基于高斯混合模型双向聚类重采样和随机森林构建 DLBCL 早期复发预测模型*

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

【提 要】 **目的** 应用一种可以同时解决少数类和多数类类间和类内不平衡问题的类别不平衡处理方法,并将其与随机森林(random forest, RF)分类器结合实现对弥漫大 B 细胞淋巴瘤(diffuse large B-cell lymphoma, DLBCL)患者早期复发的预测,为 DLBLC 患者的治疗提供参考。**方法** 首先使用一种基于高斯混合模型双向聚类重采样的类别不平衡处理方法(Gaussian mixture model,GMM-GMM)处理数据,并与随机过采样(random over sampling,ROS)、合成少数类过采样技术(synthetic minority over-sampling technique,SMOTE)、Borderline-1 SMOTE、Borderline-2 SMOTE、GMM 上采样、GMM 下采样、SMOTE+RUS、SMOTE+GMM 和 GMM+RUS 进行比较,然后以 RF 作为分类器验证 10 种类别不平衡方法的性能,之后为验证 RF 的性能,在处理后的数据集上使用 logistic 回归和决策树(decision tree, DT)作为对照,最后从区分度和校准度两方面对模型进行评价。**结果** 在本文所有模型中,采用 GMM-GMM 的 RF 模型取得了相对最优的分类性能(accuracy=0.79, AUC=0.87, sensitivity=0.71, specificity =0.87, G-means=0.79, MSE=0.21)。**结论** GMM-GMM 优于其他传统的重采样方法,结合 RF 用于 DLBCL 患者早期复发的预测取得了相对较好的分类结果,可以很好地实现对 DLBCL 患者早期复发的预测。

【关键词】 类别不平衡 高斯混合模型聚类重采样 随机森林 复发预测 弥漫大 B 细胞淋巴瘤
【中图分类号】 R195.1 【文献标识码】 A **DOI** 10.11783/j.issn.1002-3674.2025.01.002

Early Recurrence Prediction Model for DLBCL based on Gaussian Mixture Model Bi-directional Clustering Resampling and Random Forest

Wang Junxia, Zhang Yanbo, Yu Hongmei, et al (*Department of Health Statistics, School of Public Health, Shanxi Medical Uiversity, Taiyuan* 030001)

【**Abstract**】 **Objective** We apply a class imbalance treatment method that can solve the between-class imbalance problem and the within-class imbalance problem of the minority class and the majority class at the same time. And combining it with RF classifier to achieve early recurrence prediction in DLBLC patients, which provided a reference for the treatment of DLBLC patients. **Methods** Firstly, we apply a class imbalance processing method based on Gaussian mixture model bi-directional clustering resampling to process the data. And compared with ROS, SMOTE, Borderline-1 SMOTE, Borderline-2 SMOTE, GMM oversampling, GMM undersampling, SMOTE+RUS, SMOTE+GMM and GMM+RUS. Afterwards, in order to verify the performance of RF, we use logistic regression and decision tree models as controls. Finally, the evaluation of the model is carried out in terms of discrimination and calibration. **Results** The RF model with GMM-GMM resampling achieved relatively optimal classification performance(accuracy=0.79, AUC=0.87, sensitivity=0.71, specificity=0.87, G-means=0.79, MSE=0.21). **Conclusion** GMM-GMM is superior to other traditional resampling methods, and combining it with the RF model for the prediction of early recurrence in DLBCL patients has achieved relatively good classification results, which can well realize the prediction of early recurrence in DLBCL patients.

