基于 SMOTE-ENN 和深度森林的弥漫大 B 细胞淋巴瘤复发风险预测*

乔 宇[1] 张岩波[1,2,3] 余红梅[1,2,3] 曹红艳[1,2,3] 周 洁[4] 王俊霞[1] 张高源[1] 于 凯[1] 王雪嫚[1] 郭玉娇[1] 赵志强[5△] 罗艳虹[1,2,3△]

【提 要】 **目的** 对山西省某肿瘤医院血液科 2011—2020 年被确诊为弥漫性大 B 细胞淋巴瘤(diffuse large B-cell lymphoma, DLBCL)并经过治疗达到完全缓解(complete response, CR)的 498 例患者构建 2 年内的复发风险预测模型,为患者的临床治疗提供参考。**方法** 第一步使用最小绝对收缩和选择算子(least absolute shrinkage and selection operator, LASSO)特征选择算法并结合临床医师意见筛选出对 DLBCL 达到 CR 的患者两年复发率影响较大的 21 个变量因素,第二步用 SMOTE(synthetic minority oversampling technique)与 SMOTE-ENN(synthetic minority oversampling technique and edited nearest neighbor)两种不平衡方法处理数据,将原始未处理数据和两种不平衡方法处理后的数据分别使用 7 种分类器进行模型预测。第三步用深度森林(deep forest, DF)做复发风险预测模型。第四步使用准确率(accuracy)、查准率(precision)、灵敏度/召回率(sensitivity/recall)、特异度(specificity)、F1 值(F1-score)和 G 均值(G-means)比较模型分类性能,采用 Brier 分数(Brier score, BS)评价模型校准度。**结果** SMOTE-ENN 不平衡方法下的深度森林算法表现最好(accuracy=0.932, precision=0.949, recall=0.944, specificity=0.910, F1-score=0.946, G-means=0.926, Brier score=0.068)。**结论** 本文使用 SMOTE-ENN 不平衡方法与深度森林分类器结合的方法,对完全缓解的 DLBCL 患者两年复发进行预测,模型达到预期效果。

【关键词】 弥漫性大 B 细胞淋巴瘤 不平衡数据 复发预测 深度森林
【中图分类号】 R195.1 【文献标识码】 A **DOI** 10.11783/j.issn.1002-3674.2025.01.012

Prediction of Recurrence Risk of Diffuse Large B-cell Lymphoma based on SMOTE-ENN and Deep Forest

Qiao Yu, Zhang Yanbo, Yu Hongmei, et al (*Department of Health Statistic, School of Public Health, Shanxi Medical University, Taiyuan* 030001)

【**Abstract**】 **Objective** To construct a 2-year relapse risk prediction model for 498 patients diagnosed with diffuse large B-cell lymphoma (DLBCL) who achieved complete response (CR) following treatment at the hematology department of a cancer hospital in Shanxi Province between 2011 and 2020, providing a reference for clinical management. **Methods** The least absolute shrinkage and selection operator(LASSO) feature selection algorithm, combined with clinical expertise, was first used to identify 21 significant variables influencing the 2-year relapse rate in DLBCL patients with CR. To address data imbalance, synthetic minority oversampling technique(SMOTE) and synthetic minority oversampling technique and edited nearest neighbor (SMOTE-ENN) were applied. Relapse predictions were conducted using seven classifiers on both the original and balanced datasets. The deep forest (DF) algorithm was then employed to build the relapse risk prediction model. Model performance was evaluated using accuracy, precision, sensitivity/recall, specificity, F1-score, and G-means, while calibration was assessed using the Brier score. **Results** The deep forest algorithm, when combined with the SMOTE-ENN method for data imbalance, achieved the best performance (accuracy=0.932, precision=0.949, recall=0.944, specificity=0.910, F1-score=0.946, G-means=0.926, Brier score=0.068). **Conclusion** This study successfully combines the SMOTE-ENN technique with the deep forest classifier to predict 2-year relapse risk in DLBCL patients who achieved CR. The model demonstrates excellent performance and meets expectations.

