ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 6,Nov. 2021,pp. 1237－1249
DOI:10. 16337/j. 1004-9037. 2021. 06. 017

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于惩罚逻辑回归的乳腺癌预测

胡雪梅[1,2]，谢　英[1,2]，蒋慧凤[3]

(1.重庆工商大学数学与统计学院,重庆 400067; 2.重庆工商大学经济社会应用统计重庆市重点实验室，重庆 400067; 3.重庆工商大学长江上游经济研究中心,重庆 400067)

摘　要：本文采用惩罚逻辑回归方法,利用威斯康星大学的乳腺癌数据对乳腺肿瘤进行预测。首先选取与乳腺癌相关的10个指标作为自变量,接着采用逻辑回归、LASSO惩罚逻辑回归、L_2惩罚逻辑回归和弹性网惩罚逻辑回归作为分类器,利用75%的数据集作为训练集建立模型,最后利用25%的测试集、混淆矩阵和ROC曲线评估不同模型的预测精度。结果表明,LASSO惩罚逻辑回归的预测表现最好,预测精度达到97.18%;弹性网惩罚逻辑回归的预测表现随着α的增大发生变化,特别当$\alpha=0.9$时,预测精度达到97.18%,与LASSO惩罚逻辑回归的预测表现一样好;L_2惩罚逻辑回归的预测表现排第3,逻辑回归表现最差。因此,在乳腺肿瘤诊断中可借助LASSO惩罚逻辑回归和弹性网惩罚逻辑回归提高诊断精度。

关键词：乳腺癌;逻辑回归;LASSO惩罚逻辑回归;L_2惩罚逻辑回归;弹性网惩罚逻辑回归

中图分类号：TP181;R737.9　　**文献标志码**：A

Prediction of Breast Cancer Based on Penalized Logistic Regression

HU Xuemei[1,2], XIE Ying[1,2], JIANG Huifeng[3]

(1.School of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing 400067, China; 2. Chongqing Key Laboratory of Economic and Social Applied Statistics, Chongqing Technology and Business University, Chongqing 400067, China; 3. Research Center for Economy of Upper Reaches of the Yangtse River, Chongqing Technology and Business University, Chongqing 400067, China)

Abstract: In this paper, we mainly apply the breast cancer data from University of Wisconsin System to predict breast cancer using penalized logistic regression. Firstly, the ten indicators related to breast cancer are selected as the predictor variables. Then, logistic regression, the LASSO penalized logistic regression, the L_2 penalized logistic regression and the elastic net penalized logistic regression are used as the four classifiers. 75% of the data set is used as the training set to build models. Finally, 25% test set, a confusion matrix and a ROC curve are used to evaluate their prediction accuracy. The results show that the LASSO penalized logistic regression performs best, whose prediction accuracy reaches 97.18%. The prediction performance of the elastic net penalized logistic regression changes with the increase of α, especially when $\alpha=0.9$, the corresponding prediction accuracy is 97.18%, as good as that of LASSO

基金项目:重庆市第五批高等学校优秀人才支持计划(68021900601)资助项目;重庆市科委基础研究与前沿探索一般项目(cstc. 2018jcyjA2073)资助项目;重庆市统计学研究生导师团队(yds183002)资助项目;重庆市教委科学技术研究计划重大项目(KJZD-M202100801)资助项目;重庆市社会科学规划项目(2019WT59)资助项目;社会经济应用统计重庆市重点实验室平台开放项目(KFJJ2018066)资助项目;重庆工商大学数理统计团队(ZDPTTD201906)资助项目。

收稿日期:2020-12-03;**修订日期**:2021-05-26

penalized logistic regression. The L_2 penalized logistic regression ranks the third and logistic regression performs the worst in prediction performance. Therefore, for the diagnosis of breast tumors, doctors can apply the LASSO penalized logistic regression and the elastic net penalized logistic regression to improve the diagnostic accuracy.

Key words: breast cancer; logistic regression; the LASSO penalized logistic regression; the L_2 penalized logistic regression; the elastic net penalized logistic regression

引　言

乳腺癌目前是全球排列第一的肿瘤疾病,居女性恶性肿瘤发病率之首。2012年,170万名女性被诊断出患有乳腺癌,约52.2万名患者死亡。根据世界卫生组织国际癌症研究机构2018年发布的报告表示,乳腺癌病患数在各种癌症种类中排名第5,有将近210万女性乳腺癌新发病例,大约62.7万人死亡,而且随着人们饮食习惯和生活方式等方面的改变,乳腺癌恶性肿瘤的发病率会呈递增趋势。与美国相比,中国的乳腺癌患者生存率偏低,这与中国人口基数过多、有经验的影像科医生缺乏、难以实施大范围早期筛查有关。癌症早期被发现时容易治愈,因此准确筛查癌症早期患者至关重要。

目前乳腺癌的诊断方法主要有X射线诊断[1]、CT扫描、临床触诊、超声波显像检查、核磁共振成像术、近红外线扫描、钼靶和细针穿刺细胞病理学检查等。乳房X光是一种测试方法,但也存在缺点,经常会导致假阳性结果,导致不必要的活检和手术,在乳房X光片上看到具有可疑的异常细胞时,需要通过手术去除异常细胞,然而大部分肿瘤在手术中被发现是良性的,这意味着每年都有数千名妇女无端承受手术痛苦、昂贵费用和术后疤痕等。传统的诊断方法可能会由于低劣的图像质量以及临床医生的视觉疲劳或疏忽等导致漏诊或误诊。现在可以借助计算机技术辅助诊断帮助医生和乳腺癌患者。

深度学习是机器学习中一个非常接近人工智能(Artificial intelligence,AI)的领域,具有强大的能力和灵活性,能将大千世界表示为嵌套的层次概念体系:(1)无监督学习用于每一层网络的Pre-train;(2)每次用无监督学习只训练一层,将其训练结果作为其高一层的输入;(3)用监督学习去调整所有层。通过组合低层特征形成更加抽象的高层表示属性类别或特征,发现数据的分布式特征表示,建立模拟人脑进行分析学习的神经网络,模仿人脑机制解释图像、声音和文本等数据。深度学习目前在计算机视觉、图像处理、语音识别和自然语言处理等领域应用很成功,在病理成像等医学图像模式的分类和检测方面表现不俗。机器学习涉及大量统计理论,与统计推断联系密切,故也称为统计学习。机器学习通过设计自动"学习"算法,从数据中自动分析获得规律并对未知数据进行预测,已成功应用于很多领域。例如,从检测信用卡交易欺诈的数据挖掘程序,到获取户阅读兴趣的信息过滤系统,再到能在高速公路上自动行驶的汽车等。机器学习[2-6]还可以有效辅助医生诊断,帮助医生筛查乳腺癌等病症。例如,Huang等[7]开发了计算机辅助诊断(CAD)系统,用支持向量机对118个乳腺肿瘤作良性和恶性分类,分类效果很好;Montazeri等[8]利用朴素贝叶斯、树随机森林、支持向量机和多层感知器等机器学习方法结合10折交叉验证预测不同乳腺癌的生存率,并用准确度、灵敏度和ROC曲线下的面积(Area under curve, AUC)等评价几种方法,得出树随机森林效果更好(96%、96%、93%);Xia等[9]利用卷积神经网络对不同类型的乳腺癌细胞Mueller偏振成像图进行分类,准确率达到了88.3%等。

