Advances in Applied Mathematics 应用数学进展, 2026, 15(4), 777-790
Published Online April 2026 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2026.154200

Hans汉斯

癌症预后预测的机器学习集成框架的综合分析

巩琪琪

青岛大学数学与统计学院，山东 青岛

收稿日期：2026年3月27日；录用日期：2026年4月21日；发布日期：2026年4月28日

摘　要

本文基于TCGA项目的基因拷贝数变异、RNAseq基因表达、DNA甲基化等多维组学数据，结合机器学习算法构建癌症患者预后生存预测模型。首先对组学数据进行预处理，提取患者生存时间；随后采用主成分分析、偏最小二乘法等方法降维，并通过mRMR算法筛选低冗余、高生物学意义的特征子集；最后应用支持向量机、Logistic回归等算法构建分类模型，经交叉验证与多指标评估模型性能。实验结果表明，模型性能与降维及分类算法选择密切相关，其中基于偏最小二乘法降维的模型表现最优，证实患者标签信息对关键特征提取的重要性；Kaplan-Meier曲线进一步验证了模型有效性。本文构建的预测模型可为临床决策提供科学依据，助力肿瘤精准医疗发展，改善患者预后状况与生存质量，具有较高的理论意义与潜在临床应用价值。

关键词

数据降维，特征选择，机器学习，TCGA

Comprehensive Analysis of Machine Learning Ensemble Frameworks for Cancer Prognosis Prediction

Qiqi Gong

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

Received: March 27, 2026; accepted: April 21, 2026; published: April 28, 2026

Abstract

Based on multi-dimensional omics data such as gene copy number variations, RNAseq gene expression, and DNA methylation from The Cancer Genome Atlas (TCGA) project, this study constructs a prognostic survival prediction model for cancer patients by integrating machine learning algorithms. First, the omics data are preprocessed to extract patients' survival time; subsequently,

文章引用: 巩琪琪. 癌症预后预测的机器学习集成框架的综合分析[J]. 应用数学进展, 2026, 15(4): 777-790.
DOI: 10.12677/aam.2026.154200

dimensionality reduction methods including principal component analysis (PCA), non-negative matrix factorization (NMF), and partial least squares (PLS) are employed, followed by screening of low-redundancy and biologically meaningful feature subsets using the mRMR algorithm; finally, classification models are built using algorithms such as support vector machine (SVM), random forest (RF), and Logistic regression (LR), with model performance evaluated through cross-validation and multiple metrics. Experimental results indicate that model performance is closely related to the selection of dimensionality reduction and classification algorithms. Among them, models based on PLS dimensionality reduction achieve the optimal performance, confirming the importance of patient label information for extracting key features; Kaplan-Meier curves further verify the model's effectiveness. The constructed prediction model can provide a scientific basis for clinical decision-making, facilitate the development of tumor precision medicine, improve patients' prognostic outcomes and quality of life, and thus possesses significant theoretical significance and potential clinical application value.

Keywords

Data Dimensionality Reduction, Feature Selection, Machine Learning, TCGA

1. 引言

1.1. 研究背景与选题意义

恶性肿瘤(以下简称“癌症”)作为一类严重威胁人类健康与生命的恶性疾病。对癌症的有效治疗及预后预测，不仅关乎个体生命质量的提升，更是全球健康领域实现预期寿命提升、减轻疾病负担的关键环节。在此背景下，深入探索癌症的内在规律，借助先进技术手段构建科学的预测模型[1]，对于推动肿瘤个体化治疗与精准医疗[2]发展具有重要意义。

尽管近几十年来，肿瘤领域在早期筛查[3]、分子分型[4]、手术治疗、靶向药物及免疫治疗等方面均取得显著进展，但癌症的整体治愈率与长期生存率仍未达理想水平。尤为值得注意的是，即便接受相同治疗方案，患者的生存结局仍可能存在显著差异[5]，这一现象深刻反映了肿瘤疾病的复杂性与个体化特征。这种个体差异的背后，涉及多重复杂因素。

因此，如何综合考量多维度因素[6]，构建有效的预测模型[7]以实现对患者长期生存情况的精准预测，已成为精准医疗[8]及肿瘤个体化治疗领域亟待解决的重要研究课题。

1.2. 技术路线

基于 TCGA 数据，本研究构建癌症患者长期生存预测模型[9]的技术路线[4]如下：

(1) 数据预处理：包括缺失值处理、数据标准化等；(2) 特征筛选与降维：采用方差阈值剔除低信息量特征，结合 PCA、NMF、PLS 多种降维方法[10]进行比较；(3) 分类模型构建：选用支持向量机(SVM)、逻辑回归(LR)，2 类典型模型进行预测；(4) 模型评估与比较[11]：基于 F1 值、ROC-AUC 指标对不同模型与降维方法的组合进行系统评估。

2. 数据预处理与特征选择

本章基于 UCSC 和 Xena 平台下载的各种癌症的多组学数据、临床信息以及表型特征等数据，完成

数据清洗、特征筛选与压缩以及平衡样本数量等一系列预处理操作，为后续建立预后分析模型奠定基础。

2.1. 数据集概述

为了建立癌预后模型，本文从 UCSCXena 网站中选取膀胱癌(BLCA)、乳腺癌(BRCA)、结肠腺癌(COAD)、肺腺癌(LUAD)、肺癌(LUNG)和肉瘤(SARC)等多种癌症类型进行实验。预后模型的训练数据采用如下组学数据：基因水平拷贝数(Copy Number, gene-level)、RNAseq 基因表达(gene expression RNAseq)、DNA 甲基化(DNA methylation)和生存表型数据(phenotype-Curated survival data)。

