一种不平衡数据流集成分类模型

欧阳震诤[1],罗建书[1],胡东敏[2],吴泉源[2]
(1. 国防科技大学理学院,湖南长沙 410073;2. 国防科技大学计算机学院,湖南长沙 410073)

摘 要: 针对不平衡数据流的分类问题,结合基于权重的集成分类器与抽样技术,本文提出了一种处理不平衡数据流集成分类器模型.理论分析与实验验证表明,该集成分类器具有更低的计算复杂度,更能适应存在概念漂移的不平衡数据流挖掘分类,其整体分类性能优于基于权重的集成分类器模型,能明显提升少数类的分类精度.

关键词: 分类; 集成分类器; 不平衡数据流; 概念漂移
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112 (2010) 01-0184-06

An Ensemble Classifier Framework for Mining Imbalanced Data Streams

OUYANG Zhen-zheng[1], LUO Jian-shu[1], HU Dong-min[2], WU Quan-yuan[2]
(1. *Science School, National University of Defense Technology, Changsha, Hunan* 410073, *China*;
2. *Computer School, National University of Defense Technology, Changsha, Hunan* 410073, *China*)

Abstract: Many real world data streams mining applications involve learning from imbalanced data streams, where such applications expect to have a higher predictive accuracy over the minority class, however most classification model assume relatively balanced data streams, they cannot handle imbalanced distribution. In this paper, we propose a novel ensemble classifier framework (IMDWE) for mining concept-drifting data streams with imbalanced distribution by using weighted ensemble classifier framework sampling technique including over-sampling and under-sampling. Our empirical study shows that the IMDWE is superior and have improves both the efficiency in learning the model and the accuracy in performing classification over the minority class.

Key words: classification; ensemble classifier; imbalanced data streams; concept drift

1 引言

分类技术是数据流挖掘研究领域的重要课题,一个高效的数据流分类算法应能在有效处理概念漂移的同时保持相当好的分类精度.近年来,研究人员在该领域做了大量卓有成效的工作,集成分类器方法是一种被广泛采用的方法,Wang[1]等从理论上证明了集成分类器的性能要优于单个分类器.在集成分类器方法中,基于权重的集成分类器方法(Weight Ensemble Classifier,简称WE)[1~4]被普遍认为是具有较高分类精度的方法,它们能很好的处理数据流分类中的概念漂移问题.然而,集成分类器方法与目前多数数据流分类器的设计一样,它们是基于数据流中类的分布是大致平衡这一假设的,设计者通常假定训练数据集中各类所包含的样本数大致相当,而这一基本假设在许多现实数据流应用问题中并不成立,不平衡数据流在许多实际应用中经常碰到,如信用卡的欺诈辨识、网络入侵检测、Web 挖掘、信息检索等等.在这些应用中,少数类的分类辨识更加重要.而目前大部分分类方法虽然整体上具有较高的分类精度,可是对少数类的辨识率却很低[5],因此适当降低多数类的分类精度,以换取更高的少数类的分类精度就成为了不平衡数据流挖掘分类的主要目标.为能有效处理带概念漂移的不平衡数据流挖掘分类问题,本文基于 WE 模型,提出了一种不平衡数据流集成分类器模型 IMDWE.

2 相关工作

2.1 不平衡数据集分类的基本方法

在机器学习领域,鉴于不平衡学习分类的重要现实意义,研究者对该问题进行了大量研究,当前研究主要集中于数据层的处理、分类算法的改进、设计以及分类器性能评价标准设计等几个方面.

从数据层面的处理方法来看,基本目标都是如何使得少数类与多数类的样本数趋于平衡,常用的方法是过抽样(over-sampling)与欠抽样(under-sampling)或者是两种方法的结合.过抽样方法通过增加训练集中少数类的样本来提高分类器的性能,而对多数类样本不做删减,

收稿日期:2009-05-15;修回日期:2009-08-10

最简单的办法是复制少数类样本,改进的算法是在少数类中插值样本,比较著名的是 SMOTE[6].过抽样由于增加了训练集样本的规模,会导致构建分类器的时间增加.欠抽样与过抽样相反,它通过减少多数类样本的数量,从而提高少数类的分类性能,但是当随机去掉一些多数类样本时,可能造成多数类样本的一些重要信息的丢失[7].

