单位代码 10635

学　　号 112021314201624

西南大學

专业学位硕士论文

基于集成学习的乳腺癌诊断预测模型研究

论文作者：管靖宇

指导教师：刘贤宁

专业学位类别：应用统计硕士

专业领域：应用统计

提交论文日期：　2023　年　6 月 11 日

论文答辩日期：　2023　年　5 月 27 日

学位授予单位：西南大学

中 国 • 重 庆

2023 年 6 月

目　录

基于集成学习的乳腺癌诊断预测模型研究

应用统计专业 硕士研究生 管靖宇

指导老师 刘贤宁 教授

摘要

随着社会经济的逐渐发展和物质生活水平的不断提高，我国已全面建成小康社会，人们对身体健康的重视程度日渐提升。而乳腺癌作为全球发病率较高的恶性肿瘤，对女性健康造成了极大的影响。传统的乳腺癌检测方法是基于“金标准”方法，该方法包括三项测试：临床检查、放射成像和病理检查。这种传统方法是基于回归过程来指示癌症的存在，而新的机器学习技术和算法是基于模型设计的。随着科学技术的更新迭代和统计学科的不断发展，对于乳腺癌的诊断已经不是停留在传统的数据表面之上，更是需要去挖掘数据背后隐藏的信息，并从数据中发现更多的统计学规律并为医生的诊断做出具有参考价值的辅助。

本文针对乳腺癌患者的肿瘤诊断结果，在 UCI 数据库威斯康辛乳腺癌患者（诊断）数据集的基础上，建立了集成学习模型，最大程度地预测乳腺癌患者的诊断结果。首先通过探索性数据分析，对数据集中各变量的分布情况与对因变量的影响程度进行了描述性分析，最终剔除了样本中存在的离群值并认为特征细胞核平均面积、细胞核平均凹点数、细胞核平均凹度和细胞核平均周长对因变量的影响较大；而特征细胞核平均分形维数、细胞核平均对称性以及癌细胞核平均平滑度对因变量影响程度较小。在特征选择上，分别使用过滤式算法中的 mRMR 算法、封装式算法中的 ReliefF 算法以及嵌入式算法中的 Lasso 算法对数据集中 30 个特征进行特征选择。三种方法分别选择了权重靠前的 8、9、8 个变量，且三种方法选择的变量重复程度较低，认为方法之间具有对比价值。

在变量确定的基础上，使用集成学习中的随机森林、Adaboost、XGBoost 模型结合 mRMR、ReliefF 以及 Lasso 三种变量选择方法。通过训练数据集建立了乳腺癌诊断预测模型，并使用建立的模型在测试数据集上进行预测。最终结果认为，在模型内部之间，随机森林模型之中结合 ReliefF 特征选择算法效果最好，其准确率达到了 0.9035，；Adaboost 模型结合 Lasso 变量选择算法效果最好，其准确率达到了 0.9298；XGBoost 模型结合 ReliefF 变量选择算法效果最好，其准

确率达到了 0.9473。

在对比不同模型内部的预测效果之后，对于三种模型结合最优特征之间的比较。得到结论：如果医务人员和研究人员更加注重诊断的准确率，应使用 ReliefF-XGBoost 模型来获得更好的效果；如果需求是尽可能发现恶性肿瘤患者，那么使用 ReliefF-RandomForest 模型是更好的选择。最后综合对比三个模型的 AUC 值，认为综合水平 ReliefF-XGBoost 模型最好，其 AUC 值分别为 0.964,0.928 和 0.928。

通过本文的实证分析。不仅为乳腺癌相关医学研究人员作出诊断提供了数据支持，还扩充了医疗诊断研究的方法使用。并且在有限的资源下丰富了生命科学与集成学习结合领域的研究。

关键词：乳腺癌预测；变量选择；集成学习；数据挖掘；

Research on Diagnosis and Prediction Model of Breast Cancer Based on Ensemble Learning

Jingyu Guan

Directed by Professor Prof Xianning Liu

Abstract

With the gradual development of social economy and the continuous improvement of the material standard of living, our country has built a well-off society in an all-round way, people pay more and more attention to health. As a malignant tumor with a high incidence in the world, breast cancer has a great impact on women's health. Traditional breast cancer detection methods are based on the "gold standard" method, which consists of three tests: clinical examination, radiographic imaging and pathological examination. While this traditional approach is based on regression processes to indicate the presence of cancer, new machine learning techniques and algorithms are designed based on models. With the update and iteration of science and technology and the continuous development of statistical disciplines, the diagnosis of breast cancer is no longer on the surface of traditional data, but more importantly, it is necessary to dig the hidden information behind the data, find more statistical rules from the data and provide valuable references for doctors' diagnosis.

In this thesis, based on the UCI database Wisconsin Breast Cancer Patients (Diagnosis) data set, an ensemble learning model was established to predict the diagnosis results of breast cancer patients to the greatest extent. Firstly, through exploratory data analysis, the distribution of variables in the data set and the degree of influence on dependent variables were described. Finally, the outliers in the samples were eliminated and it was considered that the average area of characteristic nuclei, the average concave number of nuclei, the average concave degree of nuclei and the average perimeter of nuclei had great influence on dependent variables. The mean fractal dimension of characteristic nuclei, mean symmetry of nuclei and mean smoothness of cancer cells had little effect on the dependent variables. In feature selection, mRMR algorithm in filter algorithm, ReliefF algorithm in package algorithm

and Lasso algorithm in embedded algorithm are used respectively to select 30 features in the data set. The three methods selected 8, 9 and 8 variables with the top weight respectively, and the degree of repetition of the variables selected by the three methods was low, so it was considered that the methods had comparative value.

On the basis of variable determination, random forest, Adaboost and XGBoost models in integrated learning are used in combination with three variable selection methods, mRMR, ReliefF and Lasso. A predictive model of breast cancer diagnosis was established through training data set, and the model was used to predict on test data set. The final results show that, among the internal models, the random forest model combined with ReliefF feature selection algorithm has the best effect, and its accuracy rate reaches 0.9035. The Adaboost model combined with Lasso variable selection algorithm has the best effect, with an accuracy of 0.9298. The XGBoost model combined with ReliefF variable selection algorithm has the best effect, with an accuracy of 0.9473.

After comparing the prediction effect of different models, the comparison between the three models combined with the optimal features is made. Conclusions: If medical personnel and researchers pay more attention to the accuracy of diagnosis, Relief-XGboost model should be used to obtain better results; If the need is to find patients with malignant tumors as much as possible, then using the Relief-RandomForest model is a better choice. Finally, the AUC values of the three models were compared comprehensively, and the Relief-XGboost model with comprehensive level was considered to be the best, whose AUC values were 0.964,0.928 and 0.928, respectively.

Through the empirical analysis of this thesis. It not only provides data support for breast cancer related medical researchers to make diagnosis, but also expands the use of methods in medical diagnosis research. Moreover, it enriches the research on the combination of life science and integrated learning with limited resources.

Keywords: Breast cancer prediction; Variable selection; Integrated learning; Data mining;

1 绪论

1.1 研究背景

随着社会经济的逐渐发展和物质生活水平的不断提高，我国已全面建成小康社会，人们对身体健康的重视程度日渐提升。而乳腺癌作为全球发病率较高的恶性肿瘤，对女性健康造成了极大的影响。在癌症的临床诊断中，医生往往通过患者的各项生理指标和影像数据结合经验并做出诊断。但是随着科学技术的更新迭代和统计学科的不断发展，对于癌症的诊断已经不是停留在传统的数据表面之上，更是需要去挖掘数据背后隐藏的信息，并从数据中发现更多的统计学规律并为医生的诊断做出具有参考价值的辅助。

而作为癌症之一的乳腺癌是女性最常见的恶性肿瘤之一。乳腺癌是乳腺导管上皮细胞在各种内外致癌因素的作用下，细胞失去正常特性而异常增生，以致超过自我修复的限度而发生癌变的疾病。从 20 世纪 90 年代以来,在世界多个国家乳腺癌一直是严重威胁女性健康的恶性肿瘤之一,且发病率逐年上升[31]。令人惊奇的是在经济发达的欧美国家中,随着发病率的增加,但死亡率却相反的逐年减少。但近十年来我国的乳腺癌患者的发病率和死亡率都是不断上升的,并且这个上升的增速已经是高于世界的平均增加速度,尤其以京、津、沪等一线城市更为明显,乳腺癌已严重威胁到了我国女性的健康。根据世界卫生组织国际癌症研究机构（IARC）发布的全球最新癌症报告，2020 年，全球有 1000 万人因癌症去世，这个数字在中国是 300 万，几乎相当于一个市的人口数量。而乳腺癌便是其中不可忽视的一种癌症。2020 年，全球乳腺癌新增 226 万例，首次超过了肺癌，成为了新发病例第一的癌症；其中单是女性群体 ，乳腺癌的新发比例就达到了惊人的 24.5%，其占比远远高于排名第二的直结肠癌 9.4%。此外，单是乳腺癌的死亡率就超过了 30%。最近几年，我国乳腺癌的发病率正在持续增加，2018 年一项关于乳腺癌的研究报告中提到，我国乳腺癌的新发病例达到了 38 万人，同时死亡人数占比达到了 27%左右。乳腺癌的存在，已经成为了我国甚至全世界健康和社会的巨大负担[29]。

目前临床研究的重点和难点是如何针对乳腺癌进行早期的诊断和早期治疗，从而提高患者的生存期和生活质量。随着医学的不断发展，目前临床辅助诊断技

术在不断地发展，其中针对乳腺癌的诊断主要包括三大技术类型：医学影像学诊断、分子生物学诊断以及病理活检等[30]。

因为科技水平的提高，产生并且被收集到的数据总量正在快速膨胀扩大。对于海量的数据，如何利用好这些数据是解决问题的关键。而集成学习正好可以充分利用这些数据，集成学习与大数据分析的结合使其产生了巨大的价值。许多公司也进行了相关方面的研究：生物大数据公司 Deep Genomics 就已经在预测基因组上的变化会对人体的疾病产生怎样的影响的研究中加入了集成学习模型的使用。微软公司也把在糖尿病的管理方案设计中使用了集成学习技术，利用技术进行实时监测，同时收集病人实时数据并上传，利用集成学习模型预测分析患者的血糖数据并推测其影响因素，最终得到个性化后的解决方案。

在生物医学信息爆发的背景下，医学大数据分析正在成为一项越来越重要的工作。如何利用已知的各项数据，去得到尽可能准确的结果，这正是当前面临的巨大挑战。而集成学习技术，无疑在其中扮演了重要角色，通过这项技术，可以发现医生发现不了的规律，可能会改变未来疾病诊断的标准，彻底改变医生、患者乃至整个生态。

1.2 文献综述

近年以来，针对乳腺癌的预测或者诊断，国内外众多学者都进行了相关的研究，这些研究为现代医学的发展作出了巨大的贡献。现对于乳腺癌的研究主要研究内容整理归纳如下：

1.2.1 乳腺癌预测现状

传统的癌症检测方法是基于“金标准”方法，该方法包括三项测试：临床检查、放射成像和病理检查[45]。这种传统方法是基于回归过程来指示癌症的存在，而新的机器学习技术和算法是基于模型设计的。该模型是为预测看不见的数据而设计的，并在其训练和测试阶段提供了良好的预期结果[32]。机器学习过程基于三种主要策略，包括预处理、特征选择或提取和分类。特征提取是机器学习过程的主要部分，实际上有助于癌症的诊断和预后，这个过程可以详细阐述癌症在良性和恶性肿瘤中的作用。

在国外研究中，Youness Khourdifi[33]等(2018)对于癌症的预测，作者比较使用

了随机森林、朴素贝叶斯、支持向量机和 K 最近邻等非线性算法来预测癌症。作者结合生物信息学和医学分类技术在选择最佳分类器的基础上对数据挖掘算法进行比较，最终认为选择支持向量机（SVM）比其他算法更适合，其准确率为97.9%；Sara AlGhunaim[34]等（2019）考虑了乳腺癌基因表达（GE）和 DNA 甲基化（DM）两种数据，选择了支持向量机（SVM）、决策树和随机森林三种不同的分类算法创建九个有助于预测癌症的模型，最终认为 Spark 环境中的缩放 SVM 分类器优于其他分类器，因为它在 GE 数据集中实现了最高的精度和最低的错误率。Bichen Zheng[35]（2013）开发了一种 K-means 和支持向量机（K-SVM）算法的混合算法并利用 K-means 算法分别识别良恶性肿瘤的隐藏模式，最终在乳腺癌相关数据集上认为该方法将准确性提高到 97.38%。Habib Dhahri[36]等（2019）使用主成分分析的方法将乳腺癌数据集进行了降维，在此基础上使用 KNN 算法和朴素贝叶斯算法对乳腺癌患者诊断情况进行预测，最终认为使用 KNN 算法得到的模型能够达到最佳的效果。Hiba Asri[37]等（2016）使用 C4.5 决策树对癌症数据集进行了分析研究，并利用准确性、精密度、敏感性和特异性方面等评价指标对预测模型进行了评估。Moloud Abdar[38]等（2018）使用堆叠和投票（Voting）作为分类器组合的嵌套集成方法检测乳腺良性肿瘤和恶性肿瘤，最终认为使用所提出的两层嵌套集成模型优于单个分类器和大多数分类器。

国内对于乳腺癌诊断研究相较于国外有一定的滞后，研究的学者和主要内容如下：李勇[1]等（2020）提出了基于 C-Adaboost 模型的集成学习算法，对乳腺癌疾病进行预测，并认为此模型相较于传统机器学习模型的性能提高了 19.5%。原瑞霞[7]（2018）利用泊松回归模型和负二项回归模型对乳腺癌发病率和死亡率进行分析，并使用 ARIMA 模型预测中国未来十年的乳腺癌发病率将会持续上升，但死亡率会有所下降。赖胜圣[6]等（2019）构建基于序列前向选择算法（SFS）与支持向量机算法（SVM）分类器融合的乳腺癌预测模型，最终得到评价指标为 AUC 为 98.39%，ACC 为 97.35%，相较于传统支持向量机方法有所提高。武莉茹[2]（2020）提出一种 REL1_FW 特征选择算法对乳腺癌数据进行特征筛选并结合五种常用的分类算法进行性能比较，认为 SVM 算法的准确率是最高的。刁继尧[4]（2019）基于大数据 spark 平台，采用 Cox 回归、支持向量机 SVM 算法对乳腺癌患者数据进行了预测分析，最终得到支持向量机模型的预测精度达到了 74.6%。李星睿[3]（2020）提出基于权重和密度的聚类算法，通过计算每个特征对簇内距离和簇间距离的贡献率对 LuminalB 型乳腺癌进行了分形并进行了基因分

析和病理分析。齐惠颖[5]等（2019）通过对 TCGA 数据库中乳腺癌的基因表达、拷贝数变异、DNA 甲基化和蛋白质表达 4 种组学数据的融合，使用随机森林算法建立预测模型在测试数据集上该模型对乳腺癌分类预测的精确率为 97.22%，召回率为 98.13%。通过 AUC 值对比不同类型组合组学数据的预测性能，融合多组学数据的 AUC 值为 0.8393，性能最好。王玲玲[8]（2021）采用单因素 COX 风险比例回归模型探讨影响乳腺癌特异性生存的独立因子，并通过 Kaplan-Meier 方法评估了乳腺癌特异性生存机器相关因素，得到最终的 AUC 值为 0.829。Tseng ChinDar[56]等研究了乳腺癌患者放射治疗(RT)后常见的并发症，使用最小绝对收缩和选择算子(LASSO)和五种分类算法来提高预测能力，最终认为使用 LASSO 和支持向量机(SVM)来评估并发症的风险，以提高乳腺癌患者放疗后并发症的预测能力，并降低并发症风险评估的成本。Mahesh Vijayalakshmi[58]提出了一种以乳腺癌症生物标记物为属性的模式识别方法和集成分类方法，以准确检测癌症的存在。通过融合朴素贝叶斯、径向基函数神经网络和基于多数投票规则的线性判别分析分类器的决策，对乳腺癌 Coimbra 数据集的样本进行了评估。实验结果表明，与单个分类器相比，融合分类器的系统性能得到了提高。Naji Mohammed[59]等将五种机器学习算法：支持向量机、随机森林、逻辑回归、决策树 C4.5 和 K-Nearest Neighbors 应用于乳腺癌诊断数据集，最终认为支持向量机优于所有其他分类器，并获得了最高的准确率。

1.2.2 集成学习研究现状

集成方法论的思想是通过集成多个模型来建立预测模型，集成方法可以用于提高模型的预测性能[39]。集成学习就是在此基础上将数据的特征提取、数学模型的建立和数据挖掘使用一个特定框架来连通。首先，集成学习特征算法通过多种变换将特征提取出来。再在这些特征的基础上，使用指定的集成学习算法来得到弱预测结果[40]。最后，集成学习融合了来自上述结果的信息性知识，以自适应的方式通过投票方案实现知识发现和更好的预测性能来自统计学和人工智能等多个学科的研究人员提供了大量对于集成学习方法的研究。主要包括以下研究成果。

近年来，外国学者在集成学习领域提出了大量的研究成果。Merler[41]等（2007）开发了 P-AdaBoost 算法，P-AdaBoost 不是以顺序的方式更新与实例相关的“权重”，而是分两个阶段工作。在第一阶段，AdaBoost 算法以其顺序、标准的方式运行有限的步骤。在第二阶段中，使用从第一阶段估计的权重并行训练分类器。

P-AdaBoost 产生了标准 AdaBooster 模型的近似值，该模型可以在计算节点网络上轻松有效地分布。Zhang[42]（2008）提出了一种新的通过重新采样版本的 Adaboost 来增强。在局部 Boosting 算法中，为每个训练实例计算局部误差，然后使用该局部误差来更新该实例被用于下一迭代的训练集的概率。Diao[43]等（2014）发展了一种特征选择的思想，通过在将集合预测转换为训练样本后将分类器视为特征，来促进基本模型的选择。Yu[44]等（2017）提出了一种混合增量集成学习（HIEL）方法，该方法同时考虑特征空间和样本空间，用于分类器选择，以处理噪声数据集。TianQ[49]（2016）基于 GBDT 算法提出了 XGBoost 算法，在 GBDT 的基础上大幅度提高了模型的识别精度与运行速度。

