计算机科学
COMPUTER SCIENCE

# 基于相似网络融合算法的癌症亚型预测

张晓茜, 李东喜

引用本文

张晓茜, 李东喜. 基于相似网络融合算法的癌症亚型预测[J]. 计算机科学, 2024, 51(6A): 230500006-7.

ZHANG Xiaoxi, LI Dongxi. Cancer Subtype Prediction Based on Similar Network Fusion Algorithm[J]. Computer Science, 2024, 51(6A): 230500006-7.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230500006

# 基于相似网络融合算法的癌症亚型预测

张晓茜[1]　李东喜[2]
1 太原理工大学数学学院　太原 030600
2 太原理工大学大数据学院　太原 030600
(1518998276@qq.com)

**摘　要**　从基因表达数据中挖掘基因之间的相互作用关系,构建基因调控网络,是生物信息学中重要的研究课题之一。但目前流行的神经网络在其架构中仅考虑基因之间的交互和关联,不考虑患者之间的交互和关联。为此,提出了一种基于加权基因相似网络和样本相似网络融合算法的癌症亚型预测模型,即 WGCSS(Weighted Genetic Correlation network and Sample Similarity network)。该方法实现了特征空间和样本空间信息的融合,同时考虑了基因之间和样本之间的相互作用关系,并使用图卷积网络进行预测。在两个空间中聚合信息会导致严重的过度平滑问题,为此在该模型中引入残差层以缓解过度平滑问题。该方法通过聚合两个空间中的数据信息,可以使得癌症亚型预测的结果更加准确。为了验证方法的泛化性能,使用了乳腺浸润癌(BRCA)、多形性胶质母细胞瘤(GBM)和肺癌(LUNG)数据集进行分析,由此产生的高分类精度结果可以表明该方法的优越性。另外,还对 3 类数据集进行了生存分析,证明该方法在 3 个癌症数据集上癌症亚型的生存曲线存在显著差异。

**关键词**:加权基因相似网络;样本相似网络;残差图卷积网络;L1 正则;癌症亚型预测
**中图分类号**　TP399

## Cancer Subtype Prediction Based on Similar Network Fusion Algorithm

ZHANG Xiaoxi[1] and LI Dongxi[2]
1 College of Mathematics, Taiyuan University of Technology, Taiyuan, Shanxi 030060, China
2 College of Big Data, Taiyuan University of Technology, Taiyuan, Shanxi 030060, China

**Abstract**　Mining the interaction relationship between genes from gene expression data and construct gene regulatory network is one of the important research topics in bioinformatics. However, the current popular neural network only considers the interaction and association between genes in its architecture, and does not consider the interaction and association between patients. Therefore, a cancer subtype prediction model based on the fusion algorithm of weighted gene similarity network and sample similarity network, namely WGCSS, is proposed in this paper. In this method, the fusion of feature space and sample space information is realized, and the interaction between genes and samples is considered, and the graph convolutional network is used for prediction. Aggregating information in two spaces will lead to a serious oversmoothing problem. Therefore, a residual layer is introduced in the model to alleviate the oversmoothing problem. This method can make the prediction of cancer subtypes more accurate by aggregating the data information in the two spaces. To verify the generalization performance of the method, datasets of invasive breast carcinoma(BRCA), glioblastoma multiforme(GBM), and LUNG(LUNG) are used for analysis, and the resulting high classification accuracy demonstrates the superiority of the method. Survival analysis is also performed on three types of data sets, and it is proved that the method has significant differences in the survival curves of cancer subtypes in three cancer datasets.

**Keywords**　Weighted gene similarity network, Sample similarity network, Residual graph convolutional network, L1 regular, Cancer subtype prediction

## 1　引言

癌症的发生从微观方面来讲是一种异常细胞不受控制地分裂,并可通过血液和淋巴系统侵入附近组织和身体其他部位的疾病。在过去的十年中,一些大规模的癌症基因组学项目已经发表了来自数千名癌症患者的基因组、表观基因组、转录组和蛋白质组的数据。这些项目包括癌症基因组图谱(TCGA)、国际癌症基因组联盟(ICGC)和全基因组泛癌症分析(PCAWG)。癌症微阵列技术在基因组研究、计算生物学、统计学和机器学习等学科中开辟了广泛的多学科研究领域。在微阵列癌症数据领域开展研究,对于癌症患者的诊断、癌症亚型的识别和区分具有重要意义[1]。

基金项目:国家自然科学基金项目(11571009);山西省应用基础研究项目(201901D111086);山西省重点研发计划项目(202102020101004);山西省回国留学人员科研资助项目(2022-074)
This work was supported by the National Natural Science Foundation of China (11571009), Basic Research Project of Shanxi Province (201901D111086), Key Research and Development Project of Shanxi Province(202102020101004) and Research Support Program of Shanxi Province for Returned Overseas Students(2022-074).
通信作者:李东喜(dxli0426@126.com)

