CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 36 卷第 5 期 2014 年 5 月
Vol. 36, No. 5, May. 2014

文章编号:1007-130X(2014)05-0977-09

基于特征漂移的数据流集成分类方法*

张育培,刘树慧
(郑州大学信息工程学院,河南 郑州 450052)

摘　要:为构建更加有效的隐含概念漂移数据流分类器,依据不同数据特征对分类关键程度不同的理论,提出基于特征漂移的数据流集成分类方法(ECFD)。首先,给出了特征漂移的概念及其与概念漂移的关系;然后,利用互信息理论提出一种适合数据流的无监督特征选择技术(UFF),从而析取关键特征子集以检测特征漂移;最后,选用具有概念漂移处理能力的基础分类算法,在关键特征子集上建立异构集成分类器,该方法展示了一种隐含概念漂移高维数据流分类的新思路。大量实验结果显示,尤其在高维数据流中,该方法在精度、运行速度及可扩展性方面都有较好的表现。

关键词:特征选择;特征漂移;概念漂移;数据流;互信息;集成分类器
中图分类号:TP274　　**文献标志码**:A
doi:10.3969/j.issn.1007-130X.2014.05.032

Ensemble classification based on feature drifting in data streams

ZHANG Yu-pei, LIU Shu-hui
(School of Information Engineering, Zhengzhou University, Zhengzhou 450052, China)

Abstract: In order to construct an effective classifier for data streams with concept drifting, according to the theory that different data feature has different critical degree for classification, a method of Ensemble Classifier for Feature Drifting in data streams (ECFD) is proposed. Firstly, the definite of feature drifting and the relationship between feature drifting and concept drifting is given. Secondly, mutual information theory is used to propose an Unsupervised Feature Filter (UFF) technique, so that critical feature subsets are extracted to detect feature drifting. Finally, the basic classified algorithms with the capability of handling concept drifting is chosen to construct heterogeneous ensemble classifier on the basis of critical feature subsets. This method exhibits a new idea of way to high-dimensional data streams with hidden concept drifting. Experimental results show that the method has strong appearance in accuracy, speed and scalability, especially for high-dimensional data streams.

Key words: feature selection; feature drifting; concept drifting; data stream; mutual information; ensemble classifier

1　引言

近年来,隐含概念漂移的数据流分类问题引起了人们重视,并得到广泛研究。文献[1,2]对数据流模型、存在的问题以及概念漂移定义做了详细描述,为研究工作提供充分有力的概念模型支持。数据流分类工作,尤其高速高维数据流,需应对“无限性”、“高速性”、“大数据量”和“概念漂移”。虽已有大量研究工作及成果,但仍然缺少有效处理概念漂移且高效的分类方法。

数据流分类工作大致分为两类:单个增量模型

* **收稿日期**:2012-12-17;**修回日期**:2013-02-18
通信地址:450052 河南省郑州市大学路 75 号郑州大学信息工程学院
Address: School of Information Engineering, Zhengzhou University, 75 Daxue Rd, Zhengzhou 450052, Henan, P. R. China

和自适应集成分类器。单个增量模型是通过维持和持续更新一个单分类器去应对数据流中的概念漂移[3],但学习速度慢且分类精度低,而且难以处理高速高维数据流。自适应集成分类器利用加权集成分类器在隐含概念漂移数据流分类中具有更高精度的特性[4],将概念漂移对分类结果的影响削弱在共同决策中[5,6],并利用最新或最有效分类器替换过时分类器[7]以确保分类器的先进性。然而以往集成分类器,大多基于完整数据而难以处理高维数据流且时间复杂度高。另外,特征选择技术是高维数据分类及文本分类领域的一个重要研究方向,能够有效缩减数据维度、提高分类精度并增强结果可理解性,主要可以分为滤波器法[8,9]、嵌入式法[10]及结合法。滤波器法快速简单但精度低而嵌入式法精确却复杂,故而产生了两者结合法。文献[11]提出基于特征选择的随机森林分类方法,文献[12]提出针对文本分类的 DFS 特征过滤算法,文献[13]提出基于支持向量机的特征选择和分类方法。但是,都没有考虑数据流特性因而不适于数据流分类。

本文首先描述了特征漂移概念及其与概念漂移的关系;然后提出无监督特征选择技术 UFF(Unsupervised Feature Filter),利用相邻特征集的不同来判定特征漂移发生;最后选用具有概念漂移处理能力的概念自适应快速决策树 CAFDT(Concept Adaptive Fast Decision Tree)和在线贝叶斯(OnlineNB)为基础算法,依选定的特征子集构建基础分类器,进而建立异构集成分类器,提出了基于特征漂移的数据流集成分类方法 ECFD(Ensemble Classifier for Feature Drifting)。大量对比实验结果表明,ECFD 具有较低复杂度,且在精度、运行速度及可扩展性上都有较强的表现。

2 相关概念及原理

2.1 问题描述

数据流是按时间顺序不断到来的数据序列,可形式化表示为:$S=\{d_1,d_2,\cdots,d_n,\cdots\}$,其中 $d_i=[f_1,f_2,\cdots,f_p]$ 是维度为 p 的数据点,而 d_i 所对应的已知类标号为 $c\in\{c_1,c_2,\cdots,c_k\}$。数据流分类任务是根据先验事件构建模型 M 且 d_i 的类标号 $c_i=M(d_i)$,使得 S 新到数据点 d_{i+1} 的分类概率 $P(M(d_{k+1})=c_{k+1})\geqslant 1/2$。当维度 p 非常大时,可以选择最具有数据信息的特征子集 $CFS\subset\{f_1,f_2,\cdots,f_p\}$ 来构建数据流分类模型 M,从而降低时间复杂度。同时,若 S 中两段数据 S_m 和 S_{m+1} 具有不同的模型 M,即 $M_{S_m}\neq M_{S_{m+1}}$,则利用 M_{S_m} 按时间顺序对 S_{m+1} 段的数据分类是不正确的,称此时发生概念漂移[2]。

