Advances in Applied Mathematics 应用数学进展, 2022, 11(6), 4022-4031
Published Online June 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.116430

Hans汉斯

基于加权共表达网络的肺癌的预后预测模型构建

刘晓东，宋　凯

青岛大学数学与统计学院，山东　青岛

收稿日期：2022年5月27日；录用日期：2022年6月19日；发布日期：2022年6月29日

摘　要

肺癌是我国恶性肿瘤当中发病率第一，死亡率第一的恶性肿瘤。目前我国肺癌患者的五年平均生存率仍不足20%，对于肺癌患者的诊疗效果仍有待提升。所以有必要进一步研究肺癌患者的预后影响因素，建立预后预测模型，预测患者预后风险及生存情况，可以帮助临床医生更好地判断患者的预后情况，并发现新的疾病相关因素。本文基于机器学习算法以及TCGA数据库中肺癌患者的多组学数据，以探究肺癌患者生存期是否超过五年为目标，利用加权共表达网络算法找到影响肺癌患者预后生存的关键特征基因。通过结合分类算法和加权共表达网络算法来构建预后预测模型，并使用AUC值对模型的分类效果进行评估，最终得到使用基于KNN回归构建的预后模型效果最好，能够较为准确地判断肺癌患者生存时间是否超过5年。

关键词

加权共表达网络，多组学基因数据，机器学习，预后预测

Construction of Lung Cancer Prognostic Prediction Model Based on Weighted Co-Expression Network

Xiaodong Liu, Kai Song

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

Received: May 27th, 2022; accepted: Jun. 19th, 2022; published: Jun. 29th, 2022

Abstract

Lung cancer is the malignant tumor with the highest incidence and the highest mortality rate in

文章引用: 刘晓东, 宋凯. 基于加权共表达网络的肺癌的预后预测模型构建[J]. 应用数学进展, 2022, 11(6): 4022-4031.
DOI: 10.12677/aam.2022.116430

my country. At present, the five-year average survival rate of lung cancer patients in my country is still less than 20%, and the diagnosis and treatment of lung cancer patients still need to be improved. Therefore, it is necessary to further study the prognostic factors of lung cancer patients, and establish a prognostic prediction model to predict the prognostic risk and survival of patients, which can help clinicians better judge the prognosis of patients and discover new disease-related factors. Based on the machine learning algorithm and the multi-omics data of lung cancer patients in the TCGA database, this paper aims to explore whether the survival time of lung cancer patients exceeds five years, and uses the weighted co-expression network algorithm to find the key feature genes that affect the prognosis and survival of lung cancer patients. By combining the classification algorithm and the weighted co-expression network algorithm, the prognosis prediction model is constructed, and the AUC value is used to evaluate the classification effect of the model. Finally, the prognostic model constructed based on KNN regression has the best effect and can more accurately judge the survival of lung cancer patients whether the time is more than 5 years.

Keywords

Weighted Co-Expression Network, Multi-Omics Genetic Data, Machine Learning, Prognosis Prediction

1. 引言

一般来说，不同组织部位的癌症往往是由不同的基因突变引起的，但即使是发生在同一组织部位的癌症，引发癌症的基因突变也可以是不同的。因此，除了传统的临床所见特征如 TNM 分期[1]、年龄、家族史等，如何通过组学特征对癌症患者的预后情况进行预测也成为了癌症预后分析的热点研究[2] [3]。随着基因组学、转录组学技术快速发展，积累了越来越多的基因组学数据，如 TCGA、ICGC 等大型基因组学数据库，提供了匹配正常细胞基因组的数千个肿瘤的基因组的体细胞突变、体细胞拷贝数变异、DNA 甲基化、mRNA 表达及患者的临床资料等海量数据。

加权基因共表达网络分析(WGCNA, Weighted Correlation Network Analysis)是用来描述不同样品之间基因关联模式的系统生物学方法，可以用来鉴定高度协同变化的基因集，并根据基因集的内连性和基因集与表型之间的关联鉴定候补生物标记基因或治疗靶点。

我们应用加权共表达网络在肺癌中构建差异表达基因的共表达模块，分析模型特征内基因表达与临床特征的相关性，识别枢纽基因(hub gene)。主要的分析步骤包括以下四方面：建立基因加权相关网络；鉴定共表达基因模块；进行模块与临床特征的相关性分析；挑选模块的关键基因。