【**Key words**】 Diffuse large B-cell lymphoma; Unbalanced data; Recurrence prediction; Deep forest

弥漫性大 B 细胞淋巴瘤(diffuse large B-cell lymphoma, DLBCL)是非霍奇金淋巴瘤(non-hodgkin lymphoma, NHL)最常见的亚型[1],占成人非霍奇金淋巴瘤病例的 30%~40%[2],具有高度异质性,在临床和生物学特征、治疗反应和预后方面存在显著差异[3-4]。DLBCL 生存期表现为复发率降低、长期无进展生存期和总生存期随时间推移而改善[5]。达到至少 2 年的持续缓解是 DLBCL 长期预后良好的指标[6]。目前,已有近 20 年历史的 R-CHOP(利妥昔单抗、环磷酰胺、阿霉素、长春新碱和泼尼松)方案仍然是一线 DLBCL 患者的治疗方案,根据患者的年龄、疾病分期和生物学特性,约 60% 的患者可以通过一线

*基金项目:山西省科技厅应用基础研究计划面上项目(202103021224245);2024 年山西省高等学校教学改革创新项目(J20240531);山西省 2024 年度研究生教育创新计划项目(2024JG088);国家自然科学基金青年科学基金(81502897;82273742;82173631);山西医科大学博士启动基金(BS2017029)
1.山西医科大学公共卫生学院卫生统计教研室(030001)
2.重大疾病风险评估山西省重点实验室
3.煤炭环境致病与防治教育部重点实验室
4.山西省肿瘤医院核医学 PET/CT 中心
5.山西省肿瘤医院血液科
△通信作者:罗艳虹, E-mail: lifearena@163.com;赵志强, E-mail: zqzhao69@163.com

R-CHOP治愈[7]。但是对R-CHOP疗法耐受或者病情缓解后复发的患者,其预后较差,难以获得长期无进展生存期[8-9]。临床上目前有针对复发难治性DLBCL的二线治疗方案,能有效提高这类患者生存质量。因此,对疾病进展的准确判断至关重要,精确预测DLBCL进展阶段对指导临床决策和提高治疗效果具有重要意义。

DLBCL复发风险目前常用预测模型有支持向量机[10-11](support vector machine, SVM)、随机森林[12](random forest, RF)以及logistic回归[13]等传统机器学习模型。目前,在数据量足够大的前提下,深度神经网络(deep neural network, DNN)已经实现了远远超过传统机器学习方法的精确度[14]。深度森林(deep forest, DF)受DNN的启示发展而来,既具有传统机器学习模型对小样本、非线性、高维度数据预测性能好的优点,又有DNN所不具备的数据量需求小、模型简单及超参数少、易调参的特点,适合数据集较小的二分类问题。因此,本文采用DF的方法构建模型。

此外,DLBCL早期复发风险比例较低,不平衡率为2.7,因而出现数据不平衡的问题,需要对数据进行重采样以平衡数据集。本文选择采用SMOTE-ENN(synthetic minority oversampling technique and edited nearest neighbor)方法平衡数据集。SMOTE-ENN为过采样和欠采样的方法,可以在增加少数类样本数量的同时,限制多数类样本的删除,从而更好地平衡数据集,保持信息完整性,提高模型性能。

材料与方法

1.数据来源

本研究根据《中国弥漫大B细胞淋巴瘤诊断与治疗指南(2013年版)》[15],回顾性收集山西省某医院2011—2020年被确诊为DLBCL的患者,共498例,随访截止日期为2022年1月。其中在达到完全缓解(complete response, CR)后2年内复发者共136例。通过电子病历分别收集患者各项数据。

2.方法及原理

(1)最小绝对收缩和选择算子(least absolute shrinkage and selection operator,LASSO)特征选择方法

由于DLBCL患者的临床资料维度高、异质性大和冗余特征多,直接使用分类器预测复发风险会导致大量无关变量纳入,影响分类效率,因此首先使用LASSO进行变量筛选。

LASSO算法是一种同时进行特征选择和正则化的线性回归分析方法。其基本思想是通过加入$L1$范数作为惩罚项,根据变量的重要程度赋予其不同大小的惩罚,获得一个特征变量较小的模型,从而有效地避免过拟合[16-17]。

(2)类别不平衡数据

由于DLBCL的复发导致患者在达到CR后的两年内生存率严重降低,两年内复发但生存的人数较少,相比之下未复发的患者数目是复发但生存人数的2~3倍,从而造成“类别不平衡”问题。因此,需要对数据进行处理使其平衡。

