ADASYN 与类别逆比例加权法在阿尔茨海默病不平衡数据中的应用*

杨 慧[1] 易付良[1] 陈杜荣[1] 秦 瑶[1] 韩红娟[1] 崔 靖[1] 白文琳[1] 马艺菲[1] 张 荣[1] 余红梅[1,2△]

【提 要】 **目的** 利用自适应合成抽样(adaptive synthetic sampling,ADASYN)与类别逆比例加权法处理类别不平衡数据,结合分类器构建模型对阿尔茨海默病(alzheimer's disease,AD)患者疾病进程进行分类预测。**方法** 数据源自阿尔茨海默病神经影像学计划(Alzheimer's disease neuroimaging initiative,ADNI),经随机森林填补缺失值,弹性网络筛选特征子集后,利用 ADASYN 与类别逆比例加权法处理类别不平衡数据。分别结合随机森林(random forest,RF)、支持向量机(support vector machine,SVM)构建四种模型:ADASYN-RF、ADASYN-SVM、加权随机森林(weighted random forest,WRF)、加权支持向量机(weighted support vector machine,WSVM),与 RF、SVM 比较分类性能。模型评价指标为宏观平均精确率(macro-average of precision,macro-P)、宏观平均召回率(macro-average of recall,macro-R)、宏观平均 F1 值(macro-average of F1-score,macro-F1)、准确率(accuracy,ACC)、Kappa 值和 AUC(area under the ROC curve)。**结果** ADASYN-RF的分类性能最优(Kappa 值为 0.938,AUC 为 0.980),ADASYN-SVM 次之。利用 ADASYN-RF 预测得到的重要分类特征分别为 CDRSB、LDELTOTAL、MMSE,在临床上均可得到证实。**结论** ADASYN 与类别逆比例加权法都能辅助提升分类器性能,但 ADASYN 算法更优。

【关键词】 类别不平衡 ADASYN 加权法 阿尔茨海默病 分类
【中图分类号】 R195 【文献标识码】 A **DOI** 10.11783/j.issn.1002-3674.2024.02.003

ADASYN and Category Inverse Proportion Weighting Method to Imbalanced Data of Alzheimer's Disease

Yang Hui, Yi Fuliang, Chen Durong, et al (*Department of Health Statistics, School of Public Health, Shanxi Medical University* (030000), *Taiyuan*)

【**Abstract**】 **Objective** The adaptive synthetic sampling(ADASYN) algorithm and category inverse proportion weighting method weighting method were used to balance the datasets, then multi-classification prediction of cognitive normal(CN), mild cognitive impairment(MCI), and Alzheimer's disease(AD) combined with classifiers were performed. **Methods** Data were obtained from the Alzheimer's Disease Neuroimaging Initiative(ADNI) database, which was filled in missing values by random forest(RF), and feature subsets were selected by elastic net(EN). We chose ADASYN algorithm and category inverse proportion weighting method processing the category imbalance data, and four models were constructed by combining RF and support vector machine(SVM) respectively: ADASYN-RF, ADASYN-SVM, weighted random forest(WRF), and weighted support vector machine(WSVM). We evaluated the classification performance by macro-P, macro-R, macro-F1, ACC, Kappa value and area under the receiver operating characteristics curve(AUC). **Results** ADASYN-RF had the best classification performance(Kappa = 0.938, AUC = 0.980), followed by ADASYN-SVM. The most important classification features obtained using ADASYN-RF were CDRSB, LDELTOTAL, and MMSE, which have been clinically validated. **Conclusions** Both the ADASYN algorithm and the category inverse proportion weighting method can assist in improving classifier performance, and the ADASYN algorithm is superior.

