类别严重不均衡应用的在线数据流学习算法

赵强利[1,2] 蒋艳凰[2]
(湖南商学院计算机与信息工程学院 长沙 410205)[1]
(国防科技大学高性能计算国家重点实验室 长沙 410073)[2]

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。对于类别分布严重不均衡的应用,集成式数据流挖掘中数据块的学习方式导致样本数多的类别的分类精度高,样本数少的类别的分类精度低的问题,现有算法无法满足此类应用的需求。针对上述问题,对基于回忆机制的集成式数据流学习算法MAE(Memorizing based Adaptive Ensemble)进行改进,提出面向类别严重不均衡应用的在线数据流学习算法UMAE(Unbalanced data Learning based on MAE)。UMAE算法为每个类别设置了一个样本滑动窗口,对于新到达的数据块,其样本依据自身的类别分别进入相应的滑动窗口,最后利用各类别滑动窗口内的样本构建用于在线学习的数据块。与5种典型的数据流挖掘算法的比较结果表明,UMAE算法在满足实时性的同时,不仅整体分类精度高,而且对于样本数很少的小类别的分类精度有大幅度提高;对于异常检测等类别分布严重不均衡的应用,UMAE算法的实用性明显优于其他算法。

关键词 在线学习,数据流挖掘,回忆与遗忘机制,不均衡数据学习
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.06.044

Online Data Stream Mining for Seriously Unbalanced Applications

ZHAO Qiang-li[1,2] JIANG Yan-huang[2]
(School of Computer and Information Engineering, Hunan University of Commerce, Changsha 410205, China)[1]
(State Key Laboratory of High Performance Computing, National University of Defense Technology, Changsha 410073, China)[2]

Abstract Using ensemble of classifiers on sequential blocks of training instances is a popular strategy for data stream mining with concept drifts. Yet for the seriously unbalanced applications where the number of examples for each class in the data blocks is totally different, traditional data block creation will result in low accuracy for the small classes with much less number of instances. This paper provided an updating algorithm UMAE (Unbalanced data learning based on MAE) for seriously unbalanced applications based on MAE (Memorizing based Adaptive Ensemble). UMAE sets an equal-sized sliding window for each class. When each data block comes, each example in the data block comes into the corresponding sliding window based on its classes. During the learning process, a new data block will be created by using the instances in the current sliding windows. This new data block is adopted to generate a new classifier. Compared with five traditional data stream mining approaches, the results show that UMAE achieves high accuracy for seriously unbalanced applications, especially for the small classes with much less number of instances in the applications.

Keywords Online learning, Data stream mining, Recalling and forgetting mechanisms, Unbalanced data learning

1 引言

数据流挖掘是一种在线学习方法[1-2],即训练数据持续不断地到来,学习系统在原来学习结果的基础上不断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。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,而且能够适应概念的不断变化。

自适应集成学习是目前数据流挖掘的重要方法[3-7]。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)[3]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,则每产生一个新的决策树就利用启发式的方法替换集成分类器中的一个基分类器。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且它们的重要性相同,因此导致算法对突变的概念漂移适应性差。AWE(Accuracy-Weighted Ensembles)[4]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,根据权重对各基分类器的预测结果进行加权平均。权重的引入提高

到稿日期:2016-05-18 返修日期:2016-08-28 本文受国家自然科学基金(61272141,61120106005,61472136),国防科技大学高性能计算国家重点实验室基金(201513-02)资助。
赵强利(1973—),博士,讲师,主要研究方向为机器学习,E-mail:zhao-qiangli@163.com(通信作者)。

