基于多元区域集划分的工业数据流概念漂移检测

韩光洁,赵腾飞,刘 立,张 帆,徐政伟
(河海大学信息学部物联网工程学院,江苏常州 213022)

摘 要: 为了快速适应非平稳环境中工业数据流的分布变化,需要在非结构化和噪声干扰的数据中准确、实时的完成概念漂移的检测. 本文提出了一种基于多元区域集划分的工业数据流概念漂移检测算法(Concept Drift detection-Multivariate region set Partition,CDMP). 首先基于实例模糊密度进行多元区域集划分,根据划分的若干模糊分区集合,识别概念漂移发生的区域. 概念漂移的持续发生会显著降低基于多元区域集构建的模型的分类性能,CDMP通过构建多元历史模型池来保留具有多样性的历史模型,以降低模型调整或再训练造成的性能损耗,同时保证概念漂移检测中准确性. CDMP在不同数据集上进行了性能测试. 实验结果表明,CDMP实现了对历史模型多样性的保留和重用,能够在不同噪声水平的工业物联网环境中实现对重现型、突发型等多类型概念漂移的准确检测.

关键词: 工业物联网;概念漂移;多元区域集;实例模糊分区;多样性历史模型

基金项目: 国家自然科学基金(No.62002099);江苏省自然科学基金(No.BK20200184);常州市科技项目(No.CJ20220052);机器人学国家重点实验室联合开放基金(No.2022-KF-22-10)

中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2023)07-1906-11

电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20221362

Concept Drift Detection of Industrial Data Flow Based on Multivariate Region Set Partition

HAN Guang-jie, ZHAO Teng-fei, LIU Li, ZHANG Fan, XU Zheng-wei
(*College of Internet of Things Engineering*, *Hohai University*, *Changzhou*, *Jiangsu* 213022, *China*)

Abstract: To adapt to the rapidly changing distribution patterns generated in non-stationary industrial environments, it has become necessary to accurately and timely detect concept drift in unstructured and noisy data streams. In this study, a concept drift detection-multivariate region set partition (CDMP) algorithm for industrial data streams is proposed. The CDMP algorithm first performs multivariate region set partition based on the fuzzy density of data instances, and identifies the region in which concept drift occurs through a set of fuzzy partitions. The persistent occurrence of concept drift can significantly degrade the classification performance of models built on multivariate region sets. To address this issue, CDMP builds a historical model pool that retains diverse historical models, thus reducing the performance loss caused by model adjustment or retraining while ensuring the accuracy of concept drift detection. CDMP's performance is tested on different datasets. Experimental results show that CDMP preserves and reuses historical models with diversity, and can accurately detect different types of concept drift, including reoccurring and sudden drift, in industrial IoT environments with different levels of noise interference.

Key words: industrial Internet of things; concept drift; multivariate region sets; instance fuzzy partition; diverse history models

Foundation Item(s): National Natural Science Foundation of China (No.62002099); Natural Science Foundation of Jiangsu Province of China (No.BK20200184); Changzhou Foundation of Science and Technology (No.CJ20220052); Open Fund of State Key Laboratory of Robotics (No.2022-KF-22-10)

1 引言

现代工业环境下,设备生成数据的环境易随时间发生变化,导致底层分布在实际应用中不稳定[1],这种由潜在因素影响而导致的目标概念发生改变的现象

收稿日期:2022-11-30;修回日期:2023-04-03;责任编辑:覃怀银

称为概念漂移. 在短时间内识别非平稳环境中数据流的分布变化,完成对概念漂移的快速适应,是确保数据驱动的机器学习模型可靠性的关键[2,3]. 目前,围绕概念漂移检测的相关研究聚焦于集成学习策略[4,5],该策略需要创建多个基础模型,通过综合各个基础模型的检测结果,保证概念检测的准确性. 此类研究重点关注模型的时效性,判断概念漂移发生致使模型失效时,启用当前到达的数据新建模型. 但是,每当发生概念漂移便触发全新的模型训练任务,不利于时间敏感的工业系统在新场景下的快速响应[6]. 由于新的概念可由旧的概念在一段持续时间内增量形成,旧的概念甚至可在未来时刻重复出现. 如图1所示,t_0时刻的概念在t_{n-1}时刻又再次出现,t_1时刻出现的概念是由t_0时刻的概念渐变生成. 这表明不同概念间可能存在相似性,历史模型可能具备一定对当前概念的解释能力. 因此,本文提出一种基于多元区域集划分的工业数据流概念漂移检测算法(Concept Drift detection-Multivariate region set Partition,CDMP),通过对数据实例的多元区域集划分和保留并重用不同概念分布下生成的历史模型,从而实现对概念漂移的检测与自适应.

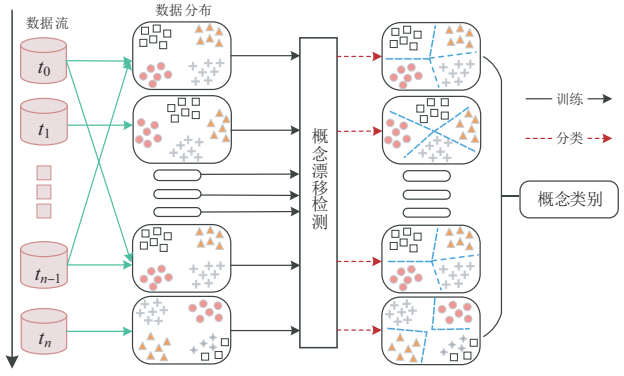

图1　工业数据流中的概念漂移

2　相关工作

2.1　概念漂移概述

概念漂移是动态的工业数据流的固有属性,在通常情况下无法预测[7],其具体表现为,在给定的时间间隔[0,t]内,存在样本$S_{0,t}=\{e_0,e_1,\cdots,e_t\}$,其中,$e_i=(\boldsymbol{X}_i,y_i)$为单个的数据实例,$\boldsymbol{X}_i$为特征向量,$y_i$为类别标签,$S_{0,t}$遵循一个特定的分布$F_{0,t}(X,y)$. 若$F_{0,t}(X,y)\neq F_{t+1,\infty}(X,y)$,则概念漂移将在$t+1$时刻发生. 因此,概念漂移的定义可表示为

$$\exists t: P_t(X,y)\neq P_{t+1}(X,y) \tag{1}$$

根据概念漂移发生的速度,概念漂移可以划分为突变型、渐变型、增量型和重现型四种类型[8].

2.2　多元区域集

多元区域集是由一组不同的数据流相互集合而成,假设存在一组数据流集合$S=\{S_1,S_2,\cdots,S_k\}$,其中集合中数据流的数据分布各不相同,则这种由不同数据分布的数据流一起组成的集合称为一个多元区域集.

多元区域集由于其构成特性,可以最大限度的维护数据的多样性,从而更好的解析出数据实例各部分的空间特征. 针对工业环境下数据流无限性和高噪声的问题,构建多元区域集可以有效控制数据流的边界范围,从噪声环境中筛选出离群的数据实例,从数据层面减少了噪声带来的干扰. 此外,多元区域集还可以对历史环境数据和特征进行保留,提高了数据的复用性.

2.3　概念漂移检测

当前,针对概念漂移的检测算法主要围绕基于主动检测和被动自适应两种策略进行研究.

