基于多随机经验核的弥漫大 B 细胞淋巴瘤复发预测*

李雪玲[1,2] 赵艳琳[1,2] 张岩波[1,2,3] 余红梅[1,2,3] 周 洁[4] 李 琼[1,2] 王俊霞[1,2]
乔 宇[1,2] 张高源[1,2] 赵志强[5△] 罗艳虹[1,2,3△]

【提 要】 **目的** 基于多随机经验核分类器构建弥漫大 B 细胞淋巴瘤完全缓解后两年内复发情况的预测模型,为患者的治疗提供决策依据。**方法** 利用山西省某三甲医院 2010-2020 年电子病历库中符合本研究要求的 445 名患者信息,基于五种常见类别不平衡处理方法以及多随机经验核分类器构建复发预测模型,并与五种分类器进行比较。**结果** 基于 SMOTE Tomek Links+多随机经验核分类器的复发预测模型取得了最优的分类性能(accuracy=0.89,precision=0.87,recall=0.92,f1-Score=0.89,brier score=0.11)。**结论** 对 DLBCL 实际数据集,本文使用 SMOTE Tomek links 处理不平衡数据并构建多随机经验核模型,模型性能达到最优的同时计算复杂度也不高,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 经验核映射 类别不平衡
【中图分类号】 R195.1 【文献标识码】 A **DOI** 10.11783/j.issn.1002-3674.2024.03.003

Precise Prediction of Diffuse Large B-Cell Lymphoma based on Multiple Random Empirical Kernel Learning Machine

Li Xueling, Zhan Yanlin, Zhang Yanbo, et al (*Department of Health Statistics, School of Public Health, Shanxi Medical University*(030001), *Taiyuan*)

【**Abstract**】 **Objectives** To construct a prediction model of relapse in diffuse large B-cell lymphoma within two years after complete remission based on multiple randomized empirical kernel learning machine to provide a basis for patient treatment decisions. **Methods** Using the information of 445 patients who met the requirements of this study in the electronic medical record database of a tertiary hospital in Shanxi Province from 2010 to 2020, a relapse prediction model was constructed based on five common categories of imbalance treatment methods and a multiple stochastic empirical kernel learning machine, and compared with the five classifiers. **Results** The recurrence prediction model based on SMOTE Tomek Links + multiple randomized empirical kernel learning machine achieved optimal classification performance (accuracy = 0.89, precision = 0.87, recall = 0.92, f1-Score = 0.89, brier score = 0.11). **Conclusion** For the actual DLBCL dataset, in this paper, we used SMOTE Tomek links to process the imbalance data and construct a multiple randomized empirical kernel learning machine, which achieves the optimal model performance with low computational complexity and can provide a powerful reference for DLBCL recurrence prediction.

【**Key words**】 Diffuse large B-cell lymphoma; Recurrence prediction; Empirical kernel mapping; Category imbalance

弥漫大 B 细胞淋巴瘤(diffuse large B-cell lymphoma,DLBCL)每年有将近 15 万新确诊病例[1]。联合利妥昔单抗治疗以来,患者的生存率有所提高,50% 到 70% 的患者在接受该治疗方案后可以治愈,但仍存在约三分之一的患者对该方案耐药[2],进而使患者在达到完全缓解后的两年内复发,他们的生存率降低到 10%~20%[3-4]。所以有必要利用现有数据构建一个预测精度较高的复发预测模型,来预测达到完全缓解的 DLBCL 患者两年内是否复发,为临床医师决策提供参考,以便临床医师进行精准、个性化治疗,降低患者复发可能。

DLBCL 患者数据具有高维、异质性较大、冗杂的特点。使用单一分类器构建的模型,有着性能差、过拟合等问题;集成模型虽能提高分类准确率,但其内部原理不易解释;而将多个或多种核函数组合起来的多核学习算法,提高了分类准确性的同时,结果易于解释。近年来,多核算法已被广泛应用于疾病预测[5-7]。传统的多核学习都是隐式核映射,由于隐式核映射对部分算法的内核化有所限制[8],研究人员引入了经验核映射[9],经验核映射是通过显式的映射将样本映射到核空间。然而传统多经验核分类器(multiple empirical kernel learning,MEKL)构造核空间时计算复杂度较高且构造的核空间具有较高维度。本文应用的多随机经验核分类器(multiple random empirical kernel learning machine,MREKLM)[10]是基于随机投影技术的随机经验核映射(random empirical kevnel learning machine,REKM),将样本以较低的计算复杂性映射到多个低维的经验特征空间。

