乳腺癌预测模型构建研究*

阮旭凌 刘 琦 郭志恒 晏峻峰*

（湖南中医药大学信息科学与工程学院 长沙 410208）

〔摘要〕 在 kaggle 网站乳腺癌诊断数据集基础上，应用机器学习算法构建预测模型，阐述基本原理、模型建立方法，分析实验结果，指出降维处理后训练的预测模型分类准确率提高，其中基于 XGBoost 构建的预测模型分类效果最佳。

〔关键词〕 乳腺癌；降维；LDA；XGBoost；分类

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.05.007

Study on the Construction of Breast Cancer Prediction Model *RUAN Xuling, LIU Qi, GUO Zhiheng, YAN Junfeng, School of Informatics, Hunan University of Chinese Medicine, Changsha 410208, China*

〔**Abstract**〕 Based on the breast cancer diagnosis data set of the kaggle website, machine learning algorithms are used to construct prediction models, the basic principles and model building methods are expounded, and the experimental results are analyzed. The classification accuracy of the trained prediction model is improved after dimensionality reduction, and the classification effect of the prediction model built by XGBoost is better.

〔**Keywords**〕 breast cancer; dimension reduction; LDA; XGBoost; classification

1 引言

近期世界卫生组织国际癌症研究机构发布全球最新癌症数据，显示乳腺癌已成全球第 1 大癌症[1]，不仅女性会患上乳腺癌，男性乳腺癌患者数量也不容小觑，乳腺癌防治形势的严峻性需要引起全社会重视，因此做到早发现、早确诊、早治疗对于患者具有重要意义。随着医疗技术发展，大量数据分析技术应用于乳腺癌疾病诊断。吴辰文、李长生和王伟等[2]提出基于加权特征向量核函数的支持向量机算法对乳腺癌进行分类预测；刘军、彭慧娴和黄斌等[3]提出 BP－GamysBoost 算法模型运用于乳腺癌疾病诊断；于凌涛、夏永强和闫昱晟等[4]采用卷积神经网络分类乳腺癌病理图像；冯照石和范祺[5]通过双重加权朴素贝叶斯算法预测乳腺癌患者复发率；余莹、樊重俊和朱人杰等[6]基于系统聚类和支持向量机算法模型对乳腺癌不同肿瘤特征进行诊断。上述几种算法虽然都取得良好效果，但分类预测精度有待提高。由于乳腺癌产生机制具有不确定性，且不同患者身心素质差异较大，因此需要针对个人制定个性化治疗方案以提高治疗效果[7]。在

〔修回日期〕 2022－03－11

〔作者简介〕 阮旭凌，硕士研究生；通信作者：晏峻峰，教授，博士生导师。

〔基金项目〕 科技创新 2030——“新一代人工智能”重大项目（批准号：2018AAA0102100）；湖南省教育厅科研重点项目“基于动态不确定因果图的证素辨证模型研究”（项目编号：18A219）；湖南省研究生创新课题“中医目诊信息数据元标准研究”（项目编号：CX2018B479）。

乳腺癌个性化治疗过程中，患者需要检查多项指标，如何在众多信息中迅速找到对患者病因有重要影响的指标是解决问题的关键。在乳腺癌原始检验指标中常存在冗余信息和噪声数据，如维度较大会导致问题的复杂性和时间的复杂度扩大，对数据特征属性进行降维能够降低系统计算量，结合机器学习算法构建预测模型，以此来诊断乳腺癌病发率，能够较好地辅助医生对患者进行乳腺癌早期诊断，具有重要现实意义。

2 数据集

2.1 数据集描述

本实验采用 kaggle 数据库提供的乳腺癌诊断数据集（Breast Cancer Wisconsin（Diagnostic）Data Set）。其中良性样本（标签为 0）308 例，恶性样本（标签为 1）190 例，每例样本中有 30 个特征属性，1 个类别标签和 1 个样本 id，类别标签表明患者乳腺细胞核属于良性还是恶性，特征属性描述细胞核的 10 个实际值特征，是鉴别患者是否被诊断为乳腺癌的关键指标。本实验将数据集的 70% 作为训练样本，剩余 30% 作为测试样本，见表 1。

