基于正则化多项式回归的癌症患者免疫检查点阻断响应预测

王小玉[1],奚晨曦[2],李钧涛[2]

(1.郑州工商学院 基础教学部,郑州 451400;
2.河南师范大学 数学与信息科学学院,河南 新乡 453007)

摘　要:探究特征之间的非线性相互作用关系,构建了用于预测癌症患者免疫检查点阻断响应的正则化多项式逻辑斯蒂回归模型.无进展生存期和总生存期的 Kaplan-Meier 曲线被用来筛选单个特征和联合特征,并据此构建了二次多项式判别函数.结合多项式负对数损失函数和弹性网络惩罚函数,提出了一种正则化多项式逻辑斯蒂回归模型,并通过特征拓维将其转化为线性模型求解.在泛癌、黑色素瘤、非小细胞肺癌和其他癌症数据集上与其他 6 种方法进行比较,结果表明所提方法取得了更高的免疫检查点阻断响应精度、F_1 分数和 AUC 值.

关键词:癌症;正则化;多项式回归;免疫检查点阻断

中图分类号:O224　　**文献标志码**:A　　**文章编号**:1000-2367(2024)04-0094-07

癌症免疫疗法利用免疫系统来识别和消除癌症,是当前癌症诊疗的有效方法[1].免疫检查点阻断(immune checkpoint blockade,ICB)可阻止在免疫细胞上表达的调节途径,以改善抗肿瘤免疫响应,现已成为最有效的一种癌症免疫疗法[2].检查点分子的细胞毒性 T 淋巴细胞相关蛋白 4(cytotoxic T-lymphocyte-associated protein 4,CTLA 4)和程序性细胞死亡 1 受体(programmed cell death 1 receptor,PD-1)/程序性细胞凋亡受体配体 1(programmed cell death receptor ligand 1,PD-L1)抑制剂可以成功提高包括黑色素瘤和非小细胞肺癌在内的癌症晚期患者的生存率[3].然而,只有极少数的癌症患者对 ICB 有响应,大多数患者并没有获得临床益处[4],这就使得预测癌症患者的 ICB 响应成为当前生物医学、数学和人工智能等领域的共同关注焦点.

从生物学的角度看,ICB 响应由多种因素(特征)的共同作用引起[5-9].血液中性粒细胞与淋巴细胞比值和嗜酸性粒细胞水平的变化与免疫检查点阻断治疗响应相关[5].不同年龄的恶性肿瘤患者对药物的耐受程度不同,进而影响 ICB 响应[6].微卫星不稳定性状态与 ICB 的高响应率有关,可以作为预后和预测标志物,并且与年龄和药物类型共同影响治疗结果[7-8].在结直肠癌患者的 ICB 治疗中,PD-L1 与微卫星不稳定性状态、C 反应蛋白以及血液中性粒细胞与淋巴细胞比值共同影响 ICB 疗效[9].如何利用生物学特征去预测癌症患者的 ICB 响应是一个挑战.

统计机器学习的方法已被成功应用于癌症患者的 ICB 响应预测[10-13].ANAGNOSTOU 等[10]整合了校正的肿瘤突变负荷、肿瘤中激活受体酪氨酸激酶、吸烟相关的突变特征和人类白细胞抗原状态等特征,提出了用于癌症病人的 ICB 响应预测的综合多变量模型.WANG 等[11]通过整合生物通路数据和单细胞测序数据,

收稿日期:2023-07-31;**修回日期**:2023-09-07.

基金项目:国家自然科学基金(61203293);河南省科技攻关计划(242102211023).

作者简介(通信作者):李钧涛(1978－),男,河南社旗人,河南师范大学教授,博士,研究方向为数据驱动建模与控制,统计学习.E-mail:juntaolimail@126.com.

引用本文:王小玉,奚晨曦,李钧涛.基于正则化多项式回归的癌症患者免疫检查点阻断响应预测[J].河南师范大学学报(自然科学版),2024,52(4):94-100.(Wang Xiaoyu,Xi Chenxi,Li Juntao.Prediction of immune checkpoint blockade response of cancer patients based on regularized polynomial regression[J].Journal of Henan Normal University(Natural Science Edition),2024,52(4):94-100.DOI:10.16366/j.cnki.1000-2367.2023.07.31.0001.)