本文基于五种常见类别不平衡处理方法以及多随机经验核分类器构建复发预测模型,并与五种分类器

*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);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

进行比较。

资料来源

参考《中国弥漫大 B 细胞淋巴瘤诊断与治疗指南 2013 版》[11]，本研究回顾性地收集山西省某三甲医院 2010 年 4 月到 2020 年 12 月被诊断为 DLBCL 的 585 名患者信息，将其中初次化疗后达到完全缓解的 445 例患者纳入研究，收集了患者的一般信息、疾病现状、实验室检查以及用药信息等 52 个变量，收集到的数据缺失率为 11.67%，采用均值插补进行补全。

首先对数据进行标准化处理，结合 LASSO 算法[12]及临床专家意见筛选出对患者复发有意义的变量。使用 LASSO 算法进行变量筛选，共筛出性别、疾病分期、结外受累数目、生发中心 B 细胞样型（germinal center B-cell-like，GCB）、不良反应数量、乳酸脱氢酶（lactate dehydrogenase，LDH）、β_2-*MG*、肿瘤长度等 19 个变量，根据文献[13-14]及临床医师意见，Karnofsky 功能状态评分标准（Karnofsky performance status，KPS）得分、国际预后指数（international prognostic index，IPI）评分也与 DLBCL 患者复发相关，因此将其与 LASSO 算法筛选出的变量一起用于 DLBCL 复发模型的构建。各变量赋值情况见表 1。

方法与原理

1. 类别不平衡

本研究纳入的 445 例病例中，两年内复发的有 120 人，数据不平衡率为 2.7，存在类别不平衡。若无视数据的不平衡，会降低算法的性能[15]。而且，本研究拟构建的模型中，少数类为弥漫大 B 淋巴瘤患者的复发，因此对少数类的准确分类尤为重要，否则会延误最佳治疗时间。由于本文样本量小，数据平衡未采用欠采样方法，而是采用了过采样与混合采样。过采样方法是增多少数类样本数量使数据达到平衡。混合采样是同时使用过采样与欠采样，该方法的提出是为了解决模型过拟合和多数类样本间信息丢失问题。

以下是本文使用的数据平衡方法：

（1）合成少数类过采样技术（synthetic minority over-sampling technique，SMOTE）[16]

SMOTE 流程如下：首先计算原有少数类样本到同类别样本的欧式距离并找到距离最小的 k 个近邻样本；然后根据采样倍率 N%，在少数类样本的 k 个近邻样本中随机选择 N 个样本进行随机线性插补，生成新的少数类样本；最后合并新生成的少数类样本与原有少数类样本，使得数据集平衡。

（2）Borderline-SMOTE[17]

Borderline-SMOTE 分为 Borderline-1 SMOTE 和 Borderline-2 SMOTE。该算法首先对少数类样本分类，根据少数类样本的近邻样本的类别，将样本分为噪声样本、安全样本以及危险样本。然后利用危险样本进行新样本的生成，其中 Borderline-1 SMOTE 是在少数类样本中选择 k 个近邻样本得到生成新样本，Borderline-2 SMOTE 是在任意样本中选择的。最后合并新生成的少数类样本与原有少数类样本。