基于主动检测的漂移检测方法主要通过一系列检测机制对数据流分布的稳定性进行检测. 常见的方法包括基于滑动窗口和基于模型性能的检测等. 文献[9]通过采用不同的滑动窗口有效识别并分析了概念漂移的类型和检测过程的关键信息. 文献[10]利用模糊集合理论,允许滑动窗口保持重叠周期来实现更高的检测精度,并对到来的新数据判断是否符合置信度来推断其知识模式. 文献[11]提出基于非平稳环境的增量学习算法(Learn++. NonStationary Environments,Learn++. NSE),通过基础分类器对新数据预测错误率进行加权来缓解分类器频繁触发的问题. 以上方法虽然能够在数据流非平稳状态下避免不必要的检测,但在模型学习过程中可能会出现概念漂移位点的误检、漏检等情况,这使得在线学习模型的泛化能力较低.

基于被动自适应的漂移检测方法主要通过假设环境随时发生变化,同时对自身模型不断进行调整和学习. 常见的方法包括基于集成学习和基于决策树的方法等. 文献[12]提出一种基于块的动态更新集成算法(Dynamic Updated Ensemble,DUE),解决了概念漂移自适应和不平衡数据流学习问题. 文献[13]使用决策树内部节点的在线错误率来检测实例空间局部区域中的漂移数据,通过叶节点中的漂移检测器指示出数据特征空间中的漂移区域. 文献[14]提出一种自适应随机森林(Adaptive Random Forest,ARF)算法,利用概念漂移警告时训练出一个备用的背景树,从而在发生概念漂移时用背景树替换历史的基础树. 以上方法虽然在一定程度上提高了模型对概念漂移的响应速度,但难以提取重要的历史信息,对历史模型没有做到很好的重用,这可能影响模型的检测效率和收敛性能.

3 基于多元区域集划分的数据流概念漂移检测

如图2所示，CDMP算法主要分为多元区域集划分、历史模型池构建和概念漂移检测三个阶段.

3.1 基于集成学习策略的概念漂移检测与自适应

多元区域划分的整体结构如图3所示. CDMP首先将数据流中的数据实例基于实例模糊密度划分成若干模糊分区集合. 随后，针对划分的多元区域进行概念漂移区域识别，实现概念漂移检测的数据检索，使得基于多元区域集中数据实例构建的机器学习模型能够解析实例各部分的空间特征.

3.1.1 基于密度峰值聚类的实例模糊分区

K-means聚类、高斯混合模型和层次聚类等聚类算法具有速度快、计算简单等优点，但均不能检测出混合度较高、非球型的簇[15]. 基于密度的峰值聚类算法不仅能检测非球型簇，还具有抗噪音特性，适用于任何形状的簇和具有噪声的工业场景[16,17].

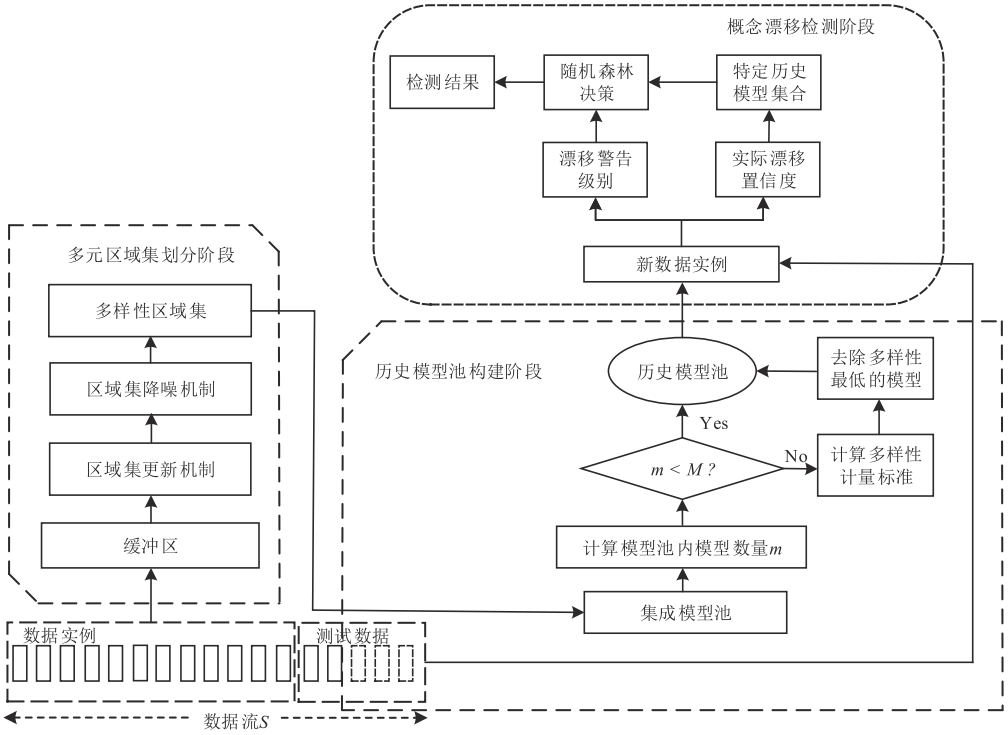

图2 算法框架图

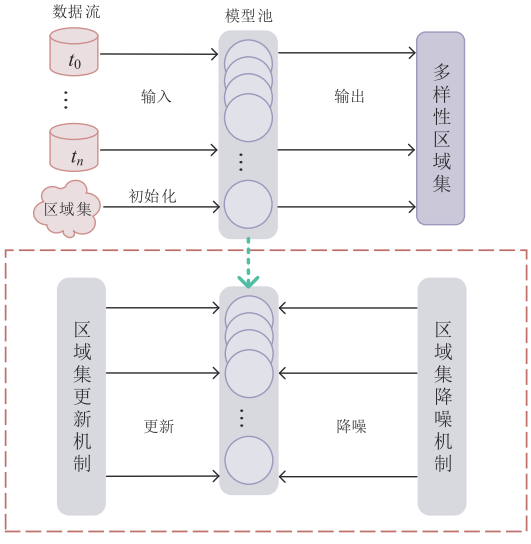

图3 多元区域集分区框架图

CDMP将模糊算子引入到密度峰值聚类算法DPC[18]中，利用模糊计算中的模糊性和不确定性，解决实例的模糊分区问题. 通过计算每个数据实例x_i的局部密度ρ_i及其距离最近的高密度点的距离$\delta_i(i=1,2,\cdots,n,n$为数据实例个数)，确定$\delta-\rho$决策图进而找出聚类中心. 则$\rho_i$和$\delta_i$的计算公式分别为

$$\rho_i=f\left(\sum_{j=1}^{n}\chi\left(d_{ij},d_c\right)\right) \tag{2}$$

$$\delta_i=\min_{j:\rho_i>\rho_j}\left(d_{ij}\right) \tag{3}$$

其中，d_{ij}为x_i和x_j的欧几里得距离，d_c为截止距离，函数f为单调非递减函数，$\chi\left(d_{ij},d_c\right)$可由高斯核函数来代替. CDMP将基于核函数的密度峰值转换成模糊密度峰值. 算子$\oplus$的f对偶算子定义如下：设$f:X\to Y$为一个映射函数，其中，$\oplus$为域中定义的二元运算符，在f的取值范

围内定义唯一模糊算子$\otimes$,使得式(4)成立:

$$f\left(x_1 \oplus x_2\right)=f\left(x_1\right)\otimes f\left(x_2\right) \tag{4}$$

则算子$\otimes$称为算子$\oplus$的f对偶算子.

