

DOI:10.13232/j.cnki.jnju.2020.01.008

一种用于数据流自适应分类的主动学习方法

张银芳¹, 于洪^{1*}, 王国胤¹, 谢永芳²

(1. 重庆邮电大学计算智能重庆市重点实验室, 重庆, 400065; 2. 中南大学信息科学与工程学院, 长沙, 410083)

摘要:概念漂移会导致数据流分类模型的分类能力随时间发展而下降,这就要求分类模型有自适应的能力. 现有的大多数自适应概念漂移的数据流分类模型往往假设数据输入分类模型得到预测标签之后就可以得到其真实标签,但这种假设在某些情况下是不合理的,因为数据标记往往成本高、耗时长. 因此,针对数据流少量标签的问题,在考虑主动学习可能出现采样偏差的情况下,结合不确定性主动学习策略以及边界点和离群点检测方法(Boundary and Outlier Detection, BOD),提出一种新的主动学习方法 ALBOD(Active Learning Based on Boundary and Outlier Detection). 比较实验的结果表明,在概念漂移发生的情况下,与 100% 标记算法 OzaBagAdwin(OBA)和 HoeffdingAdaptiveTree(HAT)相比,ALBOD 主动学习方法只需要平均 20% 左右的标签就可以使分类器保持同等分类精度,说明新方法 ALBOD 有良好的主动学习能力.

关键词:数据流,概念漂移,主动学习,自适应分类

中图分类号:TP391

文献标识码:A

An active learning method for data stream adaptive classification

Zhang Yinfang¹, Yu Hong^{1*}, Wang Guoyin¹, Xie Yongfang²

(1. Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China; 2. School of Information Science and Engineering, Central South University, Changsha, 410083, China)

Abstract: Concept drift will cause the ability of data stream classification model to decrease with time, which requires the classification model with the ability of self-adaptation. However, most of the existing data stream classification models adapting to concept drift ignore the limited label of data stream. They usually assume that the coming data input to classification model can get the real label after the predicted label obtained. But, this assumption is unreasonable in some cases, as labeling data tends to be costly and time-consuming. In this paper, a new active learning method, ALBOD (Active Learning based on Boundary and Outlier Detection), is proposed to solve the problem of the scarcity of data stream labels. Our method considers the problem of sampling bias which may occur in the process of active learning by combining the uncertainty active learning method with the BOD (Boundary and Outlier Detection) method. The first criterion is the most common active learning criterion, which selects instances that are the most uncertain in terms of class membership. The latter curbs the sampling bias by using the fact that boundary samples and outliers can reflect the feature space. Compared to the 100% label algorithm OzaBagAdwin (OBA) and HoeffdingAdaptiveTree (HAT), the proposed algorithm ALBOD can make the classification model maintain the same accuracy learning an average of about 20% labels under concept drift. The

基金项目:国家自然科学基金(61876027, 61751312, 61533020)

收稿日期:2019-08-01

* 通讯联系人, E-mail: yuhong@cqupt.edu.cn

experiments show that though ALBOD is a sample combination of the above two criterions, it has a good active learning ability.

Key words: data stream, concept drift, active learning, adaptive classification

概念漂移是现实数据流应用中普遍存在的问题,1986年由Schlimmer and Granger^[1]首次提出.现在预测分析和机器学习中的概念漂移大多指目标变量的统计特性随时间以不可预见的方式变化的现象.目前已有大量关于概念漂移的研究^[2].概念漂移会导致数据流分类模型的能力随时间的变化而下降,所以数据流的分类模型要保持其分类能力就必须能适应数据流中的概念漂移.

如何适应数据流中的概念漂移有两种策略:一种是消极策略,不管概念漂移有没有发生,定时更新模型;一种是积极主动策略,以检测概念漂移为触发器,检测到概念漂移之后进行模型的更新^[3].现在通常采用积极主动策略以减少模型的更新次数,节约成本.

现有的以检测概念漂移为基础的数据流自适应分类模型大多是基于监督学习的,它们基于一种假设:数据流中的数据输入分类模型得到预测标签之后可以获得其全部的真实标签,然后在此基础上检测概念漂移,更新分类模型^[4-5].但数据的真实标签通常由用户或者专业人员人工标注,耗时耗力,说明现有的多数数据流自适应分类模型忽略了一个实际问题,即少量标签问题.