表 1　445 例患者临床特征及赋值情况

变量	赋值或单位	例数	构成比（%）
性别	1：男	233	52.36
	2：女	212	47.64
分期	1：Ⅰ期	52	11.69
	2：Ⅱ期	147	33.03
	3：Ⅲ期	92	20.67
	4：Ⅳ期	154	34.61
结外受累数目	0	253	56.85
	1	57	12.81
	2	53	11.91
	3	32	7.19
	4	18	4.04
	5	14	3.15
	6	15	3.37
	7	3	0.67
GCB	1：是	169	37.98
	0：否	276	62.02
IPI≥3	1：是	81	18.20
	0：否	364	81.80
KPS 得分	1：≥80	319	71.69
	0：<80	126	28.31
不良反应数量	0	140	31.46
	1	116	26.07
	2	110	24.72
	3	59	13.26
	4	19	4.27
	5	1	0.22
LDH	1：异常	182	40.90
	0：正常	263	59.10
β_2-MG	1：异常	115	25.84
	0：正常	330	74.16
乙肝	1：阳性	47	10.56
	0：阴性	398	89.44
胃累及	1：是	66	14.83
	0：否	379	85.17
鼻累及	1：是	13	2.92
	0：否	432	97.08
腋下累及	1：是	52	11.69
	0：否	393	88.31
BCL6	1：异常	200	44.94
	0：正常	245	55.06
Ki-67>=80	1：异常	246	55.28
	0：正常	199	44.72
CD20	1：异常	382	85.84
	0：正常	63	14.16
CTOP-E	1：是	26	5.84
	0：否	419	94.16
R-CHOP	1：是	61	13.71
	0：否	384	86.29
R-CTOP	1：是	98	22.02
	0：否	347	77.98
其他用药	1：是	29	6.52
	0：否	416	93.48
肿瘤长度	cm		

注：肿瘤长度是连续型变量

(3)自适应综合过采样(adaptive synthetic sampling,ADASYN)[18]

ADASYN 为自适应合成抽样方法,该算法通过对不同少数类样本赋予不同权重,计算生成相应的新样本。若少数类近邻样本中同一类别样本数量少,则赋予权重高,待合成样本数量相应多。

(4)SMOTE+Tomek Links[19]

SMOTE+Tomek Links 由 SMOTE 和 Tomek Links 组成。Tomek Links 是欠采样方法,若某一样本的最近邻为其他类别样本,则这两个样本组成一对 Tomek Links。该方法用 SMOTE 过采样生成少数类样本后,通过移除 Tomek Links 来剔除噪声点或者边界点,使得任一样本的最近邻为同一类别样本。

2. 分类模型

(1)单一分类器模型

logistic 回归是经典分类算法,由于该算法简单而有效,被广泛应用于分类问题。通过逻辑回归函数计算新样本的后验概率,依据概率值的大小对样本进行分类。

支持向量机(support vector machine,SVM)是经典的分类算法,对于非线性程度大、小样本、高维模式下的分类问题十分有效。该算法原理是寻找一个最大边际超平面,这个超平面能够有效提升分类效果的同时,使算法也具有较好的泛化性能。

(2)简单集成方法

随机森林(random forest,RF)是以决策树为基分类器的集成学习算法。其基本思想是利用 Bootstrap 方法从原始训练集中有放回地随机抽取 K 个子样本,然后对 K 个子样本分别建立决策树模型,对不同决策树模型得到的预测结果,使用多数投票法得到最终分类结果。

(3)多核学习

原始空间中的非线性样本,可以通过选择不同的核函数将其映射到新特征空间中,使其线性可分。二分类问题中,给定训练样本 $\{(x_i,y_i)\}_{i=1}^{N}$,$y_i \in \{+1,-1\}$,可以通过核函数映射:$x_i \to \Phi(x_i)$,得到 $\{(\Phi(x_i),y_i)\}_{i=1}^{N}, y_i \in \{+1,-1\}$。多核学习是将多个或多种核函数组合起来替代单一的核函数,它可以有效处理含有异构信息或分布不规则的数据[20]。在多核学习中,最重要的问题是核函数的组合权重。

$$k(\Phi(x_i),y_i) = \sum_{l=1}^{K} \beta_l k_l(\Phi(x_i),y_i)$$

其中 $\beta_l \geqslant 0$,K 为选取核函数的数目,核函数 $k_l(\Phi(x_i),y_i)$ 对应某一特性空间中的点积。因此,只有满足内积形式的分类器才能采用隐性核映射进行核化。

本文中用作对比的多核学习模型是利用固定准则的多核学习(rule-based MKL,RBMKL)[21]。

(4)多经验核分类器

采用经验核映射构建经验特征空间的多核学习被称为多经验核分类器。使用经验核映射,大多数算法可以直接被内核化,这样分析经验特征空间的结构更加容易。经验核映射将训练样本由显性形式的核函数从输入空间映射到高维特征空间。经验核映射表示为:

$$\tilde{\Phi}^e(x) = \tilde{\Lambda}^{-1/2}\tilde{Q}^T [\ker(x,x_1),\cdots,\ker(x,x_N)]^T$$

