Operations Research and Fuzziology 运筹与模糊学, 2023, 13(5), 5464-5475
Published Online October 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.135546

Hans汉斯

机器学习算法在乳腺癌预测中的应用

郭昱君

南京信息工程大学数学与统计学院，江苏 南京

收稿日期：2023年9月5日；录用日期：2023年10月16日；发布日期：2023年10月24日

摘　要

乳腺癌是世界上女性最常见的恶性肿瘤，治愈乳腺癌的关键在于早期的诊断和治疗。及时诊断肿瘤对临床治疗具有重要意义，因此，找到一种能够准确识别肿瘤类型并尽早进行治疗的算法变得尤为关键。本文介绍了在威斯康星州诊断乳腺癌数据集上使用了lasso算法进行特征筛选，然后基于这些特征训练了随机森林分类器来预测乳腺癌的良性或恶性。结果显示，预测模型的准确率为95.32%，召回率为92.06%，F1分数为93.55%，通过这些指标的综合评估，证明这种方法可以有效地进行乳腺癌良恶性的预测，具有潜在的应用价值。总的来说，文中提供了一种有力的方法，可以对癌症数据进行预测，并优化分类器的性能。这种方法可以帮助医生更好地诊断乳腺癌，促进更好的治疗和预防，对乳腺癌的研究具有重要的意义。

关键词

机器学习，Lasso，随机森林，ROC曲线

Application of Machine Learning Algorithms in Breast Cancer Prediction

Yujun Guo

School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing Jiangsu

Received: Sep. 5th, 2023; accepted: Oct. 16th, 2023; published: Oct. 24th, 2023

Abstract

Breast cancer is the most common malignant tumor in women worldwide, and early diagnosis and treatment are key to curing breast cancer. Timely detection of tumors is of great significance for clinical treatment, so finding an algorithm that can accurately identify tumor types and start treatment early is crucial. This article introduces the use of the lasso algorithm for feature selec-

文章引用: 郭昱君. 机器学习算法在乳腺癌预测中的应用[J]. 运筹与模糊学, 2023, 13(5): 5464-5475.
DOI: 10.12677/orf.2023.135546

tion on a breast cancer diagnostic dataset in Wisconsin. Based on these features, a random forest classifier was trained to predict the benign or malignant nature of breast cancer. The results showed an accuracy of 95.32%, a recall rate of 92.06%, and an F1 score of 93.55% for the predictive model. Through a comprehensive evaluation of these metrics, it is proven that this method can effectively predict the benign or malignant nature of breast cancer and has potential practical value. In summary, the article provides a powerful method for predicting cancer data and optimizing the performance of classifiers. This approach can help doctors better diagnose breast cancer, promote better treatment and prevention, and has significant implications for breast cancer research.

Keywords

Machine Learning, Lasso, Random Forest, ROC Curve

1. 引言

癌症是人类致死率居全球第一位的疾病。癌症会在人体的某个部位形成一个被称为“肿瘤”的组织。这些肿瘤可能会长在人体的任何一个器官，比如大脑，胸部，肾脏等等。年龄、性别、遗传、生活质量等是影响肿瘤发生和扩散的主要因素。医生可从病人身上取下活体组织切片或从手术中取得局部组织，来判断其是否患癌。一旦诊断出肿瘤，就可以采用外科手术，化学疗法和放射疗法相结合的方法来治疗。大部分癌症都是可以治疗的，而且随着科学和技术的发展，已经研制出了很多专门用于某些癌症的特殊药物。但是一旦肿瘤扩散转移到身体的其他部位，侵入或摧毁其它细胞和组织时，则很难治愈，严重时可致患者死亡。肿瘤可分为良性肿瘤和恶性肿瘤。良性肿瘤的肿瘤细胞不会出现转移或者侵入到周围组织的现象，这类肿瘤对人体一般没有危险，不会导致死亡。恶性肿瘤则更加的危险，恶性肿瘤的起因是人体内某些细胞失去了正常的增殖能力，分裂和侵袭其它细胞。大部分的恶性肿瘤都会对身体的正常功能造成一定的影响，主要表现为癌细胞会在淋巴系统中扩散，然后继续生长，破坏正常的组织，使新生的血管供自己所需，从而引起贫血。恶性肿瘤严重威胁着人们的健康。早期发现癌症并且进行治疗是非常关键的，因此，研究者们试图建立智能系统帮助医生对癌症进行早期诊断。

2018 年，中国肿瘤研究中心发表了一份中国肿瘤病死率与生存率统计，据统计数据显示，2014 年全国新发的恶性肿瘤病例约有 380.4 万例，死亡病例高达 229.6 万例，其中乳腺癌在女性癌症中发病率占第一，每年有约 27.9 万的新发病例，而且在城镇居民中的比例更高[1]。乳腺癌的死亡率在过去 20 年里并没有发生改变。近年来，其发病率稳步上升，但是随着对癌症的早期发现和治疗技术的进步，乳腺癌的死亡率在不断降低。虽然诊断和治疗手段已经取得了不错的进展，但乳腺癌依然是导致女性死亡率最高的疾病。在对早期的乳腺癌进行诊断时，经常由有经验的医生根据乳腺 X 线摄影照片来对乳房肿瘤类别进行预测[2]。但是由于医生的主观因素，往往会出现漏诊、误诊的情况，因此，提高乳腺癌肿瘤的诊断准确率就显得尤为重要。

乳腺癌的早期诊断与治疗对提高肿瘤的治愈率、降低病死率、降低患者的经济负担具有重要意义。因此，对乳腺癌的研究有着非常重要的意义。常规的乳腺超声检查和 CT 等检查方法会受到医生主观判断的影响，由于医师的诊断经验及知识水平等因素，会对结果产生较大的影响。计算机辅助诊断[3]疾病有助于提高医生诊断的敏感性和特异性，提高诊断的准确率。因此，采用机器学习算法建立辅助诊断模