从分类算法的改进、设计来看,目前主要集中于几个方面:一是通过调整不同类样本的错分代价来给训练集中的样本加权,重构训练集[8~10];二是改进传统算法,设计代价敏感的分类算法[11];三是多分类器的集成学习.

在机器学习领域,目前不平衡学习分类的研究对象主要是静态不平衡数据集,比较普遍的做法是综合利用抽样技术与集成方法,既利用过抽样或欠抽样来提高分类器对少数类的分类性能,又利用集成的优点来提高整体分类性能[12~14].

2.2 不平衡数据流分类性能的评价标准

多类别的分类问题通常可以简化为二分类问题,在二分类问题中,称少数类为正类(positive class),多数类为负类(negative class).下面主要就不平衡数据流中的二分类问题进行讨论.

表 1 混合矩阵

	被分为正类	被分为负类
实际为正类	TP	FN
实际为负类	FP	TN

数据流的分类问题研究中,分类精度是一个主要的性能评价指标,然而单一分类精度的评价标准对于不平衡数据流来说是不合适的,到目前为止,机器学习领域中对于不平衡数据集分类问题中常用的标准有:ROC 曲线分析以及基于混合矩阵(confusion metric,如表 1 所示)的如查全率(recall)、查准率(precision)、F-Value 值以及 G-mean 等[15~17].查全率(recall)、查准率(precision)、F-Value 值以及 G-mean 的计算公式如下:

$$precision = TP/(TP + FP) \tag{1}$$

$$recall = TP/(TP + FN) \tag{2}$$

$$F\text{-}value = \frac{(1+\beta^2)\cdot recall \cdot precision}{\beta^2 \cdot recall + precision} \tag{3}$$

$$G\text{-}mean = \sqrt{\frac{TP}{TP+FN}\cdot\frac{TN}{TN+FP}} \tag{4}$$

F-Value 的定义中,参数 β 是可调的,通常取值为 1,由公式(3)可以看出,只有当查全率(recall)与查准率(precision)都比较大时,F-Value 值才会增大,因此 F-Value 值的大小能正确反映正类的分类性能.G-mean 反映的是分类器的总体分类性能,它是正类的分类精度 $TP/(TP+FN)$与负类的分类精度 $TN/(TN+FP)$二者乘积的平方根,只有正类与负类的分类精度都高时,G-mean 才会比较高.

2.3 WE 集成分类器

令 t 表示任一时间戳,C_t 表示在该时间戳到达的数据块,则数据流模型可描述为$\{\cdots, C_{t-1}, C_t, C_{t+1}, \cdots\}$.从数据流中连续采样得到 n 个数据块$\{D_1, D_2, \cdots, D_n\}$,$D_n$ 是最近的数据块,下一时刻到来的数据块记为 D_{n+1}.WE 集成分类器选择某种分类算法 F(如 VFDT、SVM、Naive Bayes 等)对每个数据块进行学习,得到 n 个基础分类器$f_i = F(D_i)(i = 1,2,\cdots,n)$,然后根据不同的方法[1~4]对每个基础分类器 f_i 赋予权重 w_i,最后组合各个基础分类器构成一个整体分类器 f_{WE}来对 D_{n+1}中数据进行预测分类,f_{WE}的计算公式为公式(5):

$$f_{WE}(x) = \sum_{i=1}^{n} w_i f_i(x) \tag{5}$$

其中 $\sum_{i=1}^{n} w_i = 1, \quad 0 \leqslant w_i \leqslant 1$.

在 D_{n+1}中数据类标号未知的情况下,WE 集成分类器认为 D_n 的类分布与 D_{n+1}的类分布最为接近.WE 采用 D_n 作为各个基础分类器测试集来求取各它们相应的权重,相对于各个基础分类器,WE 集成分类器能很好的提高分类精度.由于各个基础分类器相对简单,大部分基础分类器的构建复杂度都是超线性的(super-linear),因此构建 WE 集成分类器也是高效的,并且 WE 集成分类器本身就可以使其能够并行扩展和在线分类大数据集,其也能很好的处理数据流中的概念漂移问题.