当定义了函数f,则可确定算子$\oplus$的f对偶算子$\otimes$. 从模糊方法的角度来看式(2),通过使用合适的函数f,可在相应的基于核的密度峰值计算中使用算子$\oplus$的f对偶算子,计算模糊密度峰值. CDMP将密度峰值看作特殊的模糊峰值,即将模糊算子作用于数据实例对应的隶属度函数而获得的模糊耦合的结果,解决区域划分过程中的歧义和不确定性. 因此,由数据实例x_i的相邻实例可得到其模糊密度峰值$\tilde{\rho}_i$的计算结果:

$$\begin{aligned}\tilde{\rho}_i=&f\left(\alpha\left(x_i,x_1\right)\right)\otimes f\left(\alpha\left(x_i,x_2\right)\right)\otimes\\&\cdots\otimes f\left(\alpha\left(x_i,x_k\right)\right)\end{aligned} \tag{5}$$

其中,$f(\cdot)$为关于$\alpha\left(x_i,x_j\right)$的隶属度函数. 当$f$和$\otimes$确定之后,$x_i$的模糊密度可等效视作$\oplus$和$\alpha\left(x_i,x_j\right)$的耦合. 计算每个数据实例的模糊密度后,对模糊密度进行降序排列即可确定模糊峰值. $\alpha\left(x_i,x_j\right)$反映了点$x_i$和$x_j$之间的相似程度,由高斯核确定:

$$\alpha\left(x_i,x_j\right)=\exp\left\{-\frac{\left\|x_i-x_j\right\|}{2\sigma^2}\right\} \tag{6}$$

根据式(3)可知,可以通过引入丰富的模糊算子来解决数据的不确定性以完成数据集的模糊分区. 由于非参数型模糊算子Zadeh和Product在模糊集合合并时隶属度不会发生变化,而参数型的模糊算子引入了权重参数ω,通过控制隶属度来达到控制模糊集合权重的作用,这种方式可以在模糊计算中提供更多的灵活性. 因此,CDMP算法引入其中一种参数型模糊算子Hamacher-S模算子,具体计算如式(7)所示:

$$a\cup b=\frac{a+b-(2-\omega)ab}{1-(1-\omega)ab},\omega\in(0,1) \tag{7}$$

其中,a,b代表不同的模糊集合.

基于模糊密度峰值划分出的区域集合记为$B=\left\{B_1,B_2,\cdots,B_P\right\}$. 剩余的数据实例被划分为与其最近的邻居点相同的区域,即令q_i为$\tilde{\rho}_i(i=1,2,\cdots,n)$的降序排列. 由此在具有比$x_i$更大模糊密度的数据实例中,最接近$x_i$的数据实例数量定义为$n_{q_i}$:

$$n_{q_i}=\begin{cases}\arg\min\left\{d_{q_iq_j}\right\}, & i\geqslant 2\\ 0, & i=1\end{cases} \tag{8}$$

CDMP对所有的数据点进行初始化,即$l_i(i=1,2,\cdots,n)$,有

$$l_i=\begin{cases}j, & x_i\in B,j=1,2,\cdots,p\\ -1, & \text{其他}\end{cases} \tag{9}$$

其中,$l_{q_i}=-1$时,可以根据$l_{q_i}=l_{n_{q_i}}$更新标签信息l_i.

多元区域集的划分方法是基于密度峰值来计算的,其区域中心均局部密度较大,并且距离其他较大密度点的距离较远. 实际情况下,若一个数据集中有多个密度峰值点且其分布相似时,选择合适的区域中心变得较为困难,为避免选择噪声作为区域中心,CDMP计算出距离所有数据最近的较大密度实例的距离的标准偏差,选择离最近大密度实例的距离大于或者等于加权标准偏差的实例作为区域中心. 区域中心E和去除噪声的区域中心R的确定方式如下:

$$E=\delta_i\geqslant\lambda\sigma\left(\delta_i\right) \tag{10}$$

$$R=E\left(\rho_i\right)\geqslant\mu\left(\rho_i\right) \tag{11}$$

其中,δ_i是距最近的较大密度实例的距离,$\sigma\left(\delta_i\right)$是距离所有较大密度实例的距离的标准偏差,$\lambda$为权重,$\rho_i$为局部密度,$\mu\left(\rho_i\right)$是所有局部密度的平均值.

考虑工业数据流中因噪声干扰易出现离群值的情况,CDMP在计算每个数据实例的密度值后,将与最近高密度实例距离较远,但其局部密度小的实例看作离群值噪声,设置阈值参数将其去除,以实现降噪的目的.

3.1.2　多元区域集的概念漂移区域识别

在完成上述的模糊分区后,可以构建出一个基于实例模糊密度峰值的多元区域集合. 然而,如何适应概念漂移以及如何解决模型的持续在线更新仍是一个问题. CDMP从统计学角度出发,通过设定缓冲区对概念漂移场景下的区域集进行自适应更新,使数据实例分区得到快速响应的同时对漂移区域做出精确识别,提高多元区域集在不同类型概念漂移场景下的自适应性.

对于区域集合$B=\left\{B_1,B_2,\cdots,B_P\right\}$,定义用于实例区域划分的$\phi$级区域$B_p^{\phi}$,其中$\phi$为区域中数据的样本比例,即$\Pr\left(x_i\in B_p^{\phi}\right)=\phi$.

根据中心极限定理,假设当前的数据实例数量为m,当m值足够大时,$m^{B_i^{\phi}}$和区域样本比例ϕ的分布遵循近似正态分布. 当m值趋近于无限大时,$m^{B_p^{\phi}}$的近似正态分布的均值和方差分别为$m\cdot\phi$和$m\cdot\phi\cdot(1-\phi)$,与二项分布相同. 类似的,对近似正态分布的区域实例比例来说,均值和方差分别为ϕ和$\phi\cdot(1-\phi)/m$. 但是,近似正态分布对较小的m值来说并不准确,根据经验法则,只有当$m\cdot\phi>10$和$m\cdot(1-\phi)>10$时使用正态分布才足够准确. 即数据实例初始化的数量m应满足:

$$m\geqslant\max\left\{\frac{10}{\phi},\frac{10}{1-\phi}\right\} \tag{12}$$

在数据流持续到达的情况下,如果在一段时间内

数据流未发生概念漂移,则可认为接下来的τ个数据实例无论是位于哪个区域中,均可以视为一个伯努利过程. 因此,CDMP把连续到达区域B_t^{ϕ}的τ个实例定义为

$$D_{t+1,t+\tau}^{B_t^{\phi}}=\left\{d_i: i\in\mathbb{Z}_{t+1\leqslant i\leqslant t+\tau}^{+}\text{且}\left\|\tilde{\rho}_i-\tilde{\rho}_t\right\|<\rho_p\right\} \tag{13}$$

其中ρ_p为阈值密度. 随后,为每个区域找到其边界区域内最高密度的实例,用ρ_b来表示其密度,密度高于ρ_b的实例会分配为属于此区域,相反则会被考虑可能为漂移的实例.

