文章编号:1671-6833(2025)06-0058-08

预测 ICI 治疗响应的凹惩罚 Logistic 回归模型

穆晓霞[1], 张红梅[2], 宋学坤[3], 李钧涛[4]

(1.河南师范大学 计算机与信息工程学院,河南 新乡 453007;2.东北林业大学 生命科学学院,黑龙江 哈尔滨 150006;3.河南中医药大学 信息技术学院,河南 郑州 450046;4.河南师范大学 数学与统计学院,河南 新乡 453007)

摘 要:为提升黑色素瘤患者对免疫检查点抑制剂(ICI)治疗响应的预测准确性,提出了一种整合批量 RNA 测序和单细胞 RNA 测序数据的新方法。首先,通过皮尔逊相关性分析构建患者-细胞相关性矩阵,采用 Louvain 算法对单细胞 RNA 测序数据进行细胞分群;其次利用 CellChat 工具量化细胞群在免疫响应相关通路中的重要性;最后,通过引入基于细胞间通信网络构建的细胞群重要性评价准则,并结合群极小极大凹惩罚,提出了二重群极小极大凹惩罚 Logistic 回归模型(DMCPLR)。在 GSE35640 数据集上的实验表明,DMCPLR 模型的预测准确率达到 80.18%,精确率、召回率和 $F1$ 分数分别为 82.24%,89.71%和 85.11%,显著优于包括 Lasso 回归和随机森林在内的 14 种对比方法的性能,同时,将致命错误率降至 8.30%。消融分析实验证实,细胞群权重机制和 L2 正则化项的引入能够提高模型的性能。

关键词:黑色素瘤; 免疫检查点抑制剂; 批量 RNA 测序和单细胞 RNA 测序数据; 数据整合; 细胞间通信

中图分类号:R739.5;TP181 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2025.06.013

黑色素瘤因其高转移性和治疗耐药性成为最具侵袭性的皮肤癌。根据 Arnold 等[1]于 2020 年进行的全球黑色素瘤模型研究,预计到 2040 年,黑色素瘤将增加至 510 000 个新病例并导致 96 000 例死亡,亟需有效治疗策略。近年来,免疫检查点抑制剂(immune checkpoint inhibitor,ICI)通过激活 T 细胞免疫应答,显著改善临床治疗效果,使患者 5 年生存率从 5%提升至 30%[2]。然而,耐药性仍导致多数患者响应有限。

美国食品药品监督管理局已批准程序性死亡配体 1(PD-L1)表达水平、微卫星高度不稳定性/错配修复缺失及肿瘤突变负荷(tumer mutational burden,TMB)作为免疫治疗响应的预测标志物。然而,PD-L1 表达具有时空异质性,且其免疫组化检测结果受蛋白动态表达影响[3]。近期研究表明,RNA 测序技术相较于免疫组化,能更有效地预测黑色素瘤患者对 ICI 治疗的响应[4]。此外,TMB 评估易受肿瘤纯度干扰,可能导致结果偏差[5]。Litchfield 等[6]发现,传统标志物只能解释约 60%的 ICI 响应,上述研究表明了开发新预测模型的必要性。

人工智能方法在肿瘤免疫治疗预测领域取得显著进展。Tabari 等[7]首次将影像组学与血清乳酸脱氢酶(lactate dehydrogenase,LDH)结合构建了 LDH-LR 模型预测 ICI 治疗响应。Guo 等[8]提出了二阶多项式正则化逻辑回归模型用于预测黑色素瘤患者对 ICI 治疗的响应。Ligero 等[9]使用整合放射组学特征与临床特征的弹性网模型预测免疫治疗响应。特别地,Chowell 等[10]通过整合基因组学、分子学、人口学和临床数据提出了一种 RF16 模型,与基于 TMB 的方法相比,这个模型显著提高了 ICI 治疗响应预测的准确性。

转录组特征不仅能有效区分预后差异的患者亚群,还可指导免疫治疗的患者选择[11]。批量 RNA 测序数据已通过 Logistic 回归模型证实其在免疫治疗响应预测中的可靠性[12]。值得注意的是,单细胞测序技术可精确解析细胞异质性,通过识别特定亚

收稿日期:2025-05-12;**修订日期**:2025-06-19

基金项目:国家自然科学基金资助项目(61203293);河南省科技攻关项目(242102211023)

作者简介:穆晓霞(1980—),女,河南内黄人,河南师范大学副教授,主要从事机器学习、数学建模等方面的研究,E-mail:mxx@htu.edu.cn。

通信作者:李钧涛(1978—),男,河南社旗人,河南师范大学教授,博士,主要从事机器学习等方面的研究,E-mail:juntaolimail@126.com。

引用本文:穆晓霞,张红梅,宋学坤,等. 预测 ICI 治疗响应的凹惩罚 Logistic 回归模型[J]. 郑州大学学报(工学版),2025,46(6):58-65. (MU X X, ZHANG H M, SONG X K, et al. A concave-penalized Logistic regression model for predicting ICI treatment respense[J]. Journal of Zhengzhou University (Engineering Science), 2025,46(6):58-65.