3 IMDWE 集成分类器的设计与分析

在未来数据类标号未知的情况下,WE 集成分类器在保持较高分类精度的同时,也能很好的处理数据流中的概念漂移问题,但由于其设计考虑是基于数据流中类的分布是大致平衡这一假设,其不能有效处理不平衡数据流中的分类问题.基于 WE 集成分类器,综合利用抽样技术,我们提出了一种集成分类器模型——IMDWE 集成分类器模型.

3.1 IMDWE 集成分类器

设数据流数据是块状到达的,数据流中数据的类标识只有两个,分别为正类与负类,正类样本远少于负类样本.从数据流中连续采样得到 m 个大小固定数据块$\{C_1, C_2, \cdots, C_m\}$,$C_m$ 是最近的数据块,下一时刻到来的数据块记为 C_{m+1},IMDWE 集成分类器的详细构造过程如下(如图 1 所示):

①过抽样.PIO 总线负责按照某种规则将$\{C_1, C_2, \cdots, C_m\}$数据块中的正类样本块 $POS_1, POS_2, \cdots, POS_m$ 中样本采入正类样本集合 $ALLP$ 中,POS_i 为 C_i 的所有正类样本集合,最简单的办法是将 $POS_1, POS_2, \cdots, POS_m$ 中样本全部采样加入正类样本集合 $ALLP$ 中,则$|ALLP| = \sum_{i=1}^{m} |POS_i|$.PIO 总线负责对 $POS_1, POS_2, \cdots, POS_m$

中的每个样本记录其所在数据块的时间戳,如 m 值等.此时$\{C_1, C_2, \cdots, C_m\}$数据块中只剩负类样本,形成样本块$\{NEG_1, NEG_2, \cdots, NEG_m\}$.

②欠抽样.设定值 M,$0 < M \leqslant 0.9$,一般取 $M = 0.5$(根据具体应用问题确定),称 M 为正类比例,计算分块上限值 B,B 按公式(6)计算.

$$B = \left\lfloor \frac{\min\limits_{1 \leqslant i \leqslant m} |NEG_i|}{\left[\frac{1-M}{M} \cdot |ALLP|\right]} \right\rfloor > 1 \tag{6}$$

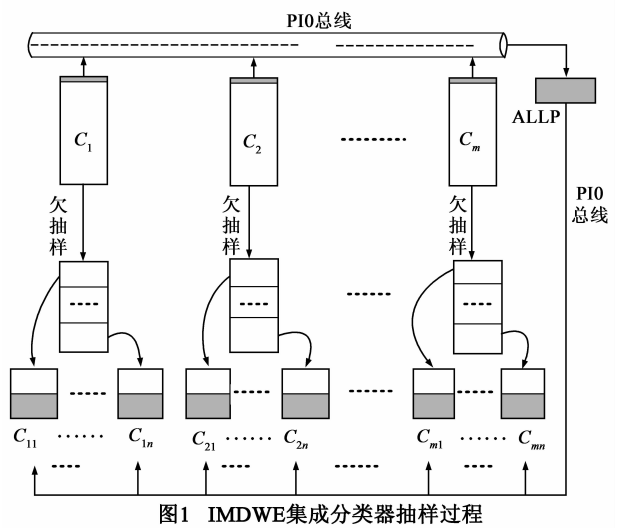

图1 IMDWE集成分类器抽样过程

取正整数 n,$1 \leqslant n \leqslant B$,一般取 $n = B$.对每个数据块 $NEG_1, NEG_2, \cdots, NEG_m$ 进行随机欠抽样,得到 m 个数据块 $NEG'_1, NEG'_2, \cdots, NEG'_m$,每块中的样本数相等,且

