Journal of Computer Applications
计算机应用,2017,37(7):1994 - 1998
ISSN 1001-9081
CODEN JYIIDU
2017-07-10
http://www.joca.cn

文章编号:1001-9081(2017)07-1994-05
DOI:10.11772/j.issn.1001-9081.2017.07.1994

基于主动学习不平衡多分类 AdaBoost 算法的心脏病分类

王莉莉[1,2*],付忠良[1,2],陶　攀[1,2],胡　鑫[1,2]
(1. 中国科学院 成都计算机应用研究所,成都 610041;　2. 中国科学院大学,北京 100049)
(*通信作者电子邮箱 wanglili8773@163.com)

摘　要:针对不平衡分类中小类样本识别率低问题,提出一种基于主动学习不平衡多分类 AdaBoost 改进算法。首先,利用主动学习方法通过多次迭代抽样,选取少量的、对分类器最有价值的样本作为训练集;然后,基于不确定性动态间隔的样本选择策略,降低训练集的不平衡性;最后,利用代价敏感方法对多分类 AdaBoost 算法进行改进,对不同的类别给予不同的错分代价,调整样本权重更新速度,强迫弱分类器"关注"小类样本。在临床经胸超声心动图(TTE)测量数据集上的实验分析表明:与多分类支持向量机(SVM)相比,心脏病总体识别率提升了 5.9%,G-mean 指标提升了18.2%,瓣膜病(VHD)识别率提升了 0.8%,感染性心内膜炎(IE)(小类)识别率提升了 12.7%,冠心病(CAD)(小类)识别率提升了 79.73%;与 SMOTE-Boost 相比,总体识别率提升了 6.11%,G-mean 指标提升了 0.64%,VHD 识别率提升了 11.07%,先心病(CHD)识别率提升了3.69%。在 TTE 数据集和 4 个 UCI 数据集上的实验结果表明,该算法在不平稳多分类时能有效提高小类样本识别率,并且保证其他类别识别率不会大幅度降低,综合提升分类器性能。

关键词:主动学习;不平衡分类;多分类 AdaBoost;多类别分类;心脏病分类
中图分类号:TP391.4;TP181　　**文献标志码**:A

Heart disease classification based on active imbalance multi-class AdaBoost algorithm

WANG Lili[1,2*], FU Zhongliang[1,2], TAO Pan[1,2], HU Xin[1,2]
(1. *Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu Sichuan* 610041, *China*;
2. *University of Chinese Academy of Sciences, Beijing* 100049, *China*)

Abstract: An imbalance multi-class AdaBoost algorithm with active learning was proposed to improve the recognition accuracy of minority class in imbalance classification. Firstly, active learning was adopted to select the most informative samples for classifiers through multiple iterations of sampling. Secondly, a new sample selection strategy based on uncertainty of dynamic margin was proposed to tackle the problem of data imbalance in the multi-class case. Finally, the cost sensitive method was adopted to improve the multi-class AdaBoost algorithm: giving different class with different misclassification cost, adjusting sample weight update speed, and forcing weak learners to "concern" minority class. The experimental results on clinical TransThoracic Echocardiography (TTE) data set illustrate that, when compared with multi-class Support Vector Machine (SVM), the total recognition accuracy of heart disease increases by 5.9%, G-mean improves by 18.2%, the recognition accuracy of Valvular Heart Disease (VHD) improves by 0.8%, the recognition accuracy of Infective Endocarditis (IE) (minority class) improves by 12.7% and the recognition accuracy of Coronary Artery Disease (CAD) (minority class) improves by 79.73%; compared with SMOTE-Boost, the total recognition accuracy of heart disease increases by 6.11%, the G-mean improves by 0.64%, the recognition accuracy of VHD improves by 11.07%, the recognition accuracy of Congenital Heart Disease (CHD) improves by 3.67%. The experiment results on TTE data and 4 UCI data sets illustrate that when used in imbalanced multi-class classification, the proposed algorithm can improve the recognition accuracy of minority class effectively, and upgrade the overall classifier performance while guaranteeing the recognition accuracy of other classes not to be decreased dramatically.