现已有的研究大多应用单一组学数据进行分析，本文创新点主要是基于多组学多模态的数据，有助于挖掘各模态数据的特征信息，填补单一组学数据中的信息缺失，综合分析患者的相关指标与预后因子之间的关系。同时在 WGCNA 的基础上运用 lasso-logistic 等机器学习方法，大大提高了筛选目标特征相关基因的精准率。

2. 相关研究综述

基因组学与影像组学均常用于肿瘤的精准诊疗[4] [5] [6]。基因组学以分子生物学和信息网络技术从

整体上探索全基因组在生命活动中的作用及规律。基因组学可辅助研究肿瘤发生发展的分子机制、发现致癌基因的突变位点和表达通路、分子标记物，但费用和技术要求较高，且获取分子标记物时只能单点单次采集。而肿瘤细胞不仅外部形态不同，内部组织特性也不同，加之所处环境的多样性，肿瘤诊断必须考虑肿瘤差异性和异质性信息[7]，仅凭基因组学难以全面描述肿瘤组织的时空异质性。影像组学则是从影像中提取定量特征，用于肿瘤的早期诊断、分期和预后预测等[8] [9] [10]。影像组学在诊疗过程中能随时跟踪肿瘤的发展状况，但特异性较差，获取重要影像标志物时，生物和分子机制的解释性不足。肿瘤识别是典型的小样本问题，基于影像提取大量底层特征进行肿瘤分类识别时，计算时间长，分类识别精度也有限。另外，临床中要求所提取的特征是稳定、可重复的，但不同的采集设备、参数设置以及特征选择方法均会使所获取的影像特征存在差异，故基于大量底层特征获取影像标记物时，需克服图像采集和特征计算标准化的问题。

国内外在癌症预后领域已有不少研究成果，杨娟等[11]采集 128 例肝内胆管癌患者临床资料，使用 Logistic 回归分析法筛选预后危险因素以便预测患者术后复发风险，Zhu 等[12]设计了监督式主成分回归法，将基因表达数据与病理图像数据进行融合，用于肺癌的生存期预测。这种前端融合一定程度提升了生存预测性能，但仅融合了数据低维的原始特征，包含大量噪声和冗余信息。Mobadersany 等[13]综合考虑组学信息和病理图像信息，对于癌症生存期的预测提出了生存卷积神经网络(SCNN)法，用于多形性胶质细胞瘤的生存期预测。Lai 等[14]开发了一种结合基因表达异质性数据和临床数据的双峰深度神经网络预测肺癌患者的 5 年总体生存状态(AUC = 0.816)。

随着各种癌症组学数据的积累，诸多研究工作开始尝试从多模态、多任务的角度出发，融合多组学和病理图像数据，进一步改进癌症生存期预测。Zhang 等[15]提出了一个基于多核学习框架的多模态数据前端融合模型，成功融合基因表达、拷贝数变异、甲基化等多种组学数据，进行了癌症预后的预测。多模态数据融合算法充分考虑并有效提取了模态间的关联性，因此这类基于多模态数据融合的模型在癌症生存期预测中均取得了不错的效果。

3. 基于加权共表达网络肺癌预后模块鉴定

3.1. 肺癌预后模型结构说明

我们将建立肺癌预后模型工作分为三个模块：数据模块、特征基因选择模块和预后模型建立模块。第一个模块为数据模块，主要完成数据获取与预处理的功能。第二个模块为特征基因选择模块，在该模块中，我们要在全基因组数万个基因中剔除与肺癌患者预后相关性不大的基因，筛选出与肺癌患者预后密切相关的少数基因。第三个模块为预后模型建立模块，在该模块中，我们的主要任务是用四种机器学习算法来构建预测肺癌患者总生存时间是否超过 5 年的预后模型，并对四种模型的分类的 AUC 值进行计算和比对，挑选效果最好的预后模型。图 1 展示了构造模型的结构层次图。

