基于 SMOTE-ENN 结合改进动态集成选择算法构建 DLBCL 患者 2 年内复发预测模型*

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

【提　要】 目的　构建基于 FIRE 动态集成选择(frienemy indecision region dynamic ensemble selection, FIRE-DES)的弥漫大 B 细胞淋巴瘤(diffuse large B-cell lymphoma, DLBCL)患者治疗达到完全缓解后两年内复发情况的预测模型,为患者的治疗提供决策依据。**方法**　收集山西省某三甲医院 2010 年 1 月至 2020 年 1 月经治疗后达到完全缓解的 498 名患者信息,构建基于四种常见类别不平衡处理方法的 FIRE-DES 复发预测模型,并与传统的五种单一分类器与两种集成分类器进行比较。**结果**　四种类别不平衡算法中 SMOTE-ENN(synthetic minority oversampling technique and edited nearest neighbor)算法取得了最优分类性能,在此基础上采用 DESP(dynamic ensemble selection performance)、KNORAU(K-nearest oracle union)和 META-DES(meta-learning for dynamic ensemble selection)动态集成选择算法的分类效果明显优于传统的单一分类器以及集成分类器模型,基于 FIRE 改进的 DESP、KNORAU 和 META-DES 动态选择算法的分类效果在其基础上实现了进一步提升,且 FIRE-META-DES 取得了最优的分类性能(准确率 = 0.909,精确率 = 0.906,召回率 = 0.967,ROC 曲线下面积 = 0.879,F1-Score = 0.936,Brier Score = 0.088)。**结论**　针对 DLBCL 实际数据集,本文 SMOTE-ENN+FIRE-META-DES 的复发预测模型在性能上达到最优,可为 DLBCL 复发预测提供有力参考。

【关键词】 弥漫大 B 细胞淋巴瘤　复发预测　类别不平衡　动态集成选择
【中图分类号】 R195.1　　**【文献标识码】** A　　**DOI** 10.11783/j.issn.1002-3674.2025.01.009

Recurrence Prediction Model of DLBCL Patients within 2 Years based on SMOTE - ENN Combined with Improved Dynamic Ensemble Selection Algorithm

Zhang Gaoyuan, Zhao Ruiqing, Zhang Yanbo, et al (*Department of Health Statistic, School of Public Health, Shanxi Medical University, Taiyuan* 030001)

【Abstract】 Objective　The prediction model of recurrence within two years after complete remission of diffuse large B-cell lymphoma(DLBCL) patients was constructed based on frienemy indecision region dynamic ensemble selection(FIRE-DES) to provide decision-making basis for the treatment of patients. **Methods**　To collect data of 498 patients who achieved complete response after treatment from January 2010 to January 2020 in a Grade-A hospital in Shanxi Province. A FIRE-DES combination prediction model based on four common category-disequilibrium treatment methods was constructed and compared with five traditional single classifiers and two integrated classifiers. **Results**　Among the four categories of unbalance algorithms, synthetic minority oversampling technique and edited nearest neighbor(SMOTE-ENN) algorithm has obtained the optimal classification performance. On this basis, the classification effect of dynamic ensemble selection performance(DESP), K-nearest oracle union(KNORAU) and meta-learning for dynamic ensemble selection(META-DES) dynamic integration selection algorithms is obviously superior to the traditional single classifier and ensemble classifier model. The classification effect of the improved DESP, KNORAU and META-DES dynamic selection algorithms based on Frienemy Indecision Region is further improved. The classification performance of FIRE-META-DES was the best(Accuracy = 0.909, Precision = 0.906, Recall = 0.967, AUC = 0.879, F1-Score = 0.936, Brier Score = 0.088). **Conclusion**　Aiming at the actual DLBCL data set, SMOTE-ENN + FIRE-META-DES combined prediction model for recurrence used in this paper achieves the optimal performance and low computational complexity, which can provide a strong reference for DLBCL recurrence prediction.