Key words: active learning; imbalance classification; multi-class AdaBoost; multi-class classification; heart disease classification

0　引言

不平衡数据集通常指类别间数量相差较大的数据集,称具有少量样本的那些类为小类,而具有大量样本的那些类为大类,传统分类方法追求的是整体识别率,而对小类的识别率一般较低。目前,对于不平衡数据集的分类问题的研究主要分为两大类:一类是改变训练集样本分布,降低不平衡程度;另一类是适当修改现有算法,使之适应不平衡分类问题。

收稿日期:2017-01-12;**修回日期**:2017-02-27。　**基金项目**:四川省科技支撑计划项目(2016JZ0035);中国科学院西部之光项目。
作者简介:王莉莉(1987—),女,河南周口人,博士研究生,主要研究方向:机器学习、模式识别、数据挖掘;　付忠良(1967—),男,重庆合川人,教授,硕士,主要研究方向:机器学习、模式识别;　陶攀(1988—),男,河南安阳人,博士研究生,主要研究方向:机器学习、数据挖掘;　胡鑫(1987—),男,贵州遵义人,硕士研究生,主要研究方向:数据仓库、数据挖掘。

降低不平衡度的方法包括训练集重采样方法和训练集划分方法。SMOTE(Synthetic Minority Over-sampling Technique)算法[1]是一种简单有效的上采样方法,首先为每个小类样本随机选出几个邻近样本,并在该样本与这些邻近样本的连线上随机取点,生成无重复的新的小类样本。Japkowicz 等[2]的实验研究了不平衡数据对经典算法的影响,包括决策树 C4.5、BP(Back Propagation)神经网络和支持向量机(Support Vector Machine, SVM)等,由于支持向量机对分类性能影响较大的是少数的支持向量,因此该方法对数据不平衡度相对不敏感。Chen 等[3]通过修剪大类的支持向量,使支持向量个数平衡,提高稀有类的识别率。Chan 等[4]将大类样本按照合理的类别样本分布比例随机地划分成一系列不相交子集,并分别与小类样本融合,组成一系列平衡的分类子问题,训练成子分类器,最后通过元学习将这些子分类器集成组合分类器。Lu 等[5-6]采用最小最大模块化 SVM 模型,提出了"部分对部分"任务分解策略,控制每个子问题的规模和平衡度,并根据先验知识和训练集的样本分布,制定有效的分解规则。SMOTEBoost(Synthetic Minority Over-sampling Technique and Boosting)算法[7]的每次迭代使用 SMOTE 生成新的样本,不再使用 AdaBoost(Adaptive Boosting)集成算法中的权值调整规则,使 Boosting 算法更专注于小类样本中的难分样本。这些方法虽然能有效地提升小类的识别率,但同时也忽略了很多潜在有用的大类样本信息,造成大类识别率降低。

算法自适应方法包括分类器集成[8]和代价敏感学习[9]等。Zhou 等[10]提出了代价敏感神经网络与分类器集成相结合的方法,实验表明,分类器集成对二分类不平衡问题和多分类不平衡问题同样有效。文献[11]提出了 AdaBoost. MLR 算法解决多类别分类问题,对识别率较低的类别给予较高的错分代价,提高"难分"类别的识别精度,但提升幅度有限。AdaCost[12]算法在 AdaBoost 算法[13]的权值更新规则中引入错分代价因子,提高了小类样本的查全率和查准率。这些方法在模型训练过程中引入了错分代价因子,强迫分类器"关注"错分代价较高的小类样本,能在确保大类样本识别率的前提下,提升小类样本的识别率,但由于类别不平衡性和模型自身的原因,对小类识别率的提升是有限的。

多分类 SVM 算法[14]采用一对一、一对其余的方法将二分类 SVM 扩展到多类别分类问题中,在不平衡数据分类中表现出很好的分类性能。文献[15]采用主动学习方法构造平衡的训练集,并提出了一种基于 SVM 的主动学习样本选择策略,实验表明,主动学习方法能用较少的样本获得较高的分类性能,但是主动学习需要迭代多次选择最有价值的样本,进行多次模型训练,而 SVM 的非线性模型优化过程对计算和存储要求太高。AdaC2. M1 算法[16]针对多类别不平衡分类问题,提出了基于代价敏感的 AdaBoost 集成学习方法,采用遗传算法搜索各个类别的错分代价,实验表明代价敏感方法很难较好地适用于多类别不平衡分类问题。

