

基于多分类器的数据流中的 概念漂移挖掘

孙岳¹ 毛国君¹ 刘旭¹ 刘椿年¹

摘要 数据流中概念漂移的检测是当前数据挖掘领域的重要研究分支,近年来得到了广泛的关注.本文提出了一种称为 M.ID4 的数据流挖掘算法.它是在大容量数据流挖掘中,通过尽量少的训练样本来实现概念漂移检测的快速方法.利用多分类器综合技术, M.ID4 实现了数据流中概念漂移的增量式检测和挖掘.实验结果表明, M.ID4 算法在处理数据流的概念漂移上表现出比已有同类算法更高的精确度和适应性.

关键词 数据挖掘, 数据流, 概念漂移

中图分类号 TP311

Mining Concept Drifts from Data Streams Based on Multi-classifiers

SUN Yue¹ MAO Guo-Jun¹ LIU Xu¹

LIU Chun-Nian¹

Abstract Mining concept drifts from data streams is one of the most important fields in data mining. In this paper, a new mine algorithm called M.ID4 is proposed, which aims at quickly detecting drifted concepts from a large volume of data stream by using a small training data set. M.ID4 uses ensemble multi-classifiers to mine concept changes from the data streams, and its every classifier in the ensemble is an improved ID4 algorithm with an incremental way. The experimental results show that M.ID4 algorithm is of higher accuracy and better adaptability to quick drifted concepts than the popular algorithms.

Key words Data mining, data stream, concept drift

在现实生活中,有一类问题是数据所蕴含的概念随着时间而变化的.由于数据随时间的持续变化,因此必然会导致概念模型的更新,进而引起概念漂移(Concept drift)问题.1996年, Widmer 等提出了概念漂移的问题^[1],之后许多学者对此问题进行了研究^[2-5].近年来,随着数据流(Data stream)挖掘研究的深入,数据流中的概念漂移问题备受瞩目.

在该领域中所开展的工作主要有:1997年, Salganicoff 等提出了基于懒惰学习技术的概念漂移检测算法 PECS^[6];1998年, Harries 等给出了利用上下文聚类技术来实现概念漂移对应的隐藏信息的识别算法 SPLICE^[4];2000年, Domingos 等对决策树算法进行了改进,建立了适应于概念漂移检测的决策树学习算法 VFDT^[7],并且 Gama 等在之后的系列工作中,对 VFDT 树做了进一步的改进^[8-10];2000年, Street 等提出了基于多分类器集成学习的概念漂移检测算法 SEA^[11];2003年, Wang 等对多分类器中的权值变化和裁减问题进行了讨论^[12];2003年, Kolter 等人提出了增量

式的 DWM 方法^[13],并在 2005 年建立了 AddExp 算法^[14];2004年, Rushing 等提出 CBEA 方法^[15],集中讨论了一种基于聚类方法的多分类器裁减问题;2004年, Chu 等将流行的 Boosting 技术用于数据流的概念检测中,提出了自适应多分类器综合挖掘方法^[16];2004年, Chu 等考虑了噪音数据流中自适应学习问题^[17];2004年, Fan 等给出了基于 4 分类器的数据流中概念漂移挖掘的算法^[18]. SEA 和 AddExp 是目前发现的用于概念漂移检测的引用率较高的两个算法.前者是典型的批处理式方法,后者是一个增量式的概念漂移挖掘算法.批处理式挖掘方法需要当数据累计到一定数量时才能进行概念的挖掘,因此对突变概念漂移的监测效果不可能很好.增量式算法被期望来增强对突变概念的挖掘能力.然而,像 AddExp 这样的静态设置新增分类器权值参数的方法也没有很好地解决该问题.

本文提出了一个基于动态自适应修改决策权值参数的增量式的多分类器算法 M.ID4,增强了对突变的概念漂移挖掘的快速反应能力,并且给出了 M.ID4 模型的错误边界的理论分析.相比已有的典型模型及算法 SEA 和 AddExp,实验表明了 M.ID4 算法的有效性.

1 基于多分类器的概念漂移检测问题和 M.ID4 算法

本节首先给出 M.ID4 算法的伪代码描述,并对它所采用的主要技术进行归纳和分析,然后给出它的理论边界证明.

