计算机科学
COMPUTER SCIENCE

标签稀疏场景下任意数据流在线学习方法

张帅, 周鹏, 张燕平

引用本文

张帅, 周鹏, 张燕平. 标签稀疏场景下任意数据流在线学习方法[J]. 计算机科学, 2025, 52(6): 139-150.

ZHANG Shuai, ZHOU Peng, ZHANG Yanping. Online Capricious Data Stream Learning with Sparse Labels [J]. Computer Science, 2025, 52(6): 139-150.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240300155

标签稀疏场景下任意数据流在线学习方法

张 帅 周 鹏 张燕平
安徽大学计算机科学与技术学院 合肥 230601
(zs0920ahu@163.com)

摘 要 随着数据体量的剧增,机器学习方法已逐渐由传统的静态学习模式转向面向流式数据的在线学习模式。任意数据流是指数据实例随着时间以流的方式逐个到达的同时,其特征空间可能会发生任意变化,即旧的特征可能随时消失,新的特征也可能随时出现。例如,在环境检测领域,新增传感器或旧传感器突然异常会使得数据流的特征空间发生任意变化。此外,现有面向数据流的在线学习方法大多假设可以获取所有数据实例的真实标签。然而,在真实应用中,由于人工标注数据的代价高昂,数据标签大多是稀疏的。为了解决标签稀疏场景下任意数据流的在线学习问题,提出一种基于被动-主动学习的在线学习算法 PAACDS(Passive Aggressive Active Learning for Capricious Data Streams)以及它的变体 PAACDS-I。首先,利用在线主动学习方法选择有价值的数据实例,使得可以在最小的监督下建立优越的预测模型。随后,在获得所选择数据实例的查询标签后,结合在线被动-主动更新规则和边界最大化原则来更新基于任意数据流中共享和新增特征空间的动态分类器。最后,将所提算法与现有的最先进方法在 12 个数据集上进行了比较,大量的实验对比和分析验证了所提算法在任意数据流标签稀疏场景下的有效性。

关键词:在线学习;任意数据流;动态特征空间;主动学习;稀疏标签
中图分类号 TP391

Online Capricious Data Stream Learning with Sparse Labels

ZHANG Shuai,ZHOU Peng and ZHANG Yanping
School of Computer Science and Technology,Anhui University,Hefei 230601,China

Abstract With the dramatic increase in data volume,machine learning methods have gradually transitioned from traditional static learning to online learning modes that are designed for streaming data. Capricious data streams refer to data instances arriving over time in a sequential manner,where the feature space can potentially undergo capricious changes. It means that old features may disappear at any time,while new features may emerge. For example,in the field of environmental monitoring,the addition of new sensors or sudden anomalies in existing sensors can cause arbitrary changes in the feature space of the data stream. Furthermore,existing online learning methods for data streams often assume access to the true labels of all data instances. However,in real-world applications,data labeling is often sparse due to the high cost of manual annotation. Therefore,to address the problem of online learning in capricious data streams with sparse labels,a passive-active learning-based online learning algorithm called PAACDS(Passive Aggressive Active Learning for Capricious Data Streams),along with its variant PAACDS-I,is proposed. Firstly,an online active learning method is utilized to select valuable data instances,allowing the construction of superior prediction models with minimal supervision. Subsequently,after obtaining the queried labels for the selected data instances,the dynamic classifier,which encompasses the shared and newly added feature spaces in the capricious data streams,is updated using online passive-active update rules and the principle of boundary maximization. Finally,the proposed algorithm is compared to existing state-of-the-art methods on twelve datasets. Extensive experimental comparisons and analyses validate the effectiveness of the proposed algorithm in scenarios involving capricious data streams and sparse labels.

Keywords Online learning,Capricious data streams,Dynamic feature space,Active learning,Sparse label

到稿日期:2024-03-25 返修日期:2024-09-11
基金项目:国家自然科学基金面上项目(62376001);安徽省自然科学基金面上项目(2308085MF215)
This work was supported by the National Natural Science Foundation of China(62376001) and Natural Science Foundation of Anhui Province,China(2308085MF215).
通信作者:周鹏(doodzhou@ahu.edu.cn)

1 引言

在线学习是一种机器学习方法，适用于按顺序到达的数据。学习器旨在每一步中学习和更新未来数据的最佳预测模型。在线学习中，学习者对问题进行预测，然后揭示真实答案，根据学习者的预测与实际答案之间的差异产生损失。学习者的主要目标是通过最小化累积损失来提高预测准确性[1]。随着数据量呈指数增长，处理流式数据的在线学习方法得到了广泛应用[2-4]。现有面向数据流的在线学习方法大多假设每个数据流具有固定的特征空间，即随着数据实例的持续到达，特征空间并不发生变化。然而，在实际应用中，由于各种原因，特征空间可能会出现丢失或新增的情况。例如，在生态系统检测[5]、实时网络入侵[6]和环境数据监测[7]等领域，当传感器损坏或网络异常时，特征空间会随着数据流实例而出现突然丢失的情况。同时，当新增传感器生效时，数据流实例的特征空间又会出现突然新增的情况。

任意数据流[8]是指数据实例随着时间的推移连续生成和到达的同时，特征空间发生任意的变动，例如特征的消失和新增。如图 1 所示，假设时间戳t_1和t_2处的数据流实例分别为x_1和x_2。相比数据样本x_1，x_2出现了特征空间的缺失，在图中用空白部分表示。x_2中绿色部分表示相比x_1新增的特征空间。灰色部分表示与先前时刻一致的共享特征空间。此外，在标签稀疏场景下，并不能总是获得所有样本的真实标签信息。对此，在学习过程中，可以有选择地查询其中某些样本的真实标签。图 1 中，Q 表示查询标签(Query the Label)，NQ 表示不查询标签(Not Query the Label)。

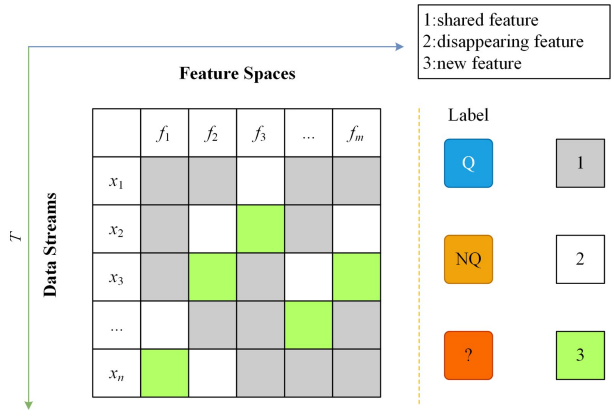


图 1 标签稀疏场景下任意数据流(电子版为彩图)

Fig. 1 Capricious data streams with sparse labels

从任意变化的数据流中进行在线学习的核心挑战是如何设计并实现高效的动态分类器。传统的在线学习算法不仅无法处理特征空间任意变化的数据流，而且对标签稀疏场景下的在线学习也难以胜任，因为它们通常假设特征空间不变，且所有的数据实例真实标签都是可以获取的。面向任意数据流的在线学习算法，学习者需要不断调整模型来适应新出现的特征。例如，在生态系统检测中，当新的环境指标被引入时，学习者需要适应这些新的特征，以提高检测准确性。类似地，在实时网络入侵检测中，当新的攻击类型出现时，学习者需要学习新的特征模式以及如何与传统特征进行关联。因此，动态特征空间且标签稀疏场景下的任意数据流在线学习给传统在线学习算法带来了巨大挑战，亟需开发新的方法来进行有效的处理。

任意数据流具有 3 个关键特征[9]。1)特征空间的缺失是任意和不规则的，先前出现的特征可能随着实例数量的增加而消失。例如，在医疗领域中，患者个人资料可能来自不同的医疗设备和提供者[10]。由于受到个体的医疗特征和设备故障等因素的影响，收集到的特征往往是不完整的，可能会任意地缺失。2)任意数据流中的类别分布通常是动态不平衡的。例如，在异常检测系统中，异常样本只占很小比例，而大多数样本是正常的[11]。这种类别分布的不平衡性会对在线学习算法产生影响，因为模型需要有效地适应不同类别之间的分布变化，以准确地识别和处理异常情况。3)任意数据流经常出现概念漂移，即由于数据流的任意变化而导致的突然或渐进性变化[12-13]。概念漂移可能是由外部环境的变化、数据源的变更或系统故障引起的。在线学习算法需要能够实时检测和适应概念漂移，以保持模型的准确性和鲁棒性。为了解决上述第一个问题，即特征空间任意的缺失或新增问题，本文提出了一种新的标签稀疏场景下任意数据流在线学习方法。

本文的主要贡献如下：

1)提出了一种新的在线学习方法，用于处理具有稀疏标签的任意数据流，在这种数据流中，数据流的体积和特征维度同时增加，并且需要主动地获取数据流的标签；

2)将主动查询策略和被动-主动更新策略相结合的思想扩展到处理具有稀疏标签任意数据流的二分类任务；

3)基于 12 个合成数据集的大量实验，验证了本文算法的有效性。

本文第 2 章介绍了相关工作；第 3 章对 PAACDS 算法进行了详细介绍；第 4 章在 12 个数据集上进行实验并对结果加以分析；最后总结全文并展望未来。

2 相关工作

2.1 静态特征空间的在线学习方法

为了应对数据流面临的挑战，目前对数据流的处理一般采用在线学习的方法。多年来，研究人员和学者们开发了各种算法用于在线学习。目前，在线学习主要有两种方式，分别是一阶算法和二阶算法。

