



基于动态策略的多源迁移学习数据流分类研究

周 胜, 刘三民

(安徽工程大学 计算机与信息学院, 安徽 芜湖 241000)

摘 要: 为解决数据流分类中的概念漂移和噪声问题, 提出一种基于样本确定性的多源迁移学习方法。该方法存储多源领域上由训练得到的分类器, 求出各源领域分类器对目标领域数据块中每个样本的类别后验概率和样本确定性值。在此基础上, 将样本确定性值满足当前阈值限制的源领域分类器与目标领域分类器进行在线集成, 从而将多个源领域的知识迁移到目标领域。实验结果表明, 该方法能够有效消除噪声数据流给不确定分类器带来的不利影响, 与基于准确率选择集成的多源迁移学习方法相比, 具有更高的分类准确率和抗噪稳定性。

关键词: 数据流分类; 多源迁移学习; 类别后验概率; 样本确定性; 集成学习

开放科学(资源服务)标志码(OSID):



中文引用格式: 周胜, 刘三民. 基于动态策略的多源迁移学习数据流分类研究[J]. 计算机工程, 2020, 46(5): 139-143, 149.

英文引用格式: ZHOU Sheng, LIU Sanmin. Research on multi-source transfer learning data streams classification based on dynamic strategy[J]. Computer Engineering, 2020, 46(5): 139-143, 149.

Research on Multi-Source Transfer Learning Data Streams Classification Based on Dynamic Strategy

ZHOU Sheng, LIU Sanmin

(College of Computer and Information, Anhui Polytechnic University, Wuhu, Anhui 241000, China)

[Abstract] To address concept drift and noise in data stream classification, this paper proposes a multi-source transfer learning method based on sample certainty. First, the method stores classifiers trained in the multi-source domain. Then the method calculates the category posterior probability and sample certainty of each source domain classifier to each sample in the target domain data block. On this basis, the source domain classifiers of which the sample certainty satisfies the current threshold limit are integrated with target domain classifiers online, so as to transfer the knowledge of multi-source domains to the target domain. Experimental results show that the proposed method can effectively eliminate the adverse effects of noisy data streams on uncertain classifiers, and has better classification accuracy and anti-noise stability than the multi-source transfer learning methods based on accuracy selection integration.

[Key words] data streams classification; multi-source transfer learning; category posterior probability; sample certainty; ensemble learning

DOI: 10.19678/j.issn.1000-3428.0054753

0 概述

随着大数据技术的不断发展, 数据流挖掘在天气预报、金融预测、电子商务等领域得到广泛应用。传统的数据流分类方法需要大量的标签样本来训练分类模型, 而获取大量标签样本费时费力, 且这类方法无法有效解决数据流中的概念漂移和噪声问题。

目前, 将多源迁移学习^[1-2]与集成学习^[3-4]相结合应用于数据流分类的研究得到广泛关注, 该类

方法通过将训练得到的多个源领域分类器进行集成来对目标领域新到样本进行分类, 能够解决源领域和目标领域相似度较低以及目标领域标注样本不足的问题^[5-7]。根据多源领域分类器的不同集成方式, 可以将多源领域分类器集成分为多源领域分类器直接集成和多源领域分类器选择集成两类。

多源领域分类器直接集成的方法将所有源领域分类器进行集成。文献[8]建立一种多源迁移学习算法 OMS-TL, 该算法根据二部图实现对目标领域

基金项目: 安徽省自然科学基金(1608085MF147)。

作者简介: 周 胜(1993—), 男, 硕士研究生, 主研方向为机器学习、迁移学习; 刘三民, 副教授、博士。

收稿日期: 2019-04-28 修回日期: 2019-05-28 E-mail: 934485488@qq.com

样本的预测,通过对数据样本进行重用来实现迁移学习。文献[9]设计 OTLMS 算法,该算法将源领域分类器和目标领域分类器进行组合以构建预测分类器,解决了与目标学习任务无关的噪声源数据影响分类精度的问题。文献[10]通过将目标特征空间分成源域的同构和异构 2 个部分,将基分类器进行加权组合获得多个源领域分类器,并将它们组合成一个集成模型来解决多源异构迁移学习问题。但是,每个源领域都可能不包含目标域的完整类别信息,当多个源领域之间存在较大差异时,将所有源领域分类器进行集成会导致分类性能降低。

