· 临床与预防 ·

基于多组学数据构建低级别胶质瘤患者预后预测模型

刘毅蓉[1],任月[1],秦阳[1],武舒琪[1],赵晋芳[1,2],罗天娥[1,2]

1. 山西医科大学公共卫生学院卫生统计学教研室,山西 太原 030001;2. 煤炭环境致病与防治教育部重点实验室

摘要:目的　探讨综合聚类方法在识别低级别胶质瘤(LGG)亚型和预后预测中的应用。**方法**　采用集合了十种聚类算法的综合聚类算法(MOVICS)对 TCGA 下载的 LGG 患者多组学数据进行整合,得到聚类亚型;通过多因素 Cox 回归分析 LGG 的预后因素。使用 mRNA 数据构建随机森林分类预测模型来评估分类性能,用 CGGA 数据集进行外部验证。**结果**　LGG 患者经过聚类分为两型,两组生存率差异具有统计学意义($\chi^2 = 54.410$, $P < 0.001$)。多因素 Cox 回归分析结果表明,年龄($HR = 1.053$, 95% CI: 1.037 ~ 1.069)、癌症分级($HR = 2.733$, 95% CI: 1.836 ~ 4.069)和聚类亚型($HR = 3.210$, 95% CI: 2.216 ~ 4.650)都是 LGG 的预后因素,Nomogram 图、校准曲线和 ROC 曲线结果表明模型的预测性能良好。十折交叉验证 RF 模型的平均预测准确率为 87.81%,训练集、内部验证集和两个外部验证集的 C 指数分别为 0.717、0.721、0.574 和 0.572,Brier 评分分别为 0.044、0.066、0.179 和 0.128,两个外部验证数据集的生存差异均有统计学意义($P < 0.05$)。**结论**　综合聚类方法能够有效识别 LGG 亚型,其亚型是 LGG 患者的预后因素,并在外部数据集 CGGA 中得到验证,可为 LGG 的临床个性化治疗提供重要的理论依据。

关键词:多组学聚类;低级别胶质瘤;预后预测;随机森林

中图分类号:R739.4　　**文献标志码:**A　　**文章编号:**1003 - 8507(2024)14 - 2669 - 06

DOI:10.20043/j.cnki.MPM.202404262

Prognostic prediction model for patients with low - grade gliomas based on multi - omics data

LIU Yi - rong*, REN Yue, QIN Yang, WU Shu - qi, ZHAO Jin - fang, LUO Tian - e

Department of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan, Shanxi 030001, China

Abstract: Objective　To explore the application of integrated clustering methods in identifying low - grade glioma subtypes and prognostic prediction. **Methods**　A comprehensive clustering algorithm (MOVICS), which pools ten clustering algorithms, was used to integrate the multi - omics data of LGG patients downloaded from TCGA to obtain cluster subtypes; prognostic factors of LGG were analyzed by multifactorial Cox regression. A random forest classification prediction model was constructed using mRNA data to evaluate the classification performance and externally validated with the CGGA dataset. **Results**　LGG patients were clustered into two subtypes, and the difference in survival between the two groups was statistically significant ($\chi^2 = 54.410$, $P < 0.001$). The results of multifactorial Cox regression analysis showed that age ($HR = 1.053$, 95% CI: 1.037 - 1.069), cancer grade ($HR = 2.733$, 95% CI: 1.836 - 4.069) and cluster typing ($HR = 3.210$, 95% CI: 2.216 - 4.650) were all prognostic factors for LGG, and the results of Nomogram plots, calibration curves and ROC curves indicated good predictive performance of the model. The average prediction accuracy of the ten - fold cross - validated RF model was 87.81%, and the C - indexes of the training set, the internal validation set, and the two external validation sets were 0.717, 0.721, 0.574, and 0.572, and the Brier scores were 0.044, 0.066, 0.179, and 0.128, and the differences in the survival of the two external validation datasets were all statistically significance ($P < 0.05$). **Conclusion**　The comprehensive clustering method can effectively identify LGG subtypes, which are prognostic factors for LGG patients, and has been validated in an external dataset, CGGA, which can provide an important theoretical basis for clinical personalized treatment of LGG.