在转移性黑色素瘤中构建了 11 个免疫细胞簇的调控网络,并基于调控网络中的配体和受体的逻辑斯蒂回归模型来预测 ICB 治疗响应.SUNG 等[12]通过 LASSO 方法提取特征基因并对其进行基因本体分析,结合随机森林(random forest,RF)、朴素贝叶斯(naive Bayes,NB)、神经网络和支持向量机(support vector machine,SVM)4 种机器学习算法为转录组测序数据构建模型,以预测胃癌患者的 ICB 响应.CHOWELL 等[13]通过整合与免疫治疗疗效相关的多种生物学特征,开发了一个具有 16 个输入特征的集成学习 RF 分类器(以下称为 RF16)以改进免疫检查点阻断在多种癌症类型中的预测.

然而,上述的 ICB 响应预测方法[10-13]没有考虑特征之间的非线性相互作用关系.本文利用 Kaplan-Meier(K-M)曲线筛选单个特征和联合特征,并据此构建正则化多项式逻辑斯蒂回归模型,以提高癌症患者的 ICB 响应预测精度.

1　问题陈述

本文旨在提高癌症患者的 ICB 响应预测精度,为此收集了具有 16 种不同癌症类型的 1 479 名接受 ICB 治疗患者的完整临床数据集(泛癌数据集)[13].这些患者接受 PD-1/PD-L1 抑制剂、CTLA-4 阻断剂或两种免疫治疗药物的联合治疗.在泛癌数据集中数据集类别不平衡,有 409 名患者的肿瘤对免疫治疗有响应,1 070 名患者的肿瘤对免疫治疗没有响应.完全缓解或部分缓解的患者被归类为响应者(responders,R);经历疾病稳定或疾病进展的患者被归类为无响应者(non-responders,NR).进一步将泛癌数据集划分为 3 个子数据集,分别记为黑色素瘤、非小细胞肺癌和其他癌症.这 4 个数据集包含的样本数量与所含癌症类别见表 1.所有数据集均包括免疫治疗药物、微卫星不稳定性状态、患者在免疫治疗前是否接受化疗、拷贝数变异分数、HLA-I 进化分化评分、肿瘤突变负荷、肿瘤分期、HLA-I25 杂合性缺失状态、体重指数、性别、血液中性粒细胞与淋巴细胞比值、年龄以及血液中白蛋白、血小板和血红蛋白的水平等 15 个特征.为消除不同数据之间的量级差异以及避免数据中的噪声和异常值对分析结果的影响,本文对数据进行了标准化处理.

表 1　接受 ICB 治疗患者的 4 个数据集描述

Tab. 1　Description of four datasets of patients receiving ICB treatment

数据集	总样本	响应样本	无响应样本	包含癌症类型
泛癌	1 479	409	1 070	膀胱癌、乳腺癌、食管癌、结直肠癌、黑色素瘤、子宫内膜癌、胃癌、头肝胆癌、间皮瘤、非小细胞肺癌、卵巢癌、胰腺癌、肾脏癌、肉瘤、小细胞肺癌、颈癌
黑色素瘤	186	86	100	黑色素瘤
非小细胞肺癌	538	145	393	非小细胞肺癌
其他癌症	755	178	577	膀胱癌、乳腺癌、食管癌、结直肠癌、子宫内膜癌、胃癌、头肝胆癌、间皮瘤、卵巢癌、胰腺癌、肾脏癌、肉瘤、小细胞肺癌、颈癌

本文使用与参考文献[13]相同的训练数据集与测试数据集,即对每个数据集按癌症类型随机取 4/5 的样本作为训练数据集,而剩余的 1/5 的样本作为测试数据集.令 $\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$ 代表训练数据集,在泛癌数据集,黑色素瘤数据集,非小细胞肺癌数据集和其他癌症数据集中 n 分别取 1 184,149,430 和 605,(x_i,y_i) 的取值随数据集的不同而变化,其中 $x_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(15)})$ 代表第 i 个患者 15 个特征的表达值,$x^{(1)},x^{(2)},\cdots,x^{(15)}$ 分别代表免疫治疗药物、微卫星不稳定性状态、体重指数、肿瘤突变负荷、肿瘤分期、HLA-I 进化分化评分、性别、患者在免疫治疗前是否接受化疗、血液中性粒细胞与淋巴细胞比值、年龄以及血液中白蛋白、血小板、血红蛋白的水平、拷贝数变异分数与 HLA-I25 杂合性缺失状态.y_i 表示与 x_i 对应的类别标签,如果第 i 个患者对 ICB 治疗有响应,则 $y_i=1$,否则,$y_i=0$.