针对多源领域分类器直接集成方法的不足,众多国内外学者提出使用多源领域分类器选择集成的方法来解决数据流分类问题。文献[11]基于局部分类精度提出一种多源迁移学习算法 LC-MSOTL,该算法将局部分类精度最高的源领域分类器和目标领域分类器进行加权集成,实验结果显示,该方法具有较高的准确率。针对概念漂移数据流分类中的概念重现问题,文献[12]提出一种重现概念漂移数据流分类算法 RC-OTL,其根据领域相似度挑选最合适的源领域分类器,实验结果表明,RC-OTL 算法能够有效克服“负迁移”问题。文献[13]构建一种多源迭代自适应算法 MSIDA,该算法根据贪婪思想实现最佳源域的选择,同时通过创建额外的伪标记实例来解决样本标注问题。文献[14]通过求解每个源领域对应的权值向量,并用该权值向量表示源领域和目标领域之间的相似度,从而显著地提高了迁移学习的效率以及分类性能。文献[15]提出一种 OHTWC 算法,其通过计算异构域中同现数据的异构相似性,解决了异构域上的数据流分类问题。文献[16]通过调整每个源领域分类器对应的权重,同时用目标领域分类器更换权重最大的源领域分类器,在出现概念漂移后快速恢复集成分类器的分类准确率,最终使集成分类器的分类性能得到改善。

在数据流分类任务中,多源领域分类器选择集成技术具有明显优势,但在多源迁移学习过程中,从多个源领域分类器中如何挑选最合适的分类器进行迁移集成是急需解决的问题。本文提出一种基于样本确定性的动态分类器选择方法,以高效选择最合适的源领域分类器并解决数据流中的概念漂移和噪声问题。

1 基本概念

为便于分析和理解,对本文涉及的基本概念进行定义:

定义 1 数据流指按时间顺序依次到达的 n 个样本的集合,即:

$$S = \{ (x_t, y_t) \mid t = 1, 2, \dots, n \}, (x_t, y_t) \in X \times Y, X \in R^m, Y = \{-1, +1\}$$

其中, X 表示 m 维的特征空间, Y 指样本类别。

定义 2 概念漂移指数据样本产生的联合概率分布函数随时间的推进发生无法预知的变化,即:

$$p_t(x, y) \neq p_{t+1}(x, y)$$

其中, x 表示样本特征信息, y 表示样本所属类别。

定义 3 信息熵用来衡量某随机事件发生的不确定性,其计算如下:

$$E(X) = - \sum_{i=1}^k p_i \lg p_i \quad (1)$$

其中, X 为随机变量, k 为随机变量 X 所有可能发生的事件数目, p_i 为对应事件发生的概率。

对信息熵计算公式分析可知,当概率为 1 时,信息熵最小为 0,相当于确定性事件。对于只有 2 种事件的随机变量,当概率为 0.5 时,信息熵取得最大值,表明无法对当前事件进行有意义的判断。随机事件的信息熵与概率之间的关系如图 1 所示。

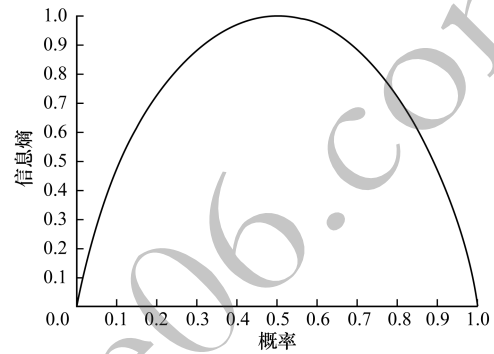


图 1 信息熵与概率之间的关系

Fig.1 Relationship between information entropy and probability

定义 4 样本确定性指分类器对待分类样本 x 预测结果的确定性程度。本文讨论的是二分类问题,结合信息熵的定义可得样本确定性计算公式如下:

$$sc(x) = 1 - E(x) = 1 + \sum_{i=1}^2 p_i \lg p_i \quad (2)$$

其中, p_i 为样本 x 对应的类别后验概率。

通过样本确定性进行数据流分类,可以有效避免不确定分类器对噪声数据流带来的不利影响,同时提高分类准确率。

2 弃权分类器

针对概念漂移数据流分类问题,文献[17-19]提出基于弃权分类器的数据流分类方法,以挑选最合适的基分类器。弃权是指未能达到特定规则要求的基分类器不参与决策,当发生概念漂移时,该方法通过允许基分类器对新实例丢失信心时放弃预测,同时允许选择的基分类器参与决策,从而探索集成的多样性,弃权分类器的结构如图 2 所示。首先计算多个基分类器对新到样本的确定性值,然后强制未能达到阈值限制的基分类器放弃预测,最后采用多数投票的方法对新到样本进行集成决策。该方法强