1.1 M.ID4 算法描述

M.ID4 算法的主要思想为:首先采用 ID4 作为基础分类算法,基础的分类器根据新到来的每个样本进行增量式的学习,然后利用多分类器集成技术得出全局分类结果.当多分类器 E 不适应当前数据变化时,就采用加入新分类器或者淘汰旧分类器的方法来快速适应数据的变化. M.ID4 算法的核心是多分类器 E 的更新过程,图 1(见下页)给出了它的伪代码描述,其对应符号如下:

- $\{\mathbf{x}, y\}$: 一个训练样本,其中 \mathbf{x} 为特征向量, y 为该样本的类标号;一个含有 n 个训练样本的数据流被表示成 $\{\mathbf{x}, y\}_n^1$.
- 决策值集合 $C = \{c_1, c_2, \dots, c_j, \dots, c_n\}$: C 界定了 $\{\mathbf{x}, y\}_n^1$ 中类标号对应的值域.
- 集成分类器 $E = \{E_1, E_2, \dots, E_i, \dots, E_m\}$: E_i 对应一个基础分类器.
- 基础分类器权值 $\omega = \{\omega_1, \omega_2, \dots, \omega_m\}$: ω_i 对应 E_i 的权值.
- 基础分类器权值和 W_t : t 时刻所有基础分类器权值和,即 $W_t = \sum_{j=1}^m \omega_j$.
- δ : 分类器数目阈值,当 E 中当前的分类器数 m 超过 δ 时将强制删除最差的分类器.
- β : 衰减因子,控制分类器对训练样本数据不匹配时的衰减幅度, $\beta \in [0, 1)$.
- σ_{c_j} : 决策值 c_j 所对应的决策估计值,每个 σ_{c_j} 反映了 E 对决策 $c_j \in C$ 的支持程度.

收稿日期 2006-09-20 收修稿日期 2007-07-09

Received September 20, 2006; in revised form July 9, 2007

国家自然科学基金(60496322, 60496327)资助

Supported by National Natural Science Foundation of China (60496322, 60496327)

1. 北京市多媒体与智能软件重点实验室 北京工业大学计算机学院 北京 100022
1. Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology, School of Computer Science, Beijing University of Technology, Beijing 100022

DOI: 10.3724/SP.J.1004.2008.00093

- θ : 分类器保留阈值, 当一个分类器的权值小于该阈值时将被删除。

M_ID4_Update ($\{\mathbf{x}, \mathbf{y}\}_1^n, E, m, \beta, \theta, \delta$)

```

1. FOR each  $\{\mathbf{x}, \mathbf{y}\} \in \{\mathbf{x}, \mathbf{y}\}_1^n$  BEGIN
2.   FOR all  $j \sigma_{c_j} \leftarrow 0$ ;
3.   IF ( $m = 0$ ) BEGIN
4.     构建新分类器  $E_{m+1}$ , 并插入到多分类器  $E$  中;
5.      $\omega_{m+1} = 1$ ;  $m = m + 1$ ;
6.   END
7.   FOR each  $E_i \in E$  BEGIN
8.      $c_j = E_i$  对  $\mathbf{x}$  的分类结果;
9.     IF ( $c_j \neq y$ )  $w_i \leftarrow \beta \times \omega_i$ ; //假如决策不正确就衰减
10.     $\sigma_{c_j} = \sigma_{c_j} + \omega_i$ ; // $\sigma_{c_j}$  在主程序中初始化为零
11.    根据  $\{\mathbf{x}, \mathbf{y}\}$  更新  $E_i$ ;
12.  END
13.  FOR all  $c_j \in C$  取具有最大  $\sigma_{c_j}$  对应的决策  $c_j$  作为该样本
     $E$  的整体决策;
14.  IF ( $c_j \neq y$ ) BEGIN
15.    IF ( $m = \delta$ ) BEGIN
16.      选择一个具有最小权值的分类器并删除;
17.       $m = \delta - 1$ ; // $m$  记录  $E$  中当前的分类器数, 在
        主程序中初始化为零
18.    END
19.    构建新分类器  $E_{m+1}$ , 并插入到多分类器  $E$  中;
20.     $\omega_{m+1} = \frac{1}{m} W_t = \frac{1}{m} \sum_{i=1}^m \omega_i$  且  $m = m + 1$ ;
21.  END
22.  FOR each  $E_i \in E$  IF ( $\omega_i < \theta$ ) BEGIN
23.    删除分类器  $E_i$ ;  $m = m - 1$ ;
24.  END
25. END

