软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(10):2567−2580 [doi: 10.13328/j.cnki.jos.004747]

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

具有回忆和遗忘机制的数据流挖掘模型与算法*

赵强利[1], 蒋艳凰[2], 卢宇彤[2]

[1](湖南商学院 计算机与信息工程学院,湖南 长沙 410205)

[2](高性能计算国家重点实验室(国防科学技术大学),湖南 长沙 410073)

通讯作者: 赵强利, E-mail: zhao-qiangli@163.com; http://www.hnuc.edu.cn

摘 要: 集成式数据流挖掘是对存在概念漂移的数据流进行学习的重要方法.针对传统集成式数据流挖掘存在的缺陷,将人类的回忆和遗忘机制引入到数据流挖掘中,提出基于记忆的数据流挖掘模型 MDSM(memorizing based data stream mining).该模型将基分类器看作是系统获得的知识,通过“回忆与遗忘”机制,不仅使历史上有用的基分类器因记忆强度高而保存在“记忆库”中,提高预测的稳定性,而且从“记忆库”中选取当前分类效果好的基分类器参与集成预测,以提高对概念变化的适应能力.基于 MDSM 模型,提出了一种集成式数据流挖掘算法 MAE(memorizing based adaptive ensemble),该算法利用 Ebbinghaus 遗忘曲线对系统的遗忘机制进行设计,并利用选择性集成来模拟人类的“回忆”机制.与 4 种典型的数据流挖掘算法进行比较,结果表明:MAE 算法分类精度高,对概念漂移的整体适应能力强,尤其对重复出现的概念漂移以及实际应用中存在的复杂概念漂移具有很好的适应能力.不仅能够快速适应新的概念变化,并且能够有效抵御随机的概念波动对系统性能的影响.

关键词: 数据流挖掘;概念漂移;回忆与遗忘;Ebbinghaus 遗忘曲线;选择性集成

中图法分类号: TP181

中文引用格式: 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

英文引用格式: 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). http://www.jos.org.cn/1000-9825/4747.htm

Ensemble Model and Algorithm with Recalling and Forgetting Mechanisms for Data Stream Mining

ZHAO Qiang-Li[1], JIANG Yan-Huang[2], LU Yu-Tong[2]

[1](School of Computer and Information Engineering, Hu’nan University of Commerce, Changsha 410205, China)

[2](State Key Laboratory of High Performance Computing (National University of Defense Technology), Changsha 410073, China)

Abstract: Using ensemble of classifiers on sequential chunks of training instances is a popular strategy for data stream mining with concept drifts. Aiming at the limitations of existing approaches, this paper introduces human recalling and forgetting mechanisms into a data stream mining system, and proposes a memorizing based data stream mining (MDSM) model. The model considers base classifiers as learned knowledge. Through “recalling and forgetting” mechanism, most useful classifiers in the past will be reserved in a “memory repository”, which improves the stability under random concept drifts. The best classifiers for the current data chunk are selected for prediction, which achieves high adaptability for different concept drifts. Based on MSDM, the paper puts forward a new algorithm MAE (memorizing based adaptive ensemble). MAE uses Ebbinghaus forgetting curve as forgetting mechanism and adopts ensemble pruning to emulate the “recalling” mechanism. Compared with four traditional data stream mining approaches, the results show that MAE achieves high and stable accuracy with moderate training time. The results also proved that MAE has good adaptability for different kinds of concept drifts, especially for the applications with recurring or complex concept drifts.

* 基金项目: 国家自然科学基金(61272141, 60905032, 61120106005, 61273232)
收稿时间: 2014-07-31; 修改时间: 2014-09-03; 定稿时间: 2014-10-21

Key words: data stream mining; concept drift; recalling and forgetting; Ebbinghaus forgetting curve; ensemble pruning

分类是机器学习领域的重要应用方向,传统的分类问题主要采用静态批量处理的学习方式,即,一次性将所有的训练数据提交给学习系统.随着各应用领域数据获取能力越来越强,学习系统的动态学习能力日益重要.本文所讨论的数据流挖掘就是一种动态学习,即:训练数据持续不断地到来,学习系统如何在原来学习结果的基础上不断地对新产生的训练数据进行学习,并在实时性和预测能力方面满足应用的需求.社会网络挖掘、垃圾邮件分类、遥感数据识别、能源应用分析等领域都日益需要数据流挖掘技术[1].

数据流挖掘具有两个显著的特点[1,2]:一是高速产生样本数据,需要实时处理;二是数据所蕴含的概念随着时间发生变化,即,存在概念漂移,例如概念的突变、渐变、重复性变化等.一个好的数据流挖掘系统不仅需要实时处理不断到来的数据,而且要能够适应概念的不断变化.目前,已有的数据流挖掘算法大致可分为 3 类:滑动窗口、漂移检测和自适应集成学习.

- 滑动窗口学习[3,4]

通过移动时间窗口将训练数据集限制为最近出现的样本,并利用批量学习方法对窗口内的样本数据进行学习,获得新的分类器,预测时直接使用最近生成的分类器.对于较小的窗口,滑动窗口策略能够迅速反映出概念的变化,但是因学习的数据量较小,分类精度通常较低;对于较大的窗口,该策略则难以适应概念的快速变化.为此,一些学者提出:可启发式地动态调整窗口的大小[4],以便在学习精度和对概念变化的适应能力方面达到均衡.

- 漂移检测[5−7]

在学习系统中设计了一个概念漂移检测器,用于检测样本标识的分布是否发生变化.一旦发现存在概念漂移并达到某一阈值,则丢弃当前的分类器,并对预警窗口内的数据集进行学习,重新生成新的分类器.漂移检测技术适用于概念突变的应用,对于某些渐变的概念漂移,由于一直触发不了预警,导致检测不到概念漂移的存在.此外,较低的阈值使得算法对噪声数据十分敏感,从而降低了预测精度.

- 自适应集成学习是目前数据流挖掘的重要方法[8−12]

该方法将顺序到达的数据流划分成数据块,对每个数据块学习一个基分类器;系统保存一定数目的基分类器,并利用所保存的基分类器对新样本进行集成预测.SEA(streaming ensemble algorithm)[7]是最早的集成式数据流挖掘算法,它对每个数据块学习一个 C4.5 决策树,如果保存的基分类器数目达到规定的上限,则每产生一棵新的决策树,就利用启发式的方法从集成分类器中删除一个基分类器.SEA 算法采用大多数投票法对未知数据进行集成预测,由于所有基分类器都参与预测,而且它们的重要性相同,导致算法对突变的概念漂移适应性差.AWE(accuracy-weighted ensembles)[9]是数据流集成学习中具有代表性的算法,该算法根据各基分类器对当前数据块的分类精度为它们设置权重值,在替换基分类器时,直接删除权重最小的基分类器;在预测阶段,则根据权重对各基分类器的预测结果进行加权平均.相对于 SEA,这种权重设置方法提高了对概念变化的适应能力.ACE (adaptive classifier ensemble)[10]在传统自适应集成的基础上增加了概念漂移监测器,提高了对概念突变的适应能力,如果没有监测到概念变化,则采用加权投票的方式进行集成预测;如果监测到概念变化,则等到预警窗口充满时,重新学习一个新的分类器用于预测.最近提出的 AUE(accuracy updated ensemble)[11,12]算法采用与 AWE 算法类似的权重策略,但是每个基分类器都具有增量学习能力,可对新的数据块进行增量式学习,这种算法的缺陷是要求基分类器学习模型具有增量学习的能力.