一阶算法利用一阶导数的信息来更新模型参数。两种常用的一阶算法是感知器算法[14]和在线梯度下降[15](Online Gradient Descent，OGD)算法。这些算法在许多学习任务中得到了广泛应用，并且已被证明在不同领域中是有效的。感知器算法是一种经典的在线学习算法，特别适用于二分类问题。它根据输入样本的特征和标签之间的差异，通过迭代地更新模型参数，以逐步调整模型的决策边界。感知器算法简单而高效，在实际应用中被广泛使用，例如文本分类、图像识别和垃圾邮件过滤等。OGD 算法是一种常见的在线学习优化算法，其基本思想是通过迭代地沿着负梯度方向更新模型参数，来最小化损失函数。在线梯度下降算法具有较好的收敛性和灵活性，适用于大规模数据集和高维特征空间，在推荐系统、自然语言处理和图像处理等领域中得到了广泛应用。

二阶在线学习算法利用二阶导数信息更好地探索特征之间的底层结构，提高了收敛性和优化速度。这些算法，如正态采样法[16]、置信加权学习和软置信加权算法[17]等，已被证明在捕捉复杂关系和利用数据的底层结构方面非常有效。其中，正态采样算法是一种基于高斯采样的第二阶在线学习算法。它利用高斯分布对权重向量进行建模，并通过在线学习过程中的高斯采样来逼近真实的权重分布。软置信加权算法是一种软置信加权算法，用于在线学习。它通过引入软边界来解决置信度加权算法中的硬边界问题，并通过最小化损失函数来更新模型参数。软置信加权算法在处理噪声和概念漂移时具有较好的鲁棒性，并能够自适应地调整置信区间的大小。Chen 等[18]提出了一种基于二阶投影二次平均的在线学习方法，以有效处理高吞吐量的数据流。通过充分利用正则化的双平均优化、二阶信息和最优投影算子，该方法能够快速收敛并得到相对优化的解决方案。

2.2　动态特征空间的在线学习方法

在动态特征空间中进行在线学习是一项具有挑战性的任务。动态特征空间指数据流的特征空间在时间上不断变化，这种变化可能是由于新特征的引入、旧特征的变化或特征的删除导致的。这种变化使得在线学习算法需要灵活适应新的特征，并对旧特征的变化进行适应。在动态特征空间中进行在线学习面临着多重挑战。

Zhang 等[19]首次提出一种用于处理梯形数据流的在线学习算法(Online Learning with Streaming Features，OLSF)。它将当前训练实例的特征分为历史特征和新特征，并通过不同的更新规则来更新这两部分特征。通过这种方式，OLSF 能够适应特征空间的增量变化。它在历史特征和新特征上分别学习分类器，并将它们整合在一起进行最终的预测。本文研究的问题不仅是特征空间任意变动，而且包含标签稀疏场景下的数据流。相比于 OLSF，本文算法的应用场景更加广泛和接近实际的应用场景。Gu 等[20]提出了一种新的在线学习算法(Learning with Incremental Instances and Features，IIF)，用于从梯形数据流中学习分类模型。该算法采用了高度动态的模型更新策略，可以从不断增加的训练数据和扩展的特征空间中进行学习。首先，将每轮获取的数据流进行划分，并为这些不同的划分部分构建相应的分类器。然后，为了实现各个分类器之间的有效信息交互，利用单一的全局损失函数来捕捉它们之间的关系。最后，采用集成的思想来得到最终的分类模型。Yu 等[21]提出了一种用于解决多源数据流分类中概念漂移问题的方法。在多源数据流分类领域，现有的方法往往忽略了不同数据流之间的动态关系，导致了负面传递问题。为了解决这个问题，该算法提出了一种双阶段的方法：在初始化阶段，利用自适应协变量漂移适应算法构建了一个初始化的集成模型，减轻了协变量漂移并学习了动态相关性；在在线处理阶段，采用高斯混合模型的加权策略来处理异步漂移。总体而言，该方法通过自适应学习动态相关性和协变量漂移适应，显著提升了多源数据流分类问题的解决效果。此外，Beyazit 等[22]提出了一种处理具有变化特征空间的数据流的算法(Online Learning from Varying Features，OLVF)。OLVF 通过将实例分类器和训练实例动态投影到共享的特征子空间中，同时学习特征空间的分类器，实现对具有变化特征空间的数据的分类。其根据投影置信度来更新实例分类器和特征空间分类器，并应用特征稀疏性方法来降低模型复杂度。虽然 OLVF 算法能够在单次遍历中学习来自不同特征空间的数据，并提升学习性能和模型的泛化能力，但是它的应用场景是有监督的，而本文提出的算法是应用于标签稀疏场景下，两者应用场景不同，并且真实世界中，数据标签往往无法全部获取，所以本文算法适应性更强。He 等[23]提出了一种针对任意数据流的生成学习算法(Generative Learning with Streaming Capricious，GLSC)来处理具有变化特征空间的数据流，其中每个到达的数据实例可以任意携带新特征或停止携带部分旧特征。具体而言，GLSC 在一个通用特征空间上训练学习器，建立旧特征和新特征之间的关系，以便在新特征空间中利用在旧特征空间中学到的模式。GLSC 与本文研究的问题最相关，都是研究任意数据流，但 GLSC 采用的是生成图模型，需要重构新旧特征空间的关系，每次计算需要耗费很大的开销，所以整体时间开销很大，而本文采用的方法整体开销较小。

在现实世界的许多应用中，由于需要人工标注数据，获取真实标签的成本往往很高，因此无法获得所有数据实例的真实标签。这类问题被称为具有稀疏标签的在线学习。为了解决这个问题，很多算法被提出。例如，Gu 等[24]提出了一种用于解决具有标签稀缺性的增量特征空间学习问题的算法(Feature Spaces Learning with Label Scarcity，FLLS)，它利用主动学习策略选择有价值的实例进行标注，以最小的监督构建优越的预测模型。Liu 等[25]提出了一种在线主动学习算法，将主动查询策略和被动侵略式(Passive Aggressive，PA)更新策略结合起来，用于处理梯形数据流上的二分类和多分类在线分类任务。Gu 等和 Liu 等提出的算法主要是解决梯形数据流中的标签稀疏问题，其对特征空间变化进行了有规则变化的假设，而本文算法不需要对特征空间变化进行假设，即特征空间可以任意地变化。此外，Cheng 等[26]提出了一种用于数据流分类的主动广义学习方法，采用了多目标进化优化的策略。具体而言，将新到达的实例存储在数据块中，通过快速的局部漂移检测来识别潜在的漂移。然后，采用多目标进化算法选择最有价值的候选实例进行标记，其中代表性实例的数量由相邻数据块的稳定性确定。总体而言，该方法提供了一种有效应对数据流中概念漂移的方案，提高了分类性能并降低了成本。Gu 等[27]提出了一种处理在线学习中增量特征空间和强化学习反馈的新算法及其两个变体。该算法利用探索-开发策略进行标签猜测，并提出了考虑强化学习反馈和猜测标签的新损失函数。同时，通过采用被动攻击规则和结构风险最小化原则更新特征，设计了一个适应动态特征空间和数据量增长的多类分类模型。Din 等[28]提出了一种可靠的自适应基于原型的学习方法，用于处理具有有限标签的不断演化数据流。该算法提出了一种自适应的半监督学习框架，采用基于原型的数据表示方法，动态更新原型的重要性，并快速适应局部概念漂移。该方法还利用自适应原型来检测概念演化，并通过引入主动学习方法，减少手动标记的需求。以上 3 种算法虽然都可以有效地处理稀疏标签问题，但是无法

解决任意数据问题，即特征空间和数据实例同时变化的场景。

最近，动态特征空间的在线学习吸引了很多学者的关注，许多理论方法和成功的现实应用得到了广泛研究和开发。然而，它们要么假定特征空间是有规则的变化，要么假定数据实例具有完整的标签，但在现实应用中，这些特殊情况是非常少的。数据实例的特征空间可能发生任意的变化，数据实例的真实标签也可能并不完整，即只有少部分的真实标签。为了解决这些问题，本文提出了一种新的标签稀疏场景下任意数据流在线学习。

3 本文方法

本文的核心思想是在第 t 次迭代中，首先通过当前分类器 w_t 计算其预测边界 q_t。$Z_t \in \{0,1\}$ 是伯努利随机变量，其决定是否查询该数据实例的标签，如果伯努利随机变量 Z_t 的概率是 1，则查询该数据实例 x_t 的真实标签 y_t。同时，分类器根据数据实例 x_t 返回其预测值 $\hat{y}_t$。然后，根据已经查询到的数据实例 x_t 的真实标签 y_t 和预测值 $\hat{y}_t$，分类器根据损失函数计算它们之间的损失，其反映了预测值和真实值之间的差异。如果伯努利随机变量 Z_t 的概率不是 1，则不查询该数据实例 x_t 的真实标签 y_t。最后，根据 B（选择特征的比例）截断 w_t，使模型具有稀疏性。算法流程如图 2 所示。