【**Key words**】 Category imbalance; Adaptive synthetic sampling; Weighting method; Alzheimer's disease; Classification

阿尔茨海默病(Alzheimer's disease,AD)是一种不可逆的进展性神经系统疾病,表现为认知功能减退,精神行为障碍及日常生活能力的进行性减退。人们常认为痴呆是老年人的常态化事件,但是研究表明,2050 年全球预计罹患痴呆的 1.53 亿人中 60%~80%是由 AD 引起[1]。我国 AD 患者数量居全球之首,而目前尚无早期治疗方法,只能通过长期服用药物进行缓解,给家庭和社会带来沉重的负担。轻度认知障碍(mild cognitive impairment,MCI)通常被认为是介于正常认知(normal cognition,NC)与 AD 的中间状态,是早期干预的关键时间窗和治疗关键期。如何准确识别这三个时期,对 AD 患者进行早筛查、早诊断、早干预至关重要。

医学上较为常用的 AD 诊断方法包括临床症状、神经心理学测试、基因、影像学检查、生物标记物等。多模态联合诊断通常会比单模态提供更高的分类准确率,有文献报道:支链转氨酶蛋白和谷氨酸两种生物标志物能准确区分正常认知和 AD,但在结合蒙特利尔认知评估量表(montreal cognitive assessment,MoCA)后,可进一步提高总体敏感性和特异性[2]。

* 基金项目:国家自然科学基金资助项目(81973154);山西省基础研究计划自由探索类青年项目(202103021223242);山西省研究生教育创新项目(2023KY406)
1. 山西医科大学公共卫生学院卫生统计教研室(030000)
2. 重大疾病风险评估山西省重点实验室
△通信作者:余红梅,E-mail:yu@sxmu.edu.cn

目前包括 AD 在内的医学数据大多具有非均衡分布的特性，即至少有一个类只占数据的少数。传统机器学习的分类研究一般基于均衡数据，假设所有分类错误带来的代价相同，当其应用于非均衡数据时，算法会倾向于识别多数类样本，导致少数类样本的识别精度偏低[3]。实际上我们更关注非均衡数据的少数类样本，所以需要采取一些措施来提高模型对于少数类样本的识别能力。Liu 等通过采取自适应合成抽样（adaptive synthetic sampling，ADASYN）对不平衡数据集进行重采样，再应用传统的分类方法进行分类[4]；李卫平等通过对训练数据属于各类别数目的倒数加权改进 KNN 算法，从而解决类别敏感问题[5]。两种方法都在提高少数类识别精度的同时提高了不平衡数据的整体分类性能，有效的解决了数据不平衡问题。

因此，本文拟结合包含人口学信息、神经心理学测试、基因、影像学检查、生物标记物的 AD 数据，通过将 ADASYN 与类别逆比例加权法分别与传统分类法—随机森林（random forest，RF）、支持向量机（support vector machine，SVM）结合构建 ADASYN-RF、ADASYN-SVM、加权随机森林（weighted random forest，WRF）、加权支持向量机（weighted support vector machine，WSVM）四种分类预测模型，同时与 RF、SVM 进行比较，选出较优模型用于临床辅助诊断。

资料与方法

1. 资料来源

本文采用的数据源自于阿尔茨海默病神经影像学计划（Alzheimer′s disease neuroimaging initiative，ADNI）（http://adni.loni.usc.edu），其利用临床、神经心理学测试、生物化学标志物、影像学与遗传学等数据监测 MCI 和 AD 的进展。

2. 研究内容

本研究纳入的变量包括人口学资料（年龄、性别、教育水平和婚姻状态）、基因（APOE）、生物学标志物（Aβ、总 tau 蛋白和磷酸化 tau 蛋白）、神经心理学测试及 MRI 体积五个方面的变量，详见表 1。