国内学者也对于集成学习方法提出了非常丰富的成果。曹莹[9]等（2013）对 AdaBoost 训练误差与泛化误差进行分析，解释算法能够提高学习精度的原因并以此提出了 Adaboost 算法未来发展的方向。常甜甜[10]（2010）针对多源数据分类问题,提出分组特征多核 SVM。该方法将不同源的数据进行分组,每组特征分别采用不同的核函数,将这些核函数的凸组合作为新的核函数,并将基于该新核的 SVM 问题转化为半定规划问题来求解.实验结果表明,该方法可以有效地提高分类器的检测性能。付忠良[16]（2008）在基于大数定理对弱学习定理进行解释与证明基础之上,对 AdaBoost 的有效性进行了分析。指出 AdaBoost 采取的样本权值调整策略其目的是确保正确分类样本分布的均匀性。 陈启伟[13]等（2017）提出一种基于 Ext-GBDT 集成的类别不平衡信用评分模型，在 UCI 德国信用数据集上,以 AUC 和代价敏感错误率作为评价指标得到很好的效果。钟熙[11]等（2019）提出了一种基于 kmean++聚类方法来增加普通朴素贝叶斯分类器的集成差异性,从而提升了朴素贝叶斯的泛化性能。樊鹏[14]（2016）提出了一种基于优化的 xgboost-LMT 模型并使用 ACROA 算法对 xgboost 框架进行参数寻优,并将基于反正切 Lasso 惩罚函数的 LMT 算法引入 xgboost 框架,作为框架的基分类器(Based Learner)。莫赞[12]等（2018）提出一种新的针对二类不平衡数据集的分类方法：对抗生成网络-自适应增强-决策树(GAN-AdaBoost-DT)算法。在信用卡诈骗数据集上的 AUC 值提高了 5.9%。孙茂伟[15]等（2016）采用高斯过程回归算法建立集成学习模型的基学习器,并在 Bagging 算法对训练样本重采样生成基学习器训练子集的基础上,采用基于正则化互信息的特征排序指标进行基学习器的输入特征抽取,实现有监督的特征扰动,从而改善学习器的差异度。Ravi Shankar[57]等学者应用一种特殊的选项树和聚类的混合分层策略来创建大多数癌症风险评估方案。使

用分组、聚类和估计等统计工具来发现癌症患者的能力倾向。

1.2.3 文献述评

通过对关于乳腺癌预测与诊断研究的了解，国内外相关学者主要是通过对乳腺癌细胞的病理数据或者相关图像进行研究的，说明了使用相关乳腺癌数据建立集成学习模型并进行预测的研究是具有意义的，这也是大部分研究人员所重点关注的内容。本文即在此基础上，从众多算法或者模型中找到适合乳腺癌诊断预测的模型，提高在乳腺癌诊断预测上的准确程度，为医学研究人员和后续学者的研究提供具有价值的信息。

1.3 研究内容及意义

1.3.1 研究内容

本文主要研究内容是结合乳腺癌病理数据集建立合适的数学模型并进行测试评价。从介绍集成学习方法的理论开始，通过不同的特征选择算法对乳腺癌数据集进行特征选择后再使用不同的集成学习方法建立数学模型。并对建立的模型使用数据进行验证与评估，最终得到更加适合于乳腺癌疾病诊断与预测的特征选择算法和分类集成学习模型的组合。 本文主要的章节结构划分为以下六个章节：

第一章为绪论，主要分为研究背景、国内外研究现状、研究内容及意义三个部分。此章主要说明当前社会乳腺癌的严峻情况以及对乳腺癌疾病诊断预测研究的重要意义。并在结合当前国内外学者对其乳腺癌预测的研究下，说明本文的主要研究内容以及结构安排。

第二章为相关理论概述，本章系统阐述了本文所使用的主要理论以及文中使用到的 mRMR、ReliefF 以及 Lasso 三种特征选择算法和决策树、随机森林、Adaboost 等集成学习模型，并说明各种算法的具体数学推导过程，为后续算法和模型的使用提供了理论基础。

第三章为数据描述统计与数据清洗，本章主要针对收集到的乳腺癌细胞病理数据进行数据清洗，如剔除掉数据中的离群点并对缺失数据进行相关处理；并对清洗完成的数据中的各项指标进行描述统计并得出相关结论。

第四章为数据的特征选择，本章主要是在得到清洗完成的数据的基础上。使

用 mRMR、ReliefF 和 Lasso 算法对数据的进行特征选择，并统计各种算法最终得到的更小维度的相关指标，为后续的建立集成学习模型准备好数据基础。

第五章为基于集成学习的乳腺癌预测模型的建立。这一部分是针对上一章的通过不同特征选择算法筛选后的数据集，按比例划分训练集和测试集，再使用决策树、随机森林、Adaboost 三种集成学习建模方法在训练集上建立模型。并使用建立的多个模型在测试集上进行评估，评价主要通过使用混淆矩阵、准确率、ROC 曲线、AUC 值等方法进行量化。

第六章为结论与展望。本章主要是总结本文的研究，并针对研究结论提出研究中存在的问题和还未解决的缺陷。并在此基础上为后续的研究方向提出了改进的方向和建议。

1.3.2 研究意义

本文的选题意义在于：

为乳腺癌相关医学研究人员作出诊断提供了数据支持。在乳腺癌形式日渐严峻、医学技术更加发达的今天，利用合理的集成学习方法，通过对乳腺癌患者部分生理指标数据进行分析处理，从而得到对患者乳腺癌情况判断的结论，并基于提出的各项方案作出了对比，为医学研究人员选择合适的诊断方法提供了理论依据。

扩充了医疗诊断研究的方法使用。在传统的医疗诊断研究中，医生常用的数据分析软件（例如 SPSS）对临床文本数据或者医学图像数据进行简单分析。这些软件都是基于统计学方法，没有挖掘数据中隐藏的关系。集成学习方法的使用可以寻找具体数据之间的关系，对研究患者病情提供了一种可行方案。

丰富了生命科学与集成学习结合领域的研究。21 世纪是生命科学的新时代，统计学在生命科学领域的进步与发展贡献了巨大的技术支持，它不但具有关键的基础研究价值，还有光明的产业化前景。本选题不仅丰富了相关学科与统计学科之间的研究，同时也可以减少可能产生的决策失误，从而提高社会资源配置效率。

2 相关理论概述

2.1 集成学习理论

集成学习（Ensemble Learning）是使用一种或多种特定的策略集成多个基础模型，再使用群体决策的方法增加决策准确率的算法。集成学习的关键就是选择哪种学习器以及如何集成多个基学习器也就是集成策略的选择。集成学习也有多分类器系统（multi-classifier system）和基于委员会的学习(committee-based learning)等别称。集成学习方法通过个体学习器不同的生成方法主要分为两类，分别是相互强依赖个体学习器的方法和可同时生成的并行化方法；前者中的Boosting 最为常用，后者则是 Bagging 和随机森林（Random Forest）[17]。

Boosting 算法的思想是将一系列弱学习器组合提升为强学习器。这一系列算法的工作机制如下：在原始训练集的基础上训练得到一个基学习器，根据这个基学习器的性能对训练集中样本的分布作出调整，使得基学习器更加关注在之前训练中分类错误的样本，然后使用调整后的训练集样本得到下一个基学习器；重复上述步骤，直到基学习器达到给定的数量T，最终根据某种权值组合这T个基学习器。在分类问题中，在同一个训练集样本中，训练一个强学习器（精准的分类规则）要比训练基学习器（粗糙的分类规则）困难更多。所以 Boosting 算法选择从基学习器开始，通过不断地训练学习，训练出一系列基学习器，再通过加权组合合成一个强学习器。Boosting 方法通常使用改变训练数据概率分布也就是其训练集数据的权值分布，再对于不同的训练数据分布进行调用[18]。

一个泛化性能强的集成学习模型的特点是模型中的各个基学习器是最大程度相互独立的；虽然现实建模中不可能做到完全独立，但需要使得各个基学习器之间的差异性达到最大。Bagging 是直接基于自助采样法（Bootstrap Sampling）的并行式集成学习方法，最初由 Leo Breiman 于 1996 年提出。与 Boosting 方法不同的是，Boosting 算法注重于各个基学习器之间的依赖关系，而 Bagging 算法更强调各个基学习器之间没有这种依赖关系，且可以并行拟合。Bagging 是每个基学习器之间的并行计算最后综合预测，在分类任务中可以使用简单投票法，在回归任务中可以使用简单平均法。在分类任务中如果出现两个或者多个类别得到的投票结果相同，最容易得方法是从结果相同的类别中随机选择一个，当然也可以通过学习器投票的置信度来确定最终的选择。Bagging 算法的主要步骤如下：

（1）从原始样本集中抽取训练集。每轮从原始样本集中使用 Bootstraping 的方法抽取 n 个训练样本（在训练集中，有些样本可能被多次抽取到，而有些样本可能一次都没有被抽中）。共进行 k 轮抽取，得到 k 个相互之间独立的训练集。

（2）在每个训练集上训练得到一个模型，即在 k 个训练集上一共可以得到 k 个模型。根据具体问题的特征应该采用适合的分类以及回归方法，如决策树、感知器等。

（3）对于分类问题，将上步得到的 k 个模型采用投票的方式得到分类结果；对于回归问题，计算上述模型的均值作为最后的结果。

2.1.1 决策树

决策树（Decision Tree）是一种基础的回归与分类方法。它是一种描述对实例进行分类的树形结构，决策树由结点（Node）和有向边（Directed Edge）组成。结点中包括了两种类型的结点，分别是内部结点（Internal Node）和叶结点（Leaf Node）。内部结点表示的是一个特征或属性，叶结点表示的是一个类。

决策树的分类步骤是从根节点开始，对实例或样本的其中一个特征进行测试，再在测试结果的基础上将实例划分到其子结点；这里的各个子结点分别对应了选择特征的若干个取值。对所有实例或样本递归以上步骤进行划分，直到达到叶结点[18]。最终把实例或样本划分到某个叶结点的类中。在决定选取的特征的准则上，通常有三种准则：信息增益（Information Gain）、信息增益比（Information Gain Ratio）以及基尼指数。

（1）信息增益

介绍信息增益之前，先给出熵和条件熵的定义。熵（Entropy）在概率统计与信息论中是用来度量随机变量的不确定性。设 X 是一个取有限个值的离散随机变量，其概率分布为：

$$P(X=x_i)=p_i, i=1,2,\cdots,n$$

则随机变量 X 的熵定义为：

$$H(X)=-\sum_{i=1}^{n} p_i \log p_i \tag{2.1}$$

在式(2.1)中，若 $p_i=0$，则定义 $0\log 0=0$，式中的对数以 2 为底或以自然对数为底。

设有随机变量(X,Y)，其联合概率分布为：

$$P(X=x_i,Y=y_j)=p_{ij},i=1,2,\cdots,n;j=1,2,\cdots,m$$

则条件熵$H(Y|X)$表示在已知随机变量X的条件下随机变量Y的不确定性。定义为：

$$H(Y|X)=\sum_{i=1}^{n}p_iH(Y|X=x_i)$$

在此基础上，特征A对训练数据集D的信息增益$g(D,A)$定义为：

$$g(D,A)=H(D)-H(D|A)$$

熵表示的是特征对于数据集进行分类的不确定性，而条件熵表示的是给定条件下特征对数据集进行分类的不确定性，而它们的差即表示信息增益，即信息增益代表的是某个特征使得数据集分类后的不确定性减少的程度。因此，在决策树的构建中，信息增益是一个有效的特征选取准则。

（2）信息增益比

在使用信息增益对数据集进行划分的时候，有时会出现偏向选择取值较多的特征。为了解决这一问题，通常使用信息增益比作为准则。则特征A对训练数据集D的信息增益比$g_R(D,A)$定义为：

$$g_R=(D,A)=\frac{g(D,A)}{H_A(D)} \tag{2.2}$$

式(2.2)中，$H_A(D)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\log_2\frac{|D_i|}{|D|}$，$n$特征$A$取值的个数。

（3）基尼指数

基尼指数与熵相似，都表示了集合的不确定性。在分类问题中，假设有K个类，样本点属于第k类的概率为p_k，则概率分布的基尼指数定义为：

$$Gini(p)=\sum_{k=1}^{K}p_k(1-p_k)=1-\sum_{k=1}^{K}p_k^2$$

而对于给定的样本集合D，其基尼指数为：

$$Gini(D)=1-\sum_{K=1}^{K}\left(\frac{|C_k|}{|D|}\right)^2 \tag{2.3}$$

式(2.3)中的C_k是D中属于第k类的样本子集,基尼指数$Gini(D,A)$表示经$A=a$分割后集合D的不确定性，基尼指数越大，样本集合的不确定性也就越大。

在确定了特征选取准则之后，就可以构建一棵决策树了，使用不同的准则对应了不同的决策树生成算法，其中应用广泛的有 C4.5 和 ID3 生成算法。在 ID3 生成算法中是使用信息增益作为准则选取特征，并在此基础上递归地训练决策树模型。ID3 算法的具体步骤是从根结点开始计算每一个特则的信息增益，并选择其中信息增益值最大的特征作为此个结点，再在此特征取值的基础上构建更多子结点；之后便在子结点上重复以上步骤最终构建决策树；停止条件设置为所有特征已被建立结点或者达到设定的重复次数，最后得到一棵决策树。ID3 算法是将最大似然估计的思想融合到决策树的构建之中。而 C4.5 算法与 ID3 算法相似，只是将特征选取准则由信息增益替换为信息增益比，这可以有效解决决策树生成过程中可能出现的过拟合问题。

决策树的生成过程中很容易出现过拟合的问题，因为决策树是不断递归直到不能再继续分类下去为止，这样产生的树往往和训练数据完美贴合，但是对训练集之外的数据分类效果不理想。此时则需要对决策树进行剪枝操作。决策树的剪枝往往通过极小化决策树整体的损失函数或代价函数来实现。即在决策树复杂度和拟合程度中选择一个合适的点，使得损失函数或代价函数最小。这里的损失函数等价于正则化的极大似然估计。

决策树的生成包括了准则的选择，决策树的生成和决策树的剪枝。决策树在集成学习中通常被用来作为基学习器，在后续的各种集成学习模型中充当一个基础且重要的角色。

2.1.2 随机森林算法

随机森林（Random Forert）是 1995 年由贝尔实验室的 Tin Kam Ho 所提出的随机决策森林（Random Decision Forests）而来的。这个方法是结合 Breimans 的“Bootstrap aggregating”想法和 Ho 的“Random Subspace Method”以建造的决策树的集合。随机森林以上文中的决策树算法为基础，将决策树作为模型的基学习器并将各个基学习器以 Bagging 结构进行集成，并在其中决策树的训练生成过程中添加了随机属性选择的思想。随机森林算法具有模型精炼、实现容易以及计算量小等优点，因此被认为是最能代表集成学习技术水平的模型算法。随机森林的思想正是在 Bagging 思想的基础上增加了一些细微改动，但是完美解决了 Bagging 算法中基学习器的“多样性”只能使通过对初始训练集采样的缺陷，随机森林中基学习器的多样性不仅来自样本扰动，还来自属性扰动，这使得最终集成的泛化性

能可通过个体学习器之间差异度的增加而进一步提升[17]。随机森林算法的具体步骤是：

（1）从原始训练数据中随机选取 n 个数据作为训练数据输入，通常 n 远小于全体训练数据 N，这样就存在一部分“袋外数据”始终不被取到，它们可以直接用于测试误差（而无需单独的测试集或验证集）。

（2）准备好需要训练的数据之后，训练一棵决策树的方法是每一个结点从全体特征集 M 中选取 m 个特征进行构建，为了模型的有效性，选择的参数需要满足条件 m 远远小于 M 。

（3）在随机森林算法中，构建决策树使用的准则通常是基尼系数。并采用同一种分裂准则来构建决策树的各个结点，设置停止条件为对所有特征分类完毕或者达到设定的树的深度。（该步骤与构建单一决策树相同）

（4）递归进行第二步和第三步，每次循环都会生成一棵相应决策树，多棵决策树就形成了一个随机森林模型。

（5）得到模型之后，模型对待测集的样本进行输入，模型中多棵决策树同时进行了输入，使用取最多数的方法在投票结果中选择出最终的预测结果。

随机森林算法的优点是：两个随机性（随机采样，样本最优特征选择）的引入，使得随机森林抗噪声能力强，方差小泛化能力强，不易陷入过拟合；每颗树是独立的，故模型易于高度并行训练，且训练速度快，适合处理大数据；创建随机森林时，是对参数的无偏估计；训练后，可以给出各个特征对于输出结果的重要性排序，并且能够检测到特征间的相互影响；引入了不少随机性在模型里，故模型对部分特征缺失不敏感。

2.1.3 Adaboost 算法

上文中介绍了 Boosting 算法的思想是将一系列弱学习器组合使得其成为一个更精确的强学习器。其中最具代表性便是 Adaboost 模型。Boosting 算法的关键问题有两个，首先是训练集数据的在每一轮训练中的权重或分布怎么更新；其次是弱分类器集如何组合为一个强分类器的策略。在 Adaboost 算法中，设定的规则是将上一次训练中分类错误的样本权值提高，以提高对这些样本的关注程度，相应地减少那些分类正确的样本的权重以降低模型对其的关注程度。所以问题才能通过多个基学习器分而治之。对于第二个问题，Adaboost 算法采用的解决方法是加权多数表决，即提高训练中分类误差率小的弱分类器的权值，使得这些分类

器在最终表决时作用增大；同时减小训练中分类误差率大的弱分类器的权值，使其在最终表决中作用减小[18]。下面给出 Adaboost 算法具体步骤。

对于训练数据集$T=\{(x_1,y_1),(x_2,y_2),\cdots(x_N,y_N)\}$，$x_i\in\chi\subseteq\mathbb{R}^n, y_i\in\{-1,+1\}$。

（1）对训练数据集权值赋初值。

$$D_1=(\omega_{11},\cdots,\omega_{1i},\cdots,\omega_{1N}),\omega_{1i}=\frac{1}{N},i=1,2,\cdots N$$

（2）对$m=1,2,\cdots,M$

（a）在当前具有权值D_m的训练数据集进行训练，以获得弱分类器

$$G_m(x):\chi\to\{-1,+1\}$$

（b）对于弱分类器$G_m(x)$计算其在训练集上的分类误差率

$$e_m=\sum_{i=1}^{N}P(G_m(x_i)\neq y_i)=\sum_{i=1}^{N}\omega_{mi}(G_m(x_i)\neq y_i)$$

（c）计算$G_m(x)$的系数

$$\alpha_m=\frac{1}{2}\log\frac{1-e_m}{e_m}$$

（d）对训练中的权值分布进行更新

$$D_{m+1}=(\omega_{m+1,1},\cdots\omega_{m+1,i},\cdots\omega_{m+1,N})$$

$$\omega_{m+1,i}=\frac{\omega_{mi}}{Z_m}\exp(-\alpha_m y_i G_m(x_i)),i=1,2,\cdots N \tag{2.4}$$