了 AWE 对概念漂移的适应能力。为了进一步提高对概念漂移的敏感性，K. Nishida 等提出了 ACE(Adaptive Classifier Ensemble)[5]算法。ACE 也是一个加权的集成式数据流挖掘算法，但是它引入了一个概念漂移监测器：如果没有监测到概念变化，则采用与 AWE 算法类似的方式更新基分类器；但是一旦监测到概念变化，算法立即进行基分类器的淘汰和学习，从而加快了对于概念突变的反应速度。由于概念漂移监测器的阈值常常难以设定，因此 ACE 算法对不同应用的适应性较差。MAE(Memorizing based Adaptive Ensemble)[6-7]算法将人类学习过程中的记忆与遗忘机制引入到集成式数据流挖掘，将学习得到的基分类器看作是人类学习过程中新获取的知识，并通过历史重要程度对知识的记忆强度进行更新，记忆强度低的基分类器逐渐被系统遗忘(删除)，这种记忆与遗忘机制使得系统在对概念变化的反应速度和已有知识的有效利用之间达到了有效平衡，从而提高了对复杂概念漂移的适应能力。

现实生活中的很多应用都存在类别分布严重不均衡的问题，即每次到达的数据块中经常出现有些类别的样本数目太少甚至缺失，而有些类别的样本数目比例很高的情况。一般当数据集中不同类别的数据量相差超过一个数量级时就可以认为类别严重不均衡，入侵检测、故障预测等异常检测类的应用均属于类别分布严重不均衡的应用，例如在 KDDCUP99 的入侵检测数据中，DOS 攻击与其他类型攻击的数据量存在着 2～4 个数量级的差距。传统的集成式数据挖掘均未考虑这种类别分布严重不均衡的现象，基分类器学习算法如神经网络、C4.5 决策树(剪枝)算法等，都倾向于忽视较少的数据类别，而集成学习的投票算法加剧了这个问题，使得算法在面对异常检测这类问题时异常检出率较低，甚至完全无法使用。通过引入记忆库，MAE 算法的记忆与遗忘机制保存了部分历史基分类器，从而能够减少类别不均衡问题带来的影响，但是对于总体上数据类别严重不均衡的情况，在学习时仍然力不从心。

本文在我们前期提出的 MAE 算法的基础上，对类别严重不均衡问题的数据流学习进行深入研究，提出一种针对不均衡问题的集成式数据流挖掘算法 UMAE(Unbalanced data Learning based on MAE)。在 UMAE 算法中，设计了一种新的基分类器训练集生成方法，即为每个类别设置了一个滑动窗口，对于新到达的数据块，不是直接将其用于在线学习，而是使数据块内的样本按照自身的类别分别进入相应的滑动窗口，最后基于各类别滑动窗口内的样本构建用于在线学习的训练数据块。对于样本数目少的类别，其滑动窗口更新慢，而对于样本数目多的类别，其滑动窗口更新快，不同的更新速度使得参与基分类器学习时每个类别的样本数目基本持平，从而将类别不均衡问题转化为类别相对均衡的问题，提高了算法对类别不均衡问题的学习能力。

2 传统集成式数据流挖掘的缺陷

传统集成式数据流挖掘算法(如 SEA，AWE，ACE 等)将顺序到达的数据流划分成大小相等的数据块，将数据流 DS(Data Stream)看成是按数据块的方式到达，每到达一个数据块 DB(Data Block)后，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 放入集成分类器 ES(Ensemble Set)中，然后利用数据块 DB 对 ES 中的基分类器进行评估。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。

图 1 示出了传统集成式数据流挖掘中基分类器训练集的生成过程：随着数据流的不断到达，一旦收到的数据中样本数目到达设定的数据块的大小，则这些数据构成一个数据块，然后以该数据块作为训练样本集对其进行学习，获得一个新的基分类器。这种基分类器的学习方式存在如下问题：如果应用本身为类别不均衡问题，或在某一时间段内样本发生突然的变化，导致数据块 DB 内各类别对应的样本数目差别明显，甚至有些类别没有样本，则用这种数据块作为训练样本集获得的基分类器 c 对应用的预测性能较差。同时当用该数据块对 ES 中的基分类器进行评估时，评估值并不能真正反映基分类器的实际分类效果，而且有可能将好的基分类器删除，从而导致集成式分类的效果并不理想。