【**Key words**】 Class imbalance; Gaussian mixture model clustering oversampling; Random forest; Recurrence prediction; Diffuse large B-cell lymphoma

弥漫大 B 细胞淋巴瘤(diffuse large B-cell lymphoma, DLBCL)的发病率约占非霍奇金淋巴瘤(non-Hodgkin lymphoma, NHL)的 45.8%[1],在我国,每年约有 8.4 万人罹患 DLBCL,其中死亡人数约 4.7 万人,是发病率增长速度最快的恶性肿瘤之一[2]。现有研究表明,一线的治疗方案可以使 70%的 DLBCL 患者达到完全缓解(complete remission, CR),但仍有 30%在达到完全缓解后复发,最终发展为难治性疾病[3]。DLBCL 患者初次达到完全缓解后预后相对较好,但复发后预后极差,从而使得该病的死亡率较高,因此,能够准确识别出早期复发的 DLBCL 患者并给予及时有效的治疗具有重要的临床意义。

由于 DLBCL 患者的数据资料存在类别不平衡,如果直接使用分类模型进行预测会导致早期复发的 DLBCL 患者有更高的误分类率,所以在使用分类模型预测前,应先考虑处理数据的类别不平衡。现阶段,解

*基金项目:山西省科技厅应用基础研究计划面上项目(202103021224245);国家自然科学基金青年科学基金(81502897;82273742;82173631);山西省 2024 年度研究生教育创新计划项目(2024JG088);2024 年山西省高等学校教学改革创新项目(J20240531);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

决类别不平衡问题的方法是重采样、成本敏感学习和集成学习[4-5]。由于重采样方法简单易于实现，所以应用广泛。然而随机过采样（random over samping，ROS）、合成少数类过采样技术（synthetic minority over-sampling technique，SMOTE）、Borderline-1 SMOTE、Borderline-2 SMOTE 和随机降采样（random under samping，RUS）等传统的重采样算法只解决了类间不平衡问题，并未考虑类内不平衡问题。但类间和类内不平衡问题都会增加分类器的误分类率[6]。为了解决类内不平衡问题，我们考虑使用聚类重采样。最初 K-means 用于聚类重采样，但 K-means 对数据聚类时，类的形状不灵活只能为圆形，且对于两个簇重叠部分的点的分配缺乏评估方案[7]。高斯混合模型（Gaussian mixture model，GMM）替换 K-means 进行聚类重采样，使用 GMM 聚类可以给出每个样本分配到每个簇的概率，相比 K-means 更加灵活[8]，同时又因为它是一种生成模型，可以直接用来生成新样本[7]。Wei 等人[8]使用 GMM 对少数类聚类进行上采样，Hongpo Zhang 等人[9]使用 SMOTE 对少数类上采样，使用 GMM 对多数类聚类进行下采样，模型性能都有了一定的提升。因此，本文基于聚类重采样的思想，应用了一种基于高斯混合模型双向聚类重采样的类别不平衡处理方法，将其整体命名为 GMM-GMM，同时解决了少数类和多数类的类间和类内不平衡问题，并将其与随机森林（random forest，RF）结合用于 DLBCL 患者早期复发的预测，帮助临床医生及时发现 DLBCL 的早期复发患者并对其进行进一步的个性化、精准化治疗，降低早期复发率，延长 DLBCL 患者的整体生存时间。

资料与方法

1.数据来源与处理

本文参考文献[10]，回顾性收集了山西省某三甲肿瘤医院 2011 年 1 月—2020 年 1 月被诊断为 DLBCL 的患者电子病历信息，汇集了该医院核医学中心、电子病案库和患者随访中心的数据，整理为一般信息、用药信息、实验室检查、疾病现状等共计 38 个指标。最终纳入研究的为初次化疗后获得完全缓解的 498 例 DLBCL 患者。

2.方法及原理

（1）数据预处理

数据预处理包括：缺失值填补、数据标准化和特征选择。

缺失值填补：使用众数填补分类变量的缺失值，使用均值填补连续变量的缺失值。

数据标准化：使用 z-score 标准化方法：$(x-\text{mean}(x))/\text{std}(x)$ 对填补后的数据进行标准化处理。

特征选择：使用自适应 LASSO 法（least absolute shrinkage and selection operator）进行变量筛选。该方法主要采用不同的权重对系数进行二次惩罚，作用是根据变量的重要程度赋予其不同大小的惩罚，从而筛选出对早期复发结局有重要影响的变量[11]。