型，并对乳腺癌诊断的规则进行归纳总结，从而协助医生对乳腺癌疑似患者进行迅速而又精确的诊断。逻辑回归、决策树、KNN、支持向量机、朴素贝叶斯等多种经典的分类算法都可以对乳腺癌临床数据进行训练。由于临床数据规模巨大、数据类型繁多，在使用机器学习方法对乳腺癌数据进行研究前，有必要对数据进行预处理的操作，比如对缺失值进行简单处理，以及特征选取等等。然而，现有的乳腺癌临床资料存在着不完备、维度较多且含有非必要的特征等问题。医学临床数据的真实性和质量是研究乳腺癌的关键，也是做出正确决策的重要依据。

综上所述，对乳腺癌良恶性分类的研究不仅有助于协助医生的诊断，还可以在早期发现乳腺癌且增加治疗方案以及提高临床疗效，降低患者死亡率，对全球乳腺癌的防治有着重要的科学意义。本文运用合适的算法筛选出能够用于乳腺癌早期预警的最佳特征子集，并在此基础上构建辅助诊断模型，为临床医生提供更多的参考。

2. 方法

2.1. Lasso

Lasso 方法最早由 Robert Tibshiran 于 1996 年提出[4]，文章发表在“统计四大”之一的皇家统计学会期刊上，尽管至今已有二十多年，但依然有着广泛的应用，由其发展出的方法层出不穷[5]。首先了解必要知识背景，即线性模型：

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\epsilon, E(\epsilon)=0, Cov(\epsilon)=\sigma^2 I \tag{1}$$

考虑线性模型的参数 $\boldsymbol{\beta}$ 和 σ^2 的估计问题，这里 $\boldsymbol{y}$ 是 $n\times 1$ 观测向量，$\boldsymbol{X}$ 是 $n\times p$ 的设计矩阵，$\boldsymbol{\beta}$ 是 $p\times 1$ 未知参数向量，ϵ 为随机误差，σ^2 为误差的方差。

估计参数向量的基本方法是最小二乘法，其思想是使得误差向量 $\epsilon=\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}$ 尽可能的小，也就是使

$$Q(\boldsymbol{\beta})=\|\epsilon\|^2=\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2=(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \tag{2}$$

到达最小。最后，使得上式达到最小值的解为

$$\hat{\beta}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{3}$$

向量的范数：

向量的 1-范数：$\|\boldsymbol{X}\|_1=\sum_{i=1}^{n}|x_i|$ 向量内各元素的绝对值之和

向量的 2-范数 $\|\boldsymbol{X}\|_2=\left(\sum_{i=1}^{n}x_i^2\right)^{\frac{1}{2}}=\sqrt{\sum_{i=1}^{n}x_i^2}$ 元素的平方和再开平方

Lasso 就是在目标函数 $Q(\boldsymbol{\beta})$ 后面加了一个 1-范数

$$Q(\boldsymbol{\beta})=\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2+\lambda\|\boldsymbol{\beta}\|_1 \rightleftharpoons \arg\min\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 \text{ s.t.} \sum|\beta_j|\le s \tag{4}$$

高维数据即 $p>n$ 的情况，现在随着数据采集能力的提高，特征数采集越来越多，但是其中可能有很多特征是不重要的，系数很小，如果用岭回归，不重要的变量也会被估出来，而用 Lasso 方法，就可以把这些不重要变量的系数压缩为 0，既实现了较为准确的参数估计，也实现了变量降维。

LASSO 的计算相对复杂。由于惩罚项中含有绝对值，函数的导数是连续不光滑的，所以无法进行求导并使用梯度下降优化。坐标下降法是每次选择一个维度的参数进行一维优化，然后不断的迭代对多个维度进行更新直到函数收敛。因此可以使用坐标下降法计算回归系数。

RSS 部分：

$$RSS(w)=\sum_{i=1}^{m}\left(y_i-\sum_{j=1}^{n}x_{ij}\beta_j\right)^2 \tag{5}$$

求导：

$$\begin{aligned}\frac{\partial RSS(\beta)}{\partial w_k}&=-2\sum_{i=1}^{m}x_{ik}\left(y_i-\sum_{j=1}^{n}x_{ij}\beta_j\right)\\&=-2\sum_{i=1}^{m}\left(x_{ik}y_i-x_{ik}\sum_{j=1,j\neq k}^{n}x_{ij}\beta_j-x_{ik}^2\beta_k\right)\\&=-2\sum_{i=1}^{m}x_{ik}\left(y_i-\sum_{j=1,j\neq k}^{n}x_{ij}\beta_j\right)+2\beta_k\sum_{i=1}^{m}x_{ik}^2\end{aligned} \tag{6}$$

令 $p_k=\sum_{i=1}^{m}x_{ik}\left(y_i-\sum_{j=1,j\neq k}^{n}x_{ij}\beta_j\right)$，$z_k=\sum_{i=1}^{m}mx_{ik}^2$ 得到：

$$\frac{\partial RSS(\beta)}{\partial \beta_j}=-2p_k+2z_k\beta_k\ \frac{\partial RSS(\beta)}{\partial \beta_j}=-2p_k+2z_k\beta_k \tag{7}$$

正则项：

$$\lambda\frac{\partial\sum_{i=1}^{n}|\beta_j|}{\partial\beta_k}=\begin{cases}-\lambda & \beta_k<0\\ [-\lambda,\lambda] & \beta_k=0\\ \lambda & \beta_k>0\end{cases} \tag{8}$$