$$|NEG'_i| = n \cdot \left[\frac{1-M}{M} \cdot |ALLP|\right] \tag{7}$$

③合成训练集.将每个 $NEG'_1, NEG'_2, \cdots, NEG'_m$ 随机分为 n 个不相交的子集,得到 mn 个数据块 C'_{11}, $C'_{12}, \cdots, C'_{1n}, C'_{21}, C'_{22}, \cdots, C'_{2n}, \cdots, C'_{m1}, C'_{m2}, \cdots, C'_{mn}$,再将这 mn 个数据块通过 PIO 总线分别与 $ALLP$ 进行合成,得到 mn 个数据块$\{C_{11}, C_{12}, \cdots, C_{1n}, C_{21}, C_{22}, \cdots, C_{2n}, \cdots, C_{m1}, C_{m2}, \cdots, C_{mn}\}$,且 $|C_{ij}| = |ALLP| + |C'_{ij}|$,$1 \leqslant i \leqslant m$,$1 \leqslant j \leqslant n$,此时每个块中正类样本与负类样本是平衡的,其中正类样本所占比例 b 为公式(8)所示:

$$M \leqslant b \leqslant \frac{M \cdot |ALLP|}{|ALLP| - M} \tag{8}$$

④WE 集成.从分类算法库 Q 中选择一种分类学习算法 F,算法 F 分别在数据块$\{C_{11}, C_{12}, \cdots, C_{1n}, C_{21}, C_{22}, \cdots, C_{2n}, \cdots, C_{m1}, C_{m2}, \cdots, C_{mn}\}$上进行学习得到 mn 个基础分类器 $f_{ij} = F(C_{ij})$,$i = 1,2,\cdots,m$,$j = 1,2,\cdots,n$,采用 C_{mn}作为各基础分类器的测试集,然后根据某种的权重分配方法(见 3.2 节)对每个基础分类器 f_{ij} 赋予权重 w_{ij},最后组合各个基础分类器构成一个整体分类器 f_{WE}来对 C_{m+1}中数据进行预测分类,f_{WE}的计算公式为公式(9):

$$f_{WE}(x) = \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} f_{ij}(x) \tag{9}$$

其中 $\sum\limits_{i=1}^{m} \sum\limits_{j=1}^{n} w_{ij} = 1$, $0 \leqslant w_{ij} \leqslant 1$.

这样我们就得到了 IMDWE 集成分类器.

当公式(6)中 $B \leqslant 1$ 或者 $n|C_{ij}| > |C_i|$时,表明 PIO 总线从$\{C_1, C_2, \cdots, C_m\}$中抽取的正类样本数需要调整,此时 PIO 总线可以按照某种取样规则对样本序列$\{POS_1, POS_2, \cdots, POS_m\}$进行抽样,考虑到数据流中的概念漂移情况导致的复杂情况,我们采用与 WE 集成分类器类似的假设,认为时间戳 $i(1 \leqslant i \leqslant m)$越靠近 m,则 C_i 的类分布与 C_{m+1}的类分布越接近,相应的从 POS_i 中抽取的样本要多,POS_i 的采样比率 u_i 就越高,比如可按 $u_i = u_{i+1} - \triangle u$($\triangle u$ 为一常数)这样的递减规则去抽样,或将时间戳最小的正类移除,如不从 POS_1, POS_2 中进行抽样.

在步骤①的过抽样过程中,不推荐采用简单复制样本与类似于 SMOTE 的样本插值方法,一方面简单复制某个块的正类样本,没有增加正类的任何新的信息,还可能导致在这个块上的基础分类器过学习(over-fitting).另一方面采用 SMOTE 方法对某个块内的样本进行插值生成新的样本,也可能导致这个块上的基础分类器过学习,并且最近的研究指出,采取集成加简单随机欠抽样的方法比 SMOTE 加集成的方法更有效[13,14],算法复杂度更低,更适合于高速的数据流挖掘.

3.2 权重的确定策略

在 IMDWE 集成分类器的 WE 集成学习过程中,在求每个基础分类器 f_{ij}的对应权重 w_{ij}时,采用 C_{mn}作为各个基础分类器的测试集,且根据分类性能的不同目标,采取下列不同的权重确定策略.

