◆计算科学◆

一种基于局部分类精度的概念漂移数据流分类算法*

张　玲,马士伦,黎利辉,文益民**
(桂林电子科技大学,广西图像与图形智能处理重点实验室,广西桂林　541004)

摘要:概念漂移数据流分类是一个极具挑战性的问题。当新概念出现时,该概念下的学习样本过少,无法对分类器进行及时调整,进而导致分类精度不高。为了解决该问题,本文提出一种基于局部分类精度的概念漂移数据流分类算法——LA-MS-CDC。第一,LA-MS-CDC 将 k-means 聚类和局部分类精度算法结合,从分类器池中挑选出最优源领域分类器;第二,将最优源领域分类器与目标领域分类器加权集成,进而对样本分类;第三,根据分类样本的真实标签分别计算各分类器的损失,并对目标领域和源领域的分类器权重进行更新;第四,再利用该分类样本对目标领域分类器、最优源领域分类器进行更新;最后,完成分类器池的更新。在公开数据集上的实验结果表明,LA-MS-CDC 能够有效地将源领域知识迁移到目标领域,与现有方法相比,其分类效果具有显著性提升。算法代码可在 https://gitee.com/ymw12345/LAMSCDC 上获取。

关键词:概念漂移;多源在线迁移学习;局部分类精度;集成学习;多样性

中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1005-9164(2024)01-0100-10
DOI:10.13656/j.cnki.gxkx.20240417.010

随着云计算和物联网等新技术的飞速发展,数据流分类算法被应用于许多领域,如信用卡欺诈检测、交通监控和医疗决策辅助等。数据流中数据的分布可能随时间、环境的变化而改变,这种变化被称为概念漂移[1]。概念漂移问题给机器学习带来了巨大的挑战,在存在概念漂移的数据流分类任务中,如果模型不能尽早识别新概念且动态地调整分类器,则模型分类准确率会迅速下降。数据流分类问题中,在新概念出现的初期仅能获取到少量样本,用这些少量样本训练出的模型分类效果较差,但是基于数据流分类中不同概念之间的数据依然存在可学习性[2]这一共识,故用迁移学习来提升数据流分类的性能就成目前研究的热点[3]。

迁移学习根据源领域的数量可以分为单源迁移学习和多源迁移学习[4]。单源迁移学习源领域数量单一,存在信息不充分的问题。多源迁移学习能够有效地利用来自不同分布的相关源领域的知识来提高目标领域分类任务的性能[5]。例如,Eaton 等[6]提出

收稿日期:2023-03-08　　修回日期:2023-03-31
*国家自然科学基金项目(62366011),广西重点研发计划项目(桂科 AB21220023)和广西图像图形与智能处理重点实验室项目(GIIP2306)资助。
【第一作者简介】
张　玲(1997—),女,在读硕士研究生,主要从事数据流分类研究,E-mail:1103650837@qqcom。
【**通信作者简介】
文益民(1969—),男,博士,教授,主要从事机器学习与模式识别、图像处理与理解等研究,E-mail:ymwen@gueteducn。
【引用本文】
张玲,马士伦,黎利辉,等.一种基于局部分类精度的概念漂移数据流分类算法[J].广西科学,2024,31(1):100-109.
ZHANG L,MA S L,LI L H,et al. A Concept Drift Data Stream Classification Algorithm Based on Local Classification Accuracy [J]. Guangxi Sciences,2024,31(1):100-109.

TransferBoost 算法,该算法合并多个源领域与目标领域的样本,并根据样本的迁移能力为各源领域的每个样本设置权重。Chattopadhyay 等[7] 基于源领域和目标领域的样本分布具有相似性这一特点,提出 CP-MDA 算法。Gao 等[8] 提出 LWE 算法,该算法在每个源领域上分别训练一个组件分类器,并对每个组件分类器赋予不同的权值,然后根据权值加权集成。Ge 等[9] 提出 OMS-TL 算法,该算法将离线的 MSTL 算法改进为在线算法。Wu 等[10] 提出的 OTLMS 算法根据每个源领域分类器对目标领域样本的分类准确度为每个源领域分类器设置一个权重,并根据分类性能自适应调节权重大小。Kang 等[11] 提出适用于多分类任务的 OMTL-MC 算法,OMTL-MC 算法采用两段集成的思想,从多个源领域迁移知识来辅助目标领域的学习任务。文益民等[12] 提出 RCOTL 算法,该算法通过计算拟分类样本与分类器池中各个分类器之间的领域相似度,改善了从源领域到目标领域的知识迁移,解决了重现概念数据流分类中的"概念漂移检测滞后"和"负迁移"的问题。秦一休等[13] 针对 CDOL 算法的不足,提出 CMOL 算法,该算法使用将权重最大的分类器从分类器池中替换出去的策略,保证了分类器池中的分类器多样性最大化。Sun 等[14] 提出 DTEL 集成模型,该算法利用源领域中所包含的知识实现概念漂移自适应训练。吕艳霞等[15] 在 DTEL 集成模型的基础上,基于准确率调整范围参数的模型更新策略提出了 HAEL 集成框架,该算法引入注意力机制,使其更倾向于使用目标领域样本构建和更新分类模型。Zhao 等[16] 采用最小二乘支持向量机作为组件分类器,提出了 Condor SVM 算法,该算法采用偏差正则化技术以利用源领域的分类器来进行迁移学习,并根据源领域分类器对目标样本的分类性能来调整各源领域分类器的权重。赵鹏等[17] 采用决策树作为组件分类器,提出了 Condor Forest 算法,该算法采用乘法权重更新法来更新模型库中每个模型的权重,并挑选权重最大的模型作为可重复使用的模型。然而,上述多源迁移学习算法在源领域与目标领域相似度低时,迁移效果不好。唐诗淇等[18] 针对这个问题,提出了一种基于局部分类精度的多源在线迁移学习算法 LC-MSOTL,LC-MSOTL 维护一个源领域分类器池,通过计算局部分类精度,从源领域分类器池中挑选出一个分类器来提升目标领域的分类性能。但是,在动态数据流环境中,LC-MSOTL 由于缺少分类器池更新机制以及所用的最近邻算法对拟分类样本周围邻居查找不够准确等原因,其分类效果并不理想。