通过高通量测序技术[2]获得的基因表达的数据结构复杂，信息冗余度高，基因之间的相关性很强。传统的生物学研究方法难以有效地处理基因表达数据，所以从基因表达数据中挖掘基因之间的相互作用关系，构建基因调控网络，成为生物信息学中重要的研究课题之一。Chen 等[3]运用皮尔逊相关系数的方法建立了基因调控网络，该方法需要不断尝试设置最优阈值，以此来保留调控网络中相关性强的基因间关系。加权基因共表达网络分析(Weighted Geneco-expression Network Analysis，WGCNA)方法基于表达模式类似的分子可能参与特定生物学功能的理论，最初由 Zhang 和 Horvath[5]提出。它被广泛研究并用于预测新的基因功能[4]，发现新的疾病生物标记物，以及检测癌症中的遗传变异。

近年来，人们提出了不同的计算方法来检测癌症亚型。这些方法通常建立在特征工程的基础上，对患者进行聚类或分类。由于生物数据高维且样本量小，早期的方法在一定程度上减少了样本的特征，并利用这些特征来聚类癌症亚型。Li 等[6]将 L1 正则化惩罚添加到目标函数中，该方法在处理噪声和异常值方面是非常有效的。Guo 等[7]提出了两步 L1 正则化方法对微阵列数据进行分类，该文定义了一种新的 L1 正则化特征选择方法，以去除不相关和冗余的特征并选择重要的特征。在我们的工作中，为了处理小样本的维度灾难，引入了 L1 正则化的特征选择，然后使用深度学习技术对多类微阵列癌症数据进行分类。

随着深度学习技术的发展，一些诸如卷积神经网络(CNNs)[8-10]、堆栈自编码器(SAE)[11]、生成对抗模型(GANs)、卷积自编码器(CAE)、变分自编码器(VAE)[12]等的深度学习方法已经应用于癌症研究领域[13-14]。基于分类的方法通常是在一些已知亚型标签的癌症样本上训练一个模型，然后使用该模型预测新的癌症样本的亚型。根据特征提取和分类的差异，基于分类的方法分为两类：两阶段方法和端到端方法。两阶段方法需要先对基因表达数据进行特征选择，然后再进行分类预测。端到端的方法可以同时进行选择和分类。DeepType[15]将癌症样本数据输入多层神经网络用以降低数据维数，同时将监督分类与非监督聚类相结合，用聚类结构确定癌症相关数据的表示形式。Dai 等[16]设计的端到端的深度学习方法 ERGCN，使用癌症样本相似性网络和残差图卷积网络来对癌症亚型进行分类，得到了较优的分类结果。本文使用两阶段的分类方法，将特征选择与端到端的分类方法相结合，试图进一步提升癌症预测的准确性。目前流行的深度神经网络在其架构中仅考虑基因之间的交互和关联，不考虑患者之间的交互和关联。考虑患者之间的关系是有益的，因为它有助于一起分析和研究相似的患者队列。通过将患者视为节点，将他们的相互作用关系视为边，图表提供了一种自然的方式来表示患者之间的相互作用。

为了解决这一问题，我们提出一种新的预测模型：基于加权基因相似网络和样本相似网络融合算法的癌症亚型预测模型(WGCSS)。在数据预处理中，为了减低基因数据的维度，采用 L1 正则进行特征选择；在构建基因相似网络时，模型采用自适应选择策略进行阈值选择；在构建样本相似网络后，本文融合了这两个相似网络中的信息；在本文中使用图卷积网络模型进行预测，为了避免训练过程中的过度平滑，在模型中构建了残差层。实验结果表明本文模型优于现有方法。

## 2 提出的方法

首先，采用 L1 正则进行特征选择，它通过在线性回归的最小二乘误差中加入惩罚项来应用收缩策略，其将不相关的特征赋零系数，仅考虑非零系数变量。该方法解决了模型的过拟合问题，减小了预测误差，简化了模型的复杂度，提升了计算稳定性。然后，基于 WGCNA 计算癌症患者基因表达之间的相似度，构建一个以基因为节点的加权基因相似网络。样本相似网络是在样本空间中基于皮尔逊相关系数构建的，本文的基因图和样本图都是提前构造的，这可以大大降低计算的复杂度，保持样本之间关系的稳定性。最后，采用残差图卷积神经网络算法进行分层消息传递，通过网络扩散节点的特征信息，学习每个节点的特征表示。残差层的构建主要是为了防止在两个空间聚合信息时的过度平滑问题。WGCSS 根据患者的特征表征来预测癌症亚型，该模型由两层卷积层、两层残差层和三层线性层构成。图 1 给出了 WGCSS 模型概述。