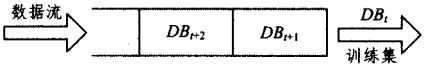

图 1 传统集成式数据流挖掘的训练集生成

我们前期提出的 MAE 算法将人类记忆的特点引入到集成式数据流挖掘中，算法本身具有类似于人类的回忆与遗忘的功能。MAE 算法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，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。MAE 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 分类效果最好的集成分类器 ES，ES 中的基分类器均是与当前数据块相关而被系统“回忆”起来的历史知识。然后根据“回忆”的结果对 MS 中的基分类器进行重新评估，每个基分类器的评估值表示该基分类器在系统“记忆库”中的记忆强度。本次被回忆起的基分类器的记忆强度得到增强，没有被回忆起的基分类器的记忆强度则衰减。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。

MAE 算法的记忆与遗忘机制的作用包括：1)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，避免某个概念变化大的数据块导致有用的基分类器被意外删除；2)通过“回忆”机制从“记忆库”中选择对预测当前数据块最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。相比其他传统集成式学习方法，MAE 算法能够获得更好且更为稳定的预测性能，但是 MAE 算法仍然不能解决应用本身的数据严重不均衡问题以及数据块内某个类别的样本缺失问题。

3 UMAE 算法

针对 MAE 算法在处理类别不均衡问题时存在的缺陷，本文在 MAE 算法的基础上提出一种不均衡数据的集成式数据流挖掘算法 UMAE(Unbalanced data Learning based on MAE)。

UMAE 算法改变了基分类器的训练样本集获取方式，它

为每个类别设置了一个滑动窗口，对于新到的数据块内的样本，根据自身的类别分别进入相应的滑动窗口，最后利用各类别的滑动窗口内的样本构建用于学习的数据块以参与数据流在线学习。

图 2 示出了 UMAE 算法中基于类别滑动窗口的训练集生成示意图。假设类别的数目为 r，每当数据流中一个新的数据块 DB 到达时，DB 中属于不同类别的样本则进入该类别所对应的滑动窗口中。每个类别的滑动窗口大小相同，当滑动窗口充满后，每进入一个新的样本，则按进入滑动窗口的时间顺序删除最早进入窗口内的样本。

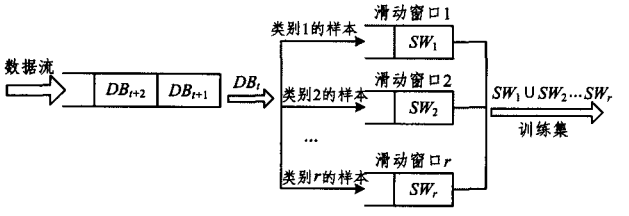

图 2　基于类别滑动窗口的训练集生成

MAE 算法直接采用 DB 作为新的基分类器的训练样本。与 MAE 算法不同，在 UMAE 算法中，每到来一个新的数据块 DB，先根据 DB 中的样本信息对每个类别的滑动窗口进行更新，令更新后的滑动窗口内的数据集分别为 SW_1，SW_2，…，SW_r。新的训练样本集由这 r 个滑动窗口内的样本重新构建。一种简单的构建方法是用这 r 个滑动窗口内所有样本的集合作为新的训练样本集，即：

$$c=Learn(SW_1 \cup SW_2 \cup \cdots \cup SW_r), MS=MS \cup \{c\} \quad (1)$$

其中，c 为新获得的基分类器，MS 为系统的记忆库。

为了保证每个类别的样本数目均衡，将各类别所对应的滑动窗口的大小设为相等。同时考虑到两种极端的情况：1)DB中每个类别的样本数目完全相同，这时由于各类别的样本数目完全均衡，有无滑动窗口的效果相同，滑动窗口的大小可设置为数据块 DB 大小的$\frac{1}{r}$，从而使 DB 内的样本数目均在新的训练集中；2)DB 中所有样本属于同一个类别，这时需要将滑动窗口的大小设置为与数据块 DB 的大小一样才能使新到的样本均在训练样本集中。因此，滑动窗口的大小 s 可设置为：

$$\frac{1}{r} \cdot |DB| \leqslant s \leqslant |DB| \quad (2)$$

当应用的类别不均衡问题严重时，可以将 s 的值设置得较大一些，在解决类别不均衡问题的同时使新的样本尽量得到学习；当应用属于类别较为均衡的问题时，可将 s 设置得较小一些，从而使训练集内的样本更新更快，适应能力更强。