（2）类别不平衡

类别不平衡包括类之间的不平衡和类内部的不平衡。类间不平衡对应的情况是样本量小的少数类与样本量大的多数类的样本例数不同的情况；类内不平衡对应的情况是，类内样本由许多不同的簇组成，这些簇的大小也不同[12]。

在本文纳入的初次化疗后获得完全缓解的 498 例患者中，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，数据存在类间不平衡。使用 GMM 对 136 例样本聚类后每个簇的样本数量分别为 8、25、28、14、18、16、7、20；使用 GMM 对 362 例样本聚类后每个簇的样本数量分别为 30、65、63、34、50、37、83，因此，少数类和多数类样本聚类后簇的大小不同，数据存在类内不平衡。

GMM 是基于多变量正态分布，假设所有的数据点都是由有限个高斯分布生成，即由 k 个混合成分组成，每个混合成分对应一个高斯分布[7]。高斯分布在现实应用广泛，又因为它是一种生成模型，因此可以用来生成新样本[7]。GMM 定义为：

$$P(x)=\sum_{i=1}^{k}\alpha_i P(x|\mu_i,\delta_i)$$

该分布由 k 个混合成分组成，可以认为是 k 个单一高斯概率密度函数的加权平均，$\sum\alpha_i>0$ 为相应的“混合系数”，$\sum_{i=1}^{k}\alpha_i=1$。μ_i,δ_i 分别是第 i 个高斯混合成分的均数和方差。

本文基于聚类重采样的思想，应用了一种基于高斯混合模型双向聚类重采样的类别不平衡处理方法，即选用 GMM 对少数类样本聚类后逆各亚簇样本数量权重生成新的少数类样本，对多数类聚类后通过删除各亚簇部分样本实现与少数类平衡。将少数类样本和多数类样本重新抽样到一个统一数量 $I_{Resample}$[13]。$I_{Resample}$ 的定义为：

$$I_{Resample}=int(\frac{N}{C})$$

其中 N 为数据集中样本总数，C 为类数。

这个模型 GMM-GMM 首先计算 $I_{Resample}$，少数类样本数量小于 $I_{Resample}$，然后我们使用 GMM 对少数类聚类后进行上采样与 $I_{Resample}$ 平衡；多数类样本数量大于 $I_{Resample}$，我们使用 GMM 将多数类聚类成 C_1 簇（C_1 为多数类样本的最优聚类数），然后从每个簇中选择 $\frac{I_{Resample}}{C_1}$ 的数据，将它们合并成新的多数类数据。此时，多数类样本数量与 $I_{Resample}$ 相平衡，从而得到一个平衡