①在追求 IMDWE 集成分类器整体分类性能的前提下,采用 G-mean 评价标准来评价 IMDWE 集成分类器的整体分类性能,对于权重 w_{ij}的确定,参照文献[1]的确定办法,详细描述如下:

把数据块 C_{mn}中样本表示为向量形式$(\boldsymbol{x}, y)$,$\boldsymbol{x}$ 为向量,$y \in$ {正类,负类},设 $f_{ij}(y|\boldsymbol{x})$是基础分类器 f_{ij}预测 $\boldsymbol{x}$ 为类 y 的概率,y 的类分布为 $p(y)$,则基础分类器 f_{ij}对样本$(\boldsymbol{x}, y)$的分类误差为 $1 - f_{ij}(y|\boldsymbol{x})$,基础分类器 f_{ij}在测试集上的均方误差为:

$$\mathrm{MSE}_{ij} = \frac{1}{|C_{mn}|} \sum_{(x,y) \in C_{mn}} (1 - f_{ij}(y|\boldsymbol{x}))^2 \tag{10}$$

另一方面,我们希望 $\boldsymbol{x}$ 被分为类 y 的概率接近 y 的类分布为 $p(y)$,因此随机对 $\boldsymbol{x}$ 进行预测分类的均方误

差为:

$$\mathrm{MSE}_r = \sum_y p(y)(1-p(y))^2 \tag{11}$$

简单起见,不妨就取 p(正类) = M, p(负类) = 1 − M. $\mathrm{MSE}_r = M(1-M)$.

IMDWE 的各基础分类器应比随机猜想具有更高的分类精度,因此对于在测试集上的均方误差不小于 MSE_r 的基础分类器丢弃,其对应权重为零.若令

$$e_{ij} = \mathrm{MSE}_r - \mathrm{MSE}_{ij} \tag{12}$$

则基础分类器 f_{ij} 的权重 w_{ij} 为

$$w_{ij} = \begin{cases} \dfrac{e_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} e_{ij}}, & e_{ij} > 0 \\ 0, & e_{ij} \leqslant 0 \end{cases} \tag{13}$$

表 2　混合代价矩阵

	被分为正类的代价	被分为负类的代价
实际为正类	0	FNcost
实际为负类	FPcost	0

②在错误分类代价极为敏感的应用中,如信用卡欺诈,把一个欺诈误认为是正常的将产生不可接受的损失,此时就需要最大化提高分类器对于正类的分类精度,此时采用 F-Value 评价标准来评价 IMDWE 集成分类器的分类性能.我们引入错分类的代价[8~10],具体见表 2,其中负类错分代价 FPcost(指实际为负类而被错分正类的代价),正类错分代价 FNcost(指实际为正类而被错分负类的代价),而无错分类的代价为零,FPcost、FNcost 的值根据具体应用问题来确定,FPcos < FNcost,基于此 IMDWE 采取如下的权重确定策略.

设样本($\boldsymbol{x}$, y)被分类为 y' 的代价为 $J_{y,y'}(\boldsymbol{x})$,根据混合代价矩阵(表 2),基础分类器器 f_{ij} 预测在测试集上的总体代价为:

$$ct_{ij} = \sum_{(x,y)\in C_{mn}} \sum_{\gamma} J_{y,\gamma}(\boldsymbol{x}) \cdot f_{ij}(\gamma \mid \boldsymbol{x}) \tag{14}$$

另一方面,随机对 $\boldsymbol{x}$ 进行预测分类的代价为:

$$ct_r = (1-M) \cdot FNCost + M \cdot FP\ Cost \tag{15}$$

基础分类器的错分类代价应比随机猜想低,因此对于在测试集上的错分类代价不小于 $|C_{mn}|\, ct_r$ 的基础分类器丢弃,其对应权重为零.若令

$$\delta_{ij} = |C_{mn}|\, ct_r - ct_{ij} \tag{16}$$

则基础分类器 f_{ij} 的权重 w_{ij} 为

$$w_{ij} = \begin{cases} \dfrac{\delta_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} \delta_{ij}}, & \delta_{ij} > 0 \\ 0, & \delta_{ij} \leqslant 0 \end{cases} \tag{17}$$

3.3　复杂度分析

设分类学习算法 F 在大小为 s 的数据块的构建基础分类器算法复杂度为 $O(f(s))$,而求取权重的算法复杂度与测试集的大小成线性关系的,于是 WE 集成分类器构建基础分类的复杂度为 $O(mf(s) + Ks)$, $m \gg K$.在构建 IMDWE 集成分类器过程中,由于将 WE 集成分类器中的训练样本数据块的大小进行了 n 等分,因此 IMDWE 集成分类器构建各个基础分类器的复杂度为 $O(nmf(s/n) + Ks/n)$,对大多数基础分类器算法,其复杂度 $O(f(s))$是远大于线性复杂度的,因此 IMDWE 集成分类器与 WE 集成分类器相比起来拥有更低的计算复杂度.