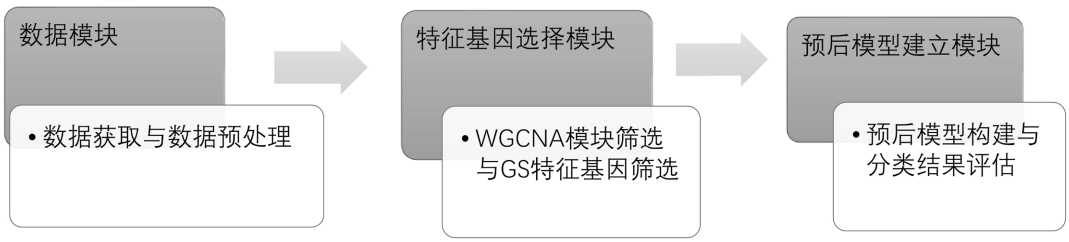

Figure 1. Model hierarchy diagram
图 1. 模型的层次结构图

3.2. 实验数据获取与预处理

TCGA (The Cancer Genome Atlas，癌症基因图谱)是由美国国家癌症研究所和美国国家人类基因组研究所于 2006 年联合启动的项目。该项目收录了人类癌症的各种数据，如临床数据、miRNA 表达数据、甲基化等数据。

本文所采用的数据来自于 TCGA 数据库，主要使用肺癌患者的基因表达量、拷贝数变异、外显子基因表达量和临床数据，其中临床数据主要使用生存信息。从 UCSC 官网(https://xenabrowser.net/heatmap/)中下载 TCGA 肺癌基因表达量、拷贝数变异、外显子基因表达量和临床数据。在对数据进行分析之前，我们首先剔除掉缺失关键信息的样本，并且只采用同时含有三个组学信息的癌症患者样本，最终得到 995 个样本的基因数据，清洗数据以后对数据进行标准化处理，同时根据临床数据中的总生存时间(OS.time)将患者按生存时间是否超过 5 年分为 OSfive = 1 和 OSfive = 0 两类，其中 OSfive 表示患者五年生存期，OSfive = 1 即为患者生存时间超过 5 年。

3.3. 加权共表达网络分析

3.3.1. 加权共表达网络概述

加权共表达网络(Weighted Gene Correlation Network Analysis, WGCNA)可以将基因网络根据表达相似性划分成不同的模块，并分析模块与特定表型之间的相关关系。在 WGCNA 算法[16]中，我们用节点代表基因，用节点与节点之间的连线代表基因表达相关性，邻接度用来表示节点之间的关系强弱，网络内的所有基因的邻接度可以形成一个邻接矩阵。在无尺度网络中，只有极少数节点与很多节点有关，因此，一个模块中有许多基因但关键基因只有几个，我们选择最佳加权系数使基因服从无尺度网络分布。

3.3.2. 加权共表达网络构建

传统方法中描述两个基因之间的关联程度一般会指定一个筛选阈值，但这种方法会丢失基因的变化趋势信息，为了解决这些问题，加权共表达网络对基因表达值之间的相关系数取 β 次幂，对于基因 i 和 j，相关系数为 r_{ij}，取 β 次幂后得到 a_{ij}，可以得到 $a_{ij} = r_{ij}^{\beta}$，最终将基因间相关性的强弱的差别放大，这样的好处是使强弱关系更为分明有利于后续聚类识别。

加权共表达网络的构建基于 RStudio 软件中的“WGCNA”函数包[17]来实现，加载“WGCNA”包，使用“goodSamplesGenes()”函数检查数据是否有异常基因，使用“hclust()”函数绘制聚类图，再应用“cutree()”函数剔除离群样本，获得较为一致的基因表达数据。然后使用“pickSoftThreshold()”函数进行自动网络拓扑分析，依据无尺度网络原则确定软阈值参数(β)，根据图 2 选择合适的 β 保证拟合优度在 0.9 以上，将相关性矩阵转化为邻接矩阵，构建一个趋向无尺度网络的加权共表达网络。

3.3.3. 肺癌预后模块的选择

将选择的 β 值代入“blockwiseModules()”函数，设置最小模块基因数、模块合并阈值等参数，划分模块并合并相似模块。然后计算模块特征向量和临床性状之间相关系数矩阵，并对相关系数进行检验。图 3 将相关系数矩阵进行热力图可视化，在图 4 中挑选 p 值小于 0.05 且相关性较高的模块作为备选模块。

3.3.4. 肺癌预后特征基因的选择

基因的模块身份(Module Membership, MM)用于描述基因在所有样本中的表达谱与某个特征向量基因表达谱的相关性，即对 module eigengene 进行相关性分析就可以得到 MM 值，所以 MM 值本质上是一个相关系数，如果基因和某个 module 的 MM 值为 0，说明二者根本不相关，该基因不属于这个 module，如果 MM 的绝对值接近 1，说明基因与该 module 相关性很高。