式(2.4)中的Z_m为规范化因子，加入是为了它D_{m+1}成为一个概率分布，其定义为

$$Z_m=\sum_{i=1}^{N}\omega_{mi}\exp(-\alpha_m y_i G_m(x_i))$$

（3）将弱分类器线性组合

$$f(x)=\sum_{m=1}^{M}\alpha_m G_m(x)$$

得到最终分类器

$$G(x)=sign(f(x))=sign\left(\sum_{m=1}^{M}\alpha_m G_m(x)\right)$$

以上便是 Adaboost 算法的完整过程，Adaboost 具有泛化错误率低，易编码，可以应用在大部分分类器上，无需要参数的调整等优点。

2.1.4 XGBoost 算法

XGBoost（Xtreme Gradient Boosting）算法是由陈天奇[49]在 2016 年提出的一个端对端的梯度提升树系统。XGBoost 算法整体结构与 GBDT（Gradient Boosting Decision Tree）算法的整体结构一致，都是在训练出一棵树的基础上，再训练出下一棵树，预测它与真实值之间的差距，通过不断训练用来弥补差距的树，最终使用树的组合实现对真实分布的模拟。但 XGBoost 的独特之处在其损失函数中不仅使用一阶导数计算了梯度下降的方向，还进一步考虑了二阶导数表示了梯度下降的趋势。数学原理外 XGBoost 算法最大的改进是大幅提升了模型的计算速度，树的构建中最耗时的部分就是为确定最佳分裂点而进行的特征值排序，XGBoost 算法在训练前会先将特征进行排序，存储为 Block 结构，此后重复使用这些结构，从而减少计算量。XGBoost 算法非常善于捕捉复杂数据集之间的相互依赖关系，能从大规模数据集中获取有效模型。

XGBoost 将若干个决策树弱分类器模型组合成为一个强分类器，换言之就是将多个决策树的结果求和得到一个最终结果，即：

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{2.5}$$

式(2.5)中，$\hat{y}_i$ 表示的是第 i 个样本的预测值，K 表示树的棵树，$f_k(x_i)$ 表示的是一棵决策树模型的最终输出结果。在此基础上，XGBoost 依然是通过最小化损失函数来确定最终模型的，其损失函数的定义如下：

$$Loss(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2.6}$$

式(2.6)中 $l(y_i, \hat{y}_i)$ 表示的是样本 x_i 的训练误差，$\Omega(f_k)$ 表示的是第 k 棵树的正则惩罚项，其定义如下：

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \tag{2.7}$$

式(2.7)中 T 为叶子节点数，ω 为叶子节点得分数，γ, λ 为参数，

由于加入第 k 棵树之间，前 $(k-1)$ 棵树已经训练完成，前面 $(k-1)$ 棵树的训练误差和正则项都变为了常数项，假设其为 C,则此时损失函数为：

$$Loss(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i + f_t(x_i)) + \gamma T + \frac{1}{2}\lambda \|\omega\|^2 + C$$

当采用均方误差作为损失函数时，将上公式二阶泰勒展开，得到：

$$\begin{aligned} Loss^{t}(\theta) &= \sum_{i=1}^{n} l(y_i, \hat{y}_i + f_t(x_i)) + \gamma T + \frac{1}{2}\lambda \left\| \omega \right\|^2 + C \\ &\approx \sum_{i=1}^{n} \left[g_i f_i(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \end{aligned} \tag{2.8}$$

式(2.8)中：

$$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}), h_i = \frac{1}{2}\partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$$

此时目标函数可以写作：

$$L^{t}(\theta) = \sum_{j=1}^{T} \left[\left(\sum_{i \in I} g_i \right) \omega_j + \frac{1}{2} \left(\sum_{i \in I} h_i + \lambda \right) \omega_j^2 \right] + \gamma T \tag{2.9}$$

式(2.9)中，ω_j 表示第 j 个叶子节点的得分。对其求偏导可以得出最优的叶子节点分数 ω_j^* 和最小的损失函数 $Loss^*$，其计算方式如下：

$$\omega_j^* = -\frac{G_j}{H_j + \lambda} \tag{2.10}$$

$$Loss^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T$$

式(2.10)中，$G_j = \sum_{i \in I} g_i, H_j = \sum_{i \in I} h_i$ 。

2.2 特征选择算法

需要注意的是，在建立上述介绍的集成学习模型的时候，所参与的特征变量并不是越多越好。模型中包含的变量越多，则模型就越容易出现过拟合的情况，所以在实践建模前，需要结合数学模型在尽可能保留样本信息的情况下选择比较优秀的部分变量，用这一部分“精挑细选”后的变量，才可以最大程度地发挥模型的实际价值。以下便是本文中使用到的几种变量筛选方法：

2.2.1 mRMR 算法

mRMR 算法（Max-Relevance and Min-Redundancy）是使用广泛，性能优异的基于空间搜索的特征选择算法。其中，第一个 mR 表示具有最大相关性，也就是特征与类别间的相关性尽可能大，第二个 MR 表示具有最小冗余性，即特征与特征之间的相关性尽可能小。互信息在 mRMR 算法中作为度量准则来衡量其中的

相关性和冗余性，相较于过去只考虑特征之间冗余度的算法 mRMR 算法拥有更佳的性能[46]。

互信息（Mutual Information）是 mRMR 算法中的基础度量准则。它可以理解为一个随机变量对另一个随机变量的信息量大小的度量，亦或理解为一个随机变量因为已知另一个随机变量而降低的不确定性。给定两个随机变量 x 和随机变量 y，其各自的概率密度函数和联合密度函数为 $p(x), p(y), p(x,y)$。其互信息定义为：

$$I(x;y)=\iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy$$

由定义可知，互信息不能为负值，且互信息越大，则说明两个变量之间的关联度就越大。

对于某一目标类 c，计算每一特征与目标类的互信息 $I(x,c)$，直到找到含有 m 个特征的特征子集 S，使得这 m 个特征和类别 c 的相关性最大，也就互信息平均数的值最大。也就是：

$$\max D(S,c)\quad D=\frac{1}{|S|}\sum_{x_i\in S}I(x_i;c)$$

根据最大相关性选得了含有 m 个特征的平均互信息最大的集合 S 很可能具有很多的冗余，这些特征之间的依赖关系有可能是非常大的。如果两个特征呈高度依赖的关系时，那么在去掉其中一个特征之后，模型的判断能力并不会发生非常大的变化，一些学者研究了一些直接或间接的方法来减少特征之间的冗余性，然后选择冗余最小的特征。其定义如下：

$$\min R(S), R=\frac{1}{|S|^2}\sum_{x_i,x_j\in S}I(x_i,x_j)$$

综合上面两个约束条件的准则算法即成为最小冗余最大相关度算法（mRMR）。结合两个约束条件，得到 mRMR 算法的约束条件为：

$$mRMR=\max\left[\frac{1}{|S|}\sum_{x_i\in S}I(x_i,c)-\frac{1}{|S|^2}\sum_{x_i,x_j\in S}I(x_i,x_j)\right]$$

2.2.2 ReliefF 算法

ReliefF 算法最早由 Kononeill 提出，最初局限于两类数据的分类问题，后续扩展到可以解决多类别数据。ReliefF 算法是一种特征权重算法(Feature weighting algorithms)，即在每个特征或者类别相关程度的基础上赋予不同的权重，并剔除小于设定阈值的特征。Relief 算法使用特征对近距离个例的区别能力衡量特征与类别之间的相关程度[23]。

算法基于训练集 D 采用随机的方法抽取一个样本，并从类别相同的其他样本中选择另一个最近邻样本，再从类别不同的样本中找到另一个最近邻样本。其更新更新特征权重方法如下：如果是相同类别的样本在某一个特征的距离小于不同类别样本的距离，就说明此特征对于区别两类样本是有帮助的，应该对其权重进行增加；相反，如果相同类别之间的距离更大，就说明此特征并不能有效地区分多种类别之间的样本，相应就对该特征的权重进行降低。递归进行以上步骤，直到达到停止条件以获得每个特征的权重平均值。其中将样本之间的距离定义为：

$$diff(A,R_1,R_2)=\frac{\left|R_1[A]-R_2[A]\right|}{\max(A)-\min(A)} \tag{2.11}$$

式(2.11)中 $diff(A,R_1,R_2)$ 表示的样本 R_1 和 R_2 在特征 A 上的差。通过距离计算公式，在训练集中反复抽样，通过每次抽样得到的样本计算各个特征的距离，并以计算的距离值不断更新权值，直至达到规定的抽样次数。

权重的大小即表示每个特征分类性能的大小。分类性能越强的特征其权重值就会更大；相反，分类性能越弱的特征其权重就会越小。ReliefF 算法的运行时间因为随着样本的抽样次数和原始特征个数的增加线性增加，所以可以有效地提高其运行的速率。

2.2.3 Lasso 算法

LASSO（Least Absolute Shrinkage and Selection Operato）由 1996 年 Robert Tibshirani 首次提出[22]。该方法是一种压缩估计。它通过加入一个惩罚函数得到一个更加精炼的模型，使得它压缩一些系数，同时设定一些系数为零。通过这样的思想，Lasso 算法在变量选择中也有非常显著的作用。Lasso 在原来的标准线性回归的代价函数上了一个惩罚系数 λ 的 ω 向量的 L_1-范数作为惩罚性，即：

$$\mathrm{Cost}(\omega)=\sum_{i=1}^{N}(y_i-\omega^T x_i)^2+\lambda\|\omega\|_1$$

并求使得代价函数最小时ω的大小，即：

$$\omega=\arg\min_{\omega}\left(\sum_{i=1}^{N}(y_i-\omega^T x_i)^2+\lambda\|\omega\|_1\right)$$

需要注意的是，上述过程中添加的是向量的 L1-范数，其中存在着绝对值，这就使得不能得到处处都可导的代价函数，所以ω的解析解不可以使用求导的方法计算得到。通常需要使用坐标下降法[50]（coordinate descent）来求得最优解。其主要步骤如下：

（1）权重系数ω赋初值，实际操作中通过初始为零向量；

（2）对每一个权重系数进行遍历，每一次令其中某一个权重系数作为变量，除此之外的权重系数取上一次循环中的结果，计算出当前一个变量情况下整个模型最优的解。

当第k次迭代时，ω_m^k表示的是第k次迭代，第m个权重系数，更新权重的方法如下：

$$\begin{cases}\omega_1^k=\arg\min_{\omega_1}(\mathrm{Cost}(\omega_1,\omega_2^{k-1},\cdots,\omega_{m-1}^{k-1},\omega_m^{k-1}))\\ \omega_2^k=\arg\min_{\omega_2}(\mathrm{Cost}(\omega_1^k,\omega_2,\cdots,\omega_{m-1}^{k-1},\omega_m^{k-1}))\\ \vdots\\ \omega_m^k=\arg\min_{\omega_m}(\mathrm{Cost}(\omega_1^k,\omega_2^{k-1},\cdots,\omega_{m-1}^{k-1},\omega_m))\end{cases}$$

（3）第二步即为完整的一次迭代。并通过不断重复迭代，直到达到了设定的迭代次数或权重向量的各个系数变化很小时，结束所有迭代[24]。

3 数据的描述性统计

3.1 数据的来源及指标

3.1.1 数据来源

本文所使用的研究数据选自 UCI 机器学习库平台。UCI 数据库是加州大学欧文分校（University of California Irvine）创建的用于机器学习的数据库，收集了机器学习领域的相关通用标准测试数据集，这些数据集在许多论文和研究中被拿来验证机器学习算法性能。

文章选取的 UCI 数据库中包含的威斯康星州（诊断）乳腺癌数据集（Breast Cancer Wisconsin (Diagnostic) Data Set）。该数据集是由威斯康星大学麦迪逊临床科学中心普通外科系的威廉 H.沃尔伯格博士收集整理得到的。数据集共包含了共 569 例样本的乳腺癌诊断结果与 30 个相关特征。其中特征是根据乳房肿块的细针抽吸物 （FNA） 的数字化图像计算的。它们描述了图像中存在的细胞核的特征。具体的特征说明如下表所示：

表 3-1 乳腺癌数据集变量取值范围及描述表

变量名称	变量取值	变量描述
diagnosis	B/M	诊断结论
radius_mean	6.981-28.11	平均半径
texture_mean	9.71-39.28	平均纹理
perimeter_mean	43.79-188.5	平均周长
area_mean	143.5-2501	平均面积
smoothness_mean	0.0524-0.1634	平均平滑度
compactness_mean	0.0194-0.3454	平均密实度
concavity_mean	0-0.4268	平均凹度
concave points_mean	0-0.2012	平均凹点
symmetry_mean	0.106-0.304	平均对称性
fractal_dimension_mean	0.0499-0.0974	平均分形维数
radius_se	0.1115-2.873	半径标准差
texture_se	0.3602-4.882	纹理标准差

续表 3-1

perimeter_se	0.757-21.98	周长标准差
area_se	6.802-542.2	面积标准差
smoothness_se	0.0017-0.0311	平滑度标准差
compactness_se	0.0023-0.1354	密实度标准差
concavity_se	0-0.396	凹度标准差
concave points_se	0-0.0528	凹点数量标准差
symmetry_se	0.0079-0.0789	对称性标准差
fractal_dimension_se	0.0009-0.0298	分形维数标准差
radius_worst	7.93-36.04	最坏半径
texture_worst	12.02-49.54	最坏纹理度
perimeter_worst	50.41-251.2	最坏周长
area_worst	185.2-4254	最坏面积
smoothness_worst	0.0712-0.2226	最坏的平滑度
compactness_worst	0.0273-1.058	最坏密实度
concavity_worst	0-1.252	最坏凹度
concave_points_worst	0-0.291	最坏凹点数量
symmetry_worst	0.1565-0.6638	最坏对称性
fractal_dimension_worst	0.055-0.2075	最坏分形维数

从表中可以看出，数据集的 30 个特征其实是通过观察乳腺癌患者癌细胞的细胞核的十个生理形态得到的，分别是癌细胞核半径、纹理度、周长、面积、平滑度、密实度、凹度、凹点数量、对称性和分形维数，通过对这十个细胞核形态特征采集其平均值，标准差已经最坏值三个角度共得到数据集中的 30 个特征指标。这三十个指标作为自变量的同时，使用乳腺癌诊断结果作为因变量，数据集中采用的 B 和 M 作为因变量特征的不同水平，其分别代表了乳腺癌患者的诊断结果为良性或者是恶性。

3.1.2 部分指标解释

在数据集中的部分指标为医学中乳腺癌研究的专业术语，这里将其中部分指标进行阐述，分别对其中乳腺癌细胞核的纹理（texture）、平滑度（smoothness）、密实度（compactness）、凹度（concavity）、凹点数量（concave points）、对称性

（symmetry）以及分形维数（fractal dimension）进行简单解释[48]。如下表所示：

表 3-2 部分特征解释表

特征	解释
纹理（texture）	通过寻找乳腺癌细胞图像中各个分量像素中灰度强度的方差的大小来定义
平滑度（smoothness）	通过测量细胞核一条径向线和围绕它的线的平均长度来定义
密实度（compactness）	周长和面积的组合，计算公式为（周长^2/面积-1.0），这个无量纲化指标会随着边界的不规则性而增加。然而，细长细胞核的这种形状测量也会增加，这并不一定表明恶性肿瘤的可能性增加。对于小细胞，由于样本数字化带来的精度下降，该特征也向上偏移
凹度（concavity）	周长和面积的组合，计算公式为（周长^2/面积-1.0），这个无量纲化指标会随着边界的不规则性而增加。然而，细长细胞核的这种形状测量也会增加，这并不一定表明恶性肿瘤的可能性增加。对于小细胞，由于样本数字化带来的精度下降，该特征也向上偏移
凹点数量（concave points）	类似于凹度，但只测量轮廓凹度的数量，而不是大小
对称性（symmetry）	在找到细胞平面图最长轴的基础上测量垂直于长轴到细胞边界的两个方向上的线之间的长度差
分形维数（fractal dimension）	细胞核的周长是使用越来越大的“标尺”来测量的。随着标尺尺寸的增加，测量精度的降低，观测到的周长也会减小。将这些值绘制为对数尺度的值，并测量向下的斜率，可以近似分形维数

3.2 变量总体分布情况

在拿到一个数据集的时候，不能直接就对数据放到数学模型中进行建模。在正式进行建模之前，还应该对数据整体情况进行描述性的统计，这有助大致了解数据结构并针对于后续的研究选择合适的方向或者合理的变量。对于本文使用的威斯康星州（诊断）乳腺癌数据集，使用 Python 语言对其进行描述性统计和探

索性分析。

首先对于因变量乳腺癌患者的诊断情况进行统计，其因变量分布情况如下图所示：

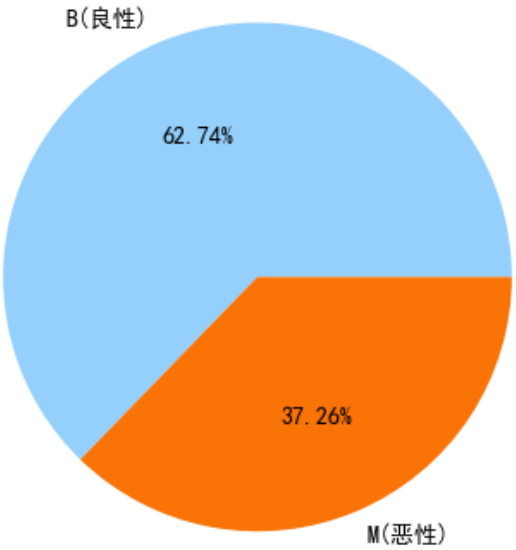

图 3-1 乳腺癌数据集因变量分布情况饼状图

从图中可以看到，数据集 569 条样本数据中，良性肿瘤患者（水平为 B）占 62.74%，共计 352 例。恶性肿瘤患者（水平为 M）占比为 37.26%，共计 208 例。整体情况来看，良性肿瘤患者的数量是略多于恶性肿瘤患者的数量。

对因变量的分布情况有所了解之后，后续对自变量的分布情况进行统计。因数据特征是乳腺癌细胞的 10 个病理特征并每个病理特征对应了 3 个不同的统计量，故将所以病理特征以 3 个为一组，对其绘制箱线图，以此来观察自变量的分布情况。对于细胞核的半径、纹理、周长以及面积指标，其分布情况如下图所示：