```

图 1 M_ID4 模型的更新过程

Fig. 1 M_ID4 update procedure

从图 1 中可以看出, M_ID4 算法对每一个到达的样本, 依次做如下工作:

- 1) 基础分类器的更新与权值衰减 (见图 1 的 7~12): 更新每个 E_i , 当预测结果错误时, 将对应的权重进行衰减;
- 2) 整体决策和集成分类器 E 的调整 (见图 1 的 13~21): 使用集成分类器 E 对样本进行预测, 当预测结果错误时, 建立新的分类器;
- 3) 删除失效的基础分类器 (见图 1 的 22~24): 通过阈值 θ 对没有保留价值的基础分类器进行删除。

1.2 算法的理论分析

在不考虑噪音的情况下, 我们通过错误边界模型来对算法进行分析。

引理. 令 ω_i^t 表示算法在时间点 t 的第 i 个基础分类器 E_i 的权值, E 中的基础分类器数目为 m , $W_t = \sum_{j=1}^m \omega_j^t$. 假设该问题有 n 个决策值 (类标号), 那么 E 在 t 时刻的整体决策 c_j 是错误的, 则 c_j 对应的决策评估值 $\sigma_{c_j} \geq W_t/n$.

证明. 根据 M_ID4_Update 算法, 有: 1) 在 t 时刻, $W_t = \sum_{j=1}^m \omega_j^t = \sum_{j=1}^n \sigma_{c_j}$; 2) 因为算法中的整体决策对应的 σ_{c_j} 是所有 σ_{c_i} ($i = 1, 2, \dots, n$) 中最大的, 因此 $\sigma_{c_j} \geq \sum_{i=1}^n \sigma_{c_i}/n = W_t/n$. 故引理成立。□

定理. 设 ω_i^t 、 m 、 n 和 W_t 的约定同引理, β 为衰减因子, 若 $m > n/(1-\beta)$, 则对于任意的时间点 $t_1 < t_2$, 且当 $\beta < 1$ 时, M_ID4 算法在时刻 t_1 和时刻 t_2 之间的错误数 μ 为

$$\mu \leq \frac{\log \frac{W_{t_2}}{W_{t_1}}}{\log(1 - \frac{1-\beta}{n} + \frac{1}{m})} \quad (1)$$

证明. 假设在 $[t_1, t_2]$ 的时间点 $t, t+1, \dots, t+\mu-1, t+\mu$, 算法进行整体预测发生错误, 那么按 M_ID4_Update 算法在这些时间点将进行相应权值衰减, 并构建一个新的权值为 W_t/m 的分类器来适应新的情况. 因此, 从 $t+1$ 时刻开始, 可以通过 W_t 和相应的约定参数 β 、 m 和 n 来讨论它们的 $W_{t+1}, W_{t+2}, \dots, W_{t+i}, \dots, W_{t+\mu}$ 的边界。

对于 $t+1$ 时刻而言, 根据上面引理, W_{t+1} 最少要有 W_t/n 部分要衰减, 即不衰减的部分小于等于 $(1 - (1/n))W_t$. 因此, $W_{t+1} \leq (1 - (1/n))W_t + W_t\beta/n + W_t/m$, 其中 $W_t\beta/n$ 和 W_t/m 是对衰减部分和增加新分类器部分的权值累计估计。

整理后, 可以得到: $W_{t+1} \leq (1 - \frac{1-\beta}{n} + \frac{1}{m})W_t$.