表 1 特征属性

特征属性	特征属性
中心到病灶周围点的平均值	紧实度（周长平均值^2/面积 - 1.0）标准误差
灰度值标准差	轮廓凹陷严重程度标准误差
肿瘤核心区大小平均值	轮廓凹陷数量的标准误差
面积平均值	对称性标准误差
光滑度（半径长度局部变化）平均值	分形维数标准误差
紧实度（周长平均值^2/面积 - 1.0）	最坏（大）半径
轮廓凹陷严重程度平均值	灰度值标准差最差（大）标准差
轮廓凹陷数量平均值	最坏（大）周长
对称性平均值	最坏（大）面积
分形维数	最坏（大）平滑度（半径长度局部变化）

续表 1

中心到周界点距离平均值的标准误差	最坏（大）紧实度（周长平均值^2/面积 - 1.0）
灰度值标准差的标准误差	最坏（大）轮廓凹陷严重程度
周长标准误差	最坏（大）轮廓凹陷数量
面积标准误差	最坏（大）对称性
平滑度（半径长度局部变化）标准误差	最坏（大）分形维数

2.2 数据预处理

经过对数据集的筛查，未出现数据缺失、数据异常、数据重复等情况，且正负样本数据较均衡。本数据集的诊断类别为字符串，为方便对照将诊断结果量化为 2 种，其中 0 代表乳腺细胞核为良性，1 代表乳腺细胞核为恶性。在数据维度较高时，为提取有用特征方便计算，通常进行数据降维。原始数据集有特征属性 30 项，在这些指标中常存在冗余信息，增加了系统计算量和时间复杂度。通过线性判别式分析（Linear Discriminant Analysis，LDA）方法合并特征属性，减少数据冗余引起的误差，快速缩小数据维度，降低系统复杂性。

3 基本原理与模型建立

3.1 LDA 方法

3.1.1 定义 LDA 是基于分类模型进行特征属性合并的方法，是一种有监督的降维方式[8]。其基本原理是将带有类别标签的数据投影到维度更低的空间中，以便按类区分投影后的点。投影后空间中同类点之间距离会更接近、方差更小，而不同类之间的方差越大[9]。降维能够减少冗余信息引起的误差，提高识别准确率。

3.1.2 LDA 降维步骤 假设存在数据集 $D = \{(x1, y1), (x2, y2), \ldots, ((xm, ym))\}$，其中有 n 维向量样本 x_i，N_j（$j = 0, 1$）为第 j 类样本的个数，X_j（$j = 0, 1$）为第 j 类样本的集合，LDA 降维步骤[10]如下：

首先计算每个样本的均值向量 μ_j 和协方差矩阵 Σ_j：

$$\mu_i = \frac{1}{N_j}\sum_{x\in X_j} x(j=0,1) \quad (1)$$

$$\Sigma_j = \sum_{x\in X_j}(x-\mu_j)(x-\mu_j)^T(j=0,1) \quad (2)$$

然后计算类内散度矩阵 S_W 和类间散度矩阵 S_b：

$$S_W = \sum_0 + \sum_1 = \sum_{x\in X_0}(x-\mu_0)(x-\mu_0)^T + \sum_{x\in X_1}(x-\mu_1)(x-\mu_1)^T \quad (3)$$

$$S_b = (\mu_0-\mu_1)(\mu_0-\mu_1)^T \quad (4)$$

给定投影的直线为一个向量 w，则对于任意样本 x_i，其在给定直线 w 的投影为 w^Tx_i，对于两个类别的中心点 μ_0 和 μ_1 在 w 的投影为 $w^T\mu_0$ 和 $w^T\mu_1$。根据 LDA 原理要最大化不同类别之间的距离，即要使 $\|w^T\mu_0 - w^T\mu_1\|^2$ 尽可能大，同时最小化同类别之间的距离，即最小化 $w^T\sum_0 w + w^T\sum_1 w$。根据上述公式，可以得出优化目标为：

$$\underbrace{argmax}_{w} J(W) = \frac{w^T S_b w}{w^T S_w w} \quad (5)$$

求得最优投影直线 w，对样本集中的每个样本特征 x_i 进行转化得到新的输出样本 w^Tx_i，即可完成降维。

3.1.3 30 个特征属性对分类结果的重要性计算 在本实验中，利用 XGBoost 模型计算 30 个特征属性对分类结果的重要性，这些重要性得分可通过 python 中的成员变量 feature_ importances_ 得到，见图 1。取出排名前 2 位的特征属性 radius_ worst 和 concave points_ mean 进行正负样本分布可视化，两种类别的数据有重合部分，边界数据混杂，因此需进行降维，见图 2。