2.2. 数据预处理

本文以乳腺癌数据为例。

Step 1. 构建患者生存标签

建立患者预后模型，首先需要确定患者是否被治愈。医学上以患者被治疗五年(1825 天)内不复发癌症作为患者是否被治愈的标准。因此可以根据生存时间构建患者的标签，并将其划分为两类：

长期生存者(标签为 1)：生存时间 OS.time ≥ 1825 天(即超过 5 年)

短期生存者(标签为−1)：生存时间 OS.time < 1825 且 OS = 1 (即死亡)

对于 OS 未知或生存状态不明确的样本，予以剔除。最终得到 301 例长期生存者与 131 例短期生存者，形成具有明确二分类标签的分析样本数据集合。

Step 2. 缺失值处理

由于乳腺癌的组学数据中存在缺失值，无法直接使用。针对这个问题，需要对缺失值进行处理。本文对数据缺失值采取如下策略。

(1) 剔除缺失率超过 20%的特征变量；

(2) 对其余缺失值采用均值填充策略进行补全。

四类组学中，仅 DNA 甲基化的表达数据存在较大比例缺失，该组学数据缺失值总数为 80,077,746 项。基于此，删除了 90,007 个缺失值超过 20.0%的特征列。

Step 3. 提取公共子集

由于不同组学数据的病人的样本编号不完全相同，预测模型的建立需要提取所有数据中具有相同编号的样本数据，确保分析对象在各数据类型中均有完整记录。基于此，对不同组学数据的样本编号取交集，共保留 249 个具有完整组学与生存表型信息的乳腺癌患者样本，其中包含 301 例长时间生存样本和 131 例子短时间生存样本。

Step 4. 特征筛选

经过上述处理，不同组学的特征个数远远大于样本数目，特征数过多一方面增加了模型的复杂度，另一方面特征多也引入了更多的噪声数据，对后续建立的模型的性能产生影响。因此，需要从原始特征中选择一些有效特征以降低数据维度，提高模型性能。由于低方差特征通常缺乏区分能力，因此可以通过计算。

各个特征的方差，设置筛选阈值，选择方差大于阈值的特征作为有效特征。

通过下面特征方差的箱线图 1 可以看出，筛选后的特征去除了过多的冗余和噪声特征。

Step 5. 标准化处理

不同的数据类型之间的量纲差异会对预测模型的性能产生非常大的影响，为了解决这个问题，采用“Z-score”标准化方法对预处理后的数据进行标准化处理。Z-score 通过将癌症的组学数据转换为均值为 0、标准差为 1 的分布，消除不同特征间因量纲或量级差异带来的干扰，使数据具备横向可比性。