研究分析表明,单一地使用重采样方法改变训练集样本分布,虽然能提升小类样本的识别率,但也会大幅度降低大类样本的识别率;单一地使用代价敏感方法虽然保证了大类样本识别率不会降低,但对小类样本的识别率提升是有限的,因此,本文采用主动学习方法,选择最有潜在价值的样本,充分利用稀有的小类样本,降低数据集的不平衡性,并结合代价敏感方法,在多分类 AdaBoost 算法弱分类器的迭代训练中,对小类样本给予较高的错分代价,对大类样本给予较低的错分代价,动态调整样本权值更新速度,实现主动学习方法、代价敏感方法和多分类 AdaBoost 方法的融合,在保证大类样本识别率不会降低的前提下,大幅度提高小类样本的识别率。

1　传统多分类 AdaBoost 算法

AdaBoost 算法是目前应用最广泛的机器学习方法之一,基本思想是将若干个弱分类器按照某种规则组合起来,集成为一个分类能力很强的强分类器,最初应用于二分类问题,多类别分类问题是二分类问题的扩展。假设训练样本集 $X = \{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$,其中 $y_i \in \{1,2,\cdots,K\}$。对于样本 x_i,若 $l = y_i$,则 $Y_i(l) = 1$,否则 $Y_i(l) = -1$。集成学习算法通常指通过某种方式得到 T 个弱分类器 $h_t(x): X \times Y \to \mathbf{R}$,弱分类器权重 α_t,然后进行组合得到强分类器,即

$$f(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{1}$$

多分类 AdaBoost 算法如下。

输入:训练样本集 $X = \{(x_1,Y_1),\cdots,(x_m,Y_m)\}$,样本权重 D,弱分类器 $h(x): X \times Y \to \mathbf{R}$,迭代次数 T。

初始化:$D_1(i,l) = 1/(mK)$,其中 $i = 1,2,\cdots,m, l = 1,2,\cdots,K$。

For $t = 1,2,\cdots,T$:

1) 根据样本分布 D_t,训练弱分类器 $h_t(x): X \times Y \to \mathbf{R}$;

2) 计算弱分类器权重 α_t;

3) 更新权重 $D_{t+1}(i,l) = [D_t(i,l)\exp(-\alpha_t Y_i(l) h_t(x_i,l))]/Z_t$,其中 $Z_t = \sum_{i,l} D_t(i,l)\exp(-\alpha_t Y_i(l) h_t(x_i,l))$;

输出:强分类器 $f(x,l) = \mathrm{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x,l)\right)$。

AdaBoost 算法中 T 个不同的弱分类器是通过改变数据分布来实现的,样本权重更新是根据弱分类器对样本的分类情况来确定的,具体来说,如果弱分类器对某个样本的某个标签分类正确,则对应的权重减少,如果分类错误,则对应权重增加,权重减少与增加的速度只与弱分类器有关,每次权重改变的速度是一样的。

2　主动学习不平衡多分类方法

在传统的主动学习任务中,往往选择对分类器最有价值的样本加入训练集参与训练,以更新分类器,但在不平衡多分类问题中,如果仍采用传统的样本选择方法,可能会导致训练集中大类的样本一直更新,而小类的样本一直得不到更新,即分类器的更新存在不平衡性。针对这个问题,本文提出一种新的基于不确定性动态间隔的样本选择策略,从原始训练集中挑选那些更有意义的样本,选择数量最小但信息量最大的子集作为最终训练集,降低类别之间数据的不平衡性。

