ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 4,Jul. 2021,pp. 684－696
DOI:10. 16337/j. 1004-9037. 2021. 04. 006

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

集成学习机制下的鼻炎辅助诊断模型

杨晶东[1]，孟一飞[1]，荀镕基[1]，余少卿[2]

(1. 上海理工大学光电信息与计算机工程学院,上海 200093;2. 同济大学附属同济医院耳鼻咽喉头颈外科，上海 200065)

摘　要：鼻炎(Rhinitis)是上呼吸道常见的慢性炎症,具有多种证型和体征。鼻炎临床分类具有样本类型多、类别不平衡特征,属于多输出分类范畴,常出现少数类样本识别率低、综合分类精度差的问题。为此,本文提出异质集成结构分类算法,将鼻炎多输出分类转化为多标签和多类别分类,采用集成学习算法构建异质集成分类器。该方法可根据子数据集中单一类标的不平衡度,自动调节集成森林基学习器数量和深度,有效减少不均衡样本对分类的影响,提高多数类和少数类的总体分类精度,进而提升集成模型的泛化能力。针对临床461例鼻炎样本进行交叉验证分类实验,本文分类模型灵敏度为74.9%,特异性为86.5%,准确度为92.0%,F1为0.783,AUC为0.953。与6种典型模型相比,本文模型具有更好的评估性能,更适合于鼻炎的早期临床诊断。

关键词：变应性鼻炎;集成学习;基学习器;多标签分类;异质结构

中图分类号：TP181　　**文献标志码**：A

Computer-Aided Diagnosis of Rhinitis's Disease Based on Ensemble Learning

YANG Jingdong[1], MENG Yifei[1], XUN Rongji[1], YU Shaoqing[2]

(1. School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; 2. Department of Otorhinolaryngology, Head and Neck Surgery, Tongji Hospital of Tongji University, Shanghai 200065, China)

Abstract: Rhinitis is a common chronic inflammation of the upper respiratory tract with a variety of symptoms and signs. The clinical classification of rhinitis is characterized by different types of instances and class imbalance, and belongs to multiple output classification. Low recognition rate and poor generalization performance often occur for minority class instances. Therefore, this article proposes a novel classification model based on heterogeneous integrated frame, which translates the multi-output classification of rhinitis to multi-label and multi-class classification, then builds a heterogeneous integrated classifier by ensemble learning algorithm. The proposed model can automatically adjust the number and depth of integrated forest learners according to the imbalance ratio of single class label in a subset. As a result, it can effectively reduce influence of class imbalance and improve classification performance of majority and minority class concurrently, further to enhance generalization of integrated classifiers. We conduct cross-validation classification experiments on 461 cases of clinical rhinitis. The outcomes show that the evaluation indicators of the proposed model, such as sensitivity, specificity, accuracy, F1 and AUC, are 74.9%, 86.5%,

基金项目:国家自然科学基金(81973749,8187040043)资助项目;上海市卫生健康委先进适宜技术推广(2019SY071)资助项目;上海市科委中医引导类(18401903600)资助项目;上海市卫计委科研面上(201740093)资助项目。
收稿日期:2020-08-29;**修订日期**:2020-11-26

92.0%, 0.783 and 0.953, respectively. In comparison to other baseline methods, it achieves better evaluation performance and is more suitable for the early clinical diagnosis of rhinitis.

Key words: allergic rhinitis; ensemble learning; base learner; multi-label classification; heterogeneous structure

引　言

鼻炎(Rhinitis)是普遍存在的一种呼吸系统疾病,严重影响患者的正常工作和生活。据统计,全球有20%～30% 的普通人被过敏症状困扰,2015年全球哮喘患者已达3亿人,变应性鼻炎(Allergic rhinitis, AR)患者达5亿人。在患病初期,咽喉有明显干痒感,逐渐变为烧灼与刺痛感。若未及时干预,初期轻症急性鼻炎会转化为慢性重症鼻炎,不仅治疗周期长,治疗效果也难以保证。因此,AR初期诊断和预防对于后期的有效治疗和控制具有重要意义。

近年来许多学者将机器学习算法应用于鼻炎诊断[1],辅助医生提高鼻炎诊断效率。Demirjian等[2]采用改进贝叶斯理论对AR发生概率预测,证明了免疫球蛋白(IgE)和嗜酸红细胞(Eosinophil)是一种重要影响因素。李少华等[3]采用层次聚类分析,得出5种鼻炎常见病症与其特征相关性。黄嘉韵[4]提出了建立CART决策树和使用关联规则算法对鼻炎5种症型建立了辅助诊断模型,并发现了一种对症治疗规律,具有较好的准确率和可解释性。

文献[5]采用遗传编程算法(Genetic programming,GP)对多种医疗病症自主分类,通过将原始数据输送进GP模型,并自主选择训练特征,分类准确率超越了决策树和贝叶斯方法。但GP模型训练过程不具可解释性,不能直接应用于风险较高的诊断和临床研究。Liu等[6]提出了采用自动超参数优化(AutoHPO)深度神经网络模型(DNN)解决医疗数据类别不平衡问题,总体精度上优于随机森林(RF)和AdaBoost算法。机器学习算法应用于AR诊断过程,虽然取得了较好的效果,但仍存在方法局限性,如样本数量不均衡[7]、数据属性缺失、维度过高等问题[8]。

样本不均衡问题常出现于临床样本诊断过程。模型分类更偏向多数类样本,导致多数类分类精度过高,少数类精度过低。而临床医学中往往更关注少数类的分类精度,如罕见病的漏诊率或误诊率。解决样本不均衡问题通常包括数据采样、算法适应[9]和特征选择[10]等方法。过采样方法包括随机过采样和启发式过采样[11],采用增加少数类样本,与多数类平衡,但容易增加无效样本。欠采样方法包括随机欠采样和基于最优子集搜索欠采样[12],通过减少多数类样本,与少数类样本平衡,但容易丢失样本重要特征。混合采样方法采用先合成样本、再剔除噪声,综合考虑欠采样和过采样方法特点。代价敏感学习采用向损失函数引入代价敏感学习因子,判断少数类与多数类错分代价。当多数类被错分时,损失函数增加;少数类被错分时,损失函数减少,从而提高少数类样本的分类精度。还有一些学者引入集成学习方法解决样本不均衡问题。如在过采样或者代价敏感学习方法中融入集成学习的基分类器,如代价敏感学习boosting方法AdaCost[13],或过采样bagging方法SMOTE [14]。此外在不均衡样本集上做特征选择也能有助于提升模型分类能力。通过选择去除掉冗余特征,保留典型的特征子集。Ksiazek等[15]在不均衡肝细胞癌诊断中采用了遗传算法(GA)实现了特征筛选和模型参数优化,具有较好的分类精度。综上所述,本文构建一种异质集成分类模型实现AR多输出分类。本文主要贡献如下:

(1) 提出一种异质集成分类模型,实现不同证型鼻炎样本的多输出分类,提高了少数类样本的分类精度。

(2) 根据鼻炎样本分布,设计一种不平衡度计算方法,增强样本均衡化,降低类别不平衡对分类的

影响。

(3) 提出一种自适应超参数优化方法,动态搜索集成RFs数量和深度,提高最优超参数搜索效率。

1 分类模型框架

1.1 基于包外估计的多类别分类

Easy ensemble (EE)[16]是一种对不均衡样本实现均衡分类算法,将欠采样技术和集成学习相融合,通过多次随机采样,充分利用单次欠采样外的遗漏数据,使训练数据集均衡化。本文变应性鼻炎病症有分度、分型两类输出,属于多类别分类问题。采用基于包外估计(Out-of-bag, OOB)EE集成分类算法OOBEE,将全部样本作为训练数据,采用Extra-tree(ET)模型作为基分类器,对所有训练数据均衡化处理,实现对不平衡小样本预测。OOBEE算法流程图如图1所示。OOBEE从多数类中抽取与少数类相等的样本,并组合重复使用的少数类样本构建多组基分类器,通过加权投票方法获得集成分类器,以减少样本不均衡对分类的影响。

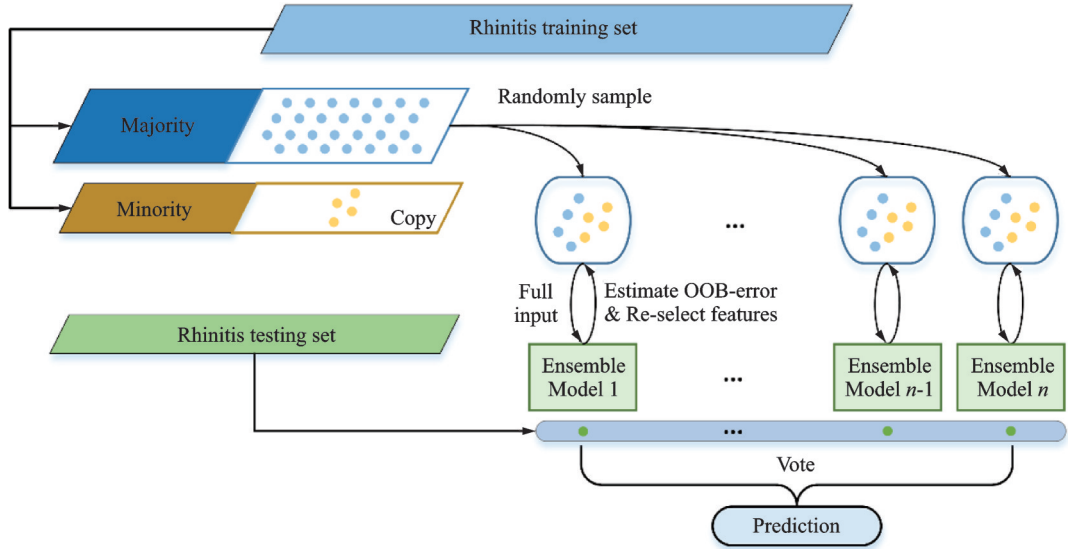

图1 集成学习OOBEE算法流程图

Fig.1 Flow chart of integrated learning model of OOBEE

该方法数学描述为:假设训练样本集$S_r=\{(x,y)\}$,做T次欠采样,采用Bootstrap随机采样法从多数类样本集S_r^+中得到一个子集S_{rk}^+,并且S_{rk}^+数量和少数类样本S_r^-相同,使用ET算法对$S_{rk}^+ \cup S_r^-$训练多组个体模型。

$$N_k=\operatorname{sgn}\left\{\sum_{j=1}^{n_k}\alpha_{k,j}h_{k,j}(x)-\theta_k\right\} \tag{1}$$

式中:$h_{k,j}(x)$为第j个ET子分类器,$\alpha_{k,j}$为$h_{k,j}(x)$权重,θ_k为子训练集的实际类别。ET中随机分裂特征数为m,由全体训练样本的计算得出。m依据特征重要程度(Variable importance, VI)选取

$$\mathrm{VI}=\frac{\sum\mathrm{Error}_{\mathrm{oob}_1}-\sum\mathrm{Error}_{\mathrm{oob}_2}}{n\mathrm{Trees}} \tag{2}$$

式中:oob_1代表所有鼻炎测试样本,oob_2代表加入噪声的测试样本。最终模型描述为

$$N(x)=\operatorname{sgn}\left\{\sum_{k=1}^{T}\sum_{j=1}^{n_k}\alpha_{k,j}h_{k,j}(x,m)-\sum_{k=1}^{T}\theta_k\right\} \tag{3}$$

由于AdaBoost[17-18]对小样本噪声敏感,难以获得最优解,因此不适合分析可能存在的少量错误样本。RF算法仅采用全部样本的36.78%作为包外估计,损失了部分训练数据,ET算法是随机选择最佳分叉属性和特征分裂数,将全部样本作为训练或包外估计OOB。该方法采用ET算法作为基分类器,使集成分类器方差更小,在小样本分类中具有更好的泛化能力,同时有利于提升鼻炎样本分度和分型的准确率。因此,本文采用ET算法作为多类别分类的基分类器。