具体而言，对于特征中的每个数据点 x，其标准化后的 z 值计算公式为：

$$z = \frac{x - \mu}{\sigma}$$

其中，μ 为该特征的均值，σ 为该特征的标准差。

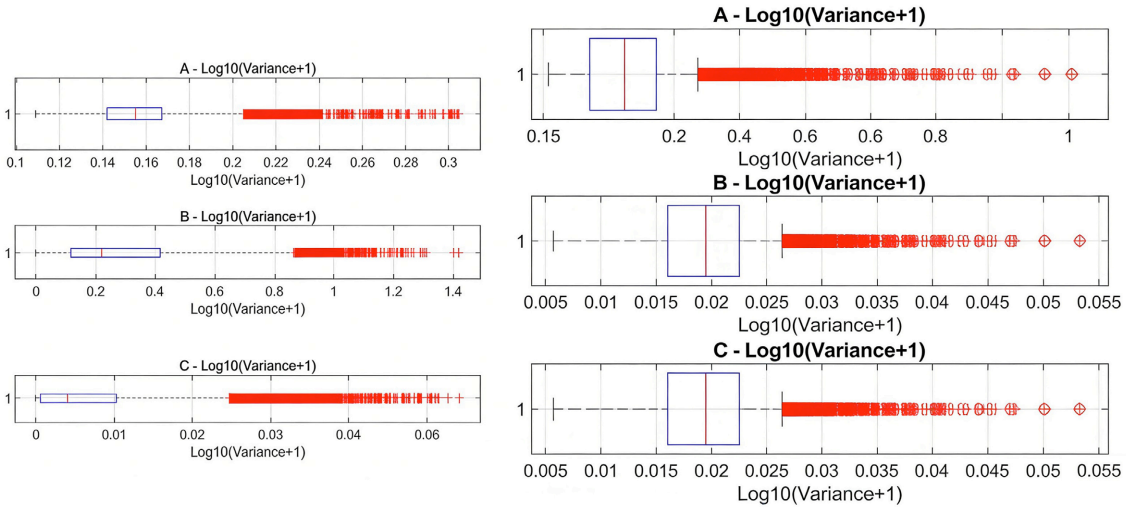

Figure 1. Original variance and filtered variance
图 1. 原始方差与筛选后的方差

数据预处理汇总

在下表中，本文汇总了不同癌症的组学数据经过上述预处理方法后的结果(如表 1、表 2 及图 2 所示)。

Table 1. Summary of data preprocessing
表 1. 数据预处理汇总

癌症名称	组学数据	原样本	标签	原特征	缺失特征	缺失 20%	方差阈值	处理后特征	共同样本
BLCA	A	436	236	8		0		2	215
	B	408	236	24,776	0	0	0.4642	12,386	215
	C	426	236	20,530	0	0	0.7689	10,265	215
	D	434	236	485,577	38,967,773	89,826	0.0259	122,640	215
BRCA	A	1236	432	8		0		2	249
	B	1080	432	24,776	0	0	0.4300	12,387	249
	C	1218	432	20,530	0	0	0.6624	10,265	249
	D	888	432	485,577	80,077,746	90,007	0.0189	128,651	249
COAD	A	545	165	8		0		2	98
	B	451	165	24,776	0	0	0.2961	12,388	98
	C	329	165	20,530	0	0	0.4330	10,265	98
	D	337	165	485,577	30,308,078	90,003	0.0145	142,930	98
LUAD	A	641	292	8		0		2	195
	B	516	292	24,776	0	0	0.4799	12,386	195

续表

LUAD	C	576	292	20,530	0	0	0.6166	10,265	195
	D	492	292	485,577	44,244,317	89,941	0.0131	117,479	195
LUNG	A	1145	560	8		0		2	386
	B	516	560	24,776	0	0	0.5131	12,388	386
	C	576	560	20,530	0	0	0.7100	10,265	386
	D	492	560	485,577	81,502,766	89,896	0.0178	108,546	386
SARC	A	271	149	8		0		2	141
	B	257	149	24,776	0	0	0.5004	12,386	141
	C	265	149	20,530	0	0	0.8269	10,265	141
	D	269	149	485,577	24,300,119	90,341	0.0304	128,629	141

Table 2. Summary of the number of cancer samples
表 2. 癌症样本数目汇总

癌症名称	长时生存	短时生存	可打标签样本	长时生存	短时生存	共同样本个数
BLCA	48	188	236	47	168	215
BRCA	301	131	432	179	70	249
COAD	54	111	165	41	57	98
LUAD	65	227	292	48	147	195
LUNG	149	411	560	109	277	386
SARC	60	89	149	57	84	141

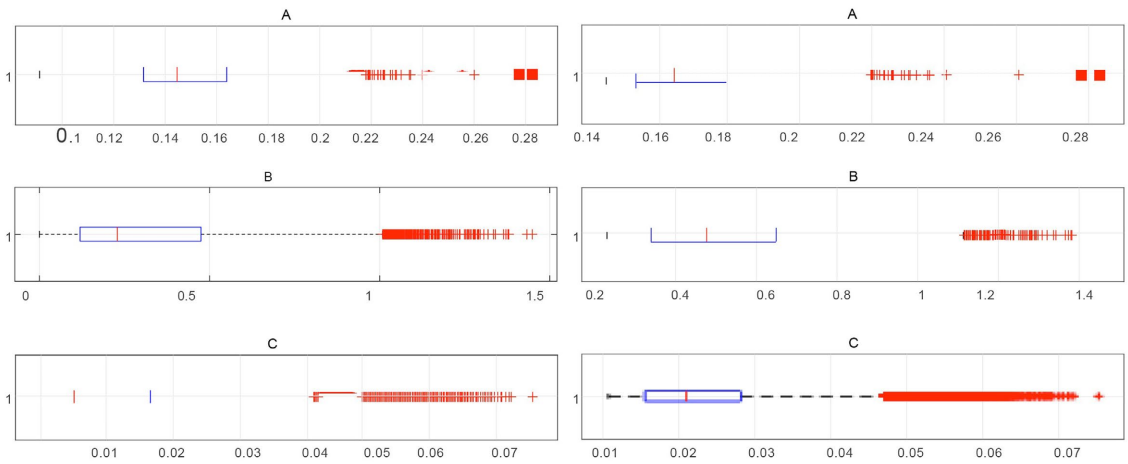

Figure 2. Comparison of BRCA features after variance filtering
图 2. BRCA 特征经过方差筛选后的对比图

2.3. 降维和特征选择

经过上述的数据预处理，我们已经得到了一个新的数据集合，容易发现组学特征数量仍然显著高于样本数量，这会产生严重“维数灾难”问题。这不仅显著增加模型训练的计算复杂度，而且容易使模型在训练集上过拟合，降低泛化性能。为了缓解这个问题，提升模型的稳定性和预测准确性，需要对不同组学数据的特征进行降维和筛选。

2.3.1. 特征降维

尽管数据经过了预处理，但剩余的组学数据仍然属于高维数据，其不同特征间仍存在冗余，本质维度可能较低。因此，可以通过降维技术最大程度保留数据中蕴含的关键信息的同时将高维数据转化为低维表示。具体来说，对数据进行降维能有效剔除原始数据中的冗余信息与噪声，提炼出更贴合任务需求的精简数据，为后续机器学习模型提供更优质的输入，从而提升模型的性能；此外，降维后的数据维度降低，意味着所需的存储空间和计算资源随之减少，从而可以加快模型的训练速度。

PCA 降维流程

PCA 是一种经典的线性无监督降维方法，其核心思想是通过正交变换将原始特征映射到一组新的主成分上，这些主成分是原始数据协方差矩阵的特征向量。我们在三类组学数据上分别应用了 PCA，为了充分保留样本数据信息，保留累计方差信息的比例被设置为 95%。