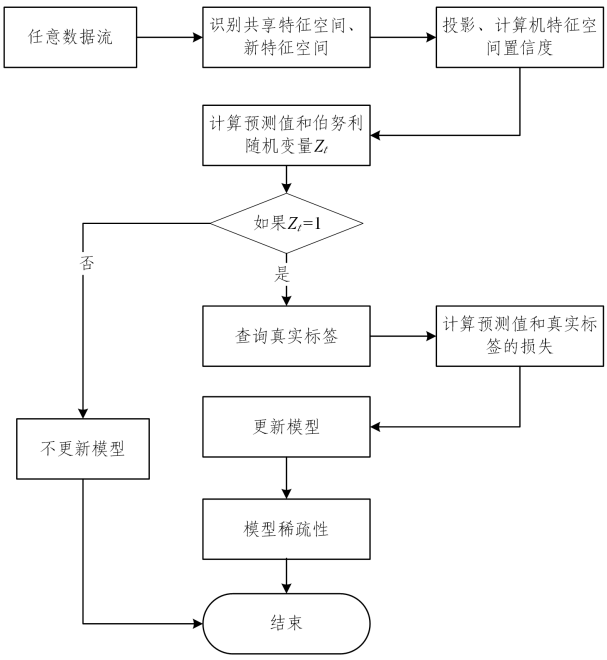


图 2 算法总体流程图

Fig. 2 Overall flowchart of algorithm

3.1 基于不确定性的特征自适应加权策略

为了增强学习过程并优化模型，使用一种基于不确定性的自适应加权策略来动态学习任意数据流。这个策略的原理是，不确定性较高的特征往往包含更有价值的信息，有助于模型的优化。本文利用迭代过程中特征的方差来评估它们的不确定性。方差较大的特征表示不确定性较高，可以为优化模型提供更多的信息。值得注意的是，方差仅取决于实例本身，并不受外部因素的影响，因此将方差作为不确定性的代理，用于评估特征的信息量。

具体而言，首先将训练实例投影到特征空间，然后计算特征信息的累积平均值作为特征空间的置信水平。该置信水平在预测过程中有着重要作用，它代表了特征空间在预测结果中所占的权重大小。在第 t 次迭代中，用 h_t^i 表示实例 x_t 中与第 i 个特征相关联的信息量。因此，共享特征空间的置信水平 p_t^s 和新特征空间的置信水平 p_t^n 可以定义为：

$$p_t^s = \sum_{i=1}^{d_t^s} \frac{h_t^i}{\sum_{j=1}^{d_t^s} h_t^j} \tag{1}$$

$$p_t^n = \sum_{i=1}^{d_t^n} \frac{h_t^i}{\sum_{j=1}^{d_t^n} h_t^j} \tag{2}$$

其中，d_t^s 表示共享特征空间的维度，d_t^n 表示新特征空间的维度。

3.2 基于边界的主动学习策略

在第 t 轮中，当数据实例 x_t 到达时，首先通过当前分类器 w_t 计算其预测边界。由于数据流的特征空间不断变化，因此 x_t 和 w_t 的维度可能不同。故重新定义预测边界：

$$q_t = p_t^s w_t^s x_t^s + p_t^n w_t^n x_t^n \tag{3}$$

其中，q_t 为预测置信度。是否查询数据实例 x_t 的真实标签由 $Z_t \in \{0,1\}$ 决定，其中 Z_t 是伯努利随机变量，$Z_t = 1$ 的概率为 $\delta/(\delta + |q_t|)$，$\delta \geqslant 1$ 是平滑参数。

上述策略的思想受到了基于边界的主动学习的启发，根据数据实例的重要性来决定是否查询其标签。这个重要性由当前分类器的预测边界 q_t 确定。然而，在在线主动学习中，可能面临当前分类器不可靠的问题，特别是在最初几个迭代时，由于数据实例是以在线的方式到达的，因此没有足够的数据实例来训练一个可靠的分类器。如果直接基于 q_t 决定是否查询标签，可能会做出错误的决策。为了解决上述问题，类似于传统方法[29]中的策略，在模型中引入了伯努利随机变量 $Z_t \in \{0,1\}$，它可以通过一定的概率来摇动分类器模型的决策，即使当前分类器做出了错误的预测，由于随机性，分类器模型也仍然可能做出正确的选择。此外，在许多现实世界的应用中，如个性化推荐、医学诊断和恶意 URL 检测等，使用伯努利随机变量已被证明是有效和可靠的。

Z_t 的值存在两种情况：1）如果 $Z_t = 0$，PAACDS 算法将不会向预测模型询问数据实例的真实标签，因此模型保持不变；2）如果 $Z_t = 1$，数据实例 x_t 的标签将由一个数据库揭示，PAACDS 将遭受瞬时的预测损失 $l_t(w_t, (x_t, y_t))$，其中 $y_t \in \{-1, +1\}$。它能够利用新获得的实例标签对 (x_t, y_t) 的潜在信息来更新分类模型。

为了适应任意变化数据流的动态特性，本文修改了 Hinge 损失函数来训练分类器 w_t。因此，第 t 次迭代中分类器的预测损失 $l_t(w_t, (x_t, y_t))$ 定义为：

$$l_t = \max\{0, 1 - y_t(p_t^s w_t^s x_t^s + p_t^n w_t^n x_t^n)\} \tag{4}$$

3.3 基于软间隔推导的更新规则

为了更新权重 w_t 与 (x_t, y_t)，采用了在线被动-主动学习的思想。具体而言，在第 t 轮，新更新的分类器 $w_{t+1} \in R^{d_t}$ 将由两部分组成，可以写为 $w_{t+1} = [w_{t+1}^s, w_{t+1}^n]$，其中 $w_{t+1}^s \in$

R^{d_t-1}用于更新共享特征并继承w_t的信息，$w_{t+1}^n \in R^{d_t-d_{t-1}}$用于更新新的特征。因此，$w_t$的更新可以通过优化式(5)转化为$w_{t+1}^s$和$w_{t+1}^n$的更新问题：

$$\begin{cases} w_{t+1} = \arg\min\limits_{w=[w^s,w^n],} \dfrac{1}{2} \| w^s - w_t^s \|^2 + \dfrac{1}{2} \| w^n \|^2 \\ \text{s.t. } l(\boldsymbol{w},(x_t,y_t)) = 0 \end{cases} \tag{5}$$

其中，$\boldsymbol{w} \in R^{d_t}$；$l(\boldsymbol{w},(x_t,y_t))$是$\boldsymbol{w}$在数据实例$x_t$上的预测损失，表示为$l(\boldsymbol{w},(x_t,y_t)) = \max\{0,1-y_t(p_t^s w_t^s x_t^s + p_t^n w_t^n x_t^n)\}$。这种更新是通过构建和调整每个数据实例的权重来最小化累积损失。

然而，这种方法对噪声非常敏感，可能导致过拟合。为了解决过拟合问题，本文引入软间隔技术。软间隔技术的核心思想是允许一定程度上的分类错误，以换取更好的泛化能力和鲁棒性。引入松弛变量(Slack Variables)，用于描述数据点可以位于错误的一侧或在超平面附近的情况。通过最小化目标函数，软间隔支持向量机在保持尽可能大的间隔的同时，允许一定数量的分类错误。因此，本文引入了一个松弛变量ξ，$\xi \in [0,1]$，允许一定程度的误分类。许多在线学习算法结合了上述约束，以实现最大化准确性和最小化损失之间的平衡。它们将权重的学习形式化为一个优化问题，目标是找到最优权重，这样既能最小化损失，又考虑到了软间隔约束。因此，本文的学习任务可以表述为：

$$\begin{aligned} & w_{t+1} = \arg\min_{w=[w^s,w^n],} \frac{1}{2} \| w^s - w_t^s \|^2 + \frac{1}{2} \| w^n \|^2 + C\xi \\ & \text{s.t. } l(\boldsymbol{w},(x_t,y_t)) \leqslant \xi, \xi \geqslant 0 \end{aligned} \tag{6}$$

为了解决上述不等式的优化问题，本文通过使用带有KKT(Karush-Kuhn-Tucker)条件的 Lagrangian 函数来解决上述优化问题，并得到以下更新规则：

$$\begin{cases} w^s = w_t^s + \tau p_t^s y_t x_t^s \\ w^n = \tau p_t^n y_t x_t^n \\ \eta = C - \tau \end{cases} \tag{7}$$

根据这些条件，可以进一步得到问题的表达式并求出τ：

$$L(\tau) = \frac{1}{2}\tau^2 (p_t^s)^2 \| x_t^s \|^2 + \frac{1}{2}\tau^2 (p_t^n)^2 \| x_t^n \|^2 + \tau(1 - y_t(p_t^s w_t^s x_t^s + p_t^n w_t^n x_t^n)) \tag{8}$$

$$\tau = \min\left\{ C, \frac{l_t}{(p_t^s)^2 \| x_t^s \|^2 + (p_t^n)^2 \| x_t^n \|^2} \right\} \tag{9}$$

式(6)中的目标函数与ξ呈线性比例关系，虽然可以解决过拟合问题，但它在数值稳定性上较差，并且对显著错误的惩罚较小，从而影响分类准确性。为了解决这个问题，本文使用松弛变量ξ的平方来改进优化问题的形式，其新的问题表述如下所示：

$$\begin{cases} w_{t+1} = \arg\min\limits_{w=[w^s,w^n],} \dfrac{1}{2} \| w^s - w_t^s \|^2 + \dfrac{1}{2} \| w^n \|^2 + C\xi^2 \\ \text{s.t. } l(\boldsymbol{w},(x_t,y_t)) \leqslant \xi, \xi \geqslant 0 \end{cases} \tag{10}$$

同样使用带有 KKT 条件的 Lagrangian 函数来解决上述不等式的优化问题，并得到以下更新规则：

$$\begin{cases} w^s = w_t^s + \tau p_t^s y_t x_t^s \\ w^n = \tau p_t^n y_t x_t^n \\ 2C\xi = \tau \end{cases} \tag{11}$$

将式(11)代入式(10)，可以得到以下公式：

$$L(\tau) = \frac{1}{2}\tau^2 (p_t^s)^2 \| x_t^s \|^2 + \frac{1}{2}\tau^2 (p_t^n)^2 \| x_t^n \|^2 + \tau(1 - y_t(p_t^s w_t^s x_t^s + p_t^n w_t^n x_t^n) - \frac{\tau}{4C}) \tag{12}$$

$$\tau = \min\left\{ C, \frac{l_t}{(p_t^s)^2 \| x_t^s \|^2 + (p_t^n)^2 \| x_t^n \|^2 + \dfrac{1}{2C}} \right\} \tag{13}$$