从机器学习的角度来看,ICB 响应预测可以转化为二分类问题,即找到决策函数

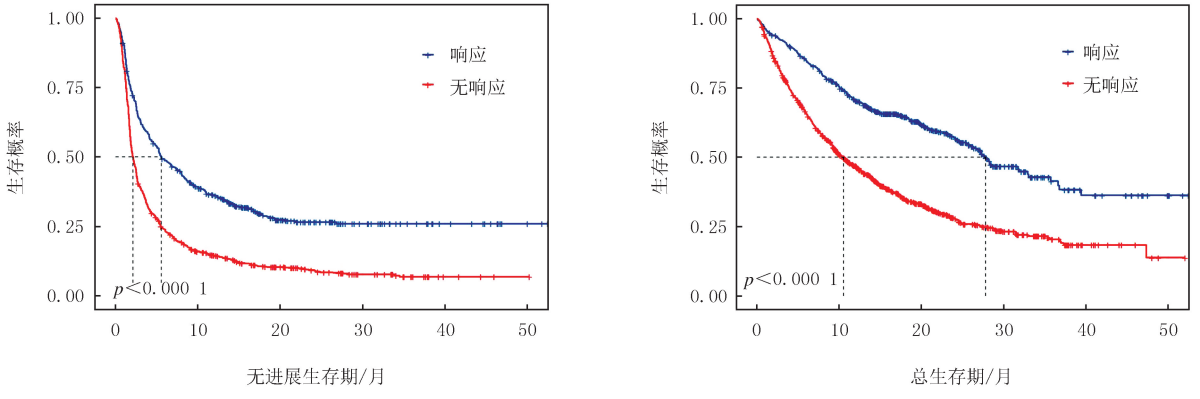
$$D(x)=\begin{cases}1, d(x)\geq\theta,\\ 0, d(x)<\theta,\end{cases}\tag{1}$$

来预测测试数据集中的类别标签,其中 $d(x)$ 为判别函数,θ 为阈值.

2 主要结果

2.1 非线性决策函数

为了初步筛选特征并探究非线性特征对患者生存期的影响,本文基于无进展生存期和总生存期数据对15个特征分别进行K-M生存分析,在无进展生存期和总生存期上,拷贝数变异分数与HLA-I25杂合性缺失状态均不可以显著区分响应组与非响应组存活概率,因此将这两个特征剔除.药物和微卫星不稳定性状态与其他特征共同影响ICB治疗结果[6-9],因此本文进一步分析药物和微卫星不稳定性状态分别与其余13个特征联合的非线性特征的K-M生存曲线.联合的非线性特征即特征之间对应表达值的乘积,简称联合特征.图1为联合特征的K-M生存曲线示例,根据图1(a)与图1(b),微卫星不稳定性状态和治疗前是否接受化疗的联合特征在总生存期和无进展生存期上都能够显著区分响应组与非响应组的存活概率.此外,本文分析了所有联合特征的K-M生存曲线,发现几乎所有联合特征都能显著区分响应组与非响应组的存活概率.因此,药物和微卫星不稳定性状态与其他13个特征的联合特征可以作为ICB响应预测的特征.



(a)微卫星不稳定性状态和治疗前是否化疗的联合特征的无进展生存期K-M生存曲线

(b)微卫星不稳定性状态和治疗前是否化疗的联合特征的总生存期K-M生存曲线

图1 联合特征的K-M生存曲线分析

Fig.1 The K-M survival curve analysis of combined features

通过上述分析,构建如下二次多项式函数作为判别函数

$$d(x)=b+\sum_{i=1}^{13}w^{(i)}x^{(i)}+\sum_{i=1}^{2}\sum_{j=i+1}^{13}w^{(i,j)}x^{(i)}x^{(j)},\tag{2}$$

其中:b 为偏移量,$w^{(i)}$ 和 $w^{(i,j)}$ 为回归系数,$x^{(i)}$ 表示患者的第 i 个特征的对应值.