如果在$\{1,2,\cdots,t+\tau\}$时段之间没有漂移发生,则$m^{B_t^{\phi}}$的基数$m^{B_t^{\phi}}=|D_{t+1,t+\tau}^{B_t^{\phi}}|$,遵循二项分布,即

$$m_{t+1,t+\tau}^{B_t^{\phi}}\sim B(\tau,\phi) \tag{14}$$

因此,在$t+\tau$时刻预测可得到的$\hat{m}_{t+1,t+\tau}^{B_t^{\phi}}$概率满足:

$$P=\left(m_{t+1,t+\tau}^{B_t^{\phi}}=\hat{m}_{t+1,t+\tau}^{B_t^{\phi}}\right)=b\left(\hat{u}_t,\tau,\phi\right) \tag{15}$$

其中,$b\left(\hat{u}_t,\tau,\phi\right)$是二项分布的概率密度函数,若$P=\left(m_{t+1,t+\tau}^{B_t^{\phi}}=\hat{m}_{t+1,t+\tau}^{B_t^{\phi}}\right)<\varepsilon$,则概念漂移有小概率发生,且报告$d_t$漂移等级,其中$\varepsilon=1-p$-Value,用于控制漂移敏感度,$p$-Value为统计学中的假设几率,表示出现该样本或比该样本更极端的结果的概率之和.

考虑在$\{t+1,t+2,\cdots,t+\tau\}$时段内没有实例进入$B_t^{\phi}$的概率记为$F(o,\tau,\phi)$,此时区域集状态可被描述为

$$F(o,\tau,\phi)<\varepsilon\Rightarrow(1-\phi)^{\tau}<\varepsilon \tag{16}$$

且当$\left\|\tilde{\rho}_i-\tilde{\rho}_t\right\|\geqslant\rho_p$时,有

$$F_{\tau+1}(o,\tau+1,\phi)=F(o,\tau,\phi)\cdot(1-\phi) \tag{17}$$

当区域B_t^{ϕ}需要更新时,则需存在一个时间段$\tau_0(t<\tau_0<t+\tau)$满足如下条件:

$$\exists\tau_0\in\mathbb{Z}^{+}\text{ s.t. }\left\|\tilde{\rho}_i-\tilde{\rho}_t\right\|<\rho_p\text{且}F(o,\tau_0,\phi)\geqslant\varepsilon \tag{18}$$

式(18)表明,概念漂移区域检测是基于从1到$t+\tau$时刻所有已到达实例而不仅由区域所包含的实例来判断. 具体来说,若一个数据实例$d_{t_{500}}$在$t=500$时刻到达区域B_t^{ϕ},则此基于缓存的499个实例$\left\{d_{t_1},d_{t_2},\cdots,d_{t_{499}}\right\}$划分为缓冲区域,假设此区域实例比例为$\phi=0.1$且设漂移敏感等级$\varepsilon=0.01$,若在接下来的$\tau=\left\lceil\log_{(1-\phi)}\varepsilon\right\rceil=\left\lceil\log_{(0.9)}0.01\right\rceil=44$个实例没有落在此区域,即$t\in\{501,\cdots,545\}$内没有数据实例出现,则$d_{t_{500}}$会被报告为漂移实例. 因此区域会重新计算模糊密度峰值,进而完成区域的更新. 综上所述,多元区域集分区、更新与维护的伪代码如算法1所示.

算法1 多元区域集的分区、更新与维护

输入:数据实例n,数据流S
模糊算子参数值ω
高斯核宽度参数σ和阈值T
漂移敏感等级ε
输出:多元区域集B

1. $\alpha\left(x_i,x_j\right)\leftarrow\sigma$
2. $\tilde{\rho}_i\leftarrow f\left(\alpha\left(x_i,x_j\right)\right)$
3. $\delta_i\leftarrow d_{ij}$
4. 初始化$\tilde{\rho}_i,\delta_i$; //根据以上公式确定区域集数量
5. $E\leftarrow T$; //根据阈值T确定中心数据点
6. $B\leftarrow$ Init(); //区域初始化完成
7. $\phi\leftarrow B$; //确定区域集比例
8. if $m<\max\left\{\frac{10}{\phi},\frac{10}{1-\phi}\right\}$,then //实例数量不够,等待足够实例
9. else for all $B_p\in B$ do
10. if $F>\varepsilon$ then //缓冲区判断主动更新机制
11. Update(); //执行主动更新机制
12. else do //执行被动更新机制
13. end if

样性保留策略和模型价值度量机制,实现不同类型概念漂移的集成检测,同时保证模型分类性能. 多元区域的建模整体框架如图4所示. 在图4中,$f_k,\cdots,f_m$表示历史模型池中第k到第m个历史模型,$f.$表示从历史模型池中选择出的与输入数据流相似的概念模型,作为集成历史模型对数据流进行检测.

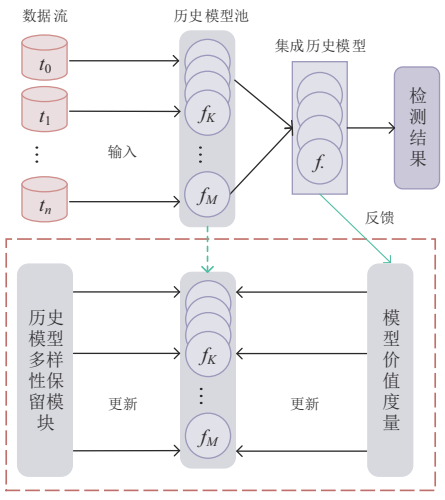

图4 多元区域集建模示意图

3.2 基于多元历史模型池的概念漂移检测

CDMP通过模糊分区机制构建出多元区域集,支持模型的多样性的同时,还可以对概念漂移区域做出自适应的识别与更新. 为了避免每次都重建多元区域集模型带来的性能和时间消耗,CDMP采用历史模型池多

3.2.1 多元历史模型池构建

CDMP采用随机森林(Random Forest,RF)算法作为集成学习策略,对历史模型池进行构建和保留多样性的模型. RF算法具有判断特征重要程度和相互影响的能力,因此被用来判断不同概念漂移下的数据实例

分类[19]. CDMP采用在线重采样的方式,选择较大的特征数量去分割构建决策树,并在构建的随机森林中对保留了多样性的模型进行训练,将得到的模型加入到历史模型池中. 通过保留有价值的历史模型,存储旧概念,CDMP使模型检测效率和性能得到了进一步提升.

图5为历史模型池多样性保留策略示意图,图中,f_n表示第n个数据实例所训练的模型,m表示当前模型池中的模型数量,M表示模型池中所能容纳的最大模型数量,f_k'表示加入到历史模型池中的模型,f_n'表示在原有模型池中的模型. 在模型未到达预设的模型池的最大模型数量M时,模型将会直接加入历史模型池;当模型池数量达到最大值M后,将会对模型池中多样性进行计算,淘汰多样性最差的模型,并通过不断反馈所获得的模型价值度量标准对历史模型池进行更新,以保证历史模型池中各模型均具有显著代表性.