这样整体的偏导数:

$$\begin{aligned}\frac{\partial f(\beta)}{\partial\beta_k}&=2z_k\beta_k-2p_k+\begin{cases}-\lambda & \beta_k<0\\ [-\lambda,\lambda] & \beta_k=0\\ \lambda & \beta_k>0\end{cases}\\&=\begin{cases}2z_k\beta_k-2p_k-\lambda & \beta_k<0\\ [-2p_k-\lambda,-2p_k+\lambda] & \beta_j=0\\ 2z_k\beta_k-2p_k+\lambda & \beta_k>0\end{cases}\end{aligned} \tag{9}$$

令 $\frac{\partial f(\beta)}{\partial\beta_k}=0$ 得到

$$\widehat{\beta_k}=\begin{cases}(p_k+\lambda/2)/z_k & p_k<-\lambda/2\\ 0 & -\lambda/2\le p_k\le\lambda/2\\ (p_k-\lambda/2)/z_k & p_k>\lambda/2\end{cases} \tag{10}$$

通过上面的公式我们便可以每次选取一维进行优化并不断迭代得到最优回归系数。

最小角回归法(LARS)是 Bradley Efron 于 2004 年的论文《Least Angle Regression》中提出的一种用于高维数据的回归算法[6]。在每一步中，它都会找到与目标最相关的特征。当多个特征具有相等的相关性时，它不是沿着相同的特征继续进行，而是沿着特征之间角平分线的方向进行。首先计算所有的相关性，刚开始最大的加入点就是 y 本身，相关为 1，然后选择最大的相关性为第一个加入点，然后游走，不停的计算相关系数，当减小到等于第二个相关系数时，把第二个变量加入，然后按照这两个角平分线继续游走，以此类推加入第三个，第四个，直到全部加入，然后游走到相关系数为 0，最终寻找到最优点。

2.2. 随机森林

随机森林是一种机器学习算法[7]。机器学习中有一种大类叫集成学习(Ensemble Learning)，集成学习的基本思想就是将多个分类器组合，从而实现一个预测效果更好的集成分类器。集成算法大致可以分为：Bagging，Boosting 和 Stacking 三大类型。随机森林采用 Bagging 的思想，所谓的 Bagging 就是每次有放回地从训练集中取 n 个训练样本，组成新的训练集，利用新的训练集，训练得到 M 个子模型，对于分类问题，采用投票的方法，得票最多子模型的分类类别为最终的类别，随机森林以决策树为基本单元，通过集成大量的决策树，就构成了随机森林。

基尼指数(Gini index)：随机森林使用“基尼指数”来选择划分属性。基尼指数越小，则数据集的纯度越高。对于给定的集合 D，其基尼指数为

$$\mathrm{Gini}(D)=1-\sum_{k=1}^{k}\left(\frac{|C_k|}{|D|}\right)^2 \tag{11}$$

这里，C_k 是 D 中属于第 k 类的样本子集，k 是类的个数。

2.3. 特征选择

在实际的机器学习中，特征的数量通常很多，而且这些特征之间并不是线性的，而是相互依存的，这就造成了对特征的分析和训练所需要的时间变得更长，造成了“维度灾难”，模型变得更复杂，模型的推广能力变得更差。为此，我们拟通过去除与分类无关或冗余的特征，降低特征数目，并对其进行有效的降维，以达到提高预测精度和降低计算时间的目的[8]。在不影响类别分布，不降低分类精度，并具有较好的稳定性和自适应能力的条件下，提取尽可能小的特征子集。特征选择就是根据某个最优标准，从特征空间中选取具有较高分类能力的子集。特征选取方法有过滤法，缠绕法，嵌入法[9]。

过滤法是利用样本数据自身的固有属性，比如距离、相关性等，来作为评价，模型学习中具有重要意义或相关性，且与分类算法无关。它的优势在于，它只与数据有关，并且具有快速的计算速度。能够快速地获得了特征子集合。但其缺点是该方法与分类算法无关，并且如果没有注意到所选择的特征子集中的相互关系，就会在最终的特征子集中产生大量的冗余特征，增大了运算的复杂性，从而降低了算法的效率。

缠绕法，它是按照分类算法的某个性能指标，对一个特征或者一个特征子集进行评价，一般以分类准确率作为一个评估函数，选取最佳的特征子集为最后的特征集。因为缠绕法与不同的分类算法相结合，所以缠绕法的分类效果要好得多。但是缠绕法在选取特征的时候要反复运用分类法，提高了运算的复杂性，而且筛选出来的基因也不具有明显的生物意义。

嵌入法在分类算法中完整地嵌入特征选择过程，嵌入法的操作速度要比缠绕法快得多，并且在选取的过程中和理论上都有很大的差异。但是和缠绕方法一样，嵌入方法取决于学习算法，所以泛化能力较差。

以上是对机器学习中常用的特征选择算法的介绍，与之相对应的还有统计领域中常用的特征选择算法。在统计学中，普遍的统计模型是通过数学统计模型来建立的。将变量之间的函数关系用数学方程式的形式表达出来，通常计算出模型的残差平方和大小用来评估模型的契合度情况。过拟合模型的复杂性一般都很高，因此采用“惩罚”概念，并通过增加惩罚项来限制参数空间的大小，从而减少模型的规模。

3. 模型建立

3.1. 数据来源

本文的数据来源于威斯康星州诊断乳腺癌数据集(WDBC) [10]，这个数据集可以从 UCI 数据库获得，数据集共 569 条数据，32 列，其中有 30 个特征，剩余两列分别是 ID 和 Diagnostic，主要针对的是细胞核特征，