上述集成式数据流挖掘算法均存在如下问题:

(1) 对基分类器的评估未考虑基分类器的历史重要性:这些算法都仅根据各基分类器对当前数据块的分类精度来评估基分类器,并确定删除哪些基分类器,这种评估方法忽略了基分类器的历史重要性;

(2) 小的概念波动容易导致有用的基分类器被删除:历史上重要的基分类器很可能因一次小的概念波动导致其评估值很差而被删除,对于概念不断频繁波动的实际应用,保留下来的基分类器往往不是全局占优的基分类器,从而难以获得好的预测结果.

本文针对当前自适应集成学习存在的缺陷,将人类的“回忆和遗忘”机制引入到数据流挖掘中,提出基于记忆的集成式数据流挖掘模型 MDSM.该模型不仅使历史上有用的基分类器不会因随机的概念波动被意外删除,而且选择最为有效的基分类器集合进行集成预测,从而能够综合提高预测的稳定性和预测精度.基于 MDSM 模型,本文提出了一种新的集成式数据流挖掘算法 MAE.该算法设计了一种基于 Ebbinghaus 遗忘曲线的基分类器评估方法,并利用选择性集成模拟人类的“回忆”机制.与传统的集成式数据流挖掘算法相比,MAE 算法不仅预测精度高、实时性好、能够很好地适应各种不同类型的概念漂移,尤其是对于概念频繁波动的实际应用,其预测精度明显优于传统的集成式数据流挖掘算法.

本文第 1 节介绍传统集成式数据流挖掘的缺陷.第 2 节详细阐述 MDSM 模型及其思想.第 3 节对 MAE 算法进行描述.第 4 节为实验结果及其分析.最后进行总结并讨论未来的研究方向.

1 传统集成式数据流挖掘的缺陷

首先,我们对传统集成式数据流挖掘进行分析.传统集成式数据流挖掘将顺序到达的数据流划分成数据块,每到达一个数据块 DB(data block),则利用批量学习的方法对该数据块进行学习,获得一个新的基分类器 c,并将 c 放入系统的集成基分类器库 ES(ensemble set)中.

$$c=Learn(DB),ES=ES\cup\{c\}.$$

然后,利用数据块 DB 对系统 ES 中的基分类器进行评估:

$$\boldsymbol{W}=Evaluate(ES,DB) \tag{1}$$

$\boldsymbol{W}$ 为评估结果.一般而言,$\boldsymbol{W}$ 为一个向量,ES 中的每个基分类器 c_i 在向量 $\boldsymbol{W}$ 中均对应着一个评估值 w_i.需要注意的是:在传统的集成式数据流挖掘算法中,$\boldsymbol{W}$ 仅与当前的基分类器库 ES 和数据块 DB 相关,与其他历史信息(如以前的评估值等)无关.如果 ES 中的基分类器数目达到规定的上限 k,则删除评估值最低的基分类器,使其满足:

$$|ES|\leqslant k.$$

在有预测任务时,系统直接利用集成基分类器库 ES 中的全部基分类器对新样本进行集成预测,对于未知样本 X,预测结果为

$$P(X)=Ensemble(ES,\boldsymbol{W},X) \tag{2}$$

传统的集成式数据流挖掘方法具有如下缺陷:(1) 评估过程中没有考虑基分类器的历史评价信息(见公式(1)),当 DB 为短时波动数据块时,则容易导致有用的基分类器因评估值很低而被删除;(2) 在预测过程中,直接对 ES 中的所有基分类器进行集成预测,由于 ES 中可能存在分类效果较差的基分类器,它们参与集成反而对集成预测结果产生负面的影响.

2 基于记忆的数据流挖掘模型

2.1 人类记忆的特点

1885 年,德国心理学家 Ebbinghaus 首先开展了人类记忆的研究[13],发现:如果不对所学知识进行复习回忆,人类很快会遗忘所学内容,有效的回忆能够增强对所学知识的记忆.Ebbinghaus 的研究结果可形象地用 Ebbinghaus 遗忘曲线表示,如图 1 所示,其中,纵坐标表示人类对所学知识的记忆强度,用百分比表示;横坐标为每次回忆所学知识的时间间隔.从图 1 可以看出:如果没有对所学知识进行复习回忆,那么人类对所学知识的记忆强度随着时间的推移将会呈指数衰减;每次对知识的复习回忆都能增强对该知识的记忆强度,并且对该知识的记忆变得更加稳定且不易被忘记.

Ebbinghaus 于 1885 年出版了《论记忆》专著,使得“记忆”成为心理学研究的重要领域.在后续的研究中[14],人们发现:回忆总是具有事件相关性,人类在对新信息进行学习的过程中,以前学习过的相关或类似的知识很容易被回忆起来;在使用知识解决实际问题时,也总是回忆起与该问题相关的知识,并利用这些知识去解决所面临

的问题.在每个人的生活过程中,所学知识不断地积累,每次回忆起的知识总是很小的一部分,而不可能将所学的所有知识都回忆起来.

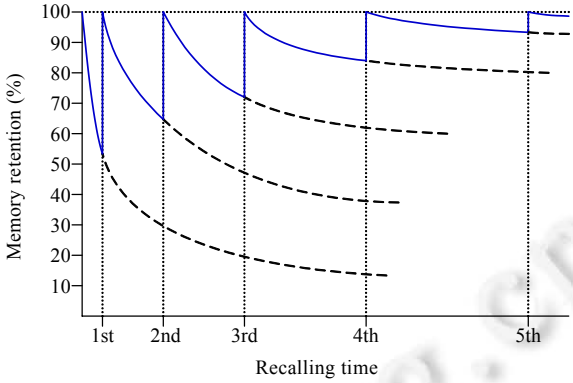

Fig.1 Ebbinghaus forgetting curve

图 1 Ebbinghaus 遗忘曲线

根据上述分析,我们可以得到如下人类记忆的特点:

(1) 记忆强度随着时间衰减:如果人类不对所学知识进行有意识的复习回忆,那么对该知识的记忆强度将随着时间呈指数衰减,并逐渐被遗忘;

(2) 回忆可增强对知识的记忆:每次对所学知识的回忆都能增强其记忆强度,并使相应的记忆更加稳定且不易忘记;

(3) 回忆的事件相关性:在对知识的回忆过程中,并不是把自己所学的所有知识都回忆起来并加以应用,而仅仅是与所处理事件相关的知识才会被回忆起来.