Keywords: Multi - omics clustering; Low - grade glioma; Prognosis prediction; Random forest

低级别胶质瘤(low - grade glioma, LGG)具有高度异变性,极易发展为最恶性的胶质母细胞瘤(Glioblastoma, GBM),对患者生存提出巨大考验[1]。由于 LGG 患者预后差异较大,不能仅根据组织学亚型进行预测,因此探讨更好的亚型预测方法十分关键。随着高通量测序技术与基因组芯片技术的飞速

基金项目:山西省自然科学基金(201801D121210)

作者简介:刘毅蓉(1998—),女,硕士在读,研究方向:机器学习在疾病分类和预后预测中的应用

通信作者:罗天娥,E - mail:luotiane1977@163.com

发展,获得了多种癌症患者的基因组学、转录组学和蛋白质组学等组学数据[2],可以反映癌症的多种生物学过程[3]。聚类可以对组学数据进行整合分析,在疾病亚型、精准医疗、药物研究等方面具有非常重要的现实意义[4-5]。整合多组学数据有利于对生物医学数据进行全面深入的研究,补充单一组学中缺失或不可靠的信息[6]。由于组学数据通常具有高维度、样本量少、高噪声[7]的特点,研究中选择合适的聚类算法十分关键[8]。

MOVICS 是一种综合聚类方法,通过综合 iClusterBayes、Mocluster、CIMLR、IntNMF、ConsensusClustering、LRAcluster、COCA、PINSplus、SNF 和 NEMO 十种聚类算法,得到聚类亚型。本文使用 TCGA 的 LGG 患者多组学数据库,通过 MOVICS 模型进行聚类,利用多因素 Cox 回归分析不同亚型对 LGG 患者预后的影响,绘制 Nomogram 图、校准曲线和 ROC 曲线评价模型的预测性能;使用 mRNA 数据构建随机森林模型验证 MOVICS 聚类性能,同时用 CGGA 数据进行外部验证,为 LGG 的临床个性化治疗提供重要的理论依据。

1 资料和方法

1.1 数据来源与处理 从 UCSCXena 网站(https://xenabrowser.net)下载低级别胶质瘤(LGG)癌症患者的基因表达、DNA 甲基化、miRNA 数据以及临床信息数据(性别、年龄、癌症分级、生存时间和生存状态)。其中 count 数据经过 log 转化,DNA 甲基化数据剔除未匹配到基因的探针以及对应多个基因名称的探针,再去除重复基因。多组学数据均剔除缺失值超过 10% 的基因,剩余缺失值用中位数进行插补,数据经过归一化处理。从中国脑胶质瘤基因组图谱计划 CGGA(http://www.cgga.org.cn/)下载胶质瘤的两个基因表达数据以及临床信息数据(生存时间、生存状态),mRNA 数据经 log 转化,排除临床数据中 WHOIV 型胶质瘤,删除缺失数据,去除批次效应。从 COSMIC(https://cancer.sanger.ac.uk/)数据库获得 733 个泛癌相关基因,在 mRNA 和 DNA 甲基化数据中筛选泛癌相关基因。

1.2 MOVICS MOVICS 是 Lu 等[9]2020 年提出的综合聚类方法。它的输入是多组学数据,输出是综合聚类后最优的分子亚型。MOVICS 选择聚类预测指数(Cluster Prediction Index, CPI)和 Gap 统计量之和为最大值的聚类数作为最佳聚类数。通过综合十种聚类算法(iClusterBayes、moCluster、CIMLR、IntNMF、COCA、NEMO、ConsensusClustering、PINSPlus、SNF 和 LRA),将患者分为不同的亚型,并使用共识聚类进行组合分类,以高度鲁棒性识别每个亚型。

具体来说,如果指定 t_{max} 算法,且 $2 \leqslant t_{max} \leqslant 10$,则每个算法计算一个矩阵 $M\frac{(t)}{n \times n}$,其中 n 为样本个数,当样本 i 和 j 聚类在同一子类型中时,$M\frac{(t)}{ij}=1$;否则 $M\frac{(t)}{ij}=0$。在获得指定算法的所有结果后,MOVICS 计算共识矩阵 $CM=\sum_{t=1}^{t_{max}} M^t$,并且 $CM\epsilon[0,10]$,并通过剪影评分计算各亚型之间的样本相似性[10]。