群的配体-受体互作网络揭示免疫调控机制[13]。最新研究通过整合批量 RNA 测序和单细胞 RNA 测序数据,结合机器学习算法,显著提升了 ICI 治疗响应预测的准确性[14]。

尽管基于转录组的机器学习模型在预测 ICI 响应方面取得进展,但尚未充分挖掘单细胞分辨率下细胞通讯对疗效的影响。受文献[15-16]的启发,本文整合黑色素瘤患者的批量 RNA 测序数据,提出了二重群极小极大凹惩罚 Logistic 回归模型(DMCPLR)并将其应用于黑色素瘤患者 ICI 治疗响应预测。

1　模型构建

1.1　问题描述

本文收集了黑色素瘤患者的批量 RNA 测序数据和单细胞 RNA 测序数据,并将它们进行皮尔逊相关性整合形成新的数据集[$\boldsymbol{X},\boldsymbol{Y}$]。令 $\boldsymbol{X}=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_k,\cdots,\boldsymbol{x}_n)^{\mathrm{T}}$ 是 $n\times m$ 相关性矩阵,其中 $\boldsymbol{x}_k=(x_{k1},x_{k2},\cdots,x_{km})$,$x_{km}$ 代表第 k 个患者与第 m 个细胞之间的皮尔逊相关系数。令 $\boldsymbol{Y}=(y_1,\cdots,y_k,\cdots,y_n)^{\mathrm{T}}$ 为样本的标签。若样本对 ICI 有响应,则 y_k 为 1,否则为 0。

ICI 响应预测问题可以转化为二分类问题,即使用如下决策函数[8]来预测样本标签:

$$D(x)=\begin{cases}1, & f(x)\leqslant T_0;\\ 0, & f(x)>T_0。\end{cases}\tag{1}$$

式中:$f(x)=\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}+\boldsymbol{\beta}_0$ 为回归函数,$\boldsymbol{\beta}$ 和 $\boldsymbol{\beta}_0$ 通过 1.5 节中算法 1 得到;T_0 为阈值,根据对数概率的显著性,T_0 的值为 0.5。

1.2　数据集描述

黑色素瘤批量 RNA 测序数据集从基因表达综合数据库(gene expression omnibus,GEO)[17]下载。该数据集包含 22 名 ICI 响应者和 34 名 ICI 非响应者的 20 631 个基因表达值。黑色素瘤单细胞 RNA 测序数据从 GEO[18]下载。该数据集共包含 48 名黑色素瘤患者和 16 291 个免疫细胞。为深入分析免疫治疗的效果差异,根据患者接受治疗的时间节点(治疗前/治疗后)以及对治疗的反应状态(有响应/无响应),将数据划分为 4 组:治疗前有响应(preR)组、治疗前无响应(preNR)组、治疗后有响应(postR)组和治疗后无响应(postNR)组。上述数据集的详细描述如表 1 所示。

1.3　数据整合策略

现有基于单细胞和批量测序数据的 ICI 响应预测方法多采用特征提取与预测建模分离的两步式策略[12-13]。这种方法难以充分挖掘患者-细胞之间复杂的交互关系,限制模型对微环境状态与 ICI 响应潜在关联的刻画能力。为此,本文提出了一种整合批量 RNA 测序和单细胞 RNA 测序数据的新策略。首先,使用 R 包“Seurat”(4.3.0 版本)对数据进行预处理。特别地,单细胞 RNA 测序数据包含 55 738 个基因,将表达水平低于数据集中细胞总数 10% 的基因去除。接着,通过取两种数据中基因的交集来筛选共享基因。最后,通过计算患者-细胞间的皮尔逊相关系数构建相关矩阵(行为患者样本,列为单细胞),实现数据整合。批量 RNA 测序和单细胞 RNA 测序数据的整合过程如图 1 所示。相似矩阵中值的范围从-1 到 1,其中-1 表示细胞与样本之间完全负相关(即完全相反的相似性);0 表示无相似性;1 表示完全正相关(即完全相同的相似性)。

表 1　批量 RNA 测序和单细胞 RNA 测序数据集的详细信息

Table 1　Detailed information on bulk RNA-seq and single-cell RNA-seq datasets

分类	分组	免疫细胞数量	基因表达数量	样本数量
批量 RNA 测序数据		—	20 631	56
单细胞 RNA 测序数据	preR 组	2 725	55 738	9
	preNR 组	3 203	55 738	10
	postR 组	2 839	55 738	8
	postNR 组	7 524	55 738	21

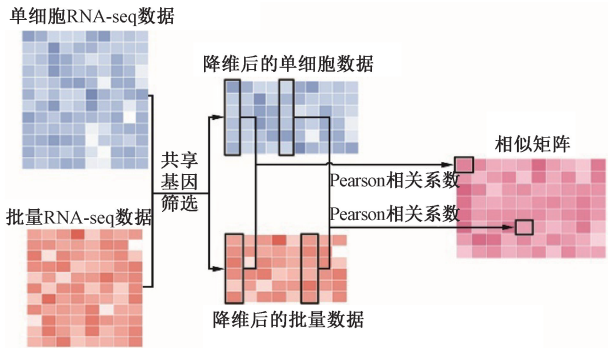


图 1　批量 RNA 测序和单细胞 RNA 测序数据的整合过程

Figure 1　The integration process of bulk RNA-seq and single-cell RNA-seq data

1.4　基于细胞-细胞通讯的细胞群重要性评价

鉴于细胞间通讯是在细胞群之间而非单个细胞之间发生,因此有必要对细胞进行预分组。本文使用 Louvain 算法对单细胞 RNA 测序数据中的细胞进行分组。首先通过 R 包“Seurat”中的“FindNeighbors”函数构建共享最近邻图。然后,通过式(2)将细胞划分为群[16]。

$$S=\frac{1}{2h}\sum_{u,v}\left(A_{uv}-\frac{k_uk_v}{2h}\right)\delta(t_ut_v)。\tag{2}$$