表 1　神经心理学测试和脑 MRI 体积

	英文名称	中文名称
神经心理学测试	CDRSB	临床痴呆量表
	ADAS-Cog11	阿尔茨海默病评定量表-认知分量表条目 11
	ADAS-Cog13	阿尔茨海默病评定量表-认知分量表条目 13
	ADASQ4	阿尔茨海默病评定量表-任务 4（单词识别）
	MMSE	简易精神状态量表
	RAVLT-immediate	瑞氏听觉和语言学习测试-即时回忆
	RAVLT-learning	瑞氏听觉和语言学习测试-学习成绩
	RAVLT-forgetting	瑞氏听觉和语言学习测试-遗忘
	RAVLT-perc-forgetting	瑞氏听觉和语言学习测试-遗忘百分比
	LDELTOTAL	延迟召回
	TRABSCOR	连线试验 B 的时间
	FAQ	功能活动量表
	MoCA	蒙特利尔认知评估量表
	mPACCdigit	改进的临床前阿尔茨海默病认知数字测试组合
	mPACCtrailsB	改进的临床前阿尔茨海默病认知试验测试组合
脑 MRI 体积	Ventricles	脑室体积
	Hippocampus	海马体体积
	Whole Brain	全脑体积
	Entorhinal	内嗅皮层体积
	Fusiform	梭状回体积
	MidTemp	颞中回体积
	Intracranial Volume（ICV）	颅内体积

3. 研究方法

（1）特征选择

首先选取 ADNI 中的基线数据，将数据中缺失值超过 80% 的变量剔除[6]，再利用随机森林填补剩余变量的缺失。为从数据集中选出最相关且非冗余的特征变量，以最大限度地提高 AD 辅助诊断模型的分类精度，使用弹性网络进行变量初筛，相关性分析进行二次筛选。弹性网络同时引入 L1 与 L2 惩罚项，既解除了筛选变量个数的限制，又提高了模型的准确度，是一种理想的医学数据变量筛选方法[7-8]。

（2）类别平衡

类别不平衡数据的处理可从数据层面和算法层面

考虑[9]。数据层面的处理方法一般是采用重采样技术改变数据集的分布,降低不平衡度来提高分类性能,主要包括过采样方法和欠采样方法。欠采样通过删除多数类样本达到平衡,但这很容易导致多数类的重要信息丢失。ADASYN 属于过采样方法中的一种[10],其以数据的密度分布为标准来自动决定需要为每个少数类样本生成的合成样本数量,减少了数据不平衡带来的偏差;还可自适应地将决策边界转移到难学习的少数类样本上,使其生成更多的合成数据[11]。

算法层面主要有单类学习方法、集成方法和代价敏感学习等,通过对传统分类算法改进使分类器对不平衡数据产生较好的分类效果。代价敏感学习在医学非均衡数据的分类及疾病诊断预测领域已较为广泛[12]。加权法基于代价敏感学习思想,根据代价矩阵中各类样本的错分代价为每类样本设置一个权重,对于我们着重关注的少数类样本设置一个较大的权重,加大了其错误分类的惩罚[13]。本研究通过类别逆比例加权法,即将权重设置为各类样本数目的反比,从而加大少数类权重,提高分类性能[13-14]。本文中研究对象包括 251 例 CN,600 例 MCI,207 例 AD,设置 $s_2=1$,$s_1/s_2=600/251$,$s_3/s_2=600/207$,s_1、s_2 和 s_3 分别为 CN、MCI 和 AD 的类样本权重,为使少数类样本的误差率更小,本文取 $s_1=2$,$s_2=1$,$s_3=3$。

(3)建模策略

①模型原理

随机森林以决策树为基学习器,过程中引入了两个随机化:随机选择样本子集和随机选择特征属性,根据每个决策树投票来确定样本的最终类别,增加了模型的多样性,提高了最终集成的泛化性能[15]。然而它会为了最小化整体错误率而更多的关注多数类的预测准确率,导致对类别不平衡数据不敏感。Chen 等提出的 WRF 针对这一问题进行了改进,提高了对类不平衡数据的分类性能[16]。

支持向量机的理论基础是结构风险最小化,基本思想是在样本空间或特征空间中,构造出最优超平面使其与不同类样本集之间的距离最大,从而达到最大的泛化能力[17]。但是 Chew 等使用 SVM 对各类别样本数有较大差异的雷达图像进行目标检测时发现存在偏差,不适用于处理类不平衡数据[18]。

因此本研究将 RF、SVM 分别与 ADASYN、类别逆比例加权法结合,构建模型:ADASYN-RF、ADASYN-SVM、WRF、WSVM,应用于 AD 不平衡分类数据分析。