具体流程如下：

(1) 对数据进行标准化处理(零均值和单位方差)；

(2) 构建协方差矩阵并计算其特征值和特征向量；

(3) 计算累计贡献率，选取前 k 个主成分使其累计方差解释率 $\geq 95\%$；

(4) 将原始数据投影到所选主成分构成的子空间中，得到降维后的特征。

PCA 方法降维结果如表 3 (以乳腺癌数据为例)。

Table 3. PCA dimensionality reduction results
表 3. PCA 降维结果

数据类型	原始样本数目	原始特征维度	保留方差比例	降维后特征维度
A	249	12,387	95%	42
B	249	10,265	95%	28
C	249	128,651	95%	20

PLS 降维流程

与上述两种无监督降维方法不同，PLS 是一种有监督的降维方法其主要目标是从原始高维预测变量中提取一组低维潜变量(Latent Variables)。这些潜变量不仅能够最大程度地捕捉原始高维变量中的变异信息，还能够有效地解释响应变量中的变异信息。

PLS 降维的基本流程如下：

(1) 收集并整理高维数据集，将其分为预测变量(自变量)和响应变量(因变量)。

(2) 构建潜变量模型，通过最大化预测变量与响应变量之间的协方差将高维预测变量线性组合成少数几个潜变量。

(3) 通过对预测变量进行正交变换，提取出新的潜变量。

(4) 重复上述过程，提取足够数量的潜变量，得到降维的特征。

PLS 的降维结果如表 4：

Table 4. PLS dimensionality reduction results
表 4. PLS 降维结果

数据类型	原始样本数目	原始特征维度	保留特征个数	降维后特征维度
A	249	12,387	100	100
B	249	10,265	100	100
C	249	128,651	100	100

降维结果汇总

接下来，为直观呈现不同癌症的组学数据在分别在主成分分析(PCA)、偏最小二乘(PLS)三种方法下的降维结果，具体针对各组学数据的样本量、特征数量及降维阈值对应的降维结果进行汇总，具体详见表 5。

Table 5. Summary of dimensionality reduction results
表 5. 降维结果汇总

癌症名称	组学数据	样本	特征	PCA 阈值	PCA	PLS 阈值	PLS
BLCA	A	215	12,386	95	85	100	100
	B	215	10,265	95	167	100	100
	C	215	122,640	95	152	100	100
BRCA	A	249	12,387	95	88	100	100
	B	249	10,265	95	180	100	100
	C	249	128,651	95	167	100	100
COAD	A	98	12,388	95	36	97	97
	B	98	10,265	95	76	97	97
	C	98	142,930	95	73	97	97
LUAD	A	195	12,386	95	72	100	100
	B	195	10,265	95	147	100	100
	C	195	117,479	95	133	100	100
LUNG	A	386	12,388	95	106	100	100
	B	386	10,265	95	275	100	100
	C	386	108,546	95	243	100	100
SARC	A	141	12,386	95	76	100	100
	B	141	10,265	95	113	100	100
	C	141	128,629	95	105	100	100

2.3.2. mRMR 特征筛选

尽管 PCA/PLS 已经有效地降低了数据维度并消除了共线性，但我们引入 mRMR 进行二次筛选是基于以下三点考虑：信息相关性的差异：PCA 主要保留数据中方差最大的成分，但这部分方差并不一定对应于对分类标签(如患者预后)最具判别力的信息。mRMR 显式地引入了标签信息，确保筛选出的特征与分类任务高度相关。冗余消除：即使经过降维，特征间仍可能存在非线性冗余。mRMR 通过最小化特征间互信息，确保入选的特征集合具有最大的独立贡献。可解释性：PCA 生成的主成分是原始分子的线性组合，难以直接对应具体的生物通路。通过 mRMR 筛选出的特征子集保留了原始特征的含义，有助于揭示潜在的分子机制，为临床转化提供直接依据。

数据在经过上述不同降维方式分别降维后，数据的维度降低到一个更易于处理的水平。接下来，进一步利用 mRMR 方法筛选出对目标变量最具影响力的特征。这一过程不仅能够提升模型的预测能力，还能为后续的分析提供更为清晰的特征视角。通过 mRMR，能够有效地识别出在保留信息的同时，还能减少特征之间的冗余，从而为后续的建模和分析奠定基础。

以下是实现 mRMR 特征检测的具体步骤：

(1) 将每个组学数据集与样本标签串联，使得每个数据集都包含特征和标签。

(2) 通过随机抽样，按照 6:4 的比例将每个组学数据集划分为训练集和测试集。以乳腺癌数据为例，划分训练集共 149 个样本、测试集共 100 个样本。

(3) 对于每个组学数据的训练集，应用 mRMR 特征选择算法，指定筛选出 60 个特征。并记录下每次选择的特征。

(4) 重复步骤 2 和步骤 3，10 次，每次都随机划分训练集和测试集，并在新的训练集上进行 mRMR 特征选择。记录选择出的 70 个特征。

(5) 比较每次选择出的特征集，并检查它们是否完全相同。

通过上述步骤可以验证所选特征的稳定性和一致性，确保了特征不是随机选择得出的。

不同癌症类型的 mRMR 特征筛选结果汇总如表 6 所示。

Table 6. Summary of mRMR characteristic screening
表 6. mRMR 特征筛选汇总

癌症名称	组学数据	样本数目	PCA	筛选后	PLS	筛选后
BLCA	A	215	85	70	100	70
	B	215	167	70	100	70
	C	215	152	70	100	70
BRCA	A	249	88	70	100	70
	B	249	180	70	100	70
	C	249	167	70	100	70
COAD	A	98	36	36	97	70
	B	98	76	70	97	70
	C	98	73	70	97	70
LUAD	A	195	72	70	100	70
	B	195	147	70	100	70
	C	195	133	70	100	70
LUNG	A	386	106	70	100	70
	B	386	275	70	100	70
	C	386	243	70	100	70
SARC	A	141	76	70	100	70
	B	141	113	70	100	70
	C	141	105	70	100	70