2.2 基于记忆的数据流挖掘模型

为了克服上述缺陷,我们将第 1 节所介绍的人类记忆的特点引入到集成式数据流挖掘中,提出基于记忆的数据流挖掘模型 MDSM(memorizing based data stream mining).该学习模型的思想是:将学习获得的基分类器看作是系统获得的知识,在系统中设定一个“记忆库”MS(memorized set),用于保存有用的知识.每到来一个新的数据块 DB,则先对 DB 进行学习获得新的基分类器 c,并将 c 放入系统的记忆库 MS 中.

$$c=Learn(DB),MS=MS\cup\{c\}.$$

于此同时,HDSM 模型将每个数据块看成是一个需要处理的“事件”,一旦新的数据块 DB 到达,与该“事件”相关的知识也被“回忆”起来.“回忆”的过程是从“记忆库”MS 中选出对 DB 分类效果最好的基分类器集合,表示它们与当前数据块相关而被系统“回忆”起来.

$$ES=Recall(MS,DB,k) \tag{3}$$

k 表示能够回忆起的最大基分类器数目.显然,被回忆起的基分类器集合 ES 和系统记忆库 MS 的子集,即,ES 满足:

$$ES\subseteq MS \text{ 且} |ES|\leqslant k.$$

然后,根据“回忆”的结果对 MS 中的基分类器进行重新评估,评估值 $\boldsymbol{W}$ 表示基分类器在系统“记忆库”中的记忆强度.本次被回忆起的基分类器,其记忆强度得到增强;没有被回忆起的基分类器,其记忆强度则会衰减.

$$\boldsymbol{W}=Evaluate(MS,ES,\boldsymbol{H}) \tag{4}$$

其中,$\boldsymbol{H}$ 为 MS 中各基分类器的历史信息.公式(4)表示根据“回忆”的结果 ES 和各基分类器的历史信息 $\boldsymbol{H}$,重新计算新的评估值 $\boldsymbol{W}$.与人类回忆与遗忘的特点相似,每个基分类器的记忆强度与其是否被“回忆起”以及时间的流逝相关.因此,与传统的集成式数据流学习不同,HDSM 模型中基分类器的评估值不仅与当前数据块 DB 相关(公

式(4)中的当前回忆结果 ES 与 DB 相关),而且与各分类器的历史信息 $\boldsymbol{H}$ 也密切相关.评估结束后,如果“记忆库”MS 中的基分类器数目超过设定的记忆容量 m,则直接删除其中记忆强度最低的基分类器,使其满足:

$$|MS|\leqslant m.$$

当有新的预测任务时,系统直接利用最近回忆起来的 ES 中的基分类器对未知样本进行集成预测.对于未知样本 X,预测结果为

$$P(X)=Ensemble(ES,X) \tag{5}$$

HDSM 模型的创新点是将人类的记忆与遗忘机制引入到集成式数据流挖掘中,一方面可以使历史上有用的基分类器能够较为稳定地保存在“记忆库”中,避免随机的概念波动导致有用的基分类器被意外删除;另一方面通过“回忆”机制,从“记忆库”中选择对预测当前数据块最为有效的基分类器参与集成预测,充分利用了数据流的时间局部性效应来提高预测精度.

图 2 给出了基于 MDSM 模型的数据流挖掘系统的组成图.与传统集成式数据流挖掘系统相比,该系统在数据获取与预处理、评估和优化、预测与应用方面的功能相同,主要不同点在于数据流挖掘部分.MDSM 采用了具有人类记忆特性的数据流挖掘方法,图 2 中,斜体文字表示相应功能所对应的人类记忆机制.例如:MDSM 模型中“基分类器评估”对应着人类的“遗忘机制”;“基分类器选择”对应着人类的“回忆机制”;“基分类器学习”则对应着人类的“知识学习”等.学习模型可以是任意监督学习方法,如决策树、神经网络、支持向量机等;系统所保存的基分类器库则对应着人类的“记忆库”.

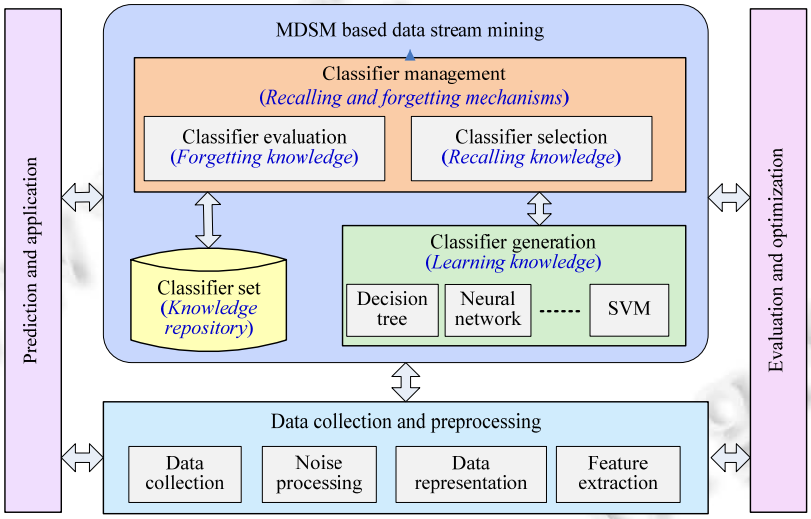

Fig.2 Structure of a MDSM based learning system

图 2 基于 MDSM 模型的数据流挖掘系统组成

3 MAE 算法

在 HDSM 模型的基础上,我们进一步提出了 MAE(memorizing based adaptive ensemble)算法.MAE 算法具有如下特点:

(1) 采用 Ebbinghaus 遗忘曲线作为 MDSM 模型的遗忘机制:Ebbinghaus 遗忘曲线是人类记忆学领域最典型的遗忘模型,MAE 算法据此设计了一种新的基分类器评估方法;

(2) 利用选择性集成来模拟人类的“回忆”机制:选择性集成是机器学习领域中提高集成分类器预测能力的重要方法,它利用校验样本集,从众多基分类器中选择部分基分类器进行集成,剔除对集成预测没有贡献的分类器,不仅能够提高预测精度,而且可以提高预测效率[15−20].MAE 算法利用选择性集成的

方法模拟人类的“回忆”机制,以当前数据块 DB 作为校验样本集,从 MS 中选择对 DB 预测能力强的基分类器组成目标集成分类器 ES,ES 中的基分类器即为系统“回忆”起的与当前数据块相关的知识.

MAE 算法引入了如下两个参数:

(1) 遗忘因子 f:每个基分类器 c 有其对应的遗忘因子 f_c,遗忘因子表明学习系统对该基分类器的记忆稳定性,取值为非负数.其值越小,表明系统对该基分类器的记忆越稳定,也越难遗忘.基分类器的遗忘因子与该基分类器被“回忆”起来的次数密切相关;

(2) 记忆强度 w:每个基分类器 c 有一个评估值 w_c,表示学习系统对该基分类器的记忆程度,取值为[0,1]区间内的非负数.其值越大,表明该基分类器在系统中越重要.基分类器的记忆强度由两个因素决定:一是基分类器的遗忘因子,该因子决定了图 1 所示中指数衰减曲线的形状;二是从最近一次被选中(如一直未被选中,则指该基分类器创建的时间)到当前的时间间隔,记忆强度随着时间的推移呈指数衰减.基分类器 c 的记忆强度计算方法如下:

$$w_c = \mathrm{e}^{-f_c\cdot(t-\tau_c)} \tag{6}$$

其中,w_c 表示基分类器 c 的记忆强度,f_c 为基分类器 c 的遗忘因子,τ_c 表示最近一次选中基分类器 c 的时间,t 为当前时间.