②参数设置

ADASYN-RF 与 ADASYN-SVM:先通过 ADASYN 算法处理类不平衡数据(通过 UBL 包中的 *AdasynClassif*( ) 函数实现,设置 *beta*=1,使其达到完全类别平衡),而后分别用分类器 RF、SVM 对平衡后的数据进行分类。

WRF:通过 *randomForest*( ) 函数实现。类权重(*classwt*)是构建 WRF 的重要参数,用以设定分类水平的权重,设置为 2∶1∶3;*ntree* 用于设置随机森林中树的数目,默认为 500;参数 *mtry* 用于设定决策树每次分支所选择的变量个数,默认为特征变量的平方根,设定为 3。

WSVM:通过 *svm*( ) 函数实现。类权重设置为 2∶1∶3;惩罚参数 *cost*=1;核函数为径向基核函数(*kernel*=radial)。

(4)模型评价

本研究将数据分为训练集(70%)和测试集(30%),其中训练集使用十折交叉验证,测试集输出模型评价指标。由于本文的数据为类别不平衡数据,宏观平均值(macro-average)更注重对少数类的识别,对类别不平衡问题更加敏感[19],所以本文使用宏观平均精确率(macro-average of precision,macro-P)、宏观平均召回率(macro-average of recall,macro-R)、宏观平均 F1 值(macro-average of F1-score,macro-F1)、准确率(accuracy,ACC)、Kappa 值和 AUC(area under the ROC curve)评价模型性能(n 为类别数)。

$$macro-P=\frac{1}{n}\sum_{i=1}^{n}P_i \tag{5}$$

$$macro-R=\frac{1}{n}\sum_{i=1}^{n}R_i \tag{6}$$

$$macro-F1=\frac{1}{n}\sum_{i=1}^{n}F1_i \tag{7}$$

结　果

1. 基本情况

本次研究共纳入 1058 名研究对象,包括 251 例 CN(23.7%),600 例 MCI(56.7%),207 例 AD(19.6%),对三组间研究对象的所有变量进行比较,发现差异均有统计学意义($P<0.05$),见表 2。

2. 特征变量选择

将变量输入弹性网络进行初筛,剔除特征系数变为 0 的变量后得到了 13 个特征变量。图 1(A)展示了经弹性网络选择后的变量重要性排序,其中 CDRSB 是最为重要的特征变量。进一步绘制相关系数图进行二次筛选,见图 1(B)。结合变量重要性排序,剔除相关系数大于 0.8 的变量(mPACCdigit 和 mPACCtrailsB),最终保留了相关性较低且较为重要的 11 个变量。