式中:h 代表图中边的总数;A_{uv} 表示细胞 u 和细胞 v 之间边的权重;k_u 和 k_v 分别是细胞 u 和细胞 v 的度;t_u 表示细胞 u 被分配到的群;δ 函数在 $t_u = t_v$ 时取值为 1,否则取值为 0。该算法通过 R 包“Seurat”中的“FindClusters”函数来实现,其中分辨率参数设置为 0.6。

本文利用 CellChat 设计群重要性评分体系。具体而言,首先利用完整的人类配体-受体相互作用数据库 CellChatDB,推断细胞间通讯情况,并将重要的配体-受体对归类到功能相关的信号通路中,进一步计算它们对相应通路的贡献。通过检索文献摘要(这些文献支持该通路的功能)中是否包含“免疫响应”一词来筛选与免疫响应相关的通路。在识别出与免疫响应相关的通路后,通过将每个配体-受体对在免疫响应相关通路中的相对贡献乘以该细胞群作为靶标的频率,来量化细胞群的重要性。

令 $\boldsymbol{C} = (c_1, \cdots, c_i, \cdots, c_p)^{\mathrm{T}}$ 为 $p \times g$ 矩阵,其中 p 为已识别的免疫响应相关通路的数量;g 表示细胞群的数量;向量 $\boldsymbol{c}_{il}$ 中的元素 c_i 代表第 l 个细胞群在第 i 条通路的重要性。设 $\boldsymbol{a}^i = (a_1^i, \cdots, a_j^i, \cdots, a_s^i)^{\mathrm{T}}$ 为 s 维向量,其中 s 为配体-受体对的数量;a_j^i 表示第 j 个配体-受体对在第 i 条通路中的相对贡献。令 $\boldsymbol{B}^i = (\boldsymbol{b}_{1l}^i, \cdots, \boldsymbol{b}_j^i, \cdots, \boldsymbol{b}_s^i)^{\mathrm{T}}$ 为 $s \times g$ 矩阵,其中向量 $\boldsymbol{b}_j^i$ 中的元素 b_j^i 表示第 l 个细胞群在第 i 条通路中作为第 j 个配体-受体对的靶标的频率。向量 $\boldsymbol{c}_i$ 使用以下公式进行计算:

$$\boldsymbol{c}_i = (\boldsymbol{B}^i)^{\mathrm{T}}\boldsymbol{a}^i = \sum_{j=1}^{s} a_j^i b_j^i 。 \tag{3}$$

第 l 个细胞群的重要性是其在每条通路中重要性的总和,可使用以下公式进行计算:

$$\boldsymbol{q}^{(l)} = \sum_{i=1}^{p} c_i^l 。 \tag{4}$$

式中:$(*)^{(l)}$ 表示向量中的 l 个分量。

1.5 二重群极小极大凹惩罚 Logistic 回归模型

针对免疫治疗响应预测中细胞群异质性强的问题,本文基于式(4)中群重要性评分,提出如下形式的惩罚函数:

$$P(\boldsymbol{\beta}) = \sum_{l=1}^{g} \frac{1}{q^{(l)}} f_{\lambda_1,b}\Big(\sum_{t=1}^{K_1} f_{\lambda_1,a}(|\beta_t^{(l)}|) \Big) + \frac{\lambda_2}{2} \|\boldsymbol{\beta}\|_2^2 。 \tag{5}$$

式中:a 和 b 为影响惩罚范围的参数;$\lambda_1 = \alpha\lambda$ 和 $\lambda_2 = (1-\alpha)\lambda$ 为正则化参数;$\|\boldsymbol{\beta}\|_2^2$ 为岭惩罚项。

$$f_{\lambda_1,a}(|\beta_t^{(l)}|) = \begin{cases} \lambda_1 |\beta_t^{(l)}| - |\beta_t^{(l)}|^2/2a, & \text{if } |\beta_t^{(l)}| \leqslant a\lambda_1; \\ (1/2)a\lambda_1^2, & \text{if } |\beta_t^{(l)}| > a\lambda_1 。 \end{cases} \tag{6}$$

式中:$f_{\lambda_1,a}(|\beta_t^{(l)}|)$ 为极小极大凹惩罚函数。

二重群极小极大凹惩罚函数为

$$\sum_{l=1}^{g} f_{\lambda_1,b}\Big(\sum_{t=1}^{K_1} f_{\lambda_1,a}(|\beta_t^{(l)}|) \Big) 。 \tag{7}$$

式中:$\boldsymbol{\beta}^i = (\boldsymbol{\beta}^{(1)}, \cdots, \boldsymbol{\beta}^{(l)}, \cdots, \boldsymbol{\beta}^{(g)})$ 表示回归系数向量;$\boldsymbol{\beta}^{(l)}$ 是其对应于第 l 个群的子向量;$\beta_t^{(l)}$ 表示第 l 个群中第 t 个细胞对应的系数;K_l 表示第 l 个群中细胞的数目。

与文献[19]相似,极小极大凹惩罚函数的导数为

$$f'_{\lambda_1,a}(|\beta_t^{(l)}|) = \begin{cases} \lambda_1 - \dfrac{|\beta_t^{(l)}|}{a}, & |\beta_t^{(l)}| \leqslant a\lambda_1; \\ 0, & |\beta_t^{(l)}| > a\lambda_1 。 \end{cases} \tag{8}$$