1.3 随机森林模型 随机森林(random forest, RF)是一种集成学习方法,通过构建多个决策树并取其结果的平均值或投票来进行预测,在每个决策树的训练过程中,采用了自助采样法对样本和特征进行随机选择,能够处理变量数量超过观测数量的数据集[11]。模型具有较高的稳定性和泛化能力,可以很好地处理过拟合问题[12],有效处理缺失和有噪声的数据。

1.4 最佳聚类数指标 CPI:基于重采样的方法将数据重复划分为训练集和测试集。每次重复时,将算法应用于训练集估计系数矩阵,在测试集上使用系数矩阵估计公共基矩阵,重复多次,计算调整后的兰德指数的平均值为 CPI[13],一般选择导致 CPI 最大值的聚类数作为最佳聚类数。Gap 统计量[14]:引入参考的测值由 MonteCarlo 采样的方法获得,通过计算标准差来矫正 Gap 统计量,一般选择 Gap 统计量最大值对应的聚类数为最佳聚类数。

1.5 分类性能指标 选择 C 指数、Brier 评分和 log-rank$P$ 值来评估模型的分类预测性能。C 指数取值在 0 和 1 之间,该值越高概率预测的准确性越高;Brier 评分取值在 0 和 1 之间,Brier 分数越低概率预测的准确性越高;检验水准 $\alpha=0.05$。

1.6 软件实现 本研究在 R 4.1.3 软件中完成,去除批次效应在 sva 包中实现,聚类分析采用 MOVICS 包,随机森林预测模型构建在 randomForest 包中实现。

2 结 果

2.1 LGG 患者多组学数据的描述 从 TCGA 数据库下载 LGG 的 mRNA、miRNA 和 DNA 甲基化数据预处理后整合得到 502 个患者,mRNA、DNA 甲基化数据筛选泛癌相关基因并归一化。CGGA 数据库的两个 mRNA 数据集筛选泛癌相关基因,去批次效应。所用数据集关键特征如表 1 所示。

2.2 综合聚类模型构建及亚型结果评价 使用 MOVICS 对 TCGA 的多组学数据进行聚类,聚类数的范围设定为[2,8],结合不同聚类数的 CPI 和 Gap 统计量结果,当聚类数为 2 时 CPI 和 Gap 统计量之和达

到最大值,模型较优,见图 1。

表 1　数据集关键特征情况

Table 1　Key features of the dataset situation

数据集	组学类型	样本量	特征数
TCGA – LGG	mRNA	502	691
	miRNA	502	1 524
	DNA 甲基化	502	671
CGGA1	mRNA	172	691
CGGA2	mRNA	420	691

最佳聚类数为 2,即 MOVICS 将 502 名 LGG 患者分为两个亚型,CS1 和 CS2,基本资料如表 2 所示。绘制 Kaplan – Meier 生存曲线,见图 2,两个亚型患者的生存率差异有统计学意义($\chi^2 = 54.410$, $P < 0.001$),且 CS2 组生存率较低,预后较差,表明多组学数据聚类可以得到不同预后的 LGG 患者。

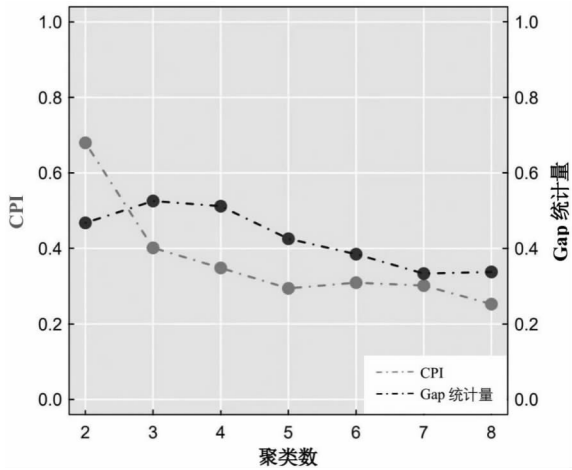

图 1　最佳聚类数

Fig. 1　Optimal number of clusters

表 2　TCGA – LGG 患者亚型的基本资料

Table 2　Basic information on TCGA – LGG patient typing

项目	CS1(n = 312)	CS2(n = 190)	χ^2/z	P 值
年龄(岁,$\bar{x} \pm s$)	40.9 ±12.4	46.6 ±14.3	4.391	< 0.001
性别[n(%)]			1.905	0.168
男性	179(57.4)	97(51.1)		
女性	133(42.6)	93(48.9)		
癌症分级[n(%)]			10.498	< 0.001
WHO G2	168(53.8)	74(38.9)		
WHO G3	144(46.2)	116(61.1)		
生存状况[n(%)]			27.604	< 0.001
存活	259(83.0)	118(62.1)		
死亡	53(17.0)	72(37.9)		