的数据集。

本文类别不平衡处理算法的具体步骤见图 1。

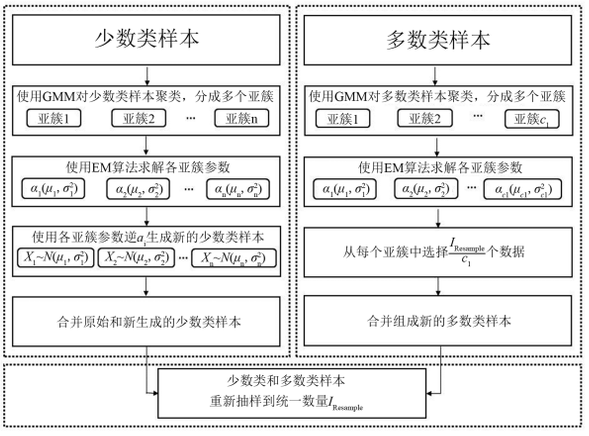

图 1 GMM-GMM 算法流程图

对于少数类样本,设新生成的少数类样本数为n'':$n''=I_{Resample}-n'$,n''为新生成的少数类样本数,n'为原始的少数类样本数。使用 GMM 将原始的少数类样本聚类成若干个亚簇,利用贝叶斯信息准则(Bayesian information criterion, BIC)[14]来确定其最优的聚类数量。利用 EM 算法求解每个亚簇的高斯分量参数,即 α_i,μ_i, $\sum_i$[15-16]。每一个亚簇需新生成的样本数量由公式(1)计算。若计算出的n'为小数,则将其四舍五入即可。使用a_i,μ_i,σ_i,n_i 生成新的少数类样本,并将其与原始的少数类样本合并,形成新的少数类数据集,与 $I_{Resample}$平衡。

$$n_i=n*\frac{\frac{1}{\alpha_i}}{\sum\frac{1}{\alpha_i}},i=1,2,\cdots,n \quad (1)$$

对于多数类样本,设多数类最优的聚类数量为 C_1。使用 GMM 将原始的多数类样本聚类成若干个亚簇,利用 BIC 来确定其最优的聚类数量 C_1。利用 EM 算法求解每个亚簇的高斯分量参数,即 α_i,μ_i, $\sum_i$。从每个簇中选择 $I_{Resample}/C_1$ 的数据,将它们合并成新的多数类数据集,与 $I_{Resample}$平衡。将平衡后的少数类数据集和多数类数据集合并,形成新的相对平衡的数据集,最终构建一个平衡数据集用于 RF 模型的输入。

(3)分类模型

逻辑回归(logistic regression, logistic)是经典的分类算法,它提供了一个 S 形逻辑函数,而不是根据数据调整一条线。给定任意问题的输出概率都由曲线求得,被广泛应用于二值分类。

决策树(decision tree, DT)是一种树形结构,其中每个内部节点表示一个属性上的判断,每个分支代表一个判断结果的输出,最后每个叶节点代表一种分类结果,可以处理分类和预测问题。

RF 是 Breiman[17]提出的一种集成算法,该算法以对原始数据集进行有放回抽样的方式对数据集进行扩充。它由多棵决策树组成,每棵决策树都是一个分类器,所以每棵决策树都对应一个分类结果,对于一个输入样本,N 棵树会有 N 个分类结果。而 RF 集成了所有的分类投票结果,将投票次数最多的类别指定为最终结果。

(4)模型评价

本文采用准确率(accuracy)、ROC 曲线下面积(area under the ROC curve, AUC)、灵敏度(sensitivity)、特异度(specificity)、G 均值(G-means)和均方误差(mean squared error, MSE)作为模型的评价指标,其中 AUC 反映模型的区分度,MSE 反映模型的校准度。混淆矩阵见表 1。

表 1 混淆矩阵

	预测为少数类	预测为多数类
实际少数类	*TP*	*FN*
实际多数类	*FP*	*TN*

$$accuracy=\frac{TP+TN}{TP+TN+FP+FN}$$

$FPR=\frac{FP}{TN+FP}$,$TPR=\frac{TP}{TP+FN}$,ROC 曲线以 FPR 为横坐标,以 TPR 为纵坐标。

$$\text{sensibility}=\frac{TP}{TP+FN}$$

$$\text{specificity}=\frac{TN}{TN+FP}$$

$$\text{G-means}=\sqrt{TPR\times TNR}$$

$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2$,$\hat{y}_i$:预测值;$y_i$:实际值;$n$:样本量。

accuracy 是评价结果准确的概率指标,即模型预测正确的数量占总量的比例;AUC 为 ROC 曲线下与坐标轴围成的面积,其值越大表示模型分类正确的可能性越大,即区分患者是否早期复发的能力越大;sensitivity 即召回率,真阳性率,预测出来的少数类占实际上少数类的比例;specificity 即真阴性率,预测出来的多数类占实际上多数类的比例;G-means 综合考虑了少数类和多数类的分类性能,只有当灵敏度和特异度两者都较高时,其值才会较高;MSE 是模型预测复发和实际复发之间的均方误差,其值越小表示模型的预测误差越小。