制不确定分类器(未能达到特定规则要求的分类器)不进行预测,即不确定分类器可以放弃参与最终决策,并将所选择的基分类器的输出结果作为分类器集合的输出。

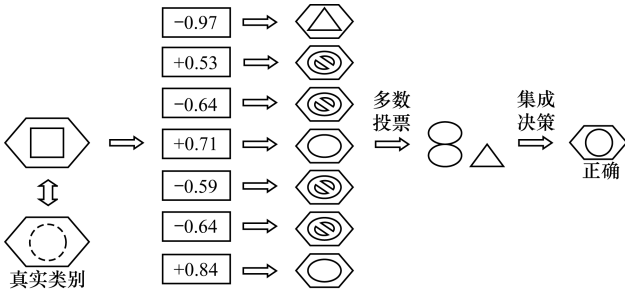


图 2 弃权分类器

Fig. 2 Waiver classifier

弃权分类器的优点在于能够挑选出达到特定规则要求的分类器,更快适应概念漂移现象,有利于消除噪声数据的影响。本文基于弃权分类器,将多个源领域分类器对目标领域数据块中每个样本的确定性值与给定的阈值进行比较,对于未能达到阈值限制的源领域分类器,强制其不参与投票。若所选择的源领域分类器集合能够正确地预测标签,则意味着已经选择了有能力的源领域分类器,可以通过降低阈值来探索其他类似能力的源领域分类器。如果所选择的源领域分类器集合做出不正确的决定,则表示可能发生了概念漂移现象,在这种情况下,需要增加阈值来排除不太合格的源领域分类器,以挑选出最适合当前流状态的源领域分类器。

3 算法描述

在样本确定性和弃权分类器的基础上,本文提出一种基于动态策略的多源迁移学习算法 DSMTL。该算法求得源领域分类器对目标领域数据块中每个样本的类别后验概率,然后根据样本确定性计算方法求得各源领域分类器对目标领域数据块中每个样本的确定性值,最后将样本确定性值满足当前阈值限制的源领域分类器与目标领域分类器进行在线集成,以对目标领域数据块进行分类,其中,使用多数投票的方式进行有关预测标签的最终决策,并根据集成决策的正确性修改阈值。DSMTL 算法的详细描述如下:

算法 1 DSMTL 算法

输入 目标域数据流 DS , 源领域分类器集合 CS , 弃权阈值 θ , 调整因子 S

输出 集成分类模型对目标领域数据块的分类准确率

1. 参数初始化 θ 、 S , 缓存 10 个规模大小相等的数据块, 并分别在数据块上训练源领域分类器
2. For $j = 1, 2, \dots$, 对后续数据块 D_j 依次循环处理
3. 基于目标领域数据块 D_j 构建目标领域分类器 f_j
4. 计算各源领域分类器对目标领域数据块中每个样本 x 的类别后验概率 p_i

5. 计算各源领域分类器对目标领域数据块中每个样本 x 的样本确定性值 sc :

$$sc(x) = 1 + \sum_{i=1}^2 p_i \ln p_i$$

6. 将样本确定性值满足阈值限制的源领域分类器 f_i 与目标领域分类器 f_j 进行在线集成, 以对目标领域数据块 D_j 进行分类

7. 根据集成决策的正确性修改弃权阈值 θ , 如果集成分类器对目标领域数据块 D_j 的分类准确率大于弃权阈值 θ , 则:

$$\theta = \theta \cdot (1 - S)$$

否则:

$$\theta = \theta \cdot (1 + S)$$

8. End For

在算法 1 中, 步骤 1 表示初始化, 步骤 3 表示基于目标领域数据块构建目标领域分类器, 步骤 4 ~ 步骤 5 表示求各源领域分类器对目标领域数据块中每个样本的确定性值, 步骤 6 表示选取样本确定性值满足阈值限制的源领域分类器, 并与目标领域分类器在线集成以对目标领域数据块进行分类, 步骤 7 表示弃权阈值修改。