表 2　研究对象基本信息（N（%）或中位数（P_{25}, P_{75}））

	CN	MCI	AD	x^2/H	P
例数	251(23.7%)	600(56.7%)	207(19.6%)	–	–
年龄（岁）	71(67,76)	72(67,77)	75(70,80)	24.95	<0.001
性别				17.69	<0.001
男	105(41.8%)	338(56.3%)	121(58.5%)		
女	146(58.2%)	262(43.7%)	86(41.5%)		
教育水平（年）	17(16,18)	16(14,18)	16(14,18)	18.98	<0.001
婚姻状态				15.60	<0.001
已婚	180(71.7%)	462(77.0%)	180(87.0%)		
单身	71(28.3%)	138(23.0%)	27(13.0%)		
APOE4 等位基因				74.26	<0.001
不携带	171(68.1%)	280(46.7%)	58(28.0%)		
携带	80(31.9%)	320(53.3%)	149(72.0%)		
CDRSB	0.0(0.0,0.0)	1.5(1.0,2.0)	4.5(3.5,5.5)	779.97	<0.001
ADAS11	5(3,7)	9(6,12)	20(15,25)	506.92	<0.001
ADAS13	8(5,11)	14(10,20)	30(25,36)	538.05	<0.001
ADASQ4	2(1,4)	5(3,7)	9(8,10)	461.81	<0.001
MMSE	29(29,30)	28(27,29)	23(21,25)	483.03	<0.001
RAVLT−immediate	47(39,56)	35(29,44)	23(18,26)	435.41	<0.001
RAVLT−learning	6(4,8)	4(3,7)	2(1,3)	262.74	<0.001
RAVLT−forgetting	3(2,5)	5(3,6)	5(3,6)	45.91	<0.001
RAVLT−perc−forgetting	29(13,50)	56(33,88)	100(86,100)	321.49	<0.001
LDELTOTAL	13(11,16)	8(5,9)	1(0,3)	645.40	<0.001
TRABSCOR	68(54,91)	94(68,128)	213(125,300)	288.67	<0.001
FAQ	0(0,0)	1(0,4)	13(8,18)	571.29	<0.001
MOCA	26(24,28)	23(21,25)	17(14,20)	436.31	<0.001
mPACCdigit	0.5(−1.5,2.3)	−4.7(−8.5,−2.2)	−16.5(−18.9,−13.8)	651.50	<0.001
mPACCtrailsB	0.7(−1.5,2.0)	−4.2(−7.5,−1.9)	−14.3(−16.7,−12.0)	642.45	<0.001
Ventricles×10^4/mm^3	2.9(2.1,3.9)	3.5(2.5,5.0)	4.8(3.3,6.2)	101.65	<0.001
Hippocampus×10^3/mm^3	7.6(7.0,8.1)	7.1(6.4,7.7)	5.8(5.3,6.5)	268.71	<0.001
WholeBrain×10^5/mm^3	10.4(9.7,11.2)	10.5(9.9,11.2)	9.8(9.2,10.7)	47.89	<0.001
Entorhinal×10^3/mm^3	3.9(3.6,4.4)	3.7(3.3,4.2)	2.9(2.5,3.5)	204.77	<0.001
Fusiform×10^4/mm^3	1.8(1.7,2.0)	1.8(1.7,2.0)	1.6(1.5,1.8)	118.86	<0.001
MidTemp×10^4/mm^3	2.1(1.9,2.3)	2.0(1.9,2.2)	1.8(1.6,1.9)	157.91	<0.001
ICV×10^6/mm^3	1.5(1.4,1.6)	1.5(1.4,1.6)	1.5(1.4,1.6)	18.32	<0.001
Aβ（pg/mL）	1042.4(924.7,1183.6)	841.5(706.0,1063.9)	651.0(556.4,735.6)	253.09	<0.001
t−tau（pg/mL）	206.4(180.3,253.2)	258.7(199.6,326.4)	358.3(298.7,436.4)	203.35	<0.001
p−tau（pg/mL）	18.9(16.1,23.9)	24.9(18.1,32.2)	35.8(29.1,44.0)	212.47	<0.001