MAE 算法对每个数据块的处理可分为 3 个主要步骤:一是知识的获取;二是知识的回忆;三是知识的遗忘.

- 知识的获取阶段.

即为基分类器学习阶段,每当新的数据块到达,首先对其进行训练,获得一个新的基分类器 c,并初始化其记忆强度和遗忘因子:

$$w_c=1, f_c=\alpha.$$

表示目前系统对基分类器 c 的记忆强度为 1,其遗忘因子为初始值 α;然后,将学习获得的基分类器 c 加入基分类器库 MS 中.

- 对数据块学习完成后,则进入知识的回忆阶段.

系统以数据块 DB 为确认样本集,对 MS 中的所有基分类器进行选择性集成,即,执行公式(3)中的 $Recall$ 操作.在 MAE 算法中,$Recall$ 操作由一种选择性集成算法实现,并得到集成分类器 ES,即:

$$ES=ensemble\text{-}prune(MS,DB,k),$$

其中,k 为 ES 中的最大分类器数目,表示能够回忆起的最大基分类器数目.回忆的结果 ES 是对新数据块 DB 预测效果好的基分类器集合,它就是当前的目标集成分类器,用于最新的预测任务.

- 最后为知识的遗忘阶段.

首先,根据选择性集成的结果 ES,对这些选中的基分类器的遗忘因子等进行更新.我们用 λ_c 表示基分类器 c 被选择性集成选中(即,被回忆起)的次数,一旦基分类器 c 被选择性集成选中,则与 c 相关的 λ_c,τ_c 和 f_c 均得到更新,其中,λ_c 直接加 1,τ_c 更新为当前时间 t,即:

$$\text{if } c\in ES \text{ then } \lambda_c=\lambda_c+1,\ \tau_c=t.$$

遗忘因子的计算方法如下:

$$f_c = \frac{\alpha}{\lambda_c+1} \tag{7}$$

然后,利用公式(6)所示的 Ebbinghaus 遗忘曲线对 MS 中的所有基分类器的记忆强度进行更新.评估结束后,判定 MS 中的基分类器的数目是否超过上限(记忆容量)m,如果基分类器的数目大于 m,则删除记忆强度最低的基分类器,保证基分类器的数目小于等于 m.删除基分类器表示无用的知识被系统“遗忘”,已删除的基分类器将不会再被系统“回忆”起.MDSM 模型中的历史信息 $\boldsymbol{H}$(见公式(4))包括了各基分类器的历史信息,在 MAE 算法中,基分类器 c 的历史信息包括其被回忆起的总次数 λ_c 及其最近一次被回忆起的时间 τ_c.

在我们的 MAE 算法中,基分类器的评估值用记忆强度表示,记忆强度取决于该基分类器的历史信息,即,基分类器在历史上被系统“回忆”起来的次数和最近一次被系统“回忆”起来的时间,因此,该评估值综合考虑了基

分类器的历史重要性和最近的预测性能.通过设定较大的基分类器库 *MS*,使得暂时分类效果差、但历史分类效果好的基分类器能够保存在 *MS* 中而不会被删除,从而提高了算法的预测稳定性.同时,利用选择性集成从 *MS* 中选择当前分类效果最好的集成分类器 *ES* 用于预测,使得算法能够快速适应概念的变化.算法 1 给出了 MAE 算法的伪代码.

算法 1. Memorizing based adaptive ensemble (MAE).

输入:S:数据流样本;

m:记忆容量,即,记忆库 MS 中的最大基分类器数目;

k:能够回忆起的最大基分类器数目,$k\leqslant m$;

1: 初始化:$MS\leftarrow\varnothing$; $\alpha\leftarrow 1$; $t=0$;

2: For all data chunks $DB_t\in S$ do //对每个数据块,执行循环

2.1: $c\leftarrow learn(DB_t)$; //对 DB_t 进行学习,获得基分类器

2.2: $f_c\leftarrow 1$; $\lambda_c=0$; $\tau_c=t$; $f_c=\alpha$; //初始化基分类器 c 的相关参数

2.3: $MS\leftarrow MS\cup\{c\}$; //将 c 加入基分类器库中

2.4: $ES=ensemble\text{-}prune(MS,DB_t,k)$; //利用选择性集成方法回忆起与 DB_t 相关的基分类器

2.5: for all classifiers $c_i\in ES$

2.5.1: $\tau_{c_i}=t$; //更新回忆时间

2.5.2: $\lambda_{c_i}=\lambda_{c_i}+1$; //更新回忆次数

2.5.3: compute the forgetting factor of c_i based on Eq.(7); //计算遗忘因子

2.6: end for

2.7: for all classifiers $c_j\in MS$

2.7.1: compute the memory retention of c_j based on Eq.(6); //计算记忆强度

2.8: end for

2.9: if $|MS|>m$ then remove the classifier with the least memory retention from MS;

//遗忘记忆强度最小的基分类器

2.10: $t=t+1$;

3: end for

数据流挖掘要求在任何时刻都能够提供最新的学习结果[8,21,22].当有新的预测任务到达时,MAE 算法则直接返回当前的目标集成分类器 *ES*,这也是对最近一个数据块分类效果最好的基分类器集合.预测过程中,对 *ES* 中所有基分类器的预测结果采用大多数投票法(majority voting)确定最终预测结果.

4 实验设置

我们通过实验对 MAE 算法的性能进行比较分析.实验数据为 12 个来自不同领域的数据集,其中,SEA,Tree,RBF 和 LED 这 4 个数据集是利用 MOA 系统生成的人工数据集[23],分别采用 MOA 系统中不同的数据产生器获得,且概念漂移的类型也互不相同,其中,SEA 和 LED 数据集还引入了部分噪声.其他 8 个数据集均为实际应用领域的数据,其中,Electricity(Elec)是数据流挖掘中广泛使用的数据集[24],为电力市场的价格数据,其价格受市场需求、能源供应、季节、天气、每天的时间段等的影响;剩余的 7 个数据集均来自 UCI 机器学习数据库[25].

表 1 给出了各数据集的简单描述.上述实验数据集不仅包括渐变、突变、重复出现等各种不同类型的概念漂移,而且包括概念变化比较单一的人工数据和概念波动复杂的实际应用数据,可以较为全面地对各种算法进行比较.

参与比较的算法包括一种滑动窗口算法 Win 以及 3 种集成式数据流挖掘算法 SEA[8],AWE[9]和 ACE[10].当每个数据块到达时,各算法首先对该数据块进行预测,然后再对其进行学习.这些算法对数据块进行预测的处理方式各不相同.