针对数据流少量标签问题,需要标记哪些数据有两种策略:随机标记^[6]与主动学习.基于不确定性^[7]是常用的主动学习策略.不确定性的衡量通常可以采用信息熵、置信度以及边界法等方法^[8-9].这类主动学习方法选择的样本通常位于决策的边界,意味着采样偏向分类器的决策边界,但概念漂移可能在任何时间发生在特征空间的任何位置,导致数据流主动学习可能出现采样偏差(sampling bias)^[10],即主动学习的样本所代表的分布与数据本身的分布发生偏离.目前已有少量考虑采样偏差的主动学习方法.Zliobaite et al^[11]提出带有随机性的不确定性策略和分裂策略.Mohamad et al^[10]提出基于不确定性和密度的双准则主动选择策略BAL(Bi-Criteria Active Learn-

ing),还提出基于能够减少未来期望误差(reduce the future expected error)选择样本的主动学习策略SAL(Stream Active Learning)^[12].

模式选择是针对一个有 N 个模式的数据集 D ,选择包含 N_s 个模式的数据子集 D_s 在一定准则下充分表示原始的数据集 D ,即期望使用子数据集 D_s 设计的分类器保持使用原始数据集 D 所设计的分类器的泛化性能,而数据量却远远小于原始数据集 $N_s \ll N$ ^[13].在数据集过大或数据量不断增大时,其优点是显而易见的.据所知,模式选择所选择的数据通常位于密集分布数据的边缘或者偏离已有的观测数据^[13-14],前者通常称边界点,而后者通常称离群点.边界点与离群点的检测有很多重要的应用,如网络入侵的识别^[15]、森林火灾的风险预测^[16]、机器故障的检测与诊断^[17]等.Li et al^[14]认为边界点和离群点代表了数据中最有效、有用、富含价值的模式.对边界点与离群点的检测方法,Chandola et al^[18]和Agrawal and Agrawal^[19]作了全面的概述.Li et al^[14]基于数据表示的相关性质提出了一种有效的边界点和离群点的检测方法BOD(Boundary and Outlier Detection algorithm).

数据流的概念漂移的过程中,数据的特征空间或决策边界发生了变化^[20].由于边界点位于密集区域的边缘,所以边界点所组成的区域在很大程度上可以反映特征空间.因此,在概念漂移的适应过程中,数据的边界点可以起很大的作用.

本文针对数据流少量标签问题,在考虑主动学习可能出现采样偏差的情况下,结合不确定性主动学习策略及边界点和离群点检测方法BOD,提出一种新的主动学习方法ALBOD(Active Learning based on Boundary and Outlier Detection).实验表明,在概念漂移的情况下,分类模型即使只使用ALBOD主动学习的少量标签进行更新,也可以表现良好的自适应能力.

本文的主要贡献如下:

(1)尝试将不确定性策略与边界点和离群点检测方法相结合,提出了一种新的主动学习方法ALBOD.

(2)实验表明,ALBOD学习的少量标签可以使数据流分类模型在概念漂移的情况下表现出良好的自适应能力.

1 研究方法

本节首先描述数据流自适应分类模型的框架,然后详细介绍分类模型置信度的计算方法以

及基于动态窗口的一致性子集的计算方法,最后给出ALBOD算法的主动学习策略.

1.1 自适应分类模型框架 建立基于主动学习的数据流自适应分类模型的主要目的是在分类模型的整个自适应过程中只使用少量的主动学习标签.整个模型的框架如图1所示:自适应分类模型 M 由多个不同 k 值的kNN分类模型组成,自适应模块分为概念漂移检测和主动学习两部分.概念漂移检测使用Haque et al^[7]的方法,检测所使用的分类器置信度计算方法见1.2节.