2.1　主动学习多分类 AdaBoost 算法

设有标注样本集为 $X = \{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$,首先抽取部分最有价值的样本作为初始训练集,且类别之间样本均衡。设 L^k、U^k 分别为第 k 次学习时的训练集和非训练集,满足 $X = L^k \cup U^k$。用多分类 AdaBoost 算法对 L^k 样本集进行训练,得到强分类器,并对样本集 U^k 进行预测,按照某种样本选择策略选择最有价值的样本加入到 L^{k+1} 中,重复上述过程直到满足停止条件。

2.1.1　样本选择策略

本文采用基于 Margin 策略的不确定性来选择待标注的样本,如式(2)所示:

$$x^* = \arg\min_x(\beta_x(f(x,l_1) - f(x,l_2))) \quad (2)$$

其中:l_1 和 l_2 分别是最具有最大和第二大值的置信度输出值,即当前分类模型最确定的两个类别,二者的差值越小说明模型对样本的不确定性越大,则对样本进行标注获得的信息量越多;β_x 为数据平衡控制因子,目的是保证类别之间的数据平衡性。

2.1.2 基于主动学习的训练集选择方法

输入:有标注样本集 $X = \{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$,其中 $y_i \in \{1,2,\cdots,K\}$,初始训练集 L^1,非训练集 $U^1 = X - L^1$。

For $k = 1,2,\cdots,iter$

1) 在训练集 L^k 上训练多分类 AdaBoost 分类器 f;

2) 统计训练集 L^k 中各类别包含样本个数,记包含样本数最少的类别为 c_1,包含样本数最大的类别为 c_2;

3) 用分类器 f 对非训练集 U^k 中样本预测,如果分类模型满足停止条件,循环终止;

4) 如果 $c_2/c_1 > thresh$,且非训练集 U^k 中样本 x 对应的类别为 c_1,则令 $\beta_x = \varepsilon$;否则 $\beta_x = 1$;对每个样本计算 $\beta_x(f(x,l_1) - f(x,l_2))$,$l_1$ 和 l_2 分别是最具有最大和第二大值的置信度输出值,选择最小的 N 个样本,记为 S;

5) 更新 $L^{k+1} = L^k \bigcup S, U^{k+1} = U^k \backslash S$;

End

输出:训练集 L。

2.2 基于代价敏感的不平衡多分类 AdaBoost 算法

分类算法总是希望平均错分代价最小,即希望式(3)最小:

$$\varepsilon_c = \sum_{i=1}^{m}\left(\sum_{l=1}^{K} cost(i,l)\delta(H(x_i) = l)\right) \quad (3)$$

对于 $\delta(\pi)$ 函数,当 π 为真时,$\delta(\pi)$ 为 1;否则为 0。在多分类 AdaBoost 算法中引入动态代价调整函数,可以得到代价敏感多分类 AdaBoost 算法。

2.2.1 改进算法流程

输入:训练样本集 $L = \{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,样本权重 D,代价调整函数 $cost$,弱分类器 $h(x): X \times Y \rightarrow \mathbf{R}$,迭代次数 T,错分代价因子 c。

初始化:$D_1(i,l) = 1/(nK)$,$cost_1(i,l) = 1$,其中 $i = 1,2,\cdots,n$,$l = 1,2,\cdots,K$;

For $t = 1,2,\cdots,T$

1) 根据样本分布 D_t,训练弱分类器 $h_t(x): X \times Y \rightarrow \mathbf{R}$;

2) 计算权值调整函数 $cost_t$:若训练集 L 为平衡数据集,则对样本 x_i,类别 l,令 $cost_t(i,l) = 1$;若训练集 L 为不平衡数据集,对样本 x_i,类别 l,若 x_i 为小类样本,且 $l = y_i$,则令 $cost_t(i,l) = c > 1$,否则 $cost_t(i,l) = 1$;

3) 计算弱分类器权重 α_t;

4) 对于样本 x_i,若 $l = y_i$,则 $Y_i(l) = 1$,否则 $Y_i(l) = -1$;

5) 更新权重:$D_{t+1}(i,l) = [D_t(i,l)\exp(-\alpha_t cost_t(i,l)Y_i(l)h_t(x_i,l))]/Z_t$,其中 $Z_t = \sum_{i,l} D_t(i,l)\exp(-\alpha_t cost_t(i,l)Y_i(l)h_t(x_i,l))$;