支持向量机只能预测类指标,不能提供类概率估计。神经网络因为噪音累积、非平稳特征和复杂维数在学习方式上有限制等原因导致预测精度不稳定,时高时低。逻辑回归不仅能预测类指标,还能得到类概率估计,并且能获得较高的预测精度。例如,胡雪梅等[10]比较了逻辑回归、支持向量机、人工

神经网络、ELMAN神经网络和基于五类统计指标的一阶自回归逻辑回归模型的预测表现,发现逻辑回归模型预测表现最好。而惩罚逻辑回归是对逻辑回归引入惩罚函数,选取重要变量作分类预测,可以进一步提高模型的拟合能力与预测精度。因此,本文采用惩罚逻辑回归来预测乳腺癌肿瘤是良性还是恶性问题。

近年来,人们提出了不同惩罚逻辑回归作分类与预测研究。例如,Park等[11]采用L_2惩罚逻辑回归探测基因的交互影响,可以识别交互结构和重要因子,并且得到合理的预测精度;Meier等[12]提出了组LASSO惩罚逻辑回归及其衍生模型:组LASSO-Ridge混合惩罚逻辑回归和组LASSO-MLE混合惩罚逻辑回归,发展了坐标下降算法,通过DNA剪接位点预测研究发现组LASSO惩罚逻辑回归是最佳预测模型;Friedman等[13]发展了具有LASSO惩罚、L_2惩罚、弹性网惩罚的广义线性模型(包含线性模型、逻辑回归和多项式回归)及其坐标下降算法,设计了glmnet程序包;Ayers等[14]在基因关联研究中选择单基因多态性(Single nucleotide polymorphisms, SNPs)作为预测变量,利用5类惩罚逻辑回归:弹性网惩罚、L_2惩罚、LASSO惩罚、MCP惩罚和正态指数伽玛分布(Normal exponential Gamma, NEG)惩罚作为两类分类器预测灵敏度和特异度;Breheny等[15]对具有分组预测变量的非凸惩罚线性回归和逻辑回归发展了组坐标下降算法,重点介绍了组LASSO、组SCAD和组MCP惩罚逻辑回归及实际数据预测分析;Li等[16]建立了基于弹性网正则化的相关性逻辑回归多标签图像分类模型,对标签间的成对相关性建模以提高多标签分类(Multi-label classification, MLC)的有效性,充分利用特征选取和标签相关的稀疏性,采用弹性网正则化进一步提高MLC的性能,并用4个多标签图像数据集证实模型的有效性;Sari等[17]分别采用LASSO惩罚逻辑回归和支持向量机作信用评分分析,两种方法的分类精度基本相同(79.2%和79.94%),但SVM的分类效果更稳定(灵敏度:79.80%和72.62%);Münch等[18]提出了分组弹性网逻辑回归,能够改进特征选取和分类表现,用3组癌症基因模拟证实分类性能和特征选择得到了改进;Garcia-Carretero等[19]发展了逐步逻辑回归、LASSO惩罚逻辑回归和弹性网惩罚逻辑回归预测高血压肥胖人群维生素D缺乏症,并采用灵敏度、特异度、误分类率和AUC评价3种方法的预测表现,结果表明,LASSO惩罚逻辑回归和弹性网惩罚逻辑回归得到的AUC显著高于逐步逻辑回归,能更准确地预测维生素D缺乏症等。

诊断乳腺肿瘤是良性还是恶性本质上是一个二分类问题。本文先从威斯康星州大学的乳腺癌数据选取10个指标作为预测变量,将诊断结果(良性:$Y=0$和恶性:$Y=1$)作为响应变量,建立4个分类器:逻辑回归和三类惩罚逻辑回归(LASSO惩罚逻辑回归,L_2惩罚逻辑回归和弹性网惩罚逻辑回归),用训练数据学习分类模型,再用测试集中的观察值X预测响应变量Y的发生概率,接着选取最佳阈值c确定预测值$\hat{Y}$,最后由所有样本和预测结果得混淆矩阵、灵敏度和特异度,绘制ROC曲线得到AUC评价不同分类模型的预测精度。LASSO惩罚逻辑回归、L_2惩罚逻辑回归和弹性网惩罚逻辑回归分别是对逻辑回归施加L_1惩罚、L_2惩罚和弹性网惩罚(介于L_1惩罚与L_2惩罚之间)。比较4个分类器的预测结果可知,LASSO惩罚逻辑回归的预测表现最好,预测精度达到97.18%;弹性网惩罚逻辑回归的预测表现随着α的增大发生变化,特别当$\alpha=0.9$时,弹性网惩罚逻辑回归的预测精度达到97.18%,与LASSO惩罚逻辑回归的预测表现一样好;L_2惩罚逻辑回归的预测表现排第3;逻辑回归的预测表现最差。因此,本文提出的LASSO惩罚逻辑回归方法和弹性网惩罚逻辑回归方法可以有效预测乳腺癌,提高诊断精度。

1 数据来源与预处理

本文分析569例乳腺癌患者诊断数据,包含10个特征。数据来自威斯康星大学UCI网站(https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29)。原始数据是569行

32列的表格:第1列是患者的ID编号,第2列是诊断结果:良性(Benign)或恶性(Malignant),第3～32列是10个特征的3种值:前10列分别对应10个特征中每个特征的平均值,中间10列分别对应10个特征中每个特征值的标准差,后10列分别对应10个特征中每个特征值的最大值(即特征值前3名的平均值,可减弱计算误差带来的影响),其中特征值描述样本图像中细胞核的形态特征,主要通过乳腺肿块的细针穿刺(Fine needle aspiration, FNA)[20]数字化图像计算得到。从诊断结果看,在569例乳腺癌患者中,良性357例,恶性212例,良性占比62.7%,恶性占比37.3%,两个类不算失衡(一般两个类的比值为9:1表示失衡,比值为99:1表示严重失衡)。本文选取数据集中细胞核特征值的平均值(均值体现样本细胞核的总体形态特征)进行分析,表1详细介绍了数据涉及的预测变量。