4 实验结果与分析

4.1 实验数据集

本文实验数据集来自 MOA^[20] 环境中的移动超平面数据集^[21], 其可以用来模拟概念漂移中的缓漂移现象, 通过调节参数能够构建不同漂移程度和噪声比的数据集。 m 维超平面满足 $\sum_{i=1}^m a_i x_i = a_0$, 其中, $a_0 = (\sum_{i=1}^m a_i) / 2$ 。基于该超平面, 当 $\sum_{i=1}^m a_i x_i \geq a_0$ 时, 则由特征值构成的样本 $(x_1, x_2, \dots, x_m)$ 代表正类样本, 否则代表负类样本。本文实验数据集有 10 个属性特征, 其中, 6 个特征发生概念漂移, 权重改变量参数取值为 1, 方向改变百分比为 5%, 按此设置生成数据集 D_1 (包含 30 000 条样本)。为验证算法在噪声环境下的性能, 分别向数据集中加入 10% 和 20% 的噪声, 形成对应数据集 D_2 和 D_3 。

4.2 结果分析

本文将所提方法与基于准确率选择集成的多源迁移学习方法 ASIMTL^[4] 进行对比。实验采用 Bayes 分类器作为基分类器, 通过批处理模式生成数据块以及训练基分类器, 其中, 源领域数据块大小为 2 000, 源领域数据块个数为 10, 训练 10 个源领域分类器, 同时基于目标域数据流形成 20 个数据块, 数据块大小设为 500, 求得各源领域分类器对目标领域数据块中每个样本的确定性值, 将样本确定性值满足阈值限制的源领域分类器与目标领域分类器进行在线集成, 从而对目标领域数据块进行分类。

实验 1 验证弃权阈值参数的影响。为验证弃权阈值对算法的影响, 本文选择 3 个不同的阈值, 将平均准确率和标准差统计量作为评价指标, 固定调整因子的值为 0.01, 实验结果如表 1 所示。

表 1 弃权阈值对 DSMTL 算法性能的影响

Table 1 Effect of waiver threshold on performance of DSMTL algorithm

弃权阈值	平均准确率	标准差统计量
0.644	0.700 8	0.041 7
0.650	0.708 6	0.028 3
0.660	0.708 2	0.032 7

从表 1 可以看出,当弃权阈值为 0.650 时,算法平均准确率和标准差统计量较优。当弃权阈值较大时,每次迭代选择的源领域投票分类器数量不够,而阈值较小时选择的源领域投票分类器可能并不适合当前流状态,弃权阈值较大或较小都会导致无法挑选出最合适的源领域分类器,从而影响算法的分类性能。

实验 2 验证 DSMTL 算法的分类性能。图 3 所示为 DSMTL 和 ASIMTL 2 种算法在数据集 D_1 上采用先测试后训练策略所得的分类准确率情况。从图 3 可以看出,在无噪声的情况下,相对于 ASIMTL 算法,DSMTL 算法的分类准确率有明显提高,其能够及时发现概念漂移并且挑选出最合适的源领域分类器以处理新出现的概念,如在数据块 2 和数据块 5 之间出现了分类准确率持续下降的情况,数据块 12 和数据块 14 之间出现了曲线低峰,但算法能够快速恢复其识别准确率,说明 DSMTL 算法能够很好地适应概念漂移情况,即通过样本确定性值选择源领域分类器的策略能够有效改善迁移效果。

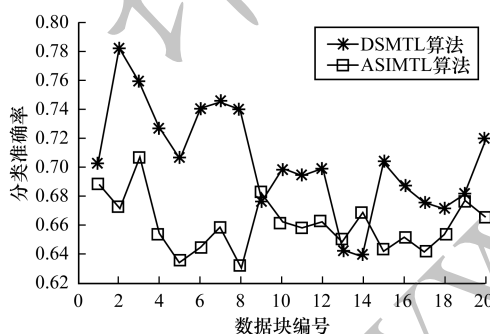
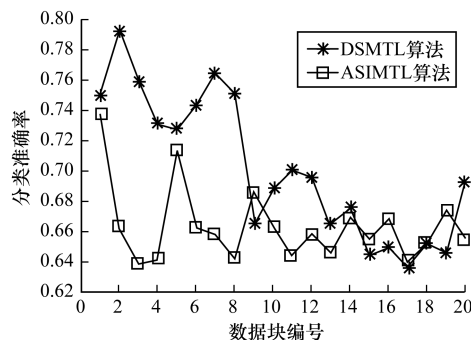
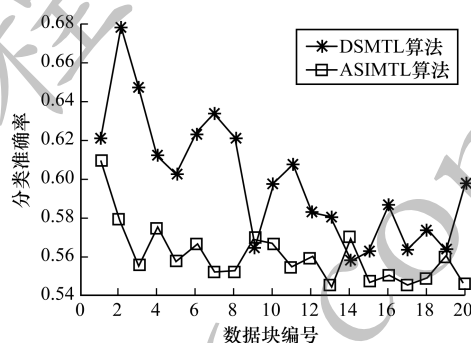
图 3 数据集 D_1 上的实验结果Fig. 3 Experimental results on dataset D_1

