面向开放特征空间的概念演化检测方法

苏 睿[1],郭虎升[1,2],王 婧[1],王文剑[2*]
(1. 山西大学计算机与信息技术学院,山西太原 030006;
2. 山西大学计算智能与中文信息处理教育部重点实验室,山西太原 030006)

摘 要: 在众多现实场景中数据以流的形式持续产生,由于流数据具有动态变化的特点,在生成过程中可能产生新的类别,也被称为概念演化. 概念演化是流数据挖掘模型预测性能衰退甚至预测失效的主要原因. 因此,能及时发现类空间变化并提醒模型做出适应性调节的概念演化检测方法受到广泛关注. 然而,目前多数概念演化检测方法基于特征空间静态不变的假设构建算法. 在现实场景中,特征空间同样具有动态性,属于开放的空间. 具体来说,随时间推移可能出现部分特征消失和新特征产生的现象,从而破坏上述假设并导致已有算法失效. 针对这一问题,本文提出一种面向开放特征空间的概念演化检测方法(Concept evolution Detection method for Open Feature space,CD_OF). 该方法通过构建微簇集成模型对新进实例分类,对于开放特征空间中的旧特征消失问题,通过利用转移矩阵将旧特征中包含的信息转换到共享特征中;对于新出现的特征,拓展共享特征空间并重构集成模型. 在此基础上,根据样本的共享邻域信息定义样本间相似度以检测概念演化,并建立动态衰减模型,以解决开放特征空间下的类消失和类循环问题. 实验结果表明,本文所提出的方法能够对开放特征空间中特征的变化作出及时响应,增强概念演化检测的能力,在特征空间变化的真实流数据中与现有方法相比,错误率降低了 1.7%~11.4%.

关键词: 概念演化;开放特征空间;特征相似性度量;共享邻域;动态衰减模型;在线学习
基金项目: 国家自然科学基金(No.U21A20513,No.62276157,No.62476157);山西省重点研发计划(No.202202020101003)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2025)10-3718-12
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250416

Concept Evolution Detection Method for Open Feature Space

SU Rui[1], GUO Hu-sheng[1,2], WANG Jing[1], WANG Wen-jian[2*]
(1. *School of Computer and Information Technology*, *Shanxi University*, *Taiyuan*, *Shanxi* 030006, *China*;
2. *Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education*, *Shanxi University*, *Taiyuan*, *Shanxi* 030006, *China*)

Abstract: In many real-world scenarios, data is continuously generated in the form of streams. Due to the dynamic nature of streaming data, new categories may emerge during the generation process, which is known as concept evolution. Concept evolution is one of the primary challenges leading to the degradation or even failure of predictive performance in stream mining models. Therefore, concept evolution detection methods capable of promptly identifying changes in the class space and alerting models to perform adaptive adjustments have attracted widespread attention. However, most of the current concept evolution detection methods construct algorithms based on the assumption that the feature space is static and unchanging. In real scenarios, the feature space is also dynamic and belongs to the open space. Specifically, over time, some features may disappear and new features may emerge, thus violating the above assumption and causing existing algorithms to fail. To address this problem, this paper proposes a concept evolution detection method for open feature space (CD_OF). The method constructs a micro-cluster ensemble model to classify incoming instances. For the problem of disappearing old features in the open feature space, the information contained in the old features is converted to the shared features through the transfer matrix; for the newly emerged features, the shared feature space is expanded and the integration model is reconstructed. On this basis, the inter-sample similarity is defined based on the shared neighborhood information of the samples to detect concept evolution, and the dynamic decay model is established to solve the class vanishing and classifications cycling problems under the open feature space. The experimental results show that the method

收稿日期:2025-05-25;录用日期:2025-10-11;责任编辑:覃怀银
*通讯作者:王文剑

proposed in this paper is able to respond to the changes of features in the open feature space in a timely manner and enhance the ability of concept evolution detection. The error rate is reduced by 1.7% to 11.4% compared to existing methods on real streaming data with feature space variations.

Key words: concept evolution; open feature space; feature similarity measure; shared neighborhood; dynamic decay model; online learning

Foundation Item(s): National Natural Science Foundation of China (No.U21A20513, No.62276157, No.62476157); Key Research and Development Program of Shanxi Province (No.202202020101003)

1　引言

在许多现实世界应用领域中,例如交通控制、社交媒体、市场分析和传感器网络,数据以流的形式不断产生[1,2],这些数据也被称为流数据. 流数据是连续高速到达的无限有序的数据项序列,具有动态性、时序性、无限性和不可再现性等特点,给数据的收集、存储、分析和处理带来了严峻的挑战[3,4]. 由于流数据存在上述特性,为提升泛化能力而设计的离线学习算法难以适应流数据的变化. 相比于离线学习算法,流数据挖掘算法还需要进一步考虑算法的实时推理能力和适应能力[5]. 其中,实时推理能力要求算法在优化过程中不能影响模型的推理过程,因此相关算法的设计不仅需要考虑计算效率与存储效率,并且还需要选择高效的在线优化策略以保证学习效果. 适应能力要求算法适应流数据随时间推移在特征空间与标签空间可能产生的动态变化. 近年来,随着机器学习与人工智能技术快速发展,大量的在线学习算法被部署在实际场景中,然而这些算法常常会受到环境动态变化的干扰导致泛化能力逐步衰退[6,7].

流数据的动态变化主要会引起概念漂移与概念演化这2种问题. 其中,概念漂移是指由流数据特征空间和标签空间联合分布的变化导致的模型预测表现衰退问题[8~11]. 概念演化是指流数据可能动态生成新的类别,而这些类别往往是已部署模型所未知的[12,13],从而导致模型预测错误,如在网络安全领域,每当发生网络攻击入侵时,计算机系统的安全就会受到威胁,而新的计算机病毒可能以不同方式对计算机系统进行攻击,从而在对网络入侵进行防护时,检测到入侵方式的类别标签会增加,这些新的类标签应及时处理,否则分类器将把与新类相关联的所有数据实例错误分类为现有类,严重影响已有分类表现. 这两种问题是动态流数据挖掘算法的主要挑战.

目前,虽然有多种概念演化和概念漂移检测适应算法,但是这些工作大多基于特征空间静态不变的假设,即假设特征空间维度不会发生变化. 然而在许多实际应用中,数据流的特征空间会随时间动态变化,如在环境监测中需要增加新的传感器并结合部分旧传感器数据以完成新物种的监测任务;在垃圾邮件监测任务中新型垃圾邮件往往具有主题标签、短URL(Uniform Resource Locator)等特征;在循环流化床工艺中合适的补偿机制需要考虑日益变化的新燃料供给和不均匀性数据等[2,14]. 因此,静态特征空间假设难以满足不断快速发展变化的真实场景. 这些包含新特征的增加、旧特征的消失等特征变化场景的空间也被称为开放特征空间.

针对上述问题,本文提出了一种面向开放特征空间的概念演化检测方法(Concept evolution Detection method for Open Feature space,CD_OF),旨在提升当流数据局部旧特征消失或局部新特征出现场景下对新类别的检测表现. CD_OF通过寻找旧特征空间到共享特征空间的转移矩阵,利用转移矩阵实时将当前特征映射到共享特征空间,以缓解旧特征消失的影响. 对于新增特征,CD_OF基于共享特征空间进行拓展,通过学习新特征空间知识重构集成模型以适应. 在对齐特征空间之后,CD_OF结合共享近邻与样本相似度2项指标检测概念演化是否发生,并通过构建动态衰减模型,以适应动态流数据空间存在的类消失和类循环现象.

