基于概率相关性的多标签数据流变化检测

石中伟[1] 文益民[1,2]

(桂林电子科技大学计算机科学与工程学院 桂林 541004)[1] (广西可信软件重点实验室 桂林 541004)[2]

摘 要 由于传统的概念漂移检测研究主要针对单标签数据流,对现实中常见的多标签数据流却缺乏足够的关注,多标签数据流概念漂移检测问题有待进一步的研究。因此,通过分析多标签数据流中存在的特殊依赖关系,提出了一种基于概率相关性的多标签数据流概念漂移检测算法。其基本思想是从概念漂移的产生原因出发,利用概率相关性近似描述数据分布来监测新旧数据分布变化,判断概念漂移是否发生。实验结果表明,提出的算法能够比较快速、准确地检测到概念漂移,并在多标签概念漂移数据流分类问题上取得了预期的学习效果。

关键词 概念漂移,多标签,数据流,概率相关性,分类

中图法分类号 TP391.4 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.8.013

Detection of Multi-label Data Streams Change Based on Probability of Relevance

SHI Zhong-wei[1] WEN Yi-min[1,2]

(School of Computer Science and Engineering, Guilin University of Electronic Technology, Guilin 541004, China)[1]

(Guangxi Key Laboratory of Trusted Software, Guilin 541004, China)[2]

Abstract Traditional detection approaches of concept drift mainly focus on single-label scenarios, however, not enough attention has been paid to the problem of mining from multi-label data streams. But applications of such data streams are common in the real world. These make it necessary to design efficient algorithms to detect concept drift for multi-label data streams. So after particularly analyzing the unique property label dependence of multi-label data streams, the paper proposed an algorithm of detecting concept drift based on the probability of relevance for multi-label data streams. The basic idea originates from the reason of concept drift and it describes the distribution of data streams by using the probability of relevance. Then, it estimates whether the concept drift occurs or not through monitoring the change of distribution between the old data and new data. The final experimental results show that the proposed algorithm can rapidly and accurately detect the concept drift and achieve prospective predictive performance for multi-label evolving stream classification.

Keywords Concept drift, Multi-label, Data streams, Probability of relevance, Classification

1 引言

随着信息技术的快速发展,数据流应用越来越广泛,如电子邮件、在线视频、股票与基金、传感器网络、社交网络等。由于数据流具有快速性、连续性和无限性的特点,这就要求学习算法能够在有限的时间和内存资源条件下完成。同时,数据流中蕴含的概念常常随着时间而变化,即出现了概念漂移。这使得建立在原有数据上的学习模型不再适应新的数据。这些都给传统的分类问题提出了巨大挑战。

学术界在数据流分类领域已经取得了一系列研究成果,包括概念漂移的定义、产生原因、检测算法和数据流的学习等。快速、准确的概念漂移检测是实现数据流学习的关键。若能有效、快速、准确地检测到概念漂移,那么就可以及时调整学习模型,以便能对不断到来的数据做出尽可能正确的判断。当前,概念漂移检测研究主要集中在单标签数据流分类领域。

然而,在生产生活中还广泛存在多标签数据流分类的例子。例如,一个用户可能对多个特定交叉领域的文章感兴趣,并且用户的兴趣会随着时间而变化;一位同时患有多种疾病的患者,在长期的治疗过程中,一些疾病被治愈,而一些疾病仍然缠身[1]。在学习的过程中,训练的模型要能够预测一个用户对同时属于多个类别的文章是否感兴趣,判断出患者当前所患的疾病;同时,要能够捕捉到用户兴趣、患者疾病的变化,以对自身进行调整,做出尽可能正确的决策。

不同于单标签数据,多标签数据中存在特殊的依赖关系,包括条件依赖(conditional dependency)和非条件依赖(uncon-

到稿日期:2014-06-10 返修日期:2014-07-22 本文受国家自然科学基金项目:基于多任务学习的复杂概念漂移数据流分类研究(61363029),广西可信软件重点实验室项目:基于多信息的旅游线路智能推荐系统(KX201311)资助。