1.2 基于动态加权RF多标签分类

常见鼻炎样本包括变应性鼻炎(AR)、鼻窦炎(RS)、上呼吸道感染(URI)和其他(OTH含鼻息肉,鼻腔肿瘤等),鼻炎预测属于多标签分类。常采用样本拆分法,选择RF作为基分类器[19],将多分类转化为多个单标签二分类。通过调整RF深度、分裂特征数,减少模型过拟合和降低特征维度。但标准RF算法基分类器参数需要人为设定。本文提出了一种自适应集成森林ARF算法,根据样本不平衡度,动态调整RF数量和深度,提高多标签分类精度和效率。传统的不均衡度(Ib)是少数类样本数量N_l与多数类样本数量N_m的比值。

$$Ib=\frac{N_l}{N_m} \tag{4}$$

该比值越接近0说明样本越不均衡,越趋近于1说明样本越均衡。但该方法无法直接应用于多类别分类,因此,本文针对鼻炎样本不均衡特性,设计不均衡度计算公式,假设全体样本个数为n,每个类标中不同种类出现频率为f_j,类别个数为c_i,每一类不均衡度b_i计算公式为

$$b_i=\frac{1}{n}\sum_{j=1}^{n}\sqrt{\frac{\min\left\{\left|f_j-\frac{n}{c_i}\right|,\frac{n}{c_i}\right\}}{\max\left\{\left|f_j-\frac{n}{c_i}\right|,\frac{n}{c_i}\right\}}} \tag{5}$$

式中:$\frac{n}{c_i}$为样本均衡时各类样本数,$\left|f_j-\frac{n}{c_i}\right|$为真实样本与均衡样本数量差。当4种鼻炎证型,类别数$c_i=2$;如包括症状程度和类型$c_i\geqslant 2$;如果不平衡度越趋近于0,说明样本数据越均衡,反之亦然。

图2描述了当样本数据为100时,二分类时不均衡度能力曲线。可见本文方法与改进的经典方法[20]不平衡度接近,但当样本类别数量趋于平衡时,本文方法比经典方法收敛更快,说明本文的不均衡度对不均衡样本更加敏感。图3给出了三分类时不平衡度评价能力等高线,样本总数为100,横轴和纵

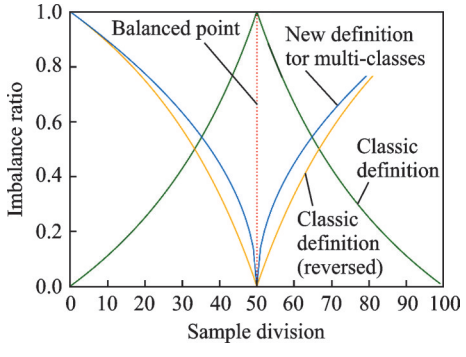

图2 不均衡度在二类别样本中的比较

Fig.2 Comparison of class imbalance ratio for binary classes

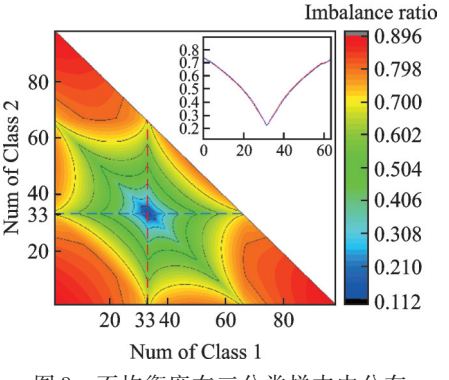

图3 不均衡度在三分类样本中分布

Fig.3 Distribution of class imbalance ratio for three classes

轴分别代表两个分类样本数量,第三分类样本数量由总数与前两类之差表示。可观察到(33,33)点不均衡度最低,而越向外围发散,不均衡度越高,等高线内部变化率也高于外围,说明本文不均衡度计算方法对于多分类的不均衡样本的敏感度较高。

本文采用自适应超参数优化ARF方法,动态调整RFs参数,其中基准参数$s(e, d)$为固定值,基分类器数量e和深度d均需搜索确定。本文通过动态网格搜索法获得多类别均衡化过程参数。在网格搜索过程中,RFs算法阈值范围为$e=[10,300]$; $d=[1,15]$。图4给出了ARF算法参数与精度动态关系图。分析可知,基分类器深度对精度影响较大,$d=12$时分类精度最好,基分类器数量对分类精度影响较小,但$e=[10,50]$过程中分类精度出现了一个较明显提升,说明当$e<60$,对分类精度影响较大。经多次测试后,参数设定为$e=70$,$d=12$。因此,ARF算法可以有效调节内嵌集的基分类器数量与训练时间的均衡,动态获得集成分类器最优精度时的基分类器匹配参数。

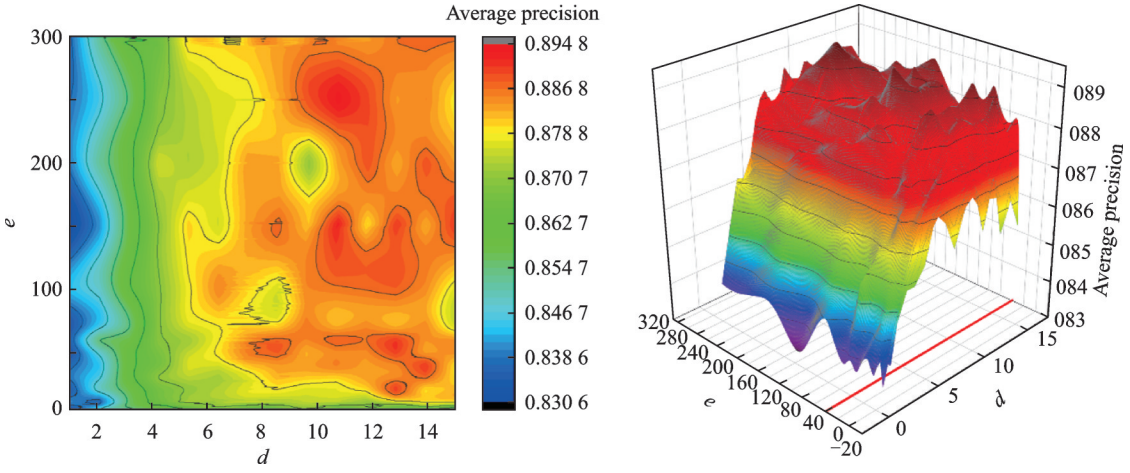

图4 ARF模型参数与精度动态关系图

Fig.4 Dynamic relationship diagram between model parameters and accuracy

ARF模型采用RF作为基学习器,采用等权重随机采样法生成训练集,每个基学习模型以等权重投票方式分类。假如一个模型测试集为X,类别数为c,基分类器数为m,则模型输出可表示为

$$Y(X)=\underset{y=1,2,\cdots,c}{\operatorname{argmax}}\left\{\sum_{i=1}^{m}I\left(f\left(X;L_i,g(s,b_i)\right)=y\right)\right\} \tag{6}$$