本文的主要贡献如下:(1)提出了一种基于共享特征空间的新特征空间对齐方法,旨在缓解特征空间变化对概念演化检测与适应的影响;(2)提出了一种融合共享近邻与样本相似度的概念演化检测方法以识别新类,构建了动态衰减模型识别消失类,提高新类检测性能的同时降低了在循环类样本的误报率.

2　相关工作

2.1　概念演化检测

对于流数据中存在的概念演化问题,现有的检测方法依据分类器构造方式,可以分为2种类型:基于聚类分析的概念演化检测方法和基于模型的概念演化检测方法.

基于聚类分析的概念演化检测方法[9,12,13,15~19]旨在利用传统的聚类算法对已知类进行建模,并据此定义决策边界. 通常假设同一类别的样本在特征空间中彼此之间更接近,而不同类别的样本则相对较远. 通过检测传入的样本实例是否位于决策边界内,来有效地识别和区分样本中的异常点. 对于异常值,Masud等人[14]

提出ECSMiner(Enhanced Classifier for data Streams with novel class Miner)方法,通过分析其在特征空间中与已知类样本和其他异常点之间的距离关系,计算异常点的q-邻域轮廓系数(q-Neighborhood Silhouette Coefficient,q-NSC),来确定该样本是否属于新类.在这种方法中,每个分类器的决策边界保持固定,这可能会把概念漂移导致的落入决策边界外的实例错误识别为异常值.Al-khateeb等人[20]提出的CLAM(Class based Micro classifier ensemble)方法在每一类初始数据上进行聚类,对于落在聚类边界外的测试实例,通过分析是否有足够的彼此接近的异常值来寻找新类.CLAM方法能够很好地区分循环类和新类,然而该方法不能区分循环类和已知类,通常将循环类标识为已知类.

基于模型的概念演化检测方法旨在找到一个分类模型来识别新的类,包括基于图的概念演化检测方法[21,22]、基于决策树的概念演化检测方法[23~25]和基于草图的概念演化检测方法[26].其中,基于图的检测方法有:SACCOS[21]方法使用相互图聚类技术,识别特征空间中局部区域内相似的数据实例;GNG[22]方法是一种基于图的自适应方法,在不断发展的数据流中进行增量学习.以样本实例作为节点,不同节点之间的连接作为边构建图.当新样本实例到达时,通过创建新节点和节点之间的连接更新图模型.基于决策树的检测方法有:AhtNODE[24]方法使用Hoeffding树学习器和ADWIN变化检测器,检测数据流中的概念漂移和概念演化.根据类的内聚性和分离性,识别落在Hoeffding树中未使用空间内的具有强内聚性的样本实例,声明新类.基于草图的概念演化检测方法有:Du等人[26]为整个训练数据创建全局草图,为数据中的每个类创建局部草图,全局草图用于检测数据流中的新类实例,局部草图用于对与已知类关联的数据实例进行分类.

2.2 开放特征空间学习

数据流学习环境的动态可变性,通常会导致特征空间随着时间的推移在特征维度以及属性上发生改变.最早与开放特征空间相关的研究是增量属性学习[27].当数据中出现新特征时,一方面在旧特征上继续训练已有的模型,另一方面在新特征上训练新的模型,最后将得到的新旧模型结合到一起.这种学习算法没有解决旧特征消失的问题.Hou等人[28]提出的FESL(Feature Evolvable Streaming Learning)方法采用最小化二乘损失的方法来学习新出现特征与旧特征之间的映射关系,用新的特征来恢复旧特征,并进一步引入集成学习思想,通过加权结合新旧模型对样本进行预测.Liu等人[29]提出面向特征继承性增减的在线分类算法,将被动-主动方法与结构风险最小化原则相结合,在原始特征空间使用被动方法训练模型,在新特征空间上使用主动方法训练模型.同时,采取矩阵补全算法,补全恢复数据流的缺失部分.He等人[30]提出了OCDS(Online learning from Capricious Data Streams)算法,根据特征之间的相关性,将不同阶段的数据特征整合到一个通用特征空间中,同时在该空间上训练一个学习器,使得在旧特征空间中学习的模型可以应用于新特征空间.

在真实场景中,特征空间变化与概念演化往往同时存在,然而现有方法仅能适应单一场景,在真实的复杂动态变化场景下难以保持方法的泛化性.

3 面向开放特征空间的概念演化检测方法

针对上述问题,本文提出CD_OF.该方法首先构建微簇集成模型对新进实例分类,对于开放特征空间中的旧特征消失问题,通过度量旧特征和共享特征之间的相似性,将旧特征中包含的信息转换到共享特征中;反之,对于新特征出现的问题,添加相应的新特征维度拓展特征空间,在拓展后的特征空间上重构集成模型.在此基础上,结合共享近邻与样本相似度检测概念演化,并建立动态衰减模型,以适应类消失和类循环.图1为该方法的整体框架示意图.

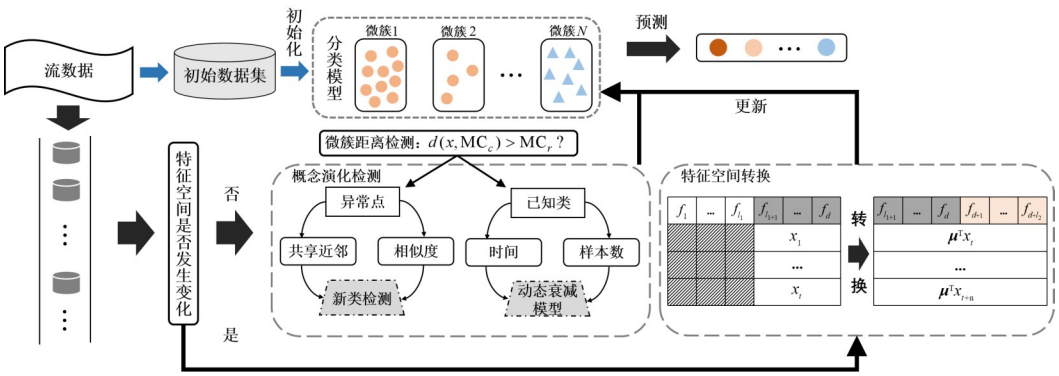


图1 CD_OF方法整体框架图

3.1 问题定义

流数据是一组大量、快速、持续到达的数据序列,一般情况下,流数据可被视为一个随时间延续而无限增长的动态数据集合. 可以将其表示如式(1)所示:

$$D=\left\{(x_i,y_i)\right\}_{i=1}^{t} \tag{1}$$

其中,(x_i,y_i)是t时刻到达的样本实例,x_i是特征向量,$x\in\mathbf{R}^d$,y_i是该样本实例对应的标签$y\in\mathbf{R}^L$.

若在流数据分类任务中t时刻发生了概念演化,可表示为标签空间的概率分布发生变化,如式(2)所示:

$$P_{t-1}(y)\neq P_t(y) \tag{2}$$

这往往是流数据中类别标签增加或者消失而导致的.

在开放特征空间下,随着时间的推移,流数据特征空间可能发生变化. 这种变化可表示为从旧特征空间到新特征空间的转换. 假设初始阶段的旧特征空间为$S_t=\{f_1,f_2,\cdots,f_d\}$,表示从最初时刻到时刻$t$的样本$x_t$特征空间,即$x_t\in\mathbf{R}^d$. 在$t+n$时刻,特征空间发生变化,空间中有$l_1$个特征消失且有$l_2$个特征出现,并且还有$d-l_1$个特征没有发生变化,因此新的特征空间表示为$S_{t+n}=\{f_{d-l_1+1},\cdots,f_d,f_{d+1},\cdots,f_{d+l_2}\}$,此时$x_{t+1}\in\mathbf{R}^{d-l_1+l_2}$. 特征空间的变化过程如图2所示.