类似地, 可以推导出

$$W_{t+2} \leq (1 - \frac{1-\beta}{n} + \frac{1}{m})W_{t+1} = (1 - \frac{1-\beta}{n} + \frac{1}{m})^2 W_t \quad (2)$$

...

$$W_{t+\mu} \leq (1 - \frac{1-\beta}{n} + \frac{1}{m})W_{t+\mu-1} = (1 - \frac{1-\beta}{n} + \frac{1}{m})^\mu W_t \quad (3)$$

将 t 和 $t+\mu$ 用 t_1 和 t_2 替换, 式 (3) 变换成

$$(1 - \frac{1-\beta}{n} + \frac{1}{m})^\mu \geq \frac{W_{t_2}}{W_{t_1}} \quad (4)$$

注意到定理的前提 $m > n/(1-\beta)$, 所以 $(1 - (1-\beta)/n + 1/m) \leq 1$. 对式 (4) 两边同时取对数, 得

$$\mu \leq \frac{\log \frac{W_{t_2}}{W_{t_1}}}{\log(1 - \frac{1-\beta}{n} + \frac{1}{m})}.$$

□

上面定理表明, 任何时间段 $[t_1, t_2]$ 的错误边界可以用起始的状态和设定的参数界定出来, 而且不难看出, 定理中的限定条件 $m > n/(1-\beta)$ 是很容易满足的。

例 1. 假设在 $[t_1, t_2]$ 时间段到来 100 个训练数据样本, 其类标号个数 $n = 2$, 假设初始时的分类器的权值和 $W_{t_1} = 15$, 结束时 $W_{t_2} = 9$, $m = 10$, 又设衰减因子 β 为 0.5, 则可以求出在当前时间段的错误数边界 $\mu = 3.14$, 因此在此时间段的错误率小于 3.14%。

2 实验研究

为了评估算法的性能, 我们在多个不同的数据集上对 M_ID4 算法的精确度和适应性进行了实验. 本节的所有实验对应的环境是: 2.8 GHz CPU 和 1 G RAM; 操作系统为 Windows; 开发环境为基于 Java 语言的 Weka 平台, 编译运行环境为 jdk 1.5.

2.1 数据集的选择

1) SEA^[11]: SEA 概念数据集的基本结构为 $\langle f_1, f_2, \dots, C \rangle$, 其中 f_1, f_2 是条件属性 (取值 $0 \sim 10$), C 是决策属性 (当样本属性满足条件 $f_1 + f_2 \leq \theta$ 时, 属于第一类, 否则属于第二类). 我们随机产生的 6 万条样本 (5 万的训练数据和 1 万的测试数据), 包含 4 个 SEA 概念、3 次漂移. 因此, 它可以被用来作为评价标准容量、非急速变化的概念漂移挖掘的公共数据集使用.

2) 超平面 (Hyperplane)^[14,19]: 一个 m 维的超平面上的样本的属性值 x_i 是随机生成并且均匀分布在 $[0, 1]$ 区间上. 当产生的随机样本满足 $\sum_{i=1}^m a_i x_i \geq a_0$ 时, 则将此样本标记为正样例; 否则标记为负样例. 我们产生 2 千条样本, 其中蕴含 4 个概念、3 次漂移. 因此, 它可以用来检测小容量样本下概念漂移挖掘的适应性问题.

3) SEA with Quick Drift: 它是在 SEA 数据集基础上改进的数据集, 主要目的是用来评价算法捕捉快速概念变化的能力. 它有 5 万个训练数据和 1 万个测试数据, 但在每个 SEA 概念中蕴藏着一个快速的概念突变. 因此, 该数据集蕴含着 4 个 SEA 概念和 4 个快速突变概念 (由大约 2 千个样本生成). 该数据集可用来评价具有突变特性的概念漂移挖掘问题.

2.2 精确度和适应性实验

实验 1. 目的是使用标准的 SEA 概念来评价 M_ID4 和 SEA 算法. 使用 SEA 数据集进行实验, 考虑不含噪音和含 5% 噪音两种情况.

实验参数: SEA 算法的块大小 $|D| = 500$; M_ID4 算法的 $\beta = 0.5$ 、 $\theta = 0.01$ 、 $\delta = 25$. 实验结果如图 2 和图 3 所示.

实验分析: 由图 2 和图 3 可知, 在 SEA 数据集上, M_ID4 算法与 SEA 算法具有大致相当的精确度. 但是, 相比 SEA 算法, M_ID4 算法更能快速捕捉到概念的变化 (如图 2 的 275×100 和 475×100 点附近). 在 5% 噪音的情况下, M_ID4 算法整体比 SEA 的精确度要好.