式中:f为指示函数,L为随机参数,g为基分类器RFs动态搜索函数,函数I为真则输出1,若为假则输出0。

鼻炎证型有4种标签分类,分别为变应性鼻炎(AR)、鼻窦炎(RS)、上呼吸道感染(URI)和其他(OTH含鼻息肉,鼻腔肿瘤等)。鼻炎样本的多标签分类ARF模型如图5所示,图中总样本集分为4组证型子集,4种分型鼻炎样本根据CIR(Calculation of imbalance ratio)值确定二分类的样本子集BS(Balanced sets)分布,分别输入到4组RFs鼻炎证型分类模型中。模型每次运行会输出预测结果与RFs包外误差,当包外误差满足优化终止条件时,可输出当前预测结果。

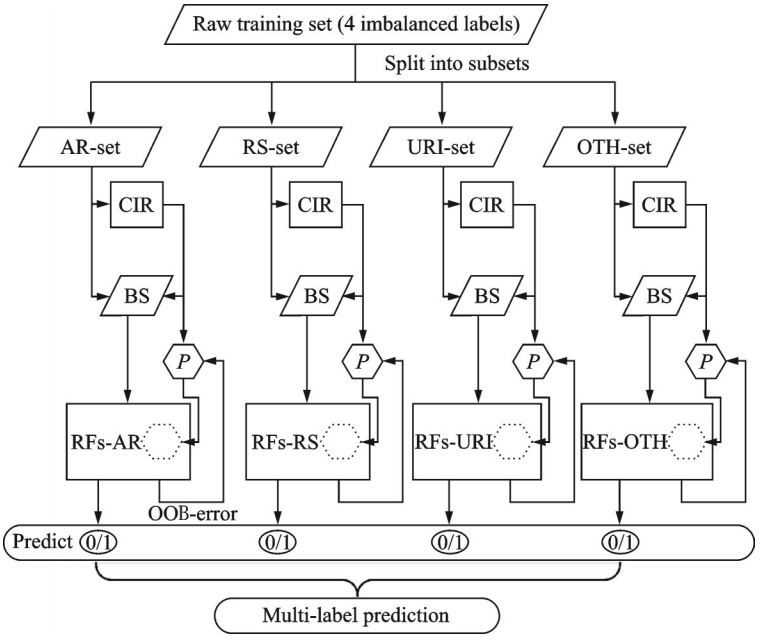

图5 ARF算法流程图

Fig.5 Flow chart of ARF model

1.3 异质集成结构的多输出分类模型

多输出分类是指从一个输入产生多个离散输出的分类模型,马忠臣等[21]总结了多输出分类类型,包括多标签分类、多输出有序分类、异质多输出分类(Heterogeneous multi-output, HGMO)。鼻炎样本包含4组常见的多标签鼻炎类型,每组又分度、分型。因此,鼻炎样本属于HGMO分类,其数学描述为

假设分类问题的输出空间包含 $m(\geqslant 2)$ 维多输出变量 $Y_1,\cdots,Y_m$,分类目标是寻求目标函数 h,使其准确学习每个输入 x 在 m 维输出变量上的相应输出 $y=(y_1,\cdots,y_m)$

$$h:\Omega_X \rightarrow \Omega_{Y_1}\times\Omega_{Y2}\times\cdots\times\Omega_{Y_m} \tag{7}$$

$$x\mapsto(y_1,\cdots,y_m) \tag{8}$$

式中:输出变量 $Y_1,\cdots,Y_m$ 具有不同类型,$y_j\in\Omega_{Y_j}$,$|\Omega_{Y_j}|\geqslant 2$;$\Omega_X$ 和 $\Omega_{Y_j}(j=1,\cdots,m)$ 分别表示输入和输出变量所属值域。

根据HGMO结构,本文提出了异质集成鼻炎分类器模型ARF-OOBEE识别多种证型鼻炎,如鼻窦炎(RS)(二元变量),变应性鼻炎(AR)严重程度或持续性(有序变量)等。图6描述了ARF-OOBEE模型示意图。该模型通过将HGMO问题转换成4标签二分类问题(Multi-label cassification)和2个多类别分类问题(Multi-class classification)。这样可有效避免多标签类型分类与多类别症状分类相互干扰,避免一个

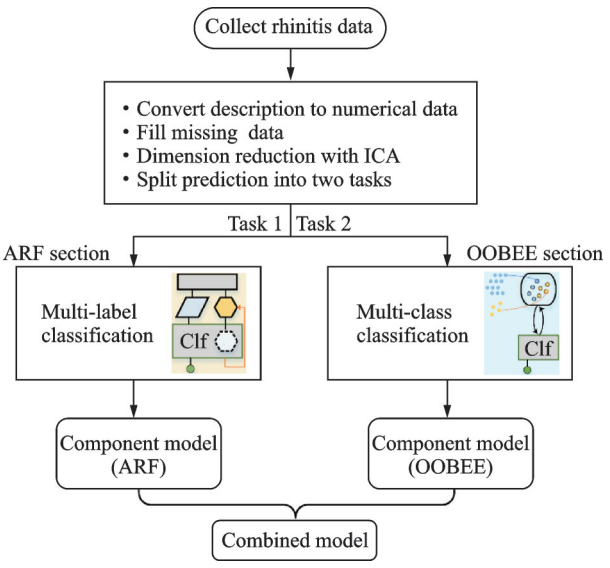

图6 ARF-OOBEE模型结构框图

Fig.6 Structure block diagram of ARF-OOBEE model

患者同时出现两组或更多的分度或分型标签。采用多种模型分别训练组件分类器,利用集成学习获得最终分类器。

图6左分支描述了动态随机森林方法ARF。ARF根据子数据集中单一类标的不平衡度,自动调节集成森林的群数和森林内的基分类器数。当出现均衡子标签时,减少森林群数,计算速度最优的单森林内的基分类器数量;当出现非均衡的子标签时,增加森林群数;最终根据验证集算出集成森林权重。ARF更有利于提高分类的准确率和均衡性,并通过动态增减训练集不均衡样本数量,实现速度与精度的动态均衡。图6右分支描述了OOBEE集成分类算法。该方法采用ET算法替代Adaboost,将全部样本作为包外估计,充分利用所有训练样本,通过欠采样集成学习方式处理多分类任务中的样本不均衡问题,避免了对不均衡样本的重复判断、少数类样本特征过于稀疏等问题,提高模型的泛化能力。