- Win 算法为最简单的滑动窗口算法,窗口的大小即为数据块的大小,该算法仅保留最近生成的基分类器,并利用它进行预测.Win 算法主要用于与集成式数据流挖掘算法进行比较;
- SEA 是最早的集成式数据流挖掘算法,在预测时,使用大多数投票法对 *ES* 中各基分类器的预测结果进行集成;
- AWE 是一种加权的集成式数据流挖掘算法,在预测时,使用加权投票法对 *ES* 中各个基分类器的预测结果进行集成;
- ACE 结合了集成式数据流挖掘和概念漂移检测,如果未检测到概念漂移,则对 *ES* 中的基分类器采用加权投票的方式进行集成;若检测到概念漂移,则等到预警窗口充满,生成新的基分类器后,重新更新权重,然后采用加权投票进行集成;
- MAE 算法每次利用选择性集成方法从记忆库 *MS* 中选取 k 个基分类器组成 *ES*,当 *MS* 中的基分类器数目小于 k 时,则 *ES* 中的基分类器与 *MS* 相同.预测时,对 *ES* 中的基分类器预测结果采用大多数投票法进行集成.

实验中,5 种算法均采用相同的批量学习(batch learning)方式对数据块进行学习,对于基分类器的学习模型,我们选择了预测精度较高、学习速度也较快的 C4.5 决策树[26].以往的集成式数据流挖掘的研究结果表明[8−12]:数据块太小,则单个基分类器的分类精度差;数据块太大,则对数据流中的概念漂移适应能力差;数据块的大小设为 500 个样本是一个较好的选择.我们的实验也验证了这一研究结论,因此,我们将所有算法的数据块大小均设为 500 个样本.

Table 1 Data sets description

表 1 数据集描述

Dataset	No.Inst	No.Attrs	No.Cls	Noise	No.Drifts	Drift type	Application area
SEA	1 000 000	3	4	10%	3	Sudden	Artificial data
Tree	100 000	10	6	0%	15	Recurring	Artificial data
RBF	1 000 000	20	4	0%	4	Gradual	Artificial data
LED	1 000 000	24	10	10%	3	Gradual	Artificial data
Adult	48 842	14	2			Unknown	Census income
Bank	41 188	20	2			Unknown	Bank marketing
Conn	67 557	42	3			Unknown	Connect-4
Cover	581 012	53	7			Unknown	Cover type
Elec	45 312	7	2			Unknown	Electricity market
EEG	14 980	14	2			Unknown	EEG eye state
Person	164 860	7	11			Unknown	Person activity
Poker	1 025 010	10	10			Unknown	Poker game

MDSQ 是一种基于排名的选择性集成方法[19,27],在该方法中,每个基分类器都有一个“签名向量(signature vector)”,每次选择的基分类器是从所有剩余的基分类器中,将其加入集成分类器后,使得集成分类器的签名向量与目标向量间的距离最小的基分类器.我们对选择性集成算法进行了比较实验,结果表明:MDSQ 算法是目前选择性集成算法中不仅速度快,而且精度好的算法[28].因此,我们选择 MDSQ 算法作为 MAE 算法的回忆机制.文献[19,27]中的 MDSQ 算法缺省是选取 20%的基分类器构成目标集成分类器,我们对文献[19,27]中的 MDSQ 算法略作修改,即:当 *MS* 中的基分类器数目小于等于 k 时,则 *MS* 中的所有基分类器都被选中;当 *MS* 中的基分类器数目大于 k 时,则从 *MS* 中选取 k 个构成 *ES*,使得 MAE 算法与其他集成式数据流挖掘算法的 *ES* 集合大小相同.

本实验中涉及的所有数据流挖掘算法和 MDSQ 选择性集成算法,我们均采用高效的 C++语言来实现,并已将它们集成在我们自主设计开发的数据挖掘开源算法库 LibEDM(library of ensemble based data mining)中[29],该算法库也集成了 Quinlan 的 C4.5 决策树算法.LibEDM 软件可直接从软件开源平台 GitHub 上下载.本次实验平台的配置为:双路四核 Intel 处理器,主频 2.2GHZ,32GB 内存,Linux 操作系统.

5 实验设置

我们首先通过实验确定目标集成分类器的大小 k;然后,从预测精度、训练时间和测试时间这 3 个方面对 5 种算法进行了比较分析.实验中,各算法均采用先预测再学习(test-then-train)的方式,即:每到达一个数据块,首先对该数据块进行预测,获得数据块的预测精度,然后再对其进行学习.

5.1 参数k的设定

为了对算法进行公平的比较和分析,我们首先通过实验确定目标集成分类器的大小 k 的取值.参数 k 决定着参与集成预测的基分类器的最大数目,它的设定对集成式数据流挖掘算法的预测性能起着重要作用.为了获得最为合适的 k 值,针对不同的 k 值获得每种算法的预测均值 $MeanAcc$,它是算法在所有数据集上的平均数据块预测精度的均值,计算方式如下:

$$MeanAcc=\frac{1}{|DS|}\sum_{i\in DS}AveAcc_i=\frac{1}{|DS|}\sum_{i\in DS}\left\{\frac{1}{N_i}\sum_{j=1}^{N_i}Acc_{ij}\right\} \tag{8}$$

其中,DS 表示数据集的集合,N_i 为第 i 个数据集中的数据块总数目,Acc_{ij} 是对第 i 个数据集的第 j 个数据块的预测精度.图 3 给出了不同 k 值情况下,各种算法的 $MeanAcc$ 实验结果.在实验中,我们将 MAE 算法的记忆容量 m 设为 k 的 5 倍,即 $m=5k$.

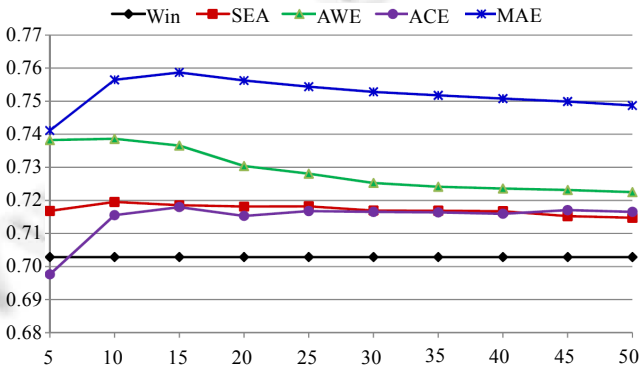

Fig.3 *MeanAcc* results of average chunk accuracies on all datasets for different k

图 3 不同 k 值情况下各种算法的 $MeanAcc$ 实验结果

从图 3 可以看出:在所有参与测试的算法中,无论 k 取什么值,MAE 算法的预测均值都是最好的.对于 AWE 和 SEA 算法,当 k 取 10 时,它们的预测性能最佳;对于 ACE 和 MAE 算法,k 取 15 时预测均值的结果略好于 k=10 时的结果.Win 算法由于每次只用最新的基分类器进行预测,其预测精度与 k 值的设置无关.从实验结果可知,k 值取[10,15]区间内的整数是较好的选择.考虑到大的 k 值会增加计算时间,在实验中,我们将参数 k 的取值设置为 10.对于 MAE 算法,其记忆容量则取 k 的 5 倍,即 m=50.