是连续型特征，每个样本的标签为乳腺良性肿瘤与恶性肿瘤，共有 357 个良性肿瘤，212 个恶性肿瘤。

在数据集中，通过对乳腺肿瘤的细针穿刺得到的数字化图像进行计算，得出特征值，该数据集的特征值反映了样本图像中细胞核的形态学特征。对于每个样本图像特征，都会计算出其平均、方差和最大值，从而得到 30 个特征。例如，1 号特征是平均半径，11 号特征是半径标准差，21 号特征是最大半径。表 1 显示出前十个特征名及其含义解释。

Table 1. Top ten features
表 1. 前十个特征

特征	解释
diagnosis	诊断标签：malignant =恶性，benign =良性
radius_mean	半径，即细胞核从中心到周边点的距离平均值
texture_mean	纹理(灰度值的标准偏差)平均值
perimeter_mean	细胞核周长平均值
area_mean	细胞核面积平均值
smoothness_mean	平滑度(半径长度的局部变化)平均值
compactness_mean	紧凑度(周长^2/面积−1.0)平均值
concavity_mean	凹度(轮廓凹部的严重程度)平均值
concave points_mean	凹点(轮廓凹部的数量)平均值
symmetry_mean	对称性平均值
fractal_dimension_mean	分形维数−1 平均值

3.2. 研究设计

本文使用 python 软件首先对特征数据用 lasso 回归筛选出部分特征，在 Scikit-learn 包中，默认情况下 Lasso 回归使用最小角回归算法进行实现。具体来说，该算法在每一轮迭代中，会选择一对训练数据与特征，并将该特征向量进行旋转，使其与目标数据向量之间的夹角最小。然后，Lasso 回归算法会将特征系数向量的范数逐步增大，并逐渐压缩特征系数的大小，直到找到最优的正则化参数，使得算法能够实现最小化代价函数。

随后用 lasso 算法得到的特征用随机森林算法对训练集进行训练，再用测试集进行测试，得到准确率，混淆矩阵，F1 值，ROC 曲线等对模型进行评价。

3.3. 数据预处理及可视化

首先对数据进行缺失值的查询，结果发现无缺失值，对所有特征数据进行标准化。数据集共 569 条数据，32 列，其中有 30 个特征，剩余两列分别是 ID 和 Diagnostic。

图 1 是样本数据分布情况，每个样本的标签为乳腺良性肿瘤与恶性肿瘤，共有 357 个良性肿瘤，212 个恶性肿瘤。

图 2 是特征的小提琴图。小提琴图是数据可视化中常用的一种图表类型。它可以用来展示数据的分布情况，特别是在比较不同组数据分布时非常有用。小提琴图主要通过展示数据的四个关键指标：最小值、最大值、中位数和四分位数来反映数据的分布情况，从而帮助人们更好地理解和分析数据，支持数据驱动的决策和分析。图中横坐标是特征的名称，纵坐标是标准化后特征的数据情况，其中黄色部分代表的是恶性数据，绿色部分代表的是良性数据，在图 2 中可以大致看出每个特征的良恶性之间的关系，例如特征 fractal_dimension_worst 和 symmetry_worst 的黄色部分和绿色部分分布很相似，可大致判断出该特征的良恶性数据代表性不强，准确的情况还需要通过检验等方法来判断。

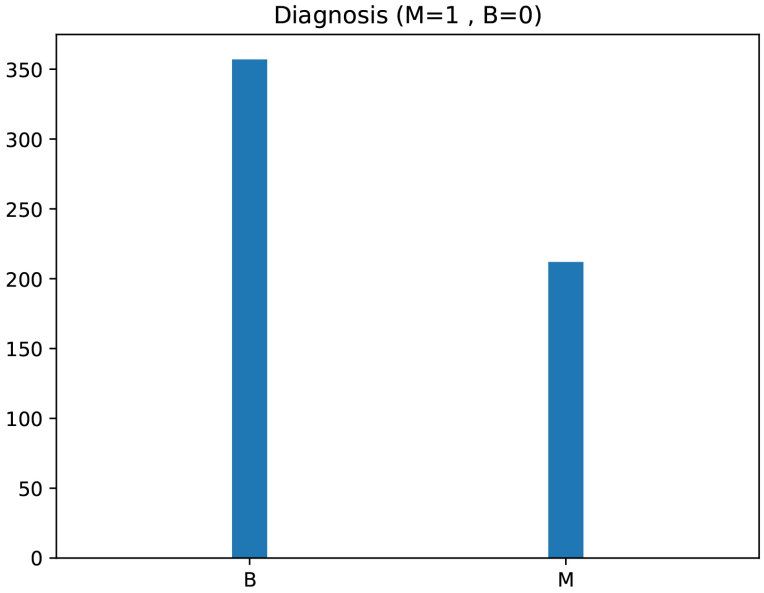

Figure 1. Data distribution histogram, “B” represents benign tumors, and “M” represents malignant tumors
图 1. 数据分布柱状图，B 代表良性肿瘤，M 代表恶性肿瘤