2.2 概念定义

定义 1 (工作窗口)为应对数据流 S 的"无限性",对按时间不断到来的数据进行分段处理,当缓存中数据点数达到所定阈值 $|W|$ 时,就对缓存中数据进行处理并清空,以便继续接收 S 新到来的数据。称该缓存为工作窗口(W),大小为 $|W|$。

定义 2 (关键度)对工作窗口 W 中数据分类时,依特征 f_i 划分之后,子集合的类别纯度越高说明 f_i 越关键,因此可以用 f_i 的信息熵来表示其关键度 CD(Critical Degree),即 $CD=H(W\mid f_i)$。关键度达到阈值的特征,称为关键特征;未达到阈值的特征,称为噪特征。

定义 3 (关键特征集)对维度为 p 的数据流 S 分类时,从 p 个特征中选出对分类起关键作用的关键特征 CF(Critical Feature),也即析取关键度相对较高的特征,组成关键特征集 CFS(Critical Feature Set),即 $CFS\subset\{f_1,f_2,\cdots,f_p\}$。

定义 4 (缓存窗口)对工作窗口 W 的数据已经完成特征选择,但还未做分类,将此类数据暂存于缓存中以等待最终处理并交回数据流 S,称这段缓存为缓存窗口(CW)。

定义 5 (特征漂移)数据流 S 中,设在相邻工作窗口 W_i 和 W_{i+1} 中,利用特征选择技术分别得到关键特征集 CFS_i 和 CFS_{i+1},若 $CFS_i\neq CFS_{i+1}$,则称 S 在数据长度为 $|W_i|+|W_{i+1}|$ 的窗口中发生了特征漂移 FD(Feature Drifting)。

定义 6 (特征数据集)工作窗口 W 中数据集 S_w,通过特征选择技术去除冗余数据属性,析取关键特征,将数据点的属性数目缩减为 $|CFS|$,将只留下关键特征的 S_w 称为特征数据集(CS_w)。

定义 7 (特征分类器)选取具有概念漂移处理能力的基础分类算法,依特征数据集 CS_w 建立分类器,称该分类器为特征分类器 FC(Feature Classifier)。多个 FC 加权集成构成集成分类器,称为面向特征漂移的数据流集成分类器 ECFD。

2.3 特征选择原理

利用特征选择技术可去除重复冗余的噪特征,降低分类时间复杂度。本文认为噪特征对 CFS 的依赖度要低于关键特征对 CFS 的依赖度,可由互

信息准则[14]来表示,为此本文给出定理1。

定理1 设条件概率 $p(f_i \mid c)=p_i$,除去特征 f_i 的 CFS 记为 $\overline{f_i}$。若 $p_i > p_j$,则互信息 $I(f_i;\overline{f_i}) > I(f_j;\overline{f_j})$。

证明 依互信息链式规则[14] $I(f_i;\overline{f_i})=I(f_i;f_j)+I(f_i;\overline{f_{ij}} \mid f_j)=I(f_i;f_j)+H(\overline{f_{ij}} \mid f_j)-H(\overline{f_{ij}} \mid f_i,f_j)$;同理,$I(f_j;\overline{f_j})=I(f_i;f_j)+H(\overline{f_{ij}} \mid f_i)-H(\overline{f_{ij}} \mid f_i,f_j)$。因此,$I(f_i;\overline{f_i})-I(f_j;\overline{f_j})=H(\overline{f_{ij}} \mid f_j)-H(\overline{f_{ij}} \mid f_i)$。

由 $p(f_i \mid c)=p_i$ 和 $p(f_j \mid c)=p_j$ 且 $p_i > p_j$,可选取 z 使得 $p(z \mid c)=p_i$,而后取 $p(f_j \mid z)=(p_i+p_j-1)/(2p_i-1)$,从而得到两个相同联合分布的马尔科夫链[14]:$f_i \to c \to \overline{f_{ij}}$ 和 $z \to c \to \overline{f_{ij}}$。因此,$H(\overline{f_{ij}} \mid z)=H(\overline{f_{ij}} \mid f_i)$。应用数据处理引理[14]得到马尔科夫链 $f_j \to z \to c \to \overline{f_{ij}}$,可知 $H(\overline{f_{ij}} \mid f_j) > H(\overline{f_{ij}} \mid z)$。综上述可知 $I(f_i;\overline{f_i}) > I(f_j;\overline{f_j})$。定理证毕。 □

由定理1可知,为分类提供更多信息的特征具有更高互信息值,即关键特征对 CFS 的依赖度更高且具有更高互信息值。因此,可以通过计算所有数据特征的互信息 $I(f_i;\overline{f_i})$ 选择出 CFS。

2.4 特征漂移与概念漂移

数据流 S 中,使用 i 和 $i+1$ 表示相邻工作窗口,若 W_i 和 W_{i+1} 的过程中发生特征漂移,也即特征选择得到的关键特征集而 $CFS_i \neq CFS_{i+1}$,则 S 在 W_i 和 W_{i+1} 之中发生概念漂移,为此本文给出定理2。