针对这些问题,本文提出了一种基于局部分类精度的概念漂移数据流分类算法——LA-MS-CDC,LA-MS-CDC 基于多源在线迁移学习框架,能有效适应概念漂移数据流环境,并且较好地解决了 LC-MSOTL 算法存在的问题。具体来说,基于支持向量差异性最大化提出的一种维护分类器池多样性最大化的方法保证了分类器池中的分类器尽可能多地映射不同概念,提高了算法对概念漂移的处理能力;根据源领域和目标领域动态权重比提出的一种选择性更新源领域分类器的方法有效避免了负迁移的发生。

1 LA-MS-CDC

样本以在线的方式到来并进行分类,本文设定 x_i 为在第 i 时刻到达的样本:$(x_i,y_i)\in X\times Y$,$X\in R^n$,$Y=\{-1,+1\}$,X 为 n 维的属性空间,Y 为样本的标记空间,y_i 为样本的标记值。

LA-MS-CDC 的流程如算法 1 所示,其基于多源在线迁移学习进行数据流分类。为了应对数据流环境中的数据分布变化,LA-MS-CDC 周期性对分类器池进行更新。在开始阶段,使用初始数据样本训练一个源领域分类器 f_s,初始化分类器池 CF、k-means 聚类簇心、源领域权重 $\alpha_{1,1}$、目标领域分类器权重 $\alpha_{2,1}$ 及目标域分类器 f_t(行 1－6),各在线分类器均为 Passive Aggressive (PA)[19] 分类器。

针对每一个样本 x_i,以在线学习的方式进行如下操作:首先找到 x_i 所属聚类簇的簇心 $c_{\min}$,并且使用样本 x_i 更新 k-means 聚类簇心(行 8－9);其次使用 Select 函数从分类器池中挑选出一个最优源领域分类器 f_{s_k},与目标领域分类器 f_t 加权集成后,对样本 x_i 进行预测(行 10－11),当分类任务结束后,获得样本 x_i 的真实标签 y_i,更新源领域权重 $\alpha_{1,i}$、目标领域权重 $\alpha_{2,i}$ 与目标领域分类器 f_t(行 12－13),如果源领域权重与目标领域权重比值大于指定的阈值 θ,则更新最优源领域分类器 f_{s_k}(行 14－16);最后将样本 (x_i,y_i) 存入数据缓存器 BD 中(行 17)。当数据缓存器 BD 中的样本达到容量上限后,将目标领域分类器加入分类器池中,完成分类器池的更新操作(行 18－22)。同时,再次初始化源领域权重 $\alpha_{1,i+1}$、目标领域权重 $\alpha_{2,i+1}$、数据缓存区 BD、目标领域分类器 f_t(行 23－25)。

算法 1 LA-MS-CDC

输入:流数据 $\{(x_i,y_i),\cdots,(x_T,y_T)\}$ 分类器池大小 M,数据缓存器 BD 的大小 N,近邻个数 K,放大系数 ζ,聚类簇数 n,源领域与目标领域权重阈值 θ。

输出:x_i 的预测标签 y_i。

1. 利用开始的数据训练一个 PA 分类器 f_{s1}
2. 初始化分类器池 $CF=\{f_{s1}\}$
3. 利用开始的数据初始化聚类簇心列表 $(c_1,c_2,\cdots,c_n)$
4. $\alpha_{1,1}=0.5,\alpha_{2,1}=0.5$
5. $f_t=$create() / * create() 函数为初始化目标域分类器 * /
6. BD=[]
7. for $i=1$ to T do
8. 获得 x_i 所属簇的簇心 $c_{\min}$
9. $(c_1,c_2,\cdots,c_n)=$update_cluster($x_i(c_1,c_2,\cdots,c_n)$) / * update_cluster() 函数为更新聚类簇心函数 * /
10. $f_{s_k}=$Select(CF,BD,x_i,K,ζ,$c_{\min}$,$(c_1,c_2,\cdots,c_n)$)
11. $y=\text{sign}(\alpha_{1,i}\prod(f_{s_k}(x_i))+\alpha_{2,i}\prod(f_t(x_i))-\frac{1}{2})$
12. $\alpha_{1,i+1}=\frac{\alpha_{1,i}s_i(w_s)}{\alpha_{1,i}s_i(w_s)+\alpha_{2,i}s_i(w_t)},\alpha_{2,i+1}=\frac{\alpha_{2,i}s_i(w_t)}{\alpha_{1,i}s_i(w_s)+\alpha_{2,i}s_i(w_t)}$
13. $f_t$update(x_i,y_i) / * update() 函数为更新目标域分类器 * /
14. if $\alpha_{1,i+1}/\alpha_{2,i+1}>\theta$
15. f_{s_k}.update(x_i,y_i) / * 更新源领域分类器 * /
16. end if
17. BD.save(x_i,y_i)
18. if | BD |=N
19. CF.add(f_t) / * 将训练好的目标域分类器加入分类器池 * /
20. if | CF |>M
21. 从 CF 中删除支持向量差异性最小的分类器
22. end if
23. $\alpha_{1,i+1}=0.5,\alpha_{2,i+1}=0.5$
24. $f_t=$create()
25. BD=[]
26. end if
27. end for
28. return