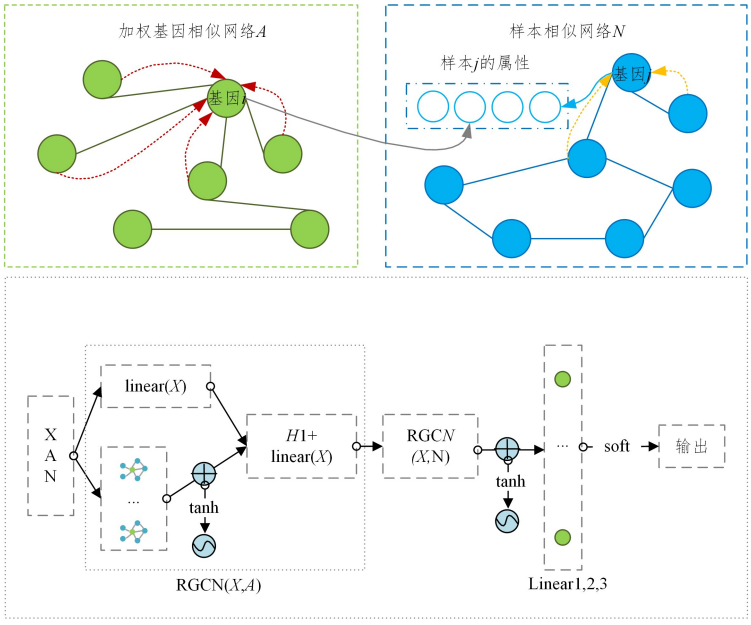



图 1 WGCSS 模型流程图

Fig. 1 Flowchart of WGCSS model

图 1 中,$\boldsymbol{X}$ 为经过特征选择后的初始矩阵,$\boldsymbol{A}$ 为加权基因相似网络的邻接矩阵,$\boldsymbol{N}$ 为样本相似网络的邻接矩阵。

### 2.1 基于 L1 正则的特征选择

L1 正则化特征选择是微阵列数据分析中引入的一种新的特征选择方法。基于 L1 的特征选择使用 LSVM 拟合数据,并返回将数据划分为类别的最佳拟合超平面。它利用局部最优解去除系数为零的特征。在本研究中,使用 L1 正则化支持向量机算法进行特征选择。

假设一个数据集 $S$ 有 $n$ 个实例,如式(1)所示:

$$S=\{(x_i,y_i)\mid x_i\in R^n,y_i\in\{-1,1\}\}_{i=1}^{k} \tag{1}$$

其中,$x_i$ 是第 $i$ 个实例,具有 $n$ 个特征和一个类标签 $y_i$。$x_i$ 的表达式如式(2)所示:

$$x_i=\{x_{i1},x_{i2},\cdots,x_{in}\} \tag{2}$$

其中,$x_{ij}$ 是实例 $x_i$ 中第 $j$ 个特征的值。

对于二分类问题,支持向量机 SVM 的原理要找到最优分离超平面 $h\times x=b$,支持向量到超平面的距离为 $\frac{2}{\|\omega\|_2^2}$,其中 $\boldsymbol{\omega}$ 是权重向量,$b$ 是偏置项,其表达式如下:

$$\min_{\omega,b}\frac{1}{2}\|\omega\|_2^2 \tag{3}$$

但须符合以下条件:

$$\{y_i(\omega x_i-b)\geqslant 1,i=\overline{1,k}\} \tag{4}$$

Bradley 和 Mangasarian[17] 提出了 L1-SVM 算法,根据所得到的稀疏解进行特征选择,如式(5)所示:

$$\min_{\omega,b,\varepsilon}\|\omega\|^1+\alpha\sum_{i=1}^{k}\varepsilon_i \tag{5}$$

其中,$\varepsilon_i$ 为松弛变量,$\alpha>0$ 为错误惩罚参数。但须符合以下条件:

$$\{y_i(\omega x_i-b)\geqslant 1-\epsilon_i,\epsilon_i\geqslant 0,i=\overline{1,k}\} \tag{6}$$

L1 正则支持向量机的优化问题如式(7)所示:

$$\min_{\omega,b}\|\omega\|^1+\alpha\sum_{i=1}^{k}\max(0,1-y_i(\omega^{\mathrm{T}}x_i+b))^2 \tag{7}$$

该算法利用控制参数 $\alpha$ 来控制数据的稀疏性。稀疏性允许矩阵中的少数特征具有较大的非零系数值。本文中设定 $\alpha=0.5$。

### 2.2 加权基因相似网络

我们引用 WGCNA 中的相似度矩阵概念,以构建加权基因相似网络。首先,计算所有基因的相似度共表达矩阵 $s_{ij}$,其中 $s_{ij}=|r(x_i,x_j)|$ 为节点 $i$ 和节点 $j$ 的基因表达谱之间相关系数的绝对值。本文采用距离相关法计算相关系数。因为距离相关系数总是正的,所以定义了一个无符号网络,其中正相关和负相关是同等的。然后,通过设置 $a_{ij}=s_{ij}{}^{\beta}$,以 $\beta$ 为软阈值幂,将相似度共表达式矩阵转化为邻接矩阵 $a_{ij}$,然后计算 TOM 拓扑重叠矩阵(使用 R 中的 pick Soft Threshold 函数,加权是指对相关性值进行幂次运算,幂次的值即为软阈值)。

### 2.3 样本相似网络

我们根据患者的相似性构建了一个患者网络。通过患者基因表达谱的 Pearson 相关系数(PCC)计算患者间的相似性(见式(8))。

$$r(x,y)=\frac{E[(x-\bar{x})(y-\bar{y})]}{\sigma_x\sigma_y}=\frac{\sum_i^n(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i^n(x_i-\bar{x})^2}\sqrt{\sum_i^n(y_i-\bar{y})^2}} \tag{8}$$