石中伟(1990—),男,硕士生,主要研究方向为机器学习与数据挖掘;文益民(1969—),男,博士,教授,CCF高级会员,主要研究方向为机器学习与数据挖掘、推荐系统、智慧旅游,E-mail:ymwen2004@aliyun.com。

ditional dependency)[2]，且这种依赖关系会随着时间而变化。这样多标签概念漂移数据流不仅包含传统的样本与标签之间依赖关系的变化，还包括标签之间依赖关系的变化。然而，一些从数据角度出发处理概念漂移的方法并没有利用多标签数据的特性，而是通过不同的策略来削弱历史数据对当前学习模型的影响，以适应不断到来的新数据。因此，本文通过对多标签数据流进行分析，借助概率相关性来描述多标签数据分布，度量数据分布中样本与标签、标签与标签之间的关联度差异，评测数据流分布差异变化，以检测概念漂移。

2 相关工作

近年来，一些学者从不同的角度对多标签数据流概念漂移检测问题进行了研究，并提出了相应的算法。Qu 等[1]假设数据流以块的形式到达，基于若干连续的多标签数据块，针对每个数据块，训练一个多标签分类器，形成一个集成分类器；当新数据块到达时，训练新的多标签分类器并替换掉当前分类性能较差的多标签分类器，以适应可能出现的新概念。Xioufis 等[3]为了处理混合概念漂移，针对每个标签，设置两个固定比率大小的滑动窗口，分别保存正、负样本，通过不断更新窗口内的样本，削弱旧样本对学习模型影响的方式，适应概念漂移。Kong 等[4]提出多颗随机树集成的分类方法，并且在随机树的每个节点增加一个衰减函数(fading function)，削减历史样本对学习模型的影响，适应不断到来的新样本。Read 等[5]提出基于多颗决策树的 Bagging 集成算法，通过监测每颗决策树分类性能变化的方法检测概念漂移，并以多标签准确率作为性能参考指标；当检测到概念漂移时，重建当前分类性能最差的决策树。

对于多标签数据流概念漂移检测问题，文献[1,3,4]并没有给出明确的检测方法，而是分别利用分类器集成、样本更新与衰减函数来逐渐削弱历史数据产生的影响，这并不能有效检测到概念漂移发生的位置以及及时对学习模型进行更新，具有一定的滞后性；文献[5]提出的方法能够主动检测概念漂移，对学习模型做出调整，并最终取得了较好的学习效果。不同于以上提及的算法，本文利用概率相关性描述多标签数据总体分布特征，监测分布特征变化，判断是否发生概念漂移。

3 相关概念和原理

3.1 概率相关性

概念漂移发生时意味着数据中蕴含的概念发生变化，由数据总体分布变化引起，具体而言是样本与标签之间的关系，为描述这种关系，Cheng 等[2]提出概率相关性：

$$f_i(x)=p(y_i=1|x),i=1,2,\cdots,m \tag{1}$$

式(1)表示对于给定的标签集合，样本 x 与某个标签的概率相关性反映出它们之间的关联程度，$f_i(x)$ 值越大关联程度越高，m 为标签的数目。

给定一组多标签数据集，评估样本与标签概率相关性的具体方法参考文献[4]：假定样本特征空间 $X\subseteq R^d$，标签集合 $L=\{l_1,\cdots,l_m\}$，多标签数据集 $DS=\{(x_1,Y_1),\cdots,(x_t,Y_t)\}$，其中，样本 $x_i=(x_i^1,\cdots,x_i^d)\in X$ 被赋予一个标签集合 $Y_i\subseteq L$，并且 $Y_i=(y_i^{(1)},\cdots,y_i^{(|L|)})\in\{0,1\}^{|L|}$，$y_i^j=1$，当且仅当 $l_j\in Y_i$；

a)对于每个类别标签 l_j，它们在所有样本的标签组合中出现的频数被保存在一个 m 维向量 $\boldsymbol{C}$ 中，即 $\boldsymbol{C}=(c_1,\cdots,c_m)$，且 $c_j=\sum_{i=1}^{t}y_i^j$；

b)用 θ 表示所有样本所属标签的数目的和，且 $\theta=\sum_{i=1}^{t}|Y_i|$，其中，$|Y_i|$ 表示样本 x_i 所属标签的数目。