由式(8)可知,惩罚率起始于初始的 Lasso 惩罚,然后逐渐放宽,直至 $|\beta_t^{(l)}| > a\lambda_1$ 时,惩罚率变为 0。群极小极大凹惩罚通过对外层惩罚 $f_{\lambda_1,b}$ 应用于内层惩罚 $f_{\lambda_1,a}$ 的总和来实现组间和组内的双层特征选择。将 b 设置为 $K_l a|\lambda/2|$,以便当 $|\beta_t^{(l)}| \leqslant a\lambda_1$ 时实现外层惩罚率为 0。此外,在惩罚函数 $P(\beta)$ 中添加了惩罚项限制参数的大小,使模型更加稳定。

通过结合式(5)中的惩罚函数 $P(\beta)$ 与负对数似然损失函数提出如下的二重群极小极大凹惩罚 Logistic 回归模型(double group minimax concave penalty logistic regression model,DMCPLR):

$$\min_{\beta} \frac{1}{n} l(\boldsymbol{\beta}) + P(\boldsymbol{\beta}) ; \tag{9}$$

$$l(\boldsymbol{\beta}) = \sum_{k=1}^{n} \sum_{l=1}^{g} \Big(\log(1 + \exp(x_k^{(l)}\boldsymbol{\beta}^{(l)} + \beta_o k)) - y_k (x_k^{(l)T}\boldsymbol{\beta}^{(l)} + \boldsymbol{\beta}_o k) \Big) 。 \tag{10}$$

式中:n 表示样本数量;$x_k^{(l)}$ 代表第 k 个患者与第 l 个群中的细胞之间的相关系数。式(10)中的系数向量 $\boldsymbol{\beta}$ 和阈值向量 $\boldsymbol{\beta}_0$ 可通过算法 1 求解。

算法 1 DMCPLR 算法。

输入:训练集 X_{train},训练数据的标签向量 $\boldsymbol{Y}_{\text{train}}$,$\lambda$ 的最小值 $\lambda_{\min}$ 和最大值 $\lambda_{\max}$,正整数 n_1;

输出:系数向量 $\boldsymbol{\beta}$,阈值向量 $\boldsymbol{\beta}_0$;

① 通过 Louvain 算法获取细胞分组;

② 通过 CellChat 推断细胞间通讯;

③ 根据公式(4)计算细胞群的重要性;

④ for $\alpha \in \{0.05, 0.10, 0.20, \ldots, 0.90, 0.95\}$ do

⑤ for λ ($\lambda_{\min}$:($\lambda_{\max} - \lambda_{\min}$)/$n_1$:$\lambda_{\max}$) do

⑥ 将 X_{train} 划分为 10 个部分,即 X_{train_i}, $i = 1, 2, \ldots, 10$;

⑦ for $i = 1:10$ do

⑧ 将 X_{train_i} 作为测试集,其余部分作为训练集;
⑨ 使用 R 语言的"grpreg"包在训练集上拟合 DMCPLR 模型,并在测试集上进行预测;
⑩ end for
⑪ 计算 10 个测试集上的平均预测误差;
⑫ end for
⑬ 确定最优的 λ;
⑭ end for
⑮ 确定最优参数对 (α^*,λ^*);
⑯ 使用 R 语言的"grpreg"包拟合 DMCPLR 模型,并在最优参数对 (α^*,λ^*) 下获取系数向量 $\boldsymbol{\beta}$ 和阈值向量 $\boldsymbol{\beta}_0$。

在进行实验之前,需要对相似矩阵进行标准化,使得 $\sum_{k=1}^{n} x_{kt}^{(l)} = 0$ 且 $\sum_{k=1}^{n} (x_{kt}^{(l)})^2 = 1(\forall l,t)$ 以确保惩罚能被平等地施加,其中 $x_{kt}^{(l)}$ 表示第 k 个患者与第 l 个群中第 t 个细胞之间的相关系数。随后,实施分层抽样,随机选取 80%(44 个)的样本作为训练集 X_{train},而剩余的样本(12 个)被指定为测试集 X_{test}。为了增强对模型性能的可靠评估,通过设置从 1 到 50 的随机种子,将预处理后的数据集随机划分 50 次。

2　实验及结果分析

2.1　黑色素瘤 ICI 治疗响应预测实验

为评估所提方法的有效性,本文在黑色素瘤 GSE35640 批量 RNA 测序数据集和整合数据集上进行了实验。实验比较了 14 种具有代表性的机器学习方法,包括 6 种经典的机器学习方法——Lasso 回归、弹性网络(elastic net, EN)、脊回归(Ridge)、高斯朴素贝叶斯(Gaussian naive Bayes,GNB)、Logistic 回归(logistic regression,LR)和支持向量机(support vecter mechine,SVM),4 种集成学习方法——随机森林(random forest, RF)、XGBoost、AdaBoost 和 LightGBM,2 种神经网络架构——深度神经网络(deep neural network,DNN)和反向传播神经网络(backpropagation neural network,BPNN),最新相关模型——LDH-LR 和 SOPRLR 模型。由于 SOPRLR 模型需要计算特征间的相互作用,而本文的数据包含数千维特征,直接应用会使特征组合爆炸。为避免这一问题,首先基于随机森林的基尼重要性评分筛选出前 50 个关键特征,随后在这些特征上应用 SOPRLR 模型。考虑到所提模型需要同时利用单细胞 RNA 测序数据和批量 RNA 测序数据进行训练和测试,因此仅在整合数据集上评估所提模型性能。为确保实验结果的可靠性,所有模型均采用相同的数据预处理流程。在求解 DMCPLR 模型的过程中,α 设置为 30。对于每个固定的 α,通过交叉验证来确定 λ,并根据实验结果确定最优参数对 (α^*,λ^*)。其余对比方法的参数配置均依据原始文献推荐策略进行调优。本文的实验均在一台配备英特尔酷睿 i7-12700H 处理器和 32 GB 内存的计算机上完成,软件版本为 R 4.3.3。