Pearson 相关系数($r(x,y)$)可以衡量两个患者之间的线性相关程度,输出范围为 $-1$ 到 1,当 $r$ 的绝对值接近 1 时,两个患者呈正相关或负相关,否则两个患者无相关性。因此,我们认为两个患者可由一条边连接,如果两个样本之间的 Pearson 系数绝对值大于阈值 $\theta$,则将邻接矩阵中对应的值设为 1,否则两个患者之间没有边连接,对应的邻接矩阵值为零。在本文中,$\theta$ 取值为 0.8。

### 2.4 残差图卷积网络

在本文的模型中,构造了两个图:$G(G)=(V(G),E(G))$ 和 $G(s)=(V(s),E(s))$。$G(G)$ 为基因相互作用图,其中 $V(G)$ 为基因集合,$E(G)$ 为表示基因间相互作用的边集合。与 $G(G)$ 不同,$G(s)$ 是样本相似度图,其中 $V(s)$ 是样本的集合,$E(G)$ 是表示样本之间相似度的边的集合。$A$ 和 $N$ 分别为 $G(G)$ 和 $G(s)$ 的邻接矩阵,$D^g$ 和 $D^s$ 分别为 $A$ 和 $N$ 的度矩阵,$X$ 为 $n\times D$ 的网络节点初始属性矩阵。$X$ 为 $n\times D$ 的网络节点初始属性矩阵。本章的 GCN 模型需要 3 个输入:两个存储节点连接的邻接矩阵和一个节点初始属性矩阵。在特征空间中的 GCN 定义如下:

$$H^{(1)}=\tanh(D^{g-\frac{1}{2}}AD^{g-\frac{1}{2}}H^{(0)}W^0) \tag{9}$$

在基于 GCN 的方法中,过度平滑是一个常见的挑战,在两个空间中聚合信息时会非常严重。因为要先在特征空间(垂直)中聚合,然后在样本空间(水平)中聚合,每个元素将聚合两次。为了防止模型过度平滑,本章同样也在 GCN 中增加了一层残差数据。为了保证输入特征的维度与一层 GCN 的节点特征维度一致,我们将初始输入特征通过一个独立的线性层直接连接到 GCN 层的输出。可以写成如下形式:

$$H^{(P)}=H^{(1)}+\tanh(linear(x^{\mathrm{T}})) \tag{10}$$

在样本空间中的 GCN 定义如下:

$$H^{(2)}=\tanh(D^{s-\frac{1}{2}}ND^{s-\frac{1}{2}}H^{(p)}W^1) \tag{11}$$

同样,在样本空间的 GCN 中也增加了一层残差数据,如下所示:

$$H^{(q)}=H^{(2)}+\tanh(linear(x^{\mathrm{T}})) \tag{12}$$

使用交叉熵损失函数来量化癌症亚型预测损失:

$$L=-\sum_{d\in y_T}\sum_{f=1}^{F}Y_{kf}\ln Z_{df} \tag{13}$$

给定训练集 $T$ 中的患者 $k$,$Y_{kf}$ 为符号函数(0 或 1),$Z_{df}$ 为癌症患者 $k$ 属于 $f$ 类的预测概率。我们将癌症亚型预测损失最小化,以优化 WGCSS。

## 3 实验和结果分析

### 3.1 实验数据集

为了验证 WGCSS 方法的有效性,使用了来自 TCGA(the Cancer Genome Atlas)的乳腺浸润性癌(BRCA)、多形性胶质母细胞瘤(GBM)和肺癌(LUNG)的肿瘤数据集,在 Wang[18] 的补充文件中下载了 3 种癌症类型的基因表达数据和生存信息。使用 R 包 TCGAbiolinks 从 TCGA 中检索癌症亚型信息。通过匹配基因表达数据的样本 ID,将其样本基因表达数据与他们的癌症亚型结合起来。最后整理出 102 例乳腺浸润性癌患者、213 例多形性胶质母细胞瘤患者和 85 例肺癌患者都有 4 种癌症亚型,详细信息如表 1 所列。

表 1 数据集说明

Table 1 Dataset descriptions

| Dataset | Number of features | Sample size | Number of classes |
|---|---|---|---|
| BRCA | 17814 | 102 | 4 |
| GBM | 12042 | 213 | 4 |
| LUNG | 12042 | 85 | 4 |

### 3.2 评价指标

外部评价指标：通过将预测分类结果与实际分类结果进行比较来评价算法的有效性。每个性能指标都有其优缺点。为了缓解这种限制，使用了诸如准确度、召回率、精度、F1 分数、混淆矩阵和宏观平均 ROC 等性能度量进行比较。这些指标的公式如下：

$$Acc=\frac{TP+TN}{TP+TN+FP+FN} \tag{14}$$

$$P=\frac{TP}{TP+FP} \tag{15}$$

$$R=\frac{TP}{TP+FN} \tag{16}$$

$$F1=\frac{2TP}{2TP+FP+FN} \tag{17}$$

其中，$TP$ 为模型预测的阳性样本为阳性，$TN$ 为模型预测的阴性样本为阴性，$FP$ 为模型预测的阴性样本为阳性，$FN$ 为模型预测的阳性样本为阴性。

我们还使用了一个内部评价指标——Davies-Boulding 指数：

$$DBI=\frac{1}{n}\sum_{i=1}^{n}\max_{i\neq j}\left(\frac{avg(c_i)+avg(c_j)}{dis(c_i,c_j)}\right) \tag{18}$$