2 实验结果与分析

2.1 数据预处理

采用上海同济大学附属同济医院临床鼻炎样本461例,其中男性261例(占56.62%),平均年龄(30.48±19.66)岁;女性200例(占43.38%),平均年龄(33.51±19.32)岁。样本含有多种数据类型,包括患者信息(性别,年龄等),医生问诊结果(是否流涕,何种变应原等),检测仪器信息(CT, IgE等)。由于输入数据源种类多,数据类型不唯一,如果采用简单的剔除缺失值样本会使样本大量减少,不利于鼻炎病症预测。本文采用了混合型缺省值填充方法,对于患者个人信息采用K近邻[22]填充相似数据均值;使用了众数插补方式填补问诊数据缺失值;对于仪器测量数据,将缺失值作为一种标签,建立RF模型,得到预测值之后进行填充。

鼻炎诊断可设定为6组类别,包括4种病症标签,AR类型2种症状类别(分度或分型),数据呈不均匀分布。因此,本文采用独热编码。表1和表2分别描述了鼻炎标签分布及不平衡度数据和鼻炎类型分布,AR标签中阳性占比较大,RS、URI、OTH阴性占比较大,Severity类型中轻症样本较多,中症次之,重症最少,Duration类型中间歇性样本较多,持续性较少,可见鼻炎AR样本最多,非AR为30例,仅占总病例6.5%,却包含3种病症标签,说明鼻炎样本分布极不均衡。鼻炎样本标签数及病历分布分别如图7、8所示,每个病例含1~5个输出,病例分布数量总计461例,分别为24、6、330、95、6。对于前4组标签型预测项,每类病历表现出阳性数量有1~3个标签,其中单证候病例354例,占总病例76.79%;兼证病例110例,占总病例21.91%;三证合一病例6例,占总病例1.3%。

表1 鼻炎标签分布及不平衡度数据

Table 1 Distribution of rhinitis labels and its class imbalance ratio

标签	AR	RS	URI	OTH
Positive	431	34	24	28
Negative	30	427	437	433
Ib/%	7.0	8.0	5.5	6.5
b/%	93.27	92.33	94.65	93.73

表2 鼻炎类型分布

Table 2 Distribution of rhinitis types

Severity			Duration		
类型	数值	比例/%	类型	数值	比例/%
Mild	299	64.9	Intermittent	290	63.0
Medium	113	24.5	Persistent	141	30.5
Severe	19	4.1	None	30	6.5
None	30	6.5			

表3给出了针对原始鼻炎样本,采用多种不平衡度计算方法对比数据,包括过采样均衡化SMOTE方法[23],ADASYN方法[24],欠采样均衡化All-KNN方法,原始样本不平衡度(RAW)以及本文方法ARF-OOBEE。可以看出,原始数据不平衡度最低为分度Severity,占比55.53%,最高为AR,占比为90.98%,说明所有待预测样本输出值均为不均衡。采用ADASYN和SMOTE方法对Types

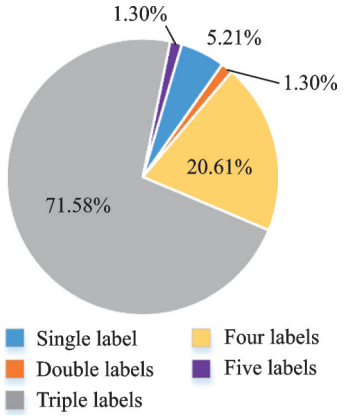

图7 预测输出的数量分布

Fig.7 Distribution of prediction output

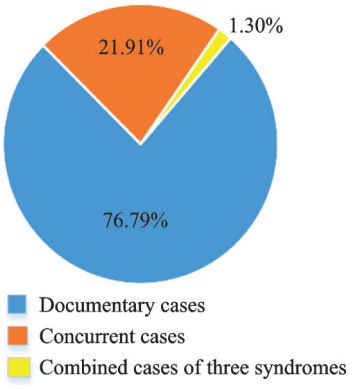

图8 鼻炎病历分布

Fig.8 Types of rhinitis among patients

预测值做均衡化处理后,与AR、Severity和Duration相比,不均衡度至少降低0.403 5,但是RS、URI、OTH不均衡度无明显变化,RS标签上的不均衡度提高了0.085 8。同样,采用SMOTE方法对URI标签均衡化后,URI不均衡度降为0,但RS、OTH不均衡度分别增加0.043 7、0.040 1。采用All-KNN欠采样后具有较高不均衡度(>0.6)。由此可见,常规类别均衡化方法仅能在指定标签上具有较好的效果,无法对多输出样本做整体均衡化。而本文方法ARF-OOBEE将6组不均衡多标签分类问题转化为4组二分类和2组多分类问题,并在组件分类模型中分别实现样本均衡化处理,较好地解决了多输出分类中样本不均衡问题。

表3 各算法样本不平衡度*b*对比

Table 3 Comparison of class imbalance ratio *b* for different methods

Method	AR	RS	URI	OTH	Severity	Duration	Average
RAW	0.932 7	0.923 3	0.946 5	0.937 3	0.676 4	0.708 7	**0.854 2**
ADASYN(Types)	0.491 5	0.964 3	0.905 8	0.862 7	0.033 7	0.187 9	**0.574 3**
SMOTE(Types)	0.333 3	0.960 7	0.698 2	0.911 1	0.166 7	0.000 0	**0.511 7**
SMOTE(URI)	0.098 4	0.922 2	0.000 0	0.935 9	0.430 2	0.273 8	**0.443 4**
All-KNN(Types)	0.814 8	0.913 6	0.814 8	0.864 2	0.636 7	0.722 2	**0.794 4**
ARF-OOBEE	0.003 5	0.007 1	0.003 4	0.004 6	0.015 7	0.013 1	**0.007 9**

2.2 维归约处理

鼻炎预测模型的原始输入特征数为66,具有不同来源的组成和数据类型。如果不做特征降维处理,会增加训练时间、噪音干扰和模型复杂度。常见特征降维方法有主成分分析法(PCA)[25],核主成分分析(KPCA),独立成分分析(ICA)[26],线性判别分析(LDA)[27]等。