4　实验分析

4.1　实验设置

算法是在 Rapidminer[18]和 Weka[19]系统基础上实现的,实验数据集采用 KDDCUP'99 数据集[20].经过分析,KDDCUP'99 数据集中 normal、neptune、smurf 三类样本所占比例为 97.3%,其他连接类型所占比例很小,将 KDDCUP'99 中 normal 视为负类,ipsweep 类型视为正类,则负类样本数所占比例为 98.73%,将此近百万多个样本进行混淆(shuffling)后再进行随机取样,构建四个块大小不同的实验数据流,第一个实验数据流为 50 个大小为 20000 的数据块,记为 stream-01.第二个实验数据流为 100 个大小为 10000 的数据块,记为 stream-02.第三个实验数据流为 200 个大小为 5000 的数据块,记为 stream-03.第四个实验数据流为 20 个大小为 40000 的数据块,记为 stream-04.分类学习算法 f 采用决策树(DT),并基于 Weka 包实现,参数取其默认值,各基础分类器的权重采用第一种确定策略.实验环境是:Intel 奔腾双核 − 1.6G 的 CPU,内存大小 2G,Java heap space JVM 设为 − Xms64m − Xmx512m,操作系统为 windowXP.

4.2　实验结果与分析

进行实验时,实验数据流 stream-01、stream-02、stream-03 在训练窗口中样本块数为 5,stream-04 为 3.为验证 IMDWE 在不同正类比例 M 下的执行情况,我们分别取了 9 个 M 值,从 0.1 到 0.9.

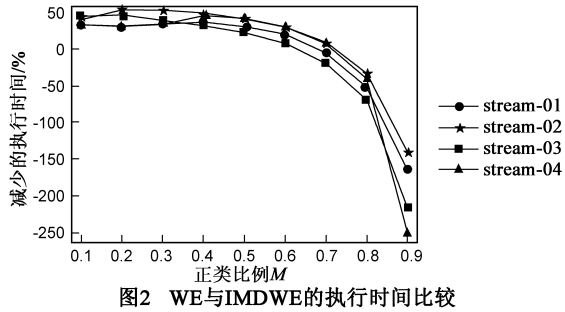

图2　WE与IMDWE的执行时间比较

从构建集成分类器的平均执行时间来看,实验比较结果表明(图 2):在正类比例 M 取值在 0.1 至 0.6 的

范围内,IMDWE 集成分类器构建分类器的执行时间比 WE 集成分类器少;对于 stream-01,平均执行时间最多减少 36.2%,最少减少 28.4%;对于 stream-02,平均执行时间最多减少 51.2%;对于 stream-03,平均执行时间最多减少 45.7%;对于 stream-04,平均执行时间最多减少 36.2%.当 M 超过 0.6 时,此时 IMDWE 平均执行时间出现增加的趋势,主要是由于训练窗口中样本个数出现明显增加($n|C_{ij}|>|C_i|$),从而导致训练时间的增加.

实验采用 G-mean 评价 IMDWE 的整体分类性能.从图 3 可以看出,对于实验取定的 9 个 M 值,IMDWE 的 G-mean 值相比 WE 的 G-mean 值都出现了明显的提升,也即 IMDWE 的整体分类性能要明显优于 WE 的整体分类性能.尤其对于 stream-01,IMDWE 的整体分类性能提升明显,最大提升为 12.4%,最小为 10.4%;而对于 stream-02、stream-04,IMDWE 的整体分类性能提升在 5.8%到 8%之间;对于 stream-03,提升度在 2.1%到 3.3%之间.