概率相关性函数 $\boldsymbol{f}(x)=(f_1(x),\cdots,f_m(x))$ 的每一个分量的值：$f_i(x)=p(y_i=1|x)=\frac{c_i}{t}$（$t$ 为数据集中样本总数）；另外，lc 表示标签组合的整体稀疏程度，即所有样本所属标签的数目的平均值，其计算方式：$lc=\frac{\theta}{t}$。

3.2 差异度量

假定：有两组序列 $\boldsymbol{\tau}=\{\tau_1,\cdots,\tau_m\}$ 和 $\boldsymbol{\tau}'=\{\tau_1',\cdots,\tau_m'\}$，其中，$\tau_i\in[0,1]$，$\tau_i'\in[0,1]$，Cheng 等[6]、Hullermeier 等[7]为度量序列 $\boldsymbol{\tau},\boldsymbol{\tau}'$ 之间差异采用统计学中基于距离的方法：

$$D(\boldsymbol{\tau}',\boldsymbol{\tau})=\#\{(i,j)|i<j,\tau_i>\tau_j\wedge\tau_i'<\tau_j'\} \tag{2}$$

式(2)表示两个序列的所有二元约束集中分量之间关系不一致的个数，体现两序列总体上的差异性。

3.3 评测差异

在数据流环境下，度量数据的分布差异会产生连续的差异值。如何评测差异值的变化波动，可借助其统计特征均值、方差等进行衡量，参考 Albert 等[8]提出的设计思路：首先用一定数量的差异值 dis 初始化窗口 W，其大小依据检测结果而不断变化；当增添新的差异值到 W 时，将其化分为任意的两个子窗口 W_1 和 W_2(见图 1)，并计算两子窗口均值的差异。若该差异超过某个阈值，则认为两窗口内差异值 dis 分布不相同，进而可认为此时数据整体分布不一致，数据流发生变化，然后时间靠前的子窗口 W_1 就会被舍弃，而靠后的子窗口 W_2 则得以保留；若不超过该阈值，则继续增添新的差异值 dis 到窗口 W。

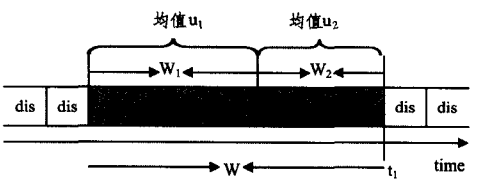

图 1 描述窗口 W 的分割过程

4 算法设计与实现

4.1 算法设计

对于概念漂移产生的原因，Kelly 等[9]认为样本与其类别的联合概率分布随时间而变化，产生概念漂移。当样本后验概率 $p(y|x)$($y=0,1$(二分类问题))发生变化时，分类规则变化，产生概念漂移；对于多标签数据，样本后验概率 $p(Y|x)$，$Y=\{y_1,\cdots,y_m\}$ 变化，产生概念漂移。因此，可利用概率相关性近似估量样本后验概率以描述数据分布，度量新旧数据分

布差异，评测差异变化，判断是否出现概念漂移。本文提出的概念漂移检测算法包含以下步骤：

a)借助概率相关性描述多标签数据的分布；

多标签数据集的分布特征可由一个二元组描述：概率相关性函数 $f(x)$ 和标签组合的稀疏度 lc，分别反映了样本与标签及标签与标签之间的关联度，即：

$$\boldsymbol{M}=(f(x),lc) \tag{3}$$

b)设定差异度量的粒度；

c)度量新旧数据分布之间的差异；

给定两组多标签数据集，利用式(3)建立对应的分布特征模型 $\boldsymbol{M}_o=(f_o(x),lc_o)$ 与 $\boldsymbol{M}_n=(f_n(x),lc_n)$，概率相关性函数 $f_o(x)$ 与 $f_n(x)$ 之间差异的度量采用对式(2)进行改进之后的方法：

$$D(f_o(x),f_n(x))=\frac{1}{lc_n(m-lc_n)}\sum_{f_n^i>f_n^j} I(f_o^i\leqslant f_o^j) \tag{4}$$