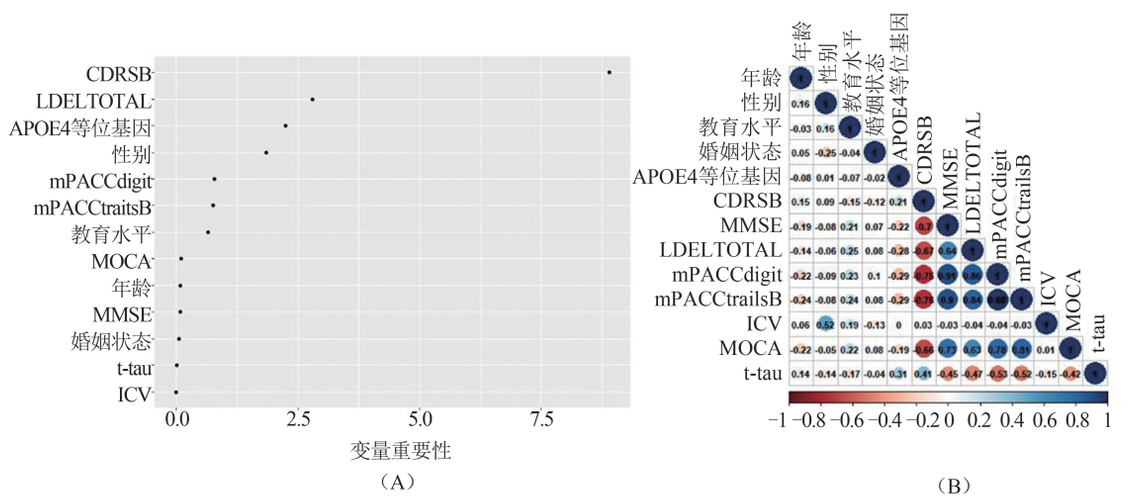

图 1　特征变量选择

3. 模型性能评价

表 3 列出了六种模型的分类性能。其中 ADASYN−RF 分类性能最优（$macro-P$ 为 0.959，$macro-R$ 为 0.959，$macro-F1$ 为 0.959，ACC 为 0.959），ADASYN−SVM 的分类性能（$macro-P$ 为 0.927，$macro-R$ 为 0.928，$macro-F1$ 为 0.927，ACC 为 0.927）仅次于 ADASYN−

RF,这两个模型的分类性能明显优于其他方法。

表 3 模型的分类性能

Model	*macro-P*	*macro-R*	*macro-F*1	*ACC*
SVM	0.879	0.858	0.868	0.880(0.839, 0.914)
RF	0.900	0.866	0.881	0.896(0.846, 0.934)
WSVM	0.862	0.881	0.871	0.880(0.839, 0.914)
WRF	0.910	0.894	0.902	0.912(0.875, 0.941)
ADASYN-SVM	0.927	0.928	0.927	0.927(0.902, 0.948)
ADASYN-RF	0.959	0.959	0.959	0.959(0.938, 0.974)

图 2 分别给出了六种模型的 Kappa 值和 AUC,结果同样显示 ADASYN-RF 的分类性能最优(Kappa 值为 0.938,AUC 为 0.980),ADASYN-SVM 的分类性能次之(Kappa 值为 0.891,AUC 为 0.965),与表 3 的结果一致,说明 ADASYN 算法可以辅助提升分类器的性能。

4. 基于 ADASYN-RF 的特征重要度

Gini 指数平均降低量(mean decrease Gini)用于计算每个变量对分类树的每个节点观测值异质性的影响,该值越大表示该变量的重要性越大。通过 ADASYN-RF 模型进行变量重要性排序,最重要的三个特征变量为 CDRSB、LDELTOTAL、MMSE,而性别、婚姻状况与 APOE4 基因的重要性程度较低,见图 3。