【Key words】 Diffuse large B-cell lymphoma; Recurrence prediction; Category imbalance; Dynamic ensemble selection

弥漫大 B 细胞淋巴瘤(diffuse large B-cell lymphoma, DLBCL)占非霍奇金淋巴瘤的 30%~58%[1],每年有将近 15 万新确诊病例[2]。尽管多数患者在接受一线化疗方案 R-CHOP 后大约 50%到 70%的患者能被治愈[3],但仍有部分患者对治疗不敏感或在达到完全缓解后的两年内复发,使得他们的生存率降低到 10%~20%[4-5]。因此,基于现有数据对获得完全缓解的 DLBCL 患者构建一个更为精确的复发预测模型,对临床医生为患者建立精准的个性化预后治疗方案具有十分重要的意义。

*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

目前疾病复发预测传统的分类算法有 logistic 回归、支持向量机(support vector machine, SVM)、多层感知器神经网络(muti-layer perception, MLP)、K-近邻(K- nearest neighbors, KNN)、决策树模型(decision tree, DT)、随机森林模型(random forest, RF)等。这些传统的分类算法只有在数据类型均衡的情况下,才能获得更好的分类效果。一旦数据出现类别不平衡或者数据异质性较大的情况,这些算法的性能往往比较差[6],因此,研究人员提出了集成学习的概念,它主要是通过集成学习方法产生若干个基分类器,然后通过某种组合策略生成强分类器,从而能够有效地提升模型的泛化能力[7]。这种方法虽然能够得到更好的分类效果,但由于要训练大量的基分类器并且每个基分类器都要进行预测,因此对存储空间的要求很高,从而降低了整体学习的效率[8]。另外,当分类器数量增多时,分类器之间的差异性也会难以控制,给分类造成负面影响。

为了弥补传统集成学习存在的缺陷,Zhou 等人[9]提出一种选择性集成框架,这种框架可以通过一定的标准对待测样本选择出分类效果较好的分类器组来做最终的集成,从而剔除某些对分类有负面影响的分类器,进而降低系统的存储空间以及提高分类的效率[8]。选择性集成可以分为两种:静态集成选择和动态集成选择(dynamic ensemble selection, DES)。静态集成选择对于不同的样本选择固定的分类器组,而 DES 技术旨在为每个新的测试样本选择一个或多个有能力的分类器。最近的工作表明,动态选择技术比静态选择技术获得更高的分类精度[10-12]。DES 首先通过训练大量分类器组成基分类器池,对于每个待测样本使用 KNN 算法从动态选择数据集提取部分样本组成待测样本能力区域(competence region, CR),CR 定义为验证集中样本 X 的 k 个最近邻居的集合,假定 CR 内的样本与待测样本的特征有较高的相似度[13],然后采用 CR 代表待测样本对基分类池中分类器进行性能评估,选择一组合适的分类器进行集成[14]。但是由于 DES 技术仅能选择将 CR 中的所有样本分类到相同类别的分类器,Oliveira 等人[15]提出通过预先选择分类器的方法来解决这个问题,该方法可以从分类器池中预先选择具有跨越 CR 的决策边界的分类,同时动态地修剪分类器池,删除局部无法正确分类的基分类器,从而提高分类性能。

因此,针对 DLBCL 患者数据高维、异质性较大、冗杂以及不平衡的特点,本文首先利用标准化和自适应 LASSO 算法[16]对数据进行预处理,采用 SMOTE(synthetic minority oversampling technique)、Borderline-1 SMOTE、Borderline-2 SMOTE 和 SMOTE-ENN(synthetic minority oversampling technique and edited nearest neighbor)算法[17-18]四种不平衡算法进行重采样以缓解数据不平衡导致的模型偏差并提高模型的泛化能力并选出效果最好的不平衡算法,在此基础上,使用基于改进动态选择算法来构建弥漫大 B 细胞淋巴瘤患者完全缓解后两年内复发情况的预测模型,从而为患者的治疗提供决策依据。