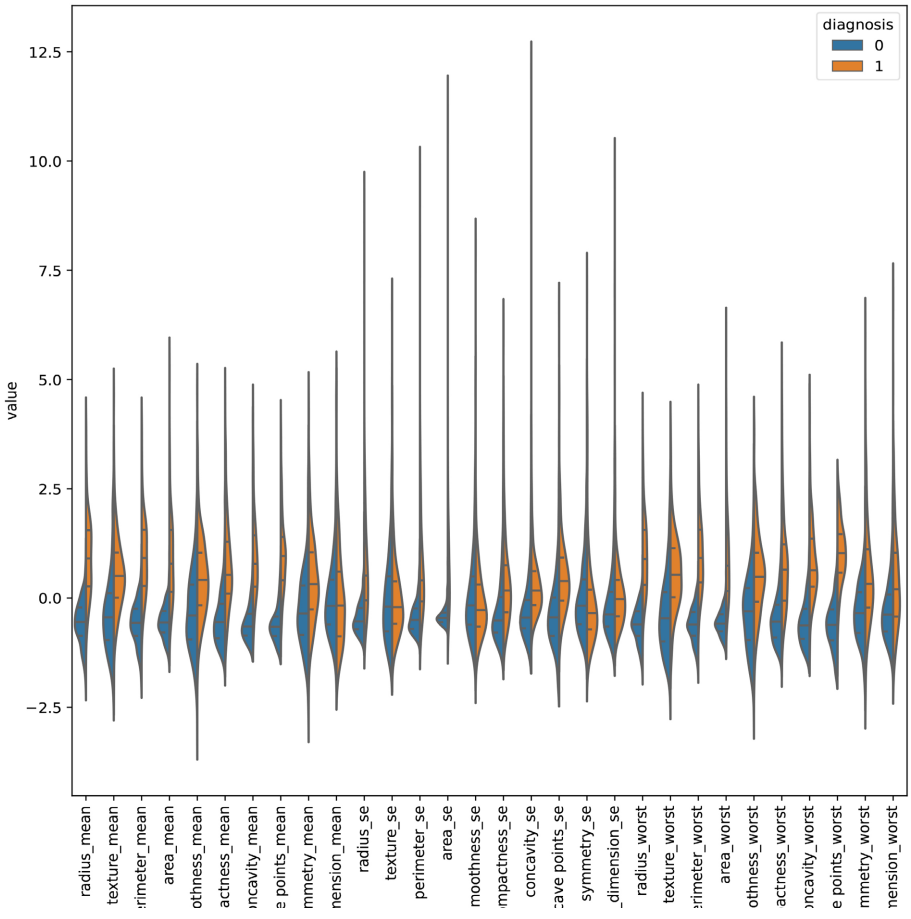

Figure 2. A violin plot illustrating the distribution of thirty features for benign and malignant tumors
图 2. 小提琴图，显示良性和恶性肿瘤三十个特征的分布情况

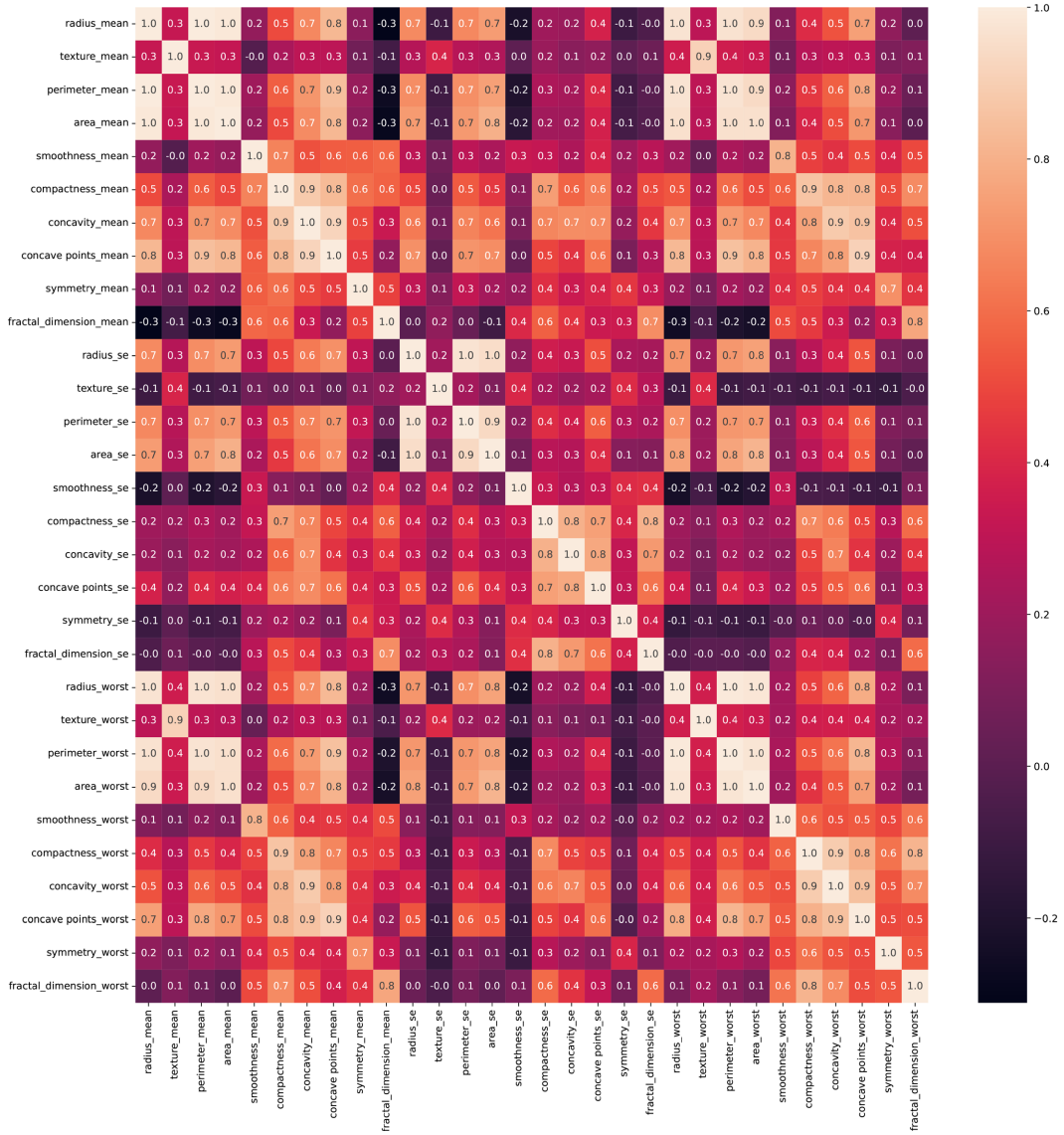

Figure 3. A heatmap that displays the correlation between features
图 3. 热力图，可显示出特征之间的相关性情况