5.2 预测精度结果分析

表 2 给出了 5 种算法的预测精度结果,结果数据为相应数据集中所有数据块预测精度的均值,最后一行为各算法在所有数据集上的结果均值 $MeanAcc$(见公式(8)).每行结果中的最佳值用黑体表示.从实验结果可以看出:MAE 算法在 Tree,Adult,Conn,Elec,Person 和 Poker 这 6 个数据集上取得了最优结果,其中,Tree 数据集为概念重复变化的人工数据集,其他 5 个数据集属于概念波动频繁、变化难以预测的实际数据集.MAE 算法在所有数据集上的均值结果也是最好的,这主要是由于 MAE 算法的记忆与遗忘机制使得基分类器不会因对某一数据块分类效果差而被立即删除,它们会因较高的记忆强度而继续保留在系统的记忆库中,当相应概念再次出现时,会被系统回忆起来.因此,MAE 算法非常适合重复出现的概念漂移.实际应用中,常常是各种概念波动混合在一起,

重复性的概念漂移是实际应用中相当常见的一种概念波动,因此,MAE 算法对于实际数据能够获得好的预测性能.例如:Person 属于典型的数据流应用,不仅样本数据之间具有时间相关性,而且每种类别之间也具有时间相关性,分类困难,但是该数据集同一类别的数据会再次出现,数据集中的样本数目较多,充分发挥了 MAE 算法中记忆库的作用,使得 MAE 算法对该数据集的预测性能明显优于其他算法.

MAE 算法在所有数据集上预测精度的均值及方差的均值结果都是最好的,即:相比其他算法,MAE 算法不仅预测能力强,而且预测的性能更加稳定.这说明,我们提出的具有回忆和遗忘机制的 MAE 算法对概念漂移具有很强的适应能力,能够适应各种不同类型的概念变化,尤其是对于概念变化重复出现和概念变化难以预测的实际数据集,预测效果明显优于其他算法.

Table 2 Average predicting accuracy and variance on each data chunk (%)

表 2 数据块平均预测精度结果(%)

Datasets	Win	SEA	AWE	ACE	MAE
SEA	84.58±2.30	87.68±2.14	**87.98±2.14**	87.14±2.89	86.54±2.28
Tree	62.37±9.59	71.36±15.41	79.65±10.65	76.71±10.81	**81.70±8.34**
RBF	83.49±5.68	93.02±5.80	**93.59±4.34**	91.30±4.79	91.09±4.42
LED	40.36±22.17	49.75±23.88	**50.97±22.80**	50.00±22.93	50.80±22.72
Adult	92.81±2.30	94.38±1.64	94.46±1.65	94.29±1.66	**94.56±1.65**
Bank	87.36±14.73	88.73±14.23	88.62±14.75	**88.96±13.87**	88.89±13.99
Conn	64.99±15.91	67.67±15.53	67.43±17.46	68.52±15.99	**71.33±13.05**
Cover	**88.79±9.30**	82.31±10.98	84.18±8.91	66.97±17.01	87.92±9.05
EEG	59.62±31.08	54.31±26.85	59.03±31.07	**64.19±34.66**	62.05±30.18
Elec	73.79±10.33	76.68±9.48	76.08±9.83	75.26±11.64	**77.28±8.79**
Person	30.24±23.93	27.81±21.66	29.19±22.96	25.55±24.64	**34.31±22.40**
Poker	75.04±17.34	69.74±23.01	75.14±19.04	69.65±16.36	**81.30±16.89**
Mean	70.29±13.72	71.95±14.22	73.86±13.80	71.55±14.77	**75.65±12.81**

AWE 算法在 SEA,RBF,LED 这 3 个数据集上取得了最优的精度结果,均值结果位于第 2 位.这 3 个数据集均为人工数据集,其中,SEA 属于概念突然发生变化的数据集;RBF,LED 属于概念逐渐变化的数据集,对于这两种概念漂移,AWE 算法采用加权投票法,使得最近产生的基分类器具有较高的权重,从而能够获得好的预测性能.

SEA 算法的均值结果排名第三.SEA 采用大多数投票法将系统保留的基分类器进行集成预测,由于集成时既没有像 MAE 那样充分利用历史上有用的基分类器,也没有像 AWE 那样为基分类器设定权重,因此,其整体预测效果比 MAE 和 AWE 差一些.

ACE 算法通过引入概念变化监测器来捕获概念漂移,由于监测器对概念变化的阈值难以设定,很难适应各种不同的概念漂移.ACE 的概念变化监测器使其对一些瞬时的变化十分敏感,从而导致预测结果的波动很大.在我们的实验中,ACE 在 Bank 和 EEG 两个数据集上的预测精度最好,但其预测精度均值比 SEA 的结果还略差一点,方差的均值在所有算法中最大,这说明 ACE 是预测性能最不稳定的算法.

Win 只使用最近生成的基分类器进行预测,相当于遗忘了以前学的所有知识,因此对于相邻数据块相似度高的情况,能够取得好的预测结果.在我们的实验中,Win 算法仅在 Cover 数据集上获得最佳结果.4 种集成式数据流挖掘算法的均值结果都优于 Win 算法,这一结果再次验证了集成式数据流挖掘的有效性.

根据表 2 中的结果,我们利用单边 t 检验对 5 种算法进行两两比较,显著性水平设为 0.05,结果见表 3.表 3 中,(a,b)位置的取值为“+”表示算法 a 的预测精度明显优于算法 b,“–”表示算法 a 的结果明显差于算法 b,“*”则表示两种算法在当前的显著性水平下预测精度没有显著差别.从表 3 可以看出:Win,SEA 和 ACE 这 3 种算法的预测精度没有显著差别.AWE 算法的预测精度明显优于 Win 和 SEA,但与 ACE 相比,没有显著性差别,这主要是 ACE 算法波动相对较大所致.我们提出的 MAE 算法的预测精度则明显优于其他 4 种算法.

Table 3 *T*-Test for accuracy results

表 3 预测精度的显著性测试结果

Algorithm	Win	SEA	AWE	ACE	MAE
Win		*	–	*	–
SEA	*		–	*	–
AWE	+	+		*	–
ACE	*	*	*		–
MAE	+	+	+	+	

图 4 和图 5 分别给出了各种算法对 Elec 和 Conn 数据集中每个数据块的预测精度结果图,可以看出:MAE 算法不仅预测精度高,而且预测结果相对比较稳定,很少出现精度大幅度下降的情况.这正是由于 MAE 算法在其“记忆库”中保存了记忆强度高的基分类器,从而增强了对概念变化的适应能力.