对象与方法

1.对象

收集山西省某三甲医院 2010 年 1 月至 2020 年 1 月经治疗达到完全缓解的 DLBCL 患者病例资料,随访截止日期为 2022 年 1 月,包括人口统计资料与实验室相关检查资料,如性别、年龄、疾病分期等 40 个变量。纳入标准:①山西省某三甲医院 2010—2020 年确诊;②在治疗后达到完全缓解;③有两次及两次以上就诊经历。排除随访中临床资料不全者。其中治疗后达到完全缓解的患者有 498 例,达到完全缓解后对其随访 2 年内复发的患者有 136 例。

2.方法

(1)自适应 LASSO 变量选择

该方法的思想是根据自变量和因变量的关系,对系数采用不同权重进行二次惩罚,惩罚表达式为 $T_j=\lambda\sum_{j=1}^{d}|\beta_j|\tau_j$,作用是对不同重要程度的变量赋予不同大小的惩罚,从而更容易筛选出重要的变量,剔除不重要的变量,从而达到降维的目的[19]。

(2)SMOTE-ENN 重采样

不同类型的样本数量存在较大差异时会造成数据集的类别不平衡,这一点在医学领域的数据集中表现得更为突出[14]。SMOTE 过采样是通过合成新的少数类样本来进行类别不平衡处理。Borderline-SMOTE[20]分为 Borderline-1 SMOTE 和 Borderline-2 SMOTE。Borderline-SMOTE 先将少数类进行处理,根据每个少数类样本的最近邻与该样本类别一致度高低分为噪音样本、危险样本、安全样本三类。Borderline-SMOTE 仅使用危险样本来生成新样本;Borderline-1 SMOTE 在生成新样本时,在 k 近邻随机选择少数类样本;Borderline-2 SMOTE 是在 k 近邻中选择任意一个样本而不关注样本类别。SMOTE-ENN 算法由 SMOTE 和 ENN 两部分组成,是在 SMOTE 算法基础上进行过采样,再用 ENN 算法对样本进行净化处理,最终获得一组数据平衡的样本集合。

(3)分类器模型

①比较模型

本文使用五种单一分类器模型作为对比模型:logistic 回归、SVM、MLP、KNN 和 DT。logistic 回归是一种被广泛应用于分类问题的线性模型,其基本思想

是将输入特征加权组合并通过 sigmoid 函数转换为输出概率。SVM 是一种基于间隔最大化的分类器，它通过寻找决策边界上的支持向量来进行分类。MLP 是一种基于人工神经元的前馈神经网络，通常包含多个隐藏层，它通过学习权重来拟合训练数据解决分类问题。KNN 是一种基于距离度量的非参数模型，它通过找到最近的 k 个训练数据点来对新数据进行分类。DT 是一种基于树结构的分类器，它通过对特征空间进行递归划分来进行分类。

AdaBoost 和 RF 是两种集成模型。AdaBoost 算法[21-22]是训练多个弱分类器并将其组合起来，最终构造出一个更强的终分类器。RF 算法[23]是采用有放回的抽样方式对原始数据集抽样，从数据集中随机抽取 k 个子样本，之后对 k 个子样本分别构建决策树模型且每个决策树对应一个分类结果，最后进行投票，以多数投票法得到最终分类结果。

②DES 算法

动态集成选择技术首先生成一个分类器池，对于每个测试样本，根据它们在测试样本的 CR 内的预测能力选择分类器集合的选择机制，经典的 DES 算法有：DESP（dynamic ensemble selection performance），KNORAU（K-nearest oracle union）、META-DES（meta-learning for dynamic ensemble selection）等。DESP 算法通过比较各个分类器之间的性能来消除不合格的分类器[24]；KNORAU 算法选择能够正确预测 CR 内单个样本的所有基本分类器集合[25]；META-DES 将基本分类器的选择视为一个元学习问题，采用元分类器为待测样本选择合适的分类器组[14]。这些 DES 算法不仅可以有效地处理噪声数据和异常值，还可以自适应地调整分类器集合来适应数据分布的变化和分类器性能的变化，利用更加灵活的方式减少学习器之间的冗余[26]。