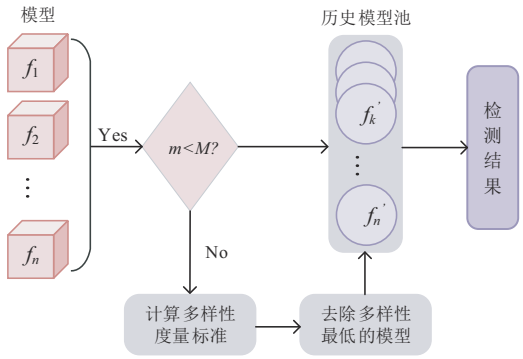

图5　历史模型池多样性保留模块

模型多样性的计算方式为

$$D(S)=1-\frac{1}{\sum_{1\leqslant i\neq j\leqslant m}1}\sum_{1\leqslant i\neq j\leqslant m}Q\left(f_i,f_j\right) \tag{19}$$

其中,$Q\left(f_i,f_j\right)$为f和f_j间的Q统计值,由式(20)计算:

$$Q\left(f_i,f_j\right)=\frac{N^{11}N^{00}-N^{01}N^{10}}{N^{11}N^{00}+N^{01}N^{10}} \tag{20}$$

其中,N^{ab}中a为f_i的分类结果,b为f_j的分类结果,1表示正确分类,0表示错误分类.

3.2.2　基于随机森林的概念漂移检测

在完成随机森林模型构建和历史模型池构建之后,使用警告检测参数δ_ε和实际漂移检测参数δ_d来进行概念漂移检测. CDMP使用经典方法ADWIN(ADaptive WINdowing)[20]作为漂移检测器,对于ADWIN,δ_ε和δ_d分别对应于警告和实际漂移的置信度. 在检测到概念漂移警告级别时将会进行区域集更新并训练新的背景树,当达到发生实际概念漂移级别的阈值时,重新度量模型的价值,并选择最具代表性的历史模型进行集成决策.

如图6为模型价值度量机制框架图. 随着概念漂移的发生,发生变化的多元区域集构建的决策树模型在集成方法中所占权重会增加,以便更精确地实现对当前概念漂移的检测. 基础决策树模型是基于多元区域集中实例训练所得,所以考虑将决策树模型的时效性和区域集的更新机制联合起来. 根据式(14)可知,当多样性区域长时间没有新的数据实例落入,则可知最新的数据实例并不在这个区域,由此区域构建的基础模型自然在新概念的集成预测中时效性较低. 具体来说,时效性由式(21)确定:

$$A_\phi^\tau\left(f_i\right)=\begin{cases}1, & \text{若}\left\|\tilde{\rho}_i-\tilde{\rho}_t\right\|<\rho_p\\ A_\phi^{\tau-1}\left(f_i\right)\times\left(1-\phi\right), & \text{其他}\end{cases} \tag{21}$$

当区域集中数据实例发生漂移时,发生变化的区域时效性将会增加,这会保证集成方法能检测到最新的变化,以提高检测概念漂移的准确率.

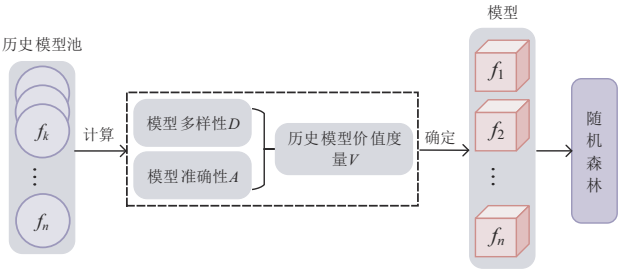

图6　模型价值度量机制

为了便于进行基础模型的价值对比,需要对多样性度量Yule 'Q统计值$Q\left(f_i,f_j\right)$的关系正比化,转化后的$Q\left(f_i,f_j\right)$计算方式如式(22)所示:

$$Q^*\left(f_i,f_j\right)=0.5\cdot\left(1-Q\left(f_i,f_j\right)\right)f_i \tag{22}$$

由于$Q\left(f_i,f_j\right)$的值域为$[-1,1]$,因此,$Q^*\left(f_i,f_j\right)$的值域为$[0,2]$,为了更好的进行价值度量,对结果进行归一化. 综上所述,一个基础模型f_i在整个集成模型M中的多样性度量值为f_i与集成模型M中的每一个基础模型由式(20)得到的$Q^*\left(f_i,f_j\right)$的平均值. 由此每个基础模型与集成模型M之间的多样性$D\left(f_i\right)$可以通过下面的公式计算:

$$D\left(f_i\right)=\sum_{f_i\in M,\forall j,f_j\in M,i\neq j}\frac{Q^*\left(f_i,f_j\right)}{|M|} \tag{23}$$

因此,当一个基础模型f_i构建成功之后,根据式(21)和式(23)分别计算的时效性$A\left(f_i\right)$和多样性$D\left(f_i\right)$,并定义f_i的真实价值$V\left(f_i\right)$由下式来计算,其中W表示模型多样性权重,且$W\in[0,1]$.

$$V\left(f_i\right)=(1-W)\cdot A\left(f_i\right)+W\cdot D\left(f_i\right) \tag{24}$$

基于上述讨论,本文算法从历史模型多样性的和

模型的权重两个方面来度量集成方法中基础模型的价值,并根据此价值确定基础模型在集成方法中的真实权重. 具体的算法伪代码如算法2所示.

算法2 历史模型池构建

输入:数据实例数n和决策树总数h
数据流S
漂移警告阈值δ_ε和漂移阈值δ_d
模型权重W
输出:集成分类器E

CDMP$(n,h,\delta_\varepsilon,\delta_d)$;
$B\leftarrow\phi$;
while HasNext(S) do;
$(x,y)\leftarrow$next(S);
for all $t\in h$ do;
根据决策树算法进行训练;
if $B<M$ then //若模型池未满,则放入模型池;
$B\leftarrow t$;
else do $D-$Count(t); //若模型池已满,则计算模型多样性并更新模型池;
end if
for all $f\in B$ do
$W\leftarrow V(f)$; //根据多样性和时效性计算权重;
if $F(\delta_\varepsilon,t,x,y)$ then //发现概念漂移警告;
Update Data(); //多元区域集更新;
end if
if $C(\delta_d,t,x,y)$ then //检测到概念漂移;
$t\leftarrow B(t)$; //使用价值度量替换价值最低的模型;
end if
end for
end while

3.3 算法复杂度分析

本节将从时间复杂度和空间复杂度两个层面对CDMP算法进行复杂度分析. 由算法1可知,多元区域集的分区、更新和维护的过程都需要时间的消耗,n个数据实例进行多元区域集分区的所需要的时间为$O(n)$,而多元区域集进行更新和维护时的阈值为T,数据流为S,所以CDMP算法的时间复杂度为$O(nTS)$.

在空间复杂度方面,CDMP算法对历史模型池的保留策略虽然将会增加额外的存储空间,但是只保留一定数量的历史模型,仍然低于对所有新数据都进行存储所带来的开销,且保留的历史模型可以对到达的新数据直接进行检测分类. 其中n个数据实例需要的存储单元为$O(n)$,额外保留的m个模型需要的存储单元为$O(m)$. 因此,整个过程的空间复杂度为$O(n+m)$. 然而,传统的概念漂移检测算法需要对所有模型和新模型进行存储. 综上所述,CDMP算法的整体复杂度为$O(nTS+n+m)$.

4 实验结果与性能分析

为了验证本文所提出CDMP算法的有效性,本文对单个区域集的快速响应性能和概念漂移检测的准确性和鲁棒性进行了实验分析.