不平衡采样技术一般分为两种:欠采样技术(按某种方式删除多数类样本)和过采样技术(按某种方式增加少数类样本)。较为经典的采样技术分别是随机欠采样和随机过采样。然而随机欠采样从数据集中删除多数类样本可能导致有效信息丢失,随机过采样又因在数据集中增加少数类样本的数量而导致过拟合。

为克服随机欠采样的不足,Wilson提出了ENN[18](edited nearest neighbors)欠采样方法。ENN的基本思想是搜索多数类样本的3个最近邻样本,并删除其中具有两个或两个以上少数类样本的多数类样本。由于多数类样本周围更多的还是同类样本,故该方法的数据平衡能力较弱。就本文数据而言,因采用的数据集不平衡率较高,且早期复发为数据集中比较重要的少数类,因此未考虑单独使用欠采样方法平衡数据集。

Chawla在2002年提出SMOTE作为经典的综合少数类过采样算法[19],该算法的主要目标是通过线性插值增加并不存在的少数类样本的数量,从而平衡数据。该方法的主要步骤:首先,采用最邻近算法,计算少数类样本集$x \in X$与X中每个样本的欧氏距离,并找出x的k个近邻。其次,从k个近邻中随机挑选一个样本x',对x与x'进行线性插值构造新样本。插值公式:

$$x_{new}=x+rand\times(x'-x)$$

其中x_{new}即为新样本,rand为0到1的随机数。但SMOTE没有差别的对少数类样本进行采样,容易出现样本重叠和样本噪声等问题。

基于以上原因,Batista等人提出了SMTOE-ENN算法[20]。SMOTE-ENN是一种在不平衡数据集中结合过采样和欠采样少数类技术的采样技术,它基于SMOTE算法,首先使用插值对少数类进行过采样,然后使用ENN方法去除冗余样本,最后生成可与机器学习算法一起使用的类别平衡数据,可兼顾过采样和欠采样的优点,避免过采样和欠采样带来的问题,以实现所需的性能[21]。SMTOE-ENN算法已被证明在多领域优于其他经典抽样方法[22-23]。

本文中不平衡方法取自Python imbalanced-learn库。

(3)深度森林

深度森林算法是Zhou等[24-25]于2017年提出的一种基于决策树分类器(decision tree classifier, DTC)

的 DNN 替代方案。深度森林借鉴神经网络的特性，且相比 DNN 有明显优势：①深度森林比 DNN 有更少的超参数，但更具稳健性；②深度森林比 DNN 在小样本上的训练效果要好。

深度森林由多粒度扫描和级联森林 2 个模块构成。

多粒度扫描：多粒度扫描是处理输入变量的模块，其核心思想是采用滑动窗口对样本采样。用尺寸为 s 的滑动窗口将 M 维原始特征向量变换成大小为 $(M-s)+1$ 维的特征向量；将得到的特征向量利用 RF 和完全随机森林（completely random forest，CRF）进行特征转换以获得类分布向量；将这些分布向量拼接成新的特征向量，用作级联森林结构的输入。

级联森林：级联森林是进行目标参数预测工作的模块，其结构类似神经网络的层状结构，每个级联层由两个 RF 和两个 CRF 组成，通过多个 RF 训练产生类别标签向量，并将类别标签向量与多粒度扫描得到的特征向量相结合，用于训练级联森林的下一级，下一层级联以上述拼接向量作为输入，重复以上过程。在每一层级联产生新的增强向量后，都在验证集上进行验证，如果验证得到准确率有所提升，则将增强向量继续传递给下一层的级联，产生新的拼接向量。直到验证性能趋于一致时，在最后一层将级联森林产生的所有类向量取平均值，类别概率最大值即为该样本最终预测结果。