③FIRE-DES 框架

FIRE-DES 框架是在 DES 的基础上进行了不确定区域检测和动态剪枝步骤，FIRE-DES 分为三个阶段[15]：

第一，分类器生成阶段。在此阶段使用训练集（T）生成分类器池 C。在 DES 中[27]产生分类器的方法主要有两种，一是产生同质分类器，二是产生异质分类器。异构分类器生成是采用不同的算法生成异构基分类器池，可以保证分类器之间的差异性。因此，本文我们考虑由 logistic 回归、SVM、MLP、KNN 和决策树组成分类器池。

第二，分类器选择阶段。在每个新的测试样本的分类中选择分类器集合。这个阶段有两个步骤。首先，不确定区域检测步骤，在该步骤中，框架评估样本 X 的 CR，以确定它是否位于不确定区域，即 CR 中具有不同类别的边界样本。如果样本 X 位于一个不确定的区域（图 1），则框架将转到下一步骤，否则预先选择所有的分类器。其次，动态剪枝步骤，DES 技术仅能选择将 CR 中的所有样本分类到相同类别的分类器（如图 1 中 c1 和 c4），而 dynamic frienemy pruning（DFP）方法是一种用于在线剪枝深度神经网络的方法，可以从分类器池中预先选择具有跨越 CR 的决策边界的分类器（如图 1 中 c2 和 c3），同时动态地修剪分类器池，删除局部无法正确分类的基分类器。

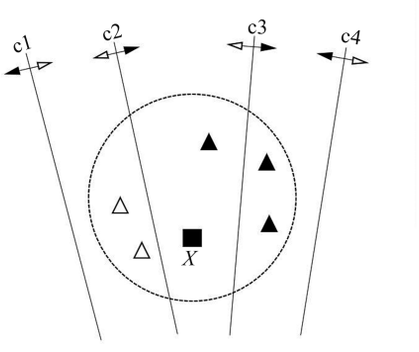

图 1　位于不确定区域中的待测样本 X

第三，分类器集成阶段。最后框架从预选的分类器中选择有能力的分类器用于样本 X 的分类，并使用一个规则将所选分类器进行组合。

（4）评价指标

本研究使用 5 折交叉验证评价模型的性能。使用准确率、精确率、召回率、ROC 曲线下面积（area under curve，AUC）、F1-Score 评估模型的整体区分度，使用 Brier Score（BS）评价它们的校准度，F1 与 BS 的公式如下。

$$F1=\frac{2*精确率*召回率}{精确率+召回率} \quad (1)$$

$$BS=\frac{1}{N}\sum_{i=1}^{n}(p_i-o_i)2 \quad (2)$$

式中，p_i 为预测的概率；o_i 为事件的实际概率；N 为预测事件数量。

结　果

1.自适应 LASSO 变量筛选

根据自适应 LASSO 回归分析结果、临床医生意见及查阅相关文献，最终筛选出性别、疾病分期、KPS 分数、IPI 得分、结外受累数量、吸烟、LDH、肿瘤长度、HBV、原发于鼻、原发于腋下、Ki-67、BCL6、CD10、CD20、C-MYC 及首次治疗方案为一线方案共 17 个变量，各变量具体信息见表 1。