UMAE 算法采用与 MAE 算法相同的记忆与遗忘机制，仍利用选择性集成的方法从基分类器库中选择被回忆起来的基分类器，并采用如下方式计算每个基分类器的遗忘因子，式(3)给出了基分类器 c 的遗忘因子计算方法：

$$f_c=\frac{\alpha}{\lambda_c+1} \quad (3)$$

其中，α 为遗忘因子的初始值，缺省为 1。λ_c 表示基分类器 c 被回忆起(即选择性集成方法选中)的次数。基分类器 c 的记忆强度计算方法如下：

$$w_c=e^{-f_c \cdot (t-\tau_c)} \quad (4)$$

其中，τ_c 表示最近一次选中基分类器 c 的时间，t 为当前时间。UMAE 的伪代码如算法 1 所示。

算法 1　Unbalanced data Learning based on MAE (UMAE)

输入：DS：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≤m

1. 初始化：MS←Ø；α←1；t=0；
2. For all data blocks $DB_t \in DS$ do //对每个数据块，执行循环
   2.1 For all instances $e \in DB_t$ do //对数据块内的每个样本，执行循环
      2.1.1 i=Class(e)；//获取样本的类别标识
      2.1.2 SW_i=Insert(SW_i,e)；//将样本加入其类别对应的滑动窗口
   2.2 SW=$SW_1 \cup SW_2 \cup \cdots \cup SW_r$//生成新的训练样本集 SW
   2.3 c←learn(SW)；//对 SW 进行学习，获得基分类器 c
   2.4 w_c←1；λ_c=0；τ_c=t；f_c=α；//初始化基分类器 c 的相关参数
   2.5 MS←MS ∪{c}；//将 c 加入基分类器库中
   2.6 ES=ensemble-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
   2.7 for all classifiers $c_i \in ES$
      2.7.1 τ_{ci}=t；//更新回忆时间
      2.7.2 λ_{ci}= λ_{ci}+1；//更新回忆次数
      2.7.3 compute the forgetting factor of c_i based on equation (3)；//计算遗忘因子
   2.8 end for
   2.9 for all classifiers $c_j \in MS$
      2.9.1 update the memory retention of c_j based on equation (4)；//计算记忆强度
   2.10 end for
   2.11 if |MS| >m then remove the classifier with the lowest memory retention from MS；
      //遗忘记忆强度最小的基分类器
   2.12 t=t+1；
3. end for

若有新的预测任务到达，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)来确定最终预测结果。

4　实验设置

网络入侵检测是典型的类别不均衡问题。采用 Kddcup99[8]数据集进行实验，该数据集包括 494021 条网络记录(样本)，包括 20 余种不同类型的攻击，这些攻击分别属于 4 种主要的攻击类型，因此样本集包括 Normal，DOS，Probe，R2L 和 u2r 5 种类别，各类别的含义及其所占的样本数目如表 1 所列。DOS 攻击的样本数目最多，占总样本数的 79.24%，其原因是这种类型的攻击通过大量发送报文迫使网络拥塞，从而使服务器无法对外提供服务。对于这种类型的攻击，检测到的记录必然很多，而且很多记录是由一次攻击产生的。Normal 属于正常状态的记录，占总样本数的 19.69%。Probe 为扫描攻击，R2L 攻击为来自远程机器的非法访问，u2r 攻击指本地机器非法获取超级用户权限。相对样本总数而言，后面 3 种类型的攻击样本所占比例均很少，分别占总样本数的 0.83%，0.23%和 0.01%，然而这 3 种攻击均会严重