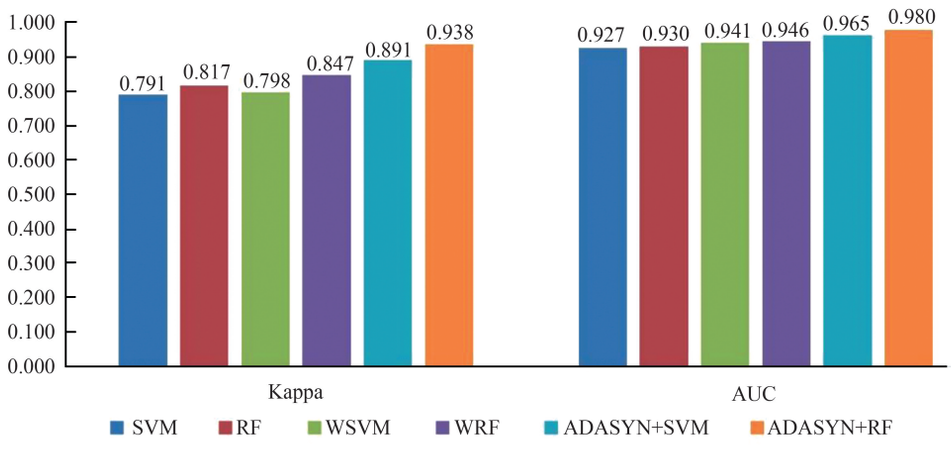

图 2 模型的 Kappa 值与 AUC

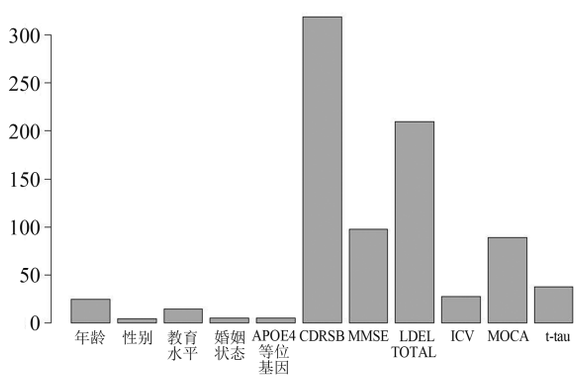

图 3 基于 Gini 指数的变量重要性排序

讨 论

本研究通过弹性网络和相关性分析进行特征选择,使用 ADASYN 算法与类别逆比例加权法两种处理类不平衡数据的方法结合 RF、SVM 构建四种分类模型,并与两种传统分类方法比较,发现 ADASYN-RF 的分类性能最优。

ADASYN 算法是人工合成少数类过采样技术(synthetic minority over sampling technique, SMOTE)的改进,SMOTE 通过对每个少数类样本确定其 K 邻近样本,然后在样本与其近邻之间的连线上进行随机线性插值来合成新的少数类样本,是较为主流的采样技术[20]。但 SMOTE 算法没有考虑少数类样本的分布,从边界和非边界少数类样本中合成新样本数量相同,无法增强决策边界。而 ADASYN 算法弥补了前者的不足,即结合样本分布,增加学习难度较大的新样本数量,以提高分类算法的性能[21]。

类别逆比例加权法的应用同样可提高分类性能,但是效果不如 ADASYN 显著。本文设置类别权重为各类样本数目的反比,是一种较为可靠的加权方法,但是少数类的错分代价很难准确估计,需要多次尝试才能找到合适的权重比,这可能是导致该方法效果略低的主要原因,之后可进一步探索对每个样本设置不同权重以提高分类性能[22]。同时由于 RF 属于集成分类器,分类性能与 SVM 相比较优,因此本研究发现 ADASYN-RF 模型较优。Galar 等人对常见的不平衡数据分类算法进行比较,发现数据采样与集成学习组合的方法分类性能最优,证实了本研究的结论[23]。

通过 ADASYN-RF 模型输出的对分类结果影响较大的三个重要特征变量分别为 CDRSB、LDELTOTAL、MMSE,三者均为神经心理学测试,说明量表对于 AD 的诊断发挥着重要作用[24]。CDRSB 是临床痴呆评分量表总得分,可评估患者的认知和功能,其对 AD 诊断的适应性和稳定性已经得到了证实[25-26]。MMSE 用于评估整体认知状态,是一种关于定向力、注意力、记忆力、计算力、语言和视觉空间能力的认知

功能测试,简单易行,作为基础筛查表以及信效度指针广泛应用于临床[27]。LDELTOTAL 为延迟回忆,已有研究表示在 AD 早期颞叶即处于病理状态导致患者记忆功能明显受损[28-29],该分值显著下降。因此,这三个变量对于 AD 的诊断有一定的临床实用性。

本研究也存在不足,首先数据来源于国外,可能会存在地域差异;其次,只选择了较为代表性的传统分类器(RF、SVM),我们下一步拟寻找合适的国内数据进行验证,拓展结果的外推性和普适性。

参 考 文 献

[1] Nichols E, Steinmetz JD, Vollset SE, et al. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. The Lancet Public Health, 2022, 7(2): e105-e125.

[2] Hudd F, Shiel A, Harris M, et al. Novel Blood Biomarkers that Correlate with Cognitive Performance and Hippocampal Volumetry: Potential for Early Diagnosis of Alzheimer´s Disease. J Alzheimers Dis, 2019, 67(3): 931-947.

[3] Longadge R, Dongre S. Class Imbalance Problem in Data Mining Review. arxiv preprint arxiv: 1305.1707, 2013.

[4] Liu J, Chen Y, Lan L, et al. Prediction of rupture risk in anterior communicating artery aneurysms with a feed-forward artificial neural network. Eur Radiol, 2018, 28(8): 3268-3275.