2.类别不平衡方法选择结果

本研究纳入的 498 例病例中，两年内复发的有 136 人，数据不平衡率约为 2.7，因此本文针对 DLBCL

数据集的类别不平衡问题分别使用 SMOTE、Borderline－1 SMOTE、Borderline－2 SMOTE 和 SMOTE－ENN 四种不平衡算法对数据进行重采样以缓解数据不平衡导致的模型偏差，并使用 logistic 回归、SVM、MLP、KNN、DT 五种传统的单一分类器以及 RF 和 Adaboost 两种传统的集成分类器建立 DLBCL 的复发预测模型，各模型的 5 折交叉验证结果均值如表 2 所示，结果表明使用 SMOTE-ENN 算法的各分类器均取得了最优的分类性能，综上，本文最终选用 SMOTE-ENN 算法进行重采样。

表 1 498 例 DLBCL 患者的 17 个变量及赋值

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	原发于鼻	1=是	14	2.8
	2=女	236	47.4		0=否	484	97.2
疾病分期	1=Ⅰ期	61	12.3	原发于腋下	1=是	53	10.6
	2=Ⅱ期	162	32.5		0=否	445	89.4
	3=Ⅲ期	104	20.9	BCL6	1=是	236	47.4
	4=Ⅳ期	170	34.1		0=否	251	50.4
	缺失	1	0.2		缺失	11	2.2
KPS≥80 分	1=是	355	71.3	Ki-67≥80	1=是	281	56.4
	0=否	135	27.1		0=否	215	43.2
	缺失	8	1.6		缺失	2	0.4
IPI≥3 分	1=是	110	22.1	CD10	1=阳性	88	17.7
	0=否	388	77.9		0=阴性	402	80.7
结外受累数量	连续变量	498	100.0		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	C-MYC	1=阳性	59	11.8
	缺失	3	0.6		0=阴性	415	83.3
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

表 2 四种类别不平衡方法与单分类器与集成分类器组合模型的性能比较

分类器	数据平衡方法	准确率	精确率	召回率	AUC	F1-score	BS
logistic	无	0.594	0.258	0.154	0.475	0.193	0.298
	SMOTE	0.640	0.607	0.713	0.643	0.656	0.235
	Borderline-1 SMOTE	0.623	0.598	0.661	0.625	0.628	0.255
	Borderline-2 SMOTE	0.607	0.579	0.669	0.609	0.621	0.288
	SMOTE-ENN	0.801	0.818	0.831	0.797	0.824	0.130
SVM	无	0.685	0.500	0.077	0.521	0.133	0.306
	SMOTE	0.628	0.589	0.748	0.632	0.659	0.231
	Borderline-1 SMOTE	0.719	0.688	0.765	0.721	0.724	0.217
	Borderline-2 SMOTE	0.669	0.643	0.704	0.671	0.672	0.291
	SMOTE-ENN	0.810	0.864	0.785	0.814	0.823	0.112
MLP	无	0.636	0.357	0.192	0.517	0.250	0.306
	SMOTE	0.686	0.647	0.765	0.689	0.701	0.219
	Borderline-1 SMOTE	0.703	0.662	0.783	0.706	0.717	0.217
	Borderline-2 SMOTE	0.674	0.629	0.783	0.678	0.699	0.264
	SMOTE-ENN	0.829	0.826	0.876	0.821	0.850	0.131
KNN	无	0.691	0.524	0.212	0.562	0.301	0.239
	SMOTE	0.678	0.648	0.722	0.679	0.683	0.221
	Borderline-1 SMOTE	0.619	0.577	0.783	0.625	0.664	0.242
	Borderline-2 SMOTE	0.603	0.563	0.774	0.609	0.652	0.271
	SMOTE-ENN	0.810	0.841	0.815	0.809	0.828	0.124
DT	无	0.582	0.328	0.312	0.509	0.319	0.418
	SMOTE	0.639	0.613	0.680	0.641	0.645	0.360
	Borderline-1 SMOTE	0.685	0.670	0.680	0.684	0.675	0.314
	Borderline-2 SMOTE	0.643	0.625	0.649	0.644	0.637	0.347
	SMOTE-ENN	0.774	0.821	0.763	0.776	0.791	0.225