定理2 数据流 S 中,特征漂移的发生必导致概念漂移的发生。

证明 首先,S 中数据段 W_i+W_{i+1} 发生特征漂移,因此由特征漂移定义知 $CFS_i \neq CFS_{i+1}$,即取得最关键特征 $top\_cri_i(\{f_1,f_2,\cdots,f_p\}) \neq top\_cri_{i+1}(\{f_1,f_2,\cdots,f_p\})$。于是由定理1可知 $\{I_1,I_2,\cdots,I_p\}_i \neq \{I_1,I_2,\cdots,I_p\}_{i+1}$,而互信息值由关键度 CD 得到,所以 $\{CD_1,CD_2,\cdots,CD_p\}_i \neq \{CD_1,CD_2,\cdots,CD_p\}_{i+1}$。其次,机器学习建立数据模型是找到对训练数据拟合的模型 $M(f_1,f_2,\cdots,f_p)$,而建立的模型 M 必定对关键度大的数据特征具有偏置性,因此 $M_i \neq M_{i+1}$。再者,由文献[2]知,若相邻数据段是由不同模型产生,则在这两段数据中发生概念漂移。故特征漂移的发生必将导致概念漂移的发生。定理证毕。 □

反之,数据流 S 中,相邻数据段 W_i 和 W_{i+1} 分别由 M_i 和 M_{i+1} 产生且 $M_i \neq M_{i+1}$,则在数据段 W_i+W_{i+1} 中发生概念漂移,但是概念漂移的发生不一定会引起特征漂移的发生,为此本文给出定理3。

定理3 数据流 S 中,多数发生概念漂移的情况会导致发生特征漂移。

证明 设数据质心为 $G(f_1,f_2,\cdots,f_p)$,数据半径为 $R(f_1,f_2,\cdots,f_p)$。数据段 W_i+W_{i+1} 中发生概念漂移,也即数据流 S 的数据分布发生改变[2]。而本文认为数据分布发生变化有两个原因:数据分布质心发生移动和数据分布半径发生变化。

(1)仅当 $G_{i+1} \neq G_i$,即 $|G_{i+1}-G_i| > 0$,则某些数据特征的关键度发生了偏移而未得到反向补偿;

(2)仅当 $R_{i+1} \neq R_i$,即 $R_{i+1}=c \times R_i$,致使数据特征质心发生分散但 $|G_{i+1}-G_i|=0$,这种情况属于局部概念漂移[2];

(3) $G_{i+1} \neq G_i$ 且 $R_{i+1} \neq R_i$,某些数据特征的关键度必定发生了变化,数据整体质心分散。

因此,概念漂移发生时,数据特征 CD 不一定发生变化。据实验统计,实际中第(3)种混合情形占85%以上,故多数概念漂移会致使某些特征关键度变化,从而使 CFS 改变,引发特征漂移。故而多数概念漂移会引发特征漂移。定理证毕。 □

由定理2和定理3得知,特征漂移是概念漂移的充分条件,且现实中多数概念漂移引发特征漂移,因此本文提出以处理特征漂移替代处理概念漂移,由数据分布半径引起的概念漂移交给基础分类器去处理。由此可降低分类复杂度和运行时间,且可即时检测到大部分概念漂移,从而提高分类精度。

3 特征选择 UFF 和 ECFD 算法

3.1 特征选择技术 UFF

由定理1可知,计算各数据特征与余下特征集的互信息值,其中互信息值较大者拥有更大的关键度,本文选取最大的 num 个特征为关键特征集。而互信息的计算需要对数据特征熵进行计算,本文采用文献[15]的熵估计法:

$$H_k=\frac{p}{n}\sum_{i=1}^{n}\log t_{ik}+GM(n)-GM(k)+\log c_p$$

其中,t_{ik} 表示数据点 d_i 和第 k 个近邻的欧几里得距离,$GM(k)$ 为双伽马函数。特别地,$H_1=\frac{p}{n}\sum_{i=1}^{n}\min_{j\neq i}(\log \|d_i-d_j\|)+GM(n)-GM(1)+\log c_p$。另外,由信息论可知,特征 f_i 和 $\overline{f_i}$ 的互信息

值 $I(f_i \mid \overline{f_i}) = H(f_i) + H(\overline{f_i}) - H(CFS)$。于是本文给出 UFF 特征度量,如公式(1)所示。

$$UFFStr = \frac{1}{n}\sum_{i=1}^{n}\log t_{ik} + \frac{p-1}{n}\sum_{i=1}^{n}\log l_{ik} \quad (1)$$

其中,t_{ik} 表示数据点 d_i 和第 k 个近邻在一维子空间的欧几里得距离;l_{ik} 为数据点 d_i 与第 k 个近邻在 $p-1$ 维子空间的欧几里得距离。由定理 1 和公式(1),同时利用数据流"无限性",通过已标记数据和已构造分类器验证精度确定关键特征的个数。本文提出特征选择算法 UFF,如算法 1 所示。

算法 1 UFF 算法

输入:数据集 S,特征集 F,已标记数据 T 和已构造分类器 M,正随机小数 ∂。

输出:关键特征集 CFS。

```
1:  while F≠null do
2:      利用公式(1)计算特征 f_i 的 UFFStr 值并按
        从小到大放入数组 UFFS;
3:  end while
4:  while UFFS≠null do
5:      num++;
6:      将 UFFS 中 num_top_highest 个特征作为关
        键特征集,利用 T 和 M 计算分类精度 P_i ;
7:      If | P_i − P_{i−1} | ≤ ∂ then
8:          CFS=num_top_highest−1;
9:      end if
10: end while
```