表1 10个预测变量

Table 1 Ten prediction variables

变量名称	变量解释
Radius (半径,X_1)	半径表示细胞核中心到边界的距离,恶性肿瘤细胞形态及大小不一致,比正常细胞和良性肿瘤细胞大,细胞核的体积增大
Texure (纹理,X_2)	良性乳腺肿瘤的边界光滑,呈现椭圆状;恶性乳腺肿瘤的边界粗糙,呈现蟹足状,形状不规则。数字化图像明暗像素分布越不均匀,说明肿瘤细胞核的边界纹理越粗糙,呈现恶性肿瘤的可能性越大
Perimeter (周长,X_3)	由于恶性肿瘤细胞形态及大小不一致,比正常细胞和良性肿瘤细胞大,可明显观测到恶性肿瘤细胞核的周长比良性或正常细胞核大
Area (面积,X_4)	恶性肿瘤细胞核的面积比良性或正常细胞核大
Smoothness (平滑度,X_5)	细胞核平滑度即似圆度,半径长度的局部变化越大,似圆度越低;细胞核的边缘越光滑,半径长度的局部变化越小,属于良性肿瘤的可能性越大;细胞核边缘越不规则的,半径长度的局部变化越大,属于恶性肿瘤的可能性越大
Compactness (紧密程度,X_6)	细胞核的紧密程度,Compactness$=$ Perimeter^2/Area $-$ 1.0
Concavity (凹度,X_7)	细胞核的凹度表示细胞核轮廓的凹陷程度,恶性肿瘤形状不规则,细胞核轮廓存在凹点,且凹陷程度比正常细胞大
Concave point (凹点,X_8)	细胞核凹点表示细胞核边缘出现的凹痕数量. 恶性肿瘤形状不规则,凹痕数量比正常细胞核或良性肿瘤多
Symmetry (对称性,X_9)	恶性肿瘤细胞核有高度异质性,形态不规则,容易出现不对称情形
Fractal dimension (分形维数,X_{10})	分形维数是一个描述分形对空间填充程度统计量,而相似维数 $D=\lg N/\lg r$ 是计算分形维数一种方法,其中N为组合成原来的图形需要的变换后的图形的份数,r代表图形缩小比例。恶性肿瘤的形状不规则,纹理粗糙且平滑度低,其分形维数略大于良性肿瘤的分形维数

表1中每个指标分别从不同方面刻画肿瘤细胞核的特征,这些特征有助于医生诊断肿瘤是恶性还是良性。表2列举了10个预测变量的重要描述统计量。

由表2可看出,细胞核的半径、纹理、周长和面积的最小值和最大值相差较大,说明良性与恶性肿瘤细胞有明显差异;细胞核的半径和纹理两个变量的均值和中位数比较接近,且与方差和标准差也比较接近,说明数据波动不大;细胞核的周长和面积两个变量的方差及标准差都比较大,数据波动大。其余6个指标的方差和标准差都非常小,说明数值波动小,比较稳定,且它们的均值和中位数相当接近,指标值在最小值和最大值之间的波动也很小。表2中10个预测变量存在相关关系,这里借助corrplot包对相关系数输出结果作可视化处理,得到如图1所示的相关系数矩阵。

表2 10个预测变量的描述统计量

Table 2 Descriptive statistics for 10 prediction variables

变量	最小值	最大值	均值	中位数	方差	标准差
X_1	6.981	28.110	14.218	13.430	12.496 97	3.535 11
X_2	9.71	39.28	18.93	18.61	16.817 22	4.100 88
X_3	43.79	188.50	92.61	86.49	593.987 68	24.371 86
X_4	143.5	2 499.0	663.2	551.1	123 574.0	351.530 94
X_5	0.062 51	0.144 70	0.096 79	0.096 86	0.000 191	0.013 82
X_6	0.019 38	0.345 40	0.105 78	0.094 45	0.002 979	0.054 58
X_7	0	0.426 80	0.092 04	0.066 51	0.006 641	0.081 49
X_8	0	0.201 20	0.050 68	0.037 00	0.001 562	0.039 52
X_9	0.116 7	0.304 0	0.183 2	0.181 2	0.000 799	0.028 27
X_{10}	0.049 96	0.097 44	0.062 79	0.061 54	0.000 050 5	0.007 11

图1中圆圈越大代表相关性(包括正相关和负相关)越强。观察图1可知,细胞核的紧密程度与凹度和凹点的相关系数分别为0.89和0.84,具有强相关性;细胞核的凹度与凹点的相关系数为0.92,也具有强相关性;细胞核的周长、面积与半径的相关系数接近1,具有更强的相关性;而数字化图像中细胞核之间的似圆度有差异,不能完全确定其相关关系,故分成两个特征指标计算。表3列出了预测变量与响应变量之间的相关系数。

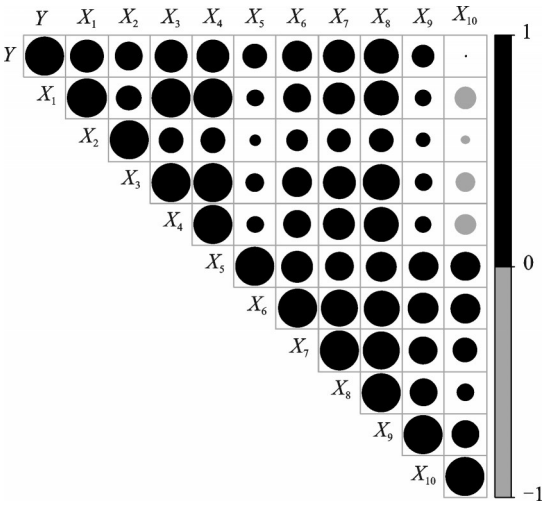

图1 10个预测变量的相关系数矩阵图

Fig.1 Correlation coefficient matrix for ten prediction variables

显然,这些预测指标存在相关性,考虑到医学诊断的特殊性,本文没有对这些相关指标作进一步处理。本文数据预处理如下:先将患者的诊断结果B和M重新编码为0和1(0表示良性,1表示恶性),再对数据集的10个预测变量作标准化处理,最后将数据集分为75%的训练集和25%的测试集,其中训练集用来学习分类模型,测试集用来检验模型的预测精度。下面利用逻辑回归与LASSO惩罚、L_2惩罚、弹性网惩罚3种惩罚逻辑回归预测乳腺肿瘤的良性与恶性,协助临床医生诊断患者肿瘤状态,减少误诊,提升诊断效率,进而提升乳腺癌患者的治愈率和存活率。

表3 预测变量与响应变量之间的相关系数

Table 3 Correlation coefficients between prediction variables and response variables

ρ_{Y,X_1}	ρ_{Y,X_2}	ρ_{Y,X_3}	ρ_{Y,X_4}	ρ_{Y,X_5}	ρ_{Y,X_6}	ρ_{Y,X_7}	ρ_{Y,X_8}	ρ_{Y,X_9}	$\rho_{Y,X_{10}}$
0.72	0.51	0.74	0.70	0.39	0.60	0.67	0.77	0.32	0.000 15

2 四种分类器

2.1 逻辑回归

逻辑回归模型是一种广义回归模型,常用于预测分类问题。二类逻辑回归(Logistic regression, LR)表示为

$$P:=P(Y=1|\boldsymbol{X};\boldsymbol{\beta})=\frac{\mathrm{e}^{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}}}{1+\mathrm{e}^{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}}}=h_{\boldsymbol{\beta}}(\boldsymbol{X}),\ 1-P=P(Y=0|\boldsymbol{X};\boldsymbol{\beta})=\frac{1}{1+\mathrm{e}^{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}}}=1-\boldsymbol{h}_{\boldsymbol{\beta}}(\boldsymbol{X}) \tag{1}$$