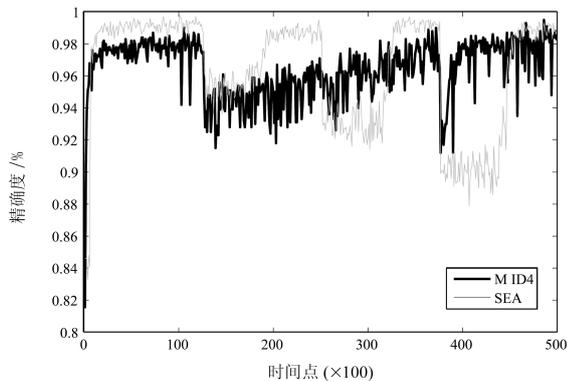


图 2 在 SEA 数据集上的精确度 (无噪音)

Fig.2 Accuracies on SEA dataset (no noise)

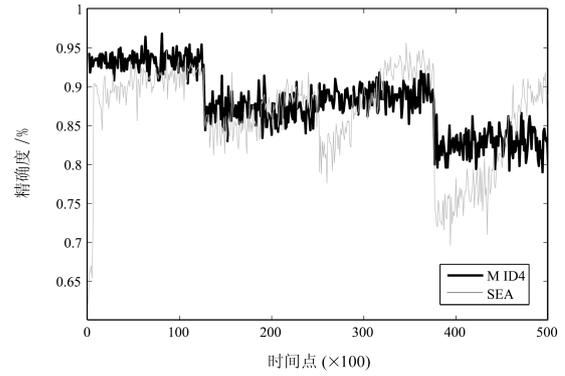


图 3 在 SEA 数据集上的精确度 (5% 噪音)

Fig.3 Accuracies on SEA dataset (5% noise)

实验 2. 目的是比较 M_ID4 算法与 SEA 和 AddExp 算法在小容量样本时对概念漂移检测的效果. 使用 Hyperplane 数据集进行实验, 考虑不含噪音和含 5% 噪音两种情况.

实验参数: SEA 算法的块大小 $|D| = 50$; AddExp 和 M_ID4 的 $\beta = 0.5$ 、 $\delta = 25$. AddExp 的新建分类器的静态调整参数 $\gamma = 0.1$, 而 M_ID4 的 $\theta = 0.01$. 实验结果如图 4 和图 5 所示.

实验分析: 由图 4 和图 5 可知, 在 Hyperplane 上, SEA 算法对概念漂移的适应能力要明显差于增量式的 M_ID4 和 AddExp 算法 (如图 4 的 100×10 点之后). 同时, M_ID4 算法在整体上要优于 AddExp 算法. 特别是, 在 5% 噪音的情况下, M_ID4 算法保持较高的精确度.

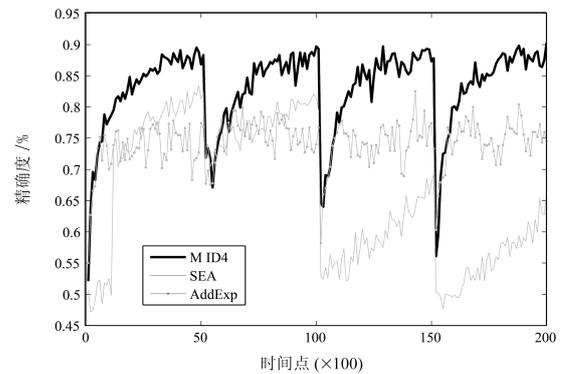


图 4 在 Hyperplane 上的精确度 (无噪音)

Fig.4 Accuracies on Moving Hyperplane (no noise)