为了验证整合数据策略的优势,本文在黑色素瘤 GSE35640 批量 RNA 测序数据集上评估了 14 种基准方法的预测性能。表 2 展示了在黑色素瘤 GSE35640 批量 RNA 测序测试数据集上 14 种模型在 50 次实验中 4 种评价指标所对应的均值及标准差,表 3 则呈现了在整合数据集上的结果。对比可

表 2　14 种模型在黑色素瘤 GSE35640 数据集上的结果

Table 2　Results of 14 models on the melanoma GSE35640 dataset

模型	准确率		精确率		召回率		*F*1 分数	
	平均值/%	标准差	平均值/%	标准差	平均值/%	标准差	平均值/%	标准差
BPNN[20]	52.33	0.134 7	59.00	0.149 2	42.33	0.134 7	51.36	0.139 5
AdaBoost[20]	52.33	0.128 2	58.35	0.158 7	51.33	0.128 2	52.41	0.136 2
DNN[21]	52.17	0.136 3	58.03	0.157 4	54.26	0.136 3	51.40	0.141 2
EN[8]	57.67	0.102 1	59.21	0.118 3	53.67	0.102 1	54.32	0.105 7
GNB[21]	56.83	0.129 1	53.61	0.138 4	47.94	0.129 1	51.33	0.129 0
Lasso[8]	54.50	0.141 9	56.65	0.160 8	52.18	0.141 9	51.68	0.152 2
LightGBM[20]	49.83	0.152 4	53.34	0.156 7	51.87	0.152 4	49.07	0.154 5
LR[20]	50.33	0.130 5	**69.69**	0.130 7	48.91	0.130 5	62.65	0.125 0
RF[21]	55.50	0.122 6	52.89	0.143 3	51.49	0.122 6	50.57	0.128 7
Ridge[8]	53.83	0.119 2	59.13	0.129 9	54.97	0.119 2	52.61	0.116 1
SVM[8]	59.17	0.922 0	43.83	0.142 9	58.17	0.922 0	48.53	0.103 7
XGBoost[20]	54.33	0.143 7	56.88	0.170 1	56.11	0.143 7	52.59	0.153 5
LDH-LR[7]	62.91	0.837 0	56.20	0.752 0	**72.16**	0.105 2	**66.74**	0.115 2
SOPRLR[8]	**68.57**	0.963 0	62.74	0.101 6	65.74	0.917 0	64.32	0.730 0

知,整合单细胞与批量 RNA 测序数据的方法性能显著提升:所有模型在整合数据集的平均准确率高于单独用批量数据;各模型整合数据集标准差普遍降低,既提高精度又增强稳定性。虽然所有模型在整合数据集性能优于批量 RNA 数据集,但 DMCPLR 模型仍居领先,相比其余模型平均预测准确率分别高出 22.18 百分点,15.18 百分点,21.85 百分点,15.68 百分点,10.51 百分点,8.51 百分点,13.18 百分点,15.85 百分点,16.85 百分点,17.18 百分点,20.35 百分点,13.85 百分点,1.82 百分点 和 9.25 百分点。

鉴于整合数据集的类别不平衡性,本文引入了额外的评价指标(精确率、召回率、*F*1 分数、混淆矩阵)。结果显示,DMCPLR 模型在平均精确率上较其他模型分别提高了 19.48 百分点,11.85 百分点,20 百分点,11.5 百分点,8.62 百分点,12.55 百分点,6.22 百分点,9.74 百分点,13.49 百分点,12.08 百分点,34.66 百分点,10.34 百分点,18.34 百分点和 14.5 百分点。类似的,在召回率与 *F*1 分数方面,DMCPLR 模型也获得了优越的性能。由于 LR 模型是所提模型的一个特例且性能不佳,因此在后续对比中不再考虑。ICI 治疗中,误将响应者预测为非响应者属于致命错误(或致患者错失关键治疗机会)。图 2 展示了 DMCPLR 模型与其他 13 种方法在整合数据集上进行 10 次实验的混淆矩阵结果。(矩阵元素数值以颜色梯度呈现,数值大则色块深)。从图 2 可以看出,在 12 个样本中,DMCPLR 模型仅出现 1 个致命错误样本,DNN、Lasso、Ridge 和 XGBoost 模型均出现 2 个致命错误样本,AdaBoost、EN、BPNN、GNB、LightGBM、SOPRLR、RF 均出现 3 个致命错误样本,SVM 和 LDH-LR 模型出现 4 个致命错误样本,这表明 DMCPLR 模型在整合数据集上的致命错误率显著低于其他方法。