式中:响应变量 $Y=1$ 表示恶性肿瘤,$Y=0$ 表示良性肿瘤,预测变量 $\boldsymbol{X}=(1,X_1,X_2,\cdots,X_{10})^{\mathrm{T}}$ 为实值随机变量,$P(Y|\boldsymbol{X})$ 表示条件概率,$\boldsymbol{\beta}=(\beta_0,\beta_1,\beta_2,\cdots,\beta_{10})^{\mathrm{T}}$ 表示预测变量对肿瘤分类的影响。注意到 $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_{10}X_{10}$,对应 logit 变换为 $g(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})=\dfrac{1}{1+\mathrm{e}^{-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}}}$,事件发生与不发生的概率比为优势比

$$\frac{P}{1-P}=\frac{\boldsymbol{h}_{\boldsymbol{\beta}}(\boldsymbol{X})}{1-\boldsymbol{h}_{\boldsymbol{\beta}}(\boldsymbol{X})}=\mathrm{e}^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_{10}X_{10}} \tag{2}$$

取对数得到

$$\ln\left(\frac{P}{1-P}\right)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_{10}X_{10} \tag{3}$$

由于 $P(Y|\boldsymbol{X};\boldsymbol{\beta})=h_{\boldsymbol{\beta}}(\boldsymbol{X})^{Y}(1-h_{\beta}(X))^{1-Y}$,故似然函数为

$$L(\boldsymbol{\beta})=\prod_{i=1}^{n}P\left(Y^{(i)}\middle|\boldsymbol{X}^{(i)};\boldsymbol{\beta}\right)=\prod_{i=1}^{n}h_{\boldsymbol{\beta}}(\boldsymbol{X}^{(i)})^{Y^{(i)}}(1-h_{\boldsymbol{\beta}}(\boldsymbol{X}^{(i)}))^{1-Y^{(i)}} \tag{4}$$

对数似然函数为

$$\ell(\boldsymbol{\beta})=\ln(L(\boldsymbol{\beta}))=\ln\left(\prod_{i=1}^{n}h_{\boldsymbol{\beta}}(\boldsymbol{X}^{(i)})^{Y^{(i)}}(1-h_{\boldsymbol{\beta}}(\boldsymbol{X}^{(i)}))^{1-Y^{(i)}}\right)=$$

$$\sum_{i=1}^{n}(Y^{(i)}\ln(h_{\boldsymbol{\beta}}(\boldsymbol{X}^{(i)}))+(1-Y^{(i)})\ln(1-h_{\boldsymbol{\beta}}(\boldsymbol{X}^{(i)}))) \tag{5}$$

对式(5)关于 $\boldsymbol{\beta}$ 求导,得到得分方程

$$\ell'(\boldsymbol{\beta})=\frac{\partial\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=\sum_{i=1}^{n}\left(Y^{(i)}\frac{1}{h_{\boldsymbol{\beta}}(\boldsymbol{X}^{(i)})}-(1-Y^{(i)})\frac{1}{1-h_{\boldsymbol{\beta}}(\boldsymbol{X}^{(i)})}\right)\frac{\partial h_{\boldsymbol{\beta}}(\boldsymbol{X}^{(i)})}{\partial\boldsymbol{\beta}}=$$

$$\sum_{i=1}^{n}\left(Y^{(i)}\frac{1}{h_{\boldsymbol{\beta}}(\boldsymbol{X}^{(i)})}-(1-Y^{(i)})\frac{1}{1-h_{\boldsymbol{\beta}}(\boldsymbol{X}^{(i)})}\right)h_{\boldsymbol{\beta}}(\boldsymbol{X}^{(i)})(1-h_{\boldsymbol{\beta}}(\boldsymbol{X}^{(i)}))\boldsymbol{X}^{(i)}=$$

$$\sum_{i=1}^{n}(Y^{(i)}-h_{\boldsymbol{\beta}}(\boldsymbol{X}^{(i)}))\boldsymbol{X}^{(i)}=0 \tag{6}$$

显然,$\partial\ell(\boldsymbol{\beta})/\partial\boldsymbol{\beta}=0$ 为非线性隐式方程。令损失函数为

$$J(\boldsymbol{\beta})=-\ell(\boldsymbol{\beta})=-\sum_{i=1}^{n}(Y^{(i)}\ln(h_{\boldsymbol{\beta}}(X^{(i)}))+(1-Y^{(i)})\ln(1-h_{\boldsymbol{\beta}}(X^{(i)}))) \tag{7}$$

则用梯度下降迭代方法得到参数

$$\beta_j:=\beta_j-\alpha\frac{\partial J(\boldsymbol{\beta})}{\partial\beta_j}\qquad j=1,2,\cdots,10 \tag{8}$$

式中 α 称为学习率或者参数 β_j 的步长变化。梯度

$$\frac{\partial J(\boldsymbol{\beta})}{\partial\beta_j}=-\sum_{i=1}^{n}(\boldsymbol{Y}^{(i)}-h_{\boldsymbol{\beta}}(\boldsymbol{X}^{(i)}))\boldsymbol{X}_j^{(i)} \tag{9}$$

注意α的取值不宜过大或过小:α过大,难以得到理想的β_j;α过小,β_j值变化很小,收敛速度变慢,需要多次迭代才能得到理想的β_j。通常取α为0.1,0.01或0.05。

牛顿迭代法也是求多项式函数根的一种常用算法。对于$J(\boldsymbol{\beta})$,牛顿迭代公式为

$$\beta_{j,n+1}=\beta_{j,n}-\frac{J'(\beta_{j,n})}{J''(\beta_{j,n})} \tag{10}$$

如果$\boldsymbol{\beta}=(\beta_1,\beta_2)$,则Hessian矩阵为

$$\boldsymbol{H}_{\ell(\beta)}=\begin{bmatrix}\dfrac{\partial^2\ell}{\partial\beta_1\partial\beta_1} & \dfrac{\partial^2\ell}{\partial\beta_1\partial\beta_2}\\ \dfrac{\partial^2\ell}{\partial\beta_2\partial\beta_1} & \dfrac{\partial^2\ell}{\partial\beta_2\partial\beta_2}\end{bmatrix} \tag{11}$$

如果$\boldsymbol{\beta}=(\beta_1,\beta_2,\beta_3)$,则Hessian矩阵为

$$\boldsymbol{H}_{\ell(\beta)}=\begin{bmatrix}\dfrac{\partial^2\ell}{\partial\beta_1\partial\beta_1} & \dfrac{\partial^2\ell}{\partial\beta_1\partial\beta_2} & \dfrac{\partial^2\ell}{\partial\beta_1\partial\beta_3}\\ \dfrac{\partial^2\ell}{\partial\beta_2\partial\beta_1} & \dfrac{\partial^2\ell}{\partial\beta_2\partial\beta_2} & \dfrac{\partial^2\ell}{\partial\beta_2\partial\beta_3}\\ \dfrac{\partial^2\ell}{\partial\beta_3\partial\beta_1} & \dfrac{\partial^2\ell}{\partial\beta_3\partial\beta_2} & \dfrac{\partial^2\ell}{\partial\beta_3\partial\beta_3}\end{bmatrix} \tag{12}$$