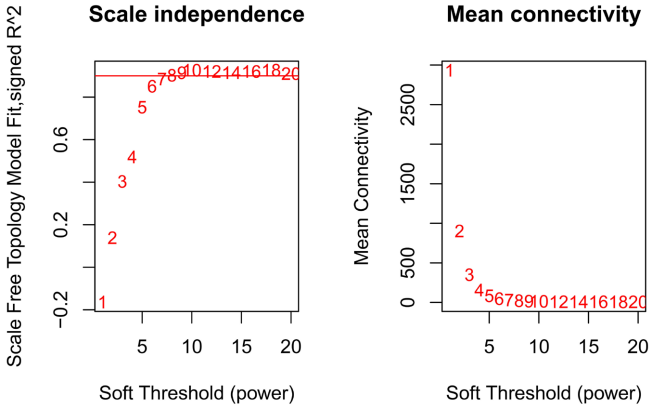

Figure 2. Determining soft threshold parameters based on scale-free network principle
图 2. 依据无尺度网络原则确定软阈值参数

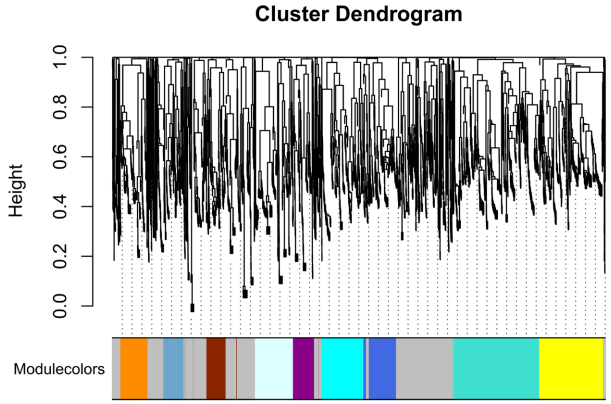

Figure 3. Module division and module merging diagram
图 3. 模块划分与模块合并图

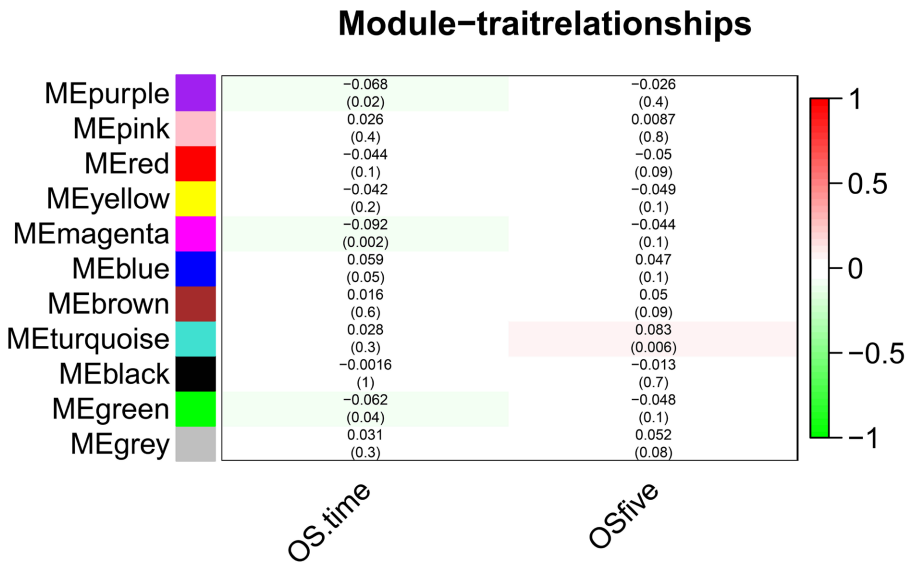

Figure 4. Module correlation coefficient matrix heatmap
图 4. 模块相关系数矩阵热力图

基因显著性(Gene Significance, GS)用于描述模块内基因与 OSfive 关联程度，将基因的表达量与对应的表型数值进行相关性分析，最终的相关系数的值就是 GS。GS 反映出基因表达量与表型数据的相关性，GS 越高说明基因在预后方面越有意义。