4.1 区域集更新性能评估

本节将主要对缓冲区的尺寸在概念漂移场景下的变化情况进行仿真及评估,以及验证每个区域子集中保留的核心数据是否真正能代表新的数据实例的概念信息. 仿真将使用与缓冲区限制相同的滑动窗口,使用KS(Kolmogorov-Smirnov)检验作为基准测试进行比较.

为了验证多样性区域集的自适应性能,将CDMP算法根据表1中描述的三种数据分布来生成数据集. 每个数据实例将会基于当前数据分布独立生成,并且为了模拟突然概念漂移和增量概念漂移,数据分布将在$t\in\{1\,251,1\,252,\cdots,1\,750\}$时进行增量型漂移,在$t=2\,500$时突变型漂移. 其中设置$\sigma_{\text{inc}}=1+0.002\times(t-1\,250)$,$\mu_{\text{inc}}=2+0.002\times(t-1\,250)$.

表1 数据分布设置

漂移状态	时间节点t范围	分布
无漂移	$\{1,\cdots,1\,250\}$	$N(0,1)\cup N(2,0.5)$
增量漂移	$\{1\,251,\cdots,1\,750\}$	$N(0,\sigma_{\text{inc}})\cup N(\mu_{\text{inc}},0.5)$
无漂移	$\{1\,751,\cdots,2\,550\}$	$N(0,2)\cup N(3,0.5)$
突变漂移	$\{2\,551,\cdots,4\,000\}$	$N(0,2)\cup N(5,0.5)$

为了维持KS检验的数据缓存区域,本实验使用了常用的概念漂移适应策略,即在检测到概念漂移告警级别时构建一个新的缓冲区,然后将会在达到实际概念漂移级别时替换旧的缓冲区. 其中警告级别的参数设置为$\alpha=0.05$,而漂移等级的参数设置为$\alpha=0.01$.

图7中展示了CDMP在突变型和增量型概念漂移中自适应的表现,以及与传统方法KS检验的对比结果,其中绿色虚线为时刻标注线. CDMP和KS检验都能在每种概念漂移发生后采取自适应措施. 从缓冲区大小可以发现,使用的传统方法在确认概念漂移发生会丢弃几乎所有的历史数据,即使这些历史数据可能是有用的. 在增量漂移发生时段,KS检验触发了两次实际概念漂移级别,使缓冲区的数据实例数量在原有基础上进一步降低,导致缓冲区可用数据实例数量过少. 相比之下,CDMP对漂移的发生更加敏感,能去除与缓冲区数据无关的数据信息,同时保留了符合当前数据分布的历史数据. 且能在漂移持续发生的情况下进行区域集的在线更新,准确识别发生概念漂移的区域.

4.2 概念漂移检测准确性分析

本文将基于大数据在线分析开源平台MOA(Massive Online Analysis)系统进行实验,此系统是用于实现概念漂移数据流相关算法以及对在线学习相关算法进

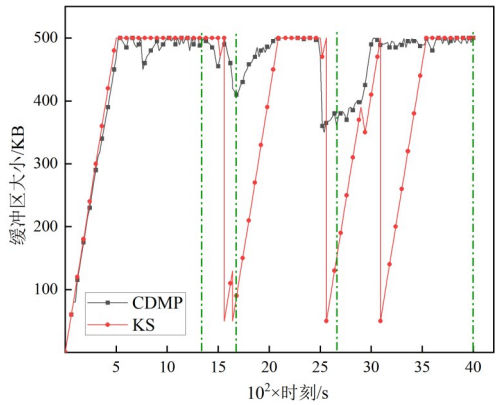

图7　缓冲区大小变化示意图

行仿真验证的实验环境.

4.2.1　数据集描述

面向数据流概念漂移检测的数据集分为真实数据集和人工数据集. 真实数据集的概念漂移不可预测,是算法中常用的数据集. 人工数据集可设置漂移位置、数量、幅度等属性,模拟出不同类型的概念漂移. 本文采用真实数据集和人工数据集相结合的形式验证算法性能和概念漂移预测准确率.

文章中使用了3个真实数据集Covertype[21]、Electricity[22]和Poker[23],以及3个人工数据集SEA[21]、Hyperplane[22]和RanTree[23],数据集各参数如表2所示.

表2　数据集参数简介

数据集	实例数/k	属性数	类别数	噪声水平/%	漂移数	漂移类型
$SEA_{G,1}$	1 000	3	2	10	9	突变型
$SEA_{S,1}$	1 000	3	2	10	3	突变型
Hyperplane	1 000	10	2	5	1	渐变增量型
$SEA_{SR,1}$	1 000	3	2	10	1	突变重现型
RanTree	100	10	6	0	15	突变重现型
Covertype	581	54	7	未知	未知	未知
Electricity	45	7	2	未知	未知	未知
Poker	829	10	10	未知	未知	未知

4.2.2　算法对比研究

在对比实验中,将CDMP与相关的算法进行对比,包括NB[24],LNSE[11],AUE2[25],ARF[14]算法,其中NB是单分类器算法,AUE2和ARF为集成分类方法,LNSE算法可以应对多种不同的概念漂移场景. 在实验过程中,首先对每个数据集随机选取部分数据分别进行预检测,从而保证数据的真实性和概念漂移类型. 此外,所有数据集均重复100次实验,平均分类准确度为去除实验结果中的最大最小值并最终取平均值所得的结果.

由表3的结果可以看出,CDMP在其中6个数据集上的平均分类准确度都显著优于其他的对比算法,但在Hyperplane和Electricity数据集上的检测准确率相比其他对比算法并未展现出明显优势. 原因在于在上述数据集包含渐进型概念漂移,而算法中区域更新策略对渐进型概念漂移的数据实例反应相对迟滞,使得CDMP集成模型对渐进型概念漂移的适应能力相对较低. ARF算法是除了CDMP算法之外表现最好的算法,原因该算法设定了两段阈值机制应对频繁发生的概念漂移. CDMP采用构建历史模型池的策略,随着数据实例的不断到达,CDMP的检测准确率最终都高于其他算法,基于历史多样性的检测机制很好的适应了数据集中重现概念漂移的出现,因此在检测中达到较高的检测准确率.

表3　平均分类准确度对比

数据集	CDMP	NB	ARF	AUE2	LNSE
$SEA_{G,1}$	**89.71**	84.66	89.56	88.95	85.83
$SEA_{S,1}$	**89.51**	83.88	89.07	89.01	86.34
Hyperplane	81.16	72.31	84.05	**89.02**	80.06
$SEA_{SR,1}$	**89.96**	86.52	89.47	89.10	85.94
RanTree	**94.96**	34.04	91.01	94.66	70.27
Covertype	**87.74**	60.27	85.54	85.49	72.66
Electricity	87.85	77.34	**90.10**	88.96	82.85
Poker	**76.01**	59.46	67.70	68.92	54.24

4.3　HHU轴承数据集性能验证

HHU轴承数据集采集于河海大学网络与安全实验室的PT700轴承故障测试平台. 该平台由一个产生扭矩的三相电动机,一个转子组件,一个传动系,齿轮箱振动传感器等组成,整体结构如图8所示.