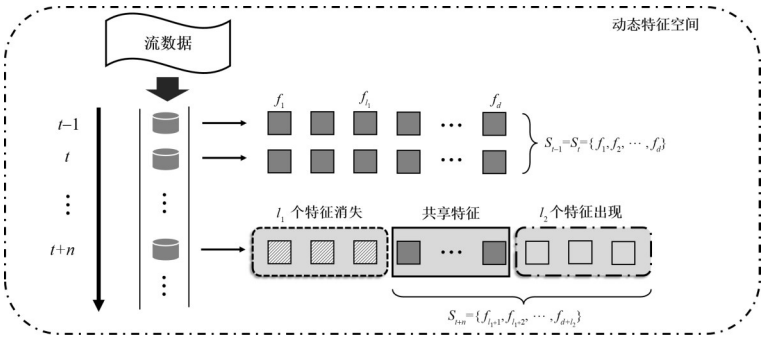


图2　开放特征空间示例图

消失特征指只在旧特征空间中存在的特征,共享特征指当前特征空间和上一阶段特征空间共同存在的特征,而新特征指只在新特征空间中存在的特征.

3.2 初始集成模型构建

假设在流数据的整个过程中,只有初始训练数据的一个子集的真实标签是已知的,而其他数据的标签是未知的. 在此训练期间,本文首先将初始有标签数据集按照数据标签类别,划分为标签不相交的集合,使得每个集合仅包含一个类的实例. 其次在每个集合上分别应用聚类算法将其划分为n个簇,利用得到的聚类对训练集中的无标签样本进行分类,直到所有无标签数据都被标记完. 最后利用标记后的数据集更新聚类簇.

由于流数据具有数据量大的特征,存储原始数据会消耗内存空间,因此计算每个聚类簇的特征并将其存储在微簇(Micro-Cluster, MC)[31]中来代替存储原始数据. 具体地,若初始训练数据有L个类,则通过K-means聚类方法生成$L\times n$个微簇. 微簇定义如式(3)所示:

$$\mathrm{MC}=(\mathrm{Ls},\mathrm{Ss},N,W) \tag{3}$$

其中,$\mathrm{Ls}=\sum_{i=1}^{N}x_i$是微簇中样本特征的线性和,$\mathrm{Ss}=\sum_{i=1}^{N}x_i^2$表示微簇中样本特征的平方和,$N$是微簇中样本的数量,$W$是微簇随时间推移的重要性,初始值设置为1. 根据微簇的定义,可计算出每个微簇的中心MC_c和半径MC_r,计算公式见式(4)、式(5):

$$\mathrm{MC}_c=\frac{\mathrm{Ls}}{N} \tag{4}$$

$$\mathrm{MC}_r=\left(\frac{\mathrm{Ss}}{N}-\frac{\mathrm{Ls}^2}{N^2}\right)^{\frac{1}{2}} \tag{5}$$

3.3 特征相似性度量与映射

新特征主要来源于真实场景中由环境因素的变化而引发的特征的变化,如在环境监测中,传感器的更换或新增,均会产生新特征. 在旧特征空间到新特征空间的变化过程中,存在部分旧特征消失和新特征产生的现象,但有一些特征始终没有发生变化,因此旧特征分布与新特征分布不完全一致. 新增特征与原有特征之间并没有确定的联系. 为了从旧特征空间转换到新特征空间,利用不变的共享特征进行辅助,以降低适应难度. 因此,假设在旧特征和共享特征之间存在确定的映射关系$\psi:\mathbf{R}^d\rightarrow\mathbf{R}^{d-l_1}$. 假设$x_t$是初始阶段缓冲窗口中的一个样本,$x_t'$是由$x_t$样本降维到$d-l_1$维度后的部分特征再进行补0得到,$x_t'$去掉尾部$l_1$个特征即为初始共享特征. 为获取当前特征分布与共享特征空间之间的最佳映射参数矩阵μ^*,本文采用局部加权线性回归方法,具体表示如式(6)所示:

$$\min_{\boldsymbol{\mu}\in R^d}\sum_{t=1}^{T_w}v_t(\boldsymbol{\mu}^{\mathrm{T}}x_t-x_t')^2 \tag{6}$$

其中,$\boldsymbol{\mu}$表示旧特征空间到共享特征空间的转移矩阵,T_w即缓冲窗口,高斯核权重v可使用x_t'与x_t计算得到,

如式(7)所示：

$$v_t=\mathrm{e}^{-\frac{(x_t-x_t')^2}{2\theta^2}} \tag{7}$$

对应时刻的最优转移矩阵求解如式(8)所示：

$$\boldsymbol{\mu}^*=\left(\sum_{t=1}^{T_w}x_t v_t x_t^{\mathrm{T}}\right)^{-1}\left(\sum_{t=1}^{T_w}x_t v_t x_t'^{\mathrm{T}}\right) \tag{8}$$

在得到转换矩阵$\hat{\boldsymbol{\mu}}$后，便可以将旧空间特征转换到共享特征空间，得到转换后的样本x'_{t+1}. 当特征空间发生变化后，利用转换矩阵$\hat{\boldsymbol{\mu}}$将新特征空间下的数据样本转换为基于旧特征重构的新样本再重新进行聚类，根据聚类结果更新微簇. 其中，对于没有对应数据的微簇，其重要性系数将不断衰减，如果重要性系数降为0，该微簇将被移除；对于具有对应数据的微簇，其聚类特征将得到更新并且重要性系数将得到提升；此外，还有数据难以被聚类到合适的微簇中，那么将新增微簇，但该微簇所属类别还需要概念演化检测器识别. 通过上述转换，特征空间与微簇将统一被转换到新特征空间中，实现特征空间对齐，既避免了特征空间变化对模型预测表现的影响，同时使模型更关注于新特征学习，有助于更好地识别新特征空间下样本类别.

综上所述，当特征空间发生变化，特征维度转换如图3所示. 假设t时刻特征空间发生了变化，有l_1个特征消失，因此先学习特征空间R^d和共享特征空间$\mathbf{R}^{d-l_1}$之间的线性映射系数矩阵$\hat{\boldsymbol{\mu}}$，其次通过转移矩阵$\hat{\boldsymbol{\mu}}$将旧微簇转换到共享特征空间. 同时，对于新样本，使得t+1时刻样本特征经转移矩阵$\hat{\boldsymbol{\mu}}$映射，包含消失的特征中所携带的旧特征空间信息，然后利用维度扩充得到新特征空间下的样本.