式(4)表示 $f_o(x)$ 与 $f_n(x)$ 的所有二元约束集中分量之间关系差异的程度，度量的策略是比较分量之间关系的一致性情况，体现了样本与标签的关联度之间关系的差异。I 是指示函数，如果条件满足，则返回值为 1；否则，返回值为 0。

另外，标签组合的稀疏度 lc_o 与 lc_n 之间差异的度量采用类似于几何学中两点距离的计算方法：

$$S(lc_o,lc_n)=(lc_n-lc_o)^2 \tag{5}$$

最后，这两组数据集的差异量化方式如下：

$$dis=D(f_o(x),f_n(x))+S(lc_o,lc_n) \tag{6}$$

d)借助差异值的均值变化判断其是否出现概念漂移。

在算法中，步骤 a)、b)与 c)的总体流程如图 2 所示，缓存 N 个样本，统计数据信息，利用概率相关性对缓存的数据总体分布的特征进行建模；对于后续不断到来的单个样本(度量粒度)，度量其与当前分布模型之间的差异，具体的量化方式使用式(6)，设定基于单个样本的差异度量有助于及时检测到概念漂移，若没有检测到概念漂移，则更新当前分布模型。步骤 d)依据 3.3 节中的方法对使用式(6)量化之后的差异值 dis 进行监测，判断是否发生概念漂移。

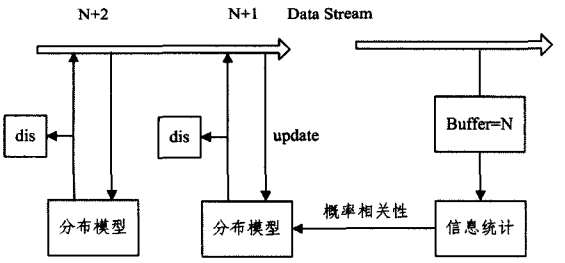

图 2 算法部分流程

4.2 算法实现

算法的伪代码描述如下。

算法 1 The outline of the proposed method

输入：数据流样本(x_i，Y_i)，窗口 W

输出：true or false(是或否发生概念漂移)

i=0；

从多标签数据流中顺序读取 N 个样本形成初始样本集合 S_0

使用 S_0 初始化当前数据分布的模型 M=(f(x)，lc)

i=N；

WHILE(数据流没有结束)

后续样本(x_i，Y_i)到达

估量(x_i，Y_i)与 M 的差异

计算：dis=D(f(x)，Y_i)+S(lc，|Y_i|)

将 dis 插入到窗口 W

监测窗口 W 内差异值 dis 的变化

函数：Monitor(W)

IF(Monitor(W)=true)

读取 N 个样本重新初始化数据分布的模型 M

RETURN true

ELSE

用(x_i，Y_i)更新当前数据分布模型 M，包括 f(x)和 lc 的更新

RETURN false

END

END

函数 Monitor(W)用来监测窗口 W 内差异值 dis 的变化，其具体过程如 3.3 节所述。返回 false，表明无新概念出现；返回 true，表明有概念漂移产生。

5 实验方案设计

5.1 实验数据

实验数据选择仿真多标签数据集，由多标签数据流生成框架[10]产生，选择 RTG(Random Tree Generation)[11]作为样本生成器，且每个属性有 5 个不同的值，每个标签有 2 个不同的类值。另外，模拟概念漂移可以改变如下参数[5,12]：z (平均每个样本所属于标签的数目)和 ld (标签依赖)，其中，