其中，$n$ 为聚类个数，$avg(c_j)$ 为第 $i$ 类样本到其聚类质心的平均距离，$dis(c_i,c_j)$ 为 $c_i$ 类中心到 $c_j$ 类中心的距离。DBI 的下限为 0，且 DBI 值越小，聚类效果越好。

### 3.3 实验设置

在特征选择阶段，设置系数性参数 $\alpha=0.5$。使用 Adam 优化器函数，学习率设置为 0.001。为了避免“选择偏差”，在所有微阵列数据集的实验中使用了 5 倍交叉验证(CV)。每个数据集的样本被随机分成 5 个大小相等的子样本，其中 4 个子样本用于训练，其余一个子样本用于测试。该过程重复 5 次，5 个子样本中的每一个都用作测试数据。所有比较方法都在由 5 倍 CV 随机生成的同一数据集上实现。在构建基因相似网络时，本文要求 $k$ 与 $p(k)$ 的相关性达到 0.85 时的 power 作为 $\beta$ 值。根据无标度拟合指数平均连接度分析 BRCA 的 $\beta$ 值为 2，GBM 的 $\beta$ 值为 6，LUNG 系统没有给出合适的 $\beta$ 值，我们手动调节软阈值并给出它的 $\beta$ 值为 3。构建样本相似网络时的 $\theta$ 取值为 0.8。

### 3.4 实验结果与分析

本文对 3 种标准的多类微阵列癌症数据集进行了实验，结果如表 2 所列(将 5 倍交叉验证测试集的平均结果作为评价指标)。表 2 显示了本文的模型在 BRCA，GBM 和 LUNG 数据集上关于准确度、召回率、精度、F1 分数和 DBI 的性能。经过特征选择，3 个数据集的特征都大规模减少。所提出的方法在乳腺浸润性癌和肺癌数据集上显示出完美的分类性能评分 1.00，在多形性胶质母细胞瘤数据集上的分类准确度为 0.97。文中方法在乳腺浸润性癌和肺癌数据集实现了 100% 的精度、召回率和 F1 度量；在多形性胶质母细胞瘤数据集中精确度、f 度量和召回率都达到了 97%。DBI 指数在 3 个数据集上都较低，证明本章的模型有较优的聚类效果。

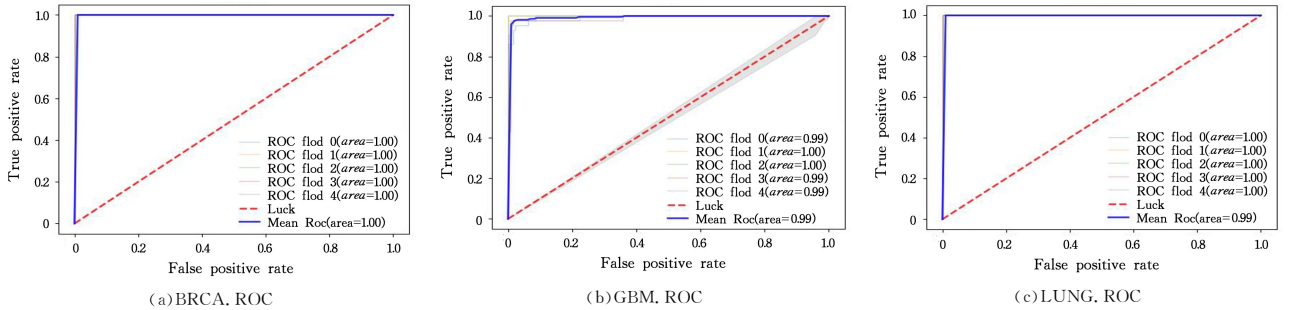
表 2 WGCSS 模型分类结果

Table 2 Classification results of WGCSS model

| 数据集 | 选择的特征数 | *Precision* | *Recall* | *F1 Score* | *Acc* | *DBI* | *AUC* |
|---|---|---|---|---|---|---|---|
| BRCA | 235 | 1.0 | 1.0 | 1.0 | 1.0 | 0.03991 | 1.0 |
| GBM | 332 | 0.97264 | 0.97715 | 0.97148 | 0.97731 | 0.12057 | 0.99154 |
| LUNG | 211 | 1.0 | 1.0 | 1.0 | 1.0 | 0.06650 | 1.0 |

通过对 3 种标准的多类微阵列癌症数据集进行实验，我们发现本文提出的算法可以提高模型的分类性能和鲁棒性。该算法在多个数据集上取得了较好的分类效果，表明其在微阵列癌症分类领域具有广阔的应用前景。