2.2 正则化多项式逻辑斯蒂回归

为了求解判别函数 $d(x)$,本文提出了正则化多项式逻辑斯蒂回归模型(regularized polynomial logistic regression,RPLR)

$$\arg\min_{w,b}\{L(w,b)+\lambda P_{\alpha}(w)\},\tag{3}$$

其中:

$$L(w,b)=-\frac{1}{n}\sum_{i=1}^{n}\{y_i[\sum_{j=1}^{13}w^{(j)}x^{(j)}+\sum_{j=1}^{2}\sum_{k=j+1}^{13}w^{(j,k)}x^{(j)}x^{(k)}+b]-\ln(1+\mathrm{e}^{\sum_{j=1}^{13}w^{(j)}x^{(j)}+\sum_{j=1}^{2}\sum_{k=j+1}^{13}w^{(j,k)}x^{(j)}x^{(k)}+b})\},$$

$$P_{\alpha}(w)=(1-\alpha)\frac{1}{2}[\sum_{i=1}^{13}(w^{(i)})^2+\sum_{i=1}^{2}\sum_{j=i+1}^{13}(w^{(i,j)})^2]+\alpha[\sum_{i=1}^{13}|w^{(i)}|+\sum_{i=1}^{2}\sum_{j=i+1}^{13}|w^{(i,j)}|],$$

分别为多项式负对数损失函数和弹性网络惩罚函数,α,λ 为正则化参数,$\alpha\in(0,1)$.

为求解最优化问题(3),令 $x^{(14)}=x^{(1)}x^{(2)}$,$x^{(15)}=x^{(1)}x^{(3)}$,…,$x^{(25)}=x^{(1)}x^{(12)}$,$x^{(26)}=x^{(2)}x^{(3)}$,$x^{(27)}=$

$x^{(2)}x^{(4)},\cdots,x^{(36)}=x^{(2)}x^{(13)}$. 在拓维特征空间上式(2)可表示为

$$d(x)=d(\bar{x})=b+\bar{w}^{\mathrm{T}}\bar{x}, \tag{4}$$

这里 $\bar{x}=(x^{(1)},x^{(2)},\cdots,x^{(36)})^{\mathrm{T}}$,$\bar{w}=(w^{(1)},w^{(2)},\cdots,w^{(36)})^{\mathrm{T}}$. 进一步,13 维特征空间上的非线性最优化问题(3)可转化为 36 维特征空间上的带有弹性网络惩罚的逻辑斯蒂回归

$$\arg\min_{\bar{w},b}\{L(\bar{w},b)+\lambda P_{\alpha}(\bar{w})\}, \tag{5}$$

其中:

$$L(\bar{w},b)=-\frac{1}{n}\sum_{i=1}^{n}[y_i(\bar{w}^{\mathrm{T}}\bar{x}_i+b)-\ln(1+\mathrm{e}^{\bar{w}^{\mathrm{T}}\bar{x}_i+b})],$$

$$P_{\alpha}(\bar{w})=(1-\alpha)\frac{1}{2}\|\bar{w}\|_{l_2}^{2}+\alpha\|\bar{w}\|_{l_1}=(1-\alpha)\frac{1}{2}\sum_{i=1}^{36}(w^{(i)})^{2}+\alpha\sum_{i=1}^{36}|w^{(i)}|,$$

分别表示负对数似然函数和弹性网络惩罚函数[14],$\bar{x}_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(36)})^{\mathrm{T}}$ 表示第 i 个患者的 36 个特征的表达值,α,λ 为正则化参数,$\alpha\in(0,1)$,$w^{(i)}$ 为第 i 个特征对应的回归系数,b 为偏移量.

最优化问题(5)是一个广义线性回归模型,可借助于 R 工具包 glmnet 进行求解.值得注意的是,同时优化模型参数和将会带来巨大的计算负担.因此,本文提前将 α 固定为 0.01,0.05,0.10,0.20,0.30,0.40,0.50,0.60,0.70,0.80,0.90,0.95,并通过对 α 的每个数值进行十折交叉验证来选择最优的 λ.考虑到数据的不平衡性,本文预先指定一个序列 $\theta=\{0.5+0.01n\mid n\in\mathbf{N},0\leq n<50\}$,在不同的数据集上对 θ 进行遍历以寻求最佳阈值 $\hat{\theta}$. 求解最优化问题(5)的算法伪代码展示在算法 1.

算法 1　RPLR 的算法伪代码

输入　数据集 $D=[\boldsymbol{X},\boldsymbol{Y}]$

　　特征矩阵 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$

　　样本标签 $\boldsymbol{Y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$

输出　系数 $\bar{w}$,b

　　测试数据集上的预测准确率:精度

1.根据式(2)得到 $\bar{D}=\{\bar{\boldsymbol{X}},\boldsymbol{Y}\}$

2.将 $\bar{\boldsymbol{X}}$ 划分为 $\bar{\boldsymbol{X}}_{\mathrm{train}}$,$\bar{\boldsymbol{X}}_{\mathrm{test}}$

　将 $\boldsymbol{Y}$ 划分为 $\boldsymbol{Y}_{\mathrm{train}}$,$\boldsymbol{Y}_{\mathrm{test}}$