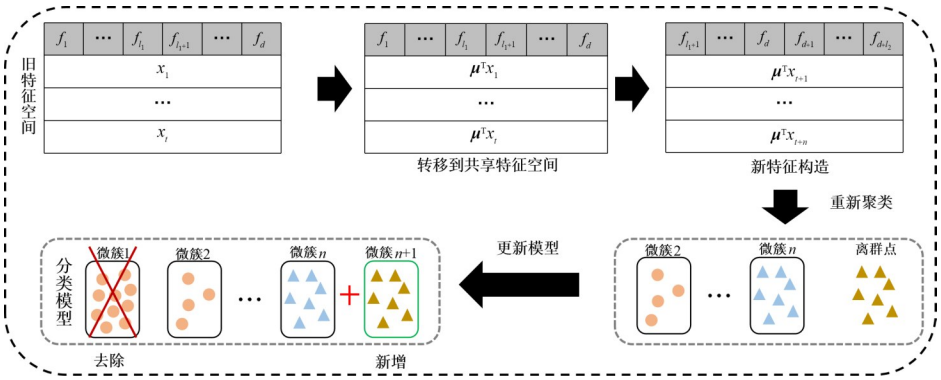


图3 空间转化过程图

3.4 特征相似性度量与映射

当新样本x_t到达并通过特征空间检测与处理后，新样本和微簇都将被转换到新的特征空间. 在转换后的样本空间中，CD_OF将对当前批次样本进行概念演化检测，以观察是否存在新类别样本. 本文利用微簇集成模型判断新样本是否为异常点，主要通过比较样本到每个微簇中心的距离与微簇半径的大小关系，对样本进行区分. 具体地，若新样本到每个微簇中心的距离大于微簇半径，则认为该样本是异常点，存入临时缓冲区B中；反之，则认为该样本属于已知类，对其进行分类，标签获取如式(9)所示：

$$y=\underset{i\in\{1,2,\cdots,k\}}{\arg\max}\frac{W_i N_i}{D(x_i,\mathrm{MC}_{c_i})} \tag{9}$$

其中，W_i是第i个微簇的重要性，N_i是第i个微簇中所含样本数量，$D\left(x_i,\mathrm{MC}_{c_i}\right)$是样本到第$i$个微簇中心的距离. 对于已知类样本，选择微簇集成中距离该样本最近的k个微簇，根据这k个微簇所含样本数量、微簇重要性以及样本到微簇中心的距离这三个指标综合判断样本标签，并更新相应微簇.

对于缓冲区中的异常点，当缓冲区B满时，通过对缓冲区内样本检测以判断是否出现新类，即是否存在概念演化. 本文基于新类实例是一组具有较强内聚性的相似数据假设，根据样本的共享邻域信息定义样本间的相似度，通过与已知类样本的相似度进行比较，从而判断样本是否属于新的类别. 其中，样本x_i与x_j的共享邻域[32]是指x_i与x_j的k近邻交集，定义如式(10)所示：

$$\mathrm{SNN}(x_i,x_j)=\mathrm{KNN}(x_i)\cap\mathrm{KNN}(x_j) \tag{10}$$

其中，$\mathrm{KNN}(x_i)$是x_i的一组最近邻，$\mathrm{KNN}(x_j)$是x_j的一组最近邻.

根据共享邻域，计算样本x_i与x_j的相似度，计算如式(11)所示：

$$\mathrm{Sim}(x_i,x_j)=\begin{cases}\mathrm{e}^{-\sum\limits_{q\in\mathrm{SNN}(x_i,x_j)}(d_{iq}+d_{jq})}\times\mathrm{e}^{\frac{\left|\mathrm{SNN}(x_i,x_j)\right|}{k}}, & \text{if } i,j\in\mathrm{SNN}(x_i,x_j)\\ 0, & \text{otherwise}\end{cases} \tag{11}$$

其中，d_{iq}表示样本x_i与共享近邻x_q之间的欧氏距离. 当x_i与x_j是彼此的k近邻时才会计算相似度，样本点之间的距离越近，不同样本所属空间的相似度越高. $\left|\mathrm{SNN}(x_i,x_j)\right|$是样本$x_i$与$x_j$共享邻域中样本点的个数，共

享邻域中样本点个数越多,样本之间的相似度越高.

若缓冲区中的一个样本与其在缓冲区中的共享近邻相似度高,但是在已知类样本中的共享近邻相似度低,则认为该点是新类实例. 具体计算如式(12)所示:

$$\gamma(x_i)=\frac{\sum\limits_{x_j\in \mathrm{KNN}(x_i)}\mathrm{Sim}(x_i,x_j)-\sum\limits_{x_p\in \mathrm{KNN}(x_i)}\mathrm{Sim}(x_i,x_p)}{\max\left(\sum\limits_{x_j\in \mathrm{KNN}(x_i)}\mathrm{Sim}(x_i,x_j),\sum\limits_{x_p\in \mathrm{KNN}(x_i)}\mathrm{Sim}(x_i,x_p)\right)} \tag{12}$$

其中,x_j是x_i在缓冲区中的最近邻,x_p表示x_i在微簇中的最近邻. $\gamma(x_i)$的取值范围是(−1,1),当$\gamma(x_i)>0$时,该样本为潜在的新类实例. 此时,CD_OF将得到的新类实例移入新类实例缓冲区中,过滤掉缓冲区中的所有新类实例点,剩余的异常点属于已知类.

若对于某个微簇,较长时间内没有新进样本落入该微簇决策边界内,则认为该微簇的重要性随着时间的推移而降低,即不能够捕获数据流中的最新概念. 当微簇的重要性近似达到0时,集成模型将消除这些过时的微簇. 具体地,使用应用指数衰减函数来计算微簇的重要性,定义如式(13)所示:

$$W=W\times \mathrm{e}^{-\lambda\frac{\sum_{j=1}^{N-1}(t_{j+1}-t_j)}{N}} \tag{13}$$

其中,W的初始值设置为1,t_{j+1}和t_j分别是微簇中第$j+1$和第j个样本出现的时刻,λ是重要性衰减参数,N表示该微簇中的样本个数. 随着该微簇中数据出现的平均时间间隔越来越大,该微簇的重要性将逐渐降低. 当W趋向于0时,从模型中消除该微簇.

3.5　CD_OF算法

CD_OF的算法伪代码如算法1所示.

算法1　面向开放特征空间的概念演化检测算法(CD_OF)

初始化:流数据$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_t,y_t),\cdots\}$;弱监督集成模型$\mathrm{MC}=\{\mathrm{MC}_1,\mathrm{MC}_2,\cdots,\mathrm{MC}_{L\times n}\}$

```
1. while 流数据未结束,第t个时刻样本x_t进入
2.     while 对于新进入的样本x_t
3.         检测新特征空间是否发生改变,S_{t-1}与S_t是否相同
4.         if S_{t-1} ≠ S_t
5.             根据式(8)求解转移矩阵μ
6.             将微簇与当前样本特征转换到新特征空间
7.         end if
8.     end while
9.     if d(x_t, MC_c) < MC_r
10.        根据式(9)对已知类进行分类
11.    else
12.        B = B ∪ x_t
13.        if |B| > T_0
14.        根据式(11)计算相似度 Sim(x_i, x_j)
15.        根据式(12)计算x_t是新类的可能性γ(x_t)
16.            if γ(x_t) > 0
17.                判定x_t属于新类
18.            end if
19.        end if
20.        根据式(13)计算指数衰减函数W
21.        if  W → 0
22.            消除微簇MC_i
23.        end if
24.    end if
25. end while
```

3.6　时间复杂度分析

本文中时间复杂度主要由2个部分构成,分别是局部线性回归和簇重新分配. 其中,在局部线性回归中,由于矩阵直接求逆具有较高的时间复杂度,因此本文先通过LU分解法(Lower-Upper Decomposition)将矩阵分解为上三角矩阵和下三角矩阵,然后利用三角矩阵求解逆矩阵. 虽然时间复杂度仍受特征空间维度d影响为$O(d^3)$,但在实际在线学习过程中,算法的计算效率可以得到良好的控制. 首先,矩阵分解过程可以在特征发生变化前做好准备,从而避免对在线学习过程的影响. 具体来说,本文中在特征变化前使用缓冲窗口T_w存储未变特征,然后使用窗口中特征求解分解结果,求解后的结果将被存储在内存当中,在特征空间发生变化后使用内存中的求解结果直接完成求逆运算,具有更高的计算效率. 其次,在在线学习过程中数据的特征维度d较小,而对于高维数据可通过特征表示方法控制特征维度大小,从而避免对实时预测过程的影响. 簇重新分配即簇更新过程,由于需要使用K-means方法重新聚类,因此时间负载度与批次大小b和聚类数$L\times n$相关,而其余统计量的计算时间复杂度均为线性,仅与批次大小相关,因此时间复杂度为$O(b+Lnb)$.