根据图 5 结果，我们挑选备选模块中 GS 绝对值大于 0.1 且 MM 绝对值大于 0.8 的基因，取备选模块中筛选出的基因的并集作为所选择的特征基因。

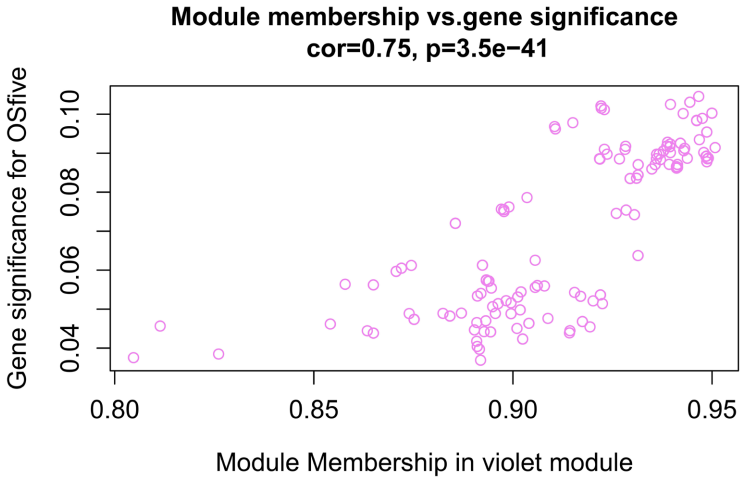

Figure 5. Module and gene expression correlation plot
图 5. 模块与基因表达相关性图

3.4. 结果

经过筛选后我们得到 51 个特征基因，考虑到所选基因之间可能存在多重共线性问题，我们用方差膨胀因子(Variance Inflation Factor, VIF)方法处理存在严重共线性的特征基因，应用 R 软件中的 car 包，使用“vif()”函数将 VIF 数值大于 100 的特征基因剔除，最终获得 39 个特征基因。在这一环节我们通过 WGCNA 算法完成了特征基因选择。

3.5. 预测模型构建

3.5.1. 逻辑回归模型

逻辑回归(Logistic Regression，LR)算法是在线性回归模型的基础上，添加 sigmoid 函数来完成映射，将连续值转化为(0, 1)之间的一个概率值，通过这个概率值我们可以解决分类问题。本文所解决的肺癌分类问题为二项分类问题，因此选择二项逻辑回归模型作为分类模型，二项逻辑回归模型的条件概率分布表示如下所示。

$$P(Y=1|x)=\frac{e^{wx+b}}{1+e^{wx+b}} \tag{1}$$

$$P(Y=0|x)=\frac{1}{1+e^{wx+b}} \tag{2}$$

其中，$x\in R^n$ 是输入变量，$Y\in\{0,1\}$ 是输出变量，w 叫做权重，b 叫做偏置量。

3.5.2. 支持向量机算法

支持向量机(Support Vector Machines, SVM)是一种强大的分类器构建方法，它的学习策略为间隔最大

化，即找寻一个可以最大化将训练数据分隔开的超平面。在样本空间中，划分超平面可以通过如下方程来描述：

$$w^{\mathrm{T}}x+b=0 \tag{3}$$

其中 w 为法向量，决定了超平面的方向；b 为位移项，决定了超平面与原点之间的距离。对于所有训练数据，w 和 b 应该满足以下两个不等式：

$$w^{\mathrm{T}}x_i+b\geq+1,\ \ y_i=+1 \tag{4}$$

$$w^{\mathrm{T}}x_i+b\leq-1,\ \ y_i=-1 \tag{5}$$

这些满足 $y_i\left(w^{\mathrm{T}}x_i+b\right)=1$ 条件的 x_i 则称为支持向量。所以支持向量机可理解为一个求解二次凸优化问题，如下所示：

$$\min\frac{1}{2}\|w\|^2 \tag{6}$$

$$\text{s.t.}\ \ y_i\left(w^{\mathrm{T}}x_i+b\right)\geq1 \tag{7}$$

3.5.3. K 近邻

K 近邻法(K-Nearest Neighbor, KNN)是一种常用于分类的算法，K 近邻利用距离来区分类型的思路是：通过离测试点最近的 k 个已知点的类型来决定测试点的类型，其中 k 是人为设定的数值，k 的取值对模型的好坏有非常重要的影响。

在训练集中，所有的数据和数据所对应的分类标签已知。每当有一个新的测试数据输入时，KNN 算法会计算测试数据与每一个训练数据之间的距离，并将距离进行排序。然后 KNN 算法会找出前 k 个与新数据点最近的训练集样本点和它们相应的标签，该测试数据对应的类别就是这 k 个数据中出现次数最多的分类。