RTG8-1 表示标签数目为 8，属性数目为 10，z 和 ld 的变化：(1.8，0%)→(1.8，10%)→(3.0，0%)→(3.0，20%)；

RTG15-1 表示标签数目为 15，属性数目为 20，z 和 ld 的变化：(1.8，0%)→(1.8，10%)→(3.0，0%)→(3.0，20%)；

RTG8-2 表示标签数目为 8，属性数目为 10，z 的变化：1.8→3.0→2.0→4.0；

RTG15-2 表示标签数目为 15，属性数目为 20，z 的变化：1.8→3.0→2.0→4.0。

样本数目 $T=1000000$，每个数据集中包含 4 个不同概念，即发生 3 次概念漂移，分别出现在 $T/4$，$2T/4$ 和 $3T/4$ 的位置。

5.2 评价指标

根据当前多标签数据流分类问题研究文献中采用的分类性能评价指标，文中选取如下指标来验证所提算法性能的实验：

1)Subset Accuracy[5,12]：基于样本的度量方法，计算多标签分类器对每个样本预测出的标签和实际标签之间的相似度([0，1])，其中，$Y_i=(y_1,\cdots,y_m)$ 表示测试样本 x_i 实际所属的标签集合，$(\hat{y}_1,\cdots,\hat{y}_m)$ 表示预测出的 x_i 所属于的标签集合，T 表示样本总数。其计算公式如下：

$$SubsetAccuracy=\frac{1}{T}\sum_{i=1}^{T}\frac{\sum_{j=1}^{m}y_j^{(i)}\wedge\hat{y}_j^{(i)}}{\sum_{j=1}^{m}y_j^{(i)}\vee\hat{y}_j^{(i)}} \quad (7)$$

2)F_1-macro[3-5,12]：基于标签的度量方法，计算每个标签的 F_1 度量值，并取所有度量值的平均值作为最终结果，其中，F_1 是准确率(Precision)和召回率(Recall)[12]的调和平均值，广泛应用于信息检索。

$$F_1\text{-}macro=\frac{1}{m}\sum_{j=1}^{m}F_1[(y_j^{(1)},\cdots,y_j^{(T)}),(\hat{y}_j^{(1)},\cdots,\hat{y}_j^{(T)})] \quad (8)$$

5.3 对比实验设计

本文采用数据流分类问题研究中通用的 prequential[13]实验方法，首先假定每个样本所属于的标签集合未知，多标签分类器对其做出预测，评价预测的结果；然后样本所属于的标签集合作为已知并用来对分类器进行更新。实现中选取 HT-ps[5,12](Multi-label Hoeffding Tree with PS classifier at the leaves)作为多标签分类器以及一个扩展的 Hoeffding 树[11]算法。

同时，为验证提出的算法的有效性，在处理数据流中的概念漂移问题上，设计如下对比实验。

(1) ADWIN-P：实验所对比的是 Read 等[5]利用 ADWIN[8]方法监测多标签分类器预测性能(Performance)的变化，以检测概念漂移。

(2)Relevance：为方便，将文中提出的多标签数据流概念漂移检测算法简称为 Relevance，因为它利用概率相关性描述多标签数据分布以检测概念漂移。

6 实验结果与分析

6.1 实验结果

为了保证实验结果的可靠性，每次实验重复执行 100 次，并取平均值作为最终的实验结果。表 1 与表 2 分别列出了在两种不同评价指标下的对比实验结果，表 3 列出了每次实验的平均执行时间。

表 1 评测指标 Subset Accuracy

	ADWIN-P	Relevance
RTG8-1	0.411±0.036	0.420±0.044
RTG15-1	0.159±0.043	0.194±0.041
RTG8-2	0.503±0.039	0.507±0.039
RTG15-2	0.211±0.059	0.234±0.047

表 2 评测指标 F1-macro

	ADWIN-P	Relevance
RTG8-1	0.352±0.038	0.365±0.047
RTG15-1	0.141±0.061	0.150±0.046
RTG8-2	0.426±0.039	0.434±0.040
RTG15-2	0.178±0.068	0.200±0.055

表 3 每次实验的平均执行时间(s)

	ADWIN-P	Relevance
RTG8-1	21±1	17±1
RTG15-1	41±2	31±1
RTG8-2	21±1	16±1
RTG15-2	37±3	28±2