4　实验与性能分析

为验证CD_OF的处理性能,本文在不同的含概念演化的标准数据集和真实数据集上进行实验验证,并与基于集成学习的CLAM[20]方法、基于霍夫丁树的Aht-NODE[24]方法、基于完全随机树的SENCForest[23]方法、基于半监督学习的OSSL[33]方法、基于聚类的EMC[17]方法和ECSMiner[14]方法进行对比.

CD_OF的微簇初始化过程中,每个类别下的聚类数量n为25,KNN近邻参数k初始化为3,缓冲区大小$|B|$初始化为100,训练数据集数量为1 000,标签总数为2. 因此在在线学习之前,模型的已知类别仅为2,在线学习过程中新类别样本将逐渐按序到达.

4.1 数据集

本文使用10个数据集进行评估,每个数据集属性如表1所示.

表1 实验数据集

数据集	属性维数	类别数	样本数	特征空间是否变化
IoT botnet	115	11	663 795	否
Forest cover	54	7	581 012	否
Kddcup99	41	23	494 021	否
Shuttle	9	7	58 000	是
Electricity	8	2	45 312	是
Poker Hand	10	10	11 250	是
Avila	10	12	20 867	是
HAR	561	6	10 299	否
Mnist	784	10	10 000	否
Page Blocks	10	5	5 473	是

IoT botnet数据集收集自真实交通数据,包含9个商业IoT设备的收集结果. 每种设备结果包含由2种僵尸网络干扰的10种攻击类型流数据和常规场景中的流数据,拥有115种属性和11种类别. 本文选用了663 795个样本.

Forest cover数据集收集自7种30 m × 30 m网格中的森林覆盖类型数据,包含54种属性和7种类别,共581 012个样本.

Kddcup99数据集收集自MIT Lincoln实验室的LAN网络流量中的TCP连接,包含23种类别和41种属性,本文选择该数据集10%版本,共494 021个样本.

Shuttle数据集包含9种属性和7种类别,共58 000个样本. 其中约有80%的数据属于类别1.

Electricity数据集收集自电力市场真实数据,包含8种属性和2种类别,共45 312个样本.

Poker Hand数据集收集自由52张标准扑克牌中随机抽取5张组成的手牌,每张手牌包含2种属性花色与点数,包含10个属性和10种类别. 本文选择11 250个样本进行测试.

Avila[34]数据集收集自800张图片并已经过标准化处理,包含10种属性和12种类别,共20 867个样本.

HAR数据集收集自30名实验人员日常活动数据,包含561种属性和6种类别,共10 299个样本.

Mnist数据集收集自250名实验人员手写数字图片,图片包含28 × 328共784种特征和10种类别,本文选择10 000个样本进行测试.

Page Blocks数据集收集自54篇不同的文档,包含10种属性和5种类别,共5 473个样本.

本文所使用的数据集均来自真实数据. 为了评估CD_OF在动态特征空间场景下的表现,本文选择Shuttle、Electricity、Poker Hand、Avila和Page Blocks数据集进行模拟. 为了模拟开放特征空间中特征随时间的出现与消失,本文对数据集进行了预处理. 首先将数据集平均划分为多个块,然后采用随机选择方法对特征进行处理. 通过随机选择方法确保每个块中特征组合方式不同,因此部分特征可能在后续块中消失或重新出现,新块中可能包含模型从未见过的新特征. 具体来说,块A和块$A+1$中的特征均从原始特征空间中随机选择,在块A中被选中特征而在块$A+1$中未被选中特征的情景模拟了特征的消失,反之,在块A中未被选中特征而在块$A+1$中被选中特征的情景模拟了新特征的出现. 随机选择方法能够有效模拟真实环境中传感器特征失效、数据采集通道变化或信息通路重构等现象,具有更强的现实相关性和动态复杂性. 非动态特征空间下使用的数据集特征维度不发生变化,旨在单独测试CD_OF中概念演化检测方法和模型适应表现.

4.2 评价指标

为验证所提CD_OF算法的性能,本文采用了3种指标从新类检测情况、模型的测试精度2个方面对模型进行评估,具体指标如下所述.

M_{new}:新类检测指标,新类实例被错误分类为已知类的百分比[35]. 具体形式如式(14)所示:

$$M_{\text{new}}=\frac{F_{\text{exi}}}{T_{\text{nov}}+F_{\text{exi}}} \tag{14}$$

其中,F_{exi}表示模型错误地将新类识别为已知类样本的数目,T_{nov}表示模型正确识别新类样本的数目.

F_{new}:新类检测指标,表示已知类实例被错误分类为新类的百分比[35]. 具体形式如式(15)所示:

$$F_{\text{new}}=\frac{F_{\text{nov}}}{F_{\text{nov}}+T_{\text{exi}}} \tag{15}$$

其中,F_{nov}表示模型错误地将已知类识别为新类样本的数目,F_{exi}表示模型正确识别已知类样本的数目.

Err:模型分类指标,通过计算模型错误识别的样本数目占总样本数目的比例表示累积错误率. 具体形式如式(16)所示:

$$\text{Err}=\frac{N-T_{\text{nov}}-T_{\text{exi}}}{N} \tag{16}$$

其中,N表示总样本数目.

4.3 实验结果与分析

本文分别从特征空间固定和特征空间变化2种环境进行分析研究. 本文所有实验采用的计算机CPU为Intel(R) Core(TM) i7-14700,内存为64 GB,操作系统为Windows 11,编程工具为Matlab.

4.3.1 特征空间不变场景

图4和图5分别为对比方法在特征空间固定不变数据集上得到的M_{new}和F_{new}测试结果. 通过图4对比实验结果,我们可以看到在M_{new}指标上,本文方法与其他对比

方法相比具有较低的M_{new}值,大多数情况下小于20%,这说明CD_OF在检测概念演化的过程中,能够以较高的准确性识别新类别的数据,避免将其错误归类为已知类别,从而有助于模型及时适应新的数据分布. 对于F_{new}指标,本文方法在这一指标上的结果都小于5%,这表明本文方法能够有效地区分已知类的实例,避免将它们错误地归类为新类,对已知类具有很好的分类效果. 上述结果表明,本文方法不仅能够有效地检测和识别新类别样本,还能够保持对已知类别样本的准确分类.