1.1 基于 k-means 聚类的局部分类精度的算法

Gao 等[8]注意到源领域与目标领域有相似的局部流形,每个源领域与目标领域都有较大的不同,但又都各自存在分布相同的局部。若在每个源领域上训练一个分类器,并把这些分类器迁移到目标域上辅助目标领域进行分类,必然会产生“负迁移”。如图 1 所示,目标领域 a3 区域与源领域 b1 区域分布相同,与 b2 区域分布相异,若使用 b1 训练的分类器对 a3 区域数据进行分类,样本就会被正确分类;若使用 b2 区域训练的分类器对 a3 区域进行分类,样本就会被错误分类。因此,对于目标领域中的样本,若能为每个样本选择合适的源领域分类器对其进行辅助分类,就能更好地实现知识迁移,从而对目标领域样本更准确地分类。

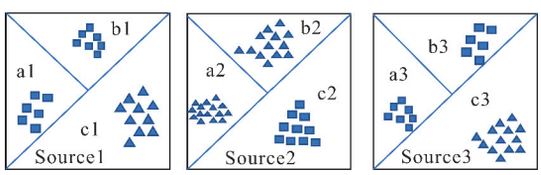

图 1 源领域与目标领域的局部相似性

Fig. 1 Local similarity between source domain and target domain

局部分类精度是分类器选择中的一种有效方法[8,20-22]。由于每个组件分类器对拟分类样本的邻域分类能力不同,因此该方法选择局部分类精度最高的源领域分类器为最优源领域分类器,其中训练该最优源领域分类器的数据与拟分类样本的邻域相似度最高。第 j 个组件分类器 f_{s_j} 针对拟分类样本 x_i 的局部分类精度 LC_j 的计算方法如公式(1)所示。

$$LC_j=\sum_{m=1}^{k}\frac{I(f_{s_j}(x_m)==y_m)}{e^{\zeta\text{dis}(x_i,x_m)}},\tag{1}$$

其中,x_i 为拟分类样本的特征向量,k 为其邻域的样本数量,(x_m,y_m) 为该邻域中第 m 个样本的特征向量和真实类别,样本 (x_m,y_m) 来自数据缓存器 BD,$\text{dis}(x_i,x_m)$ 为样本 x_m 与拟分类样本 x_i 的欧式距离,ζ 为放大系数,$I(x)$ 为指示函数,其形式如公式(2)所示:

$$I(x)=\begin{cases}1, & x=\text{true}\\ 0, & x=\text{false}\end{cases}。\tag{2}$$

LC-MSOTL 算法[18]使用 K 近邻算法挑选距离拟分类样本最近的 K 个样本作为邻居，然而，LC-MSOTL 算法在带有概念漂移的数据流环境中所找到的邻居样本并不准确，具体表现为分类器池中的各个分类器在局部邻域上计算出的局部分类精度很小甚至为零，这说明通过 K 近邻算法找到的拟分类样本的邻居样本有误，从而导致根据局部分类精度挑选出来的源领域分类器对拟分类样本的分类辅助能力很弱。

LA-MS-CDC 针对 LC-MSOTL 算法无法准确找出待分类样本的邻居样本这一问题做出改进，使用 k-means 聚类算法保留已分类数据的聚类簇信息。k-means 聚类算法是一种经典的聚类算法，其思想简单，聚类效果良好，应用广泛。相比于其他聚类算法，k-means 聚类算法具有计算速度快，资源消耗少，并且适用于在线学习环境等优点。使用 k-means 聚类算法改进后，每当一个新样本到来时，首先查找该新样本所属的簇，其次使用该新样本更新簇心信息，最后使用 K 近邻算法在该簇上寻找新样本的 K 个邻居。

动态挑选源领域分类器的 Select 函数伪代码如算法 2 所示。首先，新建一个空列表 sameList(行 1)，其次在数据缓存器 BD 中寻找与拟分类样本在同一簇的样本，并将这些样本与待分类样本间的欧式距离存入 sameList 中(行 2－7)，再从 sameList 中寻找拟分类样本 x_i 的 K 个邻居(行 10－14)，最后根据局部分类精度计算公式从分类器池 CF 中挑选出最优源领域分类器 f_{s_k} (行 15)。若找不到最优源领域分类器 f_{s_k}，分类器池中的各个分类器均对 x_i 预测标签，依据多数类投票原则确定 x_i 的标签为 y_s，再从预测为 y_s 的分类器中随机挑选一个分类器作为 f_{s_k} (行 16－18)。

算法 2 Select 函数

输入：聚类簇心 $c_1,c_2,\cdots,c_n$ 新到样本特征向量 x_i，x_i 所属簇的簇心 $c_{\min}$，分类器集合 CF，数据缓存器 BD，近邻个数 K，放大系数 ζ。

输出：最优源领域分类器 f_{s_k}。

1. sameList=[]
2. for $j=1$ to |BD| do
3. if min_cluster(x_j,($c_1,c_2,\cdots,c_n$))$==c_{\min}$ do /*函数 min_cluster()为样本 x_j 距离最近的簇心*/
4. $dis_j=\mathrm{dis}(x_i,x_j)$
5. sameList.add(dis_j)
6. end if
7. end for
8. sort(sameList)
9. Kneighbor=[]
10. for $n=1$ to |sameList| do
11. 根据 sameList[n]找到对应的邻居样本(x_n,y_n)
12. Kneighbor.add((x_n,y_n))
13. if ($n\geqslant K$) break
14. end for
15. $f_{s_k}=\underset{f_{s_j}\in CF}{\operatorname{argmax}}\sum_{n=1}^{K}\frac{I(f_{s_j}(x_n)==y_n)}{e^{\zeta\mathrm{dis}(x_i,x_n)}}$
16. if $f_{s_k}=\varnothing$
17. CF 中所有分类器按照多数类投票决定 x_i 的标签 y_s
18. 从预测为 y_s 的分类器中随机挑选一个分类器作为 f_{s_k}
19. return f_{s_k}