影响到系统的安全性，对它们的预测结果至关重要。

表 1 Kddcup99 数据集中各类别的样本分布

类别	类别说明	样本数目	样本比例/%
Normal	正常网络行为	97278	19.69
DOS	拒绝服务攻击	391458	79.24
Probe	探查/扫描攻击	4107	0.83
R2L	远程机器非授权访问	1126	0.23
u2r	非授权获取本地超级用户权限	52	0.01

参与比较的数据流学习算法包括 1 种滑动窗口算法 Win、4 种已有的集成式数据流挖掘算法（SEA[3]，AWE[4]，ACE[5] 和 MAE[6-7]）以及本文提出的不均衡样本的数据流学习算法 UMAE。当每个数据块到达时，各算法先对该数据块进行预测，获得该数据块的预测精度，然后对该数据块进行在线学习。这些算法的在线学习处理方法如下：

1）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。

2）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。

3）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。

4）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 中的基分类器直接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并生成新的基分类器后，重新更新权重，再采用加权投票进行集成。

5）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，ES 中的基分类器与 MS 相同。预测时对 ES 中的基分类器的预测结果采用大多数投票法进行集成。

6）UMAE 算法在 MAE 算法的基础上利用类别滑动窗口产生基分类器的训练样本集，是一种类别不均衡问题的在线学习方法。

C4.5 决策树[9] 不仅预测精度高，而且学习速度快，实验中 6 种算法的基分类器学习均采用该学习模型。实验中所有算法的数据流以数据块的方式到达，数据块大小均设为 500 个样本。预测时参与集成的基分类器最大数目 k 设为 10，MAE 和 UMAE 算法中基分类器库的大小（记忆容量）m 设为 k 的 5 倍，其值设为 50。UMAE 算法中滑动窗口取最小值，即其大小设为：

$$s=\left\lfloor \frac{|DB|}{r} \right\rfloor \tag{5}$$

如果新的数据块 DB 内某个类别的样本数目大于 s，则对该类别而言，仅保留最后进入其滑动窗口内的 s 个样本。

实验选择 MDSQ 算法[10] 作为 MAE 和 UMAE 算法的回忆机制，因为实验表明该算法是选择性集成算法中速度快、精度高的算法[11]。本实验中涉及的数据流挖掘算法和 MDSQ 选择性集成算法均采用 C++语言实现，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[12-13]，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。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 实验结果与分析

5.1 预测精度结果分析

表 2 列出了 6 种算法对每个类别的预测精度结果，最后一行为各算法在整个数据集上的预测精度，最优结果用黑体表示。表 2 中的预测精度结果是数据集中所有数据块预测精度的均值，即在实验中数据流 DS 以数据块的形式不断到达，算法在运行过程中采用先预测再学习的策略，即先利用当前的学习结果对数据块进行预测，获得该数块的预测精度，然后再对该数据块进行在线学习。对于第一个数据块，由于还没有学习结果，因此全部预测成正常样本，即类别 Normal。所以表 2 中的预测精度结果 Acc 为：

$$Acc=\frac{1}{|DS|}\sum_{DB_i\in DS}Acc_i \tag{6}$$

其中，$|DS|$ 表示数据流 DS 中的数据块个数，DB_i 为数据流 DS 中的第 i 个数据块，Acc_i 表示相应算法对数据块 DB_i 的分类精度。

表 2 数据块平均预测精度结果/%

类别信息		Win	SEA	AWE	ACE	MAE	UMAE
类别	样本数						
Normal	97278	95.74	88.34	93.48	91.09	**97.82**	95.82
DOS	391458	98.81	97.13	98.16	96.49	**98.90**	98.74
Probe	4107	33.11	4.31	14.19	2.95	19.65	**56.71**
R2L	1126	19.45	5.06	4.80	0.00	4.88	**38.37**
u2r	52	1.92	0.00	0.00	0.00	0.00	**76.92**
总精度	494021	97.37	94.32	96.22	94.33	**97.70**	97.58