KNN 是一种懒惰的学习算法，训练数据不需要执行任何泛化，同时，两个数据点的距离计算至关重要，不同的距离计算公式对 KNN 的模型分类效果产生巨大影响。在实际中两种最常用的距离计算公式是欧式距离和曼哈顿距离，计算公式如下：

$$\text{欧式距离：}\ \ d\left(x,y\right)=\sqrt{\sum_{k=1}^{n}\left(x_k-y_k\right)^2} \tag{8}$$

$$\text{曼哈顿距离：}\ \ d\left(x,y\right)=\sqrt{\sum_{k=1}^{n}\left|x_k-y_k\right|} \tag{9}$$

3.5.4. 随机森林

随机森林(Random Forest)是一种比较新的机器学习模型集成学习方法，随机森林对多元共线性不敏感，结果对缺失数据和非平衡数据比较稳健。随机森林是用随机的方式建立一个森林，森林由很多的决策树组成，并且每一棵决策树之间是没有关联的，得到随机森林模型后，当新样本进入时，随机森林的每一棵决策树会分别进行判断，对于分类问题通常使用投票法，得到最多票数类为最终模型输出。

若训练集大小为 N，对于随机森林中每棵树而言，随机且有放回地从训练集中抽取 N 个训练样本作为该树的训练集。随机森林算法通过随机抽取训练样本作为每棵树的训练集，可以使得构成森林的每棵树的训练集都不一样，进而减少不同树之间的相关性，使分类效果更好。

通过设计多个决策树并将它们的预测结果相结合起来，随机森林大大降低了过拟合的风险，并且使得构建出的随机森林具有很好的抗噪能力。

4. 结果

4.1. 肺癌预后模型的实现

在上一步的工作中，我们利用 WGCNA 算法从全基因组上万个基因中筛选出了与肺癌患者预后生存密切相关的 39 个特征基因。图 6 的结果是运用主成分分析(PCA)做出的主成分碎石图，选取主成分所占百分比前六的六个主成分，通过主成分分析剔除测试集中的异常样本，见图 7。我们使用机器学习算法构建预测肺癌患者总生存时间能否超过 5 年的预后模型，通过这个二分类模型可以更好的辅助医生将不同患者分入不同的危险组别之中。