2.2 消融分析

为了评估基于细胞通讯构建的细胞群重要性准则式(4)对模型性能的影响,本文对 DMCPLR 模型进行了消融分析实验。具体而言,本文去除了 DMCPLR 模型中基于细胞通讯的细胞群权重分配机制,即将所有细胞群权重统一设置为 1 进行等权重处理,同时保持其他模型结构和超参数不变,并将该模型命名为 GMCPLR(group minimax concave penalty logistic regression model)。经过相同的实验步骤后,DMCPLR 和 GMCPLR 在整合后数据集的测试集上的平均预测准确率如图 3 所示。从图 3 中可以看出,GMCPLR 在整合后数据集的测试集上的平均预测准确率低于 DMCPLR 在相应测试集上的平均预测准确率。这表明基于细胞通讯的细胞群权重分配机制能够有效提升模型预测性能,进而验证了细胞间通讯对免疫治疗响应的重要性。

为进一步探讨惩罚函数对模型性能的影响,本文在 DMCPLR 模型中去除了 L2 正则化项,并将其命名为 DMCPLR-L1。为了进行对比,本文选取 LR 模型作为基线模型。图 4 展示了在整合数据集上,DMCPLR、DMCPLR-L1 和 LR 模型在 50 次实验中的平均预测准确率、精确率、召回率和 *F*1 分数。从图 4 中可以看出,DMCPLR 模型在 4 个评价指标上均优于其他两个模型。

表 3 14 种模型在整合测试数据集上的结果

Table 3 Results of 14 models on the integrated test dataset

模型	准确率		精确率		召回率		*F*1 分数	
	平均值/%	标准差	平均值/%	标准差	平均值/%	标准差	平均值/%	标准差
DMCPLR	**80.18**	0.674 0	**82.24**	0.105 2	**89.71**	0.867 0	**85.11**	0.749 0
BPNN[20]	58.00	0.124 4	62.76	0.148 9	57.24	0.134 1	57.84	0.129 2
AdaBoost[20]	65.00	0.127 9	70.39	0.158 4	61.48	0.127 9	64.84	0.136 0
DNN[21]	58.33	0.136 0	62.24	0.157 1	59.37	0.136 0	58.23	0.140 9
EN[8]	64.50	0.990 0	70.74	0.114 7	62.40	0.990 0	64.40	0.102 5
GNB[21]	69.67	0.128 8	73.62	0.138 1	64.17	0.128 8	69.46	0.128 7
Lasso[8]	71.67	0.141 9	76.02	0.160 8	61.77	0.141 9	70.97	0.152 2
LightGBM[20]	67.00	0.146 8	72.83	0.152 0	66.21	0.146 8	67.01	0.149 9
LR[20]	64.33	0.126 6	72.50	0.126 8	64.33	0.126 6	64.93	0.121 3
RF[21]	63.33	0.118 9	68.75	0.139 0	60.37	0.118 9	63.03	0.124 8
Ridge[8]	63.00	0.115 6	70.16	0.126 0	64.19	0.115 6	63.46	0.112 6
SVM[8]	59.83	0.894 0	47.58	0.138 6	58.68	0.894 0	50.60	0.100 6
XGBoost[20]	66.33	0.139 4	71.90	0.165 0	64.95	0.139 4	66.22	0.148 9
LDH-LR[7]	78.36	0.812 0	63.90	0.729 0	76.32	0.102 0	69.54	0.111 7
SOPRLR[8]	70.93	0.934 0	67.74	0.986 0	74.32	0.890 0	70.92	0.708 0

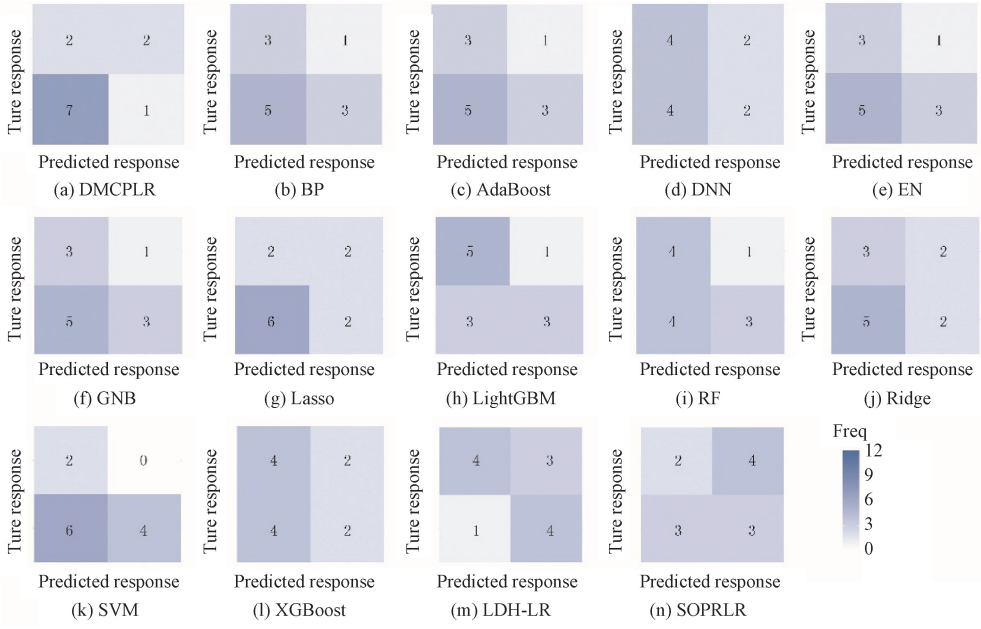


图 2　14 种方法的混淆矩阵可视化

Figure 2　Fourteen methods for visualizing confusion matrices

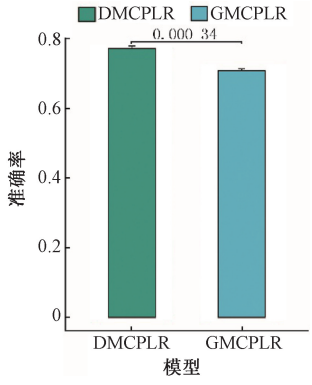


图 3　DMCPLR 和 GMCPLR 在整合数据集的测试集上的平均预测准确率箱线图

Figure 3　Boxplot of the average prediction accuracy for DMCPLR and GMCPLR on the testing sets of the integrated datasets