根据表 1、表 2 中的实验结果可知，在不同的性能指标下，算法 Relevance 的最终结果比 ADWIN-P 都要高。可见，在处理多标签概念漂移数据流时，本文提出的基于概率相关性的检测算法对概念漂移具有很好的适应性，结合已有分类方法使用时，能够保持较好的预测性能。另外，从表 3 可知，算法 Relevance 平均每次实验的执行时间要低于 ADWIN-P，在效率上同样保持一定的优势。

6.2 实验分析

以数据集 RTG8 为例，从两个不同的角度对以上实验结论进行分析：1)在概念漂移发生时，算法检测到的位置；2)在不同的时间点，算法表现出的预测性能。

表 4 列出算法 Relevance 与 ADWIN-P 检测到的 3 个概念漂移发生时的具体位置；图 3、图 4 分别直观地刻画出在不同的评测指标下，算法 Relevance 与 ADWIN-P 在不同时间点的预测性能，其中，每个时间点包含 10000 个样本，也即将整个数据流划分 100 个连续的块，依次计算评测指标的值。

由表 4 的结果可知，对于前两次概念漂移，算法 Relevance 比 ADWIN-P 提前检测出漂移点的位置；对于第 3 次概念漂移，ADWIN-P 并没有检测到数据中蕴含的概念发生变化。另外，从图 3、图 4 可知，在时间点 25 和 50 的位置，算法 Relevance 与 ADWIN-P 的图像前后波动反映出概念变化；而在时间点 75 的位置，ADWIN-P 的图像并没反映出概念变化。同样，从图 3、图 4 可知，算法 Relevance 的整体性能要优于 ADWIN-P。

表 4 检测到的漂移点位置

Methods	Positions		
	First Drift	Second Drift	Third Drift
ADWIN-P	250,208	500,224	无
Relevance	250,152	500,015	750,774

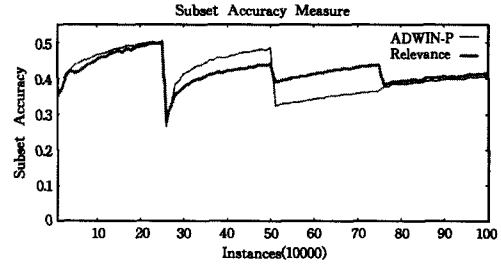

图 3 不同时间点下的 Subset Accuracy 指标值

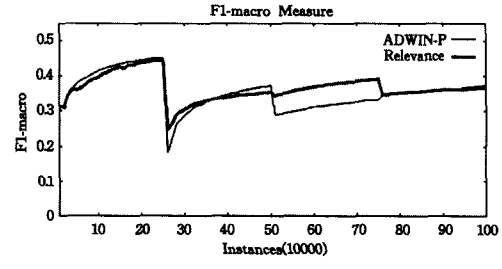

图 4 不同时间点下的 F_1-marco 指标值

因此，从以上两个角度的分析可知，算法 Relevance 在处理多标签概念漂移数据流时，能够比较快速、准确地检测到概念漂移，结合已有分类方法使用时，能够取得比较理想的学习效果。

结束语 随着多标签数据流应用在现实生活中不断涌

现，如何有效地对其进行挖掘却缺乏足够的关注。因此，本文结合多标签数据中存在的概念相关性，提出一种新的多标签数据流概念漂移检测算法。通过在多个数据集上使用不同的评测指标进行实验对比，并对实验结果从不同的角度进行分析，最终得出的结论是提出的算法能够有效处理多标签数据流中存在的概念漂移，并且应用于分类问题时能够取得较好的预测性能。

未来的工作包括：1)从信息熵的角度对多标签数据流进行分析，建立数据的特征模型，尝试设计有效的多标签数据流概念漂移检测算法；2)对于多标签数据流分类问题，分析标签数量过大可能带来的挑战，以及如何有效处理。

参 考 文 献

[1] Qu W, Zhang Y, Zhu J P, et al. Mining Multi-label Concept-Drifting Streams Using Ensemble Classifiers[C]// Proceeding of Fuzzy Systems and Knowledge Discovery. Tianjin, China, 2009, 5:275-279