1.2 LA-MS-CDC 的更新机制

1.2.1 源领域分类器选择性增量更新

增量学习是指使用不断接收的新数据持续地调整和更新模型，增量学习可以在保留大部分过去有效知识的情况下学习新知识。当概念漂移现象出现在增量学习中时，如果不针对性地调整模型的增量更新机制，新到的概念往往与模型中已经学习到的概念相冲突并影响学习的效果[23]。

LA-MS-CDC 为解决在概念漂移环境下新到数据知识与模型中已有知识相冲突的问题，当源领域权重与目标领域权重比值大于阈值 θ 时，增量更新最优源领域分类器。当 θ 很小时，说明源领域分类器对拟分类样本的分类结果贡献度低，此时的源领域分类器对拟分类样本的分类错误率较高，即源领域分类器学到的知识与当前拟分类样本的信息相冲突。因此用拟分类样本增量更新最优源领域分类器并不合理，反之亦然。

1.2.2 目标领域分类器的更新策略

目标领域分类器为在线学习器 PA，当得到样本 x_i 的预测标签后，则用样本 x_i 的预测标签和真实标签更新目标领域分类器，具体步骤如下：

①使用 hinge-loss 损失函数计算分类器损失 l_t；

②若 $l_t>0$，则对分类器进行更新 $w_{t+1}=w_t+\tau y_t x_t$，

其中 $\tau=\min(C,l_t/\|x_t\|^2)$。

1.2.3 权重 $\alpha_{1,i}$ 和 $\alpha_{2,i}$ 的更新策略

源领域权重 $\alpha_{1,i}$ 和目标领域权重 $\alpha_{2,i}$ 的更新机制:在算法开始时初始化为 0.5,每分类完一个样本则按照公式(3)和公式(4)更新[24]。

$$\alpha_{1,i+1}=\frac{\alpha_{1,i}s_i(w_s)}{\alpha_{1,i}s_i(w_s)+\alpha_{2,i}s_i(w_t)},\tag{3}$$

$$\alpha_{2,i+1}=\frac{\alpha_{2,i}s_i(w_t)}{\alpha_{1,i}s_i(w_s)+\alpha_{2,i}s_i(w_t)},\tag{4}$$

其中,$s_i(w)=\exp\{-\eta l^*(\prod(w^{\mathrm{T}}x_i),\prod(y_i))\}$,$\eta=0.5$,$l^*(z,y)=(z-y)^2$,$\prod(x)=\max(0,\min(1,\frac{x+1}{2}))$。

1.3 分类器池多样性最大化

集成学习能够提高组件分类器的泛化能力[25,26],然而当各个组件分类器的表现都相同时,集成分类器的分类性能将不会提高,因此多样的组件分类器往往可以提高集成学习的性能。在数据流场景下,分类器池多样性最大化的具体表现为在分类器池中最大限度地映射多种概念,保留尽可能多的过去有效信息。目前多样性的度量方法主要分为非成对度量和成对度量[27],因 LA-MS-CDC 算法所采用的组件分类器为在线学习分类器 PA,故本文结合组件分类器自身的特性提出了一种度量分类器池多样性的新指标,即依据支持向量差异性最大化来维护分类器池多样性最大化。

PA 分类器是一种基于感知机的在线学习分类器,它利用支持向量维护分类决策面,进而实现对样本的分类。因此,分类器池中各种分类器可以用支持向量来具体化。当分类器池中的支持向量的差异性最大时,也就说明了分类器池存储的分类决策面差异性最大,即分类器池中的分类器多样性最大。

假设 LA-MS-CDC 维护的分类器池已满。当目标领域分类器 f_t 训练完成后,第一步,将 f_t 加入分类器池中;第二步,依次删除分类器池中的一个分类器,分别计算任意两个不同组件分类器中支持向量相同的个数,并累积求和。当删除分类器池中某个分类器 f_i 后累积的支持向量个数最小时,说明分类器池多样性最大,此时 f_i 为分类器池本轮应该删除的分类器。本文采用的多样性度量方法为成对度量,f_i 的求解如公式(5)所示。

$$f_i=\underset{(CF-f_i)}{\operatorname{argmin}}\sum_{m=1}^{M}\sum_{n=m+1}^{M}\mathrm{sum}(v_m==v_n),\tag{5}$$

其中,M 为分类器池 CF 的容量,$\mathrm{sum}(v_m==v_n)$ 表示分类器 f_m 与分类器 f_n 中支持向量相同的数量。

2 实验及结果分析

2.1 数据集和实验设置

本文采用 Letter 数据集和 waveform 数据集来测试 LA-MS-CDC 的性能。Letter 数据集是来自加利福尼亚大学尔湾分校(UCI)的真实数据集,该数据集共计 20 000 个样本,包含 16 维连续属性和 26 种类别,这 26 种类别分别对应 26 个字母。waveform 数据集是人工数据集,来自数据流领域常用的开源学习框架 MOA[28]。

本文使用 Letter 数据集制作 ABCD 数据集、EFGH 数据集、NOPQ 数据集。由于原始的 Letter 数据集中并未包含概念漂移,为了更有效地对 LA-MS-CDC 算法的性能进行测试,将 Letter 数据集按照表 1 方式标注正、负类类别标签,并以此来产生概念漂移,共生成 3 种概念,ABCD 数据集的第 1−400 个样本组成概念 A,数据集的第 401−800 个样本组成概念 B,数据集的第 801−1 200 个样本组成概念 C,其中概念 A 到概念 B 为突变概念,EFGH 数据集和 NOPQ 数据集生成概念漂移的方式与 ABCD 数据集相同。