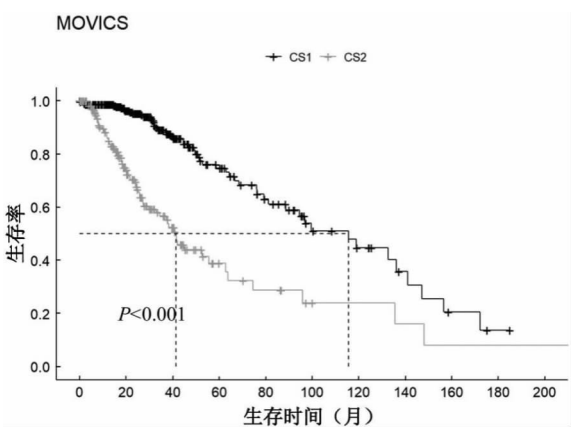

图 2　TCGA – LGG – KM 生存曲线

Fig. 2　KM survival curves for TCGA – LGG

2.3　多因素 Cox 回归分析及 Nomogram 图构建　以生存时间和生存状态作为因变量,对年龄、性别(女性 = 0,男性 = 1)、癌症分级(WHO G2 = 1,WHO G3 = 2)和 MOVICS 聚类亚型结果(CS1 = 1,CS2 = 2)进行多因素 Cox 逐步回归分析,结果显示年龄、癌症分级、聚类亚型结果具有统计学意义($P < 0.05$),可作为低级别胶质瘤患者的预后预测因素,见表 3。

基于多因素 Cox 回归分析的结果,对 LGG 患者的预后因素构建 Nomogram 图,预测 LGG 患者 1、3 和 5 年的生存率。Nomogram 图的 C 指数为 0.82。Nomogram 图中所有变量 Points 轴的分数之和在 Totalpoints 轴上显示,可以直观的估算出 LGG 患者 1、3 和 5 年的生存率,见图 3。

利用校准曲线评价 Nomogram 图,预测 LGG 患者生存率的准确性,以坐标轴原点且斜率为 1 的标准曲线作为参照。LGG 患者的 1、3 和 5 年偏差校准曲线接近标准曲线,表明预测的 LGG 患者的生存率与实际观察到的结果偏差小,有很好的一致性,见图 4。绘制 1、3 和 5 年的 ROC 曲线,LGG 患者的 AUC 均大于 0.8,表明模型有很好的预测准确性,见图 5。

2.4　随机森林分类预测模型构建　利用 TCGA 数据库中 LGG 患者的 mRNA 数据,根据滑动窗口序贯向前特征选择法筛选出 145 个基因,将这 145 个基因的

表 3 TCGA – LGG 患者多因素 Cox 回归分析结果

Table 3 Results of multifactorial Cox regression analysis in TCGA – LGG patients

变量	*b*(*S. E*)	Waldχ^2值	*P* 值	*HR*(95% *CI*)
年龄	0. 051(0. 008)	46. 221	<0. 001	1. 053(1. 037 ~ 1. 069)
癌症分级	1. 005(0. 203)	24. 522	<0. 001	2. 733(1. 836 ~ 4. 069)
亚型	1. 166(0. 189)	38. 065	<0. 001	3. 210(2. 216 ~ 4. 650)