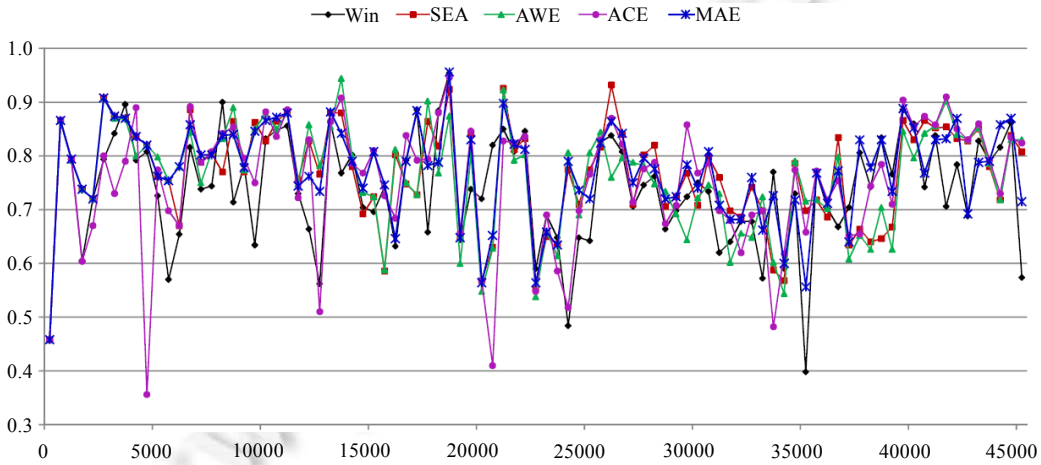

Fig.4 Chunk accuracy results of each algorithm on Elec dataset

图 4 Elec 数据集上各种算法对每个数据块的分类精度

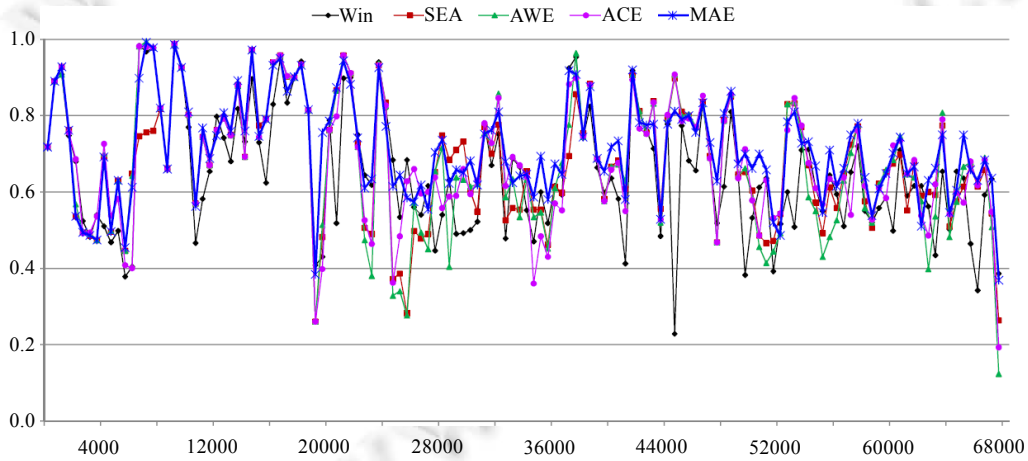

Fig.5 Chunk accuracy results of each algorithm on Conn dataset

图 5 Conn 数据集上各种算法对每个数据块的分类精度

5.3 训练时间比较

表 4 给出了 5 种算法的训练时间结果,其值为训练每个数据块所需要的时间均值,单位为 10^{-3}s.最后一行为

各种算法在所有数据集上的结果均值.从实验结果可以看出:Win 算法只是简单地训练了一个基分类器,训练时间最短;AWE 算法要为每个基分类器计算权重,训练时间比 Win 算法要长;SEA 算法在对基分类器进行评估以确定删除哪个基分类器上花费了一些时间,训练时间比 AWE 算法又长一点;ACE 算法中引入概念漂移检测的功能,较为耗时,其训练时间最长;MAE 算法的训练时间介于 SEA 与 ACE 之间.从整体的学习时间结果可知:MAE 算法对每个数据块的学习时间在 10^{-2}s 左右,能够满足实时学习的需求.

Table 4 Average chunk training time (10^{-3}s)

表 4 数据块的平均训练时间结果(10^{-3}s)

Dataset	Win	SEA	AWE	ACE	MAE
SEA	1.43	6.93	4.44	14.99	14.75
Tree	4.86	10.95	8.42	37.01	19.38
RBF	16.50	22.23	19.71	77.93	30.80
LED	5.74	12.68	9.56	23.81	22.30
Adult	2.79	6.06	5.73	28.96	10.11
Bank	1.86	7.96	5.81	13.84	10.95
Conn	4.56	9.11	7.29	21.05	14.04
Cover	6.96	12.62	10.14	80.92	20.56
EEG	2.41	5.11	3.97	13.76	7.85
Elec	2.10	7.08	4.89	20.02	11.15
Poker	2.19	7.90	5.40	48.55	16.38
Person	2.63	9.87	7.61	15.25	12.33
Mean	4.50	9.88	7.75	33.01	15.88

5.4 预测时间比较

表 5 给出了所有算法对单个数据块的平均预测时间结果,单位为 10^{-6}s.最后一行给出了各种算法在所有数据集上的结果平均值.

Table 5 Average chunk testing time (10^{-6}s)

表 5 数据块的平均预测时间结果(10^{-6}s)

Dataset	Win	SEA	AWE	ACE	MAE
SEA	0.30	2.00	2.50	4.50	2.00
Tree	0.50	3.50	3.50	21.50	3.55
RBF	0.50	3.55	3.50	33.00	3.05
LED	0.50	4.00	4.00	5.00	3.50
Adult	0.50	2.20	3.50	4.50	3.50
Bank	1.50	17.20	19.50	153.00	23.00
Conn	0.45	14.76	23.10	33.40	22.10
Cover	0.30	0.50	0.40	1.76	0.40
EGG	1.30	3.20	4.50	5.00	4.50
Elec	22.50	41.90	33.00	92.70	33.00
Person	0.30	11.50	11.70	16.50	12.30
Poker	2.50	4.70	3.10	9.10	3.30
Mean	2.60	9.08	9.36	31.66	9.52

从表 5 可以看出:Win 算法的预测时间最短,这是因为 Win 只用了一个基分类器进行预测;SEA,AWE 和 MAE 算法均采用约 10 个基分类器进行集成预测,它们的预测时间基本相等;ACE 算法的预测时间在所有算法中最长.对于所有算法,它们的数据块预测时间约为其训练时间的千分之一.

6 结 论

本文提出基于记忆的数据流挖掘模型 MDSM,并基于该模型提出了一种新的集成式数据流挖掘算法 MAE,用于验证模型的有效性.MDSM 模型的主要创新在于,它将人类的回忆与遗忘机制引入到集成式数据流挖掘中,MAE 算法的创新点在于:它利用 Ebbinghaus 遗忘曲线来实现 MDSM 系统的遗忘机制,并采用选择性集成实现系统的回忆机制.本文将 MAE 算法与 4 种典型的数据流挖掘算法 Win,SEA,AWE,ACE 进行了比较,实验结果表明:MAE 算法不仅预测精度最高,而且预测的稳定性也最好,尤其是对于重复出现的概念漂移和实际应用中的

复杂概念漂移,预测效果明显优于其他算法.MAE 算法的数据块训练时间在 10^{-2} 秒级,预测时间在 10^{-5} 秒级,能够满足应用的实时性需求.

由实验结果可知,MAE 算法对突发性的概念漂移和渐变性的概念漂移的适应能力相对差一些.如何对算法作进一步的优化,增强其对上述两种概念漂移的适应性,是我们下一步的工作重点.此外,具有回忆和遗忘机制的数据流挖掘是一种新的思路和方法,我们将展开更为全面而深入的理论分析和实验验证,以进一步论证 MDSM 模型和 MAE 算法的有效性.