2.3.3. 小结

本章聚焦数据预处理与特征选择关键环节，系统开展系列工作：

数据概述阶段，统计数据规模、缺失值与异常值分布等基础信息，全面呈现数据初始状态，为后续处理提供精准依据。

数据预处理环节，针对性解决数据问题：缺失值按特征性质与分布，采用均值、中位数或众数等方法填充；异常值通过 Z-score 法、箱线图法识别后，结合业务逻辑判断保留、修正或删除；同时开展数据标准

化与特征编码，消除量纲差异、转换非数值特征，提升数据质量与可用性，筑牢后续研究的数据基础。

特征选择阶段，先明确降维与特征选择在简化维度、去除冗余上的协同作用，再分两类开展筛选：一是传统特征筛选，通过方差阈值、相关系数等统计指标初步筛选，结合机器学习模型进一步遴选，剔除低贡献度、多重共线性特征；二是重点采用 mRMR 策略，以互信息为核心，最大化特征与目标变量相关性、最小化特征间冗余，挖掘代表性强、互补性优的特征子集。通过综合筛选，从原始高维数据中提炼优质特征集合，为后续模型训练与性能优化提供高质量数据支撑，保障研究结果的准确性与可靠性。

3. 数值实验

本章通过支持向量机(SVM)，logistc 回归(LR)两种不同的机器学习方法分别建立癌症病人预后生存模型。这是一个分类模型，模型的目的是通过上述的特征集合将病人分类为长期生存和短期生存两种类型，以此达到通过组学数据判断病人的癌症治疗后的生存情况。基于上一节对数据进行特征工程得到的数据，本节使用三种不同的机器学习方法，分别建立癌症预后生存模型。

3.1. 实验准备

3.1.1. 划分数据集

在实验中，本研究将所用的数据集采用分层抽样的方式划分为训练集与测试集，以保证训练集与测试集中各类别样本的分布比例一致，避免因数据分布偏差对模型训练与评估结果产生干扰。最终确定训练集与测试集的划分比例为 7:3，即 70%的样本用于模型训练，30%的样本用于模型性能评估。

3.1.2. 模型参数设置

由于机器学习对模型的参数非常敏感，在本节中，我们首先给出数值实验所需的实验平台、软件、用到的工具箱以及不同算法的参数设置等内容。

为确保模型性能的可比性，对于有监督学习模型，其超参数主要通过贝叶斯优化(Bayesian Optimization)结合 10 折交叉验证在训练集上进行自动调优。人工神经网络的结构则通过多次试验确定。各模型的最终参数设置如下。

支持向量机(SVM)

对于 SVM 模型，本文使用“fitcsvm”函数进行训练，并且选择高斯核(RBF 核)作为核函数来捕捉数据中潜在的非线性关系。核函数使用启发式程序自动选择核尺度，并将误分类惩罚项的参数“C”的取值集合设置为 Ω，其中

$$\Omega = \left\{2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}, 2^{1}, 2^{3}, 2^{5}, 2^{7}\right\}$$

具体的参数如表 7 所示：

Table 7. Parameter settings of SVM model
表 7. SVM 模型的参数设置

函数名	关键参数	参数设置
(fitcsvm)	KernelFunction BoxConstraint KernelScale	’gaussian’ 在集合 Ω 中进行网格搜索 ’auto’

Logistic 回归(LR)

对于 LR 模型，本文使用‘fitclinear‘函数进行训练，并使用 lasso 损失函数作为正则项，其中正则项的参数“λ”的取值集合设置为 Ω，其中

$$\Omega = \left\{2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}, 2^{1}, 2^{3}, 2^{5}, 2^{7}\right\}$$

具体的参数如下表 8 所示：

Table 8. Parameter settings for the LR model
表 8. LR 模型的参数设置

函数名	关键参数	参数设置
(fitclinear)	Regularization Lambda Learner	,lasso, 在集合 Ω 中进行网格搜索 ,logistic,

3.1.3. 模型评估

为全面、客观地比较四种模型的性能，本研究采用统一的评估指标在测试集上对所有模型进行评估，所选指标包括准确率(Accuracy)和 F1 值(F1-Score)，各指标的定义与计算方式如下：

准确率(Accuracy)：指模型正确预测的样本数占总测试样本数的比例，计算公式为：Accuracy = (TP + TN)/(TP + TN + FP + FN) * 100，其中 TP 为真正例(正类被正确预测为正类)，TN 为真负例(负类被正确预测为负类)，FP 为假正例(负类被错误预测为正类)，FN 为假负例(正类被错误预测为负类)。该指标反映了模型整体的预测正确性，但在类别不平衡数据中可能存在偏差，因此需结合其他指标综合评估。

F1 值(F1-Score)：精确率和召回率的调和平均数，计算公式为：$F1 = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$，用于综合评价模型的性能，避免单一指标的局限性，F1 值越接近 1，模型性能越优。