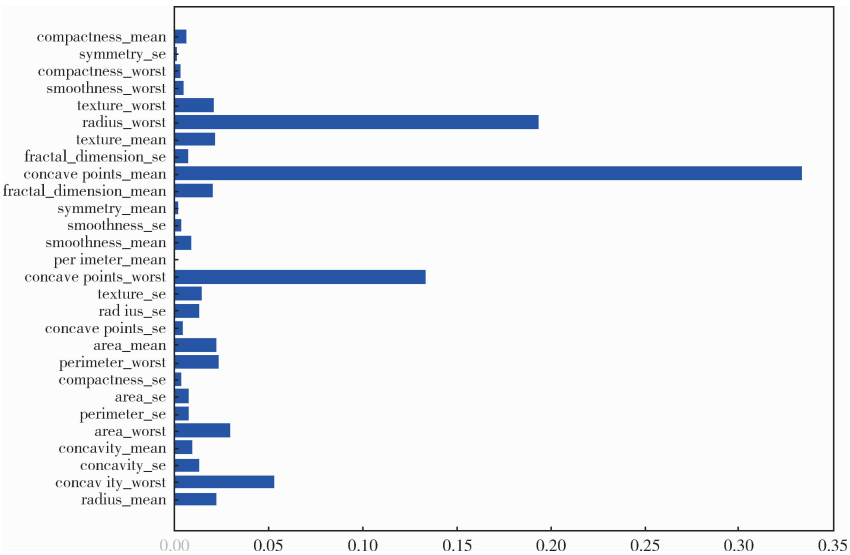

图 1 特征属性重要性分布

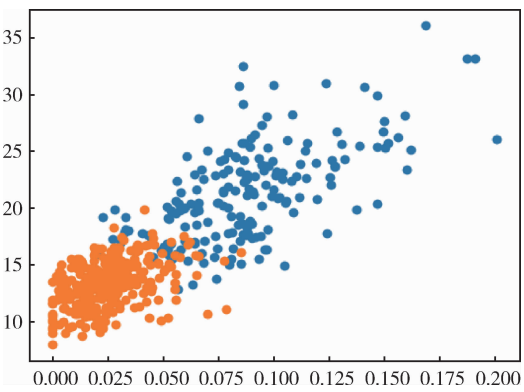

图 2 特征属性正负样本分布

3.1.4 降维效果 根据 LDA 原理可知对于二分类问题降维后的特征维数需不大于 1，因此采用 LDA 方法将数据特征从 30 维降维到 1 维，见图 3。其中所有数据被投影至同一条直线上，相同类别的数据投影点距离缩小，不同类别数据的中心间距扩大。

3.2 XGBoost 模型

3.2.1 定义 XGBoost[11] 即极端梯度提升算法，是对梯度提升（Gradient Boosting，GB）算法的改进和高效实现，除了可以是分类回归树算法（Classification And Regression Tree，CART）决策树也可

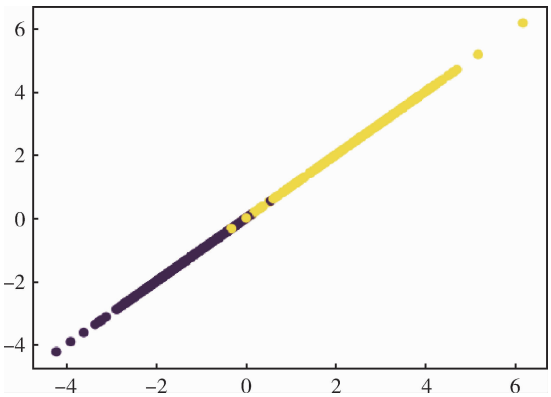

图 3 LDA 降维后效果

以是线性分类器（Gblinear）[12]；同时 XGBoost 也是集成学习模型，基本思想是将多个分类性能较低的树模型即弱分类器集成一个分类准确率较高的强分类器[13]；其支持特征抽样，能有效防止过拟合问题并减少系统计算量[14]。