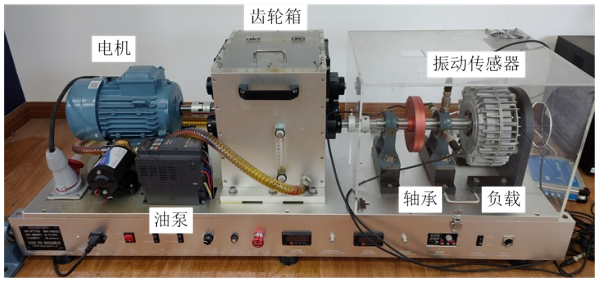

图8　PT700轴承实验台

实验的采样频率为102.4 kHz,故障类型包括内圈故障、外圈故障和滚珠故障. 实验仿真在内圈故障条件下,使用了轻度、中度、重度三种故障程度,对应的故障尺寸分别为0.2 mm、0.4 mm、0.6 mm的损伤. 该数据集中包含有环境噪声,设置信噪比阈值大小为0.01 dB,低于0.01 dB视为无噪声影响. CDMP分别对两种不同场景下的概念漂移进行性能验证,场景一为轴承在内圈故障、不同负载条件下,转速变化下引起的概念漂移,场景二为轴承在内圈故障、相同转速、故障程度变化下引起的概念漂移. 实验参数和检测结果如表4、表5

所示.

表4、表5的实验结果表明,在真实的工业场景下对数据流发生的概念漂移进行检测时,CDMP仍然表现出较高的准确度. 这进一步表明,CDMP不仅在广泛使用的实验数据集中对数据流概念漂移的检测具有较好的性能,而且在真实的工业应用场景下也具有巨大的应用潜力.

表4 不同工况下的实验参数与检测准确度

时间	负载	转速变化/rmp	样本数量/k	样本长度	准确度/%
1	Load 0	100→150	500	204 800	93.84
2	Load 1	150→200	500	204 800	92.49
3	Load 2	200→250	500	204 800	89.64
4	Load 3	250→300	500	204 800	90.26
5	Load 4	300→350	500	204 800	89.39

表5 不同故障程度下的实验参数与检测准确度

时间	转速	故障程度变化	样本数量/k	样本长度	准确度/%
1	200	轻度→中度	500	204 800	92.35
2	200	中度→重度	500	204 800	89.58

为了更好的验证CDMP的性能,在HHU数据集下对CDMP与其它概念漂移检测算法进行了对比实验,实验结果如图9所示. 对比算法除NB,LNSE,AUE2,ARF算法外,进一步加入了了AC-OE[22]和EOFWOSELM[26]两个集成学习算法. 实验结果可知,CDMP由于事先对多元区域集进行划分和历史模型池重用的策略,仍然对工业场景下HHU数据集存在的概念漂移具有较高的检测准确度和实用性.

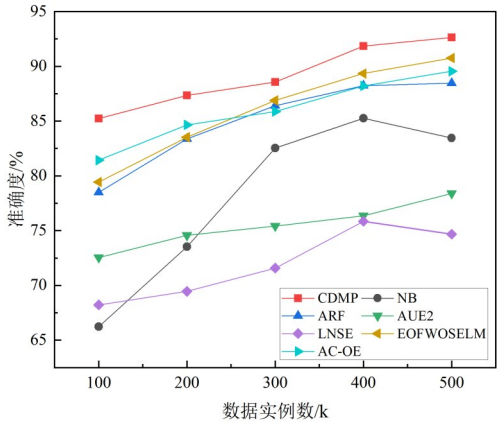

图9 HHU数据集下不同算法的检测准确度

4.4 抗噪性分析

本节对不同噪声水平下各算法的检测准确率进行验证. 表6列出了各算法在不同的噪声水平下的平均分类准确率,噪声水平表示在数据集中加入噪声的比率,设置噪声大小为5 dB. 可以看出在不同的噪声水平影响下,CDMP的分类准确率均比其他对比算法高.

表6 不同噪声水平下的分类准确度

噪声水平/%	CDMP	NB	ARF	AUE2	LNSE
5	94.01	86.08	92.13	92.01	91.22
10	89.70	80.90	89.37	89.02	86.04
15	87.72	71.01	85.64	85.67	85.63
20	83.44	65.98	80.75	80.66	81.17
25	80.51	62.08	78.56	76.18	77.12
30	76.31	60.33	73.39	70.82	72.67

图10中表示在同噪声水平下不同算法在SEA数据集上检测准确度的对比结果. 各个算法均可在无噪声的情况下保持较高的检测准确率,但当噪声水平由0%增大到20%时,对比算法的检测性能都有不同程度的降低,但本文算法能在高噪声水平的场景下保持相对更好的准确性和稳定性,原因在于在构建多样性区域集的过程中模糊密度峰值可以筛选数据实例,筛选低密度的离群值,从而实现集成方法中基础模型的高泛化性和高鲁棒性. 这是CDMP在工业高噪声水平下能保持高稳定性和高鲁棒性的根本原因.

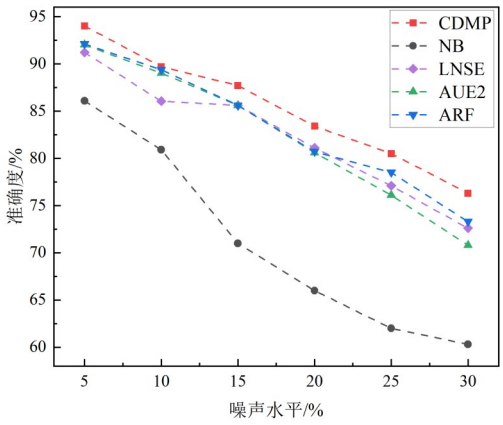

图10 不同噪声水平下的检测结果对比

4.5 消融实验

本节对CDMP所包含的模块进行了消融实验,进一步验证模糊分区机制和历史模型池的作用,共进行了两种消融实验:(1)将CDMP中模糊分区机制的模糊算子去掉,变为非模糊分区,得到CDMP-F算法;(2)将CDMP中的历史模型池去掉,直接训练模型,得到CDMP-H算法. 在SEA数据集上分别对CDMP、CDMP-F和CDMP-H算法的概念漂移检测的准确度进行实验,其结果如图11所示.

图11实验结果表明,将模糊分区机制替换为非模糊分区后,算法对概念漂移的检测准确率大幅降低,同样的,去掉算法中的历史模型池模块后,对概念漂移的

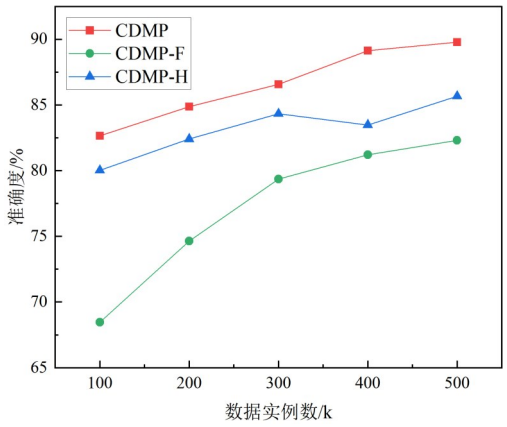

图 11　消融实验结果

识别准确度也有所降低. 可见,CDMP算法中的模糊分区机制和历史模型池模块对整个算法准确度的提升贡献至关重要.