（4）评价指标

本研究采用准确率（accuracy）、查准率（precision）、灵敏度/召回率（sensitivity/recall）、特异度（specificity）、F1 值（F1-score）和 G 均值（G-means）比较模型分类性能，即模型区分度；模型校准度采用 Brier 分数（Brier score，BS）评价。上述评价指标计算方式如下：

$$accuracy=(TP+TN)/(TP+FN+FP+TN)$$

$$precision=TP/(TP+FP)$$

$$recall=TP/(TP+FN)$$

$$specificity=TN/(FP+TN)$$

$$F1-score=2\times precision\times recall/(precision+recall)$$

$$G-means=\sqrt{recall\times specificity}$$

$$Briers\ core=\frac{1}{N}\sum_{t=1}^{N}(f_t-o_t)^2$$

其中：TP（true positive）为真阳性；FP（false positive）为假阳性；TN（true negative）为真阴性；FN（false negative）为假阴性。

accuracy 是预测结果是准确的概率，在样本严重不均衡时，这种方法就不适用；precision 和 recall 则是评价不平衡数据的常用指标，分别表示所有预测为正的样本中，预测正确的比例和正确预测为正的占全部实际为正的比例。specificity 是指正确识别负样本的数量占比，是评价分类器对负样本识别能力的指标。precision 和 recall 是一个零和博弈过程，二者此消彼长，因此引入 F1 值评分标准。F1 值是两者的调和平均值，在样本不平衡时能更好评估预测性能，较高的 F1 值能保证模型有较高的 precision 和 recall。G-means综合考虑了少数类和多数类的分类性能，必须满足多数类和少数类样本正确率的值同时高，G-means才会高。以上提及的五个指标数值越大，表示分类器的区分度越好。Brier score 是衡量概率校准的一个参数，用于评估分类任务中模型预测结果的准确性，Brier 评分越低表示概率预测越准确。公式中 N 为预测事件数量，f_t 是预测的概率，o_t 是事件 t 的实际概率。

（5）构建模型

本研究采用 Python 3.9.12 构建模型。按照 70% 训练集和 30% 测试集将数据集划分开。缺失值采用均值填补，没有缺失值的变量不列缺失项。分别采用各分类器重复采样构建模型 1000 次，取最终结果的平均值用作计算。

深度森林包从 Github（https://github.com/LAMDA-NJU/Deep-Forest）下载，对比分类器采用神经网络（neural network，NN）中的多层感知机分类器（multi-layer perceptron classifier，mlp classifier）、RF、SVM、logit、DTC 以及高斯朴素贝叶斯（Gaussian naive Bayes，GNB），六种对比模型均取自 Python scikit-learn 包。本研究流程图如图 1 所示。

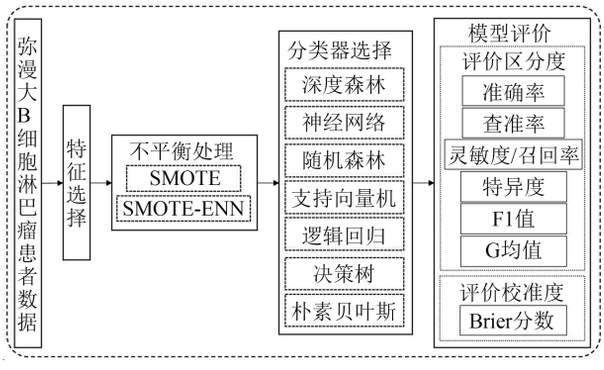

图 1 数据处理流程图

结　果

1. LASSO 特征筛选

使用 LASSO 默认参数对包括性别、年龄、KPS 分数、IPI 得分、B 症状、首发部位、首次治疗方案是否为一线方案等在内的 40 个变量进行特征筛选，共筛选出 21 个对两年复发率影响较大的变量（表 1）。首次化疗方案是否为一线方案经查阅文献以及咨询临床医师后确定，包括 CDOP、CHOP、CHOP-E、CTOP、CTOP-

E、R－CDOP、R－CEOP、R－CHOP、R－ CHOPE、R－CTOP。在上述治疗方案中，C 为 cyclophosphamide，即环磷酰胺；D 为 pegylated liposomal doxorubicin，即脂质体阿霉素；O 为 vincristine，即长春新碱；P 为 prednisone，即泼尼松；H 为 hydroxydaunorubicin，即阿霉素；E 为 etoposide，即依托泊苷；T 为 tamoxifen，即他莫昔芬；R 为 rituximab，即利妥昔单抗。