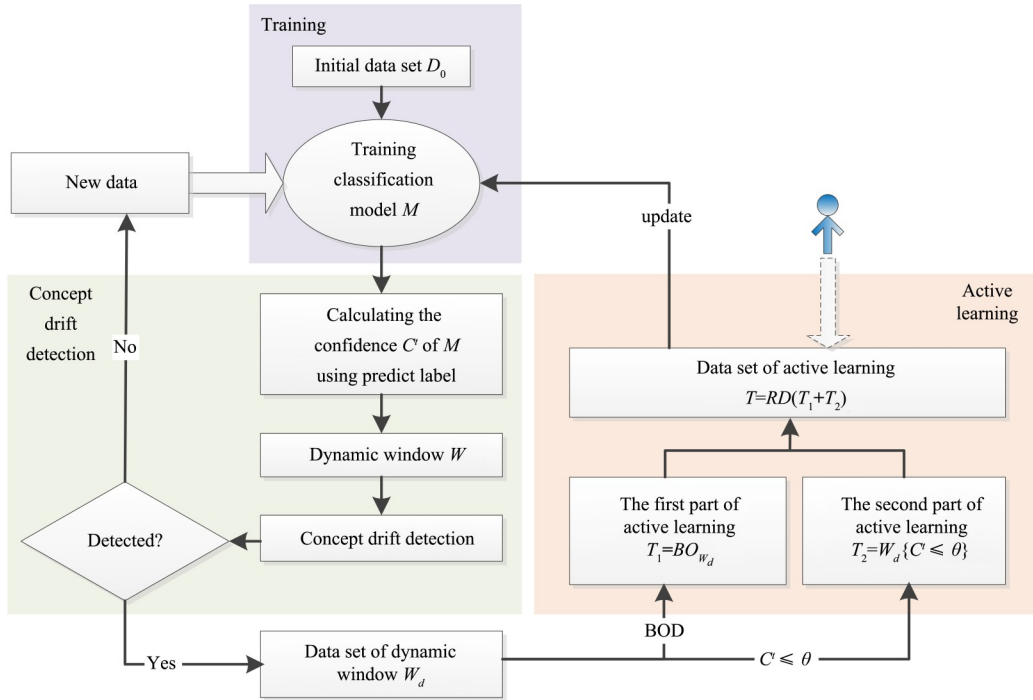


图1 基于主动学习的数据流自适应分类模型

Fig. 1 The adaptive classification model for data stream based on active learning

1.2 置信度的计算 信息熵通常用于衡量随机变量的不确定度.使用信息熵衡量分类器对样本的分类情况,信息熵越高则分类器分类错误的可能性越大,反之,分类器分类错误的可能性越小.本文的集成分类器 M 中每个分类器的信息熵的计算如式(1)和式(2)所示:

$$H(x_i)_m = - \sum_{j=1}^{class(k)} p(x_i \in class_j) \lg p(x_i \in class_j) \quad (1)$$

$$p(x_i \in class_j) = \frac{Num(class_j)}{k} \quad (2)$$

其中, $class(k)$ 表示离样本 x_i 最近的 k 个样本所包含的类别数, $Num(class_j)$ 为 k 个样本中属于类别 $class_j$ 的样本数, $p(x_i \in class_j)$ 表示 x_i 属于类别 $class_j$ 的可能性.集成分类器的信息熵为信息熵最大的单个分类器的信息熵:

$$H(x_i)_M = \max(H(x_i)_m) \quad (3)$$

得到分类器的信息熵之后,使用式(4)计算分类器的预测置信度,即分类器在 t 时刻的置信度:

$$C' = 1 - H(x_i)_M \quad (4)$$

1.3 窗口的一致性子集 给定数据集 $X = \{x_1, x_2, \dots, x_n\} \in R^{d \times n}$, d 表示特征的数量, n 表示

样本的数量. 对于样本 x , 集合 $U = \{w \in R^n | Xw = x\}$ 是一个仿射空间, 则 w 称为 x 的数据表示^[14,21]. 计算数据表示有多种方法, Roweis and Saul^[21]提出的数据表示计算方法如式(5)所示:

$$\min \sum_{i=1}^n \left| x_i - \sum_j w_{ij} x_j \right|^2, \text{ s.t. } \sum_j w_{ij} = 1 \quad (5)$$