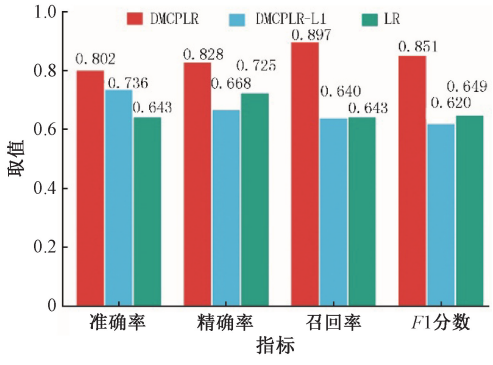
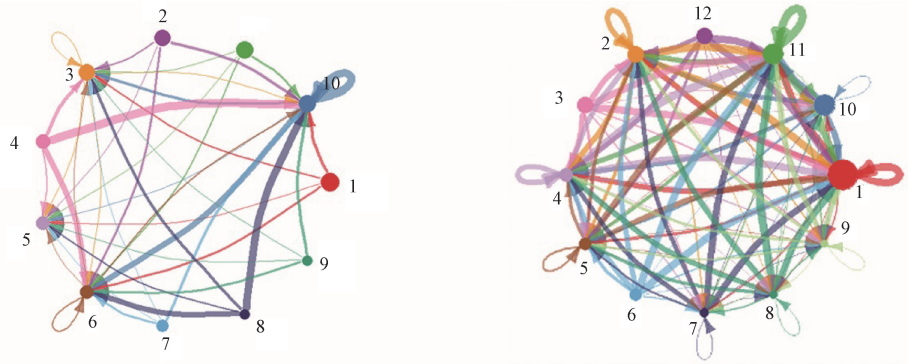


图 4　DMCPLR、DMCPLR-L1 和 LR 在整合数据集上 4 种评价指标的柱状图

Figure 4　Bar charts of four evaluation indicators DMCPLR, DMCPLR-L1, and LR on integrated datasets

2.3　细胞分群与细胞通讯实验结果

考虑到细胞群对 ICI 响应的差异性影响,使用单细胞 RNA 测序数据中推断的细胞间通信,并用 CellChat 评估细胞群的重要性。以下将以 postR 和 postNR 数据为例展示结果,preR 和 preNR 的分析类似。经过数据预处理后,postR 和 postNR 的单细胞 RNA 测序数据分别被划分为 11 和 12 个细胞群。为了研究不同细胞群对 ICI 治疗的影响,使用 CellChat 推断细胞群之间的通信。随后,在 postR 数据中识别出 37 个显著的配体-受体相互作用,这些相互作用涉及 11 个免疫响应相关的通路;在 postNR 数据中,识别出 46 个显著的配体-受体相互作用,涉及 14 个免疫响应相关的通路。由于处理过程相同,接下来的分析将仅以每个数据集中的一个信号通路为例。图 5(a)和图 5(b)分别展示了来自 postR 数据的 MHC-II 信号网络和来自 postNR 数据的 MHC-I 信号网络(图中节点的不同颜色代表不同的细胞群,节点的大小与细胞群中的细胞数量成正比。箭头从源节点指向目标节点,边的颜色与信号源一致。边的宽度表示通信概率。)。在 MHC-II 信号通路中,细胞群 3,5,6 和 10 通过多个细胞群传递信号。类似地,在 MHC-I 信号通路中,信号传递发生在细胞群 1,2,4,5,7,8,10 和 11 中。细胞群的重要性将通过其作为目标的频率来评估。



(a) 基于postR数据推断的MHC-II信号网络　(b) 基于postNR数据推断的MHC-I信号网络

图5　推断信号通路的圆形图

Figure 5　Circle plot of inferred signal pathways

3　结论

本文通过整合批量RNA测序和单细胞RNA测序数据,并引入细胞群重要性,开发了DMCPLR模型,显著提高了黑色素瘤ICI治疗响应的预测性能。实验结果表明:DMCPLR的平均预测准确率显著优于13种对照方法,同时展现出更低的假阴性率(即更少的治疗响应者被误判为非响应者);消融分析实验也证实了细胞群权重机制的引入是性能提升的关键因素。实际上,每个细胞群中单个细胞对ICI响应的影响存在差异。因此,如何量化单个细胞的重要性将是未来研究的一个关键方向。

参考文献:

[1] ARNOLD M, SINGH D, LAVERSANNE M, et al. Global burden of cutaneous melanoma in 2020 and projections to 2040[J]. JAMA Dermatology, 2022, 158(5): 495-503.

[2] GUO W N, WANG H N, LI C Y. Signal pathways of melanoma and targeted therapy[J]. Signal Transduction and Targeted Therapy, 2021, 6(1): 424.