2.平衡采样数据

本文选取的 498 例病例中，两年内复发并生存的患者有 136 例，而未复发者有 362 例，不平衡率为 2.7，存在不平衡。

为了验证平衡采样效果，分别使用 SMOTE 与 SMOTE-ENN 平衡采样，并与原始数据训练几种常规分类器重复采样 1000 次进行平均结果的比较。

表 1　LASSO 特征筛选出的对两年复发率影响较大的因素及其赋值

变量名	赋值	样本量	构成比(%)	变量名	赋值	样本量	构成比(%)
性别	1=男	262	52.6	原发于胃	1=是	68	13.7
	2=女	236	47.4		0=否	430	86.3
疾病分期	1=Ⅰ期	61	12.2	原发于鼻	1=是	14	2.8
	2=Ⅱ期	162	32.5		0=否	484	97.2
	3=Ⅲ期	104	20.9	原发于腋窝	1=是	53	10.6
	4=Ⅳ期	170	34.1		0=否	445	89.4
	缺失	2	0.4	BCL6	1=是	236	47.4
KPS≥80 分	1=是	355	71.3		0=否	251	50.4
	0=否	135	27.1		缺失	11	2.2
	缺失	8	1.6	Ki-67≥80	1=是	281	56.4
IPI≥3 分	1=是	110	22.1		0=否	215	43.2
	0=否	388	77.9		缺失	2	0.4
结外受累数量	连续变量	498	100.0	CD10	1=阳性	88	17.7
结外受累≥2	1=是	145	29.1		0=阴性	402	80.7
	0=否	353	70.9		缺失	8	1.6
吸烟	1=是	83	16.7	CD20	1=阳性	431	86.5
	0=否	415	83.3		0=阴性	63	12.7
LDH	1=升高	196	39.4		缺失	4	0.8
	0=正常	299	60.0	IRF4/MUM1	1=阳性	251	50.4
	缺失	3	0.6		0=阴性	238	47.8
	1=升高	128	25.7		缺失	9	1.8
	0=正常	365	73.3	C-MYC	1=阳性	59	11.8
	缺失	5	1.0		0=阴性	415	83.4
肿瘤长度	连续变量	498	100.0		缺失	24	4.8
HBV	1=是	49	9.8	首次治疗方案为一线方案	1=是	458	92.0
	0=否	449	90.2		0=否	40	8.0

注：KPS 评分：Karnofsky performance scale，远期生活质量评估；IPI：international prognostic index，国际预后指数；结外受累数量和肿瘤长度是连续变量；LDH：lactate dehydrogenase，为乳酸脱氢酶；HBV：hepatitis B virus，为乙型肝炎病毒；BCL6：B cell lymphoma/leukemia-6，原癌基因 B 细胞淋巴瘤/白血病-6；Ki-67：细胞增殖指数；CD10：cluster of differentiation 10，白细胞分化抗原 10；CD20：cluster of differentiation 20，白细胞分化抗原 20；IRF4/MUM1：interferon regulatory factor 4/multiple myeloma oncogene 1，干扰素调节因子 4/多发性骨髓瘤癌基因 1；C-MYC：cellular myelocytomatosis oncogene，细胞性骨髓细胞增多症癌基因。

3.各分类器的结果比较

由表 2 可知，7 种分类器中，在 SMOTE-ENN 不平衡方法下的深度森林性能综合表现最好（accuracy=0.932，precision=0.949，recall=0.944，specificity=0.910，F1-Score=0.946，G-means=0.926，Brier score=0.068）。就同一分类器的综合性能而言，无论采用哪种不平衡方式，其分类效能相比原始数据均有提升，这表明对不平衡数据的处理是有必要的。GNB 分类器在 SMOTE-ENN 不平衡采样下，precision 和 specificity 的结果表现均为最佳，但综合性能远不如深度森林。

讨　论

对不平衡数据的处理已有较多研究。有作者研究比较了使用随机欠采样（random under-sampling，RUS）、随机过采样（random over-sampling，ROS）、自适应合成采样（adaptive synthetic sampling，ADASYN）、Borderline SMOTE 和 SMOTE-ENN 五种方法来平衡脑出血患者出院生存时间的训练集，SOMTE-ENN 得到了最佳的性能表现[26]。本课题组以前也对 SMOTE、Borderline SMOTE 和 ADASYN 三种过采样法以及 SMOTE-Tomek 和 SMOTE-ENN 两种过采样与欠采样结合共 5 种不平衡方法进行过比较，对 DLBCL 完全缓解的患者进行 2 年复发预测，结果同样显示 SMOTE-ENN 方法下的类别不平衡在相同分类器中表现出最佳的分类性能[10]。Kumari 等[27] 对 SMOTE、SMOTE-KNN 以及 SMOTE-ENN 三种方法在神经网络分类器下的性能作了比较，以此预测针对马尔堡病毒的新种类先导分子，SMOTE-ENN 重采样

