transactions on Evolutionary Computation, 2008, 12(5): 542-561.
- [59] WENERSTROM B, GIRAUD-CARRIER C. Temporal data mining in dynamic feature spaces [C]//Proceedings of the 2006 International Conference on Data Mining. Las Vegas, USA, 2006: 1141-1145.
- [60] CHEN H L, CHEN M S, LIN S C. Catching the trend: a framework for clustering concept-drifting categorical data [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(5): 652-665.
- [61] MASUD M, GAO J, KHAN L, et al. Classification and novel class detection in concept-drifting data streams under time constraints [J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(6): 859-874.
- [62] ZHANG P, LI J, WANG P, et al. Enabling fast prediction for ensemble models on data streams [C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego, USA, 2011: 177-185.
- [63] KOREN Y. Collaborative filtering with temporal dynamics [J]. Communications of the ACM, 2010, 53(4): 89-97.
- [64] KLINKENBERG R, JOACHIMS T. Detecting concept drift with support vector machines [C]//Proceedings of the 17th International Conference on Machine Learning. San Francisco, USA, 2000: 487-494.
- [65] WIDMER G, KUBAT M. Learning in the presence of concept drift and hidden contexts [J]. Machine Learning, 1996, 23(1): 69-101.
- [66] FAN W. Systematic data selection to mine concept-drifting data streams [C]//Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining. Seattle, USA, 2004: 128-137.
- [67] SALGANICOFF M. Tolerating concept and sampling shift in lazy learning using prediction error context switching [J]. Artificial Intelligence Review, 1997, 11(1/2/3/4/5): 133-155.
- [68] ŽLIOBAITĖ I. Combining similarity in time and space for training set formation under concept drift [J]. Intelligent Data Analysis, 2011, 15(4): 589-611.
- [69] KLINKENBERG R. Learning drifting concepts: example selection vs. example weighting [J]. Intelligent Data Analysis, 2004, 8(3): 200-281.
- [70] KOYCHEV I. Gradual forgetting for adaption to concept drift [C]//Proceedings of the ECAI 2000 Workshop Current Issue in Spatio-Temporal Reasoning. Berlin, Germany, 2000: 101-106.
- [71] OOMMEN J, RUEDA L. Stochastic learning-based weak estimation of multinomial random variables and its applications to pattern recognition in non-stationary environments [J]. Pattern Recognition, 2006, 39(3): 328-341.
- [72] NUNEZ M, FIDALGO R, MORALES R. Learning in environments with unknown dynamics towards more robust concept learners [J]. Journal of Machine Learning Re-

- search, 2007, 8: 2595-2628.
- [73] KYOSUKE N, KOICHIRO Y, TAKASHI O. ACE: adaptive classifier-ensemble system for concept-drifting environments[C]//Proceedings of the 6th International Workshop on Multiple Classifier Systems. Heidelberg, Germany, 2005: 176-185.
- [74] MUHLBAIER M D, POLIKAR R. Multiple classifiers based incremental learning algorithm for learning in non-stationary environments[C]//Proceedings of the 6th IEEE International Conference on Machine Learning and Cybernetics. Hong Kong, China, 2007: 3618-3623.
- [75] 文益民, 王耀南, 张莹. 基于可信多数投票的快速概念漂移检测[J]. 湖南大学学报: 自然科学版, 2010, 37(6): 36-40.
- WEN Yimin, WANG Yaonan, ZHANG Ying. Fast detection of concept drifts based on confident majority voting[J]. Journal of Hunan University: Natural Sciences, 2010, 37(6): 36-40.
- [76] 关菁华, 刘大有. 一种挖掘概念漂移数据流的选择性集成算法[J]. 计算机科学, 2010, 37(1): 204-207.
- GUAN Jinghua, LIU Dayou. Selected ensemble of classifiers for handling concept-drifting data streams[J]. Computer Science, 2010, 37(1): 204-207.