图 3 是 30 个特征的热力图，热力图是数据可视化中一种常用的图表类型，它可以用来展示数据的分布情况，热力图可以通过不同的颜色呈现数据的相关程度。在该图中，颜色越深说明特征之间相关性越差，颜色越浅说明特征之间相关性越强。

3.4. 实证分析

lasso 回归会使得某些系数直接为 0，即完全忽略掉一些系数，可看作是一种自动化的特征选择。lasso 回归也有一个正则化参数 alpha，可以控制系数趋向于 0 的强度，下图 4 中展示的是不同的变量随着 alpha 惩罚后，其系数的变化，我们要保留的就是系数不为 0 的变量，alpha 值不断增大系数才变为 0 的变量在模型中越重要。可以设置更大的 alpha 值，就会看到更多的系数被压缩为 0 了。将数据按照 7:3 分为训练集和测试集，对训练集进行随机森林模型的拟合，再将测试集数据输入到训练好的模型中，对数据进行分类，并使用五折交叉验证得到了综合预测表现，通过准确性、召回率、F1 值、AUC 值等方面对算法

进行评估，得到相对较好的结果。

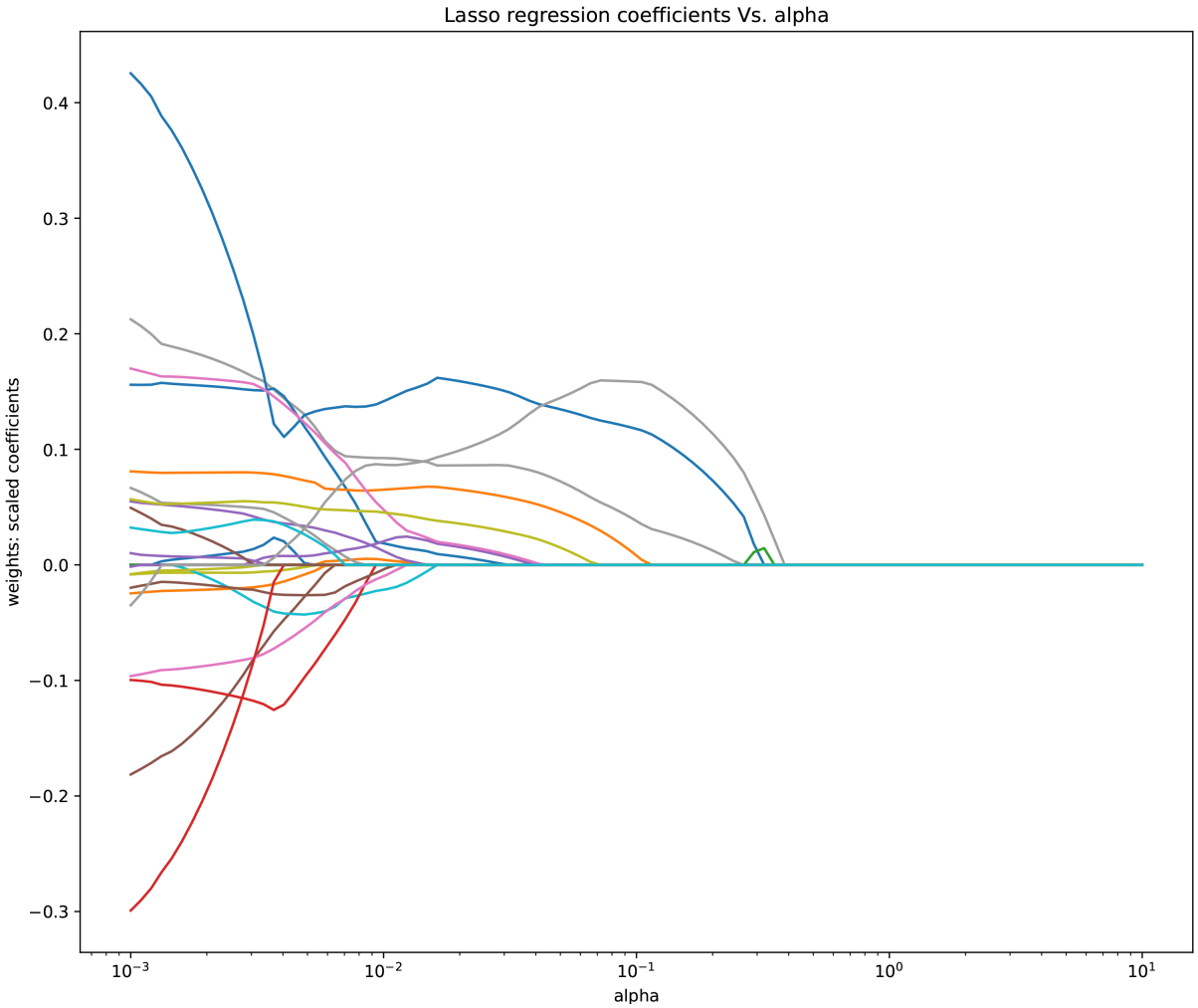

Figure 4. The coefficient changes with different alpha values in the Lasso algorithm
图 4. 在 lasso 算法中不同 alpha 值的系数变化