在评估过程中，所有指标的计算均由十次交叉验证取平均值的方法得到。

3.2. 实验结果

本节实验中，我们系统评估了所提预后模型的性能。我们组合了 2 种降维方法(PCA、PLS)与 2 种机器学习分类器(SVM、LR)，共计 4 种组合模型，并从多个维度对其进行了比较。

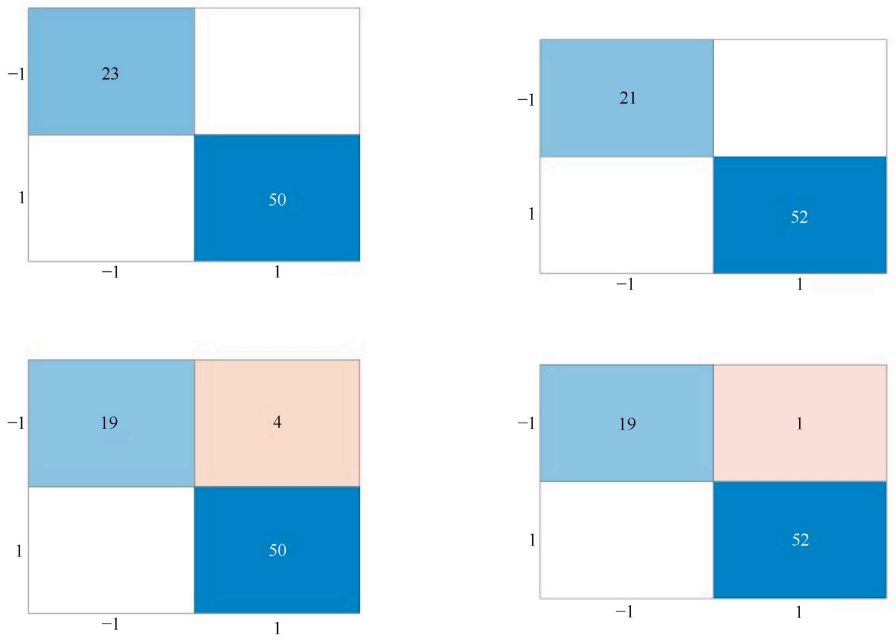

Figure 3. Confusion matrices of SVM and LR in BRCA
图 3. SVM、LR 在 BRCA 中的混淆矩阵

模型性能采用十次交叉验证进行评估，主要指标包括准确度(Accuracy)和 F1-得分(F1-Score)。各项指标的平均值详见表 9。

为更直观地展示分类结果，我们在图 3 中报告了所有组合模型的混淆矩阵。该系列图的布局规则如下：图被均匀划分为四个象限，分别对应保留的两种降维方法(左列：PCA，右列：PLS)；在每个降维方法对应的象限内，进一步展示两种核心分类器的结果(上行：SVM，下行：LR)，直观呈现“2 种降维 +2 种分类”的核心组合模型分类效果，剔除所有冗余的降维方法与分类器相关展示内容。

此外，为验证模型的稳定性，我们在图 4 中展示了所有 4 种组合模型在十次交叉验证中的准确度及其标准差。

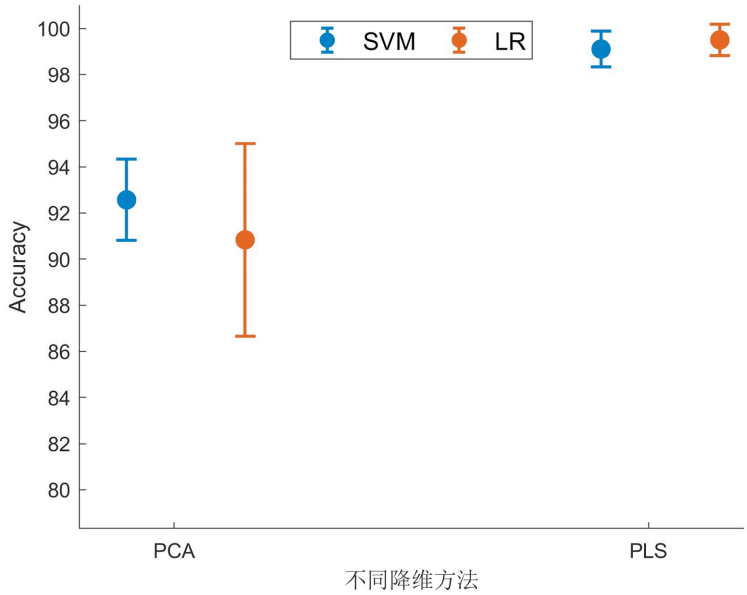

Figure 4. Accuracy and standard deviation (BRCA) of combined models with different dimensionality reduction methods and machine learning algorithms
图 4. 不同降维方法与机器学习算法组合模型的准确度和标准差(BRCA)

Table 9. Performance evaluation of combined models with different dimensionality reduction methods and machine learning algorithms (BRCA)
表 9. 不同降维方法与机器学习算法组合模型的性能评估(BRCA)

评估指标	降维方法	SVM	LR
准确度	PCA	92.57	90.83
	PLS	99.32	99.51
F1-得分	PCA	0.95	0.94
	PLS	0.99	1.00