续表 2

分类器	数据平衡方法	准确率	精确率	召回率	AUC	F1-score	BS
RF	无	0.675	0.441	0.112	0.523	0.177	0.219
	SMOTE	0.676	0.651	0.704	0.677	0.677	0.208
	Borderline-1 SMOTE	0.726	0.721	0.704	0.726	0.712	0.176
	Borderline-2 SMOTE	0.695	0.678	0.696	0.695	0.687	0.177
	SMOTE-ENN	0.836	0.835	0.883	0.830	0.858	0.106
Adaboost	无	0.607	0.331	0.231	0.506	0.267	0.342
	SMOTE	0.636	0.616	0.649	0.636	0.632	0.259
	Borderline-1 SMOTE	0.720	0.692	0.755	0.721	0.722	0.214
	Borderline-2 SMOTE	0.645	0.617	0.690	0.647	0.652	0.255
	SMOTE-ENN	0.862	0.857	0.904	0.856	0.880	0.131

3. FIRE-DES 分类结果

本文在采用自适应 LASSO 进行变量筛选和 SMOTE-ENN 算法进行重采样的基础上，进一步采用 KNORAU、DESP 和 META-DES 动态集成选择算法和基于 FIRE 改进的 KNORAU、DESP 和 META-DES 动态选择算法来构建弥漫大 B 细胞淋巴瘤完全缓解后两年内复发情况的预测模型，目标模型与对比模型的 5 折交叉验证结果均值如表 3 所示，各模型性能评价的雷达图如图 2 所示，可以得到以下结论：

（1）单一分类器算法中 MLP 分类效果总体较好，其中准确率、精确率、召回率、AUC、F1-score、BS 分别为 0.827、0.826、0.876、0.820、0.850、0.132。

（2）集成分类器 RF 和 Adaboost 的分类效果整体优于单一分类器模型，其中 Adaboost 分类效果较好，其准确率、精确率、召回率、AUC、F1-score、BS 分别为 0.862、0.857、0.904、0.856、0.880、0.131。

（3）采用 KNORAU、DESP 和 META-DES 动态集成选择算法的分类效果明显优于单一分类器以及 RF 和 Adaboost 集成分类器模型，基于 FIRE 改进的 KNORAU、DESP 和 META-DES 动态选择算法的分类效果在其基础上实现了进一步提升；且 FIRE-META-DES 取得了最优的分类效果，其准确率、精确率、召回率、AUC、F1-score、BS 分别为 0.909、0.906、0.967、0.879、0.936、0.088。

表 3　SMOTE-ENN+DES 与对比模型性能指标对比

类别	分类器	准确率	精确率	召回率	AUC	F1-score	BS
单一分类器	SMOTE-ENN+logistic	0.801	0.818	0.831	0.797	0.824	0.130
	SMOTE-ENN+SVM	0.810	0.864	0.785	0.814	0.823	0.112
	SMOTE-ENN+MLP	0.827	0.826	0.876	0.820	0.850	0.132
	SMOTE-ENN+KNN	0.810	0.841	0.815	0.809	0.828	0.124
	SMOTE-ENN+DT	0.774	0.821	0.763	0.776	0.791	0.225
集成分类器	SMOTE-ENN+RF	0.836	0.835	0.883	0.830	0.858	0.106
	SMOTE-ENN+Adaboost	0.862	0.857	0.904	0.856	0.880	0.131
DES 与 FIRE-DES	SMOTE-ENN+DESP	0.887	0.864	0.987	0.831	0.922	0.098
	SMOTE-ENN+FIRE-DESP	0.895	0.880	0.978	0.849	0.926	0.093
	SMOTE-ENN+KNORAU	0.894	0.872	0.988	0.843	0.927	0.083
	SMOTE-ENN+FIRE- KNORAU	0.902	0.881	0.989	0.854	0.931	0.092
	SMOTE-ENN+META-DES	0.879	0.877	0.955	0.838	0.914	0.102
	SMOTE-ENN+FIRE-META-DES	0.909	0.906	0.967	0.879	0.936	0.088