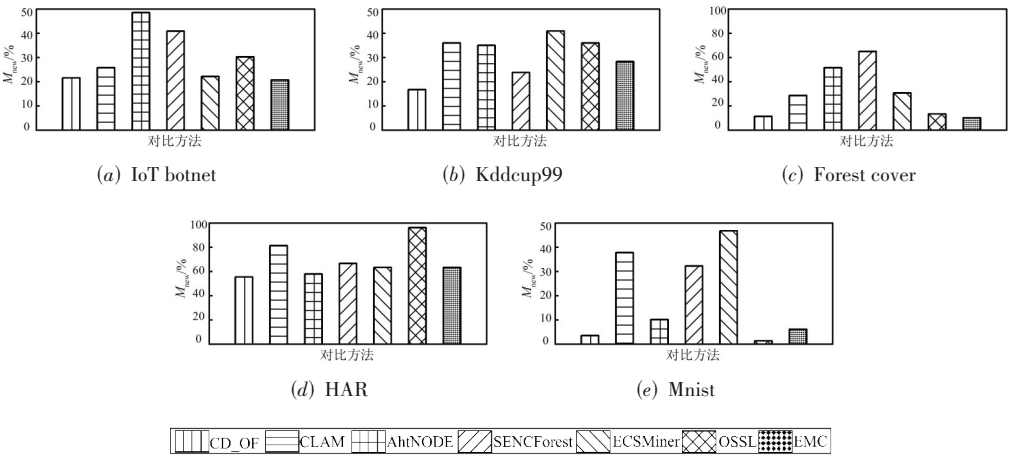


图4　在特征空间固定数据集上的M_{new}指标实验结果

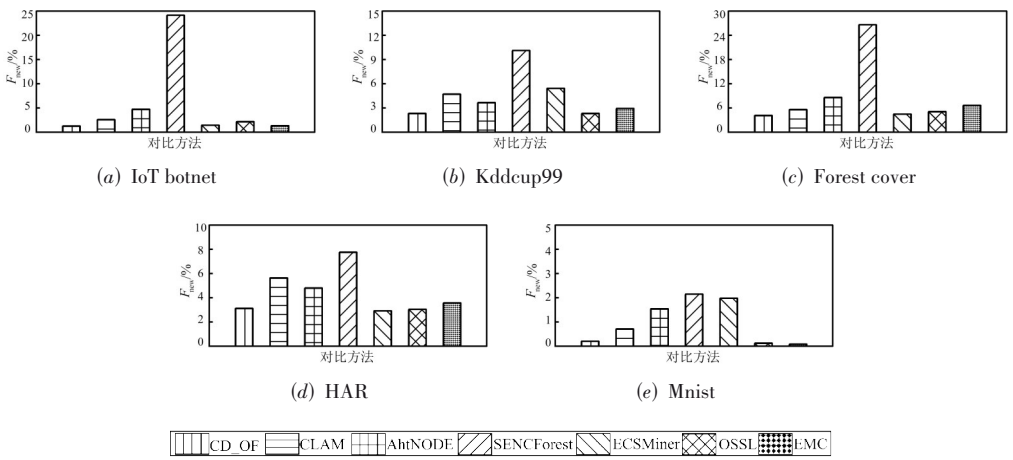


图5　在特征空间固定数据集上的F_{new}指标实验结果

图6为对比方法在特征空间固定数据集上得到的累积错误率Err值. 图6中的实验结果表明本文方法的累积错误率明显低于大部分对比方法,且在与其他对比方法的比较中也表现出微弱的优势,略低于其他对比方法. 这表明本文方法能维持较为稳定的性能并保持较为准确的分类效果.

4.3.2　特征空间变化场景

图7和图8分别为对比方法在特征空间变化数据集上得到的M_{new}和F_{new}测试结果. 图7和图8的实验结果展示了与对比方法相比,CD_OF的M_{new}值和F_{new}值都低于对比方法. 实验结果说明即使在特征空间发生变化时,本文方法将新类样本错误归类为已知类的概率相对较低,且能够较好地对已知类样本进行分类,不会将其错误地检测为新类别. 上述结果表明本文方法能够较好地处理开放特征空间下的流数据概念演化检测问题,在检测新类实例方面表现出色.

图9为对比方法在特征空间变化数据集上得到的累积错误率Err测试结果. 从图9中可以发现,CD_OF的累积错误率明显低于其他对比方法,主要是因为CD_OF能够及时检测到特征空间的变化,并有效地学习旧特征与共享特征之间的映射关系,将旧特征中携带的信息迁移到共享特征中,保证了模型性能的稳定性. 对比方法缺乏对特征空间变化的快速响应能力,无法迁移旧特征中的信息到共享特征空间. 因此,当面对

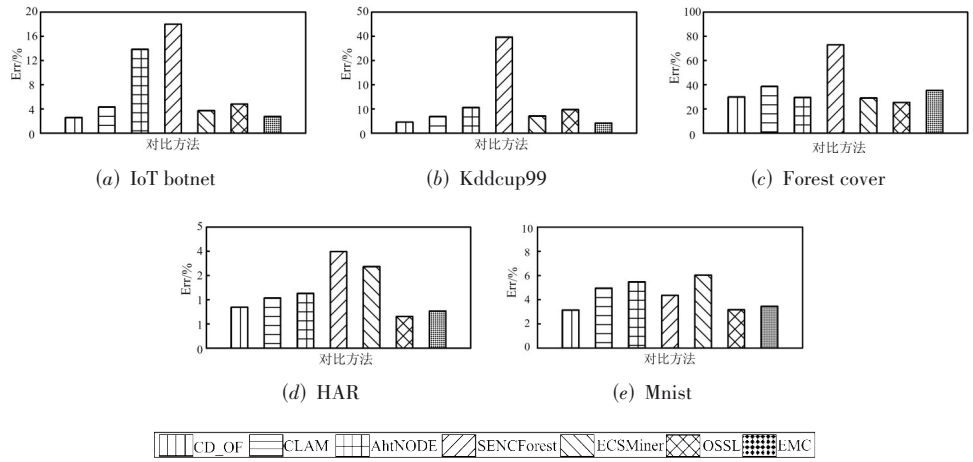


图6 在特征空间固定数据集上的Err指标实验结果

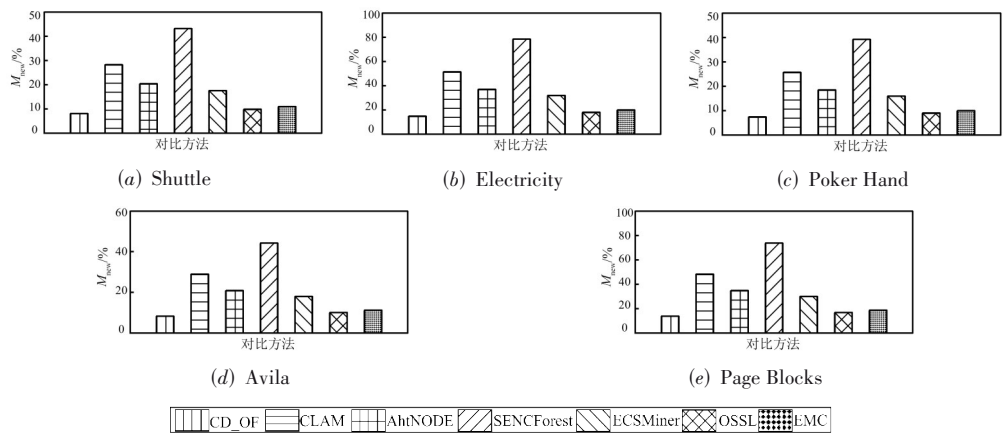


图7 不同方法在各数据集上的M_{new}比较

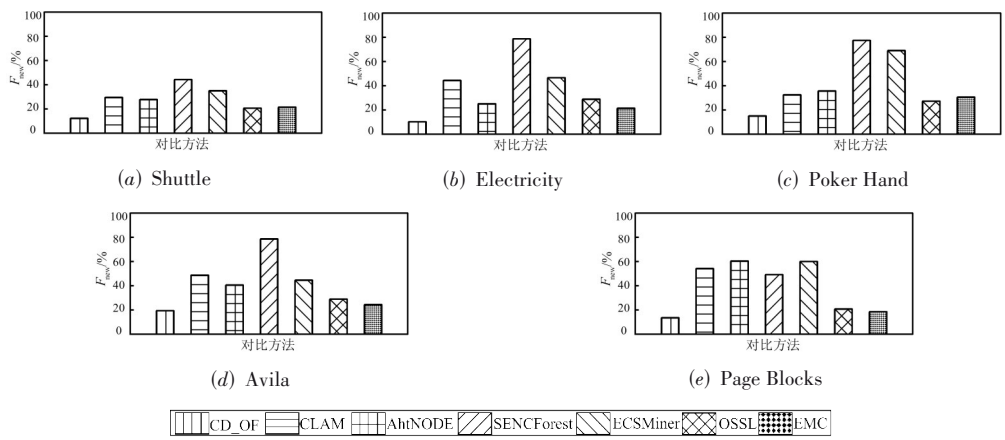


图8 不同方法在各数据集上的F_{new}比较

特征空间变化的数据流时,对比方法的累积错误率会显著增加. 本文提出的方法确保了模型能够在开放特征空间中维持高水平的性能表现,提高了模型对开放特征空间的适应能力.