End

输出:强分类器 $f(x) = \arg\max_l\left\{\sum_{t=1}^{T}\alpha_t h_t(x,l)\right\}$。

2.2.2 如何选择 α_t

由于 $Z_t = \sum_{i,l} D_t(i,l)\exp(-\alpha_t u_t(x_i,l)) \leqslant \frac{1+r_t}{2}e^{-\alpha_t} + \frac{1-r_t}{2}e^{\alpha_t}$,其中 $u_t(x_i,l) = -cost_t(i,l)Y_i(l)h_t(x_i,l)$,$r_t = \sum_{i,l} D_t(i,l)u_t(x_i,l)$。根据文献[13]中 α_t 的选择方法,可令 $\alpha_t = \frac{1}{2}\ln\left(\frac{1+r_t}{1-r_t}\right)$,此时 $Z_t \leqslant \sqrt{1-r_t^2} \leqslant 1$。所以在选择弱分类器时,可以选择能使得 $r_t = \sum_i D_t(i,l)u_t(x_i)$ 最大的若分离器 h_t,随着弱分类器个数的增加,训练误差会逐渐趋向于 0。

3 实验与分析

3.1 评价指标

实验采用 G-mean 和查准率作为性能衡量标准,令 n_i 表示属于类别 C_i 的样本总数,$cm(i,j)$ 表示类别为 C_i 的样本被判断为类别 C_j 的个数,则类别 C_i 的查准率可定义为 $R_i = cm(i,i)/n_i$,G-mean 定义为 $G\text{-}mean = \left(\prod_{i=1}^{K} R_i\right)^{\frac{1}{K}}$。

3.2 实验数据集

本文实验使用数据包括 TTE 测量数据集和 4 个 UCI(University of California Irvine)数据集,其中 TTE 数据集来源于华西医院,34 个属性,共有 2 214 个心脏疾病病例,包括感染性心内膜炎(Infective Endocarditis, IE)58 例,冠心病(Coronary Artery Disease, CAD)169 例,先心病(Congenital Heart Disease, CHD)733 例,瓣膜病(Valvular Heart Disease, VHD)1 177 例,每个病例只患有一种疾病,最大不平衡度为 20.3。详细的数据集信息如表 1 所示。

表 1 数据集信息

Tab. 1 Information of data sets

数据集	类别	样本数	属性数	最大不平衡度
TTE	4	2214(58/169/733/1177)	34	20.3
abalone	5	2014(67/203/487/568/689)	22	10.3
balance	3	625(49/288/288)	4	5.9
ecoli	5	327(20/35/52/77/143)	7	7.2
segment	7	2214(10/20/60/120/240/280/330)	34	33.0

3.3 在 TTE 数据集上的实验结果与分析

3.3.1 基于主动学习的训练集选择

实验过程中,训练集和测试集比例按照 6∶4 划分,然后在训练集上采用基于不确定性动态间隔的样本选择策略选择新的训练集。

初始训练集 L^1 的选择:在全部训练集上训练多分类 AdaBoost 分类器 f,然后调用分类器 f 对训练集中全部样本进行预测,对每个样本计算 $f(x,l_1) - f(x,l_2)$,其中 l_1 和 l_2 分别是最具有最大和第二大值的置信度输出值,对每个类别选择 $f(x,l_1) - f(x,l_2)$ 值最小的 10 个样本,共得到 40 个样本作为初始训练集;

训练集选择过程:按照 2.1.2 节的训练集选择方法,结合数据集实际情况,每次迭代选择 4 个最有价值的样本加入训练集,令 $\varepsilon = 0.005$,控制训练集平衡度,一定程度上保证每次迭代都能选择到小类样本加入训练集。训练集选择前后及测试集样本个数如表 2 所示。