其中 $\tilde{\Lambda}$ 是 $N\times N$ 个对角矩阵,$\tilde{Q}$ 是包含相应特征向量的 $N\times N$ 矩阵,$\ker(x_i,x_j)$ 表示核矩阵的第 i 行和第 j 列元素,x_i 和 x_j 为训练样本,$i,j=1,\cdots,N$。传统的多经验核处理问题存在如下局限:①构造核空间时有巨大的计算复杂度;②由于采用基于梯度的优化方法,需重复迭代多次,故而需要大量的训练时间,而且基于梯度的优化方法无法保证分类器的分类性能。

(5)多随机经验核分类器

多随机经验核分类器采用了基于随机投影技术的随机经验核映射,以较低的计算复杂性以及较快的计算速度将样本映射到多个低维的经验特征空间。即给定训练样本 $\{(x_i,y_i)\}_{i=1}^{N}, y_i \in \{1,\cdots,c\}$,随机选择 m 个随机子集 $\{Sub_1,Sub_2,\cdots,Sub_M\}$,每个子集包含 P($P<N$)个不同样本。对每一个随机子集,构造相应的随机经验核映射,将其映射到相应的特征空间。随机经验核映射 $\tilde{\Phi}_l^e(x)$ 表示为:

$$\tilde{\Phi}_l^e(x) = \tilde{\Lambda}_l^{-\frac{1}{2}}\tilde{Q}_l^T [\ker_l(x,x_l^1),\ker_l(x,x_l^2),\cdots,\ker_l(x,x_l^P)]^T$$

然后,通过采用最小范数最小二乘法和 Moore-Penrose 广义逆解优化分类器的增强权重向量。也就是说,MREKLM 用解析法优化最优决策平台得到最终的分类器。

3. 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(accuracy)、精确率(precision)、召回率(recall)、$f1$-Score 评估模型的整体分类性能,即区分度,使用 brier score 评价它们的校准性能。

$$accuracy = \frac{\text{正确分类数}}{\text{样本总数}} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$precision = \frac{\text{正确分类的少数类个数}}{\text{预测为少数类的总数}} = \frac{TP}{TP+FP}$$

$$recall = \frac{\text{正确分类的少数类个数}}{\text{真实为正类的样本数}} = \frac{TP}{TP+FN}$$

$$f1 = \frac{2\times precision\times recall}{precision+recall}$$

$$BS = \frac{1}{T}\sum_{i=1}^{T}(e_i-o_i)^2$$

其中,TP 为真阳性;TN 为真阴性;FP 为假阳性;FN 为假阴性;e_i 为事件预测的概率;o_i 为事件实际的类别;T 为预测的数量。

结　果

目标模型与对比模型的 5 折交叉验证结果均值如表 2 所示。

由表 2 可知，在多个模型中采用 SMOTE Tomek Links+MREKLM 的模型（accuracy = 0.89，precision = 0.87，recall = 0.92，f1-Score = 0.89，brier score = 0.11）取得了最优的分类性能。其中区分度高表明了对 DL-BCL 患者两年内是否复发的分类准确率优，而低的校准度表明了模型总体表现好，即通过模型预测是否复发与实际是否复发一致。对比同一不平衡数据处理方式，可以发现，使用单一分类器构建的模型，性能最差；集成模型提高了一定的分类准确性，但结果不易解释，且分类性能没有基于多核的算法高，无论是传统多核，还是经验多核；在多核算法中，MREKLM 性能优于其他多核模型，与 Fan[10] 做的验证实验结果一致。