综上所述，可以得到以下更新规则：

$$\begin{aligned} w_{t+1} &= [w_{t+1}^s, w_{t+1}^n] \\ &= [w_{t+1}^s + \tau_t p_t^s w_t^s x_t^s, \tau_t p_t^n w_t^n x_t^n] \end{aligned} \tag{14}$$

其中：

$$\tau_t = \begin{cases} \min\left\{ C, \dfrac{l_t}{(p_t^s)^2 \| x_t^s \|^2 + (p_t^n)^2 \| x_t^n \|^2} \right\}, \\ \qquad \text{OLCDS} \\ \min\left\{ C, \dfrac{l_t}{(p_t^s)^2 \| x_t^s \|^2 + (p_t^n)^2 \| x_t^n \|^2 + \dfrac{1}{2C}} \right\}, \\ \qquad \text{OLCDS-I} \end{cases} \tag{15}$$

3.4　基于相对不确定性的模型稀疏策略

数据流的无限性和高维性，使得在分类器中保留所有特征面临巨大的挑战。例如，随着数据流持续到达，内存需求和计算开销逐渐增大，这可能会导致分类器性能下降。为了解决这个问题，有必要通过截断w_t来仅选择和保留最重要的特征用于分类器的学习。一种常用的方法是在将分类器投影到$L1$球上之后截断它。然而，这种截断策略可能会引入对数据流中不频繁特征的偏差。这些特征往往具有较小的权重，并且更容易被截断。此外，高度不确定性特征的权重即便发生微小变化，也可能会导致不同的结果。因此，在保留最重要的特征的同时，避免对仅出现在少数实例中的特征引入偏差是至关重要的。

本文在特征选择过程中引入了相对不确定性，以解决这个问题。当特征表现出较高的不确定性时，优先保留它们的权重，以捕捉其潜在的重要性。首先，计算特征的相对不确定性：通过测量特征的变化范围或其他不确定性指标，计算每个特征的相对不确定性。这可以帮助评估特征的重要性和稳定性。然后，进行特征权重的投影：根据相对不确定性，对特征权重进行投影。较不确定的特征权重将更有可能被保留，以捕捉其潜在的重要性。最后，特征选择和截断：基于投影后的特征权重进行特征选择和截断。可以设定一个阈值，只保留权重高于阈值的特征，以避免对仅出现在少数实例中的特征引入偏差。在实践中，可以通过以下步骤的投影过程实现：

$$w_t = \min\left\{ 1, \frac{\lambda}{\langle w_t \cdot \boldsymbol{H}_t \rangle} \right\} w_t \tag{16}$$

其中，$\lambda > 0$是正则化参数；$\boldsymbol{H}_t = [h_t^1, h_t^2, h_t^3, \cdots, h_t^p]$表示第$t$次迭代时公共特征空间的相对不确定性向量，它由所有观察到的特征的信息内容组成。在投影之后，基于参数$B > 0 (B \in (0,1))$对分类器进行截断，保留一部分最重要的特征权重。因此，在下一次预测过程中，只有被保留的最重要的权重参与模型的计算，所以模型的稀疏性被引入。

3.5　PAACDS 算法

基于上述分析和推导，针对具有稀疏标签的任意数据流

问题，提出了基于被动-主动的任意数据流算法 PAACDS 以及它的变体 PAACDS-I，如算法 1 所示。

具体而言，对于每个时间戳 t 到达的数据实例 x_t，算法按照以下步骤进行。在步骤 3－4 中，首先根据数据实例 x_t 的特征，确定共享特征空间和新特征空间；然后将权重向量 $\boldsymbol{w}$ 和 x_t 投影到这两个特征空间中。在步骤 5－10 中，计算共享特征空间和新特征空间的置信度值，利用这些置信度值，算法预测实例 x_t 的值 $\hat{y}_t$；然后计算伯努利随机变量 Z_t，如果 $Z_t=1$，则查询数据实例的真实标签 $y_t\in\{-1,+1\}$，并根据真实标签 y_t 和预测值 $\hat{y}_t$ 计算它们的损失。为了更新模型，步骤 11－12 概述了 PAACDS 和 PAACDS-I 算法的具体更新策略。这些策略根据观察到的实例及其相关损失自适应地调整模型。在步骤 13－14 中，使用参数 B 截断权重向量 w_t。这个截断过程促进了模型的稀疏性，便于特征选择，并提高了模型的效率。如果 $Z_t\neq1$，则不查询数据实例的真实标签，不更新模型。

总之，算法 1 提供了一个全面的框架，可以在在线学习中有效处理具有稀疏标签的任意数据流问题。通过输入的实例动态调整模型，该算法实现了准确的预测，并通过特征选择保持了模型的稀疏性。

算法 1 被动-主动任意数据流算法

输入：惩罚参数 $C>0$，正则化参数 $\lambda>0$，选择特征的比例 $B\in(0,1)$，平滑参数 $\delta\geqslant1$

输出：最新的分类器 w_t

初始化：$w_1=\{0,0,\cdots,0\}\in R^{d_1}$；

1. For t=1,2,…,T do
2. 接收一个新的训练实例：$x_t\in R^{d_t}$；
3. 共享特征空间和新的特征空间：$R^s=R^{x_{t-1}}\cap R^{x_t}$，$R^n=R^{x_t}-R^{x_{t-1}}$；
4. 将 w_t，x_t 分别投影到 R^s，R^n，w_t^s，w_t^n，x_t^s，x_t^n；
5. 计算特征空间置信度：p_t^s，p_t^n；
6. 计算预测值：$\hat{y}_t=q_t=\mathrm{sign}(p_t^s w_t^s x_t^s+p_t^n w_t^n x_t^n)$；
7. 计算伯努利随机变量 Z_t 的概率：$Z_t=\dfrac{\delta}{\delta+|q_t|}$；
8. IF $Z_t=1$ Then
9. 查询真实标签：$y_t\in\{-1,+1\}$；
10. 计算损失：$l_t=(y_t,\hat{y}_t)=\max\{0,1-y_t(p_t^s w_t^s x_t^s+p_t^n w_t^n x_t^n)\}$；

//Model Update：

11. $\tau_t=\begin{cases}\min\left\{C,\dfrac{l_t}{(p_t^s)^2\|x_t^s\|^2+(p_t^n)^2\|x_t^n\|^2}\right\}, & \text{PAACDS}\\ \min\left\{C,\dfrac{l_t}{(p_t^s)^2\|x_t^s\|^2+(p_t^n)^2\|x_t^n\|^2+\frac{1}{2C}}\right\}, & \text{PAACDS-I}\end{cases}$
12. $w_{t+1}=[w_{t+1}^s,w_{t+1}^n][w_{t+1}^s+\tau_t p_t^s w_t^s x_t^s,\tau_t p_t^n w_t^n x_t^n]$；

//Model Sparsity：

13. $w_t=\min\left\{1,\dfrac{\lambda}{\langle w_t\cdot H_t\rangle}\right\}w_t$；
14. 根据 B 截断 w_t；
15. Else
16. $w_{t+1}=w_t$；
17. END IF
18. END For

3.6 算法时间复杂度分析

算法 1 给出了 PAACDS 及其变体 PAACDS-I 的伪代码。关于时间复杂度，对于单次迭代，PAACDS 和 PAACDS-I 的时间复杂度都为 $O(|w_t|+|x_t|)$。这意味着两个算法的运行时间与权重向量 w_t 的大小和输入的数据实例 x_t 呈线性关系。

PAACDS 和 PAACDS-I 旨在根据观察到的数据实例高效处理和更新模型，确保即使在大规模数据集下，计算开销也可控。PAACDS 和 PAACDS-I 的线性运行时间复杂度使其能够进行有效的在线学习，适用于实时应用程序，其中及时的模型更新至关重要。

3.7 算法理论分析

首先，为了推导出 PAACDS 算法的错误界限，先引入以下符号：$\mathcal{F}=\{t\mid t\in[T],\hat{y}_t\neq y_t\}$，$\mathcal{M}=\{t\mid t\in[T],\hat{y}_t\neq y_t,l_t(w_t;(x_t,y_t))>0\}$，其中 $[T]$ 表示为 $\{1,2,\cdots,T\}$。

引理 1 假设 $(x_1,y_1),(x_2,y_2),\cdots,(x_T,y_T)$ 是一系列训练数据，其中 $x_t\in R^{d_t}$，$y_t\in\{-1,+1\}$ 对于所有的 t 成立。学习率 $\tau=\min\left\{C,\dfrac{l_t}{(p_t^s)^2\|x_t^s\|^2+(p_t^n)^2\|x_t^n\|^2}\right\}$，对于任意的 $\boldsymbol{w}\in R^u$，以下不等式成立：

$$\sum_{t=1}^{T}2Z_t\tau_t[M_t(\gamma-|q_t|)+F_t(\gamma+|q_t|)]\leqslant\gamma^2\|u\|^2+\sum_{t=1}^{T}\tau_t^2\|x_t\|^2+\sum_{t=1}^{T}2\gamma\tau_t l_t^*(u)\tag{17}$$

其中，$F_t=\mathbb{I}_{(t\in\mathcal{F})}$，$M_t=\mathbb{I}_{(t\in\mathcal{M})}$，$\mathbb{I}$ 是一个指示函数，$l_t^*(\boldsymbol{w})=\max(0,1-y_t(p_t^s w_t^s x_t^s+p_t^n w_t^n x_t^n))$。

基于引理 1，首先证明 PAACDS 算法在线性可分情况下的期望错误界限。假设存在一个分类器 $\boldsymbol{u}\in R^{d_T}$，对于所有 $t\in[T]$，满足 $y_t(p_t^s w_t^s x_t^s+p_t^n w_t^n x_t^n)\geqslant1$。