本文采用4种特征降维方法FastICA, PCA, KPCA, LDA来对比分析,其中FastICA方法将原66维特征降至25维;PCA将原特征数量降至33维,LDA方法则将原特征数量降至10维,KPCA方法将原特征降至54维。本文使用RF算法对样本分类,根据分类后ROC曲线面积AUC值评估各算法降维效果,表4给出了上述4种降维算法后AUC值,发现FastICA方法效果最佳,达到了0.929,相较于PCA算法最大提升了5.6%。

本文采用安德森-达令检验方法(Anderson-darling test)检验鼻炎样本分布,如图9所示。假设鼻炎样本服从正态分布,当显著性水平$\alpha=0.05$时,特征临界值Critical value=0.746,而各特征统计量(Statistic)均大于临界值,因此拒绝原假设,即样本不服从正态分布。而经典降维方法PCA和LDA均符合正态分布样本。由此可见,本文采用FastICA算法对AR样本进行降维处理,该方法更适用于处理非高斯分布样本,计算简单、要求内存小、收敛速度快,且具有神经网络并行性、分布性等特点,能够从多变量统计数据中发现抽象的、本质的因素或成分。

表4 各种降维方法效果对比

Table 4 Comparison of effects of various dimensional reduction methods

方法	FastICA	PCA	KPCA	LDA
AUC	0.929	0.903	0.921	0.873

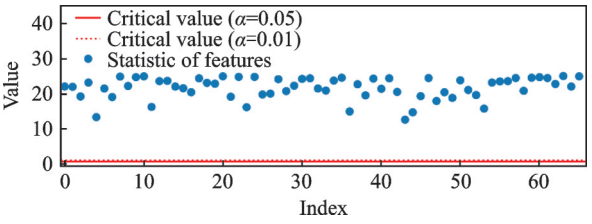

图9 原始样本的Anderson正态分布检验

Fig.9 Anderson normal distribution test of the original sample

2.3 评价指标和对比实验

为评价AR样本预测结果,选择混淆矩阵综合指标:真阳性(TP),假阴性(FN),假阳性(FP)和真阴性(TN)。并使用临床常用性能量测统计参数:精确度(Precision),灵敏度(Sensitivity),特异性(Specificity),G-Mean,F1,ROC曲线面积AUC等作为预测评估指标[28]。

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{9}$$

$$\text{Sensitivity}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{10}$$

$$\text{Specificity}=\frac{\text{TN}}{\text{TN}+\text{FP}} \tag{11}$$

$$\text{F1}=(1+\beta^2)\frac{\text{Precision}\times\text{Sensitivity}}{\beta^2\times\text{Precision}+\text{Sensitivity}}\qquad \beta=1 \tag{12}$$

$$\text{G-Mean}=\sqrt{\text{Sensitivity}\times\text{Specificity}} \tag{13}$$

本文采用6种典型集成学习分类算法进行对比实验,包括深度森林(GCForest)、堆叠集成(GA-Stacking)、代价敏感提升树(AdaCost)、随机森林(RF)、极端随机树(ET)和极端梯度提升树(XGBoost),参数设置如下:

(1) GCForest算法将两个随机森林(RF)和两颗极端随机树(ET)作为基分类器,添加进级联层中,其中每个基分类器的子树设定为100棵,最大深度12。

(2) GA-Stacking算法由遗传算法进行特征筛选,采用两点交叉,单点变异,概率均为0.8,种群规模100,迭代200次;堆叠第一级集成了RF、AdaBoost、梯度提升树(GBDT)、ET、支持向量机(SVM)和XGBoost这6种分类算法,使用10Fold分割训练数据;堆叠第二级采用逻辑回归(LR)对第一级输出预测训练。

(3) AdaCost算法为AdaBoost改进算法,集成了50个基分类器,代价参数为1.25。

(4) RF算法内部由150棵决策树构成,每个决策树最大限制深度为12,叶节点最小分裂数为2。

(5) ET算法参数设置同RF。

(6) XGBoost算法采用200个基学习器,学习率0.01,最大限制深度12。

2.4 集成模型性能分析

本文采用ARF-OOBEE模型与6种典型方法对比,分别将6类鼻炎样本按比例、随机有回放地分层划分训练集与测试集,并进行12次交叉验证,训练集与测试集比例为7:3,并分析模型评估指标均值与方差。表5给出了多种方法综合预测指标,可以发现ARF-OOBEE算法F1值为78.3%,均高于其他6种算法,其中相较于集成学习GCForest算法,F1值提升了3.2%,ARF-OOBEE算法G-Mean值为79.9%,高于其他6种算法,相较于集成学习方法XGBoost提升了1%,表明集成学习方法可以有效地提升不均衡样本分类性能。GCForest算法具有复杂的串联结构,但对于不均衡的鼻炎样本,与RF算法相比,GCForest算法灵敏度提高了2.4%,G-Mean提高1.1%,而精确度降低了4%,这说明仅仅通过增加集成复杂度,无法提升AR的分类精度,同时还会增加模型总体训练时耗。本文提出模型准确率、F1值、灵敏高于其他集成分类模型约2%~3%。说明ARF-OOBEE模型具有自适应特性,可以动态改变集成基分类器数量,对于数据不均衡样本具有较好的综合分类性能。

表5 多种分类方法的综合评价指标

Table 5 Comprehensive evaluation indicator of different classification methods 100%

Method	Accuracy	F1-Score	Precision	Sensitivity	Specificity	G-Mean
ARF-OOBEE	**0.920±0.014**	**0.783±0.061**	0.863±0.058	**0.749±0.071**	0.865±0.030	**0.799±0.054**
GCForest	0.895±0.016	0.751±0.063	0.830±0.079	0.729±0.063	0.844±0.047	0.773±0.055
GA-Stacking	0.893±0.017	0.734±0.066	0.834±0.077	0.719±0.079	0.881±0.040	0.786±0.06
AdaCost	0.877±0.018	0.743±0.058	0.774±0.067	0.735±0.068	0.844±0.049	0.782±0.058
Random Forest	0.898±0.017	0.742±0.060	**0.870±0.063**	0.705±0.065	0.847±0.045	0.762±0.058
Extra Tree	0.894±0.020	0.755±0.060	0.840±0.065	0.733±0.064	0.864±0.037	0.787±0.051
XGBoost	0.889±0.017	0.725±0.056	0.827±0.074	0.705±0.054	**0.903±0.032**	0.789±0.044