3.3.2 错分代价因子 c 的调节

本文采用 2.2.1 节中描述的改进算法,选择每个类别的查准率作为评价指标来验证改进算法对各个类别的性能影

响。在对参数 c 的最优选择实验时,错分代价调整因子 c 值在[1,30]区间内以步长 1 变化,找出较小的最合适的 c 值范围,然后在小范围内以 0.1 步长变化,寻找最合适的 c 值。以下仅列出 c = 22,23,24,25,26 时的实验结果,具体如表 3 所示,可知最可能的 c 在区间[23,24]内。

表 2　训练集选择结果(TTE)

Tab. 2　Selected results of training data sets on TTE

心脏病	原始训练集	新训练集	测试集
VHD	709	114	468
CHD	443	86	290
IE	36	36	22
CAD	95	68	74

表 3　不同参数 c 的识别率(TTE)

Tab. 3　Recognition rate of different parameter c for TTE

c 值	VHD	CHD	IE	CAD	总体
22	0.9438	0.8724	0.6702	0.7133	0.8611
23	0.9287	0.8623	0.7475	0.7326	0.8756
24	0.9362	0.8676	0.7585	0.7814	0.8804
25	0.9121	0.8831	0.7362	0.7114	0.8627
26	0.9392	0.8556	0.7045	0.6914	0.8532

3.3.3　与其他算法的对比实验结果

在对参数 c 的最优选择实验时,错分代价调整因子 c 值在[23,24]区间内以步长 0.1 变化,寻找最合适的参数 c。通过实验可知,当 c = 23.8 时,总体识别率最高可达 88.34%,VHD 识别率为 92.62%,CHD 识别率为 85.76%,IE 识别率为 76.85%,CAD 识别率为 80.14%。

在最优参数下,将本文算法与 SMOTEBoost、AdaBoost.MLR、多分类 SVM 和 ML-*KNN*(Multi-Label *K*-Nearest Neighbor)[17] 进行比较,每个类别的详细识别率和总体分类识别率如表 4 所示。

表 4　不同算法的识别率(TTE)

Tab. 4　Recognition rate of different algorithms under optimal parameters on TTE

算法	VHD	CHD	IE	CAD	总体
ML-*KNN*	0.9238	0.8724	0.3102	0.3333	0.8311
多分类 SVM	0.9188	0.8966	0.6819	0.4459	0.8631
SMOTEBoost	0.8339	0.8271	0.8316	0.8313	0.8325
AdaBoost.MLR	0.9195	0.8731	0.4682	0.4914	0.8524
本文算法(c=23.8)	0.9262	0.8576	0.7685	0.8014	0.8834

从表 4 可以看出,ML-*KNN* 算法和 AdaBoost.MLR 算法对 IE 和 CAD 的识别率很低,这是因为这两种算法的分类性能跟训练集有关;由于 SVM 模型只与少数支持向量有关,分类性能较好一些,但对 CAD 的识别率较低,对 IE 的识别率仅稍好于随机猜测;SMOTEBoost 算法虽然能提升小类样本识别率,但其他类别的样本识别率也会很大幅度地降低。本文算法相较于多分类 SVM,心脏病总体识别率提升了 5.9%,G-mean 指标提升了 18.2%,VHD 识别率提升了 0.8%,IE(小类)识别率提升了 12.7%,CAD(小类)识别率提升了 79.73%;相较于 SMOTEBoost,总体识别率提升了 6.11%,G-mean 指标提升了 0.64%,VHD 识别率提升了 11.07%,CHD 识别率提升了 3.69%。

3.4　在 UCI 数据集上的实验结果与分析

实验过程中,分别对 4 个 UCI 数据集按照 6:4 划分训练集和测试集,首先按照 2.1.2 节中的样本选择方法,产生新的训练集,然后在新的训练集上采取和 3.3.2 节类似的步骤,寻找最佳错分代价因子 c。对于每个数据集,选择前后训练集的样本个数和最佳参数 c 值如表 5 所示。

表 5　训练集选择结果和参数 c 设置(UCI)

Tab. 5　Training dataset selection results and optimal parameter c on UCI

数据集	原始训练集样本数	选择后训练集样本数	参数 c 值
abalone	1208	560	50.3
balance	375	210	15.6
ecoli	196	120	14.8
segment	636	560	24.6