每个样本点的数据表示 $w_i = (w_{i1}, \dots, w_{ij}, \dots, w_{im})$ 所对应的近邻点的表示分量 w_{ij} 有如下性质^[14]:

- (1) 如果 x 位于近邻点所组成的凸多边形的内部, 则对于任意的表示分量 w_{ij} 有 $0 < w_{ij} < 1$;
- (2) 如果 x 位于近邻点所组成的凸多边形的边界, 则存在任意的表示分量 w_{ij} 有 $w_{ij} = 0$;
- (3) 如果 x 位于近邻点所组成的凸多边形的外部, 则存在任意的表示分量 w_{ij} 有 $w_{ij} < 0$ 或者 $w_{ij} > 1$.

Li et al^[14]基于数据表示的相关性质提出了模式选择方法 BOD. 通过 BOD 算法选择出的数据称为数据集的一致性子集. 求动态窗口 W 所对应的样本集 W_d 的一致性子集 BO_{W_d} 的步骤如下:

- (1) 首先对样本集 W_d 中对应的每个样本 c_i 的特征向量 x_i 根据式(5)计算数据表示 w_i , 得到 W_d 中所有样本的数据表示矩阵 $\{w_1, \dots, w_n\}$, 其中 $w_i = (w_{i1}, \dots, w_{ij}, \dots, w_{im})$, w_{ij} 为 w_i 的表示分量.

- (2) 然后, 计算每个样本 x_i 的反向不可达量 RUR (Reverse Unreachability):

$$RUR_i = \chi(w_{ij}), j \in \{1, 2, \dots, n\} \quad (6)$$

如果 $w_{ij} \leq 0$ 则 $\chi(w_{ij}) = 1$, 否则 $\chi(w_{ij}) = 0$.

- (3) 接下来, 将每个样本的反向不可达量 RUR_i 排序. Li et al^[14]认为样本的 RUR 值越大则样本越偏离密集区域. 选择 RUR 值最大的 m 个样本作为离群点, 除去这 m 个样本的剩余样本, 则设置阈值 τ 选择其中的边界点. 由于窗口是动态变化的, 每个窗口的样本数量存在差异, 导致每个窗口的 RUR 值分布存在差异, 所以如果固定选择边界点的参数 τ 则可能导致某些窗口不能很好选择出边界点, 因此使用除去 m 个样本后剩余样本的反向不达标 RUR_{-m} 的最大值乘以百分比 p 来确定 τ 的值. 百分比 p 根据不同的数据集以及

选择样本的多少人为设定. 阈值 τ 随窗口中样本数动态的调整如式(7)所示:

$$\tau = p \times \max(RUR_{-m}) \quad (7)$$

最终, 一致性子集 BO_{W_d} 由 m 个离群点和阈值 τ 选择出的边界点组成.

1.4 主动学习 模型中的概念漂移检测使用 Haque et al^[7]的方法, 其中所使用的分类器置信度在 1.2 节已经介绍. 在检测到概念漂移之后, 记录样本置信度的窗口 W 停止增长. 接下来, 对 W 对应的样本 W_d 采用 1.3 节所介绍的方法求一致性子集 BO_{W_d} , 作为需要请求标签的第一个样本集 T_1 , 如图 2 中灰色的点所示.

根据窗口 W 存储的预测标签的置信度, 选择置信度低于 θ 的样本作为需要请求标签的第二个样本集 T_2 , 如图 2 中的绿色点所示. θ 为置信度阈值. T_1 与 T_2 中的样本去除重复组成主动学习的样本集 T . 在请求 T 中样本的标签之后, 样本集 T 加入初始训练集 D_0 , 更新分类模型 M , 得到新的分类模型 M_{new} .