对多元向量$\boldsymbol{\beta}$,如果$\nabla\ell(\boldsymbol{\beta})$表示一阶导数,则基于Hessian矩阵的牛顿迭代步骤为

$$\beta_{n+1}=\beta_n+H_{\ell(\beta)}^{-1}\nabla\ell(\boldsymbol{\beta}) \tag{13}$$

2.2 LASSO惩罚逻辑回归

LASSO惩罚是Tibshirani[21]于1996年提出的一种变量选择和收缩估计方法,主要对线性回归引入一个L_1惩罚函数(回归系数的绝对值之和小于调整参数)得到LASSO惩罚线性回归。本文的LASSO惩罚逻辑回归主要利用训练样本得到逻辑回归的负对数似然函数,并引入一个L_1惩罚,作变量选择的同时得到模型参数估计,再结合检验样本预测分类。LASSO惩罚逻辑回归的估计可以表示为

$$\hat{\beta}_{\mathrm{LASSO}}=\underset{\beta}{\operatorname{argmin}}\left\{-\ell(\boldsymbol{\beta})+\lambda\sum_{j=1}^{10}\left|\beta_j\right|\right\} \tag{14}$$

式中$\ell(\boldsymbol{\beta})$见式(5)。式(14)中系数的惩罚力度主要由调节参数λ决定并控制模型的拟合优度。当$\lambda=0$时,对应没有惩罚的逻辑回归,估计为极大似然估计;当λ增大,系数估计的压缩越大,特别当$\lambda\to\infty$时,所有系数都被压缩为0。

定义

$$\widetilde{Y}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}+\boldsymbol{W}^{-1}(Y-\widetilde{P}),Z_j=n^{-1}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{W}(\widetilde{Y}-\boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j})$$

$$\boldsymbol{X}_{-j}=(X_1,\cdots,X_{j-1},0,X_{j+1},\cdots,X_{10}),\boldsymbol{\beta}_{-j}=(\beta_1,\cdots,\beta_{j-1},0,\beta_{j+1},\cdots,\beta_{10})$$

$$\boldsymbol{W}=\mathrm{diag}\left\{W_i=\widetilde{P}_i(1-\widetilde{P}_i)\right\},\widetilde{P}=\frac{1}{1+\mathrm{e}^{-\boldsymbol{X}^{\mathrm{T}}\hat{\beta}_{\mathrm{LASSO}}}}$$

Breheny等[22]对LASSO惩罚逻辑回归引入坐标下降算法迭代得到参数估计

$$\hat{\beta}_{\mathrm{LASSO}}(Z_j,\lambda)=\frac{S(Z_j,\lambda)}{v_j} \tag{15}$$

式中:软门限算子$S(Z_j,\lambda)=\mathrm{sign}(Z_j)\left(\left|Z_j\right|-\lambda\right),v_j=n^{-1}\boldsymbol{X}_j^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X}_j$。

2.3 L_2惩罚逻辑回归

与LASSO惩罚逻辑回归不同，L_2惩罚逻辑回归是对逻辑回归的负对数似然函数引入L_2惩罚函数实现变量选择和回归系数收缩的目的。对应的惩罚似然函数为[23]

$$\ell^{\lambda}(\boldsymbol{\beta})=\ell(\boldsymbol{\beta})-\lambda\left\|\boldsymbol{\beta}\right\|_2^2 \tag{16}$$

式中：λ为调整参数，$\left\|\boldsymbol{\beta}\right\|_2^2=\sum_{j=1}^{10}\beta_j^2$为参数向量$\boldsymbol{\beta}$的2范数。调整参数$\lambda$控制$\boldsymbol{\beta}$范数的收缩范围，$\lambda$越大，收缩越大，回归系数收缩趋向于0；$\lambda$越小，收缩越小，回归系数趋向于经典逻辑回归的极大似然估计。当预测变量数目较多或者预测变量之间高度相关时会产生不稳定的参数估计，而对逻辑回归引入L_2惩罚得到的收缩估计，不仅估计方差更小，而且模型更加稳定。因此，合理选择参数λ是一个关键问题。记式(16)的极大值为$\hat{\boldsymbol{\beta}}^{\lambda}$，求解$\hat{\boldsymbol{\beta}}^{\lambda}$与极大似然估计类似，可用Newton-Raphson算法得到。$\hat{\boldsymbol{\beta}}^{\lambda}$的一阶导数为

$$U^{\lambda}(\boldsymbol{\beta})=\sum_{i=1}^{n}\boldsymbol{X}^{(i)\mathrm{T}}\{Y^{(i)}-h_{\beta}(\boldsymbol{X}^{(i)})\}-2\lambda\boldsymbol{\beta}=U(\boldsymbol{\beta})-2\lambda\boldsymbol{\beta} \tag{17}$$

$\hat{\boldsymbol{\beta}}^{\lambda}$的负二阶导数矩阵为

$$\Omega^{\lambda}(\boldsymbol{\beta})=\boldsymbol{X}^{\mathrm{T}}V(\boldsymbol{\beta})\boldsymbol{X}+2\lambda\boldsymbol{I}=\Omega(\boldsymbol{\beta})+2\lambda\boldsymbol{I} \tag{18}$$

式中：$V(\boldsymbol{\beta})=\{h_{\beta}(\boldsymbol{X}^{(i)})(1-h_{\beta}(\boldsymbol{X}^{(i)}))\}$和$\Omega(\boldsymbol{\beta})=\boldsymbol{X}^{\mathrm{T}}V(\boldsymbol{\beta})\boldsymbol{X}$。惩罚似然函数关于真实参数$\boldsymbol{\beta}_0$的一阶导数的Taylor展开式为

$$U^{\lambda}(\hat{\boldsymbol{\beta}}^{\lambda})=U^{\lambda}(\boldsymbol{\beta})+(\hat{\boldsymbol{\beta}}^{\lambda}-\boldsymbol{\beta})^{\mathrm{T}}\Omega(\boldsymbol{\beta})+o(\|\hat{\boldsymbol{\beta}}^{\lambda}-\boldsymbol{\beta}\|) \tag{19}$$

使用式(17)和(18)以及$U^{\lambda}(\hat{\boldsymbol{\beta}}^{\lambda})=0$产生$\hat{\boldsymbol{\beta}}^{\lambda}$的一阶渐近为

$$\hat{\boldsymbol{\beta}}^{\lambda}=\boldsymbol{\beta}+\{\Omega(\boldsymbol{\beta})+2\lambda\boldsymbol{I}\}^{-1}\{\Omega(\boldsymbol{\beta})-2\lambda\boldsymbol{\beta}\}=\{\Omega(\boldsymbol{\beta})+2\lambda\boldsymbol{I}\}^{-1}\{U(\boldsymbol{\beta})+\Omega(\boldsymbol{\beta})\boldsymbol{\beta}\} \tag{20}$$

类似的也可得到$\hat{\boldsymbol{\beta}}=\boldsymbol{\beta}+\Omega^{-1}(\boldsymbol{\beta})U(\beta)$。因此惩罚似然函数极大值为

$$\hat{\boldsymbol{\beta}}^{\lambda}=\{\Omega(\boldsymbol{\beta})+2\lambda\boldsymbol{I}\}^{-1}\Omega(\boldsymbol{\beta})\hat{\boldsymbol{\beta}} \tag{21}$$