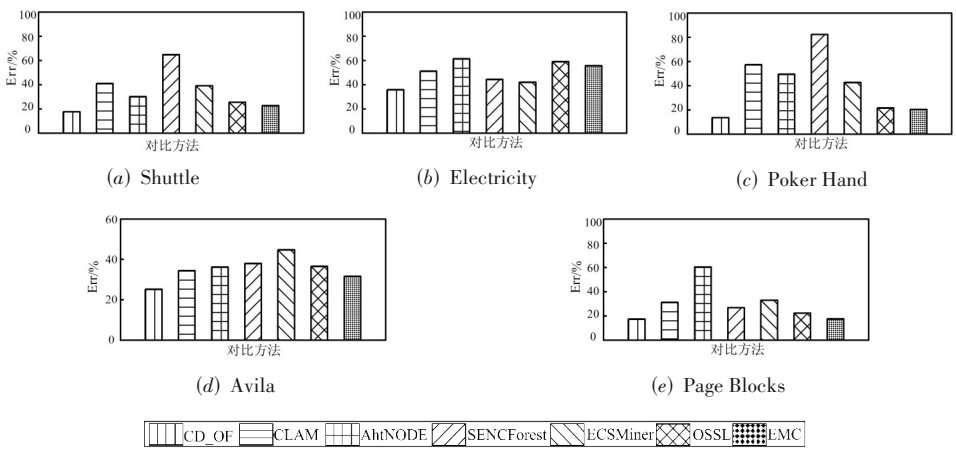
4.3.3 参数敏感性分析

在参数n敏感性实验中,设置了步长为5,从10~50的对比簇数. 实验结果如图10(a)所示. 当n=25时,在所有数据集上的平均错误率最低. 当n采用较小的数值时,难以表示类内样本之间的差异;而当n采用较大数值时,模型容易产生过拟合问题,难以应用于后续任务中.

在参数k敏感性实验中,设置了步长为1,从1~5的近邻数. 实验结果如图10(b)所示. 当k=3时,在所有数

据集上的平均错误率最低. 合适的k值对分类过程中选择微簇、微簇参数量更新和新类检测过程中共享近邻判断紧密相关. 通过上述实验,当k选择3时效果最佳,过大或过小的k取值均会对模型预测性能造成较大的影响.

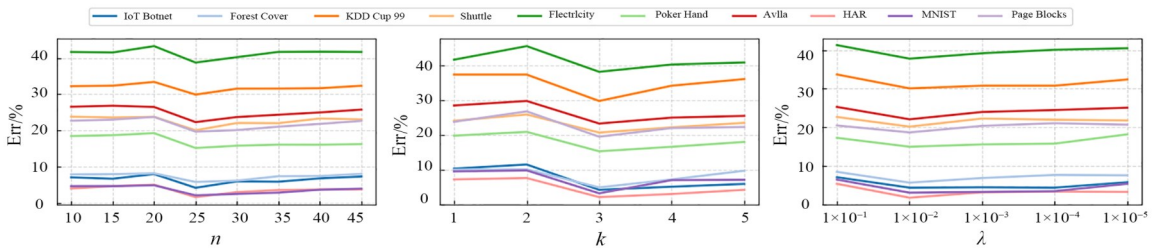
参数λ是微簇重要性衰减因子. 在敏感性实验中,我们选择了0.1,0.01,0.001,0.000 1进行对比.λ的大小表示未获取样本对微簇重要性的影响,与重要性系数的衰退成正比.λ的取值可以反映对历史数据与当前数据对模型的重要程度. 实验结果如图10(c)所示.与其余超参数相比,λ取值的变化对模型的影响较小,在λ=0.01时,模型的表现最佳.



(a) Shuttle　(b) Electricity　(c) Poker Hand

(d) Avila　(e) Page Blocks

图9　不同方法在各数据集上的Err比较



(a) 聚类数n　(b) 近邻数k　(c) 重要性衰减系数λ

图10　参数敏感性实验

4.3.4　数据内聚性分析

新类数据的轮廓系数和Davies-Bouldin指数进行计算对数据内聚性表示进行定量分析,实验结果如表2所示. 虽然一些数据集的内聚性指标低于其余数据,如实验结果在Avila数据集的表现低于Electricity数据集,但是2个数据集的错误率并没有明显差别. 这主要是因为CD_OF在每个类别中进一步聚类得到微簇能够更精确地表示数据,降低同类样本间内聚性不够强的影响.

表2　数据集内聚性测试

数据集	轮廓系数	Davies-Bouldin指数
IoT botnet	−0.46	5.56
Forest cover	−0.09	9.69
Kddcup99	0.04	5.29
Shuttle	0.39	1.05
Electricity	0.61	0.58
Poker Hand	0.08	3.30
Avila	0.01	4.93
HAR	0.06	3.48
Mnist	0.05	3.76
Page Block	−0.23	1.89

5　结论

针对传统基于静态非开放特征空间下的概念演化检测难以满足现实场景中特征维度变化情况下的概念演化检测与适应需求,本文提出一种面向开放特征空间的概念演化检测方法CD_OF. 该方法首先通过将变化后的特征空间转换到共享特征空间以对齐新旧数据分布,然后通过补齐维度从而使模型快速学习补齐后的特征,有效缓解开放特征空间动态变化的严重影响. 在此基础上,通过度量特征相似性和共享近邻覆盖范围2种信息共同判断样本点是否属于新样本,提高了模型对有较强内聚性的新类样本的检测能力. 然后,通过动态衰减模型,实现了模型对类消失和类循环数据的适应能力,增强了开放特征空间下概念演化检测的泛

化能力.

本文方法对解决开放特征空间下的概念演化检测问题进行了积极的探索. 然而,在开放特征空间中数据及特征存在更多样式的变化类型,未来我们将对这些变化情况进行深入分析从而提升已有方法的适应能力.

参考文献

[1] RAMZAN F, AYYAZ M. A comprehensive review on data stream mining techniques for data classification; and future trends[J]. EPH - International Journal of Science and Engineering, 2023, 9(3): 1-29.

[2] LU J, LIU A J, DONG F, et al. Learning under concept drift: A review[J]. IEEE Transactions on Knowledge and Data Engineering, 2019, 31(12): 2346-2363.

[3] LI J P, YU H, ZHANG Z Y, et al. Concept drift adaptation by exploiting drift type[J]. ACM Transactions on Knowledge Discovery from Data, 2024, 18(4): 1-22.

[4] 杜航原, 王文剑, 白亮. 一种基于优化模型的演化数据流聚类方法[J]. 中国科学: 信息科学, 2017, 47(11): 1464-1482.
DU H Y, WANG W J, BAI L. A novel evolving data stream clustering method based on optimization model[J]. Scientia Sinica (Informationis), 2017, 47(11): 1464-1482. (in Chinese)

[5] KSIENIEWICZ P, ZYBLEWSKI P. Stream-learn: Open-source Python library for difficult data stream batch analysis[J]. Neurocomputing, 2022, 478: 11-21.