从表 2 给出的数据流的总体精度来看，MAE 算法的整体精度最高，然后依次是 UMAE，Win，AWE，ACE 和 SEA。根据表 2 中每个类别的预测精度结果可以看出，传统的数据流学习结果对每个类别的预测能力与该类别的样本数目密切相关。对于样本数目所占比例很小的类别，其分类精度很低；对样本数目所占比例很大的类别，分类精度很高。Kddcup99 数据集中每个类别的样本数目非常不均衡，导致每个类别的分类精度差别很大。此外，由于攻击具有时间相关性，导致传统的集成式预测方法（如 SEA，AWE，ACE）的预测精度低于简单的 Win 算法。我们前期提出的 MAE 算法由于引入了人类的记忆与遗忘机制，总精度优于 Win 算法，但是其优势主要体现在 Normal 和 DOS 两种样本占多数的类别上，对 Probe，R2L，u2r 3 种小类别的预测能力也比 Win 算法低。

本文提出的 UMAE 算法为每个类别设置了一个滑动窗口，这在一定程度上缓解了样本不均衡造成的问题。从表 2 的结果可知，UMAE 算法的总精度虽比 MAE 算法略低，但是对 Probe，R2L，u2r 这 3 种小类别的预测精度却遥遥领先于其他算法。在实际应用中，常常需要对异常样本进行预测，异常样本的数目往往比正常样本少很多，但是对其进行正确预测则至关重要，是应用的关键。传统的数据流挖掘算法对这类应用的预测效果无法满足应用需求，本文提出的 UMAE 算法有效改善了对异常样本的预测能力。

相对 MAE 算法而言，UMAE 算法的总体精度虽然没有

很大的改善，但是从每个类别的精度结果来看，对于样本数目很少的 Probe，R2L 和 u2r 类别，其预测精度得到了大幅度的提高，说明 UMAE 算法对 Probe，R2L 和 u2r 这些恶意攻击的检测率大幅度提高。UMAE 是 MAE 算法的改进，若各类别的数据量越均衡，则利用多滑动窗口构造的训练集越接近于原始数据块，因此 UMAE 算法的性能也越趋近于 MAE；但是当面对类别严重不均衡的异常检测类问题时，UMAE 算法的实用性明显强于 MAE 及其他数据流挖掘算法。

5.2　训练时间结果分析

在实验中，所有的算法都采用 C++语言实现，计算效率明显优于基于 JAVA 语言的算法。表 3 列出了 6 种算法的训练时间结果，其值为训练每个数据块所需的时间均值。实验中，训练时间结果 TL 用如下方式计算：

$$TL=\frac{1}{|DS|}\sum_{DB_i\in DS}TL_i \tag{7}$$

其中，TL_i 为数据块 DB_i 到达后进行在线学习所需的时间。

表 3　平均训练时间结果/10^{-3}s

Win	SEA	AWE	ACE	MAE	UMAE
0.45	7.22	4.96	16.41	10.53	14.51

从实验结果可以看出，Win 算法只是简单地训练一个基分类器，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 算法长；SEA 算法需要对基分类器进行评估以确定删除哪个基分类器，训练时间比 AWE 算法更长；ACE 算法中引入概念漂移检测的功能，较为耗时，因此其训练时间最长。MAE 算法的训练时间介于 SEA 与 ACE 之间。UMAE 算法在 MAE 算法的基础上增加了对各类别滑动窗口的操作，学习时间比 MAE 算法略长，但是其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。