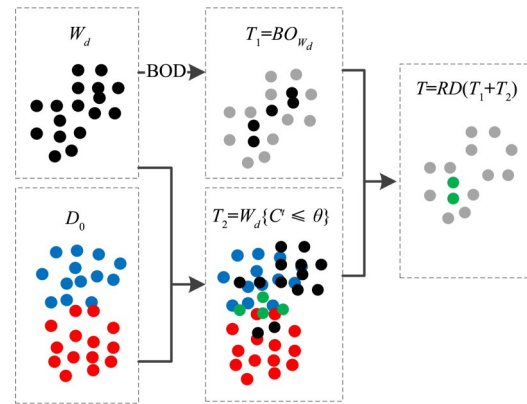


图 2 结合不确定性和 BOD 的主动学习方法

Fig. 2 Active learning combining uncertainty with BOD

在没有检测到概念漂移的情况下, W 会动态增长, 当 W 达到所设定的最大值时, 会重新初始化为空, 但分类器保持不变.

结合 BOD 方法与不确定性方法作为主动学习策略的好处是: BOD 方法得到了数据集密集区域的边界点与离群点, 这些点反映了数据集特征空间的全貌; 而不确定性方法弥补了决策边界样本点密集而 BOD 方法可能检测不到决策边界的

不足. 结合这两种主动学习方法从理论上可以缓解采样偏差的问题.

2 实验结果

2.1 数据集 表1描述了实验所使用的数据集. 合成数据集均由在线分析开源平台(Massive Online Analysis, MOA)^[22]生成. 数据集 SEAx 由 SEAGenerator 生成. SEAx 数据集是研究具有概念漂移的数据流最常用的人工数据集, 由两个类和三个 0 到 10 之间的属性组成. 数据集 SEAA 与 SEAg 的区别是: SEAA 包含突变漂移, 而 SEAg 则包含渐变漂移. 数据集 RBF0.003 由 Random-RBFGeneratorDrift 生成, 其中 0.003 表示数据生成过程中质心变化的速度. 真实数据集 Weather^[23]是由美国国家海洋和大气管理局从全世界 9000 多个气象站收集的气象测量数据, 包含以年为周期的概念漂移. 实验中, 所有数据集均作归一化处理.

表1 实验所用的数据集

Table 1 Datasets used in experiments

Datasets	Objects	Attributes	Classes
SEAA	10000	3	2
SEAg	10000	3	2
RBF0.003	100000	10	2
Weather	18159	8	2

2.2 实验分析 实验使用的对比算法 ACM (Adaptive Classification Model) 为不使用主动学习模块的自适应分类算法(见图1), 即适应概念漂移时使用动态窗口中数据的全部真实标签更新模型, Active Learning Uncertainty (ALU), Oza-BagAdwin (OBA) 和 Hoeffding Adaptive Tree (HAT) 均为 MOA^[22]中的算法. ALU 算法为只使用不确定性策略的主动学习算法, OBA 以及 HAT 算法为使用 100% 标签更新模型的算法. 选择 OBA 与 HAT 算法作为对比算法的原因是这两个算法均基于动态窗口实现, 这样可以减少由于窗口的固定设置所带来的分类精度的差别. 下文给出的实验结果均为多次运行的最好结果.

实验均使用数据集的前 100 个实例作为初始

训练集 D_0 , 剩余的数据作为测试集. 窗口的最大长度设置为 2000. 采用精确度 (Accuracy, ACC) 指标来评价算法的有效性.

2.2.1 实验 1: 与主动学习算法 ALU 的对比 在 MOA 平台上设置 ALU 算法的主动学习预算为 0.1, 记录下各数据集的分类精度, 标记百分比记录为 10%. 本文设计的主动学习算法 ALBOD 没有设置标记预算, 实验中调节参数 p 和 θ 控制标记样本在 10% 左右, 记录分类精度以及实际标记百分比. 实验结果如表 2 所示, 表中黑体字表示相对较优的分类结果.

从表 2 可以看出, 同样只学习了 10% 左右的标签, ALBOD 算法的分类效果比 ALU 算法的分类效果更好, 在真实数据集 Weather 上尤甚.