3.2.2 模型训练流程 其计算定义公式如下：

$$y_i = \varphi(x_i) = \sum_{k=1}^{K} f_k(x_i) \tag{7}$$

其中 y_i 为预测值，K 表示树的数量，x_i 表示第 i 个样本，将 K 棵树的输出加权求和即为 XGBoost 模型的最终预测值。其模型训练函数如下：

$$L(\varphi) = \sum_i l(y_i, h_i) + \sum_k \Omega(f_k) \tag{8}$$

其中第 1 项为损失函数，表示第 i 个样本的预测误差；第 2 项为正则化项，规范模型的复杂度来防止过拟合。原始数据集经过第 1 个弱分类器训练后输出结果，再根据其返回的残差不断调整下一个弱分类器，直至指定数目的分类器训练完成，最后将所有输出进行加权求和即为总预测值。

3.2.3 降维后经过 XGBoost 训练的模型（图 4）

在本实验中，利用机器学习包 sklearn 提供的 grid_ search 网格搜索[15]进行交叉验证获得最优参数来构建 XGBoost 模型，其中模型学习率 learning_ rate 为 0.1，树最大深度 max_ depth 为 2，最小叶子节点样本权重之和 min_ child_ weight 为 5，弱分类器数 n_ estimators 为 66，模型惩罚力度 gamma 为默认值 0，随机采样的每棵树的列数占比 colsample_ bytree 为 0.3，每棵树随机采样的比例 subsample 为 0.9。

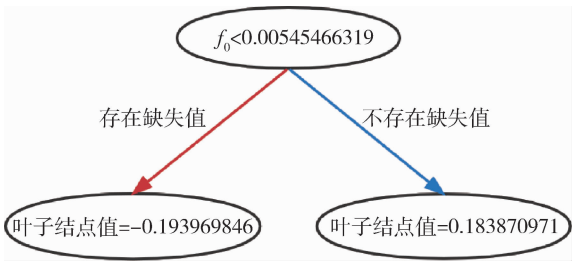

图 4 XGBoost 模型可视化

4 实验分析

4.1 评价指标

4.1.1 混淆矩阵 使用混淆矩阵将分类结果与实际样本信息进行比较，矩阵中每一行代表预测样本类别，每一列代表实际样本类别，见表 2。相关评价指标有准确率、精准率、召回率和 $F1$ 值。准确率为预测正确的正负样本数量（TP + TN）与总样本数量（TP + FN + FP + TN）的比例，是最常用的分类性能指标；精确率为预测正确的正样本数量（TP）与预测的总正样本数量（TP + FP）的比例，代表预测出为正的样本中有多少样本是真实为正的；召回率为正确预测的正样本数量（TP）与真实正样本数量（TP + FN）的比例；$F1$ - score 是目前许多实验推荐的评测指标，是精确率和召回率的调和值，当精确率和召回率接近时，$F1$ - score 值最大。

表 2 混淆矩阵

项目	预测为正样本	预测为负样本
真实为正样本	真正样本 TP	假负样本 FN
真实为负样本	假正样本 FP	真负样本 TN

4.1.2 ROC 曲线 受试者工作特征（Receiver Operating Characteristic，ROC）曲线能够客观衡量模型性能。该曲线的横坐标为假阳性率（False Positive Rate，FPR），即 N 个真实负样本中被预测为正样本的概率；纵坐标为真阳性率（True Positive Rate，TPR），即 M 个真实正样本中被预测为正样本的概率；ROC 曲线越接近左上角，表明该分类模型性能越好，若 ROC 曲线是光滑的，可以判定没有过拟合现象。曲线下面积（Area Under Curve，AUC）即

ROC 曲线下与坐标轴围成的面积，其大小通常处于 0.5～1 之间，AUC 值越大表明模型性能越好、真实性越高。

4.1.3 均方根误差 均方根误差（Root Mean Square Error，RMSE）也叫标准误差，反映预测值与真实值之间的偏差，其数值越小表示模型精度越高、稳定性越好。