3.for each $\alpha=\{0.01,0.05,0.10,0.20,0.30,0.40,0.50,0.60,0.70,0.80,0.90,0.95\}$ do

4.基于 R 语言工具包 glmnet 在训练数据集上进行 10 折交叉验证得到 λ

5.end for

6.确定最优参数对$(\hat{\alpha},\hat{\lambda})$

7.for each 阈值 $\theta=\{0.5+0.01n\mid n\in[\mathbf{N},0\leq n<50]\}$ do

8.在测试数据集上根据式(1)和式(4)预测免疫检查点阻断响应 $\bar{\boldsymbol{Y}}_{\mathrm{test}}$

9.end for

10.确定最佳阈值 $\hat{\theta}$

11.在阈值 $\hat{\theta}$ 和参数对$(\hat{\alpha},\hat{\lambda})$ 下,预测 $\bar{X}_{\mathrm{test}}$ 的标签

12.return $\bar{w}$,b,精度

3　实验

3.1　对比方法和评价指标

为了验证 RPLR 模型对癌症病人 ICB 响应预测的有效性,本文将其与 6 种方法进行对比.这 6 种方法分别为 RF16[13]、多项式逻辑斯蒂回归(multinomial logistic regression,MLR)[15]、脊回归[16]、SVM[17]、NB[18]和 LASSO 回归[19].本文采用第 2 节描述的算法求解 RPLR,采用 R 包 e1071 求解 SVM 和 NB,利用 R 包 glmnet 求解 LASSO 回归、脊回归和 MLR.

本文采用精度(J)和 F_1 分数(F_1)作为评价指标.

$$J=\frac{T_P+T_N}{T_P+F_P+T_N+F_N},F_1=\frac{2\times P\times R}{P+R},$$

其中 $P=T_P/(T_P+F_P)$,$R=T_P/(T_P+F_N)$,T_P,F_P,T_N 和 F_N 分别代表真正例,假正例,真反例和假反例.

3.2　实验结果

在泛癌、黑色素瘤、非小细胞肺癌和其他癌症数据集上将 RPLR 与其他 6 种方法进行比较.表 2 列出了 7 种方法在 4 个测试数据集上的精度,在每个测试数据集上 RPLR 都拥有最高的精度.在泛癌测试数据集

上,RPLR 的精度分别比 RF16,脊回归,LASSO 回归,MLR,SVM 和 NB 高出 2.4%,2.7%,4.1%,3.1%,3.4%和 5.1%.在黑色素瘤测试数据集上,RPLR 的精度分别比 RF16,脊回归,LASSO 回归,MLR,SVM 和 NB 高出 2.7%,10.8%,16.2%,8.1%,5.4%和 10.8%.在非小细胞肺癌测试数据集上,RPLR 的精度分别比 RF16,脊回归,LASSO 回归,MLR,SVM 和 NB 高出 2.8%,0.9%,2.8%,6.5%,4.6%和 55.6%.在其他癌症测试数据集上,RPLR 的精度分别比 RF16,脊回归,LASSO 回归,MLR,SVM 和 NB 高出 4.0%,6.7%,7.3%,6.7%,4.7%和 1.3%.

表 2 7 种方法在 4 个测试数据集上的精度

Tab. 2 Accuracy of seven methods on four test datasets

方法	泛癌	黑色素瘤	非小细胞肺癌	其他癌症	方法	泛癌	黑色素瘤	非小细胞肺癌	其他癌症
RPLR	0.772 9	0.729 7	0.814 8	0.773 3	MLR[15]	0.742 4	0.648 6	0.731 5	0.706 7
RF16[13]	0.749 2	0.702 7	0.787 0	0.733 3	SVM[17]	0.739 0	0.675 7	0.768 5	0.726 7
脊回归[16]	0.745 8	0.621 6	0.777 8	0.706 7	NB[18]	0.722 0	0.621 6	0.259 3	0.760 0
LASSO 回归[19]	0.732 2	0.567 6	0.787 0	0.700 0					