[5] 李卫平, 杨杰, 王钢. 比例逆权重 kNN 算法及其流处理应用. 计算机工程与设计, 2015, 36(12): 3355-3358.

[6] Mehrmohamadi M, Mentch LK, Clark AG, et al. Integrative modelling of tumour DNA methylation quantifies the contribution of metabolism. Nat Commun, 2016, 7(1): 13666.

[7] Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, 2005, 67(2): 301-320.

[8] 王静娴, 赵芃, 李业棉, 等. 高维生物医学数据变量筛选方法的模拟研究. 西安交通大学学报(医学版), 2021, 42(4): 628-632.

[9] 张贞梅. 面向不平衡数据的集成学习算法研究. 山东科技大学, 2019.

[10] He H, Yang B, Garcia EA, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). 2008, 1322-1328.

[11] Awal MA, Masud M, Hossain MS, et al. A Novel Bayesian Optimization-Based Machine Learning Framework for COVID-19 Detection From Inpatient Facility Data. IEEE Access, 2021, 9: 10263-10281.

[12] Kai MT. Inducing cost-sensitive trees via instance weighting. In Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery, 1998, 23-26.

[13] 李欣欣. 基于代价敏感性随机森林与支持向量机的肝硬化并发肝性脑病风险预测模型研究. 山西医科大学, 2018.

[14] 范昕炜, 杜树新, 吴铁军. 可补偿类别差异的加权支持向量机算法. 中国图象图形学报, 2003(9): 70-75.

[15] Breiman L. Random forests. Machine learning, 2001, 45(1): 5-32.

[16] Chen C, Liaw A, Breiman L. Using random forest to learn imbalanced data. University of California, Berkeley. 2004, 110(1-12): 24.

[17] Cortes C, Vapnik V. Support-vector networks. Machine learning, 1995, 20(3): 273 - 297.

[18] Chew HG, Crisp DJ, Bogner RE, et al. Target detection in radar imagery using support vector machines with training size biasing. In Proceedings of the Sixth International Conference on Control, Automation, Robotics and Vision, Singapore, 2000.

[19] Velarde-Alvarado P, Gonzalez H, Martinez-Pelaez R, et al. A Novel Framework for Generating Personalized Network Datasets for NIDS Based on Traffic Aggregation. Sensors (Basel), 2022, 22(5): 1847.

[20] Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[21] 张扬帆. 基于过采样技术的不平衡数据分类研究. 江南大学, 2019.

[22] He J, Cheng MX. Weighting Methods for Rare Event Identification From Imbalanced Datasets. Frontiers in big Data, 2021, 4: 715320.

[23] Galar M, Fernandez A, Barrenechea E, et al. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2012, 42(4): 463-484.

[24] Battista P, Salvatore C, Berlingeri M, et al. Artificial intelligence and neuropsychological measures: The case of Alzheimer´s disease. Neurosci Biobehav Rev, 2020, 114: 211-228.

[25] Coley N, Andrieu S, Jaros M, et al. Suitability of the Clinical Dementia Rating-Sum of Boxes as a single primary endpoint for Alzheimer´s disease trials. Alzheimers Dement, 2011, 7(6): 602-610.

[26] Cedarbaum JM, Jaros M, Hernandez C, et al. Rationale for use of the Clinical Dementia Rating Sum of Boxes as a primary outcome measure for Alzheimer´s disease clinical trials. Alzheimers Dement, 2013, 9(1 Suppl): S45-55.

[27] 刘瑾, 丁桃. 轻度认知功能障碍诊断标准及神经心理学量表评定的研究进展. 心理月刊, 2019, 14(20): 234-236.

[28] Fokuoh E, Xiao D, Fang W, et al. Longitudinal analysis of APOE-ε4 genotype with the logical memory delayed recall score in Alzheimer´s disease. Journal of Genetics, 2021, 100(2): 1-9.

[29] Carlesimo GA, Perri R, Caltagirone C. Category cued recall following controlled encoding as a neuropsychological tool in the diagnosis of Alzheimer´s disease: a review of the evidence. Neuropsychol Rev, 2011, 21(1): 54-65.

(责任编辑:张　悦)