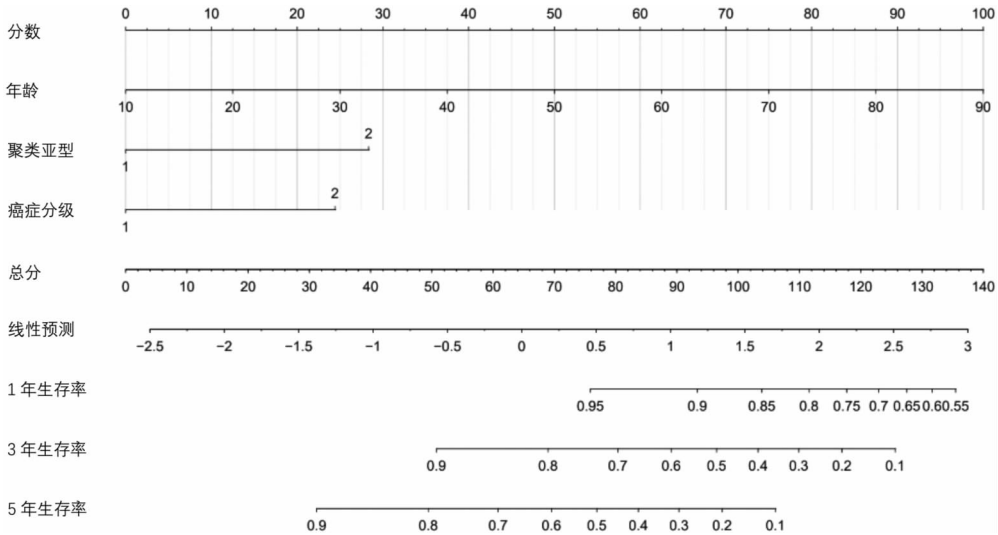

图 3 TCGA – LGG 患者多因素 Cox 回归分析 Nomogram 图

Fig. 3 Nomogram of multifactorial Cox regression analysis in TCGA – LGG patients

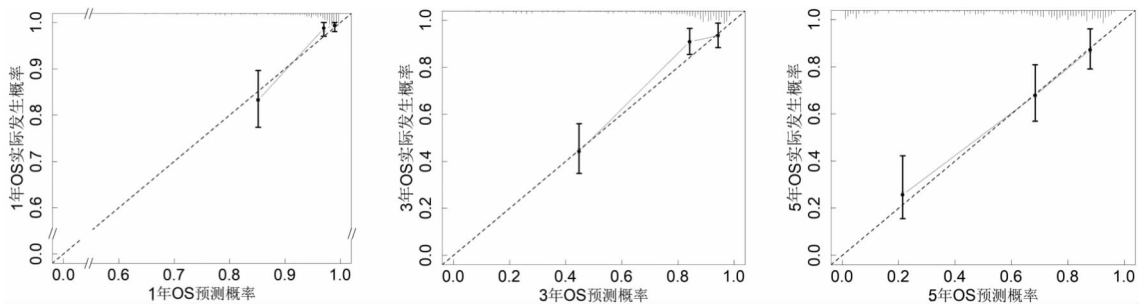

图 4 TCGA – LGG 患者多因素 Cox 回归分析 1、3 和 5 年校准曲线

Fig. 4 Multifactorial Cox regression analysis of 1, 3and 5 – year calibration curves in TCGA – LGG patients

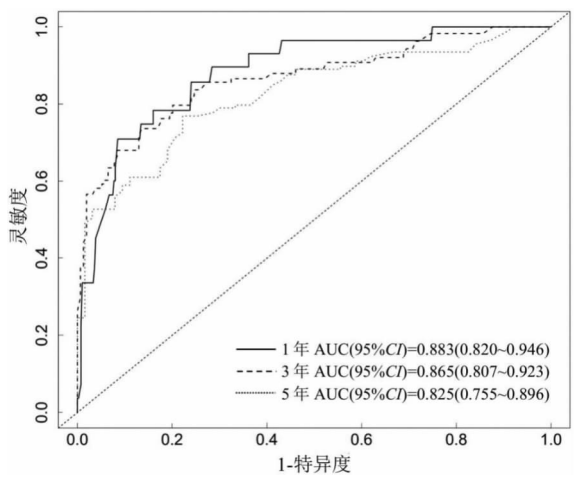

图 5 TCGA – LGG 患者多因素 Cox 回归分析 ROC 曲线

Fig. 5 Cox regression analysis of ROC curves in TCGA – LGG patients

表达量以及获得的样本聚类标签整合到一起,构建一个十折交叉验证 RF 模型。按 70% 和 30% 的比例划分训练集和内部验证集,十折交叉验证的平均预测准确率为 87. 81%。使用 CGGA 数据库中两个 LGG 患者数据集进行外部验证。