表 1 Letter 数据集概念定义

Table 1 Definition of letter dataset concept

字母集 Letter set	概念 A Concept A 1−400	概念 B Concept B 401−800	概念 C Concept C 801−1 200
A(E)(N)	−	+	+
B(F)(O)	−	+	−
C(G)(P)	+	−	+
D(H)(Q)	+	−	−

waveform 原始数据集共计 5 500 个样本,包含 21 个连续属性和 3 个类别。为使 waveform 数据集包含概念漂移,将 waveform 数据集按照表 2 所示的方式标注正、负类类别标签,共生成 3 种概念,每种概念重复 2 次,共产生 9 个概念。其中,数据集的第 1−600 个样本组成概念 E,数据集的第 601−1 200 个样本组成概念 F,数据集的第 1 201−1 800 个样本组成概念 G。

表 2　waveform 数据集概念定义

Table 2　Definition of waveform dataset concept

原始类别 Original category	概念 E Concept E 1－600	概念 F Concept F 601－1 200	概念 G Concept G 1 201－1 800
1	－	＋	－
2	＋	－	－
3	＋	＋	＋

将采用表 1 和表 2 的方式处理的数据集作为实验数据集，可以帮助证明 LA-MS-CDC 算法具有挖掘出源领域和目标领域的局部相似性的能力。基于这种局部相似性，LA-MS-CDC 能挑选出最合适的源领域分类器，从而进行更有效的迁移学习。如表 1 所示，概念 A 与概念 C 中的 B 类别均打上了负标签，如果能挖掘出概念 A 和概念 C 所具有的这种局部相似性，就可以从分类器池中挑选由概念 A 训练的分类器 f_a，从而实现对概念 C 中的 B 类别正确分类。同理，概念 B 与概念 C 中的 A 类别均打上了正标签，如果能从分类器池中挑选由概念 B 训练的分类器 f_b，则可以对概念 C 中的 A 类别正确分类。

本文实验在 3.6 GHz Intel Core i5-7700 的 CPU、16 GB RAM 的 Windows 7 计算机上进行。

2.2　实验结果分析

2.2.1　算法的有效性

为了评估 LA-MS-CDC 在多源在线迁移学习中的迁移学习能力，本文选择了两种多源在线迁移学习算法 OTLMS 算法[10]、LC-MSOTL 算法[18]以及一种在线学习算法 PA[19]作为对比算法。其中，PA 算法直接在线学习目标领域样本；LA-MS-CDC 算法、OTLMS 算法、LC-MSOTL 算法的实验方案为预先训练 2 个源领域分类器，然后再学习目标领域样本。

为了验证 LA-MS-CDC 算法在概念漂移数据流算法中的优越性，本文选择 Condor SVM 算法[16]作为对比算法。Condor SVM 算法能自适应挖掘历史数据中对当前预测有用的知识，是目前性能较优秀的概念漂移数据流算法之一。

LC-MS-CDC 算法在多源迁移学习和概念漂移数据流分类问题中的有效性实验结果见表 3。为了减弱数据流中样本出现次序的随机性对实验结果的影响，在每次实验前，每份实验数据随机打乱顺序，共打乱顺序 10 次，最终结果为 10 次实验结果的平均值，并且对表 3 的实验结果进行算法稳定性检验和累计准确率显著性差异分析。使用累计准确率的标准差表示各算法的稳定性，使用 95% 置信水平下的成对 t 检验分析 LA-MS-CDC 算法和基线方法的分类性能是否存在显著性差异。

表 3　有效性实验结果

Table 3　Experimental results of effectiveness　　**Unit：%**

数据集 Dataset	迁移学习 Transfer learning				概念漂移 Concept drift	
	PA[19]	OTLMS[10]	LC-MSOTL[18]	LA-MS-CDC	Condor SVM[16]	LA-MS-CDC
ABCD	68.31±0.01■	69.42±0.01■	88.74±0.05■	**92.52±0.12**	87.13±0.01■	**88.02±0.01**
EFGH	65.43±0.01■	66.95±0.01■	84.11±0.01□	**88.64±0.02**	84.27±0.01□	**84.36±0.02**
NOPQ	63.87±0.01■	65.75±0.01■	84.51±0.07■	**94.88±0.11**	86.24±0.01■	**87.02±0.01**
waveform	77.89±0.01■	79.03±0.01■	88.38±0.02■	**95.56±0.10**	80.73±0.01■	**85.68±0.01**

Note：bold represents the result of LC-MS-CDC. ■ represents LA-MS-CDC significantly better than the baseline method ($P<0.05$)；□ represents LA-MS-CDC significantly worse than the baseline method ($P<0.05$).

表 3 的“迁移学习”部分验证了 LA-MS-CDC 算法的迁移学习效果，“概念漂移”部分验证了 LA-MS-CDC 算法在概念漂移数据流环境中的学习性能。实验结果表明，LA-MS-CDC 算法在分类效果上优于迁移学习领域的基线方法和数据流领域的基线方法，并且在分类准确率上显著优于对比算法，这是 LA-MS-CDC 算法中选择性增量更新策略、分类器池多样性最大化策略以及局部分类精度策略共同作用的结果。

2.2.2　选择性增量更新策略的有效性

为了验证 LA-MS-CDC 算法增量更新的有效性，本文在 Letter 数据集、waveform 数据集上对比了 3 种源领域分类器增量更新方法：选择性增量更新源领域分类器（update）、不增量更新源领域分类器（no-update）、无限制增量更新源领域分类器（no_limit_

update)。3 种源领域增量更新方法在 Letter 数据集、waveform 数据集上的有效性见图 2。选择性增量更新源领域分类器(update)的分类效果最佳。update 的分类效果都优于 no-update 的分类效果。

选择性增量更新源领域分类器(update)与无限制增量更新源领域分类器(no_limit_update)的实验结果说明,LA-MS-CDC 算法的选择性更新策略能有效减少增量学习中因学习的知识冲突而影响分类效果的情况。