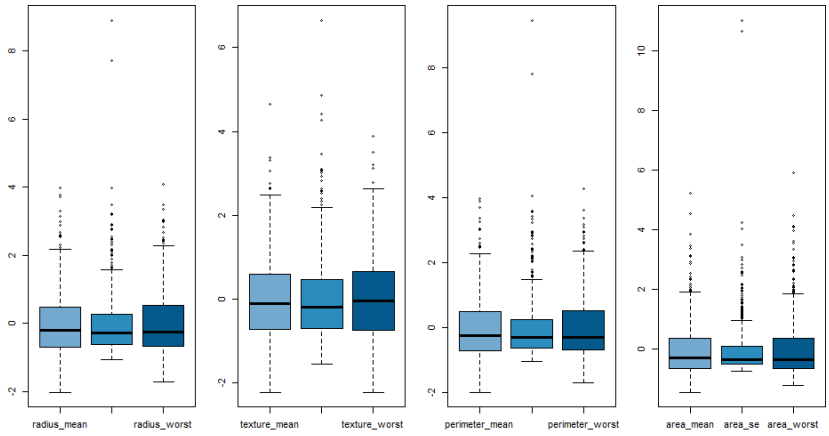

图 3-2 细胞核半径、纹理、周长、面积分布箱线图

从图 3-2 中可以发现，对于癌细胞的半径指标，在处于控制线内的数据分布较为均匀，其中平均半径、半径方差、最差半径的中位数都位于 0 左右（数据已进行标准化），大部分点位都位于 4 以下，只存在两个半径方差的点位出现了较大程度的离群；对于细胞核的纹理指标，同样是大部分点位在控制线内分布比较均匀，但在上控制线上方同样出现了大量的离群点。对于细胞核的面积指标，与前三个指标相比，面积指标的各项数据显得更为集中，且中位数都位于 0 下方，同样是存在有离群点的问题。对于数据的平滑度、密实度、凹度以及凹点四个指标，其分布情况如下图所示：

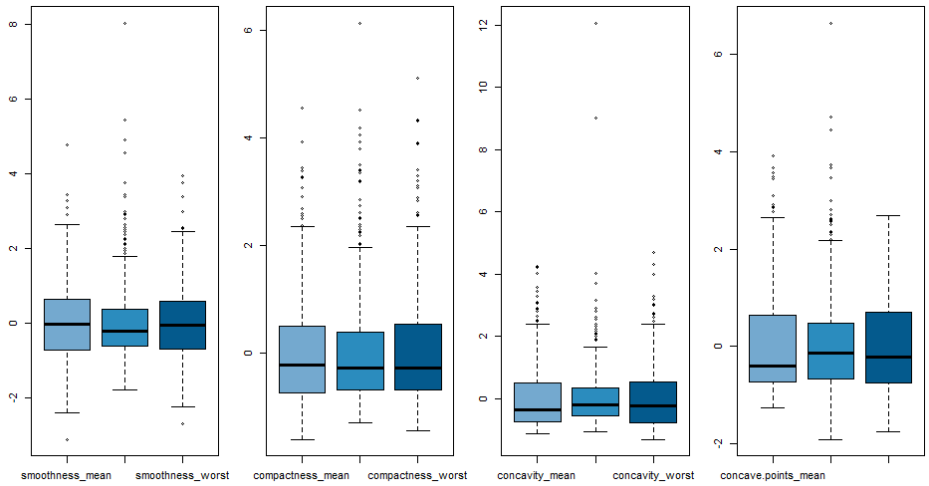

图 3-3 细胞核平滑度、密实度、凹度以及凹点分布箱线图

从图 3-3 中可以发现，对于癌细胞核的平滑度指标，平滑度均值和最差平滑度的中位数都是位于 0 刻度线之上的，平滑度方差的中位数是位于 0 刻度线之下的，与其他指标不同的是，平滑度指标出现了两个位于控制线下方的离群点；除此之外，其余三个指标的分布都不太均匀，且同样存在位于控制线上方的离群点，特别是凹度和凹点数量这两个指标的均值，其中位数离 0 刻度线有比较多的距离。对于最后两个指标对称性和分形维数，其分布情况如下图所示：

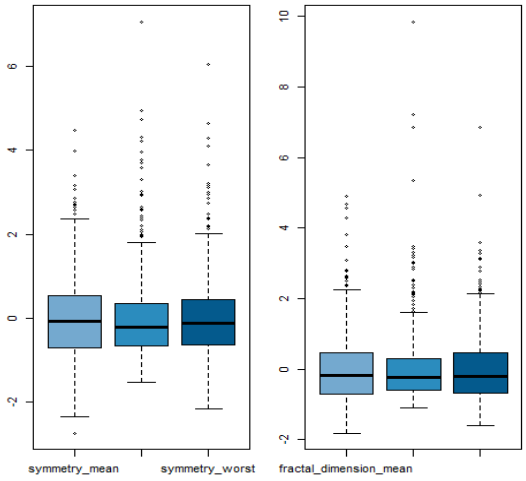

图 3-4 细胞核对称性和分形维数分布箱线图

从图 3-4 可以看出，对称性与分型维度的整体分布在控制线内的分布都较为均匀，且中位线都大致位于 0 刻度线上下，但与上述指标存在的同样问题是模型存在着一定的离群点，出平均对称性有一点离群点位于下控制线外，其余离群点都在上控制线之外。

综合上面对于 30 个自变量分布情况的观察可知，大部分指标的分布都比较均匀，少部分指标如凹度均值和凹点数量的均值，其中位数都位于 0 刻度线下方。除此之外，所有自变量都存在着一定量的离群点，且都属于上位离群点，在后续建立模型的时候要注意这些离群点对模型产生的影响，如果离群点过多降低了模型的预测效果，需要选择合适的方法对其进行处理。

3.3 单变量影响的描述性分析

单因素变量分析是数据探索性分析中非常重要的一环。通过这样的统计分析去判断一个变量对因变量是否具有影响以及是怎么样去影响因变量的，对后续变量的运用提供了十分重要的基础。本文研究的乳腺癌数据集一共包含了 10 个病理指标的 30 个特征，其中包括指标的均值、方差以及最坏值，这里分别在这三个方面选取部分指标进行单因素影响分析。

3.3.1 区分度较好的变量

以下四个自变量是通过绘制直方图和分组折线图描述分析总结认为对因变量区分程度较好的变量，分别是平均周长、平均面积、平均凹度和平均凹点数，图像如下：

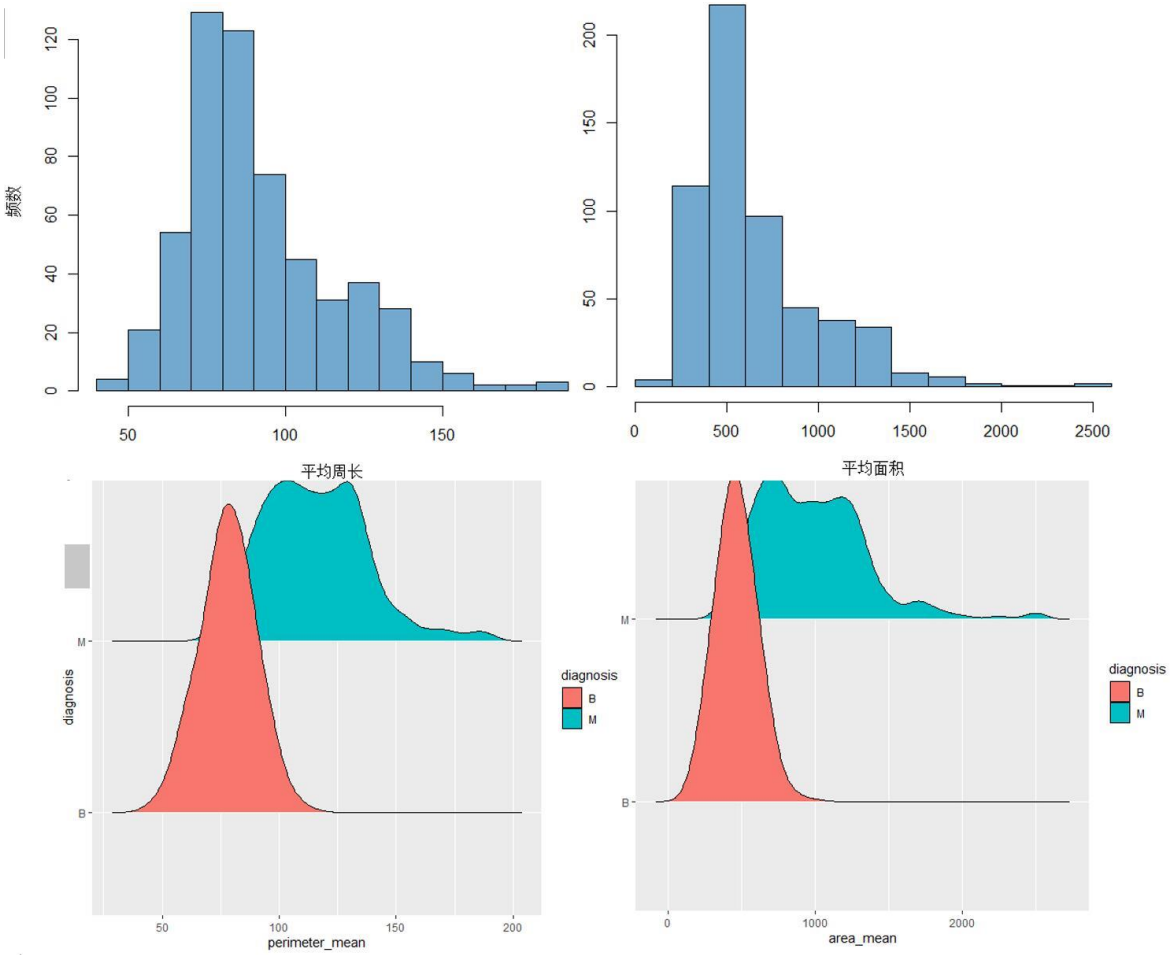

图 3-5 平均周长、平均面积直方图和分组折线图

图 3-5 左边两幅图分别描述细胞核平均周长的整体分布情况和按照因变量诊断结果分组后的分布情况。从整体分布情况的角度来看，病人的细胞核平均周长大概率集中在 50-100 之间；而从分为良性病人和恶性病人的角度来看，良性病人的癌细胞核平均周长整体是小于恶性病人的，其交叉区间大概位于 50-125 之间，即说明癌细胞核平均周长处于这个区间的病人存在着不确定性的诊断结果，而平均癌细胞周长在 50 以下的患者则不认为其存在恶性肿瘤，相反，对于癌细胞核平均周长处于 125 以上的病人，则可以认为其存在着恶性肿瘤，需要更进一步的治疗。

图 3-5 右边两幅图描述的是细胞核平均面积和整体分布情况和按照病人诊

断结果分组后的分布情况。从第一张图可以看出，大部分病人的细胞核平均面积都处于 500 左右，有很少量的样本细胞核平均面积超过了 2500，出现了十分极端的情况。将病人安装因变量的不同水平进行分组后，可以发现一个现象：即恶性肿瘤患者的细胞核平均面积包括了 0-3000 的整个范围，而良性肿瘤患者仅仅只包含了 0-1000 这个范围，从这样的现象中可以推断认为：对于乳腺癌患者来讲，如果其癌细胞核平均面积位于 0-1000 之间，则认为其有患恶性肿瘤的风险，如果其癌细胞核平均面积位于 1000 以上，则可以认为其癌细胞中有发展为恶性肿瘤的癌细胞。

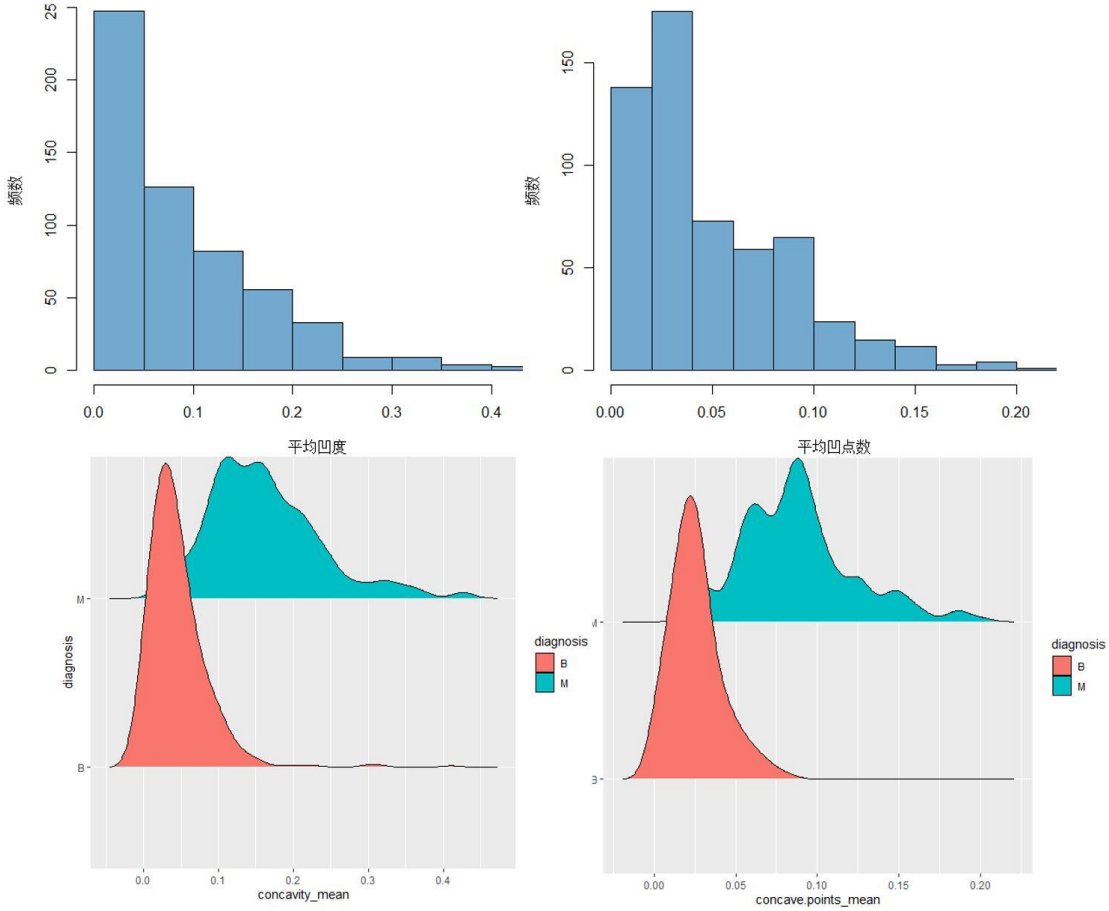

图 3-6 平均凹度、平均凹点数直方图和分组折线图

图 3-6 左边两幅图描述的是细胞核平均凹度和整体分布情况和按照病人诊断结果分组后的分布情况。与上述特征都不同的是，样本乳腺癌细胞平均凹度的整体分布完全右偏，有将近一半的数据位于 0-0.05 之间，大部分数据都集中在了 0-0.1 之间；结合图 3-18 可以发现，良性病人和恶性病人两组数据的偏离程度非常明显，也就是说，大部分的良性病例癌细胞核的平均凹度是较小，而大部分诊断为恶性的病例细胞核平均凹度较大，这说明这个特征能够很好的区分乳腺癌患者的诊断情况，在后续建模中，应该重点考虑此类特征。

图 3-6 右边两幅图描述的是细胞核平均凹点数和整体分布情况和按照病人诊断结果分组后的分布情况。和细胞核平均凹度一样，细胞核平均凹点数也呈一个明显的右偏分布，其整体位于 0-0.2 区间范围内，大部分样本位于 0-0.05 范围内。从分组数据折线图中，可以看出两组数据的重合部分比较小，这说明两种因变量水平中细胞核平均凹点数的差异比较大的，也就是细胞核平均凹点数会影响乳腺癌的最终诊断。综上认为，乳腺癌细胞核平均凹点数指标能够有效区分乳腺癌的不同类型，在后续统计建模过程中应当予以考虑。

3.3.2 区分度较差的变量

以下三个自变量是通过绘制直方图和分组折线图描述分析总结认为对因变量区分程度较好的变量，分别是平均分形维数、平均对称度和平均平滑度，其图像如下：

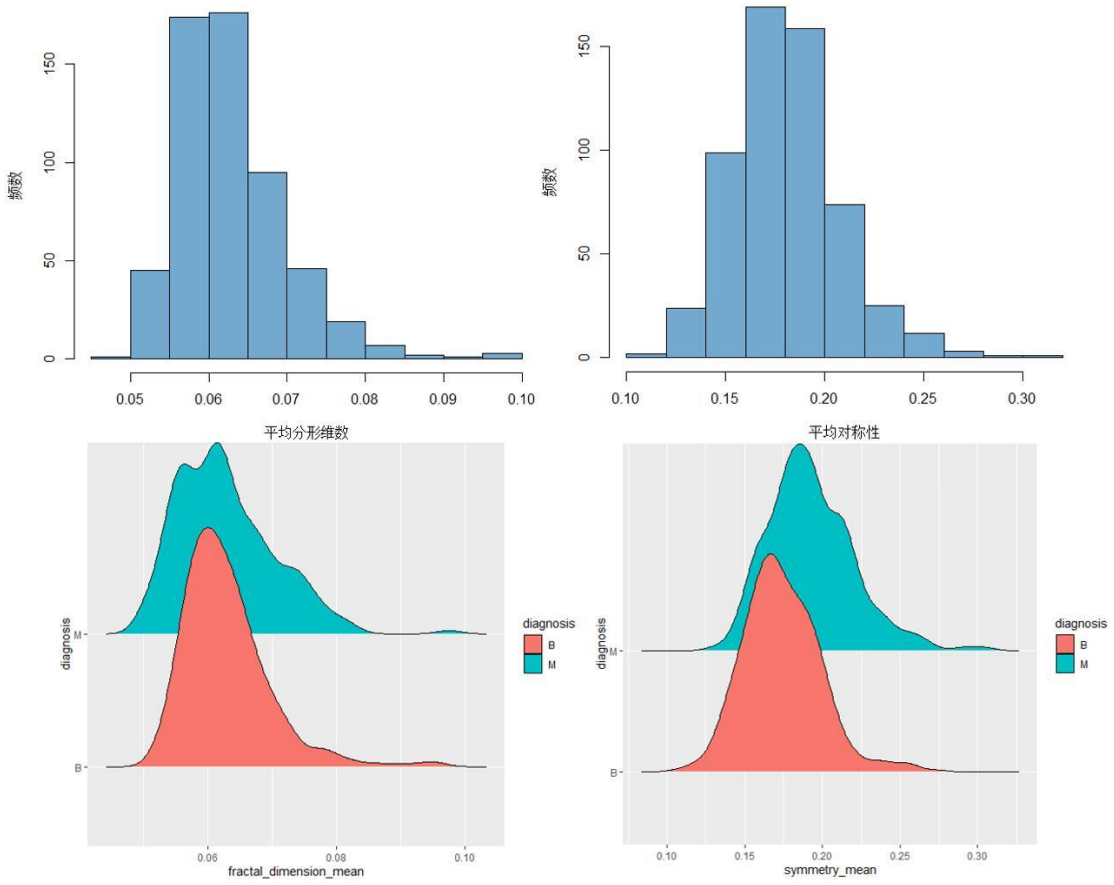

图 3-7 平均分形维数、平均对称度直方图和分组折线图

图 3-7 左边两幅图描述的是细胞核平均对称性和整体分布情况和按照病人诊断结果分组后的分布情况。从整体角度来看，乳腺癌细胞核平均对称性位于

0.10-0.30 之间，分布比较均匀，不存在偏态的情况，大部分数据集中在 0.15-0.20 之间。从分组的角度上来看，恶性病例（M 型）整体的对称性要高于良性病例（B 型），但是两组整体位于的区间范围是一致的。也就是说无法通过观察乳腺癌患者细胞核的平均对称性来判断该患者的最终诊断结果是良性或者恶性，也说明了该特征在对最终诊断中提供的信息价值不如其他特征，在后续的建模过程应综合考虑是否选择此变量。