[3] YIN X M, LIAO H, YUN H, et al. Artificial intelligence-based prediction of clinical outcome in immunotherapy and targeted therapy of lung cancer[J]. Seminars in Cancer Biology, 2022, 86: 146-159.

[4] CONROY J M, PABLA S, NESLINE M K, et al. Next generation sequencing of PD-L1 for predicting response to immune checkpoint inhibitors[J]. Journal for Immunotherapy of Cancer, 2019, 7(1): 18.

[5] 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(1): 99-111.

[6] LITCHFIELD K, READING J L, PUTTICK C, et al. Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition[J]. Cell, 2021, 184(3): 596-614. e14.

[7] TABARI A, COX M, D'AMORE B, et al. Machine learning improves the prediction of responses to immune checkpoint inhibitors in metastatic melanoma[J]. Cancers, 2023, 15(10): 2700.

[8] GUO Q H, XIANG S, LI J T. Second-order polynomial regularized logistic regression for predicting melanoma patients response to immune checkpoint inhibitors[C]// Proceedings of 2024 Chinese Intelligent Systems Conference. Cham: Springer, 2024: 610-617.

[9] LIGERO M, GARCIA-RUIZ A, VIAPLANA C, et al. Artificial intelligence combining radiomics and clinical data for predicting response to immunotherapy[J]. Annals of Oncology, 2019, 30: 476.

[10] 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(4): 499-506.

[11] ZHANG S N, LI M Y, TAN Y L, et al. Identification of mutational signature for lung adenocarcinoma prognosis and immunotherapy prediction[J]. Journal of Molecular Medicine, 2022, 100(12): 1755-1769.

[12] KONG J, HA D, LEE J H, et al. Network-based machine learning approach to predict immunotherapy response in cancer patients[J]. Nature Communications, 2022, 13(1): 3703.

[13] VALDES-MORA F, HANDLER K, LAW AMK, et al. Single-cell transcriptomics in cancer immunobiology: the future of precision oncology[J]. Frontiers in Immunology, 2018, 9: 2582.

[14] ZHANG Z, WANG Z X, CHEN Y X, et al. Integrated analysis of single-cell and bulk RNA sequencing data reveals a pan-cancer stemness signature predicting immunotherapy response[J]. Genome Medicine, 2022, 14(1): 45.

[15] SUN D C, GUAN X N, MORAN A E, et al. Identifying phenotype-associated subpopulations by integrating bulk and single-cell sequencing data[J]. Nature Biotechnology, 2022, 40(4): 527-538.
[16] LI J T, ZHANG H M, MU B Y, et al. Identifying phenotype-associated subpopulations through LP_SGL[J]. Briefings in Bioinformatics, 2023, 25(1): bbad424.
[17] BARRETT T, TROUP D B, WILHITE S E, et al. NCBI GEO: archive for high-throughput functional genomic data [J]. Nucleic Acids Research, 2009, 37(Database issue): D885-D890.
[18] ULLOA-MONTOYA F, LOUAHED J, DIZIER B, et al. Predictive gene signature in MAGE-A3 antigen-specific cancer immunotherapy[J]. Journal of Clinical Oncology, 2013, 31(19): 2388-2395.
[19] BREHENY P, HUANG J. Penalized methods for bi-level variable selection[J]. Statistics and Its Interface, 2009, 2(3): 369-380.
[20] CHEN M, LI Y X, ZHOU S M, et al. Establishment of a risk prediction model for olfactory disorders in patients with transnasal pituitary tumors by machine learning[J]. Scientific Reports, 2024, 14(1): 12514.
[21] SARKAR J P, SAHA I, SARKAR A, et al. Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers[J]. Computers in Biology and Medicine, 2021, 131: 104244.

A Concave-penalized Logistic Regression Model for Predicting ICI Treatment Respense

MU Xiaoxia[1], ZHANG Hongmei[2], SONG Xuekun[3], LI Juntao[4]

(1. College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China; 2. College of Life Sciences, Northeast Forestry University, Harbin 150006, China; 3. College of Information Technology, Henan University of Chinese Medicine, Zhengzhou 450046, China; 4. School of Mathematics and Statistics, Henan Normal University, Xinxiang 453007, China)

Abstract: To improve the accuracy of predicting the response of melanoma patients to immune checkpoint inhibitor (ICI) therapy, a new method integrating bulk RNA-seq and single-cell RNA-seq data was proposed. Firstly, a patient-cell correlation matrix was constructed through Pearson correlation analysis, and the Louvain algorithm was used to classify single-cell RNA-seq data into cell groups. The importance of cell groups in immune response related pathways was quantified using the CellChat tool. On this basis, a double group minimax concave penalty logistic regression model (DMCPLR) was proposed by introducing the cell group importance evaluation criterion constructed based on the cell-cell communication network and combining with the group minimax concave penalty. The experiments on the GSE35640 dataset showed that the prediction accuracy of the DMCPLR model reached 80.18%, with precision, recall, and $F1$ score of 82.24%, 89.71%, and 85.11%, respectively, significantly better than the performance of 14 comparison methods including Lasso regression and random forest, while reducing the fatal error rate to 8.30%. The ablation analysis experiment confirmed that the introduction of cell group weight mechanism and L2 regularization term can improve the performance of the model.

Keywords: melanoma; immune checkpoint inhibitor; bulk RNA-seq and single-cell RNA-seq data; data integration; cell-cell communication