4.2 实验结果

4.2.1 性能检验 XGBoost 模型训练完成后用测试集检验其性能。为保证评价方法的科学性、客观性、准确性，在保持模型参数不变的情况下，将原始数据集与降维后的数据集作为输入分别得出预测结果；将 Adaboost、随机森林、朴素贝叶斯算法作为性能比较分类器。据结果可知降维处理后训练的预测模型分类准确率比降维之前平均高出 2.7%，见图 5。降维后模型训练时间大幅缩短，减少了系统计算量。

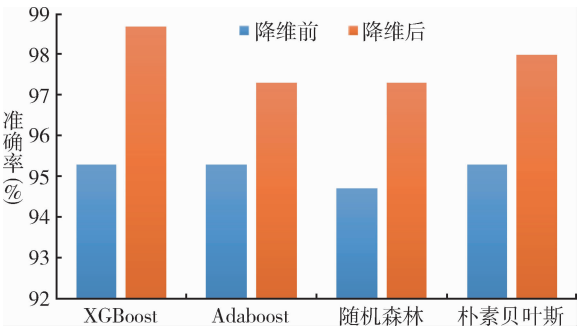

图 5 降维前后各算法模型准确率变化

4.2.2 模型性能检验结果 绘制降维后各算法模型的 ROC 曲线，其中虚线代表某个随机分类器的 ROC 曲线，模型分类性能越好，离此虚线就越远。本实验中关于乳腺癌的分类预测问题，依据评价指标结果可以看出 XGBoost 模型性能要优于其他 3 种模型，其中 XGBoost 模型的分类准确率均值达到 0.987，AUC 指标均值达到 0.984，均方根误差为 0.115，分别比 Adaboost、随机森林和朴素贝叶斯的均方根误差低 4.8%、4.8%、2.6%，见表 3、图 6。

表 3 模型性能结果

模型	准确率	AUC	RMSE
XGBoost	0.987	0.984	0.115
Adaboost	0.973	0.968	0.163
随机森林	0.973	0.968	0.163
朴素贝叶斯	0.98	0.976	0.141

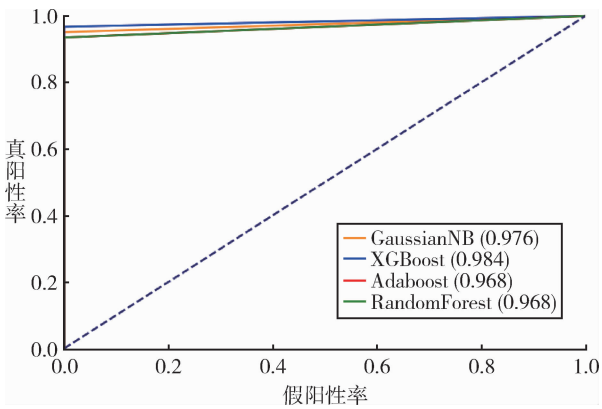

图 6 降维后各算法模型的 ROC 曲线

4.2.3 结果分析 LDA 方法在大幅降低数据维度的同时提高识别精准度、缩短模型训练时间；上述 4 种算法训练的预测模型分类效果均取得较理想分数，其中 XGBoost 模型分类效果最佳，具有可借鉴的医学价值。

5 结语

当前乳腺癌防治形势严峻，因此提高乳腺癌诊断率、早发现早治疗具有重要的现实意义。为提高乳腺癌诊断率、加快诊断速度，本实验利用 LDA 方法将原有多维数据特征合并，降低维度和系统复杂性；从 4 种算法择优选择 XGBoost 构建乳腺癌预测模型，利用网格搜索进行交叉验证提高模型准确率并防止模型过拟合；将降维前的原始数据集同样进行训练和预测，实验结果显示降维处理后训练的预测模型分类性能较降维之前更好，证明基于 LDA 和 XGBoost算法构建的乳腺癌预测模型具有良好的分

类效果，为更好地辅助医生增强医疗临床应用，提高我国医疗水平提供一定技术方法支撑。

参考文献

1 世界卫生组织．乳腺癌成全球最常见癌症［EB/OL］．［2021－06－10］．https：//baijiahao. baidu. com/s? id = 1690756461465319744&wfr = spider&for = pc.