图 3-7 右边两幅图描述的是细胞核平均分形维数和整体分布情况和按照病人诊断结果分组后的分布情况。从整体角度来看，乳腺癌细胞核平均分形维数位于 0.05-0.1 区间之内，大部分数据位于 0.055-0.07 之间，分布整体呈右偏分布。对于分组后的情况，可以发现两组中样本分布几乎一致，这说明细胞核平均分形维数这一特征完全不具有作出最终诊断的参考价值，在后续的建模过程应加以修改或者直接予以删除。

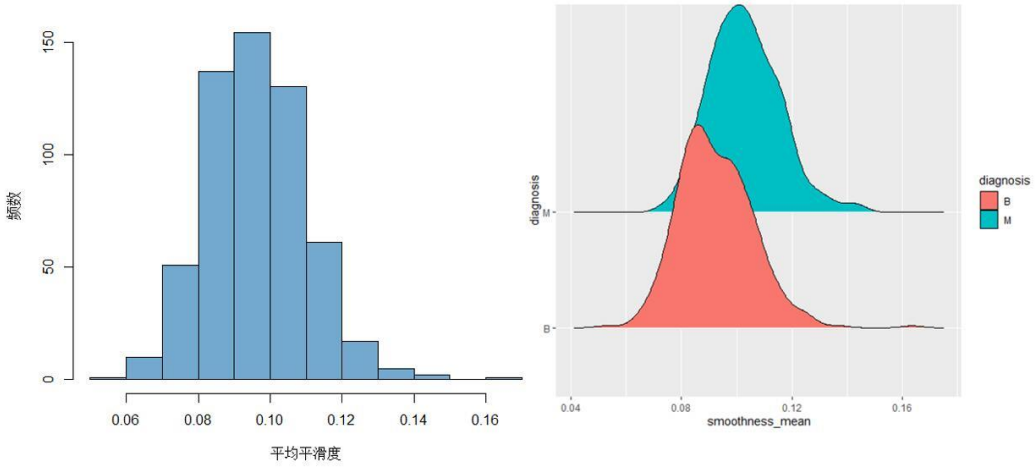

图 3-8 平均平滑度直方图和分组折线图

图 3-8 描述的是细胞核平均平滑度和整体分布情况和按照病人诊断结果分组后的分布情况。从整体角度来看，癌细胞核平均平滑度分布呈正态分布且对称程度较高，整体范围位于 0-0.18 区间，大部分数据处于 0.08-0.11 之间。从分组数据折线图的角度来看，两组数据所包括的区间范围几乎完全重合，整体趋势上良性病人的癌细胞核平均平滑度略小于恶性病人，但是由于两者所包含的区间基本相同，因此是无法通过观察癌细胞核平均平滑度这一指标对乳腺癌患者进行初步诊断。

通过对上述十个特征的影响分析，对自变量的分布情况及是否对因变量产生影响做出了基础的初步判断。根据绘制条形图和折线图，可以认为对因变量区分度较好的变量有细胞核平均面积（area mean）、细胞核平均凹点数（concave points

mean）、细胞核平均凹度（concavity mean）、细胞核平均周长（perimeter mean）；对因变量区分度较差的特征有细胞核平均分形维数（fractal dimension mean）、细胞核平均对称性（symmetry mean）、癌细胞核平均平滑度（smoothness mean）。在对各变量有了初步的了解，在后续的统计建模和变量筛选的过程中应更加注意不同变量对模型带来的影响。

3.4 数据预处理

在进行数据分析工作之前，需要对数据的整体情况进行观察统计，并对数据的一些不正常现象作出相应的处理，这才能保证在后续建模过程中得到的模型是有效且可靠的。除此之外，还需要对数据集进行一些准备工作，便于后续建模过程中提高效率。

3.4.1 异常值处理

数据的异常值通常包括缺失值和离群值等。在发现异常点的情况，一般的处理方式有删除、替换以及补全等。

在上文的描述性统计中，发现其样本数据的各个特征中存在着离群值。经过统计，发现各个特征中发现的偏离点大多数都来自于同一个样本。其中偏离较大的样本数量只有 9 条，其占总样本数的比例很小，故直接将这存在异常值的 9 个样本剔除。即后续模型的建立是建立在 560 条正常范围内的样本之上的。

3.4.2 划分训练集和测试集

对于模型的评估与选择，可以通过实验测试来对学习器的泛化误差进行评估并对模型进行选择，因此需要一个用于测试样本来测试学习器对没有见过的新样本的判别能力。因此通常在建立模型市，需要将全部数据集划分为训练集（Train Data）和测试集（Test Data）。

划分数据集的一般方法即按照一定的比例将数据集划分成为两个互斥的集合。其中划分比例一般情况下 7:3 或者 8:2,即表示使用 70%或者 80%的样本训练模型，再使用 30%或者 20%的样本数据对模型的泛化性能进行评估。本文使用 8:2 的比例对全部数据集进行划分，即训练集包括 448 条样本数据，测试集包含 112 条样本数据。

4. 数据特征的选择

4.1 特征选择

4.1.1 特征选择分类

特征选择即指在某种特定的评价标准下从数据原有的特征集中选取最优的特征子集。根据特征选择算法的整个过程与之后的建模过程是否有独立关系，特征选择算法分为了三个类别，分别包括了过滤式（filter）、封装式（wrapper）和嵌入式（embedded）[19]。

过滤式特征选择方法是利用一个独立的评价函数去度量一个特征对分类的重要程度，然后根据度量值对所有的特征进行排序，最后选择度量值最高的特征构成特征子集；封装式特征选择算法是利用某个分类学习器算法作为特征的评价标准去选择特征子集，然后再利用该子集来训练分类器；嵌入式特征选择方法是将特征选择的过程直接嵌入到分类器的学习过程中。

4.1.2 特征选择过程

特征选择主要包括了四个过程，分别是子集的生成，到子集的选择。随后是子集的评估和停止判决[51]，其运行流程如下图所示：

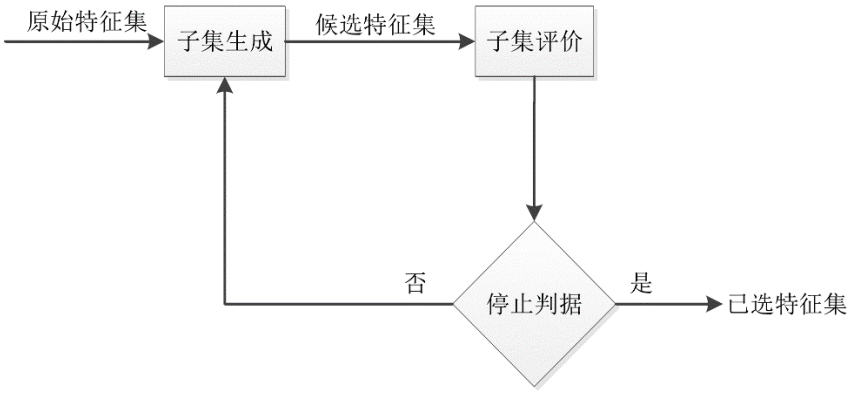

图 4-1 特征选择过程

从图中可知整个特征选择的步骤，首先对于备选特征集的若干特征生成子集，在生成的若干个子特征集中选择一个，那么这个子集就成为一个备选特征集，然后使用某一准则或评估函数对备选子集进行判断，如果某个备选子集满足了准则或评估函数设定的停止条件，则停止判断，并输出满足准则或设定要求的备选子

集作为特征选择的结果[19]。

子集的生成这一步骤是特征选择算法的开始，会影响到后面子集的评估结果。子集的评估是对生成的特征集的优劣进行评价，如果当前得到的优劣值高于了之前所生成的所有优劣值，则将当前的备选子集作为输出的最优值。

停止判决是评估当前的特征子集是否满足预设的条件。如果满足了设定的某个阈值，则结束特征选择的整个流程，否则就使用新的方法（如重新抽样）重新生成一批新的特征子集。特征算法通常使用四种条件作为停止判断，分别是：特征集的特征数超过了预先定好的特征数量；重复搜索的循环次数超过了实现设置的次数；或是当前特征子集的评估已经是最优值；以及当前评估分数已经超过了预先设定的评估分数。

4.2 乳腺癌数据集的特征选择

4.2.1 基于 mRMR 算法下的特征选择

mRMR 算法（Max-Relevance and Min-Redundancy）是一种用途十分广泛的封装式（wrapper）特征选择方法。mRMR 通过增加特征子集之间的相关性，较少特征子集之间的冗余性的原则，对特征进行排序并输出。mRMR 采用序列前项搜索策略，每次从剩余的候选特征中选出一个能够与已选特征集 S 构成当前最好的特征集。[25]因此，mRMR 在特征选择的过程中考虑了候选特征与已选特征集 S 之间的冗余，并使用代价函数评价候选特征的质量，具体的代价函数如下：

$$J(g)=I(C;g)-\frac{1}{|S|}\sum_{s\in S}I(s;g) \tag{4.1}$$

式(4.1)中 $|S|$ 表示集合 S 中的特征个数；C 表示目标类别，即数据集中乳腺癌的两种不同诊断结果。$I(C;g)$ 表示特征 g 和水平 C 之间的相关性，而 $I(s,g)$ 表示了两个特征之间的冗余性。

mRMR 算法的具体步骤[21]如下：

（1）初始化空集 S ,采用序列前向搜索策略根据互信息选出与目标类别最相关的特征，加入到已选特征集 S 中，并从候选特征集中剔除该特征。

（2）使用代价函数评价候选特征集合中的每一个特征，并取得最大值 $J(g)$ 的特征加入到已选特征集 S 中，并将其从候选特征集中删除。

（3）判断是否满足特征选择的停止条件，如果满足，则停止特征选择过程，并返回当前特征子集，否则转到步骤（2）

本文使用 Python 语言中的 mifs 库对数据进行 mRMR 特征选择[52]。Mifs 库提供了若干个参数对算法的使用进行调整。其中参数 method 表示了信息准则的选取，n_features 表示了设定的输出特征个数，其具体参数设置如下表所示：

表 4-1 Mifs 函数参数说明表

参数	取值范围	说明
method	JMI\JMIM\MRMR	选择准则，分别是联合互信息、联合互信息最大化、最大相关性最小冗余
k	（4,10）	设置用于内核密度估计的样本数。默认 5
n_features	（1，p）	设置输出的特征数
categorical	True\False	设置因变量类别，分别是离散型\连续性
n_jobs	(1,n)	设置同时运行线程数
verbose	(0,1,2)	设置输出详细程度

在默认条件下，将样本数据使用 mifs 包中的 MRMR 准则进行计算，设定输出的特征数为 20，则得到的结果如下表所示：

表 4-2 乳腺癌数据集各特征代价函数得分表

序号	特征	$J(g)$	序号	特征	$J(g)$
1	perimeter_worst	0.4448	11	smoothness_worst	-0.1769
2	texture_worst	-0.0074	12	fractal_dimension_se	-0.1221
3	compactness_worst	0.2073	13	texture_mean	-0.2131
4	radius_se	0.1558	14	radius_mean	-0.1280
5	area_se	0.0199	15	perimeter_se	-0.2147
6	concavity_wors	0.0121	16	compactness_se	-0.2183
7	symmetry_worst	-0.0622	17	concave points_se	-0.2308
8	compactness_mean	-0.0919	18	concave points_mean	-0.2197
9	smoothness_mean	-0.0686	19	radius_worst	-0.4036
10	smoothness_se	-0.0754	20	area_worst	-0.3203

因为乳腺癌数据集的结构是 10 个生理指标的 30 个数据特征，比如就乳腺癌细胞核半径这一生理指标其对应了三个特征，即平均值、标准差、最坏值，这三个特征必定是提供了重复的信息，所以，这里得到的序号前 9 个特征都刚好表示

了不同的生理指标，其中第三栏表示的是每一次循环中成功选择的变量的代价函数（即一次循环中代价函数的最大值）。将其绘制成折线图如下图所示：

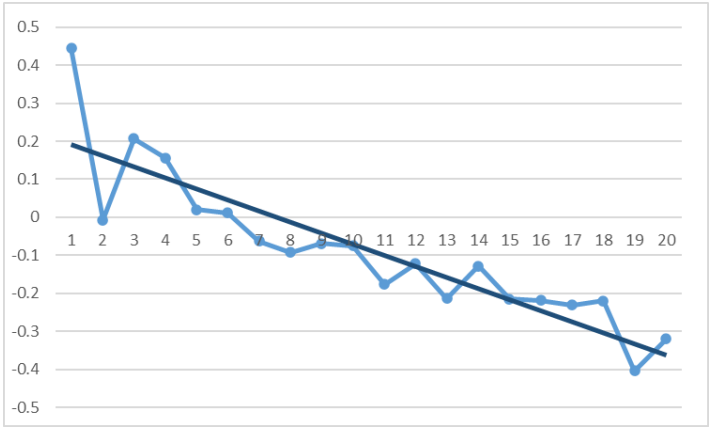

图 4-2 代价函数折线以及趋势图

从折线图中可以看出，在变量选择的整个过程中，代价函数的局部可能出现上下波动的情况，但是整体的趋势是不断减小的。这说明代价函数的趋势是由大变小，即说明备选变量和因变量之间的相关性减去备选变量和已选变量之间平均相关的差是在逐渐减小的。换句话说，越到后面选择的变量，其满足最大相关性，最小冗余度的值就越低。

前面提到，在模型程序运行的时候涉及到了一个参数 k，其在程序中的作用是设置用于内核密度估计的样本数[53]。这是因为使用 mRMR 算法时需要涉及到计算自变量和自变量之间以及自变量和因变量之间的互信息，而互信息的计算是需要涉及到样本特征的密度函数的，所有这里使用通过抽取样本的方法来估计样本的密度函数。根据程序源代码可知，这里采用的是使用的是 K 邻近（KNN，K-Nearest Neighbor）算法来进行核密度估计。

系统默认给了每次抽取 5 个样本估计样本函数 ，如果改变此参数，势必会带来最终特征排序结果的改变，这里分别取 4 到 10，最终排序结果如下。

表 4-3 k 值改变后模型最终得分表

k	4	5	6	7	8	9	10
$sum(J(g))$	-2.620	-1.713	-0.936	-1.400	-1.701	-2.066	-2.581
$\bar{J}$	-0.131	-0.086	-0.047	-0.070	-0.085	-0.103	-0.129

从表中可以看见，代价函数的总和和均值随着参数 k 的改变的趋势。可以发现 $sum(J(g))$ 先递减再递增，在 $k=6$ 时取到了最小值，说明此时整体选择的变量是比取其他值更加优秀的。此时发现模型前 9 个特征分别代表了对于乳腺癌指标

的 9 个不同生理指标，所以选择此时的前 9 个变量分别是： perimeter_worst、texture_worst、radius_se、compactness_worst、concavity_mean、area_se、symmetry_se、smoothness_mean、 symmetry_worst。

4.2.2 基于 ReliefF 算法下的特征选择

ReliefF 算法的思想属于三大类特征选择算法中的过滤式特征选择算法（filter），其主要通过样本数据学习得到每个特征对应的权值。ReliefF 算法参加权重计算的近邻样本取决于样本之间距离的大小，考虑到参与距离计算的特征会对样本的相对距离产生影响，从而影响选择哪些近邻样本，最终对特征权重的评估起作用，因此 ReliefF 算法在计算近邻的过程中实现了特征间的相互影响，将特征的相互依赖作用纳入考虑。ReliefF 算法的核心思想是根据特征在同类与异类近邻样本之间的差异对特征的区分能力进行衡量。若特征在同类样本之间差异小，而在异类样本之间差异大，则表明该特征具有较强的区分能力。

ReliefF 算法详细描述如下[20]：

输入：训练集 S，迭代数 n，近邻样本数 k，阈值 T（目标维数）

（1）对训练数据集中所有样本的每个特征的权值赋初值（一般为 0）。

（2）在训练集中随机抽取一个样本 R。

（3）对于和 R 同类别的样本集，在其中找到与选取样本 R 最近邻的 k 个样本；并在与 R 不同类别的样本集中同样找最近邻的 k 个样本。

（4）利用下面公式更新每个特征的权重 $W(f_i)$：

$$\begin{aligned} W(f_i) = W(f_i) - \sum_{j=1}^{k} diff(f_i, R, H_j)/(mk) + \\ \sum_{L \neq class(R)} \left[\frac{P(L)}{1 - P(class(R))} \sum_{j=1}^{k} diff(f_i, R, M_j(L)) \right] / (mk) \end{aligned} \tag{4.2}$$

式(4.2)中 $diff(f_i, R, H_j)$ 是两个样本 R 和 H_j 在特征 f_i 上面的差，它计算公式在上文中的(2.11)，且之中的 H_j 是和 R 相同一类的第 j 个近邻样本。$diff(f_i, R, M_j(L))$ 表示的是两个样本在特征集 $M_j(L)$ 中特征 f_i 上的差，$class(R)$ 表示的是样本 R 所在的类，$P(L)$ 表示的是第 L 类目标的概率，而 $P(class(R))$ 表示的所在类的概率。