定理 1 假设 $(x_1,y_1),(x_2,y_2),\cdots,(x_T,y_T)$ 是一系列训练数据，$x_t\in R^{d_t}$，$y_t\in\{-1,+1\}$，$\|x_t\|^2\leqslant R$ 对于所有的 t 成立。存在一个向量 $\boldsymbol{u}\in R^{d_T}$，$l_t^*(u)=0$，使得对于所有的 t 成立。假设每个查询决策使用一个伯努利分布 $(\delta/(\delta+|q_t|))$，其中 $\delta>0$，那么 PAACDS 算法的期望错误数量受到以下的限制：

$$\mathbb{E}[\sum_{t=1}^{T}F_t]\leqslant\mathbb{E}[\sum_{t=1}^{T}F_t l_t(w_t)]\leqslant\frac{R^2}{4}(\delta+\frac{1}{\delta}+2)\|\boldsymbol{u}\|^2\tag{18}$$

设 $\delta=1$，可以得到以下上界：

$$[\sum_{t=1}^{T}F_t]\leqslant\mathbb{E}[\sum_{t=1}^{T}F_t l_t(w_t)]\leqslant R^2\|\boldsymbol{u}\|^2\tag{19}$$

证明：通过结合 $l_t^*(\boldsymbol{u})=0$ 和引理 1，可以得到：

$$\sum_{t=1}^{T}2Z_t\tau_t[M_t(\gamma-|q_t|)+F_t(\gamma+|q_t|)]\leqslant\gamma^2\|\boldsymbol{u}\|^2+\sum_{t=1}^{T}\tau_t^2\|x_t\|^2\tag{20}$$

不等式(20)可以进一步表达为：

$$\gamma^2\|u\|^2\geqslant\sum_{t=1}^{T}2Z_t\tau_t[M_t(\gamma-|q_t|)+F_t(\gamma+|q_t|)]=\sum_{t=1}^{T}2Z_t\tau_t[M_t(\gamma-|q_t|-\frac{\tau_t}{2}\|x_t\|^2)+F_t(\gamma+|q_t|-\frac{\tau_t}{2}\|x_t\|^2)]$$

$$=\sum_{t=1}^{T}2Z_t\tau_t[M_t(\gamma-|q_t|-\frac{l_t(w_t)}{2})+F_t(\gamma+|q_t|-\frac{l_t(w_t)}{2})]$$

$$=\sum_{t=1}^{T}2Z_t\tau_t[M_t(\gamma-|q_t|-\frac{1-q_t}{2})+F_t(\gamma+|q_t|-\frac{1-y_tq_t}{2})]$$

$$=\sum_{t=1}^{T}2Z_t\tau_t[M_t(\gamma-|q_t|-\frac{1-|q_t|}{2})+F_t(\gamma+|q_t|-\frac{1-|q_t|}{2})]$$

$$=\sum_{t=1}^{T}2\,M_tZ_t\tau_t(\gamma-\frac{1+|q_t|}{2})+\sum_{t=1}^{T}2F_tZ_t\tau_t(-\frac{1-|q_t|}{2}) \tag{21}$$

设置 $\gamma=(\delta+1)/2$，其中 $\delta>1$，并将其代入不等式(21)，可以得到如下不等式：

$$\left(\frac{1+\delta}{2}\right)^2\|\boldsymbol{u}\|^2\geqslant\sum_{t=1}^{T}F_tZ_t\tau_t(\delta+|q_t|) \tag{22}$$

将 $l_t(w_t)/\|x_t\|^2\geqslant l_t(w_t)/R^2$ 代入不等式(22)，可以得到：

$$\left(\frac{1+\delta}{2}\right)^2\|\boldsymbol{u}\|^2\geqslant\frac{1}{R^2}\sum_{t=1}^{T}F_tZ_t\tau_t(\delta+|q_t|) \tag{23}$$

通过使用不等式(23)并进行期望值推导，可证明定理 1。

$$\mathbb{E}[\frac{1}{R^2}\sum_{t=1}^{T}F_tZ_tl_t(w_t)(\delta+|q_t|)]$$

$$=\mathbb{E}[\frac{1}{R^2}\sum_{t=1}^{T}F_tl_t(w_t)(\delta+|q_t|)\mathbb{E}(\delta_t)]$$

$$=\frac{1}{R^2}\mathbb{E}[\delta\sum_{t=1}^{T}F_tl_t(w_t)]$$

$$\leqslant(\frac{\delta+1}{2})^2\|\boldsymbol{u}\|^2 \tag{24}$$

定理 1 证毕。

4　实验及结果分析

4.1　实验设置

4.1.1　数据集

本小节在 12 个数据集上测试 PAACDS 及其变体 PAACDS-I 的有效性。所有的数据集均来自 UCI 数据库，数据集是随机选择的，主要涵盖了图像分类、经典二分类、垃圾邮件分类、信用评分和风险预测、基因研究、医疗研究等应用领域。例如，wdbc 数据集是一种常用于乳腺癌诊断的数据集，包含了 569 个乳腺肿瘤样本的相关特征数据，每个样本包含了 30 个特征，这些特征基于数字化的图像来描述乳腺细胞核的特征，包括其大小、形状、质地等。credit-A 数据集是一个经常被用于信用评分和风险预测的数据集，包含了 690 个样本，每个样本由 15 个特征组成，这些特征描述了申请人的个人和财务信息，包括年龄、性别、婚姻状况、收入、贷款金额等。spambase 数据集是一个用于垃圾邮件分类的常用数据集，用于训练和评估机器学习算法在垃圾邮件检测方面的性能，包含了来自电子邮件的 4 601 个样本，这些样本被标记为两个类别：垃圾邮件(Spam)和非垃圾邮件(Non-Spam)。每个样本由 57 个特征组成，这些特征基于电子邮件的内容和元数据，包括词频、字符频率、特殊字符的存在等。spect 数据集是一个用于图像分类的数据集，包含了 267 个样本，每个样本由 22 个特征组成，这些特征描述了患者的心脏 spect 图像的各个方面，如像素强度、形状、大小等。实验数据集的详细情况如表 1 所列。

表 1　实验数据集

Table 1　Experimental datasets

数据集	实例数	特征数
wdbc	569	30
splice	3 190	60
credit-a	690	15
svmguide3	1 243	22
spambase	4 601	57
ionosphere	351	33
spect	267	22
libras	360	90
dermatology	358	34
arrhythmia	452	279
Kr-vs-kp	3 196	36
pima	768	8

4.1.2　评价指标

本文使用两个评价指标，分别是准确率(Accuracy)和 F_β-measure。F_β-measure 作为默认的评估指标，其中 $\beta=1$。F_β-measure 结合了模型的精确度和召回率，能够综合评估算法的性能。

为了确定 PAACDS 及其变体 PAACDS-I 与对比算法在 F-measure 预测中是否存在显著差异，本文进行了一项 Friedman 检验，显著性水平为 95%。Friedman 检验用于检验多个算法在多个数据集上的性能是否有显著差异。如果拒绝了 Friedman 检验的零假设，那么可以得出结论：这些竞争算法的性能存在显著差异。

在拒绝了零假设之后，使用 Nemenyi 测试进行事后检验。Nemenyi 测试能够帮助确定哪些算法之间存在显著差异。通过进行多次配对比较，可以计算出算法之间的平均差异和显著性水平。如此，就可以获得算法之间的相对性能，并确定是否存在显著差异。

4.1.3　具体实现设置

为了模拟任意数据流，随机地从每个到达的实例 x_t 中删除特征，删除特征的比例用 α 表示。例如，$\alpha=0.5$ 中最多有 50%的特征被随机删除。在本文的实验中，α 默认为 0.5。

所有实验在每个数据集上进行了 10 次随机重复操作，并且所有的实验结果都是平均值。

4.2　实验结果

4.2.1　PAACDS 和 PAACDS-I vs. 固定特征空间的在线学习方法

本节将 PAACDS 及其变体 PAACDS-I 与传统的在线学习方法 OGD[15] 和 RDA[30] 进行比较。OGD 和 RDA 在完整的特征空间数据流上进行实验，PAACDS 和 PAACDS-I 在任意变化的数据流以及标签的查询率固定在 20%的设置下进行实验。表 3 和表 4 分别列出了这些对比算法的 F-measure 和 Accuracy 的实验结果。根据 Friedman 检验，F-measure 的 p 值为 2.14×10^{-2}，Accuracy 的 p 值为 3.22×10^{-7}。根据 Nemenyi 检验，计算的临界差值(Critical Difference，CD)为

1.48。由表 2、表 3 和表 4 可以得到以下结论。

1)PAACDS 和 PAACDS-I vs. OGD

根据统计检验,在 F-measure 上,PAACDS 与 OGD 之间没有显著差异;在 Accuracy 上,PAACDS 与 OGD 之间有显著差异。其中,在 Accuracy 上,PAACDS 和 PAACDS-I 的表现优于 OGD,且 PAACDS 的性能平均比 OGD 高出约 18%;在 F-measure 上,OGD 的表现优于 PAACDS 和 PAACDS-I,且 OGD 的性能平均比 OLCDS 高出约 26%。然而,值得一提的是,OGD 是应用在完整的特征空间上,而 PAACDS 不仅是在任意变化数据流上,而且标签的查询比例为 20%。因此,PAACDS 和 PAACDS-I 具有适应固定特征空间和特征空间任意变化且具有标签稀疏场景的独特优势。此外,由表 2 可知,本文提出的算法在 10 个数据集上有 6 个运行时间最短,这表明了本文算法中稀疏性模型的有效性。综上,PAACDS 和 PAACDS-I 具有更广泛的适用范围和应用场景。