5.3　预测时间结果分析

表 4 列出了所有算法对单个数据块的平均预测时间结果。实验中，预测时间的结果 TT 用如下方式计算：

$$TT=\frac{1}{|DS|}\sum_{DB_i\in DS}TT_i \tag{8}$$

其中，TT_i 为预测数据块 DB_i 所需的时间。

表 4　平均预测时间结果/10^{-6}s

Win	SEA	AWE	ACE	MAE	UMAE
0.50	3.55	3.50	19.50	3.55	3.50

由于每个数据块的大小相同，预测时间的结果主要取决于参与集成预测的基分类器的类型和数目。从表 4 的结果可以看出，Win 算法只有一个基分类器参与预测，其预测时间最快；SEA，AWE，MAE 和 UMAE 算法参与集成分类的基分类器数目基本相同，因此预测时间也相当；ACE 算法引入漂移检测，其产生的基分类器结构更为复杂，预测所需的时间最长。

结束语　本文针对各类别分布严重不均衡的应用，在 MAE 算法的基础上提出类别不均衡数据流在线学习算法 UMAE，该算法为每个类别设置了一个样本滑动窗口，每到达一个新的数据块，UMAE 使该数据块内的样本按照自身的类别分别进入相应的滑动窗口，最后利用各类别滑动窗口内的样本构建新基分类器的训练数据块，并利用该数据块进行数据流的在线学习。与典型的数据流挖掘算法的比较结果表明，UMAE 算法不仅整体分类精度高，而且能够大幅度提高对样本数目少的小类别的预测精度。此外，UMAE 算法的训练和预测时间能够满足在线学习的实时性需求。因此，相对于已有的数据流在线学习方法，UMAE 算法对于异常检测类的不均衡应用具有更好的实用性。

参 考 文 献

[1] SAYED-MOUCHAWEH M, LUGHOFER E. Learning in Non-Stationary Environments: Methods and Applications [M]. New York: Springer, 2012.

[2] GAMA J. Knowledge Discovery from Data Streams(1st ed) [M]. London, U. K.: Chapman & Hall, 2010.

[3] STREET W N, KIM Y. A streaming ensemble algorithm (SEA) for large-scale classification [C]// Proc. KDD '01 ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining. 2001: 377-382.

[4] WANG H, FAN W, YU P S, et al. Mining concept-drifting data streams using ensemble classifiers [C]// Proc. KDD'03 ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining. 2003: 226-235.

[5] NISHIDA K, YAMAUCHI K, OMORI T. ACE: Adaptive classifiers-ensemble system for concept-drifting environments[C]// Proc. 6th Int. Workshop Multiple Classifier Syst.. 2005: 176-185.

[6] ZHAO Q L, JIANG Y F, LU Y T. Ensemble model and algorithm with recalling and forgetting mechanisms for data stream mining[J]. Journal of Software, 2015, 26(10): 2567-2580. (in Chinese)
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[J]. 软件学报，2015，26(10)：2567-2580.

[7] JIANG Y H, ZHAO Q L, LU Y T. Adaptive Ensemble with Human Memorizing Characteristics for Data Stream Mining [J]. Mathematical Problems in Engineering, 2015, 2015: 1-10.

[8] UCI Machine Learning Repository[OL]. http://archive.ics.uci.edu/ml.

[9] QUINLAN J R. C4.5: Programs for Machine Learning[M]. USA: Morgan Kaufmann Publishers, 1993.

[10] MARTINEZ-MUNOZ G, HERNÁNDEZ-LOBATO D, SUAREZ A. An analysis of ensemble pruning techniques based on ordered aggregation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(2): 245-259.

[11] ZHAO Q L, JIANG Y H, XU M. Catergorization and Comparision of the Eusemble Pruning Algorithm[J]. Computer Engineering and Science, 202, 34(2): 134-138. (in Chinese)
赵强利，蒋艳凰，徐明. 选择性集成算法分类与比较[J]. 计算机工程与科学，2012，34(2)：134-138.

[12] ZHAO Q, JIANG Y. LibEDM: a platform for ensemble based data mining [C]// Proceedings of the IEEE International Conference on Data Mining Workshop (ICDMW'14). Shenzhen, 2014: 1250-1253.

[13] Library for Ensemble based Data Mining [OL]. https://github.com/Qiangli-Zhao/LibEDM.