为了处理数据流,本文将 UFF 算法使用于工作窗口中,以窗口数据为数据集 S。当工作窗口数据点数目达到阈值时,便启动 UFF 算法进行特征选择,同时清空工作窗口 W 以继续接收数据流 S 的数据,将 W 中的数据交予缓存窗口 CW 等待最终处理。当相邻工作窗口得到的关键特征集不相同时,就断定此时有特征漂移发生。

3.2 ECFD 算法

本文提出 ECFD 算法的目的是对隐含概念漂移的数据流进行分类,该方法从特征漂移入手,并利用特征选择技术析取关键特征集,构建特征集成分类器,从而降低时间和空间复杂度,且提高分类精度。ECFD 算法流程包含四个步骤,如图 1 所示。

(1)利用 UFF 算法对工作窗口数据进行特征选择。随着数据点进入工作窗口 W,当 $Len(W) = |W|$ 时,利用 UFF 算法对 W 中数据集进行特征选择,得到关键特征集 CFS_i,同时将 W 中数据全部移入缓存窗口 CW。

(2)判断是否有特征漂移发生。若 $CFS_i = CFS_{i-1}$,则没有特征漂移发生,即特征漂移检测

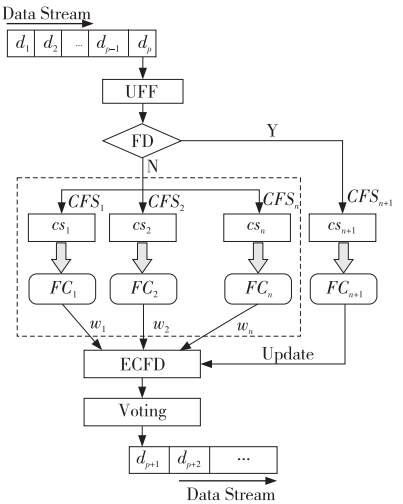


Figure 1 Algorithm of ECFD

图 1 ECFD 算法流程示意图

FD 过程返回 NO,此时需对具有 CFS_i 的特征分类器特征漂移 FC 进行再学习,使 FC 提高分类精度且获取最近数据信息;若 $CFS_i \neq CFS_{i-1}$,则有特征漂移发生,即 FD 过程返回 YES,此时需利用新得到的特征数据集 CS 训练新的特征分类器 FC_{new}。

(3)特征集成分类器 ECFD 学习与实时更新。对特征分类器 FC 的再训练,本文采用文献[16]所提出的平均距离测试泊松分布方法得到 Poisson(1)的值 num,对特征数据集 CS 中的数据拟合训练 Num 次。而对于 ECFD 的更新,首先找出其中精度最高的和最差的特征分类器,利用新得到的 CS 训练,并选用与精度最高分类器同样的基础算法,得到新的 FC_{new},然后将 FC_{new} 替换掉最差 FC。据定理 2 和定理 3 及以上分析,提出特征集成分类器 ECFD 的学习算法,如算法 2 所示。

算法 2 ECFD 学习算法

输入:数据流 S,集成分类器数目 N,工作窗口尺寸 W_l。

输出:ECFD 集成分类器。

```
1:  以 N * W_l 个数据点初始化 ECFD;
2:  while S≠null do
3:      if Len(W)< W_l then
4:          将数据点 d 加入 W;
5:      else
6:          执行 UFF 得到 FCS_i 及 CS_i;
7:          if CFS_i ≠ CFS_{i+1} then
8:              以 CS_i 评估 ECFD 得出精度最高 b-
                FC 及其类型 T 和 w-FC;
9:              以类型 T 和数据 CS_i 训练新分类
                器 n-FC;
10:             以 n-FC 替换 w-FC;
```

```
        else
            令 num=Poisson(1)[16];
            以 CS_i 再训练 ECFD 中同特征 FC num 次;
        end if
    end if
end while
```

(4)利用加权投票对数据进行分类。由于ECFD进行了特征选择,故依文献[4]的集成分类器权值计算方法和本文特征度量值 $UFFStr$ 的定义,提出如公式(2)所示的 ECFD 加权法。公式(2)以分类所用关键特征集中所有特征的特征度量值之和弥补特征选择带来的偏置性,从而使各分类对于所有数据特征相对公平;同时,使用文献[4]所提出的方法来区别各分类器权值比例。

$$W_i = \frac{|UFFStr(CFS_i)|}{(MSE_r - MSE_i) + \partial} \tag{2}$$

其中,$|UFFStr(CFS_i)|$ 表示 CFS_i 中所有特征度量值之和;$MSE_r - MSE_i$ 为文献[4]的权值方案;∂ 为非常小的随机常数。对数据进行分类时,首先据公式(2)计算 FC 权重,然后进行加权投票,本文给出 ECFD 投票分类算法,如算法 3 所示。

算法 3 ECFD 分类算法

输入:未分类数据点 d。

输出:d 的类标矩阵 $\boldsymbol{C} = [c_1 \quad c_2 \quad \cdots \quad c_k]$ 及最大类标号。

```
依据公式(2)计算 ECFD 中所有 FC 的权值 w;
for FC_i ∈ ECFD
    if FC_i 分类为 C_t then
        c_t = c_t + w_i
    end if
end for
返回向量 C 和类标号 argmax(C);
```