在分类问题中，ROC 曲线被广泛用于检验模型的性能。3 类数据集的五倍交叉平均 ROC 曲线如图 2 所示。本文模型在多形性胶质母细胞瘤数据集上所得的 AUC 评分为 99.15%，在乳腺浸润性癌和肺癌数据集上也都实现了完美的分类精度，因此这两类数据集的 AUC 评分为 100%。每次交叉验证的结果都是较优的，证明本文的模型也是较为优越的。




(a)BRCA. ROC (b)GBM. ROC (c)LUNG. ROC

图 2 五倍交叉验证后的平均 ROC 曲线

Fig. 2 Average ROC curve after five-fold cross-validation

### 3.5 对比分析

Dai 等提出的 ERGCN 模型是基于 PCC 构建的样本相似网络，其只考虑了样本空间的信息而忽略了特征空间中的基因信息。我们将经过特征选择后的基因组数据输入该模型，其结果如表 3 所列(加粗字体为 WGCSS 模型的结果)。可以看到，在 BRCA 数据集上两个模型都达到了最优，在

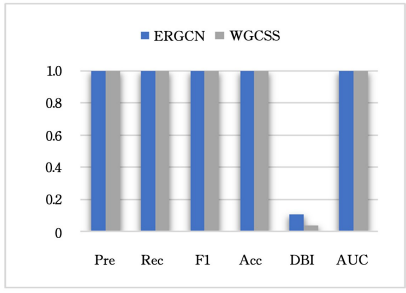
GBM 数据集上的各项评价指标上,本文模型比 ERGCN 模型提高了 2%左右;在 LUNG 数据集上的各项评价指标上,本文模型比 ERGCN 模型提高了 1.5%左右。对于 DBI 指数,本文的模型在 BRCA 数据集上比 ERGCN 模型降低了 6.76%,在 GBM 和 LUNG 数据集上各降低了约 13.13%和 8.92%,说明本文的模型具有更优的聚类效果。上述结果说明基因之间和样本之间的相互作用都包含了有价值的癌症样本分类信息。

表 3 ERGCN 模型和 WGCSS 模型的分类结果比较

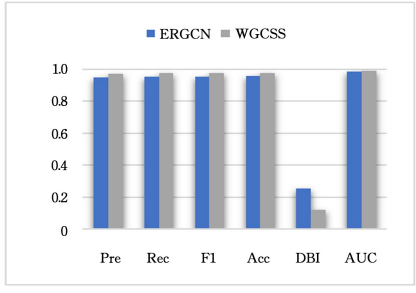
Table3 Classification comparison between ERGCN and WGCSS models

| 数据集 | 选择的特征数 | 方法 | *Precision* | *Recall* | *F1 Score* | *Acc* | *DBI* | AUC |
|---|---|---|---|---|---|---|---|---|
| BRCA | 235 | ERGCN | 1.00000 | 1.00000 | 1.00000 | 1.00000 | 0.10752 | 1.00000 |
| | | WGCSS | **1.00000** | **1.00000** | **1.00000** | **1.00000** | **0.03991** | **1.00000** |
| GBM | 332 | ERGCN | 0.95051 | 0.95274 | 0.95318 | 0.95591 | 0.25188 | 0.98728 |
| | | WGCSS | **0.97264** | **0.97715** | **0.97748** | **0.97731** | **0.12057** | **0.99154** |
| LUNG | 211 | ERGCN | 0.98271 | 0.98418 | 0.98523 | 0.98764 | 0.15572 | 0.98699 |
| | | WGCSS | **1.00000** | **1.00000** | **1.00000** | **1.00000** | **0.06650** | **1.00000** |

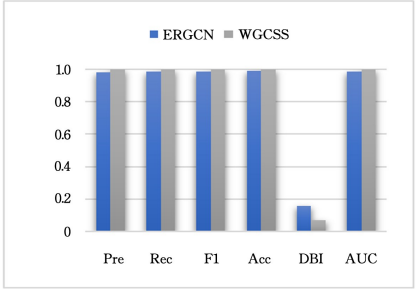
在分类精度方面,将本文模型与现有的两阶段分类模型进行比较,如表 4 所列。可以看到,文中模型在 BRCA 数据集上比 DFN Forest 模型提高了 19.14%,比 SGL-SVM 模型提高了 14.35%,比 CFN Forest 模型提高了 5.64%;在 GBM 数据集上比 DFN Forest 模型提高了 10.74%,比 SGL-SVM 模型提高了 14.63%,比 CFN Forest 模型提高了 5.99%;在 LUNG 数据集上比 CFN Forest 模型提高了 9.11%,比 SGL-SVM 模型提高了 4.27%,比 MOEDA 模型提高了 4%,比 RNBC 模型提高了 11.5%。