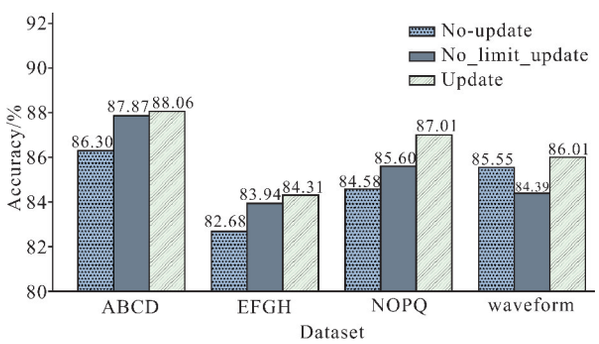

图 2　增量更新有效性对比

Fig. 2　Comparison of incremental update effectiveness

2.2.3　维护分类器池多样性最大化更新策略的有效性

为了综合评估 LA-MS-CDC 算法在维护分类器池多样性最大化的有效性,本文设置 5 种分类器池更新策略:No-delete,不限制分类器池大小;Delete-oldest,限制分类器池大小,删除最老分类器;Delete-svm,限制分类器池大小,并依据支持向量差异性最大化删除分类器;Q,限制分类器池大小,并依据 Q 统计维护分类器池多样性最大化;kw,限制分类器池大小,并依据多样性度量 Kohavi-Wolpert 方差维护分类器池多样性最大化。其中,第 4 种分类器池更新策略为成对多样性度量中的经典方法,第 5 种分类器池更新策略为非成对多样性度量中的经典方法,实验结果见图 3。使用支持向量差异性维护分类器池多样性最大化,能够得到与分类器池无限大相近的实验结果,这说明分类器池中的各分类器确实存在冗余,并且本文提出的分类器池更新策略可以最大限度地保留对后续分类任务有帮助的分类器。对比删除最老分类器的分类器池更新策略,Delete-svm 策略明显具有优越性。

2.2.4　基于 k-means 聚类的局部分类精度算法的有效性

k-means 聚类算法与局部分类精度算法相结合,

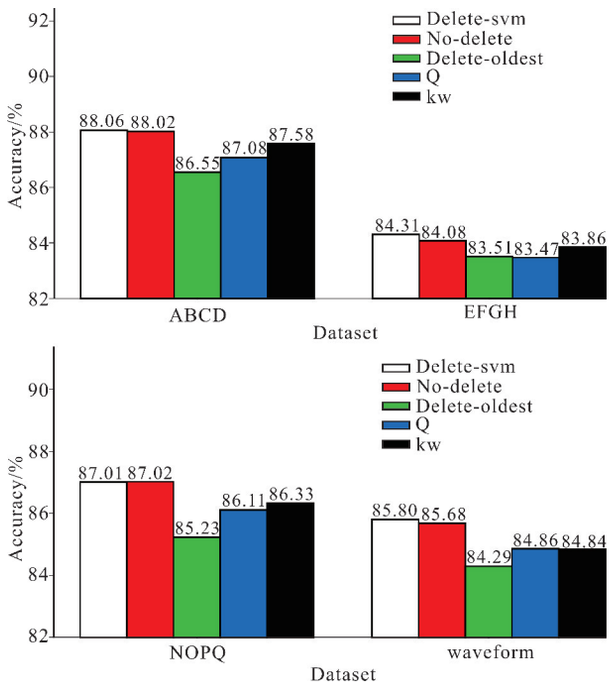

图 3　不同分类器池更新算法分类准确率对比

Fig. 3　Comparison of classification accuracy of different classifier pool update algorithms

能把过去的有效数据信息保存为簇信息,将簇信息与当前样本的信息相结合,进而实现了查找邻居的准确性。为证明基于 k-means 聚类的局部分类精度算法能更准确地查找拟分类样本的邻居样本,在 ABCD 数据集上设计对照实验。实验分为两组,一组采用基于 k-means 聚类的局部分类精度算法,一组采用 LC-MSOTL 算法中的局部分类精度算法。为了区分样本,按照样本的到来顺序给样本编号,例如待分类样本编号 523,即为第 523 个到来的样本,通过列举 4 个拟分类样本在这两种实验方案中所查找到的邻居样本编号,以及邻居样本与拟分类样本是否属于同一概念,来验证 LA-MS-CDC 算法的有效性。

由表 4 可知,对于 523 号样本,使用基于 k-means 的局部分类精度算法找到的邻居样本为 471、392、422 和 495 号样本。其中,与 523 号样本属于同一概念的样本是 471、422 和 495 号样本,与 523 号样本属于不同概念的有 392 号样本。对比 LC-MSOTL 算法的局部分类精度算法,该算法找到的邻居样本中的 392、379 号样本与待分类样本(编号 523)属于不同概念,不同概念邻居样本的数量将影响局部分类精度计算的准确性[式(1)]。同时,采用基于 k-means 的局部分类精度的实验组在 ABCD 数据集上的分类准确率为 8876%,采用 LC-MSOTL 算法的局部分类精度的实验组在ABCD数据集上的分类准确率为

表 4　查找邻居样本的有效性

Table 4　Validate the effectiveness of finding neighbor samples

待分类样本编号 Number of samples to be classified	基于 k-means 的局部分类精度 Local classification accuracy based on k-means			LC-MSOTL 算法的局部分类精度 Local classification accuracy based on LC-MSOTL		
	邻居样本编号 Neighbor number	同概念样本编号 Same concept number	不同概念样本编号 Different concept number	邻居样本编号 Neighbor number	同概念样本编号 Same concept number	不同概念样本编号 Different concept number
523	471,392,422,495	471,422,495	392	471,392,422,495,379	471,422,495	392,379
567	522,387,507,511	522,507,511	387	522,387,507,361,511	522,507,511	387,361
2953	2938,2910,2837	2938,2910,2837	None	2938, 2910, 2823, 2756, 2837	2938, 2910, 2823, 2837	2756
2997	2885,2882,2780	2885,2882	2780	2763, 2771, 2885, 2822, 2780	2885,2822	2763,2771, 2780
Accumulated accuracy		88.76%			87.05%	