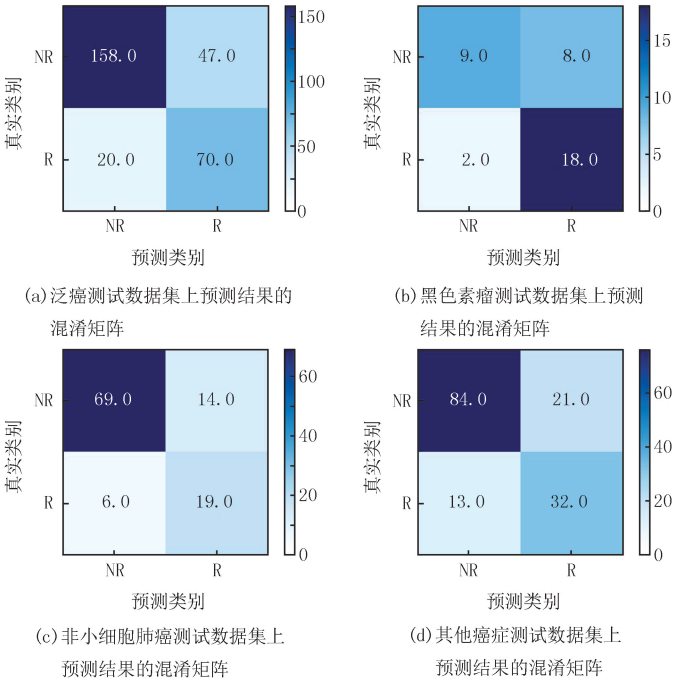
表 3 列出了 7 种方法分别在 4 个测试数据集上的 F_1.F_1 同时兼顾了分类模型的准确率和召回率,它的最大值为 1,最小值为 0,值越大意味着模型越好.结果表明,在每个测试数据集上 RPLR 的 F_1 都高于其他 6 种方法.在泛癌、黑色素瘤、非小细胞肺癌和其他癌症测试数据集上 RPLR 的 F_1 分别为 0.676 3,0.782 6,0.655 2和 0.65 3.可以看到,RPLR 的 F_1 在黑色素瘤测试数据集上表现最佳.

表 3 7 种方法在 4 个测试数据集上的 F_1

Tab. 3 F_1 of seven methods on four test datasets

方法	泛癌	黑色素瘤	非小细胞肺癌	其他癌症	方法	泛癌	黑色素瘤	非小细胞肺癌	其他癌症
RPLR	0.676 3	0.782 6	0.655 2	0.653 1	MLR[15]	0.366 7	0.580 6	0.292 7	0.266 7
RF16[13]	0.650 9	0.731 7	0.634 9	0.629 6	SVM[17]	0.363 6	0.625 0	0.074 1	0.349 2
脊回归[16]	0.347 8	0.533 3	0.200 0	0.214 3	NB[18]	0.453 3	0.631 6	0.384 6	0.550 0
LASSO 回归[19]	0.275 2	0.428 6	0.206 9	0.042 6					

为了更充分展现 RPLR 预测错误的情况发生的概率、类型和程度,图 2 展示了所提方法在 4 个测试数据集上的预测结果的混淆矩阵.混淆矩阵的色块颜色与数字对应,数字越大颜色越深.由图 2 得出,在 4 个测试数据集上预测结果的混淆矩阵主对角线的颜色深,并且非主对角线的颜色浅,表明该方法的预测性能较好.对于癌症患者来说,将 ICB 无响应判断为响应,认为付出的成本小.由图 2(a),图 2(b),图 2(c)和图 2(d),在泛癌、黑色素瘤、非小细胞肺癌和其他癌症测试数据集上 RPLR 在预测错误的样本中将响应预测为无响应的样本的个数分别为 20,2,6 和 13,分别占相应总测试样本数的 6.78%,5.41%,5.56% 和 8.67%.在 4 个测试数据集上 RPLR 在预测错误的样本中将响应的预测为无响应的样本是少数的,说明了 RPLR 在 ICB 响应预测上的合理性.