方法的表现依然最佳。以上研究使用的分类器不尽相同,但数据类型相似,经过对比选择的重采样方法均为 SMOTE-ENN,以上实验结果均表明 SMOTE-ENN 是一种较为有效的不平衡数据处理方法,能够有效地平衡数据集,提高模型的泛化能力。

表 2 采样前后各分类器性能比较

评价指标	不平衡方法	DF	NN	RF	SVM	logit	DT	GNB
accuracy	采样前	0.702	0.710	0.690	0.641	0.718	0.610	0.690
	SMOTE	0.822	0.737	0.799	0.765	0.677	0.708	0.674
	SMOTE-ENN	0.932	0.904	0.903	0.920	0.825	0.827	0.584
precision	采样前	0.378	0.418	0.379	0.391	0.482	0.306	0.436
	SMOTE	0.856	0.725	0.829	0.748	0.668	0.706	0.662
	SMOTE-ENN	0.949	0.911	0.920	0.940	0.848	0.865	**0.954**
recall	采样前	0.118	0.127	0.195	0.556	0.233	0.331	0.423
	SMOTE	0.777	0.770	0.755	0.802	0.709	0.715	0.715
	SMOTE-ENN	0.944	0.943	0.929	0.935	0.886	0.865	0.369
specificity	采样前	0.924	0.931	0.878	0.673	0.903	0.716	0.792
	SMOTE	0.868	0.705	0.844	0.728	0.646	0.701	0.634
	SMOTE-ENN	0.910	0.836	0.857	0.895	0.718	0.761	**0.965**
F1-Score	采样前	0.175	0.188	0.252	0.456	0.306	0.314	0.424
	SMOTE	0.813	0.744	0.789	0.773	0.686	0.709	0.686
	SMOTE-ENN	**0.946**	**0.926**	0.924	0.937	0.865	0.864	0.526
G-means	采样前	0.317	0.331	0.408	0.609	0.452	0.482	0.575
	SMOTE	0.820	0.735	0.797	0.764	0.676	0.707	0.672
	SMOTE-ENN	**0.926**	0.887	0.892	0.914	0.796	0.810	0.592
Brier Score	采样前	0.298	0.290	0.310	0.359	0.282	0.390	0.310
	SMOTE	0.178	0.263	0.201	0.235	0.323	0.292	0.326
	SMOTE-ENN	**0.068**	0.096	0.097	0.080	0.175	0.173	0.416

注:加粗结果表示在对应类别中模型性能最优。

深度森林作为一种非神经网络式的深度模型,相比传统 DNN 拥有显著的优势。目前,深度森林模型已在不同领域广泛应用,并且表现良好。已有的研究数据表明,深度森林在医学领域的表现也优于目前通用的机器学习模型。如 Su 等[28]通过构建深度森林模型预测抗癌反应,并与在制药行业得到广泛应用的 SVM 进行比较,结果显示深度森林拥有较高的预测精度,证明了该模型的判别能力。Zheng 等[29]基于 RF 的优势和深度森林特点,构建了改进的深度森林模型(tripartite heterogeneous network cascade deep forest, THNCDF)来进行药物-靶点相互作用(drug-target interactions, DTIs)的预测,并将其与四种目前业内较新的方法(RLS-KF[30], RF[31], DTiGEMS[32], iDTI-ESBoost[33])比较,结果显示 THNCDF 方法最优。Zhao 等[34]提出了一种基于深度森林的蛋白质定位算法,结果显示该算法性能优秀,同时参数较少,且更易训练。Guo 等[35]在基因表达数据对癌症亚型进行分类的研究中,基于深度森林提出一种深度学习模型(boosting cascade deep forest, BCDForest),与四种传统方法最近邻算法(KNN)、SVM、logit 和 RF 以及标准深度森林的分类性能进行比较,结果显示文章提出的改进的深度森林和标准深度森林在用到的 5 类癌症数据集上性能都优于传统方法。