通过图 4 可以得出，基于 PLS 降维的方法在所有分类器上均取得了最优或接近最优的性能，其各项指标都显著高于其他降维方法，并且在十次交叉验证中的结果最为稳定。这一对比初步表明，降维方法的选择对预后模型的性能有决定性影响，且其效果因后续采用的分类算法不同而存在显著差异。

以乳腺癌数据为例，按降维方法来看，PLS 降维算法的表现全面领先于其他降维算法。它在所有评估指标和所有分类器上，PLS 都取得了最佳或接近最佳的成绩。其准确度(最高 99.51%)、F1-score (最高 1.00)都接近理论极限。这再次强有力地证明了监督降维方法在利用标签信息提取与分类最相关特征方面

的巨大优势。对于生物医学分类任务，PLS 是一个非常可靠的选择。此外，PCA 也是一个稳健可靠的降维方法。PCA 作为无监督方法，虽然无法与 PLS 媲美，但整体上依然良好且稳定。它与所有分类器组合的 F1-score 都在 0.91 以上。这表明 PCA 成功捕获了数据中最重要的方差结构，这些结构对于区分样本类别仍然非常有效。

按机器学习算法来看，SVM 和 LR 的分类表现稳定且强劲。这两个算法对各种降维方法的“容忍”都很高。LR 的分类表现反映了输入特征的质量。使用最好的 PLS 特征时，它取得了所有模型中最高的 F1-score (1.00)。

按表中的评估指标来看，PLS 降维算法加持下的分类算法的各项指标全面领先。所有模型的精确度(Precision)都普遍较高(>75%)，但是召回率显示较大差异。PCA、PLS 与 SVM/LR 的组合召回率极高(>98%)，意味着几乎抓住了所有正样本。F1-Score 最直观地反映了 PLS 方法的优势。

因此，通过各项指标的分析，我们可知，PLS + LR/SVM 在 BRCA 数据集上是非常高效的组合。LR 达到了理论最佳性能。这说明了生物医学数据集的标签信息对于提取关键至关重要。任何能利用标签信息的降维或特征选择方法都可能带来性能提升。

3.3. 预后生存曲线

为了评估建立的预后生存模型的预测性能，我们在不同癌症类型的测试集上进行了生存分析。经过前文分析，我们选择 PCA 和 PLS 作为降维方法，SVM 和 LR 为机器学习算法，两两组合建立预后生存模型。

首先，使用训练好的预后生存模型计算测试集中每位患者的预测风险评分(risk score)。随后，我们以患者的五年存活时间(1825 天)为临界值，将患者划分为高风险组(预测为短期生存)与低风险组(预测为长期生存)。

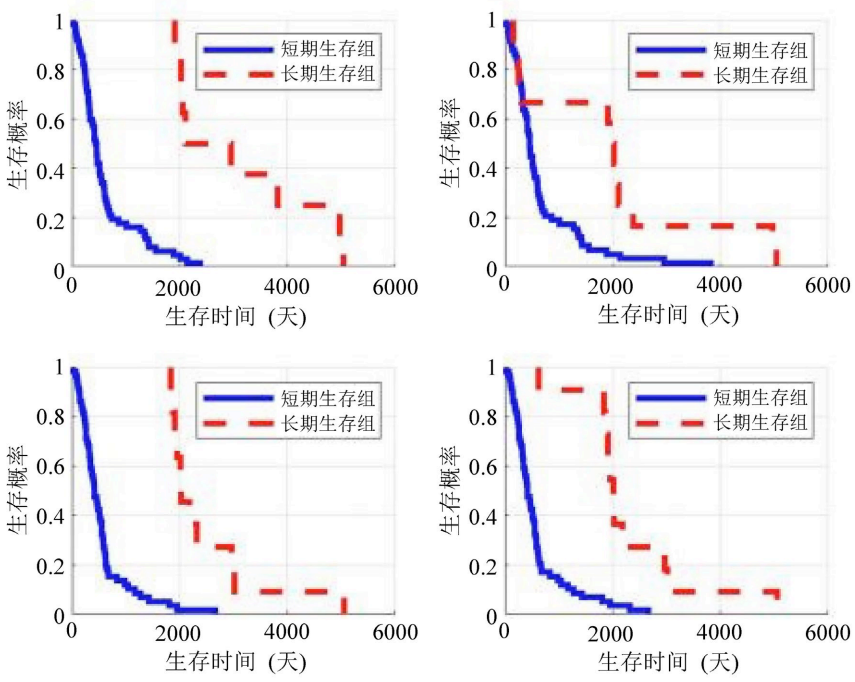

Figure 5. K-M curves of the combined model integrating dimensionality reduction and machine learning for BLCA cancer types
图 5. 降维和机器学习的组合模型在 BLCA 癌症类型上的 K-M 曲线

接下来，我们整合了从 TCGA 数据库获取的患者临床随访数据，包括总体生存状态(OS)和生存时间(OS.time)。采用 Kaplan-Meier 方法分别绘制高风险组和低风险组的生存曲线，并使用 Log-Rank 检验来评估两组患者之间的生存差异是否具有统计学意义。具体的实验结果如下图 5、图 6 所示。

根据图 5 中的 Kaplan-Meier 生存分析曲线所示，低风险组患者的总体生存率显著高于高风险组。这表明基于 PCA PLS 降维算法和 SVM，LR 机器学习算法组合构建的预后生存预测模型能够有效区分不同预后风险的患者，证明了模型具有良好的预后预测价值。