表2 与主动学习算法 ALU 对比的分类效果

Table 2 Classification performance of ALBOD and active learning algorithm ALU

Datasets	ALU ^[22]		ALBOD	
	ACC (%)	Labeled (%)	ACC (%)	Labeled (%)
SEAA	81.2	10	83.16	9.92
SEAg	81.5	10	82.95	9.79
RBF0.003	48.8	10	50.94	4.27
Weather	42.1	10	73.30	1.56

2.2.2 实验 2: 与全标记算法 OBA, HAT 和 ACM 的对比 在 MOA 平台上, OBA 和 HAT 算法使用默认参数设置, 对各数据集进行分类并记录分类精度; 运行 ACM 记录分类精度. 将 ACM 加入主动学习算法 ALBOD, 调节参数 p 和 θ 控制标记百分比, 记录对应的分类精度. 实验结果如表 3 所示. 为了保证对比的公平, 表 3 中 ALBOD 算法的分类结果为尽可能少标记而与全标记算法相当的结果, 表中黑体字是在得到相近的分类精度时 ALBOD 使用的标签数.

由表 3 可知, 算法 ACM 的分类精度与 OBA 和 HAT 算法相当时, 加入主动学习算法 ALBOD 可以让分类模型只使用少量的标记就能达到与全标记相当的结果. 例如, 在数据集 SEAA 与 SEAg 上, 所提算法只需学习 10% 左右的标签就可以让

表 3 与全标记算法 OBA, HAT 和 ACM 对比的分类效果

Table 3 Classification performance of comparing to 100% labelled algorithm OBA, HAT and ACM

Datasets	OBA ^[22]	HAT ^[22]	ACM	ALBOD
	ACC (%)	ACC (%)	ACC (%) (Labeled (%))	ACC (%) (Labeled (%))
SEAA	84.21	84.98	83.31 (100)	83.16 (9.92)
SEAg	83.51	83.44	83.29 (100)	83.36 (10.75)
RBF0.003	62.61	59.14	57.37 (100)	55.53 (22.28)
Weather	71.51	71.33	77.98 (100)	77.14 (20.82)

分类模型达到 100% 标记的同等分类精度;在数据集 Weather 上,所提算法在学习 20.82% 的标签时,其分类精度 77.14% 已经接近于 100% 标记的分类精度 77.98%,高出全标记算法 OBA 和 HAT 的分类精度;在数据集 RBF0.003 上的表现虽然稍差,但标记 22.28% 的数据时,其分类精度与 100% 标记算法 ACM 的相差也不大,这可能是由于分类模型本身设计的问题。

结合表 2 中 ALBOD 的实验数据与表 3 中 ACM 和 ALBOD 的实验数据分析可知,在数据集 SEAA 与 SEAg 上,标记样本量从 10% 左右增加到 100% 时,分类精度提高不明显;在数据集 RBF0.003 与 Weather 上,标记样本量从小于 5% 增加到 20% 左右时,分类精度有较大的提升,但此时,即使标记样本再增加到 100%,在数据集 Weather 上分类精度的提升也不明显,在数据集 RBF0.003 上的提升也不大。由此可以说明,ALBOD 学习的少量样本代表了数据集中关键的样本,可以使分类模型在只使用少量标签更新的情况下即可达到使用动态窗口中全部数据标签更新的分类精度,即让数据流分类模型在概念漂移的情况下表现出良好的自适应能力。

3 结 论

数据流挖掘因为其概念漂移的特点变得复杂,而在实际应用中,获得适应概念漂移时用于更新模型的数据标记不仅代价昂贵也更加困难,因此,本文采用主动学习的方法标记少量数据。考虑到主动学习过程中可能出现采样偏差的问题,将不确定性学习策略与边界点和离群点检测方法相结合,提出一种新的主动学习方法 ALBOD。该

方法结合了不确定性样本能够反映决策边界、边界点和离群点能够反映特征空间的优点,主动学习的少量标签样本尽可能反映数据的真实分布,并在合成数据集与真实数据集上得到了验证。实验结果表明,通过 ALBOD 学习的少量标签更新的分类模型可以有与全标记更新的分类模型相当的概念漂移适应能力。但 ALBOD 算法毕竟只学习了少量的标签,如果是密集区域且分类情况复杂,则 ALBOD 算法可能会使分类模型损失很大的精度。这将是下一步研究的问题。