表6给出了针对原始样本6类鼻炎病症数据独立分类评价指标。数据分析可知,针对多标签分类鼻炎病症AR、RS、URI、OTH预测准确度较高(>90%),而多分类鼻炎病症Severity、Duration分类准确度较低(<90%)。这是因为前者ARF模型是二分类输出,后者OOBEE模型是多类别分类。两者基分类器均为决策树,但是与多标签二分类相比,多类别分类模型中决策树分裂次数更多,分裂机制更复杂,因此,ARF多标签二分类精度高于多类别分类。此外,AR特异性值仅为59.3%,比其他病症类型明显偏低,这是由于AR型样本不平衡度(93.27%)过大,而ARF-OOBEE算法会自适应均衡化鼻炎AR类的非均衡样本,导致一部分AR样本没有参与样本子集训练,使AR二分类特异性降低,模型会将较少的阴性患者诊断为阳性,导致误诊率升高,但提高了AR多标签二分类的其他评估指标。在实际临床

表6 ARF-OOBEE算法各分类预测评价指标

Table 6 Evaluation Indicator comparison of ARF-OOBEE for different classes 100%

Classification	Accuracy	F1-Score	Precision	Sensitivity	Specificity	G-Mean
AR	0.969±0.008	0.837±0.054	0.911±0.071	0.793±0.059	0.593±0.120	0.685±0.096
RS	0.953±0.025	0.810±0.087	0.938±0.046	0.756±0.097	0.996±0.004	0.866±0.055
URI	0.973±0.008	0.805±0.060	0.908±0.099	0.778±0.097	0.995±0.006	0.878±0.052
OTH	0.960±0.012	0.782±0.084	0.875±0.074	0.735±0.088	0.992±0.004	0.852±0.051
Severity	0.867±0.011	0.747±0.039	0.807±0.019	0.730±0.046	0.843±0.012	0.778±0.033
Duration	0.800±0.021	0.715±0.043	0.740±0.042	0.704±0.041	0.773±0.031	0.732±0.040

中，可以通过医师二次核查排除，从而提升AR特异性，有效地降低鼻炎误诊率。

图10和图11分别给出了本文ARF-OOBEE方法与6种典型分类模型ROC和PR曲线数据统计。由图10中可以看出，蓝线曲线ARF-OOBEE算法从(0,0)点快速上升，增幅大于其他6种方法，说明正例样本检测精度较高。ARF-OOBEE算法AUC面积为0.953，均高于其他算法，比GCForest算法提高1.4%，比RF提高2.4%。由于ROC曲线对数据的不均衡分布不敏感，因此本文还采用了PR曲线作为辅助参考。PR曲线Precision和Recall值都关注正类样本检测率。如图11所示，ARF-OOBEE模型PR曲线位于所有方法曲线的最外围，数据变化平缓，与横轴面积超过了其他6种曲线，其中mAP=0.895，比典型集成GCForest算法多0.5%，比RF算法多2.2%，说明该模型具有较高的查准率和查全率，样本不均衡对分类影响较小，因此，本文鼻炎预测模型ARF-OOBEE具有较好的泛化性能。

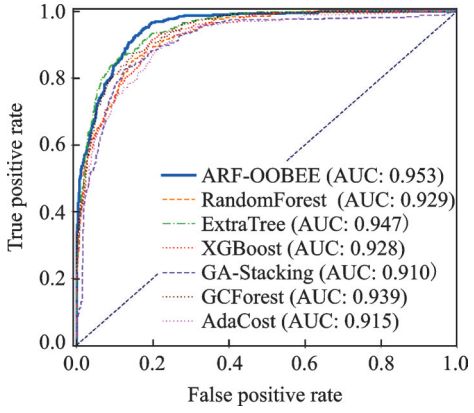

图10 多种分类器ROC曲线对比

Fig.10 ROC curve for comparison different classifiers

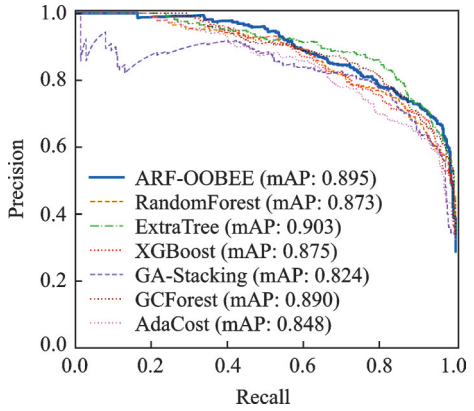

图11 多种分类器PR曲线对比

Fig.11 PR curve comparison for different classifiers

3 结束语

针对临床鼻炎样本高维度、不均衡、稀疏特征，本文构建一种异质集成分类器，采用一种有效的不平衡类分析方法，设计自适应动态子分类器，对多类型、不均衡鼻炎样本实现多输出分类。该方法可快速均衡化鼻炎样本，提高多数类和少数类分类精度。本文算法训练时耗低于GCForest算法，但高于RF算法。LR、NB两种算法分类精度与上述集成方法相似，但训练时耗较低。以后工作考虑对ARF-OOBEE模型自适应参数搜索算法中增加LR基分类器，减少训练时耗。由于ARF-OOBEE模型中含有异质基分类器的并行计算，可以通过对多核处理器的优化来提高模型训练的运算效率和分类精度。

参考文献：

[1] ALSULIMAN T, HUMAIDAN D, SLIMAN L. Machine learning and artificial intelligence in the service of medicine: Necessity or potentiality?[J]. Current Research in Translational Medicine, 2020, 68(4): 245-251.

[2] DEMIRJIAN M, RUMBYRT J S, GOWDA V C, et al. Serum IgE and eosinophil count in allergic rhinitis—Analysis using a modified Bayes' theorem[J]. Allergologia et Immunopathologia, 2012, 40(5): 281-287.

[3] 李少华，王云娜，徐庆文，等.运用聚类分析法研究区域性AR的常见证型[J].世界中西医结合杂志，2015，10(9): 1195-1197.
LI Shaohua, WANG Yunna, XU Qingwen, et al. Clustering analysis to study common syndrome pattern of reginal allergic rhinitis[J]. World Journal of Integrated Traditional Chinese and Western Medicine, 2015, 10(9), 1195-1197.