（5）重复第（2）步到第（4）步的过程 n 次，最终输出每个特征最后的权重。

（6）根据权重大小对特征进行排序，并输出对应的特征子集。

在清楚 ReliefF 具体计算步骤之后，使用程序对乳腺癌细胞核的 30 个特征进行特征筛选，得到其前 20 个特征的权重如下图（归一化后）所示：

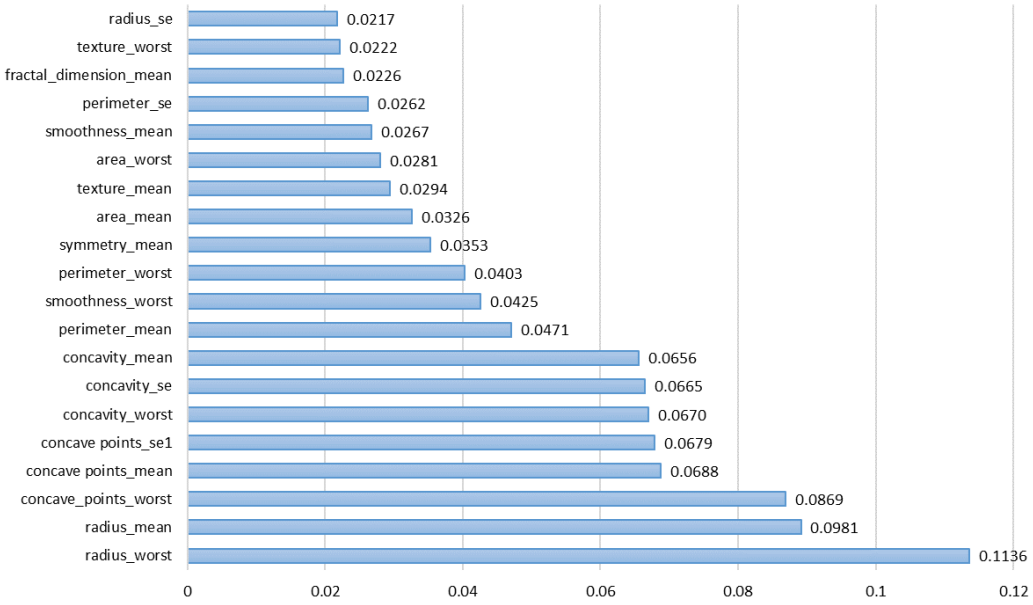

图 4-3 ReliefF 算法特征权重图

从图中可以看出，变量 radius_worst 的权重是最大的，远远超过其他变量的权重；其次是 radius_mean 和 concave_points_worst，权重都在 0.09 左右；随后是 concave points_mean、concave points_se、concavity_worst、concavity_se 这几个变量，其权重都处于 0.06-0.07 之间；再之后的变量就成有规律的递减趋势。

4.2.3 基于 Lasso 算法下的特征选择

LASSO（Least absolute shrinkage and selection operato）算法是一种非常典型的嵌入式（embedded）特征选择算法。嵌入式特征选择算法的特点是将特征选择和学习器的训练过程相互联系到一起。Lasso 算法在使用时是通过给目标函数增加一个惩罚项来达到压缩自变量权重的效果[26]。需要注意的是，惩罚项的系数 λ 往往是需要自定义的，在实际建模中很难去直接找到一个合适的 λ 来建立模型。所以需要通过确定一种准则去选择最佳的参数。目前应用最广泛的调节参数准则是 BIC 准则和 CV 准则。

在变量选择上，本文使用 C_p 准则来进行 Lasso 回归中变量的选择。C_p 准则是在残差平方和的基础上定义的一个准则，定义为[22]：

$$C_p = \frac{S_E^2}{\hat{\sigma}^2} - (n - 2p)$$

这里 $\hat{\sigma}^2$ 是全模型时得到的 σ^2 的估计：

$$\hat{\sigma}^2 = Y'(I - X(X'X)^{-1}X')Y/(n-p-1)$$

上面式子中的分子指的是部分变量的残差平方和。

将乳腺癌数据集使用 Lars 库中的 lars 函数进行变量筛选，最终得到的结果如下表所示：

表 4-4 Lasso 回归各步骤得分表

迭代次数	变量个数	RSS	C_p
0	1	133.012	1816.956
1	2	121.217	1607.551
2	3	76.939	815.968
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
31	26	30.115	22.752
32	25	30.099	20.465
33	26	30.099	22.465
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
40	31	30.018	31.005
41	30	30.018	29.002
42	31	30.018	31.000

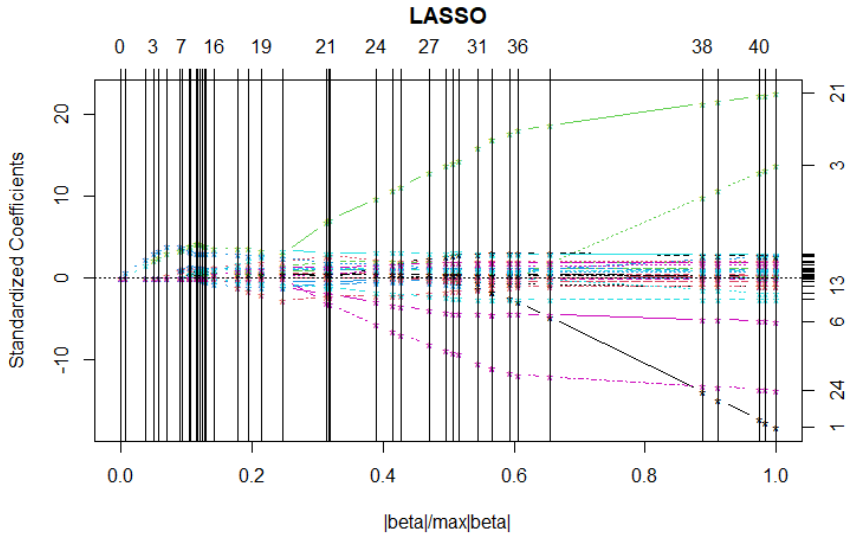

图 4-4 LARS 变量选择系数图

从表 4-4 中可以发现，当迭代次数为 32 次时，得到的 C_p 值是最小的，此时模型选择的变量个数为 25 个，这就意味着有 5 个变量的系数被压缩到了 0，即舍弃掉这 5 个变量。这五个变量分别是：perimeter_mean、area_mean、

symmetry_mean、compactness_se、compactness_worst。其余变量的权重系数如下表所示：

表 4-5 Lasso 变量选择最终变量权重表

变量	权重系数	变量	权重系数
radius_mean	-0.01593	concave.points_se	8.8216
texture_mean	0.0043	symmetry_se	1.3828
smoothness_mean	-0.1486	fractal_dimension_se	-5.8694
compactness_mean	-3.4580	radius_worst	0.1379
concavity_mean	1.5314	texture_worst	0.0076
concave.points_mean	2.0951	area_worst	-0.0008
fractal_dimension_mean	-0.5497	smoothness_worst	0.7644
radius_se	0.4592	concavity_worst	0.3745
texture_se	-0.0094	concave_points_worst	0.5899
perimeter_se	-0.0117	symmetry_worst	0.6317
area_se	-0.0013	fractal_dimension_worst	4.3400
smoothness_se	15.5439	concavity_se	-3.5469

在剩下的权重系数没有被压缩到 0 的变量中可以看到，其中较多变量的权重系数都比较小，有的甚至处于小数点后三位，故此时仅选择其中权重系数大于 1 的 9 个变量。

4.2.4 三种特征选择方法总结对比

本章通过分别使用过滤式（filter）、封装式（wrapper）和嵌入式（embedded）三种类型的特征选择方法代表：mRMR 算法、ReliefF 算法、Lasso 算法对乳腺癌（诊断）数据集进行了特征筛选，分别筛选出了 9、8 与 9 个不同的特征变量集合，其主要结果如下表所示：

表 4-6 各特征选择方法最终结果对比表

模型算法	特征集合
mRMR	perimeter_worst、texture_worst、radius_se、compactness_worst、concavity_mean、area_se、symmetry_se、smoothness_mean、symmetry_worst、
ReliefF	radius_worst、radius_mean、concave_points_worst、concave_points_mean、concave_points_se、concavity_worst、concavity_se、perimeter_worst、
Lasso	compactness_mean、concavity_mean、concave points_mean、smoothness_se、concave points_se、symmetry_se、fractal_dimension_se、fractal_dimension_worst、concavity_se

从表格中可以得知，三种方法共同选择的变量只有 2 个；被两种变量筛选方法共同选择到的变量有 3 个；仅被一种变量选择方法筛选到变量有 15 个，有 10 个变量同时被三种筛选方法所丢弃。这说明三种方法选择变量的重复率不太高，在后续的分析研究中是有对比价值的。

5 基于集成学习的乳腺癌诊断预测模型

5.1 模型评估准则

建立的数学模型需要对模型的泛化性能进行评价。对于各种模型有多种不同的评价准则进行评估。本文所研究的乳腺癌诊断问题作为一个典型的二分类问题，将使用以下的评价准则对模型性能进行量化。

（1）混淆矩阵

混淆矩阵（Confusion Matrix）即误差矩阵，其作为一种标准格式对模型的精确程度作出评价[28]，混淆矩阵表示为 n 行 n 列的矩阵。其中矩阵的每一列表示的是模型预测的两个或者多个类；每一行表示的是样本数据的实际类。在只有两个类别的情况下，混淆矩阵的具体形式如下：

表 5-1 混淆矩阵表

	预测类 1	预测类 0
真实类 1	TP	FN
真实类 0	FP	TN

其中 TP（True Positive）表示属于类 1 的样本同样被模型判断为类 1 的个数；FN（False Negative）表示属于类 1 的样本被模型错误判断到类 0 的样本个数；FP（False Positive）表示的是属于类 0 的样本被模型错误判断到了类 1 的个数；TN（True Negative）表示的属于类 0 的样本同样被模型判断为类 0 的样本个数。

（2）准确率、精确率、召回率

在建立混淆矩阵的基础上，可以计算更多的统计量作为模型的评价指标，常用的统计量有准确率、精确率以及召回率。

准确率（Accuracy）是评价模型最常用的一个指标，其表示的所有预测正确的样本数（包括正类和非正类）占所有样本的比例，准确率的计算公式如下：

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \tag{5.1}$$

精确率（Precision）也被称作查准率，即正确预测为正类的占全部预测为正类的比例。精确率是针对预测结果而言的，它表示的是预测为正类的样本中有多少是真正的正类样本。精确率的计算公式如下：

$$\text{Precison} = \frac{TP}{TP + FP} \tag{5.2}$$

召回率（Recall）也被称作查全率，即正确预测为正类的样本数占全部实际为正类的比例。召回率是针对原始样本而言的，它表示的是全体样本中的所有正类样本有多少被预测正确了。其计算公式如下：

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.3}$$

（3）F1 值

F1 值（F1 Score）是统计学中用来衡量二分类模型精确度的一种指标，其同时兼顾了分类模型的精确率和召回率，可以看作是模型精确率和召回率的一种调和平均，其最大值为 1，最小值为 0。计算公式如下：

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.4}$$

F1 值主要是解决模型结果的精确率和召回率不平和的问题，在实际评估中，模型的精确率和召回率往往是矛盾的，即当精确率高的时候，其召回率有可能会很低；当模型的召回率高的时候，其准确率可能会较低。在无法拿准是否使用精确率或者召回率来评价模型的时候，可以考虑使用 F1 值。

（4）ROC 曲线和 AUC 值

接受者操作特性曲线（Receiver Operating Characteristic curve, ROC）是指在不同判断标准下，以模型的假阳性率为横坐标，真阳性率为纵坐标，并连接各点绘制而成的连线。

假阳性率（False Positive Rate, FPR）即属于非正类的样本被模型判断为正类的概率。其计算公式为：

$$FPR = \frac{FP}{TN + FP} \tag{5.5}$$

真阳性率（True Positive Rate, TPR）即表于所有属于正类的样本被模型判断为正类的概率。其计算公式为：

$$TPR = \frac{TP}{TP + FN} \tag{5.6}$$

ROC 曲线的一个优秀特点[55]是：其不会因为测试集中正例样本和负例样本分布的变化而发生改变。真实的数据集往往伴有类不平衡（Class Imbalance）的情况，即每个类别样本的数量相差较大，或者每个类别样本数量随着时间的变化而变化，ROC 曲线就很好的解决了这个问题。

AUC 值（Area Under Curve）被定义 ROC 曲线下与坐标轴围成的面积[56]。由于有时候各模型 ROC 曲线有很多的交错，所以难以分清哪个模型的曲线是更好

的，所以可以使用 AUC 值来进行判断。并且使用 AUC 值也可以更好的比较两个模型 ROC 曲线的具体优劣程度[29]。

由于 ROC 曲线通常位于 $y=x$ 这条直线的上方，故 AUC 取值范围在 0.5 到 1 之间当 AUC 值取到 0.5 的时候，就认为此时的模型完全没有识别样本的能力，即对样本进行的随机分类；当 AUC 值取到 1 的时候，即模型对所有样本都进行正确分类。

5.2 集成学习预测模型

5.2.1 随机森林模型

本文使用使用 Python 语言 sklearn 库中的函数 RandomForestClassifier（）建立随机森林模型。带入上文中 mRMR、ReliefF 以及 Lasso 最终选择的变量，带入随机森林模型，得到三种特征选择方法的随机森林模型的测试集结果如下：

表 5-2 随机森林模型评价表

特征选择	Accuracy	Precision	Recall	F1
mRMR	0.8684	0.7714	0.7941	0.7826
ReliefF	0.9035	1.000	0.7105	0.8308
Lasso	0.8596	0.9355	0.6744	0.7838

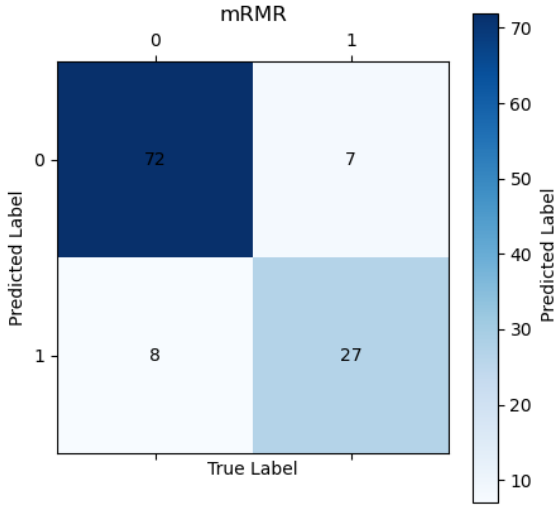

图 5-1 随机森林（mRMR）混淆矩阵

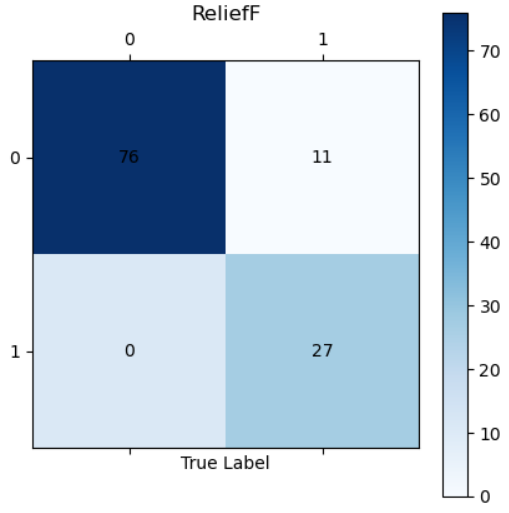

图 5-2 随机森林（ReliefF）混淆矩阵

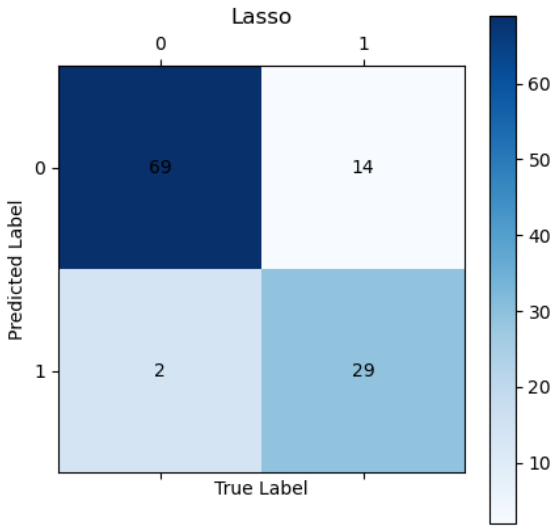

图 5-3 随机森林（Lasso）混淆矩阵

从表 5-3 中可以发现。对于准确率、查准率以及 F1 值这三个指标都是使用 ReliefF-RandomForest 最好，分别是 0.9035、1.0、0.8308。可以发现其中模型的查准率已经达到了 1,也就是预测为恶性肿瘤的患者中全部正确，说明如果只关注恶性肿瘤的诊断，此类方法是最为合适的；对于召回率指标，发现使用 mRMR 方法是最高的，这说明如果要在大概率恶性乳腺癌肿瘤患者中再次确定其肿瘤情况，更适合考虑使用 mRMR 方法筛选的变量，即更适合在相关医生高度怀疑时辅助作出判断。

绘制的模型 ROC 曲线图如下：

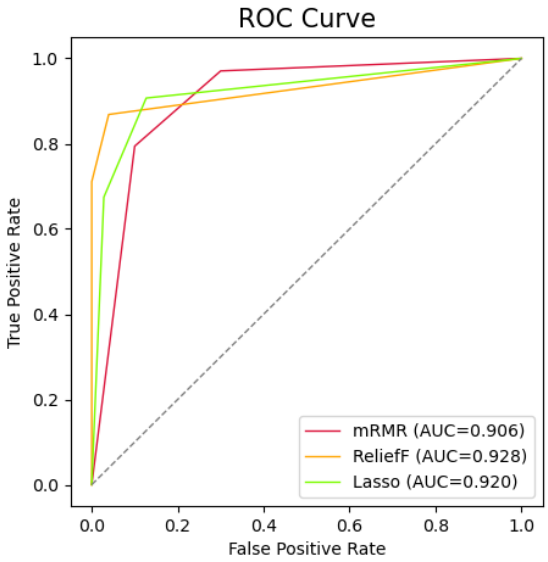

图 5-4 随机森林 ROC 曲线

从 ROC 曲线图中可知，三种模型在阈值不同时的表现各有优劣，无法通过图像给出一个精确的判断，所以在此基础上通过比较三个模型的 AUC 值来判断

优劣。图中角标显示使用 mRMR 变量的随机森林模型的 AUC 值为 0.906，使用 ReliefF 变量的随机森林模型的 AUC 值为 0.928，最后是使用 Lasso 方法选择变量的随机森林模型 AUC 值为 0.920。三个模型中，使用 ReliefF 方法和使用 Lasso 方法的模型泛化性能差距不大，但均优于使用 mRMR 的模型性能。综合上述表格中的各项指标，使用 ReliefF 结合随机森林模型的性能是最优的。

观察使用 ReliefF 方法建立的随机森林模型在验证集上的情况，从图 5-3 的混淆矩阵中可以看到，模型将 38 个乳腺癌恶性患者中的 27 名患者数据成功识别为了恶性肿瘤，将 11 名恶性肿瘤患者错误识别成了良性肿瘤患者；此外，对于另外 76 名良性乳腺癌患者，模型将其中 76 名患者都成功分类为良性肿瘤，没有良性肿瘤患者被误分类到恶性。