References:

[1] Sayed-Mouchaweh M, Lughofer E. Learning in Non-Stationary Environments: Methods and Applications, New York: Springer-Verlag, 2012. [doi: 10.1007/978-1-4419-8020-5]

[2] Gama J. Knowledge Discovery from Data Streams. London: Chapman & Hall, 2010.

[3] Cohen E, Strauss MJ. Maintaining time-decaying stream aggregates. Journal of Algorithms, 2006,59(1):19−36. [doi: 10.1016/j.jalgor.2005.01.006]

[4] Bifet A, Gavaldà R. Learning from time-changing data with adaptive windowing. In: Apte C, Skillicorn D, Liu B, Parthasarathy S, eds. Proc. of the 7th SIAM Int'l Conf. on Data Mining. Society for Industrial and Applied Mathematics, 2007. 443−448.

[5] Page ES. Continuous inspection schemes. Biometrika, 1954,41(1-2):100−115. [doi: 10.2307/2333009]

[6] Gama J, Medas P, Castillo G, Rodrigues P. Learning with drift detection. In: Proc. of the 17th SBIA Brazilian Symp. on Artificial Intelligence. 2004. 286−295. [doi: 10.1007/978-3-540-28645-5_29]

[7] Baena-García M, Campo-Ávila JD, Fidalgo R, Bifet A, Gavaldà R, Morales-Bueno R. Early drift detection method. In: Gama J, Aguilar-Ruiz JS, Klinkenberg R, eds. Proc. of the 4th Int'l Workshop Knowledge Discovery in Data Streams. 2006. 1−10. http://www.ecmlpkdd2006.org/ws-kdds.pdf

[8] Street WN, Kim Y. A streaming ensemble algorithm (SEA) for large-scale classification. In: Proc. of the 7th ACM SIGKDD Int'l Conf. on Knowledge Discovery in Data Mining. 2001. 77−382. [doi: 10.1145/502512.502568]

[9] Wang H, Fan W, Yu PS, Han J. Mining concept-drifting data streams using ensemble classifiers. In: Proc. of the 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery in Data Mining. 2003. 226−235. [doi: 10.1145/956750.956778]

[10] Nishida K, Yamauchi K, Omori T. ACE: Adaptive classifiers-ensemble system for concept-drifting environments. In: Proc. of the 6th Int'l Workshop Multiple Classifier System. 2005. 176−185. [doi: 10.1007/11494683_18]

[11] Brzezinski D, Stefanowski J. Accuracy updated ensemble for data streams with concept drift. In: Proc. of the 6th HAIS Int'l Conf. on Hybrid Artificial Intelligence Systems (HAIS 2011). 2011. 155−163. [doi: 10.1007/978-3-642-21222-2_19]

[12] Brzezinski D, Stefanowski J. Reacting to different types of concept drift: The accuracy updated ensemble algorithm. IEEE Trans. on Neural Networks and Learning System, 2014,25(1):81−94. [doi: 10.1109/TNNLS.2013.2251352]

[13] Ebbinghaus H. Memory: A contribution to experimental psychology. 2012. http://nwkpsych.rutgers. edu/~jose/courses/578_mem_learn/2012/readings/Ebbinghaus_1885.pdf

[14] Coon D. Introduction to Psychology: Exploration and Application. 7th ed., Minneapolis-St.Paul: West Publishing Company, 1995.

[15] Zhou ZH, Xin J, Tang W. Ensembling neural networks: Many could be better than all. Artificial Intelligence, 2002,137(1-2):239−263. [doi: 10.1016/S0004-3702(02)00190-X]

[16] Lazarevic A, Obradovic Z. The effective pruning of neural network classifiers. In: Proc. of the IEEE/INNS Int'l Conf. on Neural Networks (IJCNN 2001). 2001. 796−801. [doi: 10.1109/IJCNN.2001.939461]

[17] Zhao Q, Jiang Y, Xu M. A fast ensemble pruning algorithm based on pattern mining process. Data Mining and Knowledge Discovery, 2009,19(2):227−292. [doi: 10.1007/978-3-642-04180-8_19]

[18] Zhao Q, Jiang Y, Xu M. Fast ensemble pruning based on FP-tree. Ruan Jian Xue Bao/Journal of Software, 2011,22(4):709−721 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3752.htm [doi: 10.3724/SP.J.1001.2011.03752]

[19] Martinez-Munoz G, Hernández-Lobato D, Suarez A. An analysis of ensemble pruning techniques based on ordered aggregation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2009,31(2):245−259. [doi: 10.1109/TPAMI.2008.78]

[20] Zhou ZH. Ensemble Methods: Foundations and Algorithms. Boca Raton: CRC Press, 2012.

[21] Domingos P, Hulten G. A general framework for mining massive data streams. Journal of Computational and Graphical Statistics, 2003,12(4):945−949. [doi: 10.1198/1061860032544]

[22] Bifet A, Holmes G, Kirkby R, Pfahringer B. MOA: Massive online analysis. Journal of Machine Learning Research, 2010,11: 1601−1604.

[23] Bifet A, Holmes G, Kirkby R, Pfahringer B, *et al*. MOA: Massive online analysis. 2014. http://moa.cms.waikato.ac.nz/

[24] Harries M. SPLICE-2 comparative evaluation: Electricity pricing. Technical Report, 9905, New South Wales: School of Computer Science and Engineering, University of New South Wales, 1999.

[25] Lichman M. UCI machine learning repository. 2013. http://archive.ics.uci.edu/ml/

[26] Quinlan JR. C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann Publishers, 1993.

[27] Martinez-Munoz G, Suarez A. Aggregation ordering in Bagging. In: Hamza MH, ed. Proc. of the IASTED Int'l Conf. on Artificial Intelligence and Applications. Calgary: ACTA Press, 2004. 258−263.

[28] Zhao Q, Jiang Y, Xu M. Categorization and comparison of the ensemble pruning algorithms. Computer Engineering and Science, 2012,34(2):134−138 (in Chinese with English abstract). [doi: 10.3969/j.issn.1007-130X.2012.02.025]

[29] Zhao Q, Jiang Y. Library for ensemble based data mining. 2014. https://github.com/Qiangli-Zhao/LibEDM

附中文参考文献:

[18] 赵强利,蒋艳凰,徐明.基于FP-Tree的快速选择性集成算法.软件学报,2011,22(4):709−721. http://www.jos.org.cn/1000-9825/3752.htm [doi: 10.3724/SP.J.1001.2011.03752]

[28] 赵强利,蒋艳凰,徐明.选择性集成算法分类与比较.计算机工程与科学,2012,34(2):134−138. [doi: 10.3969/j.issn.1007-130X.2012.02.025]

赵强利(1973－),男,陕西宝鸡人,博士,讲师,主要研究领域为机器学习,数据挖掘.

蒋艳凰(1976－),女,博士,副研究员,主要研究领域为机器学习,高性能计算.

卢宇彤(1969－),女,博士,研究员,博士生导师,CCF会员,主要研究领域为高性能计算,大数据分析.