3.3 算法性能分析及比较

3.3.1 算法时间复杂度

设 v 为属性值的最大个数,c 为类别的最大个数,l 为树最长的路径长度,p 为数据流维度,k 为ECFD中包含特征分类器的数目。由于分类算法与以往算法只是权值公式的不同,所以这里只对ECFD学习算法进行分析。初始化不计入持续学习时间,ECFD算法主要包含以下三个步骤。

(1)特征选择。UFF算法需要计算 $UFFStr$ 值以及使用分类器对属性验证选择,其平均时间复杂度为 $O(p * |W| * v + |CFS|)$。

(2)特征漂移检测。漂移检测只需比较相邻关键特征集,平均时间复杂度为 $O(|CFS|)$。

(3)特征分类器构建。特征分类器的构建与选择的基础分类器有关,ECFD选择CVFDT和OnlineNB作为基础分类算法,这两个算法的平均时间复杂度为 $O(l * c * v * |CFS|)$ 和 $O(v * |W| * |CFS|)$,由于机器学习数据独立且随机,所以在选择算法时同样具有随机性,故可取其均值来表示,即 $O((l * c + |W|) * v * |CFS|/2)$。

因此,ECFD学习算法的时间复杂度为三者之和,合并转换后如公式(3)所示。

$$O(ECFD) = 2|CFS| + v * |CFS| * ((p * |W|)/|CFS| + l * c/2 + |W|/2) \tag{3}$$

公式(3)显示了,ECFD使用了特征选择将数据流维度 p 降低为 $|CFS|$,从而从整体上降低了时间复杂度。当 $p/|CFS| \to 1$ 时,也即数据维度通过UFF没有降低,主要表现在处理低维度数据流时,此时由于UFF技术反而增加了ECFD的学习时间;当 $p/|CFS| \to +\infty$ 时,也即数据维度通过UFF有效降低,主要表现在处理高维度数据流时,此时公式(3)随 p 增加而有所增大,但对同维度而言,整体是随 $|CFS|$ 值的减小而大大降低。总体来说,ECFD花费了特征选择的时间对数据特征进行选择,使后两步时间复杂度降低,故ECFD能够大大降低对高维数据流处理的时间复杂度。

因为ECFD是主动检测概念漂移,为了方便与比较,假设数据流 S 共有 $numW$ 个窗口,含 $numC$ 次概念漂移,则 $O_1(ECFD) = 2|CFS| + v * |CFS| * |numC| * ((p * |W|)/(|CFS| * |numC|) + (l * c)/2 + |W|/2)$,而 Bagging(NB)[17]时间为:$O_2(Bnb) = v * |CFS| * |W| * |numC|$;Bagging(CVFDT)[17]时间为:$O_3(Bcvfdt) = v * |CFS| * l * c * |numW|$。因此,$O_1/O_2$ 和 O_1/O_3 都随着 p 和 $numC$ 的增大而趋向于0,说明在高维度数据流且概念漂移次数比较多的情况下,ECFD相对Bagging算法需要的运行时间更少。

3.3.2 算法精度和抗噪性

ECFD算法主动处理概念漂移和过滤无用特征,从而减少不必要的算法运行和缩小分类器的规模,进而达到降低时间复杂度和增加分类精度的目的。

(1)概念漂移的检测与处理。ECFD方法采用先检测后处理的办法,且有特征漂移和概念漂移的双重检测,大大提高检测能力,从而提高分类精度,而目前多数方法是边训练边检测给分类器学习造

成巨大负担且检测效果不佳。

(2)对信息的提取能力。ECFD 算法采用 UFF 特征提取方法,有效地利用了数据流的特性,从而使得不再像其他大多数方法一样疲于应对数据流的特性。虽然特征提取丢弃了一些数据信息,但是同时也提高了 ECFD 算法的抗噪性,因为关键特征含有大量对分类有益的信息,而噪特征则含有大量的噪声信息。因而 ECFD 可以达到相对高的分类精度。

(3)分类策略。ECFD 的分类策略采用加权投票方式,一定程度上矫正了权值的偏置性,对优良的特征分类器更为有利,具有更好的公平性,从而更有利凸出正确类别。

总之,从理论上 ECFD 可以达到相对高的分类精度,且具有更好的抗噪性。

4 实验分析

为了对 ECFD 算法全面测试,首先对 UFF 特征选择与已有先进技术进行对比实验,然后对概念漂移检测能力进行检验并做对比,最后对算法整体性能进行测试对比。实验中,设置 ECFD 拥有 10 个特征分类器,工作窗口尺寸为 1 000。运行环境为双核 CPU 主频 2.27 GHz,内存 2 GB,使用 VS2010 平台 C++编程实现。

4.1 数据集描述

(1)人工数据集。便于与其他方法对比,本文依托 MOA 软件平台[18]分别使用仿真数据流产生器和移动超平面数据流产生器,产生含有突变概念漂移和缓慢概念漂移的数据集,大小为 100 000,并分别记为 SEA 和 HYP。

(2)真实数据集。为真实反映 ECFD 对网络数据实时动态处理情况,将入侵检测竞赛数据库 KDDCup99 模拟为数据流,大小为 494 022;为了检测 ECFD 的抗噪性,选用 UCI 的 LED 数据集,大小为100 000;另外,通过雅虎 Web 服务接口采集的提供者与商家的雅虎购物数据库,记为数据集 YSD,大小为840 000; http://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets 手写识别数据集,记为 HWD,大小为60 000。