87.05%。因此，在查找拟分类样本的邻居时，基于 k-means 聚类的局部分类精度算法与 LC-MSOTL 算法的局部分类精度算法相比，能通过簇心保存更多有效的历史数据信息，从而更准确地查找到拟分类样本的邻居样本，进而提高分类准确率。

2.2.5　参数敏感性分析

为了验证组件分类器的最近邻数 K、源领域与目标领域的权重比值 θ 对算法的影响，本文设置不同的参数值，分析各参数对 LA-MS-CDC 算法分类准确率的影响。关于最近邻数 K、源领域与目标领域权重比值 θ 的敏感性实验结果分别见图 4、图 5。

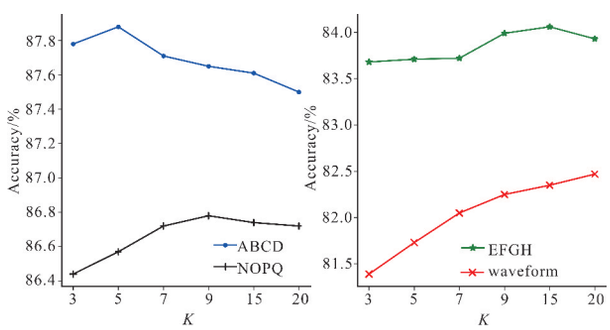

图 4　不同数据集上最近邻数 K 对分类准确率的影响

Fig. 4　Influence of nearest neighbor K on classification accuracy on different datasets

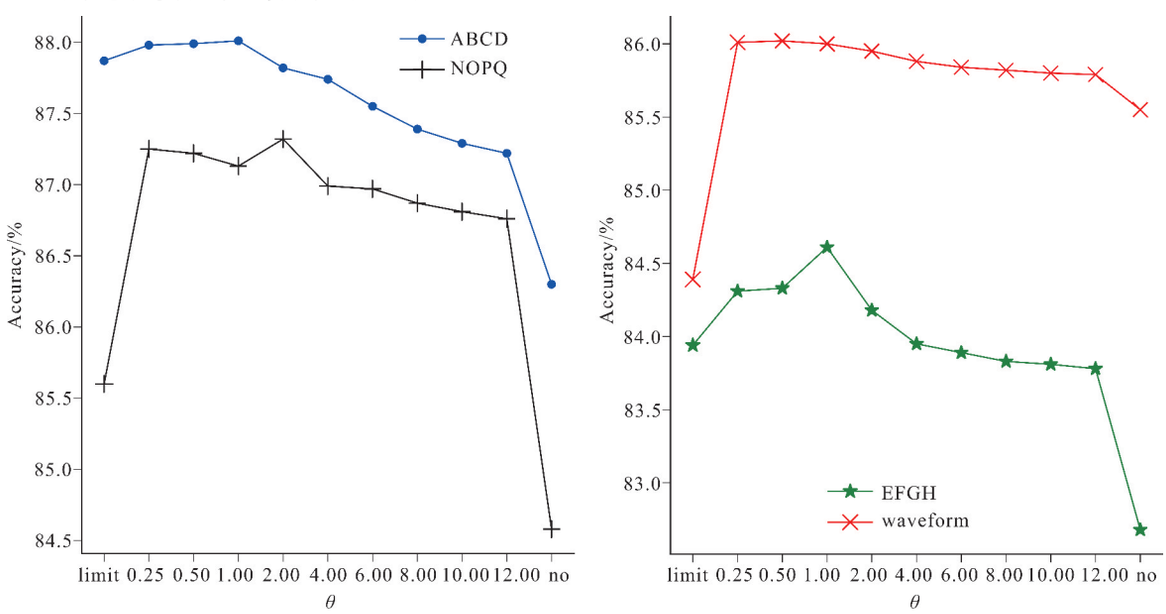

图 5　源领域与目标领域权重比值 θ 对分类准确率的影响

Fig. 5　Influence of the weight ratio θ between source domain and target domain on classification accuracy

实验结果表明最近邻数 K、源领域与目标领域权重比值 θ 在不同的数据集上有不同的取值,取值的差异性由数据集本身的特性决定。分类准确率的波峰所对应的参数值,就是该数据集下的最优参数值。在参数 θ 的敏感性分析中,横坐标 limit 代表无限制增量更新源领域分类器,no 代表不增量更新源领域分类器横坐标从左向右表示增量更新源领域分类器的条件越来越严格。更新源领域分类器的严格程度与数据集中概念的相似度有关。由图 5 可知,分类准确率随着增量更新条件的严苛程度而逐渐降低,这说明了增量更新源领域分类器能使得源领域分类器获得更好的学习效果,进而提高分类准确率。

3 结论

本文提出一种新型的基于局部分类精度的多源在线迁移学习算法 LA-MS-CDC,用于概念漂移数据流分类。LA-MS-CDC 算法首先将 k-means 聚类算法与局部分类精度算法相结合,通过提取过去的有效信息,更准确地查找到拟分类样本的邻居样本,解决了 LC-MSOTL 算法最近邻居查找不准确的问题。为进一步应对概念漂移数据流场景,LA-MS-CDC 算法使用支持向量差异性最大化维护分类器池多样性最大化,使得分类器池最大限度地映射多种概念,并用实验证明了这种多样性指标的有效性。同时,为降低增量学习中因学习的知识相冲突造成分类性能变差的问题,LA-MS-CDC 算法依据源领域与目标领域的权重比值,选择性增量更新最优源领域分类器,实验说明这种增量更新机制的有效性。在未来的工作中,挖掘源领域和目标领域的局部相似性,并将有效的源领域知识迁移到目标域上还有待进一步研究。

参考文献

[1] 文益民,刘帅,缪裕青,等. 概念漂移数据流半监督分类综述[J]软件学报,2022,33(4):1287-1314.