训练集、内部验证集和外部验证集的 C 指数、Brier 评分和 log – rank *P* 值结果如表 4 所示。训练集、内部验证集的 C 指数大于 0. 7, Brier 评分小于 0. 1,模型性能良好;两个 CGGA 数据集的 C 指数大于 0. 5, Brier 评分小于 0. 25,模型性能较好,表明使用的特征选择算法对预测 LGG 的亚型具有鲁棒性。绘制 Kaplan – Meier 生存曲线,表明 RF 模型能够很好地区分该 LGG 内部验证集和 CGGA 外部验证集的亚型类型,生存差异均有统计学意义(*P* <0. 05),见图 6。

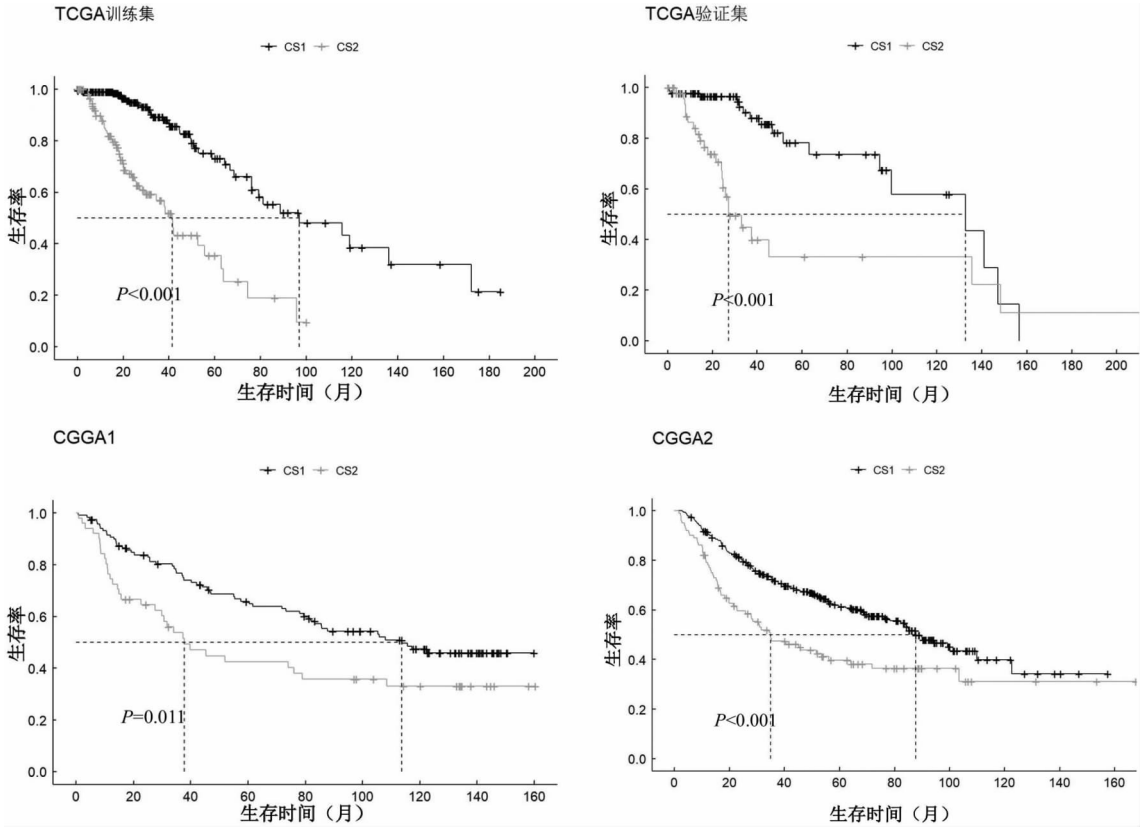

图 6　TCGA 训练集、内部验证集和 CGGA 外部验证集 LGG 患者生存曲线

Fig. 6　Survival curves of LGG patients for TCGA training set, internal validation set and CGGA external validation set

表 4　TCGA 训练集、内部验证集和 CGGA 外部验证集的模型性能

Table 4　Model performance for TCGA training set, internal validation set and CGGA external validation set test set

数据集	C 指数	Brier 评分	log－rank$P$ 值
TCGA 训练集(70%)	0.717	0.044	1e－12
TCGA 验证集(30%)	0.721	0.066	5e－05
CGGA1	0.574	0.179	1e－02
CGGA2	0.572	0.128	1e－04