4.2 结果与分析

4.2.1 特征选择

为了验证 UFF 算法的特征选择能力,分别在各数据集运行 KP-SVM[13] 与 FCBF[9] 各 50 次并计算平均结果,如表 1 所示。由于 KP-SVM 和 UFF 是具体的嵌入式方法,而 FCBF 是独立的过滤方法,所以 FCBF 要比 KP-SVM 和 UFF 高效。但是,从表 1 可以看到,KP-SVM 和 UFF 得到的关键特征均数都要比 FCBF 要小,这是因为嵌入式特征选择要比过滤器方法对特征的评估更加有效。同时,表 1 显示了 UFF 具有与 KP-SVM 相当甚至更好的特征选择能力,但 UFF 利用了数据流的特性故而适合于对数据流的分类。关键特征选择结果说明,ECFD 方法中特征选择 UFF 算法的有效性,能够去除冗余数据属性,且充分利用数据流的特性缩短了算法运行时间,更利于对高维数据流的处理,进而降低了分类器学习的时间复杂度。

Table 1 Critical feature selection

表 1 关键特征选择

数据集	特征总数目	算法	关键特征均数
SEA	3	KP-SVM	2.2
		FCFB	2.6
		UFF	2.2
HYP	10	KP-SVM	6.2
		FCFB	7.4
		UFF	6.0
KDDCup99	34	KP-SVM	2.6
		FCFB	3.1
		UFF	2.6
LED	24	KP-SVM	14.4
		FCFB	16.6
		UFF	15.8
YSD	903	KP-SVM	40.2
		FCFB	67.6
		UFF	47.4
HWD	780	KP-SVM	39.4
		FCFB	102.2
		UFF	30.8

4.2.2 概念漂移

为了验证 ECFD 方法检测概念漂移的能力,与文献[19]所提出的 PSCCD 漂移检测算法进行对比,对所有数据集进行特征漂移检测 50 次,结果如表 2 所示。从表 2 可以看到,ECFD 在 ESEA、HYP 和 YSD 数据集上误报次数高于 PSCCD,是因为 ECFD 算法是通过检测特征漂移达到概念漂移检测的目的,由定理 3 也可以得知特征漂移检测并不能完全检测概念漂移,但是少部分非特征漂移的概念漂移由分类器处理。而表 2 中 ECFD 的失

报次数要低于 PSCCD,主要是因为 ECFD 算法是每个窗口都能进检测,而 PSCCD 算法是通过统计积累检测故而会大量失报持续时间短的概念漂移。从表 2 也可以看到,ECFD 算法的概念漂移检测能力和 PSCCD 的相当,甚至更好,这是因为大多数概念漂移是由特征漂移引起的。总之,不论是人工数据集还是真实数据集,ECFD 方法都能够有效地检测概念漂移,且分类结果证明失报率在可以接受的范围内。

Table 2 Dection of concept drifting

表 2 概念漂移检测

数据集	检测次数	检测方法	误报次数	失报次数
SEA	392	PSCCD	15	86
		ECFD	17	75
HYP	416	PSCCD	24	107
		ECFD	29	62
KDDCup99	3 987	PSCCD	477	517
		ECFD	431	375
LED	261	PSCCD	18	55
		ECFD	16	57
YSD	6 777	PSCCD	666	344
		ECFD	705	341
HWD	79	PSCCD	17	19
		ECFD	7	21

4.2.3 性能比较

(1)为了清楚地看到 ECFD 方法的有效性,在数据集上以先测试后训练最后再丢弃的顺序,分别与 AWE(NB)[4]、AWE(C4.5)[4]、Bagging(NB)[17] 和 Bagging(CVFDT)[17] 进行对比。将 ECFD 及 AWE 算法和 Bagging 算法在所有数据集上运行 50 次并计算平均结果,如表 3 所示。从表 3 中可以看出,除在 KDDCup99 数据集外,ECFD 分类精度都高于其他算法,这是因为实际中多数概念漂移是由特征漂移引起的,ECFD 准确地检测特征漂移同时使用了具有概念漂移检测能力的算法构建特征分类器,所以有着比其他算法更好的概念漂移处理能力。而在 KDDCup99 数据集上也有着不错的精度,但不如 Bagging(CVFDT),主要是由于 ECFD 方法使完整数据集转为特征数据集,因而缺少一定的训练维度造成的。另外,从表 3 还可以看出,除两个人工数据集外,ECFD 方法运行时间也较其他方法快,维度越高越能体现其时间效率,这主要是因为 ECFD 使用 UFF 特征选择使得构建分类器更简单,说明 ECFD 方法更容易应对高维数据流。而在两个低维模拟数据集上,ECFD 特征选择的时间相对构建分类器来说比较大,从而该方法时效性不如 Bagging(NB)。总之,实验结果也充分证实了文中对算法分析的结论,不论是人工数据集还是真实数据集,ECFD 方法都有较高的分类精度和时间效率。

(2)为了验证 ECFD 算法在特征选择之后数据流分类中的优势,与 ACE 集成分类算法[11] 和 KP-SVM 分类算法[13] 分别在各数据集上运行 50 次,同样 ACE 算法取 10 个分类器集成,结果如图 2 所示。