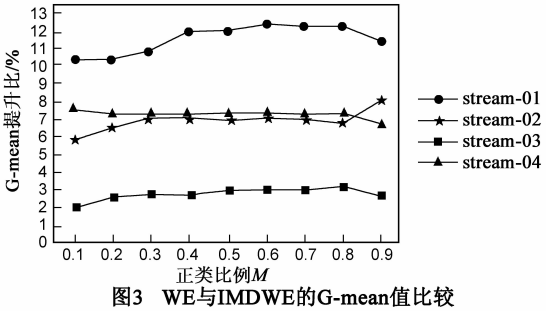

图3 WE与IMDWE的G-mean值比较

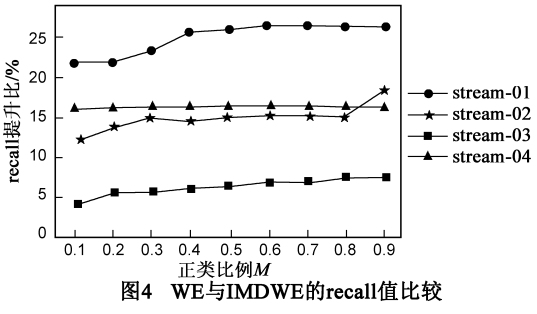

图4 WE与IMDWE的recall值比较

从提升整个正类的分类精度来看,从图 4 可以看出,相比 WE,IMDWE 对于提高正类的分类精度是明显的,尤其对于 stream-01,recall 值最大提升达到 26.7%,最小也在 22%以上,而最小的提升出现在 stream-03,相对提升度在 4.2%到 7.6%之间.从 F-Value 值来看,从图 5 可以看出,相比 WE,IMDWE 对于不同的 M 值,出现了不同的结果,但是对于 stream-01,在 M 取 0.1 到 0.5 的范围内,F-Value 值提升都在 10.8%以上,最大达到 11.8%;对于 stream-02,在 M 取 0.1 到 0.5 的范围内,F-Value 值提升最大在 5.9%,最小在 0.4%;而对于 stream-03 与 stream-04,只有在 $M=0.1$ 附近才出提升,而其他 M 值都出现降低,此时主要是 IMDWE 的 precision 值出现了降低(图 6).

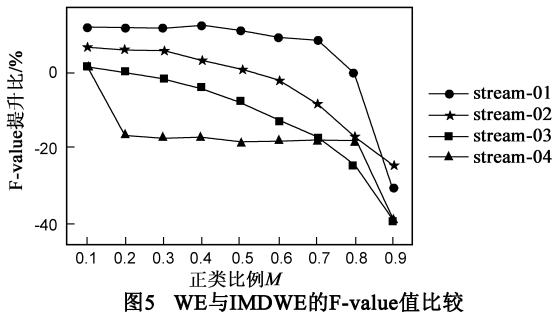

图5 WE与IMDWE的F-value值比较

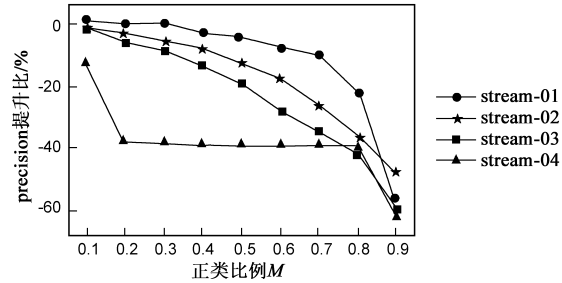

图6 WE与IMDWE的precision值比较

5 结论

目前多数数据流分类器的设计是基于数据流中类的分布是大致平衡这一假设的,而某些实际应用中对于少数类的分类性能是重点关注的.本文基于权重集成分类器,综合利用抽样技术,提出了一种处理不平衡数据流的集成分类模型——IMDWE 集成分类器模型.实验验证表明:IMDWE 集成分类器的整体分类性能优于 WE 集成分类器,能明显提高少数类的分类精度,并且构建模型的算法复杂度更低,更适合高速数据流的分类挖掘.

从实验中可以看出,相比 WE 集成分类器,IMDWE 集成分类器在提升少数类的 F-value 值时对于不同的正类比例 M 出现了不稳定性,这主要是由于 precision 值的降低过快造成的.因此根据应用问题中正负类样本比例的不同、数据流流速(块大小)的不同,如何选取适当 M 值变得非常重要,这也是我们下一步研究的方向.

参考文献:

[1] H Wang, et al. Mining concept-drifting data streams using ensemble classifiers[A]. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York: ACM Press, 2003. 226 – 235.