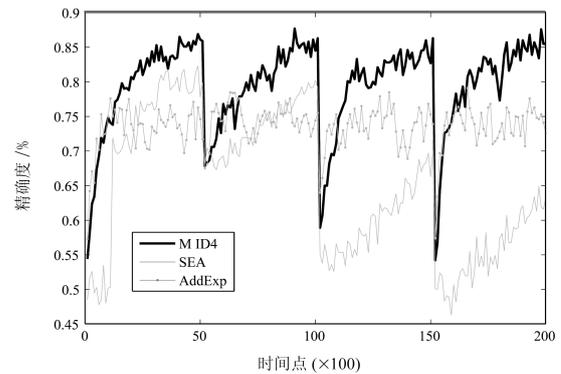


图 5 在 Hyperplane 上的精确度 (5% 噪音)

Fig.5 Accuracies on Moving Hyperplane (5% noise)

实验 3. 目的是比较 SEA 算法、AddExp 算法与 M.ID4 算法在捕获快速概念漂移的能力. 使用 SEA with Quick Drift 数据集进行实验, 考虑不含噪音和含 5% 噪音两种情况.

实验参数: SEA 算法的块大小 $|D| = 500$; AddExp 和 M.ID4 的 $\beta = 0.5$, $\delta = 25$. AddExp 的 $\gamma = 0.1$, M.ID4 的 $\theta = 0.01$. 实验结果如图 6 和图 7 所示.

实验分析: 从图 6 和图 7 中可以很明显地看出, SEA 算法无法适应快速的概念突变 (如图 6 的 30×200 和 90×200 点附近). AddExp 算法在快速漂移的数据流上的精确度也不是很理想. 我们分析, 增量方式的 AddExp 算法精确度差的原因是它使用静态方式对新分类器设置初始化权值, 这导致了新加入的分类器相对于已有的分类器出现权值过大或者过小的情况, 使整体预测出现较大偏差. 由于 M.ID4 通过动态方式初始化新分类器的权值, 使得权值在一定范围内变化, 保证了对快速概念的捕获. 在 5% 噪音的情况下, M.ID4 的精度仍然明显好于其他两个算法.

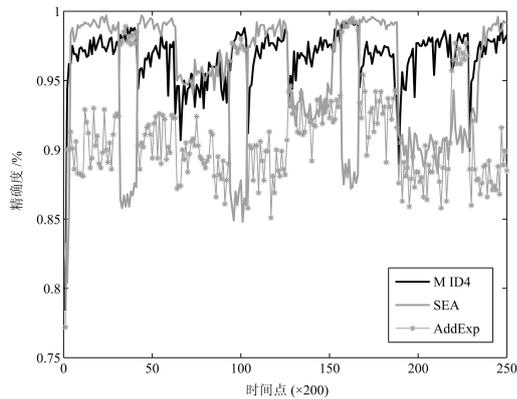


图 6 在 SEA with Quick Drift 上的精确度 (无噪音)

Fig. 6 Accuracies on SEA with Quick Drift (no noise)

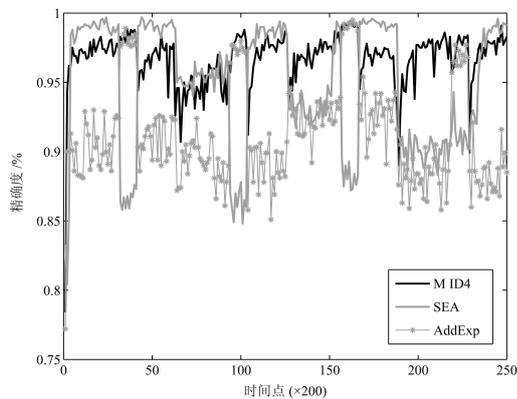


图 7 在 SEA with Quick Drift 上的精确度 (5% 噪音)

Fig. 7 Accuracies on SEA with Quick Drift (5% noise)

3 结论

数据流中的概念漂移挖掘算法是当前数据挖掘领域的一个研究热点. 已经提出的多分类器算法在某些情况下还不能很好地捕捉数据流中概念漂移的变化. 本文提出了一种基于

增量式的多分类器算法 M.ID4. 它通过多个加权分类器对样本进行分类, 并且通过动态地改变自身的模型来更好地适应当前数据流中的目标概念. 实验显示, M.ID4 算法在少量训练样本和快速概念漂移的情况下, 具有很好的精确度和适应性.