3　讨　论

LGG 具有高度的遗传异质性，传统组织病理学分级中同一类别的预后也存在异质性[15]，需要寻求其他分类方法；多组学聚类的方法适用于癌症等遗传驱动疾病的无监督聚类，目的是发现新的疾病亚型，对疾病的诊断和治疗意义重大[16]。本研究在 LGG 多组学数据的基础上通过 MOVICS 综合十种聚类结果确定 LGG 的最终分子亚型，然后通过 RF 构建预后模型，为临床更准确的癌症诊断和治疗提供了可能。

为了提高聚类的鲁棒性，MOVICS 通过借鉴共识集成[17]的思想，对不同的聚类结果进行整合，得到的共识矩阵代表了样本的成对相似性。本文采用综合聚类方法对 LGG 多组学数据进行整合，最终将 TCGA 的 502 个 LGG 患者分为两型。多因素 Cox 分析结果显示年龄、癌症分级、聚类亚型具有统计学意义（P < 0.05），可作为低级别胶质瘤患者的预后预测因素，预测 LGG 患者 1 年、3 年和 5 年总生存期的 ROC 曲线 AUC 值都在 0.82 以上，校准曲线也接近标准曲线，提示模型拥有较高的预测能力。Gao 等[18]关于过氧化物酶体相关基因特征预测 LGG 患者预后的结果中，风险评分和年龄是 LGG 患者的预后因素；Zhao 等[19]研究了铜氧化酶基因预测 LGG 患者的生存率，结果显示风险评分、年龄和癌症分级是 LGG 患者的预后因素，本研究与之有相同之处。在 LGG 患者中，年龄越大，患者的死亡风险越高，生存率越低，预后越差；在癌症分级中，WHO G3 级患者的死亡风险是 WHO G2 级患者 2.733 倍，在两亚型中，LGG 患者 CS2 组的死亡风险是 CS1 组患者的 3.210 倍，且 CS2 组患者的平均年龄为 46.6 岁，大于 CS1 组的 40.9 岁，WHO G3 级患者在 CS2 组的占比也多于 CS1 组，故 CS2 组患者的死亡风险更高，预后较差。这提示临床应对高龄、WHO G3 级患者给予一定的重视，且使用多组学数据进行聚类发现的疾病亚型，为 LGG 患者的诊断和治

疗提供了新思路。

构建的随机森林分类预测模型,在 TCGA 内部验证集表现出较好的分类性能,在两个 CGGA 外部验证集上,成功预测出样本的亚型标签,亚型之间的生存差异都具有统计学意义,预后差的亚型与 TCGA 训练集相一致,均为 CS2 组。通过内部和外部验证,说明该模型具有较好的分类预测性能,为 LGG 亚型提供了理论参考,具有一定的实际意义。

本研究也具有一定的局限性,第一,在获得聚类标签时,仅用到 mRNA 和 DNA 甲基化的泛癌基因,可能会缺失一些特有的关键基因;第二,在构建随机森林分类预测模型时,由于样本量的原因,仅使用了 mRNA,今后会试图整合多个组学数据进行构建模型;第三,分类模型验证时,仅用到 CGGA 数据库,为了检验模型的泛化能力,接下来会寻找其它的数据集进行验证,如 GEO 数据库。

利益冲突声明 本研究不存在任何利益冲突

参考文献

[1] 蔡祥,王仁东,王世佳,等. 胶质母细胞瘤恶性进展中不同细胞亚群的动态轨迹和细胞通讯网络[J]. 北京大学学报:医学版,2024,56(2):199 - 206.
Cai X, Wang RD, Wang SJ, et al. Dynamic trajectory and cell communication of different cell clusters in malignant progression of glioblastoma[J]. Journal of Peking University(Health Sciences), 2024, 56(2): 199 - 206.

[2] 兰宁,张永超,李淼,等. 结合正则化方法的区块森林在癌症多组学数据预后预测中的应用研究[J]. 中国卫生统计,2023,40(3):382 - 385.
Lan N, Zhang YC, Li M, et al. Application of block forest combined with regularization method in prognostic prediction of cancer multi - omics data[J]. Chinese Journal of Health Statistics, 2023, 40(3): 382 - 385.