[4] 黄嘉韵.基于数据挖掘的鼻鼽辨治规律的初步研究[D].广州:广州中医药大学，2015.
HUANG Jiayun. The study of biqiu about syndrome differentiation and tratment law based on digital mining[D]. Guangzhou:

Guangzhou University of Chinese Medicine,2005.

[5] CASTELLANOS-GARZÓN J A, COSTA E, JOSÉ L, et al. An evolutionary framework for machine learning applied to medical data[J]. Knowledge-Based Systems,2019, 185: 104982.

[6] LIU Tianyu, FAN Wenhui, WU Cheng. A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset[J]. Artificial Intelligence in Medicine, 2019,101(3): 101723.

[7] GAN Dan, SHEN Jiang, AN Bang, et al. Integrating TANBN with cost sensitive classification algorithm for imbalanced data in medical diagnosis[J]. Computers & Industrial Engineering, 2020, 140: 106266.

[8] PURWAR A, SINGH S K, Hybrid prediction model with missing value imputation for medical data[J]. Expert Systems with Applications, 2015, 42(13): 5621-5631.

[9] 魏建安,黄海松,康佩栋.针对不平衡数据的PSO-DEC-IFSVM分类算法[J].数据采集与处理, 2019,34(4): 723-735.
WEI Jianan,HUANG Haisong,KANG Peidong. PSO-DEC-IFSVM classification algorithm for unbalanced data[J]. Journal of Data Acquisition and Processing, 2019, 34(4): 723-735.

[10] 菅小艳,韩素青,崔彩霞.不平衡数据集上的Relief特征选择算法[J].数据采集与处理, 2016,31(4): 838-844.
JIAN Xiaoyan, HAN Suqing, CUI Caixia. Relief feature selection algorithm on unbalanced datasets[J]. Journal of Data Acquisition and Processing, 2016,31(4): 838-844.

[11] LASSOUAOUI M, BOUGHACI D , BENHAMOU B. A synergy Thompson sampling hyper-heuristic for the feature selection problem[J]. Computational Intelligence, 2020(3): 1-23.

[12] TSAI M T, WANG D W, LIAU C J, et al. Heterogeneous subset sampling[C]//Proceedings of Computing and Combinatorics, 16th Annual International Conference, COCOON 2010. Nha Trang, Vietnam: DBLP, 2010.

[13] FAN W, STOLFO S J, ZHANG J X, et al. AdaCost: Misclassification cost-sensitive boosting[C]//Proceedings of the 16th International Conference on Machine Learning. San Francisco, CA: [s.n.], 1999: 97-105.

[14] WANG S, YAO X. Diversity analysis on imbalanced data sets by using ensemble models[C]//Proceedings of IEEE Symposium on Computational Intelligence & Data Mining.[S.l.]: IEEE, 2009.

[15] KSIAZEK W, HAMMAD M A, PAWIAK P, et al. Development of novel ensemble model using stacking learning and evolutionary computation techniques for automated hepatocellular carcinoma detection[J]. Biocybernetics and Biomedical Engineering, 2020, 40(4): 1512-1524.

[16] LIU X Y, WU J, ZHOU Z H. Exploratory undersampling for class-imbalance learning[C]//Proceedings of IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics).[S.l.]: IEEE, 2009: 539-550.

[17] KIM K, CHOI H. Adjusting initial weights for Adaboost learning[C]//Proceedings of 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT). Kuta Bali, Indonesia:[s.n.], 2017: 1-5.

[18] KIM Jeonghyun, PARK Jonghyun, KANG Dongjoong. Method to improve the performance of the AdaBoost algorithm using Gaussian probability distribution[C]//Proceedings of 2008 International Conference on Control, Automation and Systems. Seoul,Korea:[s.n.], 2008: 1749-1752.

[19] PERRY T, BADER-EL-DEN M. Imbalanced classification using genetically optimized random forests[C]//Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation (GECCO Companion′15). New York, USA:[s.n.], 2015: 1453-1454.

[20] CAO L, ZHAI Yikui. An over-sampling method based on probability density estimation for imbalanced datasets classification [C]//Proceedings of the 2016 International Conference on Intelligent Information Processing (ICIIP ′16). New York, USA: [s.n.], 2016: 1-6.

[21] 马忠臣,陈松灿.多输出分类综述[J].杭州电子科技大学学报(自然科学版),2019,39(3): 1-9.
MA Zhongchen, CHEN Songcan. A survey on multi-output classification[J]. Journal of Hangzhou University of Electronic Science and Technology (Natural Science edition), 2019,39(3): 1-9.

[22] SEN S, DAS M N, CHATTERJEE R. A weighted kNN approach to estimate missing values[C]//Proceedings of 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN). Noida, India: [s.n.], 2016: 210-214.

[23] CHAWLA N V, BOWYER K W, HALL L O, et al. Smote: Synthetic minority oversampling technique[J]. Journal of Artificial Intelligence Research,2002, 16: 321-357.

[24] HE H, YANG B, GARCIA E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]// Proceedings of Neural Networks, 2008, IJCNN.[S.l.]: IEEE, 2008: 4633969.

[25] YANG Z, ZHUANG X, BIRD C, et al. Performing sparse regularization and dimension reduction simultaneously in multimodal data fusion[J]. Frontiers in Neuroscience, 2019(1): 13.

[26] HUANG Panling, XU Liang, LUO Chuan, et al. A study on noise reduction of gear pumps of wheel loaders based on the ica model[J]. International Journal of Environmental Research and Public Health, 2019, 16(6): 999.

[27] KAUR M, ARORA A S. Classification of arrhythmias with LDA and ANN using orthogonal rotations for feature reduction[J]. IJCSI International Journal of Computer Science, 2012, 19(4): 411-420.

[28] WENG C G, POON J. A new evaluation measure for imbalanced datasets[C]//Proceedings of the 7th Australasian Data Mining Conference.[S.l.]: Australian Computer Society Inc., 2008: 27-32.

作者简介：

杨晶东(1973-)，**通信作者**，男，博士，副教授，研究方向：人工智能、机器学习、机器视觉等，E-mail: eerfriend@yeah.net。

孟一飞(1998-)，男，学士，研究方向：人工智能、机器学习等，E-mail: eerfriend@yeah.net。

荀镕基(1999-)，男，学士，研究方向：人工智能、机器学习等，E-mail: 2682414501@qq.com。

余少卿(1975-)，男，博士后，主任医师，教授，博士生导师，E-mail: yu_shaoqing@163.com。

(编辑：夏道家)