此外，通过表 10 可以发现两个分类模型预测结果的对数秩检验 P 值均小于 0.05，这表明本文建立的模型可以将患者的生存状况准确分层，具有很强的预后价值，证明了预测癌症病人预后生存的潜力。

Table 10. Summary of Log-Rank tests
表 10. Log-Rank 检验汇总表

癌症类型	PCA + SVM	PCA + LR	PLS + SVM	PLS + LR
BLCA	4.24e−05	2.84e−03	1.06e−06	1.72e−05
BRCA	1.35e−11	6.22e−07	5.76e−07	7.63e−09
COAD	1.37e−03	9.02e−03	1.94e−02	4.00e−05
LUAD	3.92e−03	2.33e−02	1.12e−07	9.69e−09
LUNG	3.13e−09	6.84e−07	4.61e−11	5.52e−18
SARC	4.55e−06	3.08e−05	9.91e−03	8.11e−03

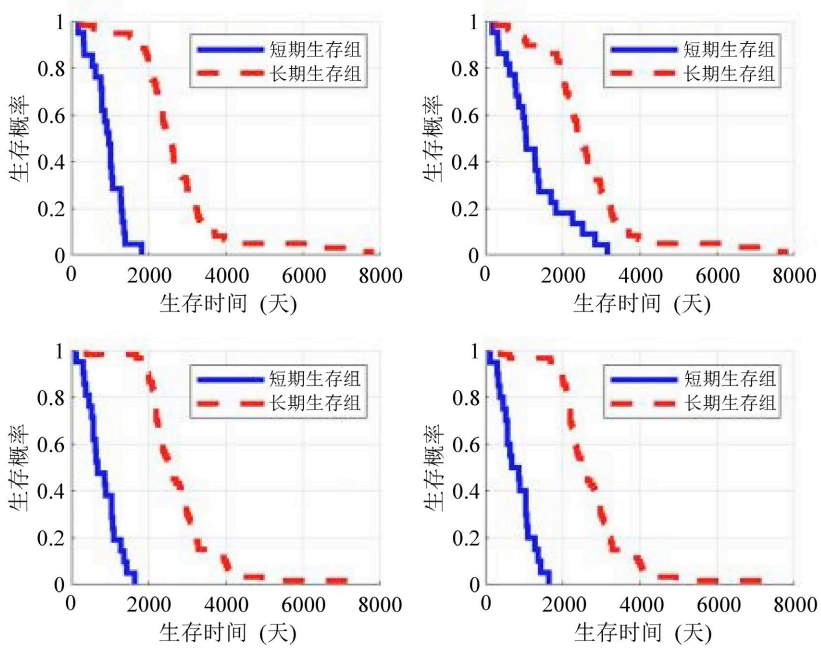

Figure 6. K-M curves of the combined model integrating dimensionality reduction and machine learning for BRCA cancer subtypes
图 6. 降维和机器学习的组合模型在 BRCA 癌症类型上的 K-M 曲线

参考文献

[1] Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., *et al*. (2024) Global Cancer Statistics 2022: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA*: *A Cancer*

Journal for Clinicians, **74**, 229-263. https://doi.org/10.3322/caac.21834

[2] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. and Jemal, A. (2018) Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA*: *A Cancer Journal for Clinicians*, **68**, 394-424. https://doi.org/10.3322/caac.21492

[3] Weinberg, R.A. and Weinberg, R.A. (2006) The Biology of Cancer. WW Norton & Company.

[4] Luo, L,. Wang, X., Lin, Y., *et al*. (2024) Deep Learning in Breast Cancer Imaging: A DECADE of progress and Future Directions. *IEEE Reviews in Biomedical Engineering*, **18**, 130-151.

[5] Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., *et al*. (2013) The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nature Genetics*, **45**, 1113-1120. https://doi.org/10.1038/ng.2764

[6] Verhage, R.J., Hazebroek, E., Boone, J., *et al*. (2009) Minimally invasive Surgery Compared to Open Procedures in Esophagectomy for Cancer: A Systematic Review of the Literature. *Minerva Chirurgica*, **64**, 135-146.

[7] Lee, Y.T., Tan, Y.J. and Oon, C.E. (2018) Molecular Targeted Therapy: Treating Cancer with Specificity. *European Journal of Pharmacology*, **834**, 188-196. https://doi.org/10.1016/j.ejphar.2018.07.034

[8] Sharma, P. and Allison, J.P. (2015) The Future of Immune Checkpoint Therapy. *Science*, **348**, 56-61. https://doi.org/10.1126/science.aaa8172

[9] Gebski, V., Garès, V., Gibbs, E. and Byth, K. (2018) Data Maturity and Follow-Up in Time-To-Event Analyses. *International Journal of Epidemiology*, **47**, 850-859. https://doi.org/10.1093/ije/dyy013

[10] Ye, T., Shao, J. and Yi, Y. (2024) Covariate-Adjusted Log-Rank Test: Guaranteed Efficiency Gain and Universal Applicability. *Biometrika*, **111**, 691-705. https://doi.org/10.1093/biomet/asad045

[11] K, M.S., Preetha, J., Reddy, K.N., Ramya, S., S, Y. and Murugan, S. (2024) Survival Analysis with Cox Proportional Hazards Model in Predicting Patient Outcomes. 2024 5*th International Conference on Electronics and Sustainable Communication Systems* (*ICESC*), Coimbatore, 7-9 August 2024, 1155-1161. https://doi.org/10.1109/icesc60852.2024.10689732