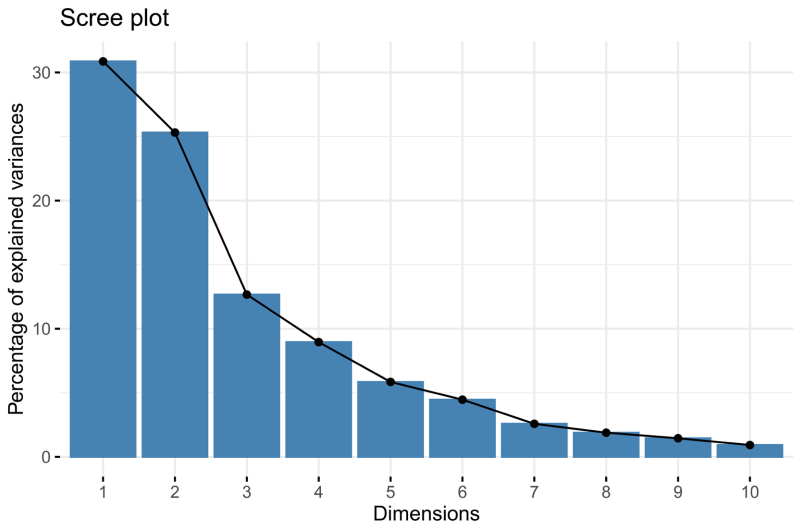

Figure 6. Scree plot
图 6. 碎石图

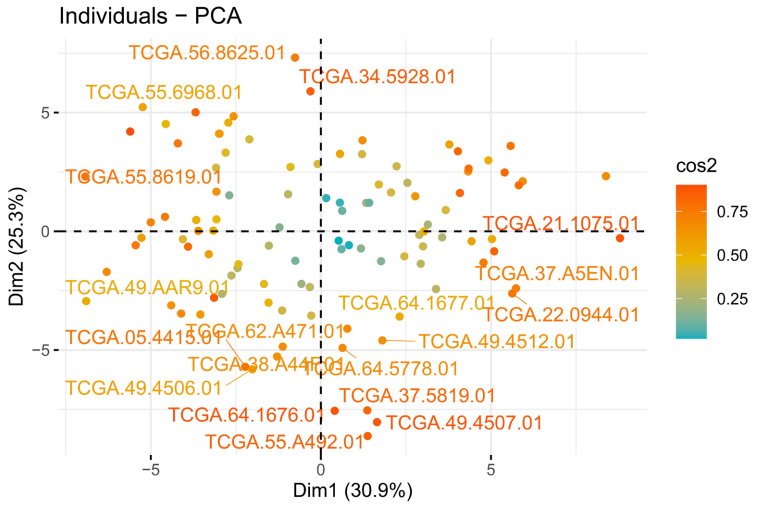

Figure 7. Variables on dimensions 1 and 2
图 7. 变量在主成分 1 和 2 上表示

在本节工作中，我们同时使用逻辑回归、支持向量机、K 近邻、随机森林四种机器学习算法构造出四个不同的肺癌患者预后模型，通过对四种预后模型效果进行比对，筛选出一个最好的预后模型。由于

肺癌患者基因数据有限，为得到可靠模型，在训练模型时使用 10 折交叉验证，通过交叉验证方法在一定程度上避免过拟合，使预后模型具有更好的泛化能力。

4.2. 预后模型分类效果

在实验过程中，我们对所有模型都采用 10 折交叉验证，通过将数据集按照 8:2 划分训练集和测试集，通过多次划分多次训练，实验结果取十次实验的平均值来提高模型的泛化能力，在这一环节我们使用 AUC 指标来评价模型的好坏。

ROC 曲线如图 8 所示，其中红色曲线为逻辑回归 ROC 曲线，蓝色曲线为支持向量机 ROC 曲线，紫色曲线为 K 近邻 ROC 曲线，绿色曲线为随机森林 ROC 曲线。

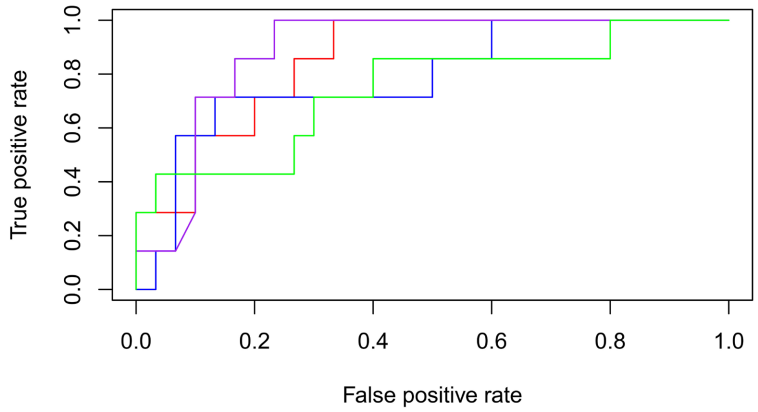

Figure 8. ROC curves of four prognostic models
图 8. 四种预后模型的 ROC 曲线

表 1 展示了四种机器学习算法在预测肺癌患者 5 年总生存时间的 AUC 值，从表 1 中可以看出，四种模型的 AUC 值都在 0.7 以上，其中 KNN 算法构建的模型效果最佳，AUC 值达到 0.888。

Table 1. Classification results of prognostic model based on characteristic gene construction
表 1. 基于特征基因构建的预后模型分类结果表

机器学习算法	AUC 值
逻辑回归	0.857
支持向量机	0.790
K 近邻	0.888
随机森林	0.743

5. 结论

本文考虑肺癌患者基因的多组学多模态数据，通过 TCGA 数据库获取相关组学数据，对数据进行预处理后，基于加权共表达网络算法筛选出来 39 个特征基因，运用机器学习算法构建了肺癌患者预后预测模型，对所有的预后模型效果进行比较，得到使用基于 KNN 算法构建的预后模型效果最好，AUC 值达到 0.888，能够较为准确地判断肺癌患者生存时间是否超过 5 年。

参考文献

[1] Deshmukh, P.R. and Phalnikar, R. (2020) TNM Cncer Stage Detection from Unstructured Pathology Reports of Breast

Cancer Patients: Proceeding of International Conference on Computational Science and Applications. Saint Petersburg, 1-4 July 2019, 411-418. https://doi.org/10.1007/978-981-15-0790-8_40

[2] Zhu, X., Yao, J. and Huang, J. (2016) Deep Convolutional Neural Network for Survival Analysis with Pathological Images. 2016 *IEEE International Conference on Bioinformatics and Biomedicine*, Shenzhen, 15-18 December 2016, 544-547. https://doi.org/10.1109/BIBM.2016.7822579