$\hat{\boldsymbol{\beta}}^{\lambda}$的渐近方差为

$$\{\Omega(\boldsymbol{\beta})+2\lambda\boldsymbol{I}\}^{-1}\Omega(\boldsymbol{\beta})\{\Omega(\boldsymbol{\beta})+2\lambda\boldsymbol{I}\}^{-1} \tag{22}$$

2.4 弹性网惩罚逻辑回归

Zou等[24]提出了弹性网(Elastic net，EN)惩罚方法。该方法是L_1惩罚和L_2惩罚的折中，适合$p\gg n$的超高维稀疏情况和具有多重共线性的模型。由于对逻辑回归引入弹性网惩罚可改进模型表现，减少预测误差，因此这里对逻辑回归的负对数似然函数引入弹性网惩罚，得到弹性网惩罚逻辑回归及其模型估计。

$$\hat{\boldsymbol{\beta}}_{\mathrm{EN}}=\arg\min_{\beta}\left\{-\ell(\boldsymbol{\beta})+\lambda_1\sum_{j=1}^{10}\left|\beta_j\right|+\lambda_2\sum_{j=1}^{10}\beta_j^2\right\} \tag{23}$$

令$\alpha=\dfrac{\lambda_1}{\lambda_1+\lambda_2}$，$\lambda=\lambda_1+\lambda_2$，则式(23)可写成

$$\hat{\boldsymbol{\beta}}_{\mathrm{EN}}=\arg\min_{\beta}\left\{-\ell(\boldsymbol{\beta})+\lambda\alpha\sum_{j=1}^{10}\left|\beta_j\right|+\lambda(1-\alpha)\sum_{j=1}^{10}\beta_j^2\right\} \tag{24}$$

当$\alpha=0$时，弹性网惩罚逻辑回归为L_2惩罚逻辑回归；当$\alpha=1$时，弹性网惩罚逻辑回归为LASSO惩罚逻辑回归。因此弹性网惩罚结合了LASSO惩罚与L_2惩罚的优点，既能进行变量选择又能消除共线性影响。

3 模型精度评估

3.1 混淆矩阵

常用混淆矩阵和ROC曲线评估预测精度。混淆矩阵分别对真实分类序列和预测分类序列统计分类模型归错类与归对类的观测值个数。本文借助caret包中的confusionMatrix函数来计算混淆矩阵。如果0表示良性,1表示恶性,则二分类混淆矩阵如表4所示。

表4 二分类混淆矩阵

Table 4 Binary confusion matrix

预测类别	真实类1(恶性:$Y=1$)	真实类2(良性:$Y=0$)
预测类1(恶性:$\hat{Y}=1$)	TP(真阳性)	FP(假阳性)
预测类2(良性:$\hat{Y}=0$)	FN(假阴性)	TN(真阴性)

下面利用混淆矩阵计算准确率、精确率、错误率、灵敏度和特异度等指标。

准确率(Accuracy)表示模型预测结果正确的样本数与所有样本数的比值,即

$$\text{准确率}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FN}+\text{FP}}$$

精确率(Precision)表示模型预测结果为正例的样本中,实际为正例的样本所占比例,即

$$\text{精确率}=\frac{\text{TP}}{\text{TP}+\text{FP}}$$

错误率(Error rate)表示模型预测结果错误的样本数与所有样本数的比值,即

$$\text{错误率}=\frac{\text{FP}+\text{FN}}{\text{TP}+\text{TN}+\text{FN}+\text{FP}}$$

灵敏度(Sensitivity)表示正确预测的正例数在实际正例样本数中的占比,即

$$\text{灵敏度}=\frac{\text{TP}}{\text{TP}+\text{FN}}$$

特异度(Specificity)表示正确预测的负例样本在实际负例样本中的占比,即

$$\text{特异度}=\frac{\text{TN}}{\text{TN}+\text{FP}}$$

这些指标无法直观判断模型效果,因此利用特异度和灵敏度两个指标绘制ROC曲线:1−特异度为x轴,表示假阳性率(False postive rate,FPR),FPR越小,误判率越低,预测正类中实际负类越小;灵敏度为y轴,表示真阳性率(True postive rate,TPR),TPR越大,命中率越高,预测正类中实际正类越多。绘制不同阈值下1−特异度和灵敏度的组合变化,ROC曲线下的阴影面积就是AUC指标。通常AUC的值越大,模型拟合效果越好。本文利用pROC包实现ROC曲线的可视化,用glmnet包构建逻辑回归、惩罚逻辑回归在坐标下降算法下的正则化路径。

3.2 几种模型的预测表现

利用乳腺癌数据集建立逻辑回归,求出各个预测变量对应的参数估计值,结果如表5所示。

表5 逻辑回归模型的参数估计值

Table 5 Parameter estimation for logistic regression model

参数	估计值	标准差	P值	影响程度
β_0	0.880 77	0.679 25	0.194 7	
β_1	−9.268 92	14.252 07	0.515 46	
β_2	2.084 49	0.370 52	1.85e−08	***
β_3	1.527 46	14.098 93	0.917 73	
β_4	13.035 94	6.989 69	0.062 18	·
β_5	1.481 10	0.558 38	0.007 99	**
β_6	−0.436 62	1.220 75	0.720 59	
β_7	0.400 20	0.779 23	0.607 55	
β_8	2.380 43	1.321 56	0.071 67	·
β_9	0.042 67	0.348 70	0.902 61	
β_{10}	−0.090 78	0.664 04	0.891 26	

注:“*”表示变量显著,*数量越多,表示显著性越强;“·”表示变量显著性不强;空白表示变量不显著。表6同。

从表5中看出显著变量并不多,但不显著变量

对诊断结果也不是完全没有影响。这里用逐步回归作变量选择,结果如表6所示。

表6 变量选择后的参数估计值

Table 6 Parameter estimation after variable selection

参数	估计值	标准差	P值	影响程度
β_0	1.004 2	0.621 5	0.106 121	
β_1	−9.393 2	4.931 8	0.056 832	·
β_2	2.084 49	0.376 7	1.66e−08	***
β_4	14.860 0	6.056 6	0.014 147	*
β_5	1.359 8	0.427 0	0.001 449	**
β_8	2.522 8	0.657 0	0.000 123	***

从表6看出选出的变量显著,比较两个度量模型拟合优度(Akaike information criterion, AIC)结果,包含全部变量的模型AIC=130.13,变量选择后的模型AIC=120.66,AIC值越低的模型效果更好,因此通过变量选择提升了模型效果。通过变量选择,效果明显变化的是变量X_4和X_8,$\hat{\beta}_4$从13.035 94上升到14.860 0,其P值从0.062 18下降到0.014 147,变得更加显著;$\hat{\beta}_8$从2.380 43上升到2.522 8,其P值从0.071 67下降到0.000 123,显著性明显提高。因此,变量选择后的预测模型为

$$\ln\left(\frac{P}{1-P}\right)=1.004\,2-9.393\,2X_1+2.084\,49X_2+14.860\,0X_4+1.359\,8X_5+2.522\,8X_8$$