结　果

1.特征选择

本文采用自适应 LASSO 共筛选出 14 个变量,此外,查阅文献[18-19]和参考临床医生的意见,将对 DLBCL 患者早期复发结局有重要影响的年龄和性别与自适应 LASSO 筛选出的 14 个变量一起用于

DLBCL 患者早期复发模型的构建。各变量的具体描述见表 2。

表 2　498 例 DLBCL 患者的 16 个变量及赋值

变量名	赋值	例数/构成比(%)
性别	1:男	262(52.6)
	2:女	236(47.4)
年龄	1:≥60 岁	223(44.8)
	0:<60 岁	275(55.2)
疾病分期	1:Ⅰ期	61(12.2)
	2:Ⅱ期	162(32.5)
	3:Ⅲ期	104(20.9)
	4:Ⅳ期	171(34.4)
IPI≥3 分	1:是	110(22.1)
	0:否	388(77.9)
结外受累数量	连续变量	498(100.0)
吸烟	1:是	83(16.7)
	0:否	415(83.3)
LDH	1:升高	196(39.4)
	0:正常	302(60.6)
肿瘤长度	连续变量	498(100.0)
原发于鼻	1:是	14(2.8)
	0:否	484(97.2)
原发于腋窝	1:是	53(10.6)
	0:否	445(89.4)
BCL6	1:是	236(47.4)
	0:否	262(52.6)
Ki-67≥80	1:是	283(56.8)
	0:否	215(43.2)
CD10	1:阳性	88(17.7)
	0:阴性	410(82.3)
CD20	1:阳性	435(87.3)
	0:阴性	63(12.7)
C-MYC	1:阳性	59(11.8)
	0:阴性	439(88.2)
首次治疗方案为一线方案*	1:是	458(92.0)
	0:否	40(8.0)

*: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。

2.模型预测性能比较

使用 ROS、SMOTE、Borderline-1 SMOTE、Borderline-2 SMOTE、GMM 上采样、GMM 下采样、SMOTE+RUS、SMOTE+GMM、GMM+RUS 和 GMM-GMM 等 10 种重采样方法来平衡数据集,然后将全部数据的 80%作为训练集,20%作为测试集,之后分别与 logistic、DT 和 RF 结合用于 DLBCL 患者早期复发的预测,每个模型运行 100 次,由于篇幅所限,本文仅给出测试集循环 100 次各指标的平均值,各模型性能对比见表 3,所有程序均在 Python 3.7 上实现。

由表 3 可知,在所有模型中,采用 GMM-GMM 的 RF 模型取得了相对最优的分类性能(accuracy = 0.79, AUC = 0.87, sensitivity = 0.71, specificity = 0.87, G-means = 0.79, MSE = 0.21)。采用 SMOTE+GMM 的 logistic 模型的 Sensitivity 相对较高,而采用 ROS 的 RF 模型的 Specificity 相对较高,但综合性能都远远不如 GMM-GMM。对比同一分类器在数据经不同重采样方法处理前后的结果可知,GMM-GMM 优于其他传统的重采样方法。综上所述,基于高斯混合模型双向聚类重采样的类别不平衡处理方法与 RF 模型结合用于 DLBCL 患者早期复发预测的性能相对较好。

讨　论

本文针对真实的 DLBCL 数据集应用了一种可以同时解决少数类和多数类类间和类内不平衡问题的类别不平衡处理方法,并将其与 RF 分类器结合用于 DLBCL 患者早期复发的预测。