5.2.2 Adaboost 模型

本文使用使用 Python 语言 sklearn 库中的函数 AdaBoostClassifier（）建立 Adaboost 模型。结合第四章特征选择中使用三种不同的特征选择方法：mRMR、ReliefF 以及 Lasso 最终选择的变量，带入 Adaboost 模型，得到三种特征选择方法的 Adaboost 模型的测试集结果如下：

表 5-3 Adaboost 模型评价表

特征选择	Accuracy	Precision	Recall	F1
mRMR	0.8596	0.7045	0.9118	0.7949
ReliefF	0.8860	0.8000	0.8235	0.8116
Lasso	0.9298	0.9000	0.9000	0.9000

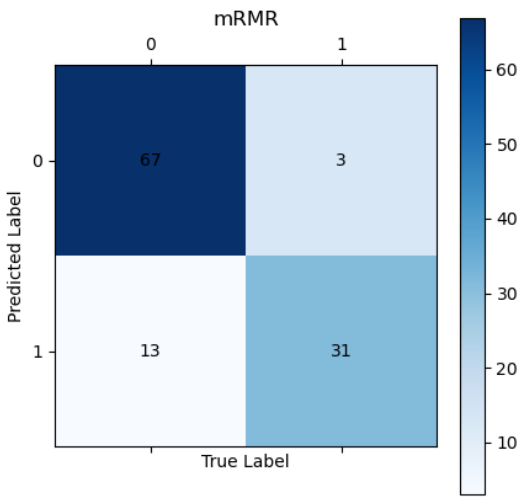

图 5-5 Adaboost（mRMR）混淆矩阵

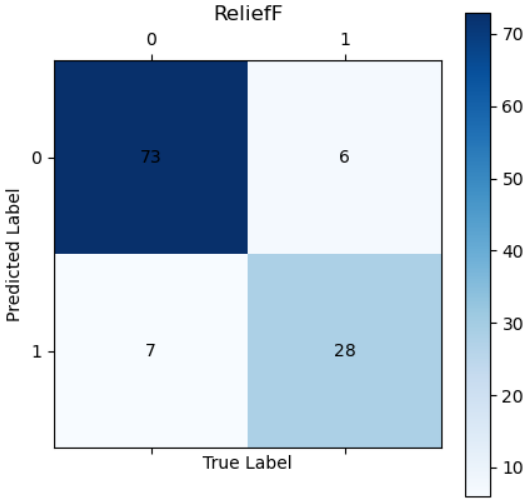

图 5-6 Adaboost（ReliefF）混淆矩阵

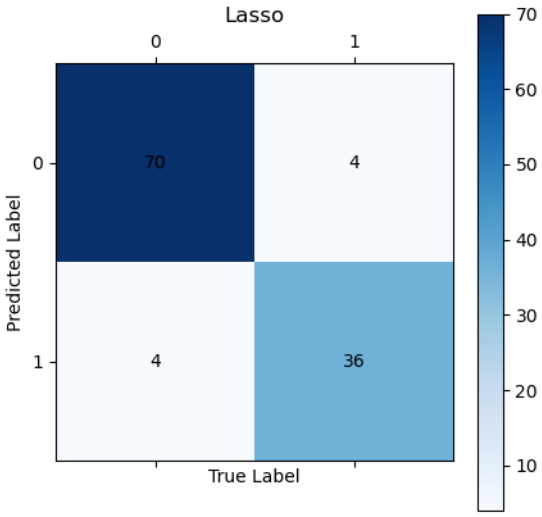

图 5-7 Adaboost（Lasso）混淆矩阵

从表 5-3 中可以发现。对于准确率、查准率以及 F1 值这三个指标，使用 Lasso 方法选择的变量要优于其他两种方法。这说明如果研究者更关注于对每一个病人都作出准确的判断，建议使用 Lasso 筛选后的变量结合 Adaboost 进行使用；对于召回率指标，发现使用 mRMR 方法是最高的，这说明如果要在大概率恶性乳腺癌肿瘤患者中再次确定其肿瘤情况，更适合考虑使用 mRMR 方法筛选的变量，即更适合在相关医生高度怀疑时辅助作出判断。

绘制三个模型的 ROC 曲线图如下：

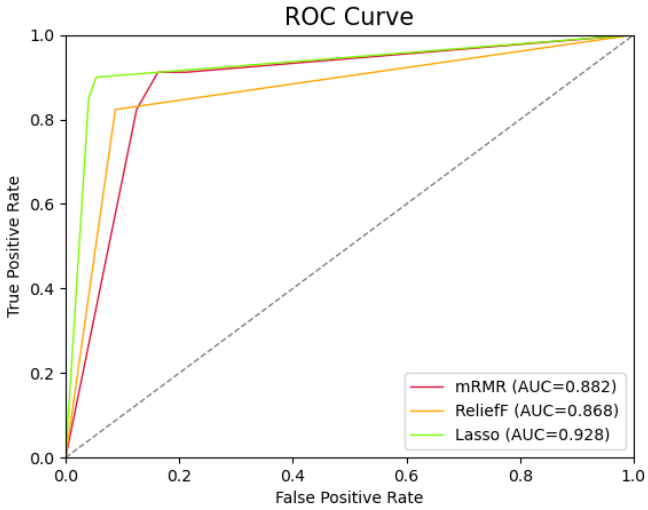

图 5-8 Adaboost 模型 ROC 曲线

在此基础上对比三个模型的 ROC 曲线，明显 Lasso 方法的曲线完全覆盖了使用 mRMR 方法和使用 ReliefF 方法筛选变量建立的模型。从模型的 AUC 值来看也可以得到相同的结果：Lasso 方法筛选变量后的模型 AUC 值为 0.928，是高

于使用 mRMR 方法的 0.882 和使用 Relief 方法的 0.868 的。所以认为在建立 Adaboost 模型进行分类时，使用 Lasso 方法筛选变量是最合适的。

观察使用 Lasso 方法建立的 Adaboost 分类模型在验证集上的情况，从图 5-7 的混淆矩阵中可以看到，模型将 40 个乳腺癌恶性患者中的 36 名患者数据成功识别为了恶性肿瘤，将 4 名恶性肿瘤患者错误识别成了良性肿瘤患者；此外，对于另外 74 名良性乳腺癌患者，模型将其中 70 名患者都成功分类为良性肿瘤，将其中 4 名患者错误划分成为恶性肿瘤。

5.2.3 XGBoost 模型

本文使用使用 R 语言 xgboost 库中的函数 xgboost（）建立随机森林模型。带入上文中 mRMR、ReliefF 以及 Lasso 最终选择的变量，带入 XGBoost 模型，得到三种特征选择方法的 XGBoost 模型的测试集结果（将 R 程序结果带入 python 绘图）如下：

表 5-4 XGBoost 模型评价表

特征选择	Accuracy	Precision	Recall	F1
mRMR	0.9035	0.7143	0.9615	0.8197
ReliefF	0.9473	0.8571	0.9231	0.8889
Lasso	0.8860	0.6857	0.9231	0.7869

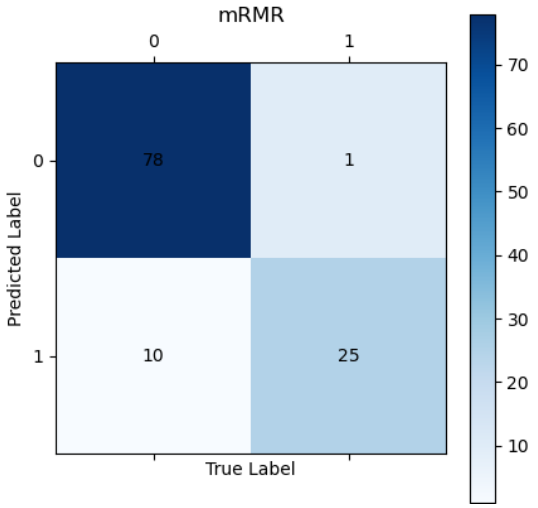

图 5-9 XGBoost（mRMR）混淆矩阵

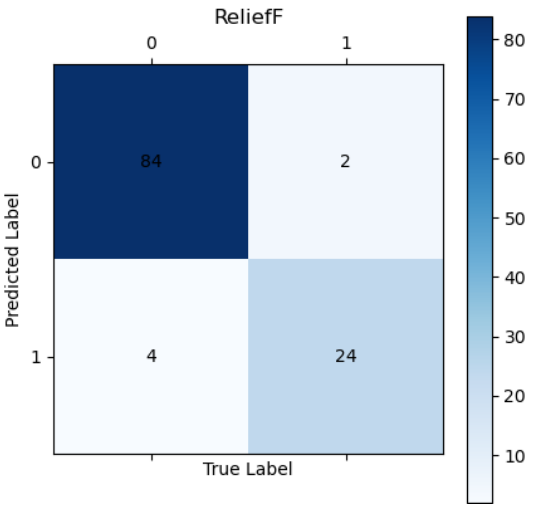

图 5-10 XGBoost（ReliefF）混淆矩阵

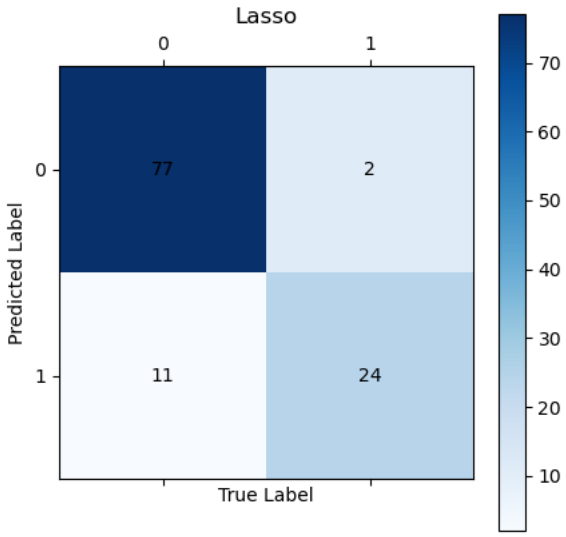

图 5-11 XGBoost（Lasso）混淆矩阵

从表 5-4 中可知，准确率最高的模型是 XGBoost-ReliefF，其准确率达到了 0.9473，表示测试集中有 94.73%的样本被分到了正确的类别之中；对于精确率可以发现，同样是 XGBoost-ReliefF 模型最为精确，且相较于 XGBoost-mRMR 和 XGBoost-Lasso 模型有比较大幅度的提升；对于召回率，最为优秀是 XGBoost-mRMR 模型，其召回率达到了 0.9615；对于 F1 值，最高的是使用 ReliefF 算法选择的变量建立的 XGBoost 模型，且 F1 值达到了 0.889。

为了更好的判断，绘制三个模型的 ROC 曲线与计算对应的 AUC 值如下：

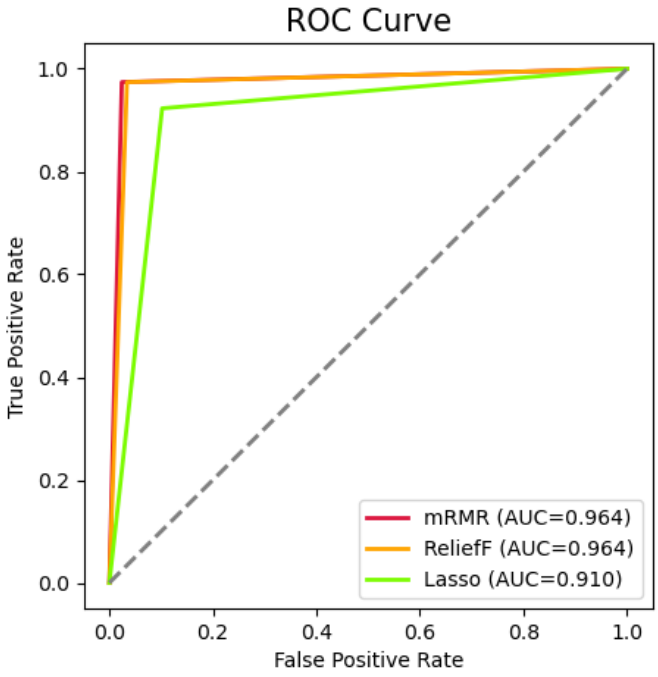

图 5-12 XGBoost 模型 ROC 曲线

从图 5-12 中看到，绿色代表的是使用 Lasso 方法选择变量后的 XGBoost 模

型的 ROC 曲线，而红色和黄色则是分别代表了使用 mRMR 方法和使用 ReliefF 方法选择变量后模型的 ROC 曲线。显然，使用 mRMR 方法和 ReliefF 法的 ROC 曲线明显优于使用 Lasso 方法。从 AUC 值的角度来看，前两者的 AUC 值都是 0.964，而 Lasso 方法的模型 AUC 值为 0.910，同样也是前者更加优秀。结合上文中 XGBoost-ReliefF 法的三个指标准确率、查准率、F1 值都是最高，所以认为 XGBoost 模型中最优秀的是结合 ReliefF 法的 XGBoost-ReliefF 模型。

图 5-10 是模型的混淆矩阵。测试集中共 26 例诊断结果为恶性肿瘤的样本，其中有 24 例样本都经过模型成功分类到了 1 类型，只有 2 例样本被错误分类到了良性肿瘤；测试集中共 88 例良性肿瘤样本有 84 例都被划分到了良性预测结果，有 4 例样本被错误划分到了恶性肿瘤的分类之中。

5.3 模型的对比

5.3.1 最佳模型之间对比

在上文中针对每一个模型内部进行了对比分析，分别在各个模型之中选择了一个最佳模型，分别是：ReliefF-RandomForest 模型、Lasso-Adaboost 模型和 ReliefF-XGBoost 模型。在上一节的基础上，现将不同模型之间进行横向对比并选择适合乳腺癌诊断预测的最佳模型。对于上述三个模型，其统计指标值如下表所示：

表 5-5 不同模型评价表

模型	Accuracy	Precision	Recall	F1
ReliefF-RandomForest	0.9035	1.000	0.7105	0.8308
Lasso-Adaboost	0.9298	0.9000	0.9000	0.9000
ReliefF-XGBoost	0.9473	0.8571	0.9231	0.8889

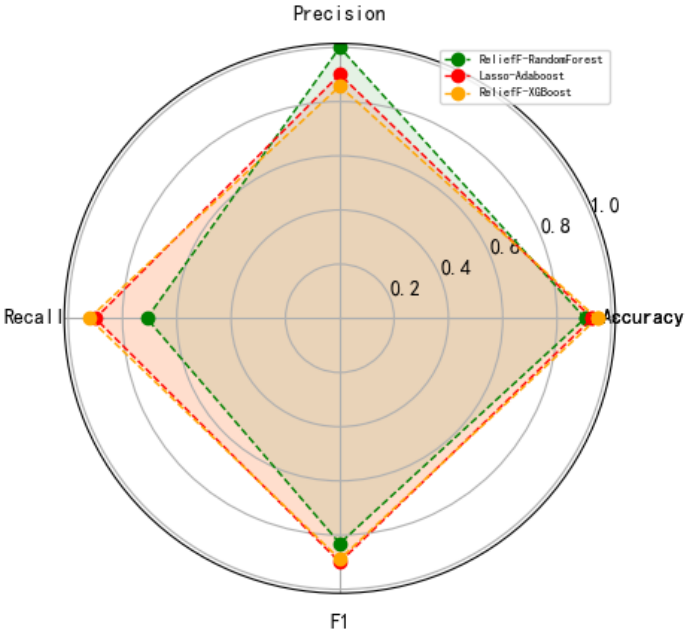

图 5-13 模型性能对比雷达图

从表 5-5 中可知三种模型的各项统计指标。对于准确率，使用 ReliefF-XGBoost 模型预测值最高，达到了 0.9473，表示对于测试集中的样本有 94.73 的样本是被归到了正确的类别中，其次是 Lasso-Adaboost 的 92.98%和 ReliefF-RandomForest 的 90.35%；对查准率，效果最好的模型是 ReliefF-RandomForest，达到了 1，明显优于其他两个模型；对于召回率，使用 ReliefF-XGBoost 模型的效果是最为优秀的，其次是 Adaboost 模型，而随机森林的表现相对差很多；最后是 F1 值，三个模型之间的差距都比较小，其中效果最好的模型是 Lasso 选择变量后的 Adaboost 模型。

从以上结果中可以认为。如果研究人员或医务人员更加注重于诊断的准确率，也就是将良性肿瘤患者和恶性肿瘤患者更好的区分开来，使用 ReliefF-XGBoost 可以获得更加好的效果，结合使用 ReliefF 筛选得到的变量来看，此时应该更加关注乳腺癌患者癌细胞核的最差半径（radius worst）、平均半径（radius mean）、最差凹点数量（concave points worst）、平均凹点数量（concave points mean）等生理指标再结合 XGBoost 模型进行判断；如果研究人员或医务人员的需求是尽可能准确地发现恶性肿瘤患者，那么使用 ReliefF-RandomForest 模型可以获得更好的效果。

为了对三个模型更好的作出综合性判断，绘制模型之间的 ROC 曲线和计算其分别的 AUC 值如下图所示：

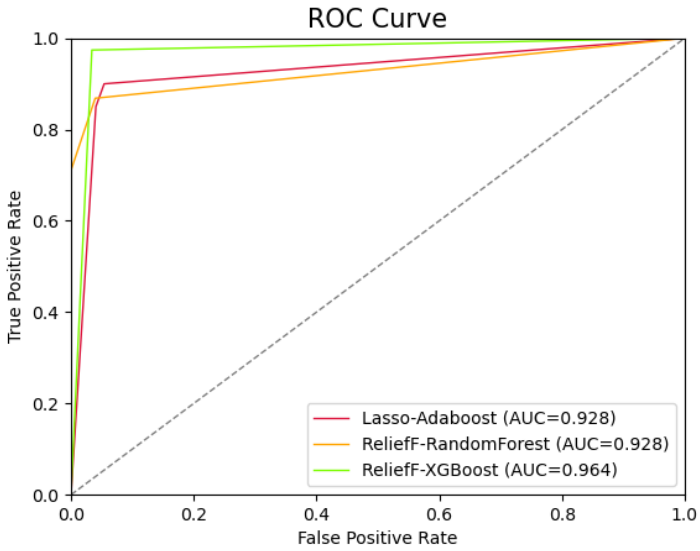

图 5-14 模型对比 ROC 曲线图

图中红、黄、绿色的线分别表示 Lasso-Adaboost、ReliefF-RandomForest 以及 ReliefF-XGBoost 模型的 ROC 曲线。绿色曲线明显优于另外两个模型，而 Adaboost 模型与随机森林模型表现接近，两者分别在不同的阈值上各有优异的表现，在实际建模中可以根据阈值的不同选择适合的模型进行乳腺癌患者的诊断预测。