由以上模型得出,X_1每增加1个单位,相应的优势比对数减少9.393 2,即乳腺肿瘤细胞核的半径与肿瘤是良性还是恶性呈负相关;X_2每增加1个单位,相应的优势比增加2.084 49,即乳腺肿瘤细胞核的纹理与肿瘤是良性还是恶性呈正相关;其余的X_4、X_5、X_8皆呈现正相关性。

利用75%的训练集数据学习预测模型,利用25%的测试集数据建立混淆矩阵和ROC曲线图检验模型预测效果。先设定一个阈值,将概率值大于这个阈值的归为1类,小于这个阈值的归为0类。最佳阈值的选取会直接影响模型预测结果,这里将预测值和实际值从低到高排序放在一起,在0和1分界值的区间中选取最佳阈值,先设定区间中的一个值得出ROC曲线图,得到最佳阈值和特异度、灵敏度等值,再根据这个阈值调整使得设定的阈值与ROC得到的最佳阈值一致,即为最终的最佳阈值。

利用10个预测变量和响应变量建模LASSO惩罚逻辑回归,用十重交叉验证选取模型,对于每一个λ值,在红点所示目标参量的均值左右,可以得到一个目标参量的置信区间。两条虚线分别指示了两个特殊的λ值:一个是lambda.min,即给出最小交叉核实误差的λ值;另一个是lambda.1se,即给出交叉核实误差最小值的1倍标准差范围的λ值。通过选取模型准确性较高的λ作为调节参数,得到LASSO惩罚逻辑回归的混淆矩阵。对逻辑回归引入L_2惩罚得到L_2惩罚逻辑回归。将LASSO惩罚和L_2惩罚进行折中得到弹性网惩罚,并选取3组不同值进行分析($\alpha=0.3$、0.5、0.9),得到不同模型的混淆矩阵,结果如表7所示。

表7 混淆矩阵

Table 7 Confusion matrix

分类器	预测类别	1(恶性)	0(良性)
LR	1(恶性)	33	4
	0(良性)	2	103
LASSO	1(恶性)	33	2
	0(良性)	2	105
L_2	1(恶性)	33	3
	0(良性)	2	104
EN ($\alpha=0.3$、0.5)	1(恶性)	33	3
	0(良性)	2	104
EN($\alpha=0.9$)	1(恶性)	33	2
	0(良性)	2	105

在表7中,$\alpha=0.3$和$\alpha=0.5$时的弹性网惩罚逻辑回归与L_2惩罚逻辑回归得到的混淆矩阵一致,$\alpha=0.9$时的弹性网惩罚与LASSO惩罚的混淆矩阵一致。通过逻辑回归的混淆矩阵计算可得:准确率$=\dfrac{33+103}{33+4+2+103}\approx0.957\,7$;精确率$=\dfrac{33}{33+4}\approx0.891\,9$;灵敏度$=\dfrac{33}{33+2}\approx0.942\,9$;特异度$=\dfrac{103}{4+103}=$

0.962 6。计算指标值如表8所示。

表8 通过混淆矩阵计算的指标值

Table 8 Index values calculated by confusion matrix

分类器	准确率	精确率	灵敏度	特异度
LR	0.957 7	0.891 9	0.942 9	0.962 6
LASSO	0.971 8	0.942 9	0.942 9	0.981 3
L_2	0.964 8	0.916 7	0.942 9	0.972 0
EN($\alpha=0.3$)	0.964 8	0.916 7	0.942 9	0.972 0
EN($\alpha=0.5$)	0.964 8	0.916 7	0.942 9	0.972 0
EN($\alpha=0.9$)	0.971 8	0.942 9	0.942 9	0.981 3

由表8可得模型的预测准确率,$\alpha=0.3$和$\alpha=0.5$时的弹性网与L_2惩罚逻辑回归的计算结果一致,$\alpha=0.9$时的弹性网惩罚与LASSO惩罚的结果一致,且每种模型的灵敏度都为0.942 9。初步判断模型的预测效果,逻辑回归的准确率为0.957 7;LASSO惩罚逻辑回归的准确率为0.971 8;L_2惩罚逻辑回归的准确率为0.964 8;弹性网惩罚逻辑回归对于不同的α值,预测表现也不同,其最佳预测准确率为0.971 8。因此,惩罚逻辑回归模型预测表现优于不加惩罚的逻辑回归模型。

ROC曲线图可直观评价预测效果,主要依据折线下AUC的值,AUC越大,分类器预测表现越好。AUC的取值范围在0.5和1之间,越接近1的预测效果越好,当AUC=0.5时,分类器基本不起作用,预测结果和随机猜测差不多;AUC在0.5~0.7时有较低准确性;AUC在0.7~0.9时有一定准确性;AUC在0.9以上有较高准确性。利用灵敏度和特异度绘制ROC曲线图如图2所示。

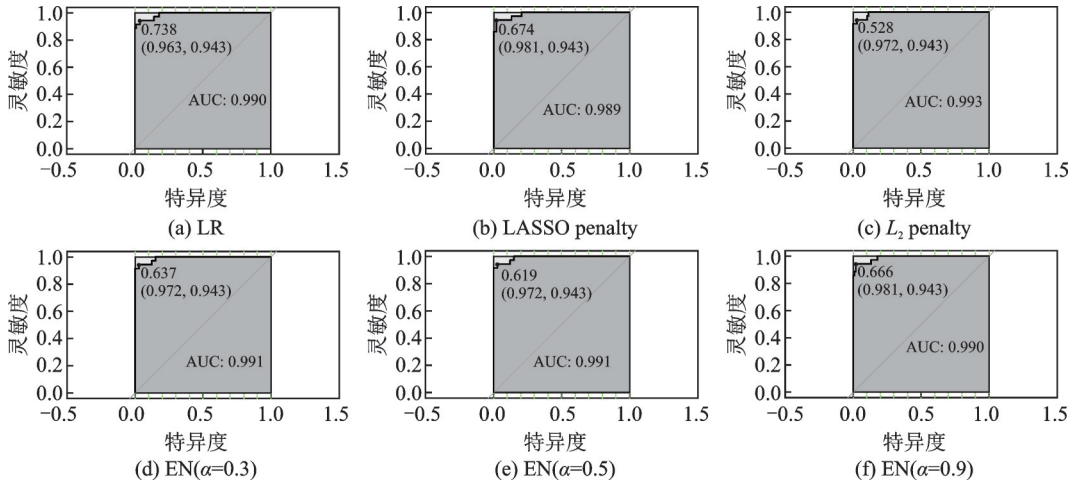

图2 几种模型的ROC曲线

Fig.2 ROC curves for several models

由图2可直接得到模型的最佳阈值、灵敏度、特异度和AUC值,结果如表9所示。

表9 通过ROC曲线得出的指标值

Table 9 Index values derived by ROC curves

分类器	阈值	灵敏度	特异度	AUC
LR	0.738	0.943	0.963	0.990
LASSO	0.674	0.943	0.981	0.989
L_2	0.528	0.943	0.972	0.993
EN($\alpha=0.3$)	0.637	0.943	0.972	0.991
EN($\alpha=0.5$)	0.619	0.943	0.972	0.991
EN($\alpha=0.9$)	0.666	0.943	0.981	0.990