表 3　不同重采样方法在测试集上的性能指标

分类器	状况	重采样方法	评价指标					
			accuracy	AUC	sensitivity	specificity	G-means	MSE
logistic	采样前类间不平衡	No-Resampling	0.64	0.69	0.62	0.65	0.63	0.36
		ROS	0.65	0.71	0.66	0.65	0.65	0.35
		SMOTE	0.67	0.72	0.70	0.65	0.67	0.33
		Borderline-1 SMOTE	0.69	0.73	0.71	0.67	0.69	0.31
		Borderline-2 SMOTE	0.66	0.70	0.67	0.64	0.66	0.34
		SMOTE+RUS	0.64	0.69	0.66	0.62	0.64	0.36
	单向类内和类间不平衡	GMM 上采样	0.75	0.82	0.72	0.78	0.75	0.25
		GMM 下采样	0.76	0.83	0.72	0.79	0.75	0.24
		GMM+RUS	0.67	0.74	0.66	0.68	0.67	0.33
		SMOTE+GMM	0.75	0.82	**0.78**	0.72	0.75	0.25
	双向类内和类间不平衡	GMM-GMM	**0.77**	**0.85**	0.75	**0.79**	**0.77**	**0.23**
DT	采样前类间不平衡	No-Resampling	0.56	0.57	0.51	0.57	0.53	0.44
		ROS	0.63	0.68	0.66	0.60	0.63	0.37
		SMOTE	0.72	0.79	0.69	0.77	0.72	0.28
		Borderline-1 SMOTE	0.73	0.80	0.67	0.79	0.73	0.27
		Borderline-2 SMOTE	0.72	0.79	0.67	0.76	0.71	0.28
		SMOTE+RUS	0.66	0.72	0.61	0.71	0.65	0.34
	单向类内和类间不平衡	GMM 上采样	0.72	0.77	0.61	0.83	0.71	0.28
		GMM 下采样	0.72	0.79	0.66	0.77	0.71	0.28
		GMM+RUS	0.70	0.78	0.64	0.76	0.69	0.30
		SMOTE+GMM	0.67	0.74	0.67	0.68	0.67	0.33
	双向类内和类间不平衡	GMM-GMM	**0.77**	**0.84**	**0.71**	**0.83**	**0.76**	**0.23**

续表 3

分类器	状况	重采样方法	评价指标					
			accuracy	AUC	sensitivity	specificity	G-means	MSE
RF	采样前类间不平衡	No-Resampling	0.63	0.59	0.37	0.73	0.51	0.37
		ROS	0.76	0.85	**0.81**	0.71	0.76	0.24
		SMOTE	0.77	0.86	0.78	0.76	0.77	0.23
		Borderline-1 SMOTE	0.76	0.85	0.77	0.76	0.76	0.24
		Borderline-2 SMOTE	0.77	0.86	0.76	0.80	0.77	0.22
		SMOTE+RUS	0.70	0.78	0.71	0.69	0.70	0.30
	单向类内和类间不平衡	GMM 上采样	0.78	0.85	0.67	0.87	0.77	0.22
		GMM 下采样	0.75	0.83	0.71	0.80	0.75	0.25
		GMM+RUS	0.71	0.80	0.60	0.81	0.70	0.29
		SMOTE+GMM	0.77	0.85	0.78	0.76	0.77	0.23
	双向类内和类间不平衡	GMM-GMM	**0.79**	**0.87**	0.71	**0.87**	**0.79**	**0.21**