References

- 1 Widmer G, Kubat M. Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 1996, **23**(1): 69–101
- 2 Widmer G, Kubat M. Effective learning in dynamic environments by explicit context tracking. In: *Proceedings of the European Conference on Machine Learning*. London, UK: 1993. 227–243
- 3 Widmer G. Tracking context changes through meta-learning. *Machine Learning*, 1997, **27**(3): 259–286
- 4 Harries M B, Sammut C, Horn K. Extracting hidden context. *Machine Learning*, 1998, **32**(2): 101–126
- 5 Widyantoro D H, Ioerger T R, Yen J. An adaptive algorithm for learning changes in user interests. In: *Proceedings of the 8th International Conference on Information and Knowledge Management*. Missouri, USA: 1999. 405–412
- 6 Salganicoff M. Tolerating concept and sampling shift in lazy learning using prediction error context switching. *Artificial Intelligence Review (Special Issue on Lazy Learning)*, 1997, **11**(1-5): 133–155
- 7 Domingos P, Hulten G. Mining high-speed data streams. In: *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Boston, USA: 2000. 71–80
- 8 Gama J, Rocha R, Medas P. Accurate decision trees for mining high-speed data streams. In: *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Washington D.C., USA: 2003. 523–528
- 9 Gama J, Medas P, Rocha R. Forest trees for on-line data. In: *Proceedings of the 2004 ACM Symposium on Applied Computing*. Nicosia, Cyprus: 2004. 632–636
- 10 Gama J, Medas P, Rodrigues P. Learning decision trees from dynamic data streams. In: *Proceedings of the 2005 ACM Symposium on Applied Computing*. New Mexico, USA: 2005. 573–577
- 11 Street W N, Kim Y S. A streaming ensemble algorithm for large-scale classification. In: *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, USA: 2001. 377–382
- 12 Wang H, Fan W, Yu P S, Han J. Mining concept-drifting data streams using ensemble classifiers. In: *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Washington D.C., USA: 2003. 226–235
- 13 Kolter J Z, Maloof M A. Dynamic weighted majority: a new ensemble method for tracking concept drift. In: *Proceedings of the 3rd IEEE Conference on Data Mining*. Los Alamitos, USA: IEEE, 2003. 123–130
- 14 Kolter J Z, Maloof M A. Using additive expert ensembles to cope with concept drift. In: *Proceedings of the 22nd International Conference on Machine Learning*. Bonn, Germany: 2005. 449–456

- 15 Rushing J, Graves S, Criswell E, Lin A. A coverage based ensemble algorithm (CBEA) for streaming data. In: Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence. Huntsville, USA: 2004. 106–112
- 16 Chu F, Zaniolo C. Fast and light boosting for adaptive mining of data streams. In: Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Sydney, Australia: 2004. 282–292
- 17 Hulten G, Spencer L, Domingos P. Mining time-changing data streams. In: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA: 2001. 97–106
- 18 Fan W. Systematic data selection to mine concept-drifting data streams. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, USA: 2004. 128–137
- 19 Chu F, Wang Y Z, Zaniolo C. An adaptive learning approach for noisy data streams. In: Proceedings of the 4th IEEE International Conference on Data Mining. Brighton, UK: IEEE, 2004. 351–354

孙 岳 北京工业大学计算机学院硕士研究生. 主要研究方向为数据挖掘与知识发现.

(**SUN Yue** Master student at College of Computer Science and Technology, Beijing University of Technology. His research interest covers data mining and knowledge discovery.)

毛国君 北京工业大学教授. 主要研究方向为数据挖掘和分布式系统. 本文通信作者. E-mail: maoguojun@bjut.edu.cn

(**MAO Guo-Jun** Professor at Beijing University of Technology. His research interest covers data mining and distribution system. Corresponding author of this paper.)

刘 旭 北京工业大学计算机学院硕士研究生. 主要研究方向为数据挖掘与知识发现.

(**LIU Xu** Master student at College of Computer Science and Technology, Beijing University of Technology. His research interest covers data mining and knowledge discovery.)

刘椿年 北京工业大学教授. 主要研究方向为人工智能, 知识工程及数据挖掘.

(**LIU Chun-Nian** Professor at Beijing University of Technology. His research interest covers artificial intelligence, knowledge engineering, and data mining.)