在此基础上，使用 ROC 曲线下面积的大小即 AUC 指来评估模型的整体性能。其中表现最好的模型是 ReliefF-XGBoost，其 AUC 值达到了 0.964；其次是 Lasso-Adaboost 和 ReliefF-RandomForest，其 AUC 值都为 0.928。由此可以认为，在考虑模型的综合性能时，可以首先选择使用。

5.3.2 最佳模型与原模型对比

以上的模型都是在建立在使用 mRMR、ReliefF 以及 Lasso 方法对变量进行筛选后建立的模型。为验证通过变量筛选后的建立的模型是要优于全部变量建立的模型，对全部共 30 个自变量分别建立 RandomForest、Adaboost 以及 XGBoost 模型，其模型性能指标如下表所示：

表 5-6 未筛选变量模型评价表

模型	Accuracy	Precision	Recall	F1	AUC 值
RandomForest	0.8509	0.7179	0.8235	0.7671	0.897
Adaboost	0.8947	0.7750	0.9118	0.8378	0.907
XGBoost	0.8684	0.6410	0.9615	0.9126	0.901

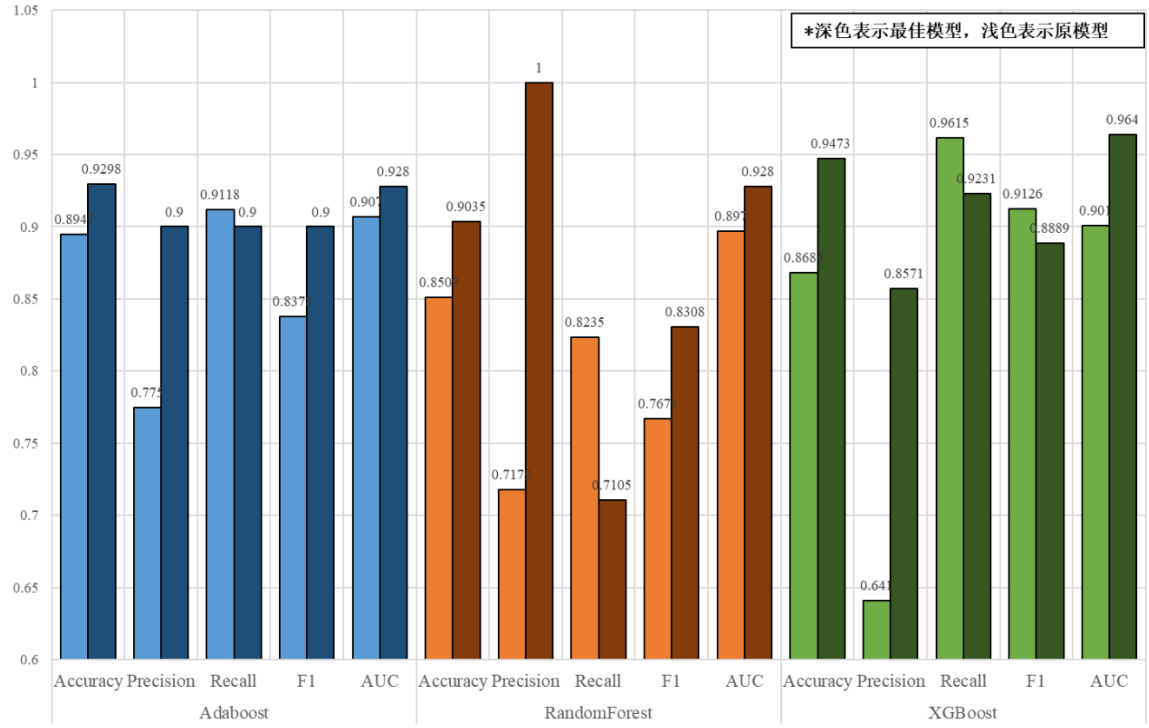

图 5-15 模型指标对比图

从评价表以及对比图中可以看出，对于 Adaboost 模型，未筛选变量建立模型的准确率、查准率和 F1 值都低于使用最佳方法筛选变量后建立的 Adaboost 模型，其中精确率和 F1 值都有较大程度的提升。而原模型的召回率略高于最佳模型的召回率；对于 RandomForest 模型，同样是最佳模型的准确率、查准率和 F1 值高于原模型，且最佳模型的准确率提升效果非常明显。而原模型的召回率是要优于最佳模型的；对于 XGBoost 模型，最佳模型的准确率、查准率都要高于原模型，其中准确率的提升是最为明显的。而原模型的特点在于其召回率和 F1 值，其两个评价指标略微高于使用 ReliefF 方法筛选变量后的 XGBoost 模型。

从整体上来看，不筛选变量直接建立的 Adaboost、RandomForest 和 XGBoost 模型的 AUC 值分别是 0.907、0.897 以及 0.901；在分别对三种模型使用 ReliefF 法或者 Lasso 法筛选变量后，其建立的模型 AUC 值提升为 0.928、0.928 以及 0.964。由此看来，使用筛选后的变量建立的模型在数据集上拥有更好的性能，也就是说上述的研究过程是有意义的。

6 结论与展望

6.1 研究结论

癌症作为一种骇人听闻的疾病，已经夺走了无数患者的生命。而乳腺癌是众多癌症中最为常见的恶性肿瘤之一，其对我国乃至全世界的女性的身体健康和生命安全造成了巨大的威胁。随着经济社会的发展以及科学技术的进步，对于癌症的研究也在不断的更新，越来越多高新技术的设备参与到对乳腺癌的诊断和预测之中。得益于这些高清的影像设备，医务人员和研究人员可以观察非常细微的乳腺癌细胞核的各种形态结构并且将其有效的量化；再加上统计学科理论的不断发展与硬件设备的不断升级，使得将统计学习中的集成学习方法与观察到的乳腺癌细胞核的各项生理指标之间相互结合，并以此为基础建立乳腺癌患者诊断和预测相关的集成学习模型。而如何去建立一个科学、有效、实用价值高的模型已经是现在急需解决的重要论题。

在此基础上，本文在选取了 UCI 数据库中包含的威斯康星州（诊断）乳腺癌数据集并使用集成学习模型中的随机森林、Adaboost、XGBoost 模型建立乳腺癌患者诊断预测模型，并在数据集特征选择和模型评估上提出了建议。主要研究结论如下：

（1）对 UCI 数据库中的威斯康星州（诊断）乳腺癌数据集整体的分布情况进行了分析，并统计其中各种生理指标特征对诊断结果的影响。通过绘制相关统计图像给出变量的描述性结果，并在后续建模中对数据集的离群样本进行了有效剔除。最终得到有效的数据集 560 例。

（2）在特征选择中，分别使用过滤式算法中的 mRMR 算法、封装式算法中的 ReliefF 算法以及嵌入式算法中的 Lasso 算法对数据集中 30 个特征进行特征选择。三种方法分别选择了权重靠前的 8、9、8 个变量，且三种变量选择方法之间的重复率较低，方法之间具有对比价值。

（3）使用随机森林、Adaboost、XGBoost 模型结合三种特征选择方法建立了乳腺癌诊断预测模型。在模型内部之间，认为随机森林模型之中结合 ReliefF 特征选择算法效果最好；Adaboost 模型结合 Lasso 变量选择算法效果最好；XGBoost 模型结合 ReliefF 变量选择算法效果最好。三者模型的诊断准确率分别达到了 0.9035,0.9298 以及 0.9473。

（4）对于三种模型结合最优特征之间的比较。得到结论：如果医务人员和研究人员更加注重诊断的准确率，应使用 ReliefF-XGBoost 模型来获得更好的效果；如果需求是尽可能发现恶性肿瘤患者，那么使用 ReliefF-RandomForest 模型是更好的选择。最后综合对比三个模型的 AUC 值，认为综合水平 ReliefF-XGBoost 模型最好，其 AUC 值分别为 0.964,0.928 和 0.928。

6.2 研究展望

本文在 UCI 数据库威斯康星州（诊断）乳腺癌数据集的基础上建立了共计 9 个集成学习模型用于乳腺癌患者的诊断预测。虽然在诊断预测上得到一定的结论，但无论是研究过程还是研究结果中还存在着一些缺点和不足，还需要再深入探讨一些问题。

（1）首先是使用的数据集样本数量较少，现在的计算设备已经支持对大容量数据集的计算，使用样本容量更高的数据集可以训练得到更加准确的模型，对于乳腺癌患者的诊断预测有更好的效果。

（2）其次是本文仅使用了集成学习中的 Adaboost、XGBoost 以及随机森林这三个经典模型。随着学科的发展还有更多高性能高速度的模型可以用于本文的研究。在后续研究中还可以尝试使用更多新模型进行诊断预测，并与原始模型对比其效果。

（3）本文在使用软件训练模型的时候使用的默认参数较多，只是在自身经验的基础上进行了小范围的调参。在统计学习中还有很多专门用于参数调整的方法，如网格搜索法等，在后续研究中还可以对模型的参数进行合理的调整，以达到更好的研究效果。

参考文献

[1] 李勇, 陈思萱, 贾海, 王霞. 基于 C-AdaBoost 模型的乳腺癌预测研究[J]. 计算机工程与科学,2020, 42(08): 1414-1422.

[2] 武莉茹. 基于多组学数据的乳腺癌生存期预测算法研究[D]. 西安电子科技大学, 2020.

[3] 李星睿. 基于机器学习的乳腺癌诊断及再分型研究[D]. 北京工业大学, 2020.

[4] 刁继尧. 基于机器学习的乳腺癌风险分析与预测研究[D]. 南京邮电大学, 2019.

[5] 齐惠颖, 江雨荷. 基于多组学数据融合构建乳腺癌生存预测模型[J]. 数据分析与知识发现, 2019, 3(08): 88-93.

[6] 赖胜圣, 刘虔铖, 余丽玲, 刘文平, 杨蕊梦, 金浩宇. 基于 SFS-SVM 的乳腺癌预测模型的构建[J]. 中国医学物理学杂志, 2019, 36(07): 826-829.

[7] 原瑞霞. 基于 GBD 大数据分析与预测中国女性乳腺癌发病与死亡趋势的研究[D]. 武汉大学, 2018.

[8] 王玲玲. 年轻女性乳腺癌生存预测列线图的构建及验证[D]. 浙江大学, 2021.

[9] 曹莹, 苗启广, 刘家辰, 高琳. AdaBoost 算法研究进展与展望[J]. 自动化学报, 2013, 39(06): 745-758.

[10] 常甜甜. 支持向量机学习算法若干问题的研究[D]. 西安电子科技大学, 2010

[11] 钟熙, 孙祥娥. 基于 Kmeans++聚类的朴素贝叶斯集成方法研究[J]. 计算机科学, 2019, 46(S1): 439-441+451.

[12] 莫赞, 盖彦蓉, 樊冠龙. 基于 GAN-AdaBoost-DT 不平衡分类算法的信用卡欺诈分类[J]. 计算机应用, 2019, 39(02): 618-622.

[13] 陈启伟, 王伟, 马迪, 毛伟. 基于 Ext-GBDT 集成的类别不平衡信用评分模型[J]. 计算机应用研究, 2018, 35(02): 421-427.

[14] 樊鹏. 基于优化的 xgboost-LMT 模型的供应商信用评价研究[D]. 广东工业大学, 2016.

[15] 孙茂伟, 杨慧中. 基于改进 Bagging 算法的高斯过程集成软测量建模[J]. 化工学报, 2016, 67(04): 1386-1391.

[16] 付忠良. 关于 AdaBoost 有效性的分析[J]. 计算机研究与发展, 2008(10): 1747-1755.

[17] 周志华. 机器学习[M]. 北京：清华大学出版社, 2016.

[18] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2019.

[19] 顾翔元. 基于信息度量的特征选择算法研究及应用[D]. 天津大学, 2020.

[20] 杜利敏. 面向不平衡数据的特征选择与半监督分类算法研究[D]. 西南交通大学, 2017.

[21] 王爱国. 微阵列基因表达数据的特征分析方法研究[D]. 合肥工业大学, 2015.

[22] 杨虎, 杨玥含. 金融大数据统计方法与实证[M]. 北京: 科学出版社, 2016.

[23] 刘博菲. 基于集成学习的 5G 潜在用户识别[D]. 大连理工大学, 2021.

[24] 蒋锋, 张婷, 周琰玲. 基于 Lasso-GRNN 神经网络模型的地方财政收入预测[J]. 统计与决策, 2018, 34(19):91-94.

[25] 菅小艳, 韩素青, 崔彩霞. 不平衡数据集上的 Relief 特征选择算法[J]. 数据采集与处理, 2016, 31(04):838-844.

[26] 王露, 龚光红. 基于 ReliefF+mRMR 特征降维算法的多特征遥感图像分类[J]. 中国体视学与图像分析, 2014, 19(03): 250-257.

[27] 李晓岚. 基于 Relief 特征选择算法的研究与应用[D]. 大连理工大学, 2013.

[28] 孔英会, 景美丽. 基于混淆矩阵和集成学习的分类方法研究[J]. 计算机工程与科学, 2012, 34(06): 111-117.

[29] 蒋尧西, 彭松. 乳腺癌的影像学诊断现状与进展[J]. 中国医药指南, 2018, 16(29): 23-24.

[30] 陆兴练. 乳腺癌诊断技术的研究进展[J]. 大医生, 2022, 7(18): 123-126.

[31] Freddie B, Jacques F, Isabelle S, et al. Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries[J]. CA: A Cancer Journal for Clinicians, 2018, 68(6): 394-424.

[32] Salod Z, Singh Y. Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol[J]. Journal of Public Health Research, 2019, 8(3). 112-118.

[33] Youness K, Mohamed B. Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification[J]. International Conference on Computer Science, 2018, 42(3): 232-239.

[34] Sara A, Heyam A. On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context[J]. IEEE Access, 2019, 2927080: 1-13.

[35] Bichen Z, Sang Y, Sarah S. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms[J]. Expert Systems With Applications, 2014, 41(4): 7-14.

[36] Dhahri H, Al M, Mahmood A, et al. Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms[J]. Journal of Healthcare Engineering, 2019: 4253641.

[37] Hiba A, Hajar M, Hassan M, et al. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis[J]. Procedia Computer Science, 2016, 83(C): 1064-1069.

[38] Moloud A, Mariam Z, Xujuan Z, et al. A new nested ensemble technique for automated diagnosis of breast cancer[J]. Pattern Recognition Letters, 2020, 132: 1-11.

[39] Rokach, Lior. Ensemble-based classifiers[J]. The Artificial Intelligence Review, 2010, 33(1-2): 1-39.

[40] Xibin D, Zhiwen Y, Wenming C, et al. A survey on ensemble learning[J]. Frontiers of

Computer Science, 2020, 14(2): 1-17.

[41] Stefano M, Bruno C, Cesare F. Parallelizing AdaBoost by weights dynamics[J]. Computational Statistics and Data Analysis, 2006, 51(5): 2487-2498.

[42] Chun-Xia Z, Jiang-She Z. A local boosting algorithm for solving classification problems[J]. Computational Statistics and Data Analysis, 2008, 52(4): 1928-1941.

[43] Diao R, Chao F, Peng T, et al. Feature selection inspired classifier ensemble reduction.[J]. IEEE Transactions on Cybernetics, 2014, 44(8): 1-10.

[44] Yu Z, Wang D, Zhao Z, Chen C, et al. Hybrid Incremental Ensemble Learning for Noisy Real-World Data Classification.[J]. IEEE Transactions on Cybernetics, 2017, 49(2): 1-14.

[45] Fatima N, Liu L, Hong S, et al. Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis[J]. IEEE Access, 2020, 8(4): 1-17.

[46] Hanchuan P, Fuhui L, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005: 27(8), 1226–1238.

[47] Roobini R, Naomi F. Performance Analysis of Different Classifiers in Prediction of Breast Cancer[J]. Indian Journal of Science and Technology, 2019, 12(8): 1-6.

[48] Street N, Wolberg H ,Mangasarian L. Nuclear feature extraction for breast tumor diagnosis[J]. Univ. of Wisconsin/Madison (United States), 1993, 1905: 861-870.

[49] Tianqi C, Carlos G. XGBoost: A Scalable Tree Boosting System.[J]. ACM Transactions on Intelligent Systems and Technology, 2016, 1: 785-794.

[50] Taleghamar H, MoghadasDastjerdi H, Czarnota G, et al. Characterizing intra-tumor regions on quantitative ultrasound parametric images to predict breast cancer response to chemotherapy at pre-treatment. [J]. Scientific Reports,2021,11(1): 14865.

[51] Daping Y, Zhidong L, Chongyu S, et al. Copy number variation in plasma as a tool for lung cancer prediction using Extreme Gradient Boosting (XGBoost) classifier [J]. Thoracic Cancer, 2020, 11(1): 1-8.

[52] Abdu D, Awad A. A Comparative analysis study of lung cancer detection and relapse prediction using XGBoost classifier[J]. IOP Conference Series: Materials Science and Engineering, 2021, 1076(1): 1-14.

[53] Jianxing W, Piyun C, Chiahung L, et al. Breast Benign and Malignant Tumors Rapidly Screening by ARFI-VTI Elastography and Random Decision Forests Based Classifier[J]. IEEE Access, 2020, 8: 54019-54034.

[54] Mengmeng S, Tao D, Xu-Qing T, et al. An Efficient Mixed-Model for Screening Differentially Expressed Genes of Breast Cancer Based on LR-RF[J]. IEEE/ACM Transactions on

Computational Biology and Bioinformatics (TCBB), 2019, 16(1): 1-8.

[55] Quist J, Taylor L, Staaf J, et al. Random Forest Modelling of High-Dimensional Mixed-Type Data for Breast Cancer Classification[J]. Cancers, 2021, 13(5): 1-15.

[56] Tseng C, Shieh C, Yujie H, et al. Using LASSO regression based SVM classification to improve the predictive performance of radiation-induced pneumonitis complication in breast cancer[J]. Journal of the Chinese Institute of Engineers, 2018, 41(8): 1-7.

[57] Ravi S, Nithish S, Nithish M, et al. Breast Cancer Prediction using Decision Tree[J]. Journal of Physics: Conference Series, 2021, 1916(1): 1-8.

[58] Mahesh V, Mohan M. An ensemble classification based approach for breast cancer prediction[J]. IOP Conference Series: Materials Science and Engineering, 2021, 1065(1): 1-10.

[59] Naji A, Filali E, Aarika K, et al. Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis[J]. Procedia Computer Science, 2021, 191: 487-492.