只有当各个类别的查准率都很高时 G-mean 才会高,因此实验采用 G-mean 指标对本文算法和 ML-*KNN*、多分类 SVM、SMOTEBoost、AdaBoost.MLR 算法进行对比,各算法的 G-mean 值如表 6 所示。

表 6　各算法的 G-mean 值对比(UCI)

Tab. 6　G-mean value comparison of different algorithms on UCI

数据集	ML-*KNN*	多分类 SVM	SMOTE-Boost	MLR	本文算法
abalone	0.000	0.0000	0.2503	0.023	0.2687
balance	0.067	0.8909	0.5866	0.000	0.6756
ecoli	0.687	0.0000	0.7925	0.000	0.8204
segment	0.896	0.9431	0.9300	0.912	0.9215

从表 6 可以看出有些数据集的 G-mean 值为 0,这是由于小类样本的查准率为 0 造成的,这也说明小类样本的分类性能影响算法的整体性能。

本文算法在 TTE、abalone 和 ecoli 数据集上取得最高的 G-mean 值;相较于多分类 SVM 算法,TTE 数据集上的 G-mean 值提升了 18.2%,相较于 SMOTEBoost 算法,G-mean 值提升了 0.64%。

4　结语

本文针对多类别不平衡分类中小类样本识别率低问题,采用主动学习思想,选择少量的最有价值的样本作为训练集,并将不平衡分类问题转化为代价敏感分类问题。在多分类 AdaBoost 算法弱分类器的迭代训练时,对小类样本给予较高的错分代价,在可行的代价选择空间内,寻找能使得分类性能最优的错分代价调整因子,调整样本权重更新速度,对多分类 AdaBoost 算法进行改进。在心脏病 TTE 测量数据集上的实验结果表明,该方法对小类样本识别率有较大幅度的提升,还能保证其他类别的识别率不会大幅降低,综合提升了分类器的性能。综合 UCI 数据集上的实验结果表明,本文算法在 TTE、abalone 和 ecoli 数据集上的 G-mean 值最高,而且训练集只需要少量的有价值的样本,模型训练效率高、速度快、识别率高,性能更优。

参考文献(References)

[1]　CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16(1): 321 - 357.

[2] JAPKOWICZ N, STEPHEN S. The class imbalance problem: a systematic study [J]. Intelligent Data Analysis, 2002, 6(5): 429 – 449.

[3] CHEN X, GERLACH B, CASASENT D. Pruning support vectors for imbalanced data classification [C]// IJCNN 2005: Proceedings of the 2005 International Joint Conference on Neural Networks. Piscataway, NJ: IEEE, 2005, 3: 1883 – 1888.

[4] CHAN P K, STOLFO S J. Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection [C]// KDD 1998: Proceedings of the 1998 ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI Press, 1998: 164 – 168.

[5] LU B L, ITO M. Task decomposition and module combination based on class relations: a modular neural network for pattern classification [J]. IEEE Transactions on Neural Networks, 1999, 10(5): 1244 – 1256.

[6] LU B L, WANG K A, UTIYAMA M, et al. A part-versus-part method for massively parallel training of support vector machines [C]// IJCNN 2004: Proceedings of the 2004 International Joint Conference on Neural Networks. Piscataway, NJ: IEEE, 2004, 1: 735:740.

[7] CHAWLA N V, LAZAREVIC A, HALL L O, et al. SMOTEBoost: improving prediction of the minority class in boosting [C]// PKDD 2003: Proceedings of the 2003 European Conference on Principles of Data Mining and Knowledge Discovery. Berlin: Springer, 2003: 107 – 119.

[8] FU Z, WANG L, ZHANG D. An improved multi-label classification ensemble learning algorithm [C]// CCPR 2014: Proceedings of the 6th Chinese Conference on Pattern Recognition. Berlin: Springer, 2014: 243 – 252.

[9] 付忠良. 多分类问题代价敏感 AdaBoost 算法[J]. 自动化学报, 2011, 37(8): 973 – 983. (FU Z L. Cost-sensitive AdaBoost algorithm for multi-class classification problems [J]. Acta Automatica Sinica, 2011, 37(8): 973 – 983.)