首先取 alpha = 10^{-1}，得到的特征为 area_mean, concave points_mean, fractal_dimension_mean, texture_se，使用随机森林算法对训练集进行拟合，随机森林是一种集成学习方法，由多个决策树组成。对于每个决策树，通过使用 Lasso 选中的特征，对数据进行划分和分类。最终，每个树的分类结果会被综合起来，以获得最终的分类结果。这种方法可以结合 Lasso 的特征选择优势和随机森林的集成优势，提高分类性能和模型的稳健性。再将测试集数据输入到训练好的模型中，测试集结果如下图 5 所示。

图 5 展示了参数 alpha = 10^{-1} 时的混淆矩阵和 ROC 曲线[11]，此时对测试集进行分类得到的准确率为 90.64%，召回率为 88.88% [12]，F1 分数为 87.50%，AUC 值为 0.98。然后取 alpha = 10^{-2}，使用随机森林算法进行分类，结果如图 6 所示。此时模型的准确率为 95.32%，召回率为 92.06%，F1 分数为 93.55%。

结果表明，所提出的模型在测试集上取得了显著的性能提升。具体而言，我们的模型达到了 95.32% 的准确率，说明模型能够高度准确地预测样本的分类标签。此外，模型召回率为 92.06%，表明模型对正类样本能够有较好的识别能力。我们进一步综合准确率和召回率计算出 F1 分数，结果为 93.55%，进一步验证了模型的优越性。

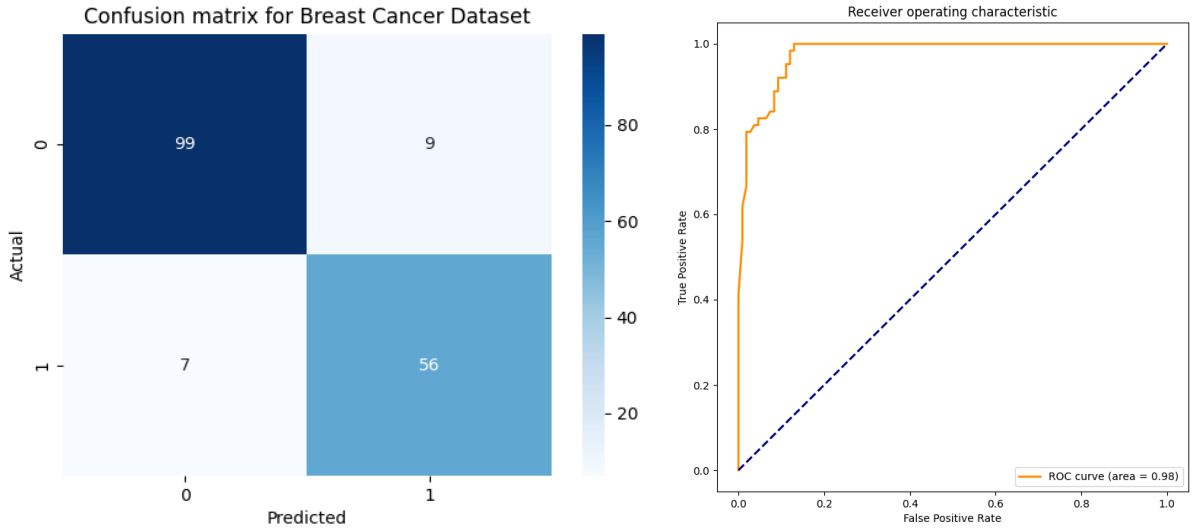

Figure 5. Confusion matrix and ROC curve in machine learning when $\text{alpha} = 10^{-1}$

图 5. 在 $\text{alpha} = 10^{-1}$ 时的混淆矩阵和 ROC 曲线

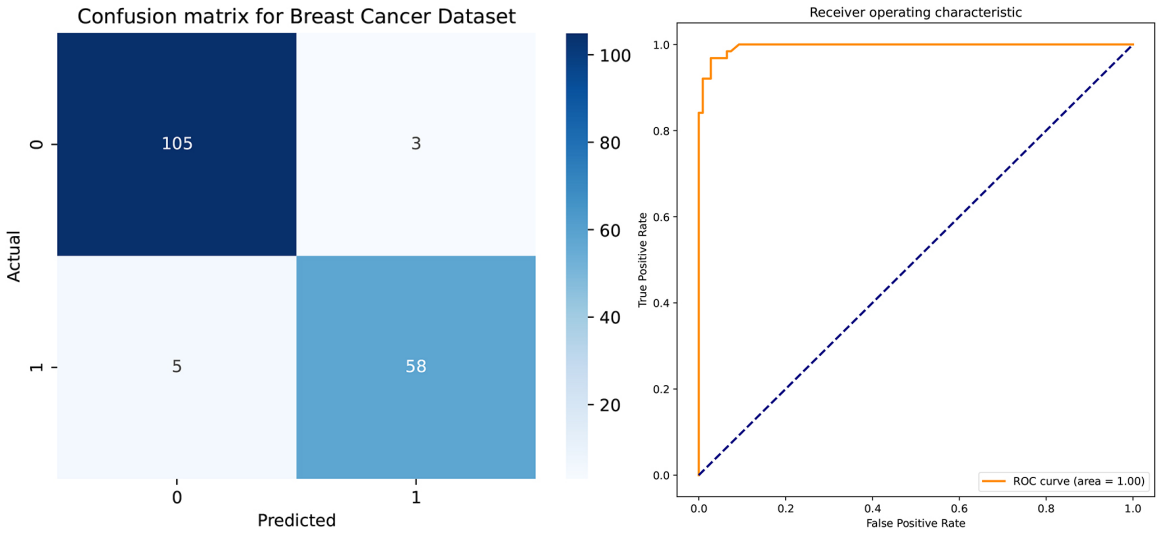

Figure 6. Confusion matrix and ROC curve in machine learning when $\text{alpha} = 10^{-2}$