图 4、图 5 是噪声环境下 2 个数据集上的实验结果,从图 4、图 5 可以看出,在有噪声的情况下,DSMTL 算法仍然具有较高的分类准确率,优于 ASIMTL 算法,说明其具备一定的抗噪性能,这是由于 DSMTL 算法使用了信息熵求样本确定性值,其消除了噪声数据的影响。但随着噪声数据的增加,DSMTL 算法分类准确率出现较大波动,这是由于噪声样本被误认为是概念漂移样本用以构建目标领域分类器,导致集成分类器对新概念无法收敛。

图 4 数据集 D_2 上的实验结果Fig. 4 Experimental results on dataset D_2 图 5 数据集 D_3 上的实验结果Fig. 5 Experimental results on dataset D_3

综上可知,DSMTL 数据流分类算法可行,分类准确率优于 ASIMTL 算法,原因是 DSMTL 算法将求得各源领域分类器对目标领域数据块中每个样本的确定性值与一个给定的阈值进行比较,只选择满足当前阈值限制的源领域投票分类器,即强制不确定的分类器不参与投票,目的是为投票步骤选择最不可能受噪声数据影响的源领域分类器。因此,该算法能够通过设置弃权阈值的方式挑选出最合适的源领域分类器,且其策略可以消除噪声数据的影响,更快地适应概念漂移现象,使分类模型面对概念漂移问题时具有更好的泛化能力。

从表 2 可以看出,DSMTL 算法性能明显优于 ASIMTL 算法,平均准确率约高出 5 个百分点,在噪声环境下仍然具有较高的准确率,说明 DSMTL 算法能够有效地处理数据流中的噪声,更快地适应概念漂移现象。当数据集由不含噪声的数据变为含有噪声的数据时,DSMTL 算法的分类准确率下降程度小于 ASIMTL 算法,说明 DSMTL 算法面对噪声数据流时具有更好的稳定性和更强的抗噪性。原因是 DSMTL 算法利用信息熵求得各源领域分类器对目标领域数据块中每个样本的确定性值,有效消除了噪声数据的影响,使分类模型保持较高的分类精度以及较好的稳定性。

表2 2种算法平均准确率和标准差统计量对比

Table 2 Comparison of average accuracy and standard deviation statistics of two algorithms

数据集	平均准确率		标准差统计量	
	ASIMTL	DSMTL	ASIMTL	DSMTL
D_1	0.660 8	0.705 4	0.014 6	0.027 9
D_2	0.602 8	0.656 2	0.017 1	0.040 8
D_3	0.543 8	0.600 2	0.011 3	0.025 6

5 结束语

本文结合弃权分类器和样本确定性计算方法,提出一种多源迁移数据流分类学习算法 DSMTL。该算法能够有效利用弃权分类器思想强制不确定分类器不进行预测,同时计算样本确定性值以从源领域分类器集合中挑选最合适的分类器,并与目标领域分类器进行在线集成。实验结果表明,DSMTL 算法能够对集成的多样性进行选择控制,消除噪声数据的影响,并解决数据流中概念变化和样本标注问题。下一步将对源领域与目标领域数据的共有特征进行识别,根据新到达样本对分类模型进行更准确的权重设置,并探究对源领域和目标领域之间的区分性进行建模的方法。