[10] ZHOU Z H, LIU X Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem [J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(1): 63 – 77.

[11] 王莉莉, 付忠良. 基于标签相关性的多标签分类 AdaBoost 算法[J]. 四川大学学报(工程科学版), 2016, 48(5): 91 – 97. (WANG L L, FU Z L. Multi-label AdaBoost algorithm based on label correlations [J]. Journal of Sichuan University (Engineering Science Edition), 2016, 48(5): 91 – 97.)

[12] FAN W, STOLFO S J, ZHANG J, et al. AdaCost: misclassification cost-sensitive boosting [C]// ICML 1999: Proceedings of the 1999 International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann, 1999: 97 – 105.

[13] SCHAPIRE R E, SINGER Y. Improved boosting algorithms using confidence-rated predictions [J]. Machine Learning, 1999, 37(3): 297 – 336.

[14] WU T F, LIN C J, WENG R C. Probability estimates for multi-class classification by pairwise coupling [J]. Journal of Machine Learning Research, 2004, 5: 975 – 1005.

[15] ERTEKIN S, HUANG J, GILES C L. Active learning for class imbalance problem [C]// SIGIR 2007: Proceedings of the 2007 International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2007: 823 – 824.

[16] SUN Y, KAMEL M S, WANG Y. Boosting for learning multiple classes with imbalanced class distribution [C]// ICDM 2006: Proceedings of the 2006 International Conference on Data Mining. Washington, DC: IEEE Computer Society, 2006: 592 – 602.

[17] ZHANG M L, ZHOU Z H. ML-KNN: a lazy learning approach to multi-label learning [J]. Pattern Recognition, 2007, 40(7): 2038 – 2048.

This work is partially supported by the Sichuan Science and Technology Support Project (2016JZ0035), West Light Foundation of Chinese Academy of Sciences.

WANG Lili, born in 1987, Ph. D. candidate. Her research interests include machine learning, pattern recognition, data mining.

FU Zhongliang, born in 1967, M. S., professor. His research interests include machine learning, pattern recognition.

TAO Pan, born in 1988, Ph. D. candidate. His research interests include machine learning, data mining.

HU Xin, born in 1987, M. S. candidate. His research interests include data warehouse, data mining.

(上接第 1976 页)

[22] WHITLEY D. An overview of evolutionary algorithms: practical issues and common pitfalls [J]. Information and Software Technology, 2001, 43(14): 817 – 831.

[23] GOLDBERG D E. The Design of Innovation: Lessons from and for Competent Genetic Algorithms [M]. Rotterdam: Springer Science & Business Media, 2013: 59 – 93.

[24] LU G. Characterising fitness landscapes with fitness-probability cloud and its applications to algorithm configuration [D]. Birmingham: University of Birmingham, 2014: 133 – 149.

[25] PAPADIMITRIOU C H, STEIGLITZ K. Combinatorial Optimization: Algorithms and Complexity [M]. Massachusetts: Courier Corporation, 2013: 156 – 182.

[26] MERZ P. Advanced fitness landscape analysis and the performance of memetic algorithms [J]. Evolutionary Computation, 2004, 12(3): 303 – 325.

[27] LU G, LI J, YAO X. Fitness landscapes and problem difficulty in evolutionary algorithms: from theory to applications [M]// Recent Advances in the Theory and Application of Fitness Landscapes. Berlin: Springer, 2014: 133 – 152.

[28] VANNESCHI L, CLERGUE M, COLLARD P, et al. Fitness clouds and problem hardness in genetic programming [C]// GECCO 2004: Proceedings of the 2004 Genetic and Evolutionary Computation Conference, LNCS 3103. Berlin: Springer, 2004: 690 – 701.

This work is partially supported by National Natural Science Foundation of China (61105062).

ZHU Chunmei, born in 1981, Ph. D. candidate. Her research interests include intelligent control, pattern recognition.

MO Hongqiang, born in 1976, Ph. D., professor. His research interests include evolutionary computation.