本研究对 DLBCL 的复发患者数据特征筛选后进行不平衡采样,分别比较了在两种不平衡方法下深度森林、NN、RF、SVM、Logit、DT 以及 GNB 共 7 种分类器的分类预测能力。结果显示,在 SMOTE-ENN 下,深度森林的综合性能表现 F1-Score = 0.946,G-means = 0.926,Brier score = 0.068,优于其他分类器。这与之前的研究结果一致[28-29,34-35],表明深度森林在医学领域具有良好的预测性能。进一步分析发现,深度森林的优势主要体现在以下两个方面:①泛化能力强:深度森林通过对多个决策树进行集成,能够有效避免过拟合问题,从而提高模型的泛化能力。②稳健性高:深度森林不依赖于特定的数据分布,因此在面对不平衡数据时,仍能保持较好的预测性能。

本研究的不足之处在于:①数据来源单一,为了进一步验证结果需要结合更大样本,多中心数据。②本文构建的 DLBCL 患者复发风险预测模型性能还有待提升。后续我们将进一步探究多种不平衡方法对分类器结果的影响,并比较多款改进后的深度森林分类器之间的预测效果差异,继续在基于深度森林改进后的分类器上寻找一款最优的分类模型。

综上所述,本研究表明,SMOTE-ENN 与深度森林的组合是预测达到 CR 的 DLBCL 患者两年内复发的有效方法。

参 考 文 献

[1] Dürig J, Uhlig J, Gerhardt A, et al. Subcutaneous rituximab in patients with diffuse large B cell lymphoma and follicular lymphoma: Final results of the non-interventional study MabSCale [J]. Cancer Medicine, 2023, 12(3): 2739-2751.

[2] Dunleavy K, Erdmann T, Lenz G. Targeting the B-cell receptor pathway in diffuse large B-cell lymphoma [J]. Cancer treatment reviews, 2018, 65: 41-46.

[3] Fei F, Zheng M, Xu Z, et al. Plasma Metabolites Forecast Occurrence and Prognosis for Patients With Diffuse Large B-Cell Lymphoma [J]. Frontiers in Oncology, 2022, 12: 894891.

[4] Schmitz R, Wright GW, Huang DW, et al. Genetics and pathogenesis of diffuse large B-cell lymphoma [J]. New England Journal of Medicine, 2018, 378(15): 1396-1407.

[5] Pfreundschuh M, Trümper L, Österborg A, et al. CHOP-like chemotherapy plus rituximab versus CHOP-like chemotherapy alone in young patients with good-prognosis diffuse large-B-cell lymphoma: a randomised controlled trial by the MabThera International Trial (MInT) Group [J]. The Lancet Oncology, 2006, 7(5): 379-391.

[6] Ekberg S, Crowther M, Harrysson S, et al. Patient trajectories after diagnosis of diffuse large B-cell lymphoma: a multistate modelling

approach to estimate the chance of lasting remission [J]. British Journal of Cancer, 2022, 127(9): 1642-1649.

[7] Liu Y, Barta SK. Diffuse large B-cell lymphoma: 2019 update on diagnosis, risk stratification, and treatment [J]. American Journal of Hematology, 2019, 94(5): 604-616.

[8] Skrabek P, Assouline S, Christofides A, et al. Emerging therapies for the treatment of relapsed or refractory diffuse large B cell lymphoma [J]. Current Oncology, 2019, 26(4): 253-265.

[9] Coiffier B, Sarkozy C. Diffuse large B-cell lymphoma: R-CHOP failure: what to do? [J]. Hematology-American Society of Hematology Education Program, 2016(1): 366-378.

[10] Xing M, Zhang Y, Yu H, et al. Predict DLBCL patients' recurrence within two years with Gaussian mixture model cluster oversampling and multi-kernel learning [J]. Computer Methods and Programs in Biomedicine, 2022, 226: 107103.

[11] Wang L, Zhao ZQ, Luo YH, et al. Classifying 2-year recurrence in patients with dlbcl using clinical variables with imbalanced data and machine learning methods [J]. Computer Methods and Programs in Biomedicine, 2020, 196: 105567.

[12] Krajnc D, Spielvogel CP, Grahovac M, et al. Automated data preparation for in vivo tumor characterization with machine learning [J]. Frontiers in Oncology, 2022, 12: 1017911.