2)PAACDS 和 PAACDS-I vs. RDA

根据统计检验,在 F-measure 上,PAACDS 与 RDA 之间没有显著差异;在 Accuracy 上,PAACDS 与 OGD 之间有显著差异。其中,在 Accuracy 上,PAACDS 和 PAACDS-I 的表现优于 OGD,且 PAACDS 的性能平均比 RDA 高出约 15%;在 F-measure 上,RDA 的表现优于 PAACDS 和 PAACDS-I,且 RDA 的性能平均比 PAACDS 高出约 27%。RDA 是一种双平均方法,它结合了随机梯度下降和正则化技术,用于处理数据流上的机器学习和优化问题。然而,类似于 OGD,RDA 也局限于具有固定特征空间的传统在线学习场景。

表 2 运行时间

Table 2 Operation time

(s)

datasets	OGD	RDA	PAACDS	PAACDS-I
wdbc	2.816	5.577	2.295	**2.100**
splice	**2.729**	10.197	15.812	10.439
credit-a	2.783	5.730	2.679	**1.489**
svmguide3	**2.752**	4.187	6.706	3.723
spambase	**2.828**	9.081	34.061	17.705
ionosphere	2.691	8.042	1.612	**1.378**
spect	2.698	5.622	1.113	**0.928**
libras	2.732	6.027	2.068	**2.038**
dermatology	2.675	7.607	1.802	**1.672**
arrhythmia	**3.060**	39.167	6.837	6.870

表 3 与传统在线学习方法的 F-measure 比较

Table 3 F-measure results compared with traditional online learning methods

Datasets	OGD	RDA	PAACDS	PAACDS-I
wdbc	**0.937±0.010**	0.894±0.006	0.793±0.025	0.786±0.027
splice	**0.789±0.004**	0.731±0.010	0.617±0.028	0.629±0.012
credit-a	**0.787±0.014**	0.753±0.012	0.617±0.048	0.605±0.054
svmguide3	0.346±0.008	0.320±0.004	0.369±0.041	**0.377±0.030**
spambase	0.676±0.003	0.656±0.007	0.684±0.009	**0.687±0.015**
ionosphere	0.714±0.011	**0.787±0.013**	0.477±0.067	0.445±0.113
spect	0.839±0.010	**0.863±0.014**	0.567±0.054	0.629±0.078
libras	0.848±0.014	**0.901±0.011**	0.647±0.077	0.622±0.053
dermatology	0.940±0.013	**0.980±0.017**	0.707±0.070	0.699±0.062
arrhythmia	**0.907±0.014**	0.893±0.021	0.655±0.050	0.652±0.052
AVG.	**0.7784**	0.7777	0.6133	0.6131
AVG. RANKS	**1.7**	2.0	3.0	3.4

表 4 与传统在线学习方法的 Accuracy 比较

Table 4 Accuracy results compared with traditional online learning methods

Datasets	OGD	RDA	PAACDS	PAACDS-I
wdbc	0.921±0.014	0.904±0.025	**0.973±0.034**	0.971±0.035
splice	0.787±0.036	0.767±0.016	0.916±0.064	**0.923±0.052**
credit-a	0.656±0.013	0.672±0.024	**0.931±0.035**	0.923±0.015
svmguide3	0.899±0.015	0.859±0.018	**0.919±0.076**	0.915±0.045
spambase	0.629±0.036	0.672±0.027	**0.953±0.064**	0.950±0.013
ionosphere	0.789±0.012	0.817±0.038	**0.910±0.023**	0.905±0.034
spect	0.648±0.031	0.648±0.014	0.899±0.046	**0.901±0.025**
libras	0.750±0.046	0.847±0.010	**0.917±0.063**	0.906±0.063
dermatology	0.889±0.016	**0.958±0.032**	0.939±0.035	0.939±0.024
arrhythmia	0.857±0.022	0.868±0.021	0.895±0.025	**0.908±0.045**
AVG.	0.7825	0.8012	**0.9252**	0.9241
AVG. RANKS	3.6	3.1	**1.4**	1.7

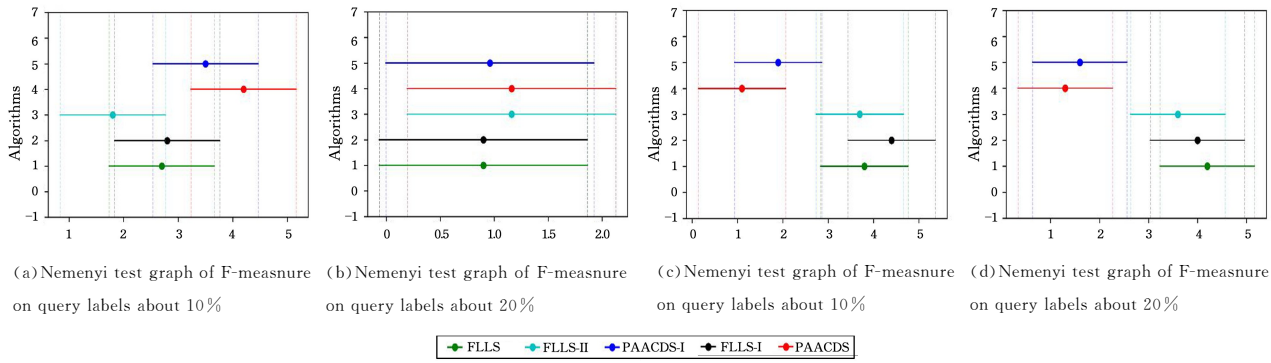
综上所述,PAACDS 和 PAACDS-I 的性能与传统的在线学习方法相比,在 F-measure 上相当,在 Accuracy 上更优。然而,这些传统的在线学习方法只能处理具有固定特征空间的数据流,而 PAACDS 和 PAACDS-I 不仅可以处理特征空间任意变化的数据流,而且可以应用于稀疏标签的场景。因此,PAACDS 和 PAACDS-I 具有更广泛的适用性和应用场景。

4.2.2 PAACDS 和 PAACDS-I vs. 标签稀疏在线学习方法

本小节从不同方面将 PAACDS 和其变体 PAACDS-I 与最先进的在线学习半监督数据流算法进行对比,包括 FLLS[22] 以及它的两个变体 FLLS-I 和 FLLS-II。在 12 个数据集上进行实验,数据集的具体信息如表 1 所列。

表 5—表 8 列出了 PAACDS 和 PAACDS-I 与对比算法 FLLS 以及它的两个变体 FLLS-I 和 FLLS-II 在不同标签查询比例下的实验结果,性能指标包括 F-measure 和准确率(Accuracy)。采用 Friedman 检验分别在度量指标 F-measure 查询标签比例 10%、度量指标 F-measure 查询标签比例 20%、度量指标 Accuracy 查询标签比例 10%、度量指标 Accuracy 查询标签比例 20%下进行比较,计算出的 p 值分别为 6.8792×10^{-3},1.0953×10^{-1},4.2958×10^{-13} 和 1.7796×10^{-11}。根据 Nemenyi 检验,计算的临界差值(CD)为 1.93。由表 5—表 8 和图 3 可以得出以下结论。

根据统计检验,在 F-measure 上,PAACDS,PAACDS-I 与 FLLS,FLLS-I,FLLS-II 之间没有显著性差异;在 Accuracy 上,PAACDS,PAACDS-I 与 FLLS,FLLS-I,FLLS-II 之间有显著性差异。根据 F-measure 实验结果,FLLS,FLLS-I,FLLS-II 相比 PAACDS 和 PAACDS-I 整体性能更好。但是,FLLS 以及它的两个变体 FLLS-I 和 FLLS-II 主要是处理梯形数据流,对数据流的特征空间变化规则进行了假设,即特征空间有规则地变化;而本文提出的 PAACDS 以及 PAACDS-I 不对数据流的特征空间进行任何的假设,即数据流的特征空间可以任意地变化,所以 PAACDS 和 PAACDS-I 适应范围更广。根据 Accuracy 实验结果,PAACDS 和 PAACDS-I 相比 FLLS,FLLS-I 和 FLLS-II,不仅在所有数据集上取得了最好的性能,而且平均值最高,平均排名最好,这充分表明 PAACDS 和 PAACDS-I 的整体性能更好,使用场景更加广泛。



(a) Nemenyi test graph of F-measnure on query labels about 10%　(b) Nemenyi test graph of F-measnure on query labels about 20%　(c) Nemenyi test graph of F-measnure on query labels about 10%　(d) Nemenyi test graph of F-measnure on query labels about 20%

图 3　PAACDS 和 PAACDS-I 与对比算法在不同标签查询比例上的统计测试图

Fig. 3　Statistical test charts of PAACDS, PAACDS-I and comparison algorithms on different label query ratios

表 5　查询标签比例固定为约 10%情况下的 F-measure 结果

Table 5　F-measure results of algorithms with the query labels ratio fixed to about 10%