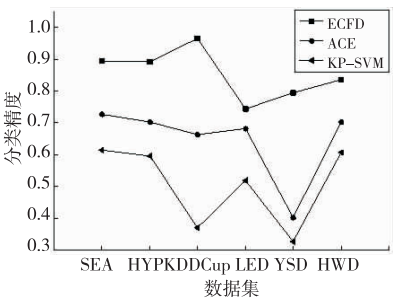


Figure 2 Result of kinds of dataset with concept drifting

图 2 含漂移各数据集

图 2 说明 ECFD 算法在各实验数据集上的分类精度均高于其他两者,而 ACE 算法的分类精度要高于 KP-SVM,这主要是因为数据集中隐含有概念漂移,而 ECFD 主动处理概念漂移,ACE 由于

Table 3 Algorithm comparison on precision and time

表 3 算法分类精度和运行时间比较

数据集	AWE(NB)		AWE(C4.5)		Bagging(NB)		Bagging(CVFDT)		ECFD	
	精度/%	时间/s	精度/%	时间/s	精度/%	时间/s	精度/%	时间/s	精度/%	时间/s
SEA	87.08	6.67	88.01	27.64	87.94	3.60	89.12	21.47	89.38	22.76
HYP	88.94	20.13	72.76	28.12	86.92	8.17	88.90	40.14	89.16	12.03
KDDCup99	96.21	285.63	95.19	288.00	92.63	230.00	97.75	209.60	96.42	141.00
LED	74.05	74.39	72.13	40.36	73.93	26.25	73.79	83.00	74.27	29.11
YSD	49.67	1435.00	55.61	764.66	44.35	802.18	57.48	1005.20	79.36	37.10
HWD	11.24	2116.20	78.40	1246.70	12.20	1286.00	24.33	1462.00	83.52	439.00

集成分类器而对概念漂移有一定应对能力,KP-SVM 没有考虑概念漂移。但是,该实验还表明 ECFD 算法运行速度要低于其他两者,这是因为 ECFD 特征选择完之后需要去寻找合适的特征集数目。因此,ECFD 可高精度处理隐含概念漂移的数据流分类。

(3)为了验证 ECFD 的抗噪性,本文选用 HYP 数据集和 KDDCup99 数据集分别加入 5%、10%、15%、20%和 25%的噪声数据,各算法分别运行 50 次并计算分类精度均值,如图 3 和图 4 所示。图 3 中随噪声数据的增多,ECFD 分类精度下降约 17%,而其他算法的下降都大于 20%;同时,图 4 中随噪声数据的增多,ECFD 分类精度下降约 10%,而其他算法的下降都大于 16%,且在噪声数据达到 20%以及更高时,ECFD 精度超过了 Bagging(CVFDT)。实验表明,ECFD 算法比其他算法具有更好的抗噪性,这是因为 ECFD 做了特征选择而使得噪特征的噪声对该算法没有大的影响。

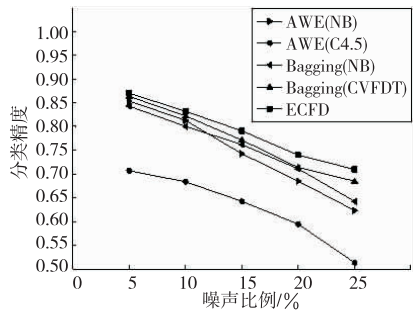


Figure 3 Result of dataset HYP

图 3 HYP 数据集

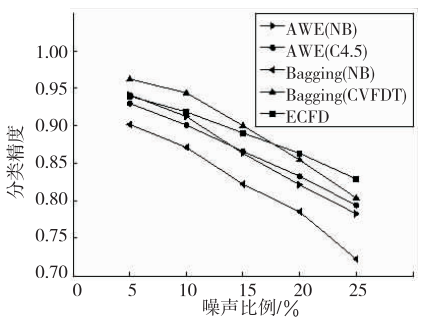


Figure 4 Result of dataset KDDCup99

图 4 KDDCup99 数据集

5 结束语

本文提出了一种基于特征漂移的数据流集成分类方法(ECFD),首先给出特征漂移的概念及其与概念漂移的关系,论证了可以通过检测特征漂移来检测概念漂移的原理;然后为应对数据流特性,利用互信息理论提出无监督特征选择 UFF 技术并检测概念漂移;最后提出了 ECFD 的学习算法,并根据改造后的权值计算方法给出 ECFD 分类算法。ECFD 充分利用数据流的特性比较成功地解决数据流难题,且特征选择算法和基础分类算法是可选的,为隐含概念漂移的数据流分类展示了一个新思路。理论分析和实验结果都表明 ECFD 算法具有更高的分类精度和更好的抗噪性。但是,对该方法的研究才刚开始,对特征选择算法的稳定性及算法框架的完整性有待研究,这将是下一步的研究方向。

参考文献:

[1] Babcock B, Babu S, Datar M, et al. Models and issues in data stream systems [C]// Proc of ACM PODS, 2002:16-24.

[2] Tsymbal A. The problem of concept drift:Definitions and related work [R]. TCD-CS-2004-15. Ireland: Trinity College Dublin, Department of Computer Science, 2004.

[3] Hulten G, Spencer L, Domingos P. Mining time-changing data streams [C]// Proc of ACM SIGKDD, 2001:97-106.

[4] Wang H, Fan W, YU P S, et al. Mining concept-drifting data streams using ensemble classifiers [C]// Proc of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003:226-235.

[5] Masud M M, Gao J, Han J, et al. Classification and novel class detection in concept-drifting data streams under time constraints[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(6):859-874.