[13] de Jesus FM, Yin Y, Mantzorou-Kyriaki E, et al. Machine learning in the differentiation of follicular lymphoma from diffuse large B-cell lymphoma with radiomic [18F] FDG PET/CT features [J]. European Journal of Nuclear Medicine and Molecular Imaging, 2022, 49(5): 1535-1543.

[14] Kriegeskorte N, Golan T. Neural network models and deep learning [J]. Current Biology, 2019, 29(7): R231-R236.

[15] 中华医学会血液学分会.中国弥漫大 B 细胞淋巴瘤诊断与治疗指南(2013 年版)[J].中华血液学杂志, 2013, 34(9): 816-819.

[16] Fonti V, Belitser E. Feature selection using LASSO [J]. VU Amsterdam Research Paper in Business Analytics, 2017, 30: 1-25.

[17] 杨耀,李四海.基于对称不确定性和 LASSO 的基因数据特征选择算法[J].信息技术与信息化,2022(1):8-11.

[18] Wilson DL. Asymptotic properties of nearest neighbor rules using edited data [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1972(3): 408-421.

[19] Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique [J]. Journal of artificial intelligence research, 2002, 16: 321-357.

[20] Batista GE, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data [J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 20-29.

[21] Nishat MM, Faisal F, Ratul IJ, et al. A comprehensive investigation of the performances of different machine learning classifiers with SMOTE-ENN oversampling technique and hyperparameter optimization for imbalanced heart failure dataset [J]. Scientific Programming, 2022, 2022: 1-17.

[22] Wang J. Prediction of postoperative recovery in patients with acoustic neuroma using machine learning and SMOTE-ENN techniques [J]. Mathematical Biosciences and Engineering, 2022, 19(10): 10407-10423.

[23] Fotouhi S, Asadi S, Kattan MW. A comprehensive data level analysis for cancer diagnosis on imbalanced data [J]. Journal of Biomedical Informatics, 2019, 90: 103089.

[24] Zhou ZH, Feng J. Deep forest: towards an alternative to deep neural networks. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence [J]. Melbourne, Australia: AAAI Press, 2017: 3553-3559.

[25] Zhou ZH, Feng J. Deep forest [J]. National Science Review, 2019, 6(1): 74-86.

[26] Tang J, Wang X, Wan H, et al. Joint modeling strategy for using electronic medical records data to build machine learning models: an example of intracerebral hemorrhage [J]. BMC Medical Informatics and Decision Making, 2022, 22(1): 278.

[27] Kumari M, Subbarao N. A hybrid resampling algorithms SMOTE and ENN based deep learning models for identification of Marburg virus inhibitors [J]. Future Medicinal Chemistry, 2022, 14(10): 701-715.

[28] Su R, Liu X, Wei L, et al. Deep-Resp-Forest: a deep forest model to predict anti-cancer drug response [J]. Methods, 2019, 166: 91-102.

[29] Zheng Y, Wu Z. Cascade Deep Forest With Heterogeneous Similarity Measures for Drug-Target Interaction Prediction [J]. Frontiers in Genetics, 2021, 12: 702259.

[30] Hao M, Wang Y, Bryant SH. Improved prediction of drug-target interactions using regularized least squares integrating with kernel fusion technique [J]. Analytica Chimica Acta, 2016, 909: 41-50.

[31] Cao DS, Zhang LX, Tan GS, et al. Computational prediction of drug-target interactions using chemical, biological, and network features [J]. Molecular Informatics, 2014, 33(10): 669-681.

[32] Thafar MA, Olayan RS, Ashoor H, et al. DTiGEMS+: drug-target interaction prediction using graph embedding, graph mining, and similarity-based techniques [J]. Journal of Cheminformatics, 2020, 12(1): 1-17.

[33] Rayhan F, Ahmed S, Shatabda S, et al. iDTI-ESBoost: identification of drug target interaction using evolutionary and structural features with boosting [J]. Scientific Reports, 2017, 7(1): 17731.

[34] Zhao L, Wang J, Nabil MM, et al. Deep forest-based prediction of protein subcellular localization [J]. Current Gene Therapy, 2018, 18(5): 268-274.

[35] Guo Y, Liu S, Li Z, et al. BCDForest: a boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data [J]. BMC Bioinformatics, 2018, 19(5): 1-13.

(责任编辑:邓　妍)