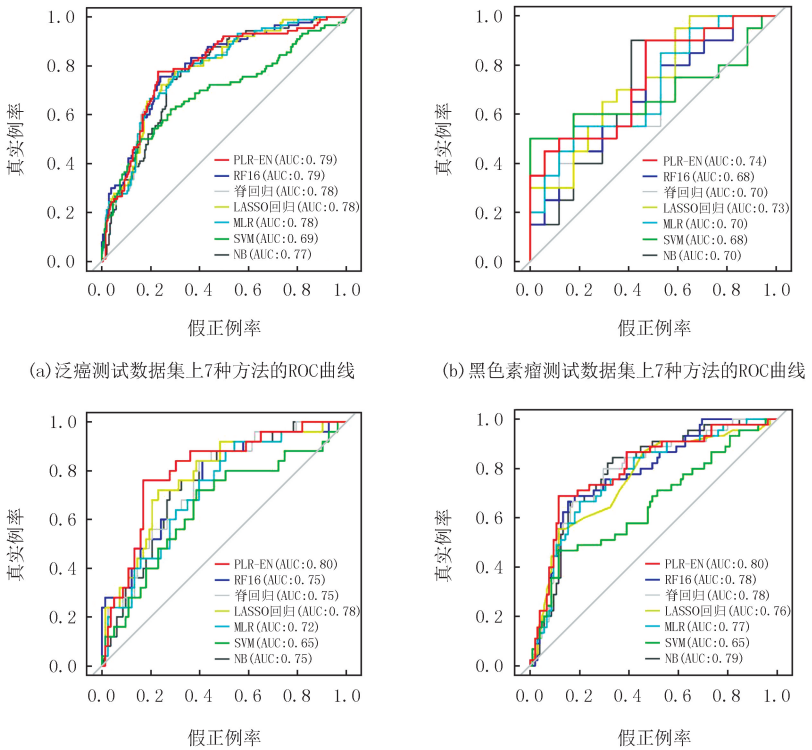
(a)泛癌测试数据集上预测结果的混淆矩阵 (b)黑色素瘤测试数据集上预测结果的混淆矩阵

(c)非小细胞肺癌测试数据集上预测结果的混淆矩阵 (d)其他癌症测试数据集上预测结果的混淆矩阵

图2 RPLR在4个测试数据集上预测结果的混淆矩阵

Fig. 2 Confusion matrix of RPLR prediction results on four test datasets

7 种方法在 4 个数据集上对应的受试者工作特征（receiver operating characteristic，ROC）曲线如图 3 所示.图 3 中不同颜色的曲线代表了不同方法的性能，并且 7 种方法在每个数据集中的 AUC 值被展示在图中.AUC 指 ROC 曲线下的面积，常被用来衡量模型的预测准确性.AUC 值的范围一般在 0.5 到 1.0 之间，越接近 1.0，意味着模型的性能越好.RPLR 在 4 个数据集上都获得了最高的 AUC 值，在泛癌测试数据集上虽然 RF16 与 RPLR 具有相同的 AUC 值，但 RPLR 的预测精度与 F_1 都高于 RF16，RPLR 表现更优.



(a)泛癌测试数据集上7种方法的ROC曲线　(b)黑色素瘤测试数据集上7种方法的ROC曲线

(c)非小细胞肺癌测试数据集上7种方法的ROC曲线　(d)其他癌症测试数据集上7种方法的ROC曲线

图3　7种方法在4个测试数据集上的ROC曲线

Fig. 3　ROC curves of seven methods on four test datasets

4　结　论

本文考虑特征之间的非线性相互作用关系，提出了用于预测癌症病人 ICB 响应的 RPLR 模型.根据 K-M 生存曲线，对单个特征和联合特征进行初步筛选，并证明联合特征可以作为 ICB 预测的特征.通过特征拓维，将 RPLR 模型转化为带有弹性网络惩罚的逻辑斯蒂回归模型进行求解.在泛癌、黑色素瘤、非小细胞肺癌和其他癌症数据集上与其他 6 种方法的对比实验结果表明，RPLR 在 7 种方法中表现出最高的预测精度，F_1 分数和 AUC 值.此外，RPLR 在预测错误的样本中只有少数样本将响应样本预测为无响应样本.

参　考　文　献

[1] DAGHER O K，SCHWAB R D，BROOKENS S K，et al.Advances in cancer immunotherapies[J].Cell，2023，186(8)：1814-1814.e1.

[2] CHEN Q，WANG C，CHEN G J，et al.Delivery strategies for immune checkpoint blockade[J].Advanced Healthcare Materials，2018，7(20)：1800424.

[3] KIM J，HONG J，LEE J，et al.Recent advances in tumor microenvironment-targeted nanomedicine delivery approaches to overcome limitations of immune checkpoint blockade-based immunotherapy[J].Journal of Controlled Release，2021，332：109-126.

[4] GANESAN S，MEHNERT J.Biomarkers for response to immune checkpoint blockade[J].Annual Review of Cancer Biology，2020，4：331-351.

[5] HWANG M，CANZONIERO J V，ROSNER S，et al.Peripheral blood immune cell dynamics reflect antitumor immune responses and predict clinical response to immunotherapy[J].Journal for Immunotherapy of Cancer，2022，10(6)：e004688.

[6] LI S，MUKHERJI R，PATEL S A，et al.Impact of age on immune checkpoint blockade tolerability across malignancies：a single institution review[J].Journal of Clinical Oncology，2018，36(15_suppl)：e15069.