[6] CACCIARELLI D, KULAHCI M. Active learning for data streams: A survey[J]. Machine Learning, 2024, 113(1): 185-239.

[7] 翟婷婷, 高阳, 朱俊武. 面向流数据分类的在线学习综述[J]. 软件学报, 2020, 31(4): 912-931.
ZHAI T T, GAO Y, ZHU J W. Survey of online learning algorithms for streaming data classification[J]. Journal of Software, 2020, 31(4): 912-931. (in Chinese)

[8] KLIKOWSKI J. Concept drift detector based on centroid distance analysis[C]//2022 International Joint Conference on Neural Networks. Piscataway: IEEE, 2022: 1-8.

[9] LI X J, ZHOU Y, JIN Z Y, et al. A classification and novel class detection algorithm for concept drift data stream based on the cohesiveness and separation index of mahalanobis distance[J]. Journal of Electrical and Computer Engineering, 2020, 2020(1): 4027423.

[10] 韩光洁, 赵腾飞, 刘立, 等. 基于多元区域集划分的工业数据流概念漂移检测[J]. 电子学报, 2023, 51(7): 1906-1916.
HAN G J, ZHAO T F, LIU L, et al. Concept drift detection of industrial data flow based on multivariate region set partition[J]. Acta Electronica Sinica, 2023, 51(7): 1906-1916. (in Chinese)

[11] 代劲, 李昊, 王国胤. 基于动态样本选择的概念漂移自适应预测方法[J]. 电子学报, 2024, 52(9): 3228-3239.
DAI J, LI H, WANG G Y. Concept drift adaptive prediction method based on dynamic sample selection[J]. Acta Electronica Sinica, 2024, 52(9): 3228-3239. (in Chinese)

[12] ZUBAROĞLU A, ATALAY V. Online embedding and clustering of evolving data streams[J]. Statistical Analysis and Data Mining: the ASA Data Science Journal, 2023, 16(1): 29-44.

[13] 王婧. 基于集成学习的概念演化检测方法研究[D]. 太原: 山西大学, 2024.
WANG J. Research on Concept Evolution Detection Method Based on Ensemble Learning[D]. Taiyuan: Shanxi University, 2024. (in Chinese)

[14] 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.

[15] 王婧, 郭虎升, 王文剑. 基于弱监督集成的概念演化自适应检测方法[J]. 吉林大学学报(信息科学版), 2024, 42(3): 406-420.
WANG J, GUO H S, WANG W J. Adaptive detection method for concept evolution based on weakly supervised ensemble[J]. Journal of Jilin University (Information Science Edition), 2024, 42(3): 406-420. (in Chinese)

[16] GUO H S, XIA H S, LI H, et al. Concept evolution detection based on noise reduction soft boundary[J]. Information Sciences, 2023, 628: 391-408.

[17] DIN S U, SHAO J M. Exploiting evolving micro-clusters for data stream classification with emerging class detection[J]. Information Sciences, 2020, 507: 404-420.

[18] GARCIA K D, DE FARIA E R, DE SÁ C R, et al. Ensemble clustering for novelty detection in data streams[C]// Discovery Science. Cham: Springer, 2019: 460-470.

[19] ZHENG X L, LI P P, HU X G, et al. Semi-supervised classification on data streams with recurring concept drift and concept evolution[J]. Knowledge-Based Systems, 2021, 215: 106749.

[20] AL-KHATEEB T, MASUD M M, AL-NAAMI K M, et al. Recurring and novel class detection using class-based ensemble for evolving data stream[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(10): 2752-2764.

[21] GAO Y, CHANDRA S, LI Y F, et al. SACCOS: A semi-

supervised framework for emerging class detection and concept drift adaption over data streams[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(3): 1416-1426.

[22] BOUGUELIA M R, NOWACZYK S, PAYBERAH A H. An adaptive algorithm for anomaly and novelty detection in evolving data streams[J]. Data Mining and Knowledge Discovery, 2018, 32(6): 1597-1633.

[23] MU X, TING K M, ZHOU Z H. Classification under streaming emerging new classes: A solution using completely-random trees[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(8): 1605-1618.

[24] GANDHI J, GANDHI V. Novel class detection with concept drift in data stream - AhtNODE[J]. International Journal of Distributed Systems and Technologies, 2020, 11(1): 15-26.

[25] 赵鹏, 周志华. 基于决策树模型重用的分布变化流数据学习[J]. 中国科学: 信息科学, 2021, 51(1): 1-12.
ZHAO P, ZHOU Z H. Learning from distribution-changing data streams via decision tree model reuse[J]. Scientia Sinica (Informationis), 2021, 51(1): 1-12. (in Chinese)

[26] ZHANG Z L, LI Y, ZHANG Z W, et al. Adaptive matrix sketching and clustering for semisupervised incremental learning[J]. IEEE Signal Processing Letters, 2018, 25(7): 1069-1073.

[27] GUAN S U, LI S C. Incremental learning with respect to new incoming input attributes[J]. Neural Processing Letters, 2001, 14(3): 241-260.

[28] HOU B J, ZHANG L J, ZHOU Z H. Learning with feature evolvable streams[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(6): 2602-2615.

[29] 刘兆清, 古仕林, 侯臣平. 面向特征继承性增减的在线分类算法[J]. 计算机研究与发展, 2022, 59(8): 1668-1682.
LIU Z Q, GU S L, HOU C P. Online classification algorithm with feature inheritably increasing and decreasing[J]. Journal of Computer Research and Development, 2022, 59(8): 1668-1682. (in Chinese)

[30] HE Y, WU B J, WU D, et al. Online learning from capricious data streams: A generative approach[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, 2019: 2491-2497.

[31] LIAO G B, ZHANG P, YIN H P, et al. A novel semi-supervised classification approach for evolving data streams[J]. Expert Systems with Applications, 2023, 215: 119273.

[32] LIU R, WANG H, YU X M. Shared-nearest-neighbor-based clustering by fast search and find of density peaks[J]. Information Sciences, 2018, 450: 200-226.

[33] UD DIN S, SHAO J M, KUMAR J, et al. Online reliable semi-supervised learning on evolving data streams[J]. Information Sciences, 2020, 525: 153-171.

[34] DE STEFANO C, MANIACI M, FONTANELLA F, et al. Reliable writer identification in medieval manuscripts through page layout features: The "Avila" Bible case[J]. Engineering Applications of Artificial Intelligence, 2018, 72: 99-110.

[35] ZAREMOODI P, KAMALI SIAHROUDI S, BEIGY H. Concept-evolution detection in non-stationary data streams: A fuzzy clustering approach[J]. Knowledge and Information Systems, 2019, 60(3): 1329-1352.

作者简介

苏　睿　男,1999年1月出生于山西省忻州市. 现为山西大学计算机与信息技术学院博士研究生. 主要研究方向为流数据挖掘、机器学习、深度学习.
E-mail: 202212407015@email.sxu.edu.cn

王　婧　女,1999年8月出生于山西省吕梁市. 现为山西大学计算机与信息技术学院硕士研究生. 主要研究方向为流数据挖掘.
E-mail: 1756645158@qq.com

郭虎升　男,1986年10月出生于山西省晋中市. 现为山西大学教授、博士生导师. 主要研究方向为数据挖掘、机器学习.
E-mail: guohusheng@sxu.edu.cn

王文剑　女,1968年10月出生于山西省太原市. 现为山西大学计算智能与中文信息处理教育部重点实验室副主任、教授、博士生导师. 主要研究方向为机器学习、数据挖掘.
E-mail: wjwang@sxu.edu.cn