[2] LIN Y H,CHANG L. An online transfer learning framework for time-varying distribution data prediction [J]. IEEE Transactions on Industrial Electronics, 2022, 69(6):6278-6287.

[3] ZHUANG F Z,QI Z Y,DUAN K Y,et al. A comprehensive survey on transfer learning [J]. Proceedings of the IEEE,2021,109(1):43-76.

[4] 庄福振,罗平,何清,等. 迁移学习研究进展[J]. 软件学报,2015,26(1):26-39.

[5] NGUYEN C V,LE K H,TRAN A M,et al. Learning for amalgamation: a multi-source transfer learning framework for sentiment classification [J] Information Sciences,2022,590:1-14.

[6] EATON E,DESJARDINS M. Selective transfer between learning tasks using task-based boosting [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2011,25(1):337-343

[7] CHATTOPADHYAY R,YE J P,PANCHANATHAN S,et al. Multi-source domain adaptation and its application to early detection of fatigue [C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Diego: ACM,2011:717-725.

[8] GAO J,FAN W,JIANG J,et al. Knowledge transfer via multiple model local structure mapping [C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas: ACM,2008:283-291.

[9] GE L,GAO J,ZHANG A D. OMS-TL:a framework of online multiple source transfer learning [C]//Proceedings of the 22nd ACM International Conference on Information and Knowledge Management. San Francisco: ACM,2013:2423-2428.

[10] WU Q Y,ZHOU X M,YAN Y G,et al. Online transfer learning by leveraging multiple source domains [J]. Knowledge and Information Systems,2017,52(3):687-707.

[11] KANG Z,YANG B,YANG S,et al. Online transfer learning with multiple source domains for multi-class classification [J]. Knowledge-Based Systems, 2020, 190:105149.

[12] 文益民,唐诗淇,冯超,等. 基于在线迁移学习的重现概念漂移数据流分类[J]. 计算机研究与发展,2016,53(8):1781-1791

[13] 秦一休,文益民,何倩. 概念漂移数据流分类中的多源在线迁移学习算法[J]. 计算机科学,2019,46(1):64-72.

[14] SUN Y,TANG K,ZHU Z X,et al. Concept drift adaptation by exploiting historical knowledge [J]. IEEE Transactions on Neural Networks and Learning Systems,2018,29(10):4822-4832.

[15] 吕艳霞,刘波男,王翠荣,等. 面向概念漂移数据流的自适应增量集成分类算法[J]. 小型微型计算机系统,2019,40(12):2624-2630.

[16] ZHAO P,CAI L W,ZHOU Z H. Handling concept drift via model reuse [J]. Machine Learning, 2020, 109:533-568.

[17] 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.

[18] 唐诗淇,文益民,秦一休. 一种基于局部分类精度的多源在线迁移学习算法[J]. 软件学报,2017,28(11):

2940-2960.
[19] CRAMMER K,DEKEL O,KESHET J,et al. Online passive-aggressive algorithms [J]. Journal of Machine Learning Research,2006,7:551-585.
[20] ZHOU Z H,LI M. Semi-supervised regression with co-training [C]//Proceedings of the 19th International Joint Conference on Artificial Intelligence. NewYork: ACM,2005:908-913.
[21] ZHOU Z H,LI M. Semi-supervised regression with co-training-style algorithms [J]. IEEE Transactions on Knowledge and Data Engineering,2007,19(11):1479-1493.
[22] WOODS K,KEGELMEYER W P,BOWYER K. Combination of multiple classifiers using local accuracy estimates [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1997,19(4):405-410.
[23] 熊旺.数据流增量学习关键技术研究[D].长沙:国防科技大学,2018.
[24] ZHAO P L,HOI S C H. OTL:a framework of online transfer learning [C]//Proceedings of the 27th International Conference on Machine Learning. Haifa: Omnipress,2010:1231-1238.
[25] HANSEN L K,SALAMON P. Neural network ensembles [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1990,12(10):993-1001.
[26] DIETTERICH T G. Ensemble learning [M]. Cambridge: The MIT Press,2002.
[27] KUNCHEVA L I,WHITAKER C J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy [J]. Machine Learning,2003,51:181-207.
[28] ALBERT B,GEOFF H,RICHARD K,et al. MOA: massive online analysis [J]. Journal of Machine Learning Research,2010,11:1601-1604.

A Concept Drift Data Stream Classification Algorithm Based on Local Classification Accuracy

ZHANG Ling,MA Shilun,LI Lihui,WEN Yimin**
(Guangxi Key Laboratory of Image and Graphic Intelligent Processing,Guilin University of Electronic Technology,Guilin,Guangxi,541004,China)

Abstract:The classification of concept drift data streams is a challenging problem. When a new concept appears,there are too few learning samples of the concept,and the classifier cannot be adjusted in time,which leads to low classification accuracy. In order to solve this problem,this article proposes a concept drift data stream classification algorithm,called LA-MS-CDC,based on local classification accuracy. Firstly,LA-MS-CDC combines k-means clustering and local classification accuracy algorithm to select the optimal source domain classifier from the classifier pool. Secondly,the optimal source domain classifier and the target domain classifier are weighted and integrated to classify the samples. Then,according to the real labels of the classification samples,the loss of each classifier is calculated respectively and the weights of the classifiers in the target domain and the source domain are updated. Then,the classification samples are used to update the target domain classifier and the optimal source domain classifier. Finally,the update of the classifier pool is completed. The experimental results on the public datasets show that LA-MS-CDC can effectively transfer the source domain knowledge to the target domain,and the classification effect of LA-MS-CDC is significantly improved compared with the existing methods. The algorithm code can be obtained on https://gitee.com/ymw12345/LAMSCDC.

Key words:concept drift;multi-source online transfer learning;local classification accuracy;ensemble learning;diversity

责任编辑:陆　雁,陈少凡