[2] M Scholz, R Klinkenberg. An ensemble classifier for drifting concepts[A]. Proceedings of the Second International Workshop on Knowledge Discovery in Data Streams [C]. Porto, Portugal: Springer, 2005. 53 – 64.

[3] Wei Fan. Systematic data selection to mine concept – drifting

data streams[A]. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York: ACM Press, 2004. 128 – 137.

[4] J Z Kolter, M A Maloof. Using additive expert ensembles to cope with concept drift[A]. Proceedings of the 22nd International Conference on Machine Learning[C]. New York: ACM Press, 2005. 449 – 456.

[5] G M Weiss, F Provost. Learning when training data are costly: the effect of class distribution on tree induction[J]. Journal of Artificial Intelligence Research, 2003, (19): 315 – 354.

[6] N V Chawla, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, (16): 321 – 357.

[7] G M Weiss. Mining with rarity: a unifying framework[J]. ACM SIGKDD Explorations, 2004, 6(1): 8 – 19.

[8] C Elkan. The foundations of cost – sensitive learning[A]. Proceedings of the 17th International Joint Conference on Artificial Intelligence[C]. Seattle, Washington, USA: Morgan Kaufmann Publishers Inc, 2001. 973 – 978.

[9] M Ciraco, M Rogalewski, G Weiss. Improving classifier utility by altering the misclassification cost ratio[A]. Proceedings of the 1st International Workshop on Utility-based Data Mining [C]. New York: ACM Press, 2005. 46 – 52.

[10] C X Ling, V S Sheng. Cost-sensitive learning and the class imbalance problem[A]. Encyclopedia of Machine Learning [M]. New York: Springer. 2008.

[11] M G Karagiannopoulos, et al. Local cost sensitive learning for handling imbalanced data sets[A]. Proceedings of the 15th Mediterranean Conference on Control & Automation[C]. Athens, Greece: IEEE Press, 2007. 1 – 6.

[12] N V Chawla, et al. SMOTEBoost: improving prediction of the minority class in boosting[A]. Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases[C]. Cavtat Dubrovnik, Croatia: Springer, 2003. 107 – 119.

[13] J V Hulse, T M Khoshgoftaar, A Napolitano. Experimental perspectives on learning from imbalanced data[A]. Proceedings of the 24th International Conference on Machine Learning[C]. New York: ACM Press, 2007. 935 – 942.

[14] C Seiffert, et al. RUSBoost: improving classification performance when training data is skewed[A]. Proceedings of the 19th International Conference on Pattern Recognition[C]. Tampa, Florida, USA: IEEE Computer Society, 2008. 1 – 4.

[15] H Han, et al. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[A]. Proceedings of International Conference on Intelligent Computing (ICIC' 05)[C]. Hefei, China: Springer, 2005. 878 – 887.

[16] M V Joshi, V Kumar, R C Agarwal. Evaluating boosting algorithms to classify rare classes: comparison and improvements [A]. Proceedings of the 2001 IEEE International Conference on Data Mining[C]. San Jose, CA, USA: IEEE Computer Society, 2001. 257 – 264.

[17] T Fawcett. An introduction to ROC analysis[J]. Pattern Recognition Letters, 2006, 27(8): 861 – 874.

[18] I Mierswa, et al. YALE: Rapid Prototyping for Complex Data Mining Tasks[A]. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York: ACM Press, 2006. 935 – 940.

[19] I H Witten, E Frank. Data Mining: Practical Machine Learning Tools and Techniques [M]. San Fransisco: Morgan Kaufmann Publishers Inc, 2005.

[20] Irvine. KDD Cup 1999 Data [OL]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, 1999 – 10 – 28/2009 – 09 – 01.

作者简介:

欧阳震诤　男,1973 年出生于湖南益阳.国防科技大学理学院讲师.主要从事数据库与数据挖掘、分布式软件等方面的研究.

E-mail: oyz21@163.com

罗建书　男,1956 年出生于湖南新化.博士,教授.国防科技大学计算数学专业博士生导师,主要从事小波分析与数据压缩、电磁拓扑与计算等方面的研究.

E-mail: ljsh3115@sina.com