注:加粗结果表示在对应类别中模型性能最优。

对于类别不平衡问题,ROS 是最初的上采样形式,其对少数类样本直接进行复制,很容易导致过拟合[20]。SMOTE 是 Chawla NV 等提出的一种改进算法,通过合成新的少数类来平衡数据集[21]。Borderline-1 SMOTE 和 Borderline-2 SMOTE 是 SMOTE 的改进,这两种类型的 SMOTE 都使用边界上的少数类样本合成新的样本数据[22]。RUS 是通过随机剔除多数类样本来平衡数据集,容易导致过拟合。但以上算法的关注点多为少数类和多数类的类间不平衡,并未考虑类内不平衡。2017 年 Last[23] 提出了基于聚类的上采样方法,即首先使用 K-means 对少数类样本聚类,然后使用 SMOTE 生成新的样本。基于同样的思想,2017 年 Xin Wei 等人将 K-means 替换成 GMM,GMM 是由 Stauffer[24] 等人提出的一种混合概率分布。2017 年 Wei 等人[8] 使用 GMM 进行聚类上采样来平衡数据,然后使用简单贝叶斯算法作为分类器对互联网协议电视用户投诉进行预测,结果优于 SMOTE、Borderline- SMOTE 和 K-means 聚类过采样。2020 年 Zhang 等人[9] 构建了 SGM-CNN 模型,即使用 SMOTE 对少数类上采样,使用 GMM 对多数类下采样,使用 CNN 神经网络对网络入侵进行监测,结果也优于传统的重采样方法。2021 年郑建华[25] 融合了 Borderlins1-SMOTE、GMM 逆权重上采样和随机下采样的混合重采样方法平衡数据,然后使用随机森林模型作为分类器,模型性能也有了一定的提升。

实验结果表明,基于高斯混合模型双向聚类重采样的类别不平衡处理方法优于其他传统的重采样算法,同时解决了少数类和多数类的类间和类内不平衡问题,再结合 RF 模型用于 DLBCL 患者早期复发的预测取得了相对较好的分类结果,可以很好地实现对 DLBCL 患者早期复发的预测。

参 考 文 献

[1] Intragumtornchai T, Bunworasate U, Wudhikarn K, et al. Non-Hodgkin lymphoma in South East Asia: An analysis of the histopathology, clinical features, and survival from Thailand [J]. Hematological Oncology, 2018, 36(1): 28-36.

[2] 张晓娟, 杜伟, 郭树霞. 弥漫性大 B 细胞淋巴瘤组织 MYC 和 Bcl-2 及 Bcl-6 检测的预后价值[J]. 中华肿瘤防治杂志, 2015, 22(15): 1193-1197.

[3] Parvez A, Tau N, Hussey D, et al. 18F-FDG PET/CT metabolic tumor parameters and radiomics features in aggressive non-Hodgkin′s lymphoma as predictors of treatment outcome and survival [J]. Annals of Nuclear Medicine, 2018, 32(2): 1-7.

[4] Tarekegn AN, Giacobini M, Michalak K. A review of methods for imbalanced multi-label classification [J]. Pattern Recognition, 2021, 118: 107965.

[5] Guo H, Jennifer Y, Li S, et al. Learning from class-imbalanced data: Review of methods and applications [J]. Expert Systems with Applications, 2017, 73(1): 220-239.

[6] Japkowicz N. Concept-Learning in the Presence of Between-Class and Within-Class Imbalances [J]. Springer-Verlag, 2001: 67-77.

[7] 周志华.机器学习[M]. 北京:清华大学出版社, 2016.

[8] Wei X,Li Z,Liu R,et al. IPTV User′s complaint prediction based on the Gaussian mixture model for imbalanced dataset[J]. Journal of Computers,2017,28(6):216-224.

[9] Zhang H, Huang L, Wu CQ, et al. An Effective Convolutional Neural Network Based on S-MOTE and Gaussian Mixture Model for Intrusion Detection in Imbalanced Dataset [J]. Computer Networks, 2020, 177(18):107315.

[10] 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.

[11] 袁宝红, 卢宇, 胡婷芳. 基于自适应 Lasso 流形规整的特征提取算法研究[J]. 湖南文理学院学报(自然科学版), 2021, 33(4): 23-26.

[12] Jia S, Huang X, Qin S. A bi-directional sampling based on K-means method for imbalance text classification[J]. 2016 IEEE/ACIS 15th International Conference on Computer and International Conference on Computer and Information Science, ICIS 2016-Proceedings, 2016: 1-6.

[13] Abdulhammed R, Musafer H, Alessa A, et al. Features Dimensionality Reduction Approaches for Machine Learning Based Network Intrusion Detection [J]. Electronics, 2019, 8(3): 1-27.

(下转第 17 页)