Datasets	FLLS	FLLS-I	FLLS-II	PAACDS	PAACDS-I
wdbc	0.837±0.030	0.761±0.019	**0.858±0.044**	0.709±0.050	0.730±0.060
splice	0.747±0.027	0.735±0.023	**0.761±0.063**	0.595±0.015	0.596±0.032
credit-a	0.683±0.019	0.712±0.010	**0.745±0.052**	0.575±0.051	0.601±0.070
svmguide3	0.630±0.024	0.606±0.020	**0.657±0.034**	0.352±0.044	0.390±0.048
spambase	0.846±0.043	0.850±0.042	**0.863±0.013**	0.633±0.040	0.628±0.036
ionosphere	**0.785±0.020**	0.764±0.013	0.758±0.035	0.437±0.091	0.386±0.085
spect	0.550±0.035	0.497±0.034	0.436±0.024	**0.598±0.131**	0.580±0.069
libras	0.570±0.010	0.619±0.014	0.620±0.020	0.604±0.113	**0.622±0.075**
dermatology	0.781±0.017	0.793±0.052	**0.795±0.053**	0.684±0.102	0.652±0.073
arrhythmia	0.622±0.029	0.609±0.024	0.633±0.063	0.592±0.071	**0.644±0.065**
Kr-vs-kp	0.745±0.019	0.700±0.013	**0.745±0.026**	0.619±0.031	0.607±0.032
pima	0.638±0.032	0.685±0.025	**0.687±0.045**	0.484±0.070	0.461±0.069
AVG.	0.703	0.694	**0.713**	0.573	0.575
AVG. RANKS	2.58	2.75	**1.75**	4.17	3.75

表 6　查询标签比例固定为约 20%情况下的 F-measure 结果

Table 6　F-measure results of algorithms with query labels ratio fixed to about 20%

Datasets	FLLS	FLLS-I	FLLS-II	PAACDS	PAACDS-I
wdbc	0.838±0.015	**0.851±0.038**	0.832±0.033	0.793±0.025	0.786±0.027
splice	0.763±0.012	0.760±0.021	**0.779±0.042**	0.617±0.028	0.629±0.012
credit-a	**0.824±0.038**	0.747±0.025	0.822±0.018	0.617±0.048	0.605±0.054
svmguide3	0.618±0.029	0.667±0.027	**0.674±0.035**	0.369±0.041	0.377±0.030
spambase	0.867±0.013	0.842±0.015	**0.869±0.020**	0.684±0.009	0.687±0.015
ionosphere	0.810±0.023	**0.837±0.017**	0.823±0.045	0.477±0.067	0.445±0.113
spect	0.469±0.034	0.614±0.034	0.533±0.032	0.567±0.054	**0.629±0.078**
libras	0.637±0.005	0.630±0.010	0.611±0.005	**0.647±0.077**	0.622±0.053
dermatology	0.808±0.011	0.831±0.016	**0.835±0.010**	0.707±0.070	0.699±0.062
arrhythmia	0.646±0.035	0.645±0.021	0.647±0.023	**0.655±0.050**	0.652±0.052
Kr-vs-kp	**0.837±0.052**	0.730±0.035	0.777±0.021	0.664±0.020	0.679±0.017
pima	0.679±0.024	**0.691±0.026**	0.661±0.029	0.484±0.071	0.479±0.054
AVG.	0.733	0.737	**0.739**	0.607	0.607
AVG. RANKS	2.50	2.42	**2.33**	3.75	4.00

表 7　查询标签比例固定为约 10%情况下的 Accuracy 结果

Table 7　Accuracy results of algorithms with query labels ratio fixed to about 10%

Datasets	FLLS	FLLS-I	FLLS-II	PAACDS	PAACDS-I
wdbc	0.830±0.035	0.740±0.052	0.820±0.043	**0.976±0.032**	0.975±0.046
splice	0.746±0.067	0.734±0.036	0.760±0.045	0.958±0.054	**0.960±0.063**
credit-a	0.670±0.056	0.713±0.063	0.742±0.023	**0.965±0.013**	0.958±0.026
svmguide3	0.675±0.033	0.624±0.065	0.672±0.013	**0.962±0.042**	0.960±0.063
spambase	0.844±0.024	0.857±0.035	0.868±0.024	**0.975±0.052**	0.973±0.073
ionosphere	0.817±0.053	0.780±0.055	0.769±0.052	**0.962±0.073**	0.947±0.063
spect	0.507±0.047	0.500±0.023	0.415±0.062	**0.951±0.039**	0.930±0.034
libras	0.561±0.063	0.586±0.024	0.606±0.022	**0.961±0.058**	0.958±0.036
dermatology	0.717±0.073	0.703±0.042	0.711±0.045	**0.962±0.035**	0.956±0.074
arrhythmia	0.596±0.030	0.596±0.053	0.613±0.053	**0.943±0.064**	0.940±0.035
Kr-vs-kp	0.791±0.013	0.716±0.024	0.751±0.013	0.957±0.011	**0.967±0.026**
pima	0.639±0.035	0.683±0.046	0.683±0.035	0.935±0.024	**0.966±0.041**
AVG.	0.699	0.686	0.701	**0.959**	0.958
AVG. RANKS	3.83	4.33	3.67	**1.25**	1.75

表 8 查询标签比例固定为约 20%情况下的 Accuracy 结果

Table 8 Accuracy results of algorithms with query labels ratio fixed to about 20%

Datasets	FLLS	FLLS-I	FLLS-II	PAACDS	PAACDS-I
wdbc	0.810±0.023	0.820±0.035	0.810±0.063	**0.973±0.034**	0.971±0.035
splice	0.762±0.062	0.759±0.027	0.778±0.074	0.916±0.064	**0.923±0.052**
credit-a	0.818±0.066	0.744±0.045	0.817±0.026	**0.931±0.035**	0.923±0.015
svmguide3	0.633±0.034	0.685±0.021	0.686±0.026	**0.919±0.076**	0.915±0.045
spambase	0.869±0.014	0.850±0.033	0.877±0.056	**0.953±0.064**	0.950±0.013
ionosphere	0.834±0.055	0.854±0.072	0.837±0.052	**0.910±0.023**	0.905±0.034
spect	0.507±0.073	0.567±0.063	0.507±0.025	0.899±0.046	**0.901±0.025**
libras	0.631±0.063	0.594±0.022	0.597±0.063	**0.917±0.063**	0.906±0.063
dermatology	0.756±0.023	0.764±0.014	0.769±0.023	**0.939±0.035**	0.939±0.024
arrhythmia	0.618±0.063	0.659±0.056	0.633±0.056	0.895±0.025	**0.908±0.045**
Kr-vs-kp	0.852±0.053	0.714±0.046	0.809±0.047	**0.942±0.035**	0.932±0.047
pima	0.674±0.025	0.686±0.028	0.664±0.033	**0.925±0.023**	0.918±0.023
AVG.	0.730	0.725	0.732	**0.927**	0.924
AVG. RANKS	4.08	4.00	3.75	**1.25**	1.67

4.2.3 PAACDS 不同标签查询比例分析

为了进一步了解 PAACDS 和 PAACDS-I 在不同标签查询比例变化时可能受到的影响，将参数 C 设置为 1，B 设置为 0.64，标签查询比例在 10%到 50%之间变化(通过调整参数 δ)。表 9 列出了在不同标签查询比例下的平均 F-measure 结果。

表 9 不同查询标签比例的 F-measure 结果

Table 9 F-measure results with different query labels ratios

Datasets	Query Labels Ratio	PAACDS	PAACDS-I	Data Set	Query Labels Ratio	PAACDS	PAACDS-I
wdbc	10%	0.709±0.050	0.730±0.060	ionosphere	10%	0.437±0.091	0.386±0.085
	20%	0.793±0.025	**0.786±0.027**		20%	0.477±0.067	0.445±0.113
	30%	0.760±0.032	0.738±0.043		30%	0.489±0.067	0.456±0.095
	40%	0.790±0.031	0.781±0.033		40%	0.512±0.040	0.493±0.056
	50%	**0.824±0.030**	0.784±0.047		50%	**0.543±0.033**	**0.519±0.059**
splice	10%	0.595±0.015	0.596±0.032	spect	10%	0.598±0.131	0.580±0.069
	20%	0.617±0.028	**0.629±0.012**		20%	0.567±0.054	**0.629±0.078**
	30%	0.604±0.013	0.564±0.014		30%	0.619±0.060	0.608±0.048
	40%	0.610±0.016	0.575±0.028		40%	0.621±0.038	0.615±0.052
	50%	**0.631±0.018**	0.586±0.019		50%	**0.627±0.048**	0.621±0.038
credit-a	10%	0.575±0.051	0.601±0.070	libras	10%	0.604±0.113	0.622±0.075
	20%	0.617±0.028	**0.605±0.054**		20%	0.647±0.077	0.622±0.053
	30%	0.631±0.025	0.584±0.049		30%	0.613±0.066	0.609±0.068
	40%	0.638±0.030	0.589±0.025		40%	0.624±0.036	0.626±0.039
	50%	**0.658±0.016**	0.599±0.027		50%	**0.633±0.052**	**0.631±0.036**
svmguide3	10%	0.352±0.044	0.390±0.048	dermatology	10%	0.684±0.102	0.652±0.073
	20%	0.369±0.041	0.377±0.030		20%	0.707±0.070	0.699±0.062
	30%	0.371±0.033	0.384±0.043		30%	0.684±0.036	0.709±0.042
	40%	0.384±0.019	0.383±0.031		40%	0.709±0.058	0.713±0.036
	50%	**0.408±0.022**	**0.391±0.021**		50%	**0.717±0.029**	**0.726±0.033**
spambase	10%	0.633±0.040	0.628±0.036	arrhythmia	10%	0.592±0.071	0.644±0.065
	20%	0.684±0.009	**0.687±0.015**		20%	0.655±0.050	0.652±0.052
	30%	0.662±0.010	0.621±0.027		30%	0.680±0.029	0.690±0.043
	40%	0.695±0.014	0.638±0.023		40%	0.687±0.036	0.704±0.027
	50%	**0.708±0.009**	0.655±0.016		50%	**0.693±0.024**	**0.707±0.025**