[7] MACHERLA S，LAKS S，NAQASH A R，et al.Emerging role of immune checkpoint blockade in pancreatic cancer[J].International Journal of Molecular Sciences，2018，19(11)：3505.

[8] ANDREEV-DRAKHLIN A，SHAH A Y，ADRIAZOLA A C，et al.Efficacy of immune checkpoint blockade in patients with advanced upper tract urothelial cancer and mismatch repair deficiency or microsatellite instability(MSI)[J].Journal of Clinical Oncology，2021，39(6_suppl)：487.

[9] KONG P F,WANG J,SONG Z,et al.Circulating lymphocytes,PD-L1 expression on tumor-infiltrating lymphocytes,and survival of colorectal cancer patients with different mismatch repair gene status[J].Journal of Cancer,2019,10(7):1745-1754.

[10] ANAGNOSTOU V,NIKNAFS N,MARRONE K,et al.Multimodal genomic features predict outcome of immune checkpoint blockade in non-small-cell lung cancer[J].Nature Cancer,2020,1:99-111.

[11] WANG J W,LI F,XU Y J,et al.Dissecting immune cell stat regulation network reveals biomarkers to predict ICB therapy responders in melanoma[J].Journal of Translational Medicine,2021,19(1):296.

[12] SUNG J Y,CHEONG J H.Machine learning predictor of immune checkpoint blockade response in gastric cancer[J].Cancers,2022,14(13):3191.

[13] CHOWELL D,YOO S K,VALERO C,et al.Improved prediction of immune checkpoint blockade efficacy across multiple cancer types[J].Nature Biotechnology,2022,40:499-506.

[14] 王小玉,陈留院,刘云卿,等.基于弹性网络的大鼠肝再生关键基因选择[J].河南师范大学学报(自然科学版),2013,41(5):26-28.
WANG X Y,CHEN L Y,LIU Y Q,et al.Selection of the key genes for the rat liver regeneration via elastic net[J].Journal of Henan Normal University(Natural Science Edition),2013,41(5):26-28.

[15] YIN M,ZENG D Y,GAO J B,et al.Robust multinomial logistic regression based on RPCA[J].IEEE Journal of Selected Topics in Signal Processing,2018,12(6):1144-1154.

[16] GHOSH D.Penalized discriminant methods for the classification of tumors from gene expression data[J].Biometrics,2003,59(4):992-1000.

[17] 李钧涛,杨瑞峰,左红亮.统计机器学习研究[J].河南师范大学学报(自然科学版),2010,38(6):35-40.
LI J T,YANG R F,ZUO H L.Statistical machine learning[J].Journal of Henan Normal University(Natural Science Edition),2010,38(6):35-40.

[18] ZHANG H,JIANG L X,YU L J.Attribute and instance weighted naive Bayes[J].Pattern Recognition,2021,111:107674.

[19] 张艳丽,尤晓琳,强薇,等.基于 LASSO 的企业财务危机预警与关键指标选择[J].河南师范大学学报(自然科学版),2016,44(3):160-165.
ZHANG Y L,YOU X L,QIANG W,et al.LASSO-based early warning of enterprise's financial crisis and the selection of key indicators[J].Journal of Henan Normal University(Natural Science Edition),2016,44(3):160-165.

Prediction of immune checkpoint blockade response of cancer patients based on regularized polynomial regression

Wang Xiaoyu[1], Xi Chenxi[2], Li Juntao[2]

(1. Department of Basic Teaching, Zhengzhou Technology and Business University, Zhengzhou 451400, China;
2. College of Mathematics and Information Science, Henan Normal University, Xinxiang 453007, China)

Abstract: This paper explored the nonlinear interaction between features and constructed a regularized polynomial logistic regression model for predicting immune checkpoint blockade response in cancer patients. The Kaplan-Meier curves for progression free survival and overall survival were used to screen for individual and combined features, a quadratic polynomial discriminant function was constructed. Combining the polynomial negative logarithmic loss function and the elastic network penalty function, a regularized polynomial logistic regression model is proposed, and it is transformed into a linear model through feature extension. Compared with other six methods for the data sets of pan cancer, Melanoma, non-small cell lung cancer and other cancers, the proposed method has achieved higher blocking response accuracy, F_1 score and AUC value of immune checkpoints.

Keywords: cancer; regularization; polynomial regression; immune checkpoint blockade

[责任编校 陈留院 赵晓华]