[3] 龙智平,王帆. 多组学整合分析的设计及统计方法在肿瘤流行病学研究中的应用[J]. 中华流行病学杂志,2020,41(5):788 - 793.
Long ZP, Wang F. Study design and statistical methods used for integrative analysis on multi - omics in cancer epidemiology[J]. Chinese Journal of Epidemiology, 2020, 41(5): 788 - 793.

[4] Liu XD, Wang WH, Zhang XL, et al. Metabolism pathway - based subtyping in endometrial cancer: An integrated study by multi - omics analysis and machine learning algorithms[J]. Molecular Therapy. Nucleic Acids, 2024, 35(2): 102155.

[5] Chakraborty S, Sharma G, Karmakar S, et al. Multi - OMICS approaches in cancer biology: New era in cancer therapy[J]. Biochim Biophys Acta Mol Basis Dis, 2024, 1870(5): 167120.

[6] 钟雅婷,林艳梅,陈定甲,等. 多组学数据整合分析和应用研究综述[J]. 计算机工程与应用,2021,57(23):1 - 17.
Zhong YT, Lin YM, Chen DJ, et al. Review on integration analysis and application of multi - omics data[J]. Computer Engineering and Applications, 2021, 57(23): 1 - 17.

[7] Cai YY, Wang SF. Deeply integrating latent consistent representations in high - noise multi - omics data for cancer subtyping[J]. Briefings in Bioinformatics, 2024, 25(2): bbae061.

[8] 高振博. 基于多组学数据的聚类方法研究[D]. 大连:大连理工大学,2020.
Gao ZB. Research on clustering method based on multi - omics data[D]. Dalian: Dalian University of Technology, 2020.

[9] Lu XF, Meng JL, Zhou YJ, et al. MOVICS: an R package for multi - omics integration and visualization in cancer subtyping[J]. Bioinformatics, 2021, 36(22/23): 5539 - 5541.

[10] Guan Y, Yue SY, Chen YD, et al. Molecular cluster mining of adrenocortical carcinoma via Multi - Omics data analysis Aids precise clinical therapy[J]. Cells (Basel, Switzerland), 2022, 11(23): 3784.

[11] Macaulay BO, Aribisala BS, Akande SA, et al. Breast cancer risk prediction in African women using Random Forest Classifier[J]. Cancer Treatmentand Research Communications, 2021, 28: 100396.

[12] Halder RK,Uddin MN, Uddin MA, et al. ML - CKDP: machine learning - based chronic kidney disease prediction with smart web application[J]. Journal of Pathology Informatics, 2024, 15: 100371.

[13] Chalise P, Fridley BL. Integrative clustering of multi - level 'omic data based on non - negative matrix factorization algorithm[J]. PLOS One, 2017, 12(5): e0176278.

[14] 王俊丰,贾晓霞,李志强. 基于 K - means 算法改进的短文本聚类研究与实现[J]. 信息技术,2019(12):76 - 80.
Wang JF, Jia XX, Li ZQ. Research and implementation of short text clustering based on improved K - means algorithm[J]. Information Technology, 2019(12): 76 - 80.

[15] Du ZX, Wang YQ, Liang JQ, et al. Association of glioma CD44 expression with glial dynamics in the tumour microenvironment and patient prognosis[J]. Computational and Structural Biotechnology Journal, 2022, 20: 5203 - 5217.

[16] Li L, Zhang WW, Sun YJ, et al. A clinical prognostic model of oxidative stress - related genes linked to tumor immune cell infiltration and the prognosis of ovarian cancer patients[J]. Heliyon, 2024, 10(7): e28442.

[17] Rappoport N, Shamir R. Multi - omic and multi - view clustering algorithms: review and cancer benchmark[J]. Nucleic Acids Research, 2018, 46(20): 10546 - 10562.

[18] Gao DD, Zhou QY, Hou DQ, et al. A novel peroxisome - related gene signature predicts clinical prognosis and is associated with immune microenvironment in low - grade glioma[J]. PeerJ, 2024, 12: e16874.

[19] Zhao ZR, Ma YH, Liu Y, et al. A cuproptosis - based prognostic model for predicting survival in low - grade glioma[J]. Aging, 2024, 16(10): 8697 - 8716.

收稿日期:2024-04-16