从表 9 中可以明显观察到 F-measure 随着标签查询比例的增加而增加，这也符合常识。原因在于，对于标签稀疏的场景，并不是所有的数据实例都有标签，只有小部分数据实例有标签，所以每当一个数据实例在 t 时刻到达，先通过伯努利随机变量 δ 计算概率 Z_t，如果 $Z_t=1$，则查询标签，否则不查询标签。通过查询标签可以实时更新模型，使其累计损失最小化，从而使得模型性能更好。因此，标签查询比例的增加，意味着有更多的训练数据用于实时更新模型，模型的性能更好。

4.2.4 参数分析

在本小节中，使用两个数据集(wdbc 和 credit-a)来研究本文提出算法的参数敏感性。该算法依赖于两个关键参数：惩罚代价参数 C 和选择特征比例 B。

分别进行两组实验来分析这些参数的影响。在第一组实验中，将 B 固定为 1，将 C 从 1×10^{-4} 变化到 1×10^{4}。在第二组实验中，将 C 固定为 1，并使用集合{0.04，0.08，0.16，0.32，0.64}中的值来调整 B。以上实验中查询标签比例都固定在 50%。通过系统地探索不同参数值的影响，了解参数 C 和 B 如何影响算法的性能。

图 4(a)和图 4(b)展示了 PAACDS 和 PAACDS-I 在不同 C 值下的性能表现。可以看出，这两个算法的性能整体上是先提升后下降。总体而言，在较大的范围内，PAACDS 和 PAACDS-I 对参数 C 并不敏感。此外，随着 C 值增大，PAACDS 和 PAACDS-I 的性能越接近。这是因为随着 C 值的增大，PAACDS 和 PAACDS-I 的 τ 值越来越接近，导致它们的模型性能也越来越接近。图 4(c)和图 4(d)展示了 PAACDS 和 PAACDS-I 在不同值下的性能表现。从整体上

看，在参数 B 值逐渐变化的情况下，两个模型的准确率基本保持不变，这表明 PAACDS 和 PAACDS-I 对参数 B 并不敏感。其次，较大的 B 值不一定会获得更好的性能，这表明本文所采用的稀疏策略不仅改善了内存的使用，而且使算法具有更好的性能。

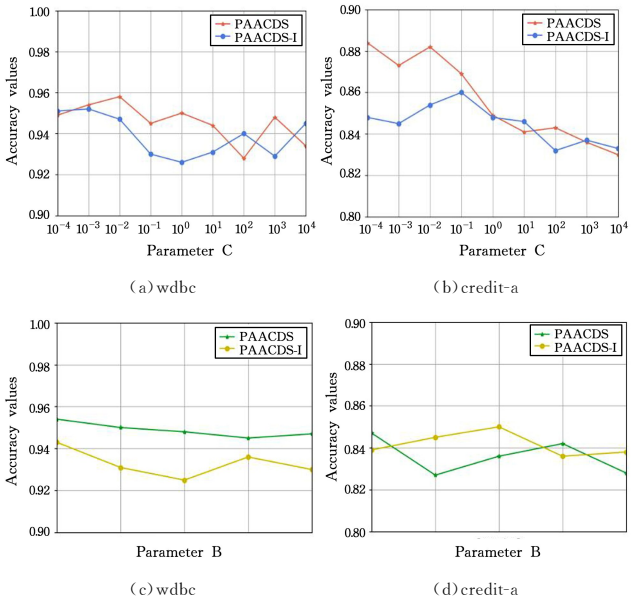


图 4　两个数据集上关于参数 C 和 B 的准确率结果

Fig. 4　Accuracy results on two datasets with respect to parameter C and B

综上所述，可知 PAACDS 和 PAACDS-I 整体上对惩罚成本参数 C 和选择特征比例 B 不敏感。

结束语　本文提出了一种基于被动-主动查询策略的任意数据流在线学习算法 PAACDS 和其变体 PAACDS-I。具体来说，首先，利用在线主动学习方法选择有价值的数据实例，从而在最小的监督下建立优越的预测模型。然后，在查询获得有价值数据实例的真实标签后，结合在线被动-主动更新规则和边界最大化原则来共同更新基于任意数据流的共享和新增特征空间的动态分类器。最后，使用投影截断技术构建一个稀疏但高效的在线学习模型。在 12 个数据集上的大量实验对比结果表明，本文提出的算法在处理标签稀疏场景下的任意数据流的问题上具有良好的性能。

本文虽然利用被动-主动更新规则和边界最大化原则解决了在标签稀疏场景下任意数据流的问题，但是该算法仅针对线性任务设计，如何将现有的算法扩展到非线性任务是未来重要的研究工作。此外，如何充分利用数据流中更多有价值的信息，如分布式信息，以及如何在不断变化的特征空间中进行聚类等问题，也是未来非常值得关注的研究方向。

参 考 文 献

[1] ZHAO P, WANG D, WU P, et al. A unified framework for sparse online learning[J]. ACM Transactions on Knowledge Discovery from Data(TKDD), 2020, 14(5): 1-20.

[2] ZHAO Q L, JIANG Y H. Online Data Stream Mining for Seriously Unbalanced Applications[J]. Computer Science, 2017, 44(6): 255-259.

[3] DE LANGE M, TUYTELAARS T. Continual prototype evolution: Learning online from non-stationary data streams[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 8250-8259.

[4] VIDHYA M, AJI S. Parallelized extreme learning machine for online data classification[J]. Applied Intelligence, 2022, 52(12): 14164-14177.

[5] FU X, SEO E, CLARKE J, et al. Link prediction under imperfect detection: Collaborative filtering for ecological networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 33(8): 3117-3128.

[6] PHADKE A, KULKARNI M, BHAWALKAR P, et al. A review of machine learning methodologies for network intrusion detection[C]// 2019 3rd International Conference on Computing Methodologies and Communication(ICCMC). IEEE, 2019: 272-275.

[7] ULLO S L, SINHA G R. Advances in smart environment monitoring systems using IoT and sensors[J]. Sensors, 2020, 20(11): 3113.

[8] HE Y, WU B, WU D, et al. Online learning from capricious data streams: a generative approach[C]// International Joint Conference on Artificial Intelligence Main Track. 2019.

[9] YOU D, XIAO J, WANG Y, et al. Online learning from incomplete and imbalanced data streams[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(10): 10650-10665.

[10] ZHANG D, JIN M, CAO P. ST-Meta Diagnosis: Meta learning with Spatial Transform for rare skin disease Diagnosis[C]// 2020 IEEE International Conference on Bioinformatics and Biomedicine(BIBM). IEEE, 2020: 2153-2160.

[11] ZHOU Y, REN H, LI Z, et al. Anomaly detection via a combination model in time series data[J]. Applied Intelligence, 2021, 51: 4874-4887.

[12] LU J, LIU A, DONG F, et al. Learning under concept drift: A review[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(12): 2346-2363.

[13] AGRAHARI S, SINGH A K. Concept drift detection in data stream mining: A literature review[J]. Journal of King Saud University-Computer and Information Sciences, 2022, 34(10): 9523-9540.

[14] LI H, FANG C, LIN Z. Accelerated first-order optimization algorithms for machine learning[C]// Proceedings of the IEEE. 2020: 2067-2082.

[15] ZINKEVICH M. Online convex programming and generalized infinitesimal gradient ascent[C]// Proceedings of the 20th International Conference on Machine Learning(ICML-03). 2003: 928-936.

[16] CRAMMER K, LEE D. Learning via gaussian herding [C]// Proceedings of the 24th International Conference on Neural Information Processing Systems. 2010: 451-459.

[17] CRAMMER K, DREDZE M, KULESZA A. Multi-class confidence weighted algorithms[C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. 2009: 496-504.

[18] CHEN Z, ZHAN H, SHENG V, et al. Projection dual averaging based second-order online learning[C]// 2022 IEEE International Conference on Data Mining(ICDM). IEEE, 2022: 51-60.

[19] ZHANG Q, ZHANG P, LONG G, et al. Online learning from trapezoidal data streams[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(10): 2709-2723.

[20] GU S, QIAN Y, HOU C. Learning with incremental instances and features[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 35(7): 9713-9727.

[21] YU E, LU J, ZHANG B, et al. Online boosting adaptive learning under concept drift for multistream classification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2024: 16522-16530.

[22] BEYAZIT E, ALAGURAJAH J, WU X. Online learning from data streams with varying feature spaces[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 3232-3239.

[23] HE Y, WU B, WU D, et al. Toward mining capricious data streams: A generative approach[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(3): 1228-1240.

[24] GU S, QIAN Y, HOU C. Incremental feature spaces learning with label scarcity[J]. ACM Transactions on Knowledge Discovery from Data(TKDD), 2022, 16(6): 1-26.

[25] LIU Y, FAN X, LI W, et al. Online passive-aggressive active learning for trapezoidal data streams[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 34(10): 6725-6739.

[26] CHENG J, ZHENG Z, GUO Y, et al. Active broad learning with multi-objective evolution for data stream classification[J]. Complex & Intelligent Systems, 2024, 10(1): 899-916.

[27] GU S, LUO T, HE M, et al. Online Learning With Incremental Feature Space and Bandit Feedback[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(12): 12902-12916.

[28] DIN S U, ULLAH A, MAWULI C B, et al. A reliable adaptive prototype-based learning for evolving data streams with limited labels[J]. Information Processing & Management, 2024, 61(1): 103532.

[29] HAO S, LU J, ZHAO P, et al. Second-order online active learning and its applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 30(7): 1338-1351.

[30] LIN X. Dual averaging method for regularized stochastic learning and online optimization[J]. The Journal of Machine Learning Research, 2010, 11: 2543-2596.

ZHANG Shuai, born in 1996, postgraduate, is a student member of CCF (No. U3918G). His main research interests include data streams and online learning.

ZHOU Peng, born in 1987, Ph. D, is a member of CCF (No. K6292M). His main research interests include data mining and machine learning.

(责任编辑:何杨)