(a)BRCA




(b)GMB




(c)LUNG

图 4 WGCSS 模型和 ERGCN 模型结果的对比图

Fig. 4 Results comparison between WGCSS and ERGCN models

表 4 不同两阶段分类模型精度的比较

Table 4 Accuracy comparison of different two-stage classification models

| Dataset | Experiments | Method | *Acc*/% |
|---|---|---|---|
| BRCA | Xu[18] | DFN Forest | 80.86 |
| | Huo[20] | SGL-SVM | 85.65 |
| | Zhong[21] | CFN Forest | 94.36 |
| | This paper | WGCSS | **100** |
| GBM | Xu[18] | DFN Forest | 86.52 |
| | Huo[20] | SGL-SVM | 82.61 |
| | Zhong[21] | CFN Forest | 91.25 |
| | This paper | WGCSS | **97.26** |
| LUNG | Chandra[22] | RNBC | 88.50 |
| | Lv[23] | MOEDA | 96.00 |
| | Huo[20] | SGL-SVM | 95.73 |
| | Zhong[21] | CFN Forest | 90.89 |
| | This paper | WGCSS | **100** |

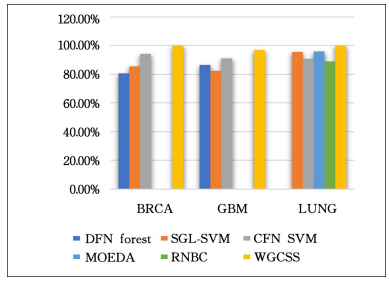


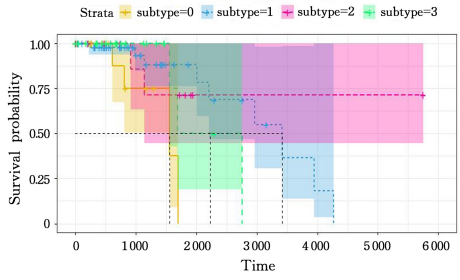

图 5 不同两阶段模型的精度比较

Fig. 5 Precision comparison of different two-stage models

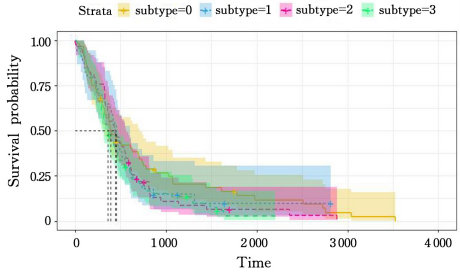
### 3.6 生存分析

为了进一步探讨所识别的亚型之间的关系,本节对 WGCSS 模型的结果进行了生存分析。从理论上讲,不同的癌症亚型应有不同的生存曲线。图 6 展示了文中模型在 BRCA,GBM 和 LUNG 数据集上的 Kaplan-Meier 生存曲线,横轴($x$ 轴)表示以天为单位的时间,纵轴($y$ 轴)表示生存的概率或生存人口的比例。图中曲线代表两组病人的生存曲线。曲线的垂直下降表示事件。曲线上的垂直刻度表示这个病人在这个时候被审查了。本文计算了不同亚型生存曲线上 log-rank 检验的 $p$ 值,还在曲线上绘制了中位生存时间。这 3 类癌症参与者在时间为 0 时,生存概率是 1.0,即所有的参与者都活着。在 BRCA 数据集上,第一种亚型的中位生存时间为 1563 天,第二种亚型为 3418 天;第三种亚型为 NA;第四种亚型为 2227 天。NA 表示第三组大多数患者无法活过中位生存时间。在该数据集上第二种亚型的生存率明显优

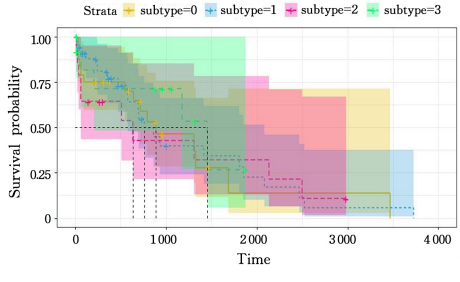
于其他 3 种亚型，且每个类别之间的中位生存时间差距都很大，表明在该数据集上的癌症亚型的生存曲线存在显著差异。对于 GBM 数据集，癌症亚型间的差异不是很明显。第一种亚型患者中位生存时间为 394 天；第二种亚型为 455 天；第三种亚型为 440 天；第四种亚型为 362 天。在 LUNG 数据集上，第一种亚型的中位生存时间为 888 天；第二种亚型为 761 天；第三种亚型为 631 天，第四类是 1456 天。在该数据集上，第四种亚型的生存率明显优于其他 3 种亚型，且每个类别之间的中位生存时间差距也都很大，表明在该数据集上的癌症亚型的生存曲线存在显著差异。在 BRCA、GBM 和 LUNG 数据集生存曲线上，log-rank 检验的 $p$ 值分别为 0.034，0.087，0.056，因此文中模型在 3 个癌症数据集上癌症亚型的生存曲线存在显著差异。




(a)BRCA 不同亚型的生存时间




(b)GBM 不同亚型的生存时间




(c)LUNG 不同亚型的生存时间

图 6 不同亚型的生存时间

Fig. 6 Survival time of different subtypes

**结束语** 本文提出了一种基于加权基因相似网络和样本相似网络融合算法的癌症亚型预测模型(WGCSS)。该模型不仅考虑了基因之间的相互作用关系，同时也考虑了样本之间的相互作用关系，通过聚合两个空间中的基因数据信息，使得预测的结果更加准确。与仅考虑样本空间或特征空间的模型相比，该模型有更优的准确率和聚类效果，说明样本之间和基因之间的相互作用都包含了有价值的癌症样本分类信息。最后进行的生存分析也表明，在该模型，癌症亚型的生存曲线是有显著差异的。本文研究为癌症亚型分类预测提供了一种新的方法，也为精准医疗提供了新的可能。

## 参 考 文 献


[1] BERGER M F, MARDIS E R. The emerging clinical relevance of genomics in cancer medicine[J]. Nature Reviews Clinical Oncology, 2018, 15(6): 353-365.