参考文献

- [1] SUN Shiliang, SHI Honglei, WU Yuanbin. A survey of multi-source domain adaptation[J]. Information Fusion, 2015, 24: 84-92.
- [2] JI Dingcheng, JIANG Yizhang, WANG Shitong. Multi-source transfer learning method by balancing both the domains and instances[J]. Acta Electronica Sinica, 2019, 47(3): 692-699. (in Chinese)
季鼎承, 蒋亦樟, 王士同. 基于域与样例平衡的多源迁移学习方法[J]. 电子学报, 2019, 47(3): 692-699.
- [3] KRAWCZYK B. Active and adaptive ensemble learning for online activity recognition from data streams[J]. Knowledge-Based Systems, 2017, 138(15): 69-78.
- [4] HOSSEINI M J, GHOLIPOUR A, BEIGY H. An ensemble of cluster-based classifiers for semi-supervised classification of non-stationary data streams[J]. Knowledge and Information Systems, 2016, 46(3): 567-597.
- [5] SHU Xing, YU Huimin, ZHENG Weiwei, et al. Classifier-designing algorithm on a small dataset based on margin Fisher criterion and transfer learning[J]. Acta Automatica Sinica, 2016, 42(9): 1313-1321. (in Chinese)
舒醒, 于慧敏, 郑伟伟, 等. 基于边际 Fisher 准则和迁移学习的小样本集分类器设计算法[J]. 自动化学报, 2016, 42(9): 1313-1321.
- [6] AGHAMALEKI J A, BAHARLOU S M. Transfer learning approach for classification and noise reduction on noisy Web data[J]. Expert Systems with Applications, 2018, 105: 221-232.
- [7] HANG Wenlong, JIANG Yizhang, LIU Jiefang, et al. Transfer affinity propagation clustering algorithm[J]. Journal of Software, 2016, 27(11): 2796-2813. (in Chinese)
杭文龙, 蒋亦樟, 刘解放, 等. 迁移近邻传播聚类算法[J]. 软件学报, 2016, 27(11): 2796-2813.
- [8] GE Liang, GAO Jing, ZHANG Aidong. OMS-TL: a framework of online multiple source transfer learning[C]//Proceedings of the 22nd ACM International Conference on Information and Knowledge Management. New York, USA: ACM Press, 2013: 2423-2428.
- [9] WU Qingyao, ZHOU Xiaoming, YAN Yuguang, et al. Online transfer learning by leveraging multiple source domains[J]. Knowledge and Information Systems, 2017, 52(3): 1-21.
- [10] WU Qingyao, WU Hanrui, ZHOU Xiaoming, et al. Online transfer learning with multiple homogeneous or heterogeneous sources[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(7): 1494-1507.
- [11] TANG Shiqi, WEN Yimin, QIN Yixiu, et al. Online transfer learning from multiple sources based on local classification accuracy[J]. Journal of Software, 2017, 28(11): 2940-2960. (in Chinese)
唐诗淇, 文益民, 秦一休, 等. 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.
- [12] WEN Yimin, TANG Shiqi, FENG Chao, et al. Online transfer learning for mining recurring concept in data stream classification[J]. Journal of Computer Research and Development, 2016, 53(8): 1781-1791. (in Chinese)
文益民, 唐诗淇, 冯超, 等. 基于在线迁移学习的双重概念漂移数据流分类[J]. 计算机研究与发展, 2016, 53(8): 1781-1791.
- [13] BHATT H S, RAJKUMAR A, ROY S. Multi-source iterative adaptation for cross-domain classification[C]//Proceedings of International Joint Conference on Artificial Intelligence. New York, USA: ACM Press, 2016: 3691-3697.
- [14] BIAN Zekang, WANG Shitong. Similarity-learning based multi-source transfer learning algorithm[J]. Control and Decision, 2017, 32(11): 1941-1948. (in Chinese)
卞则康, 王士同. 基于相似度学习的多源迁移算法[J]. 控制与决策, 2017, 32(11): 1941-1948.
- [15] YAN Yuguang, WU Qingyao, TAN Minghui, et al. Online heterogeneous transfer learning by weighted offline and online classifiers[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2016: 467-474.
- [16] QIN Yixiu, WEN Yimin, HE Qian. Multi-source online transfer learning algorithm for classification of data streams with concept drift[J]. Computer Science, 2019, 46(1): 64-72. (in Chinese)
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.

(上接第 143 页)

- [17] BLASZCZYNSKI J, STEFANOWSKI J, ZAJAC M. Ensembles of abstaining classifiers based on rule sets [C]//Proceedings of International Symposium on Methodologies for Intelligent Systems. Berlin, Germany: Springer, 2009: 382-391.
- [18] PIETRASZEK T. On the use of ROC analysis for the optimization of abstaining classifiers [J]. Machine Learning, 2007, 68(2): 137-169.
- [19] PIETRASZEK T. Classification of intrusion detection alerts using abstaining classifiers [J]. Intelligent Data Analysis, 2007, 11(3): 293-316.
- [20] HOLMES G, KIRKBY R, PFAHRINGER B. MOA: massive online analysis [EB/OL]. [2019-03-25]. <http://sourceforge.net/projects/moa-datastream>.
- [21] HULTEN G, SPENCER L, DOMINGOS P. Mining time-changing data streams [C]//Proceedings of International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2001: 97-106.

编辑 吴云芳