由表9可知,ROC曲线得出的灵敏度和特异度与混淆矩阵计算值一致,其中LASSO和$\alpha=0.9$的弹性网惩罚逻辑回归具有相同的精确率0.942 9,灵敏度0.942 9,特异度0.981 3。表9中逻辑回归的AUC=0.990;LASSO惩罚逻辑回归的AUC=0.989;L_2惩罚逻辑回归的AUC=0.993;弹性网惩罚逻辑回归$\alpha=0.9$时的AUC=

0.990,故分类器预测表现不错。

4 结束语

本文利用威斯康星大学的数据来学习不同分类模型,利用指标平均值作为预测变量建立4种分类器:逻辑回归、LASSO惩罚逻辑回归、L_2惩罚逻辑回归和弹性网惩罚逻辑回归预测乳腺癌,展示了不同的预测表现和预测精度,其中LASSO惩罚逻辑回归分类器和弹性网惩罚逻辑回归分类器预测表现更好,预测精度最高,逻辑回归预测表现最差,预测精度最低。本文收集的数据样本量不大,下一步重点研究如何在大数据集中采用这些分类器进一步提高乳腺癌的预测表现。

参考文献:

[1] 陈后金,李艳凤,彭亚辉. 多视角乳腺X线图像匹配方法综述[J]. 数据采集与处理, 2016, 31(5): 845-855.
CHEN Houjin, LI Yanfeng, PENG Yahui. Survey of multi-view matching in mammogram[J]. Journal of Data Acquisition and Processing, 2016, 31(5): 845-855.

[2] AL-OBEIDAT F, SPENCER B, ALFANDI O. Consistently accurate forecasts of temperature within buildings from sensor data using ridge and LASSO regression[J]. Future Generation Computer Systems, 2020, 110: 382-392.

[3] HUANG J H, TSAI Y C, WU P Y, et al. Predictive modeling of blood pressure during hemodialysis: A comparison of linear model, random forest, support vector regression, XGBoost, LASSO regression and ensemble method[J]. Computer Methods and Programs in Biomedicine, 2020. DOI:10.1016/j.cmpb.2020.105536.

[4] SAKUMA Y, OKAMOTO N, SAITO H, et al. A logistic regression predictive model and the outcome of patients with resected lung adenocarcinoma of 2 cm or less in size[J]. Lung Cancer, 2009, 65(1): 85-90.

[5] YI J, YANG G, ZHANG Z, et al. An improved elastic net method for traveling salesman problem[J]. Neurocomputing, 2009, 72(4/5/6): 1329-1335.

[6] 刘柳,陶大程. LASSO问题的最新算法研究[J]. 数据采集与处理, 2015, 30(1): 35-46.
LIU Liu, TAO Dacheng. Review on recent method of solving LASSO problem[J]. Journal of Data Acquisition and Processing, 2015, 30(1): 35-46.

[7] HUANG Y L, CHEN D R, JIANG Y R, et al. Computer-aided diagnosis using morphological features for classifying breast lesions on utrasound[J]. Ultrasound in Obstetrics & Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology, 2008, 32(4): 565-572.

[8] MONTAZERI M, MONTAZERI M, MONTAZERI M, et al. Machine learning models in breast cancer survival prediction [J]. Technology and Health Care, 2016, 24(1): 31-42.

[9] XIA L, YAO Y, DONG Y, et al. Mueller polarimetric microscopic images analysis based classification of breast cancer cells [J]. Optics Communications, 2020. DOI: 10.1016/j.optcom.2020.126194.

[10] 胡雪梅,蒋慧凤. 具有技术指标的逻辑回归模型预测谷歌股票的涨跌趋势[J]. 系统科学与数学, 2021, 41(3): 1-22.
HU Xuemei, JIANG Huifeng. Logistic regression model with technical indicators predicts ups and downs for google stock prices [J]. Journal of Systems Science and Mathematical Sciences, 2021, 41(3): 1-22.

[11] PARK M Y, HASTIE T. Penalized logistic regression for detecting gene interactions[J]. Biostatistics, 2008, 9(1): 30-50.

[12] MEIER L, VAN DE GEER S , BÜHLMANN P. The group LASSO for logistic regression[J]. Journal of the Royal Statistical Society Series B—Statistical Methodology, 2008, 70(1): 53-71.

[13] FRIEDMAN J, HASTIE T, NIGHTSHIRT R. Regularization paths for generalized linear models via coordinate descent[J]. Journal of Statistical Software, 2010, 33(1): 1-22.

[14] AYERS K L, CORDELL H J. SNP selection in genome-wide and candidate gene studies via penalized logistic regression[J]. Genetic Epidemiology, 2010, 34(8): 879-891.

[15] BREHENY P, HUANG J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors[J]. Statistics and Computing, 2015, 25(2): 173-187.

[16] LI Q, XIE B, YOU J, et al. Correlated logistic model with elastic net regularization for multilabel image classification[J]. IEEE

Transaction on Image Processing, 2016, 25(8): 3801-3813.

[17] SARI P D, AIDI M N, SARTONO B. Credit scoring analysis using LASSO logistic regression and support vector machine (SVM)[J]. International Journal of Engineering and Management Research, 2017, 7(4): 393-397.

[18] MÜNCH M M, PEETERS C F W, VAN DER VAART A W, et al. Adaptive group-regularized logistic elastic net regression [J]. Biostatistics, 2019. DOI: 10.1093/biostatistics/kxz062.

[19] GARCIA-CARRETERO R, VIGIL-MEDINA L, BARQUERO-PEREZ O, et al. Logistic LASSO and elastic net to characterize vitamin D deficiency in a hypertensive obese population[J]. Metabolic Syndrome and Related Disorders, 2020, 18 (2): 79-85.

[20] WOLBERG W H, STREET W N, MANGASARIAN O L. Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates[J]. Cancer Letters, 1994, 77(2/3): 163-171.

[21] TIBSHIRANI R J. Regression shrinkage and selection via the LASSO[J]. Journal of the Royal Statistical Society. Series B: Methodological, 1996, 73(1): 273-282.

[22] BREHENY P, HUANG J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection[J]. The Annals of Applied Statistics, 2011, 5(1): 232-253.

[23] 胡雪梅,刘锋. 高维统计模型的估计理论与模型识别[M]. 北京: 高等教育出版社, 2020: 241-265.
HU Xuemei, LIU Feng. Estimation theory and model identification of high dimensional statistical models[M]. Beijing: Advanced Education Press, 2020: 241-265.

[24] ZOU H, HASTIE T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society B, 2005, 67(2): 301-320.

作者简介:

胡雪梅(1978-),女,教授,博士生导师,研究方向:高维数据分析、统计学习、面板数据分析、经验似然、半参数统计和过程推断,E-mail:huxuem@163.com。

谢英(1997-),**通信作者**,女,硕士研究生,研究方向:统计方法,E-mail: 2811457187@qq.com。

蒋慧凤(1993-),女,博士研究生,研究方向:高维分类模型。

(编辑:王静)