[2] Cheng W, Hüllermeier E, Dembczynski K J. Bayes optimal multilabel classification via probabilistic classifier chains[C]// Proceedings of the 27th international conference on machine learning. Haifa, Israel, 2010:279-286

[3] Xioufis E S, Spiliopoulou M, Tsoumakas G, et al. Dealing with concept drift and class imbalance in multi-label stream classification[C]// Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence. Barcelona, Spain, 2011, 2: 1583-1588

[4] Kong X, Yu P S. An ensemble-based approach to fast classification of multi-label data streams[C]// Proceeding of 7th the International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom). Orlando, Florida, USA, 2011:95-104

[5] Read J, Bifet A, Holmes G, et al. Scalable and efficient multi-label classification for evolving data streams[J]. Machine Learning, 2012, 88(1/2):243-272

[6] Cheng W, Hüllermeier E. A Simple Instance-Based Approach to Multilabel Classification Using the Mallows Model[C]// ECML/PKKD Workshop on Learning from Multi-label Data. Bled, Slovenia, 2009:28-38

[7] Hüllermeier E, Fürnkranz J, Cheng W, et al. Label ranking by learning pairwise preferences[J]. Artificial Intelligence, 2008, 172(16):1897-1916

[8] Bifet A, Gavalda R. Learning from Time-Changing Data with Adaptive Windowing[C]// Proceeding of the SIAM International Conference on Data Mining. Minneapolis, Minnesota, USA, 2007, 7:443-448

[9] Kelly M, Hand D, Adams N. The impact of changing populations on classifier performance[C]// Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining. San Diego, CA, USA, 1999:367-371

[10] Read J, Pfahringer B, Holmes G. Generating synthetic multi-label data streams[C]// ECML/PKKD Workshop on Learning from Multi-label Data. Bled, Slovenia, 2009:69-84

[11] Domingos P, Hulten G. Mining high-speed data streams [C]// Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. New York, USA, 2000:71-80

[12] Read J, Bifet A, Holmes G, et al. Efficient multi-label classification for evolving data streams[R]. University of Waikato, 2010

[13] Gama J, Sebastião R, Rodrigues P P. Issues in evaluation of stream learning algorithms[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2009:329-338

(上接第 55 页)

[7] Wan T, Canagarajah N, Achim A. Statistical multiscale image segmentation via alpha-stable modeling[C]// Proceedings of IEEE International Conference in Image Processing. Texas, 2007:357-360

[8] Haris K, Estradiadis S N, Maglaveras N, et al. Hybrid image segmentation using watersheds and fast region merging[J]. IEEE Transactions on Image Processing, 1998, 7(12): 1684-1699

[9] Liu H, Guo Q, Xu M, et al. Fast image segmentation using region merging with a k-nearest neighbor graph[C]// Proceedings of IEEE International Conference Cybernetics and Intelligent Systems. Chengdu, 2008:179-184

[10] Shu Y, Bilodeau G A, Cheriet F. Segmentation of laparoscopic images: Integrating graph-based segmentation and multistage region merging[C]// Proceedings of the 2nd Canadian Conference on Computer and Robot Vision. Regina, 2005:429-436

[11] Moscheni F, Bhattacharjee S, Kunt M. Spatio-temporal segmentation based on region merging[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(9):897-915

[12] Ning J, Zhang L, Zhang D, et al. Interactive image segmentation by maximal similarity based region merging[J]. Pattern Recognition, 2010, 43(2):445-456

[13] Peng B, Zhang L, Zhang D. Automatic image segmentation by dynamic region merging[J]. IEEE Transactions on Image Processing, 2011, 20(12):3592-3605

[14] Cheng M, et al. Global contrast based salient region detection[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, 2011:409-416

[15] Felzenszwalb P F, Huttenlocher D P. Efficient graph based image segmentation[J]. International Journal of Computer Vision, 2004, 59(2):167-181

[16] Martin D, Fowlkes C, Tal D, et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Kauai, 2001:416-423