表 2　各模型性能指标对比

分类器	数据平衡方法	accuracy	precision	recall	f1-score	BS
logistic	无	0.72	0.47	0.24	0.72	0.19
	SMOTE	0.66	0.66	0.67	0.66	0.21
	Borderline-1 SMOTE	0.69	0.67	0.72	0.69	0.20
	Borderline-2 SMOTE	0.65	0.65	0.66	0.65	0.22
	ADASYN	0.68	0.67	0.70	0.68	0.21
	SMOTE+Tomek Links	0.68	0.68	0.70	0.68	0.20
SVM	无	0.69	0.28	0.08	0.70	0.19
	SMOTE	0.76	0.73	0.80	0.76	0.16
	Borderline-1 SMOTE	0.79	0.75	0.88	0.79	0.15
	Borderline-2 SMOTE	0.76	0.73	0.83	0.76	0.16
	ADASYN	0.74	0.71	0.83	0.74	0.17
	SMOTE+Tomek Links	0.77	0.74	0.83	0.77	0.16
RF	无	0.72	0.53	0.18	0.72	0.19
	SMOTE	0.84	0.88	0.80	0.84	0.12
	Borderline-1 SMOTE	0.84	0.86	0.80	0.84	0.12
	Borderline-2 SMOTE	0.82	0.86	0.77	0.82	0.13
	ADASYN	0.83	0.84	0.81	0.83	0.13
	SMOTE+Tomek Links	0.83	0.85	0.79	0.83	0.12
RBMKL	无	0.71	0.44	0.18	0.24	0.29
	SMOTE	0.85	0.84	0.87	0.85	0.15
	Borderline-1 SMOTE	0.86	0.84	0.88	0.86	0.14
	Borderline-2 SMOTE	0.86	0.84	0.88	0.86	0.14
	ADASYN	0.85	0.83	0.90	0.86	0.15
	SMOTE+Tomek Links	0.86	0.85	0.88	0.87	0.14
MEKL	无	0.73	0.51	0.25	0.33	0.27
	SMOTE	0.87	0.84	0.91	0.87	0.13
	Borderline-1 SMOTE	0.86	0.85	0.91	0.88	0.12
	Borderline-2 SMOTE	0.87	0.84	0.88	0.86	0.14
	ADASYN	0.87	0.84	0.92	0.88	0.13
	SMOTE+Tomek Links	0.87	0.86	0.89	0.87	0.13
MREKLM	无	0.74	0.70	0.10	0.16	0.26
	SMOTE	0.88	0.86	0.91	0.88	0.12
	Borderline-1 SMOTE	0.88	0.85	**0.93**	**0.89**	0.12
	Borderline-2 SMOTE	0.87	0.86	0.89	0.87	0.13
	ADASYN	0.88	0.84	**0.93**	**0.89**	0.12
	SMOTE+Tomek Links	**0.89**	**0.87**	0.92	**0.89**	**0.11**

注：加粗结果表示在对应类别中模型性能最优。其中，logistic 代表 logistic 回归；SVM 代表支持向量机；RF 代表随机森林。

从每个模型采样前后对比可知，未对数据进行不平衡处理时，模型的召回率都极低，意味着对复发患者的分类准确率低，不符合我们构建模型的初衷。在对数据进行不平衡处理后，模型的召回率大幅提升，证明对数据进行不平衡处理是极有意义的。在同一分类器中，SMOTE+Tomek links 处理不平衡数据的算法分类准确率取得了最优。

讨 论

本研究表明,处理不平衡数据和应用多经验核分类器构建的模型有效提高了预测精度。王梅英等[22]采用 SMOTE 算法有效解决化疗肿瘤患者下呼吸道感染中数据不平衡导致的预测误差。Liu 等[23]将 SMOTE+Tomek links 应用于 28 种不同癌症类型癌症基因组图谱中 DNA 甲基化数据中,结果表明,SMOTE +Tomek links 处理不平衡医学数据的方法是十分有效的。汪琳琳等[24]提出基于自步学习的多经验核映射集成分类器,应用于乳腺癌的辅助诊断;Fei 等[25]将基于投影参数转移的稀疏多经验核分类器应用于脑疾病的诊断,均有效提升了分类性能。

本文在 DLBCL 的实际数据集上,基于五种类别不平衡处理以及六种分类器构建模型,对 DLBCL 患者两年内复发进行了预测。解决数据类别不平衡问题后,模型对 DLBCL 复发患者的分类准确率有明显提升。其中 SMOTE Tomek links 的使用,既降低了只进行过采样的过拟合风险,又解决了只进行欠采样可能导致的信息损失问题。使用多随机经验核分类器,是一种使用显性核函数将样本从输入空间随机映射到核空间的方法,分类性能提高的同时降低了计算复杂度。

本研究的不足之处在于本研究数据来源单一。目前研究证明,使用来自多个模态的数据会提高分类性能。本文仅利用了收集到的结构化信息,未充分利用 CT 图像/PET-CT 等非结构化的信息。下一步我们拟收集非结构化信息,从多个视图出发,构建多视图多核模型来提高模型性能。

参 考 文 献

[1] Sehn LH, Salles G. Diffuse large B-cell lymphoma. New England Journal of Medicine, 2021, 384(9): 842-858.