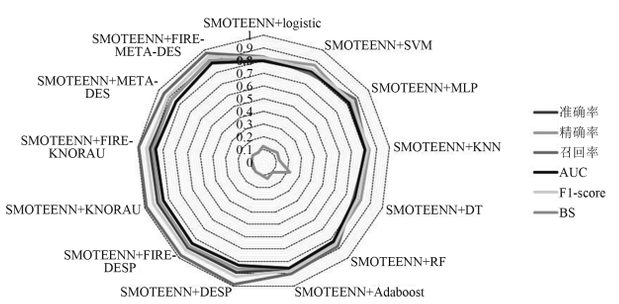

图 2　各模型性能评价的雷达图

综上，与五种传统的单一分类器以及 RF 和 Adaboost 两种传统的集成分类器相比，动态集成选择模型取得了更优的分类性能，FIRE-DES 各框架的分类性能优于 DES，在多个动态集成选择算法中 FIRE-META-DES 取得了最优的分类性能。

讨　论

DLBCL 是我国最常见的淋巴瘤类型，在病理形态、细胞遗传学和临床表现及预后方面均具有较高的侵袭性和异质性，针对 DLBCL 患者数据高维、异质性

较大、冗杂的特点,本文首先利用标准化和自适应LASSO算法对数据进行预处理来解决这个问题,接着采用四种不平衡算法进行重采样来缓解数据的不平衡并选出效果最好的SMOTE-ENN不平衡算法。Khushi等[28]对比了多种处理不平衡数据的算法,证明了SMOTE-ENN算法在PLCO数据集中分类器取得了更好性能。SMOTE-ENN在解决了数据类别不平衡问题后,既降低了过采样的过拟合风险,又解决了欠采样可能导致的信息损失问题。

本文在DLBCL的实际数据集上,使用SMOTE-ENN进行不平衡处理后,使用FIRE-DES框架下的动态集成选择算法构建DLBCL患者的复发预测模型并与传统的单一分类器和集成分类器模型对比。FIRE-DES框架检测样本是否位于不确定区域,如果样本位于不确定区域则对分类器池进行修剪,从而提升分类器整体的学习和分类效率。Lamari等[29]将SMOTEEEN算法和动态集成选择方法META-DES组合,在三个不平衡的医学数据集上验证了该模型高效的分类性能。本研究在此基础上增加了DESP和KNORAU动态集成选择方法构建DLBCL患者复发预测模型,结果显示所有的DES算法均优于用于对比的传统分类器模型,①是因为DES算法对于不同的待测样本选择一组优势分类器进行集成,解决了分类器单一问题和静态集成中不合适的弱分类器影响,使得诊断模型的性能有了明显提升;②是在FIRE-DES框架下对分类器池进行动态剪枝,使得KNORAU、DES-P和META-DES动态集成选择算法的性能均得到了提升。Oliveira等[15]在40个分类数据集上采用不同的动态选择方法对FIRE框架进行了实验。实验结果表明,所提出的FIRE-DES框架在分类精度上优于DES。综上采用FIRE-DES框架对本文所使用的DLBCL数据集有较好的适用性和有效性,且SMOTE-ENN不平衡算法和FIRE-DES框架构建的DLBCL复发预测模型有效提高了预测精度。

本研究的不足之处在于本研究基于静态数据,未充分利用动态的电子病历数据的信息。下一步我们将收集不规则时间序列数据构建分类预测模型。

参 考 文 献

[1] Tilly H, Vitolo U, Walewski J, et al. Diffuse large B-cell lymphoma (DLBCL): ESMO Clinical Practice Guidelines for diagnosis, treatment andfollow-up [J]. Annals of Oncology, 2012, 23(suppl 7): 78-82.