5　结论

工业场景下的多变工况极大影响了传统概念漂移检测的快速响应能力. 本文提出的基于多元区域集划分的工业数据概念漂移检测算法CDMP基于模糊密度峰值的聚类方法实现在线的数据实例的模糊分区,对多元区域集实施概念漂移区域识别. 并通过构建并维护历史模型池,实现了对概念漂移识别的快速响应. 实验结果表明,CDMP能在噪声干扰的工业数据流中实现对多元区域集的实时更新并准确识别概念漂移区域,具有较高的泛化性与鲁棒性. 在未来的工作中,将考虑对概念漂移区域中缓冲区的大小进行自适应调整,同时降低增量漂移对算法的适应性产生的负面影响.

参考文献

[1] YANG C E, CHEUNG Y M, DING J L, et al. Concept drift-tolerant transfer learning in dynamic environments[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(8): 3857-3871.

[2] FRAVOLINI M L, DEL CORE G, PAPA U, et al. Data-driven schemes for robust fault detection of air data system sensors[J]. IEEE Transactions on Control Systems Technology, 2019, 27(1): 234-248.

[3] XU L D, HE W, LI S C. Internet of things in industries: A survey[J]. IEEE Transactions on Industrial Informatics, 2014, 10(4): 2233-2243.

[4] CANO A, KRAWCZYK B. Kappa updated ensemble for drifting data stream mining[J]. Machine Learning, 2020, 109(1): 175-218.

[5] SIDHU P, BHATIA M P S. A novel online ensemble approach to handle concept drifting data streams: Diversified dynamic weighted majority[J]. International Journal of Machine Learning and Cybernetics, 2018, 9(1): 37-61.

[6] ZHOU H, SHE C Y, DENG Y S, et al. Machine learning for massive industrial Internet of Things[J]. IEEE Wireless Communications, 2021, 28(4): 81-87.

[7] 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.

[8] GAMA J, ŽLIOBAITĖ I, BIFET A, et al. A survey on concept drift adaptation[J]. ACM Computing Surveys, 2014, 46(4): 1-37.

[9] GUO Hu-sheng, LI Hai, REN Qian-yan, et al. Concept drift type identification based on multi-sliding windows[J]. Information Sciences, 2022, 585: 1-23.

[10] LIU A J, ZHANG G Q, LU J. Fuzzy time windowing for gradual concept drift adaptation[C]//2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). Piscataway: IEEE, 2017: 1-6.

[11] SHEN Y, DU J G, TONG J G, et al. A parallel and reverse Learn++.NSE classification algorithm[J]. IEEE Access, 2020, 8: 64157-64168.

[12] LI Zeng, HUANG Wen-chao, XIONG Yan, et al. Incremental learning imbalanced data streams with concept drift: The dynamic updated ensemble algorithm[J]. Knowledge-Based Systems, 2020, 195: 105694.

[13] IKONOMOVSKA E, GAMA J, DŽEROSKI S. Online tree-based ensembles and option trees for regression on evolving data streams[J]. Neurocomputing, 2015, 150: 458-470.

[14] GOMES H M, BIFET A, READ J, et al. Adaptive random forests for evolving data stream classification[J]. Machine Learning, 2017, 106(9): 1469-1495.

[15] 浦慧忠. k-means聚类分析算法在人工智能+个性化学习系统中的应用[J]. 智能计算机与应用, 2022, 12(8): 152-156.
PU H Z. Application research of k-means cluster analysis algorithm in artificial intelligence + personalized learning system[J]. Intelligent Computer and Applications, 2022, 12(8): 152-156. (in Chinese)

[16] SAXENA A, PRASAD M, GUPTA A, et al. A review of clustering techniques and developments[J]. Neurocomputing, 2017, 267: 664-681.

[17] XU R, WUNSCH D. Survey of clustering algorithms[J]. IEEE Transactions on Neural Networks, 2005, 16(3): 645-678.

[18] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.

[19] KRAWCZYK B, MINKU L L, GAMA J, et al. Ensemble learning for data stream analysis: A survey[J]. Information Fusion, 2017, 37: 132-156.

[20] GHOMESHI H, GABER M M, KOVALCHUK Y. A non-canonical hybrid metaheuristic approach to adaptive data stream classification[J]. Future Generation Computer Systems, 2020, 102: 127-139.
[21] LIU A J, LU J, ZHANG G Q. Diverse instance-weighting ensemble based on region drift disagreement for concept drift adaptation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(1): 293-307.
[22] 郭虎升, 丛璐, 高淑花, 等. 基于在线集成的概念漂移自适应分类方法[J/OL]. 计算机研究与发展: 1-12. [2023-04-10]. http://kns.cnki.net/kcms/detail/11.1777.TP.20220818.1639.010.html.
GUO H S, CONG L, GAO S H, et al. Adaptive classification method for concept drift based on online ensemble[J/OL]. Journal of Computer Research and Development: 1-12. [2023-04-10]. http://kns.cnki.net/kcms/detail/11.1777.TP.20220818.1639.010.html.(in Chinese)
[23] 夏源, 赵蕴龙, 范其林. 基于信息熵更新权重的数据流集成分类算法[J]. 计算机科学, 2022, 49(3): 92-98.
XIA Y, ZHAO Y L, FAN Q L. Data stream ensemble classification algorithm based on information entropy updating weight[J]. Computer Science, 2022, 49(3): 92-98. (in Chinese)
[24] KRAWCZYK B, WOZNIAK M. Weighted naïve bayes classifier with forgetting for drifting data streams[C]// 2015 IEEE International Conference on Systems, Man, and Cybernetics. Piscataway: IEEE, 2015: 2147-2152.
[25] BRZEZINSKI D, STEFANOWSKI J. Reacting to different types of concept drift: The accuracy updated ensemble algorithm[J]. IEEE Transactions on Neural Networks and Learning Systems, 2014, 25(1): 81-94.
[26] 陆克中, 陈超凡, 蔡桓, 等. 面向概念漂移和类不平衡数据流的在线分类算法[J]. 电子学报, 2022, 50(3): 585-597.
LU Ke-zhong, CHEN Chao-fan, CAI Huan, et al. Online classification algorithm for concept drift and class imbalance data stream[J]. Acta Electronica Sinica, 2022, 50(3): 585-597. (in Chinese)

作者简介

韩光洁 男, 1972 年 8 月出生于黑龙江省绥化市. 现为河海大学物联网工程学院教授, 博士生导师, 同时也是 IEEE Fellow、IET/IEE Fellow、AAIA Fellow. 从事工业物联网、智慧海洋、人工智能、网络安全等方面的研究工作.
E-mail: hanguangjie@gmail.com

赵腾飞 男, 1997 年 9 月出生于河南省周口市. 现为河海大学硕士研究生. 从事工业物联网故障诊断和迁移学习方面的研究工作.
E-mail: zhaotengfei868@163.com

刘 立 男, 1992 年 5 月出生于江苏省无锡市. 现为河海大学物联网工程学院讲师. 从事人工智能、机器学习、大数据分析方面的研究工作.
E-mail: liulihhuc@gmail.com

张 帆 女, 1996 年 7 月出生于江苏省镇江市. 2019 年毕业于河海大学物联网工程专业并获得学士学位. 现于河海大学物联网工程学院攻读博士学位. 从事工业物联网、边缘计算和机器学习方面的研究工作.
E-mail: zhangfanhhuc@gmail.com

徐政伟 男, 1994 年 3 月出生于河南省新乡市. 现于河海大学攻读计算机科学与技术专业博士学位. 从事深度学习、大数据分析及辐射源个体识别等方面的研究工作.
E-mail: xuzhengweihhu@outlook.com