[3] Zhu, X., Yao, J., Luo, X., Xiao, G., Xie, Y., Gazdar, A. and Huang, J. (2016) Lung Cancer Survival Prediction from Pathological Images and Genetic Data—An Integration Study. 2016 *IEEE* 13*th International Symposium on Biomedical Imaging*, Prague, 13-16 April 2016, 1173-1176. https://doi.org/10.1109/ISBI.2016.7493475

[4] Yao, J., Wang, S., Zhu, X. and Huang, J. (2016) Imaging Biomarker Discovery for Lung Cancer Survival Prediction. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Athens, 17-21 October 2016, 649-657. https://doi.org/10.1007/978-3-319-46723-8_75

[5] Wang, H., Xing, F., Su, H., Stromberg, A. and Yang, L. (2014) Novel Image Markers for Non-Small Cell Lung Cancer Classification and Survival Prediction. *BMC Bioinformatics*, **15**, Article No. S1. https://doi.org/10.1186/1471-2105-15-S12-S1

[6] Bøvelstad, H.M., Nygård, S., Størvold, H.L., Aldrin, M., Borgan, Ø., Frigessi, A. and Lingjærde, O.C. (2007) Predicting Survival from Microarray Data—A Comparative Study. *Bioinformatics*, **23**, 2080-2087. https://doi.org/10.1093/bioinformatics/btm305

[7] Dagogo-Jack, I. and Shaw, A.T. (2018) Tumour Heterogeneity and Resistance to Cancer Therapies. *Nature Reviews Clinical Oncology*, **15**, 81-94. https://doi.org/10.1038/nrclinonc.2017.166

[8] 陈钰莹, 黄小华, 刘念, 唐玲玲, 胡云涛. 影像组学在胰腺癌中的临床研究进展[J]. 磁共振成像, 2021, 12(8): 108-110.

[9] 侯东辉, 吴宁. 影像组学在肺癌中的应用进展[J]. 癌症进展 2019, 17(2):128-130+144.

[10] Choi, W., Oh, J.H., Riyahi, S., Liu, C.J., Jiang, F., Chen, W., White, C., Rimner, A., Mechalakos, J.G. and Deasy, J.O. (2018) Radiomics Analysis of Pulmonary Nodules in Low-Dose CT for Early Detection of Lung Cancer. *Medical Physics*, **45**, 1537-1549. https://doi.org/10.1002/mp.12820

[11] 杨娟, 母齐鸣, 谭琴. 影响肝内胆管癌手术预后的危险因素 Logistic 回归分析[J]. 江苏大学学报: 医学版, 2019, 29(2): 161-165.

[12] Zhu, X., Yao, J., Xin, L., Xiao, G. and Huang, J. (2016) Lung Cancer Survival Prediction from Pathological Images and Genetic Data—An Integration Study. *IEEE International Symposium on Biomedical Imaging*.

[13] Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D.A., Barnholtz-Sloan, J.S., Vega, J.E.V., Brat, D.J. and Cooper, L.A. (2018) Predicting Cancer Outcomes from Histology and Genomics Using Convolutional Networks. *Proceedings of the National Academy of Sciences of the United States of America*, **115**, E2970-E2979. https://doi.org/10.1073/pnas.1717139115

[14] Lai, Y.H., Chen, W.N., Hsu, T.C., Lin, C. and Wu, S. (2020) Overall Survival Prediction of Non-Small Cell Lung Cancer by Integrating Microarray and Clinical Data with Deep Learning. *Scientific Reports*, **10**, Article No. 4679. https://doi.org/10.1038/s41598-020-61588-w

[15] Zhang, Y., Ao, L., Chen, P. and Wang, M. (2016) Improve Glioblastoma Multiforme Prognosis Prediction by Using Feature Selection and Multiple Kernel Learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **13**, 825-835. https://doi.org/10.1109/TCBB.2016.2551745

[16] Zhang, B. and Horvath, S. (2005) A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, **4**, Article No. 17. https://doi.org/10.2202/1544-6115.1128

[17] Langfelder, P. and Horvath S (2008) WGCNA: An R package for Weighted Correlation Network Analysis. *BMC Bioinformatics*, **9**, Article No. 559. https://doi.org/10.1186/1471-2105-9-559