[2] Sehn LH, Salles G. Diffuse Large B-Cell Lymphoma [J]. The New England Journal of Medicine, 2021, 384(9):842-858.

[3] Coiffier B, Thieblemont C, Neste E, et al. Long-term outcome of patients in the LNH-98.5 trial, the first randomized study comparing rituximab-CHOP to standard CHOP chemotherapy in DLBCL patients: a study by the Groupe d'Etudes des Lymphomes de l'Adulte [J]. Blood, 2010, 116(12):2040-2045.

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

[5] Zou Q, Xie S, Lin Z, et al. Finding the Best Classification Threshold in Imbalanced Classification [J]. Big Data Research, 2016,5: 2-8.

[6] 叶枫,丁锋. 不平衡数据分类研究及其应用[J]. 计算机应用与软件,2018,35(1):132-136+205.

[7] 吕晓宁. 多分类器选择性集成方法研究及其应用 [D]. 大连:大连海事大学, 2019.

[8] 朱雪. 动态集成分类方法的研究 [D]. 大连:大连海事大学, 2019.

[9] Zhou ZH, Wu J, Tang W. Ensembling Neural Networks: Many Could Be Better Than All [J].Artificial Intelligence, 2002, 137(1-2):239-263.

[10] Jr A, Sabourin R, Oliveira L. Dynamic selection of classifiers: A comprehensive review [J]. Pattern Recognition, 2014, 47(11): 3665-3680.

[11] Ko A, Sabourin R, Britto AS, et al. From dynamic classifier selection to dynamic ensemble selection [J]. Pattern Recognition, 2008, 41(5): 1718-1731.

[12] Cruz RMO, Sabourin R, Cavalcanti GDC. META-DES.H: a Dynamic Ensemble Selection Technique Using Meta-learning and a Dynamic Weighting Approach [C]. The International Conference on Neural Networks, 2015.

[13] 向欣, 陆歌皓. 基于DESMID-AD动态选择的类别不平衡信用评估模型 [J]. 计算机应用研究, 2021, 38(12): 3604-3610.

[14] 刘子华, 郑汉东, 刘卫勇. 基于改进动态集成选择算法的乳腺肿块辅助诊断模型 [J]. 计算机应用研究, 2023, 40(1): 147-154.

[15] Oliveira D, Cavalcanti G, Sabourin R. Online pruning of base classifiers for Dynamic Ensemble Selection [J]. Pattern Recognition, 2017, 72: 44-58.

[16] Tibshirani R. Regression shrinkage and selection via the LASSO [J]. Journal of the Royal Statistical Society, Series B, 1996, 58(1): 02080.

[17] Wang K, Tian J, Zheng C, et al. Improving Risk Identification of Adverse Outcomes in Chronic Heart Failure Using SMOTE+ENN and Machine Learning [J]. Risk Manag Healthc Policy, 2021, 14: 2453-2463.

[18] Lu T, Huang Y, Zhao W, et al. The Metering Automation System based Intrusion Detection Using Random Forest Classifier with SMOTE+ENN [C]. 2019 IEEE 7th International Conference on Computer Science and Network Technology(ICCSNT), 2019.

[19] 张桢桢. 基于自适应LASSO的食管鳞癌生存风险预测研究 [D]. 郑州:郑州轻工业大学, 2021.

[20] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning [J]. Lecture Notes in Computer Science, 2005,3644:878-887.

[21] Tang D, Tang L, Dai R, et al. MF-Adaboost: LDoS attack detection based on multi-features and improved Adaboost [J]. Future Generation Computer Systems, 2020, 106: 347-359.

[22] Wang C, Xu R, Qiu J, et al. AdaBoost-inspired Multi-operator Ensemble Strategy for Multi-objective Evolutionary Algorithms [J]. Neurocomputing, 2020, 384: 243-255.

(下转第61页)