图 6. 在 $\text{alpha} = 10^{-2}$ 时的混淆矩阵和 ROC 曲线

随机森林特征系数[13]提供了评估特征重要性的指标，可以帮助了解模型中各个特征的贡献程度和重要性排序。特征系数表示了在随机森林中各个特征对于最终模型预测的重要程度，可以用来评估每个特征对于模型的影响力，通常特征系数越高，表示该特征在预测中的重要性越高。它衡量了特征对于模型的预测性能的贡献程度。在本研究中得到的特征系数绝对值按照从大到小排序，打印出重要特征及其权重如表 2 所示。

特征系数可以通过不同的度量方式来计算，比如基尼系数、平均不纯度减少等。特征系数的计算方式可以根据具体的随机森林实现而有所不同，但一般情况下，特征系数越高，则该特征对于模型的预测贡献越大。它可以帮助我们了解随机森林中各个特征的重要性排序，进而进行特征选择、特征工程或可视化分析等相关任务。表 2 中 coefficients 是特征系数绝对值，是用来度量特征重要性的。因此，通过输

出这些系数绝对值并从大到小排序，我们可以得知哪些特征是最为重要的，即它们在乳腺癌预测中发挥更大的作用。这对我们指导实际应用具有很大的参考价值。

Table 2. The final selected feature ranking
表 2. 最终选中的特征排序情况

features	coefficients
radius_worst	0.141582
concave points_mean	0.092434
concave points_worst	0.08667
texture_worst	0.065088
concavity_worst	0.047893
symmetry_worst	0.045319
fractal_dimension_mean	0.021503
smoothness_worst	0.020413
radius_se	0.018654
smoothness_se	0.012053
concavity_se	0.009797
texture_mean	0.004359
compactness_se	0.003984

4. 结论与讨论

肿瘤是威胁人类生命的重大疾病，对于女性而言，乳腺癌是死亡率最高的癌症之一。在早期阶段，乳腺癌能够被治疗，甚至被治愈[14]。但是当到了晚期阶段，乳腺癌就变得非常难以治疗了，它会直接造成人体的死亡。乳腺癌的治疗最关键的一点，就是要对其进行早期诊断并进行治疗，及早的发现肿瘤，这对乳腺癌的临床治疗有很大的影响。近年来，随着诸如机器学习等多种计算机技术的应用，乳腺癌的诊断与治疗取得了长足的进步。医学科研工作者一直致力于通过大数据分析和机器学习等方法，帮助临床医师从大数据中挖掘出乳腺癌的特征，从而实现对乳腺癌的快速诊断和早期治疗。本文将为乳腺癌病人的治疗提供新的思路和方法，为其治疗提供理论依据和实践依据。另一方面，癌症数据特别是遗传学数据具有多维性和高维性，这加大了癌症数据分析的难度。在对高维特征数据进行分析时，随着特征维数的增加，计算复杂度也随之增加，给计算机带来巨大的负担。但是，在高维数据中，存在着许多冗余的特征和噪声，这些多余的特征对分类准确率没有明显的影响，反而增加了算法的工作负担，因此，需要对高维的数据进行降维，并从中筛选出有效的特征。

本文针对乳腺癌疾病从两方面来进行研究，首先提出了 lasso 算法，对数据进行特征筛选与提取，随后对训练集进行随机森林模型的拟合，对数据进行分类，得到的分类结果相对较好，并使用交叉验证得到了不同数据集拆分下的综合预测表现，通过准确性、召回率、F1 值、AUC 值等方面对算法进行评估，得到相对较好的结果。

基金项目

本工作得到了国家自然科学基金项目(批准号：61902192)、江苏省高层次创新创业项目(苏人办文，编号：[2019]20)。

参考文献

[1] 郑雅文. 基于特征选择和支持向量机的乳腺癌诊断研究[D]: [硕士学位论文]. 太原: 太原理工大学, 2019.

[2] 蔡玉琴, 张璟, 张帆. 乳腺癌影像学检查现状与研究进展[J]. 中国全科医学, 2009(13): 1228-1231.

[3] 孙哲, 黎庶, 徐惠绵. 数字化乳腺 X 线计算机辅助诊断系统临床应用价值的初步探讨[J]. 中华医学杂志, 2005, 85(24): 1692-1695.

[4] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[5] 黄登香, 卢春婷. Lasso 方法在基于行为决定因素的宫颈癌早期检测中的应用[J]. 应用数学进展, 2022, 11(2): 781-789.

[6] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least Angle Regression. *Annals of Statistics*, **32**, 407-499. https://doi.org/10.1214/009053604000000067

[7] 李欣海. 随机森林模型在分类与回归分析中的应用[J]. 应用昆虫学报, 2013, 50(4): 1190-1197.

[8] 姚旭, 王晓丹, 张玉玺, 等. 特征选择方法综述[J]. 控制与决策, 2012, 27(2): 161-166.

[9] 李丽, 李霞, 郭政, 等. 两种过滤特征基因选择算法的有效性研究[J]. 生命科学研究, 2003, 7(4): 369-373.

[10] Wolberg, W., Mangasarian, O., Street, N. and Street, W. (1995) Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B

[11] Fawcett, T. (2006) An Introduction to ROC Analysis. *Pattern Recognition Letters*, **27**, 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

[12] Powers, D.M. (2011) Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies*, **2**, 37-63.

[13] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[14] 李思琪. 乳腺癌数据处理及辅助诊断建模[D]: [硕士学位论文]. 哈尔滨: 哈尔滨理工大学, 2023.