[2] JIA Q, CHU H, JIN Z, et al. High-throughput single-cell sequencing in cancer research[J]. Signal Transduction and Targeted Therapy, 2022, 7(1): 145.

[3] CHEN W, LI J, HUANG S, et al. GCEN: An easy-to-use toolkit for gene co-expression network analysis and lncRNAs annotation[J]. Current Issues in Molecular Biology, 2022, 44(4): 1479-1487.

[4] YANG R, DU Y, WANG L, et al. Weighted gene co-expression network analysis identifies CCNA2 as a treatment target of prostate cancer through inhibiting cell cycle[J]. Journal of Cancer, 2020, 11(5): 1203.

[5] ZHANG B, HORVATH S. A general framework for weighted gene co-expression network analysis[J]. Statistical Applications in Genetics and Molecular Biology, 2005, 4(1).

[6] LI C N, SHAO Y H, DENGN Y. Robust L1-norm two-dimensional linear discriminant analysis[J]. Neural Networks, 2015, 65: 92-104.

[7] GUO S, GUO D, CHEN L, et al. A L1-regularized feature selection method for local dimension reduction on microarray data [J]. Computational Biology and Bhemistry, 2017, 67: 92-101.

[8] LIU B, CHI W, LI X, et al. Evolving the pulmonary nodules diagnosis from classical approaches to deep learning-aided decision support: three decades' development course and future prospect[J]. Journal of Cancer Research and Clinical Oncology, 2020, 146: 153-185.

[9] QI LL, WU B T, TANG W, et al. Long-term follow-up of persistent pulmonary pure ground-glass nodules with deep learning-assisted nodule segmentation[J]. European Radiology, 2020, 30: 744-755.

[10] MUNIR K, FREZZA F, RIZZIA. Brain tumor segmentation using 2D-UNET convolutional neural network[J]. Deep Learning for Cancer Diagnosis, 2021, 908: 239-248.

[11] XU J, WU P, CHENY, et al. A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data[J]. BMC Bioinformatics, 2019, 20(1): 1-11.

[12] PARK K H, BATBAATAR E, PIAO Y, et al. Deep learning feature extraction approach for hematopoietic cancer subtype classification[J]. International Journal of Environmental Research and Public Health, 2021, 18(4): 2197.

[13] LOPEZ M M. Deep Learning for Brain Tumor Segmentation [M]. University of Colorado Colorado Springs, 2017.

[14] MUNIR K, ELAHI H, AYUB A, et al. Cancer diagnosis using deep learning: a bibliographic review[J]. Cancers, 2019, 11(9): 1235.

[15] CHEN R, YANG L, GOODISON S, et al. Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data[J]. Bioinformatics, 2020, 36(5): 1476-1483.

[16] DAI W, YUE W, PENG W, et al. Identifying Cancer Subtypes Using a Residual Graph Convolution Model on a Sample Simi-

larity Network[J]. Genes,2022,13(1):65.

[17] BRADLEY P S,MANGASARIAN O L. Feature selection via concave minimization and support vector machines[C]// ICML. 1998:82-90.

[18] WANG B,MEZLINI A M,DEMIR F,et al. Similarity network fusion for aggregating data types on a genomic scale[J]. Nature Methods,2014,11(3):333-337.

[19] XU J,WU P,CHEN Y,et al. A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data[J]. BMC Bioinformatics,2019,20(1):1-11.

[20] HUO Y,XIN L,KANG C,et al. SGL-SVM:a novel method for tumor classification via support vector machine with sparse group Lasso[J]. Journal of Theoretical Biology,2020,486:110098

[21] ZHONG L,MENG Q,CHEN Y. A Cascade Flexible Neural Forest Model for Cancer Subtypes Classification on Gene Expression Data[J]. Computational Intelligence and Neuroscience,2021,2021:6480456.

[22] CHANDRA B,GUPTA M. An efficient statistical feature selection approach for classification of gene expression data[J]. Journal of Biomedical Informatics,2011,44(4):529-535.

[23] LV J,PENG Q,CHEN X,et al. A multi-objective heuristic algorithm for gene expression microarray data classification[J]. Expert Systems with Applications,2016,59:13-19.

**ZHANG Xiaoxi**, born in 1997, postgraduate. Her main research interests include data mining and analysis and so on.

**LI Dongxi**, born in 1982, Ph.D, associate professor. His main research interests include data mining and biostatistics.