[2] 张静, 顾岩, 吴雪, 等. 利妥昔单抗联合 CHOP/EPOCH 方案治疗弥漫大 B 细胞淋巴瘤患者的难治复发相关因素分析. 中国实验血液学杂志, 2020, 28(6): 1912-1918.

[3] Coiffier B, Sarkozy C. Diffuse large B-cell lymphoma: R-CHOP failure—what to do? Hematology 2014, the American Society of Hematology Education Program Book, 2016, 2016(1): 366-378.

[4] Zou Q, Xie S, Lin Z, et al. Finding the best classification threshold in imbalanced classification. Big Data Research, 2016, 5: 2-8.

[5] Tao M, Song T, Du W, et al. Classifying breast cancer subtypes using multiple kernel learning based on omics data. Genes, 2019, 10(3): 200.

[6] Wilson CM, Li K, Yu X, et al. Multiple-kernel learning for genomic data mining and prediction. BMC bioinformatics, 2019, 20(1): 1-7.

[7] Shanka K, Lakshmanaprabu SK, Gupta D, et al. Optimal feature-based multi-kernel SVM approach for thyroid disease classification. The journal of supercomputing, 2020, 76(2): 1128-1143.

[8] Arriaga RI, Vempala S. An algorithmic theory of learning: Robust concepts and random projection. Machine learning, 2006, 63(2): 161-182.

[9] Xiong H, Swamy MNS, Ahmad MO. Optimizing the kernel in the empirical feature space. IEEE transactions on neural networks, 2005, 16(2): 460-474.

[10] Fan Q, Wang Z, Zha H, et al. MREKLM: a fast multiple empirical kernel learning machine. Pattern Recognition, 2017, 61: 197-209.

[11] 李军民, 管忠震, 沈悌, 等. 中国弥漫大 B 细胞淋巴瘤诊断与治疗指南(2013 年版). 中华血液学杂志, 2013, 34(9): 816-819.

[12] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 267-288.

[13] Fan S, Zhao Z, Yu H, et al. Applying probability calibration to ensemble methods to predict 2-year mortality in patients with DLBCL. BMC Medical Informatics and Decision Making, 2021, 21(1): 1-12.

[14] Vardhana SA, Sauter CS, Matasar MJ, et al. Outcomes of primary refractory diffuse large B-cell lymphoma(DLBCL) treated with salvage chemotherapy and intention to transplant in the rituximab era. British journal of haematology, 2017, 176(4): 591-599.

[15] Singh A, Purohit A. A Survey on Methods for Solving Data Imbalance Problem for Classification. International Journal of Computer Applications, 2015, 127(15): 37-41.

[16] Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: Synthetic Minority Over-samplingTechnique. Journal of Artificial Intelligence Research, 2002, 16(1): 321-357.

[17] Han H, Wang WY, Mao BH. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. International conference on intelligent computing, Springer, Berlin, Heidelberg, 2005: 878-887.

[18] He H, Bai Y, Garcia EA, et al. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. 2008 IEEE International Joint Conference on Neural Networks, 2008: 1322-1328.

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

[20] Bach FR, Lanckriet GRG, Jordan MI. Multiple kernel learning, conic duality, and the SMO algorithm. Proceedings of the twenty-first international conference on Machine learning. 2004: 6.

[21] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press, 2000.

[22] 王梅英, 杨敏, 刘佳微, 等. 基于 SMOTE 算法的化疗肿瘤患者下呼吸道感染预警模型构建. 中国感染控制杂志, 2021, 20(12): 1094-1101.

[23] Liu C, Wu J, Mirador L, et al. Classifying dna methylation imbalance data in cancer risk prediction using smote and tomek link methods.International Conference of Pioneering Computer Scientists, Engineers and Educators. Springer, Singapore, 2018: 1-9.

[24] 汪琳琳, 沈璐, 施俊等. 基于自步学习的多经验核映射集成分类器在乳腺癌超声计算机辅助诊断上的应用. 生物医学工程学杂志, 2021, 38(1): 30-38.

[25] Fei X, Wang J, Ying S, et al. Projective parameter transfer based sparse multiple empirical kernel learning Machine for diagnosis of brain disease. Neurocomputing, 2020, 413: 271-283.

(责任编辑:张 悦)