参考文献

- [1] Schlimmer J C, Granger R H Jr. Incremental learning from noisy data. *Machine Learning*, 1986, 1(3): 317—354.
- [2] 郑灿彬, 闻立杰, 王建民. 基于可扩展活动关系的过程概念漂移检测. *计算机集成制造系统*, 2018, 24(7): 1589—1597. (Zheng C B, Wen L J, Wang J M. Process concept drift detection based on extensible activity relationship. *Computer Integrated Manufacturing Systems*, 2018, 24(7): 1589—1597.)
- [3] Ditzler G, Roveri M, Alippi C, et al. Learning in nonstationary environments: a survey. *IEEE Computational Intelligence Magazine*, 2015, 10(4): 12—25.
- [4] ZareMoodi P, Beigy H, Siahroudi S K. Novel class detection in data streams using local patterns and neighborhood graph. *Neurocomputing*, 2015, 158: 234—245.
- [5] 孙艳歌, 王志海, 原继东等. 基于信息熵的数据流自适应集成分类算法. *中国科学技术大学学报*, 2017, 47(7): 575—582. (Sun Y G, Wang Z H, Yuan J D, et al. Adaptive ensemble classification algorithm for data streams based on information entropy. *Journal of*

- University of Science and Technology of China, 2017, 47(7):575—582.)
- [6] Ahmadi Z, Beigy H. Semi-supervised ensemble learning of data streams in the presence of concept drift//The 7th International Conference on Hybrid Artificial Intelligence Systems. Springer Berlin Heidelberg, 2012:526—537.
- [7] Haque A, Khan L, Baron M. Sand: semi-supervised adaptive novel class detection and classification over data stream//The 30th AAAI Conference on Artificial Intelligence. Phoenix, AZ, USA: AAAI, 2016:1652—1658.
- [8] Settles B. Active learning literature survey. computer Sciences. Technical Report 1648. Madison: University of Wisconsin-Madison, 2009:3—4.
- [9] Fu Y F, Zhu X Q, Li B. A survey on instance selection for active learning. Knowledge and Information Systems, 2013, 35(2):249—283.
- [10] Mohamad S, Bouchachia A, Sayed-Mouchaweh M. A bi-criteria active learning algorithm for dynamic data streams. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(1):74—86.
- [11] Zliobaite I, Bifet A, Pfahringer B, et al. Active learning with drifting streaming data. IEEE Transactions on Neural Networks and Learning Systems, 2013, 25(1):27—39.
- [12] Mohamad S, Sayed-Mouchaweh M, Bouchachia A. Active learning for classifying data streams with unknown number of classes. Neural Networks, 2018, 98:1—15.
- [13] Li Y, Maguire L. Selecting critical patterns based on local geometrical and statistical information. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(6):1189—1201.
- [14] Li X J, Lv J C, Yi Z. An efficient representation-based method for boundary point and outlier detection. IEEE Transactions on Neural Networks and Learning Systems, 2016, 29(1):51—62.
- [15] Ahmed M, Mahmood A N, Hu J K. A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 2016, 60:19—31.
- [16] Salehi M, Rashidi L. A survey on anomaly detection in evolving data: with application to forest fire risk prediction. ACM SIGKDD Explorations Newsletter, 2018, 20(1):13—23.
- [17] Gao Z W, Cecati C, Ding S X. A survey of fault diagnosis and fault-tolerant techniques - Part I: fault diagnosis with model-based and signal-based approaches. IEEE Transactions on Industrial Electronics, 2015, 62(6):3757—3767.
- [18] Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. ACM Computing Surveys, 2009, 41(3):15.
- [19] Agrawal S, Agrawal J. Survey on anomaly detection using data mining techniques. Procedia Computer Science, 2015, 60:708—713.
- [20] Lu J, Liu A J, Dong F, et al. Learning under concept drift: a review. IEEE Transactions on Knowledge and Data Engineering, 2018, doi: 10.1109/TKDE.2018.2876857.
- [21] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000, 290(5500):2323—2326.
- [22] Bifet A, Holmes G, Kirkby R, et al. MOA: massive online analysis. Journal of Machine Learning Research, 2010, 11:1601—1604.
- [23] Elwell R, Polikar R. Incremental learning of concept drift in nonstationary environments. IEEE Transactions on Neural Networks, 2011, 22(10):1517—1531.

(责任编辑 杨可盛)