- [77] 朱群, 张玉红, 胡学钢, 等. 一种基于双层窗口的概念漂移数据流分类算法[J]. 自动化学报, 2011, 37(9): 1078-1083.
- ZHU Qun, ZHANG Yuhong, HU Xuegang, et al. A double-window-based classification algorithm for concept drifting data streams[J]. Acta Automatica Sinica, 2011, 37(9): 1078-1083.
- [78] 孙岳, 毛国君, 刘旭, 等. 基于多分类器的数据流中的概念漂移挖掘[J]. 自动化学报, 2008, 34(1): 93-97.
- SUN Yue, MAO Guojun, LIU Xu, et al. Mining concept drifts from data streams based on multi-classifiers[J]. Acta Automatica Sinica, 2008, 34(1): 93-97.
- [79] 辛轶, 郭躬德, 陈黎飞, 等. IKnnM-DHecoc: 一种解决概念漂移问题的算法[J]. 计算机研究与发展, 2011, 48(4): 592-601.
- XIN Yi, GUO Gongde, CHEN Lifei, et al. IKnnM-DHecoc: a method for handling the problem of concept drift[J]. Journal of Computer Research and Development, 2011, 48(4): 592-601.
- [80] KUNCHEVA L I, ZLIOBAITE I. On the window size for classification in changing environments[J]. Intelligent Data Analysis, 2009, 13(6): 314-323.
- [81] 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.
- [82] KLINKENBERG R, RENZ I. Adaptive information filtering: learning in the presence of concept drifts[C]// Proceedings of the Workshop of the AAAI-98/ICML-98 on Learning for Text Categorization. Madison, USA, 1998: 33-40.
- [83] LAST M. Online classification of nonstationary data streams[J]. Intelligent Data Analysis, 2002, 6(2): 1-16.
- [84] GAMA J, MEDAS P, CASTILLO G, et al. Learning with drift detection[C]//Proceedings of the 17th Brazilian Symposium on Artificial Intelligence. Sao Luis, Brazil, 2004: 286-295.
- [85] NISHIDA K, YAMAUCHI K. Detecting concept drift using statistical testing[C]//Proceedings of the 10th International Conference on Discovery Science. Sendai, Japan, 2007: 264-269.
- [86] 罗秀, 王大玲, 冯时, 等. 一种面向周期性概念漂移的数据流分类算法[J]. 计算机研究与发展, 2009, 46(Suppl): 400-405.
- LUO Xiu, WANG Daling, FENG Shi, et al. An algorithm for classifying data stream with recurrent concept drift[J]. Journal of Computer Research and Development, 2009, 46(Suppl.): 400-405.
- [87] 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.
- [88] ALIPPI C, ROVERI M. Just-in-time adaptive classifiers—part I: detecting nonstationary changes[J]. IEEE Transactions on Neural Networks, 2008, 19(7): 1145-1153.
- [89] ALIPPI C, ROVERI M. Just-in-time adaptive classifiers—part II: designing the classifier[J]. IEEE Transactions on Neural Networks, 2008, 19(12): 2053-2064.
- [90] 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. San Jose, USA, 2011: 1675-1681.
- [91] PETER V, ABRANHAM B. Entropy-based concept drift detection[C]//Proceedings of the 6th International Conference on Data Mining. Hong Kong, China, 2006: 1113-1118.
- [92] KUNCHEVA L. Change detection in streaming multivariate data using likelihood detectors[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 6(1): 1-7.
- [93] KUBAT M. To load balancing in computer networks[J]. International Journal of Cybernetics and System, 1992, 23(3/4): 389-400.
- [94] COHEN L, AVRAHAM BAKISH G, LAST M, et al. Real-time data mining of non-stationary data streams from sensor networks[J]. Information Fusion, 2008, 9(3): 344-353.
- [95] HARRIES M, HORN K. Detecting concept drift in financial time series prediction using symbolic machine learning