[6] Zhang P, Zhu X, Tan Jian-long, et al. Classifier and cluster ensembles for mining concept drifting data streams [C]// Proc of IEEE International Conference on Data Ming, 2010:1175-1180.

[7] Sattar H, Ying Y, Zahra M, et al. Adapted one-vs-all decision tree for data stream classification [J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(5):624-637.

[8] Inza I, Larranaga P, Blanco R, et al. Filter versus wrapper gene selection approaches in DNA microarray domains[J]. Artificial Intelligence in Medicine, 2004, 31(2):91-103.

[9] Lei Y, Huan L. Feature selection for high-dimensional data: A fast correlation-based filter solution[C]// Proc of the 20th ICML'03, 2003:856-863.

[10] Hsu W H. Genetic wrappers for feature selection in decision tree induction and variable ordering in Bayesian network structure learning[J]. Information Sciences, 2004, 163(1-3):103-122.

[11] Tuv E, Borisov A, Runger G, et al. Feature selection with ensembles, artificial variables, and redundancy elimination [J]. Journal of Machine Learning Research, 2009, 10:1341-1366.

[12] Alper K U, Serkan G, A novel probabilistic feature selection method for text classification[J]. Knowledge-Based Systems,2012, 36:226-235.

[13] Maldonado S, Webber R, Basak J. Simultaneous feature selection and classification using kernel-penalized support vector machines [J]. Information Sciences, 2011, 181(1): 115-128.

[14] Cover T M, Thomas J A. Elements of information theory [M]. New York:Wiley-Interscience, 1991.

[15] Goria M, Leonenko N, Mergel V, et al. A new class of random vector entropy estimators and its applications in testing statistical hypotheses[J]. Journal of Nonparametric Statistics, 2005 17(3):277-297.

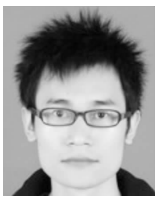
[16] Gabor J S, Maria L R. Mean distance test of poisson distribution [J]. Statistics & Probability Letters, 2004, 67(3): 241-247.

[17] Breiman L. Bagging predictors [J]. The Journal of Machine Learning Research, 1996,24(2):123-140.

[18] Bifet A, Holmes G, Kirkby R. MOA:Massive online analysis [J]. The Journal of Machine Learning Research, 2010, 11:1601-1604.

[19] Niloofar M, Sattar H, Ali H. A precise statistical approach for concept change detection in unlabeled data streams [J]. Computers and Mathematics with Applications, 2011, 62: 1655-1669.

作者简介:



张育培(1985 -),男,河南嵩县人,硕士生,CCF 会员(E200027694G),研究方向为机器学习与数据挖掘。**E-mail**:zzuiezhyp@163.com

ZHANG Yu-pei, born in 1985, MS candidate, CCF member(E200027694G), his research interests include machine learning, and data mining.

2014 中国计算机大会(CNCC2014)征文通知

第十一届 CCF 中国计算机大会(CCF China National Computer Congress,CCF CNCC 2014)将于 2014 年 10 月 23～25 日在河南郑州国际会展中心举行,承办单位是信息工程大学。CNCC 是由中国计算机学会(CCF)于 2003 年创建的系列学术会议,已在不同的城市举办十届,现每年一次。

CNCC 旨在探讨计算机及相关领域最新进展和宏观发展趋势,展示中国学术界、企业界最重要的学术、技术事件和成果,使不同领域的专业人士能够获得探讨交流的机会并获得所需信息。CNCC2014 将有逾 2000 人参会交流,有近百项科研成果进行展示,是中国 IT 领域的一次盛会。

CNCC2014 现公开征集会议论文,征文范围涵盖计算机领域各方向,要求是没有公开发表过的原创性论文。本次大会不出版会议论文集,拟挑选不超过 50 篇的优秀论文刊登在《计算机学报》上,其他录用论文将推荐到《小型微型计算机系统》、《计算机科学》、《计算机工程与应用》、《计算机工程与科学》等 CCF 会刊发表。《计算机学报》和《小型微型计算机系统》录用文章将在 2014 年 10 月发表。

征稿范围(但不限于)

(1)计算机系统结构:高性能计算、CPU 设计与多核处理器技术、计算机网络与新一代互联网、传感器网络和物联网、物理信息融合系统、对等计算与网格计算、云计算与数据中心网络、网络存储系统、网络安全、信息与内容安全。

(2)计算机软件与理论:计算机科学理论、程序设计语言与编译技术、软件测试、形式化方法、操作系统与系统软件、数据库技术、数据挖掘、内容检索、软件工程。

(3)计算机应用技术:人工智能与模式识别、机器学习、知识工程、智能控制技术、图形学与人机交互、虚拟现实与可视化技术、多媒体技术、中文信息技术、电子政务与电子商务、生物信息学。

投稿方式

稿件内容要求以中文书写,并隐去作者姓名和单位,请提交 PDF 文件

论文模板:中文论文模板

大会网站:http://cncc.ccf.org.cn(请务必正确注册邮箱,并填写详细个人信息,包括联系电话以及通讯地址,以便联络论文修改和寄发录用通知等事宜,如信息不全,将会影响论文评审。)

联系:cncc_pr@ccf.org.cn(请在邮件标题中注明"CNCC2014 征文")

重要日期

论文提交截止日期:2014 年 5 月 10 日　　录用通知发出日期:2014 年 8 月 1 日

CNCC 召开日期:2014 年 10 月 23 日～25 日