2 吴辰文，李长生，王伟，等．一种改进的 SVM 算法在乳腺癌诊断方面的应用［J］．计算机工程与科学，2017，39（3）：562－566.

3 刘军，彭慧娴，黄斌，等．基于 BP－GamysBoost 的乳腺癌诊断模型［J］．计算机与现代化，2021（4）：8－14.

4 于凌涛，夏永强，闫昱晟，等．利用卷积神经网络分类乳腺癌病理图像［J］．哈尔滨工程大学学报，2021，42（4）：567－573.

5 冯照石，范祺．双重加权朴素贝叶斯算法预测乳腺癌复发率［J］．牡丹江师范学院学报（自然科学版），2021（2）：11－15.

6 余莹，樊重俊，朱人杰，等．基于系统聚类和 SVM 模型的乳腺癌诊断研究［J］．智能计算机与应用，2020，10（11）：97－100，105.

7 常炳国，李玉琴，冯智超，等．基于主成分机器学习算法的慢性肝病的智能预测新方法［J］．计算机科学，2017，44（S2）：65－67，91.

8 董虎胜．主成分分析与线性判别分析两种数据降维算法的对比研究［J］．现代计算机（专业版），2016（29）：36－40.

9 Belhumeur P N，Hespanha J P. Eigenfaces vs. Fisherfaces：Recognition Using Class Specific Linear Projection［J］. IEEE Transactions on Pattern Analysis and Machine Intelligence，1997，19（7）：711－720.

10 Aburomman A A，Reaz M B I. Ensemble of Binary SVM Classifiers Based on PCA and LDA Feature Extraction for Intrusiondetection［C］. Xi′an：Proc. of the Advanced Information Management，Communicates，Electronic and Automation Control，2017：636－640.

11 Chen T，Guestrin C. XGBoost：A Scalable Tree Boosting System［EB/OL］．［2019－03－30］．https：//blog. csdn. net/u014033218/article/details/88921277.

12 CSDN. 机器学习算法之 Xgboost 算法［EB/OL］．［2021－06－12］．https：//blog. csdn. net/qq_ 20412595/article/details/82621744.

13 余天阳．基于 XGBoost 的固态白酒发酵产量预测［J］．计算机与数字工程，2020，48（5）：1233－1237，1251.

14 李占山，刘兆赓．基于 XGBoost 的特征选择算法［J］．通信学报，2019，40（10）：101－108.

15 邓卓，苏秉华，张凯．基于集成学习的乳腺癌分类研究［J］．中国医疗设备，2020，35（12）：59－62.

（上接第 33 页）

21 王伟帅，李阳兵，刘鑫源，等．基于 CiteSpace 对中国高原高血压研究热点与演变的知识图谱分析［J］．中国高原医学与生物学杂志，2020，41（3）：156－164.

22 Ruan T，Huang Y，Liu X，et al. QAnalysis：A Question－Answer Driven Analytic Tool on Knowledge Graphs for Leveraging Electronic Medical Records for Clinical Research［J］. BMC Med Inform Decis Mak，2019，19（1）：82.

23 Fecho K，Balhoff J，Bizon C，et al. Application of MCAT Questions as a Testing Tool and Evaluation Metric for Knowledge Graph－based Reasoning Systems［J］. Clinical and Translational Science，2021，14（5）：1719－1724.

24 李贺，刘嘉宇，李世钰，等．基于疾病知识图谱的自动问答系统优化研究［J］．数据分析与知识发现，2021，5（5）：115－126.

25 Li C，Hang S，Hu X，et al. KnowHealth：A Knowledge Graph based Question－Answer Platform for Elderly People［C］. Weihai：12th EAI International Conference on Mobile Multimedia Communications，Mobimedia，2019.

26 Zhou X，Wu B，Zhou Q. A Depth Evidence Score Fusion Algorithm for Chinese Medical Intelligence Question Answering System［J］. J Healthc Eng，2018，1（3）：1205354.

27 Hasan S，Rivera D，Wu X C，et al. Knowledge Graph－Enabled Cancer Data Analytics［J］. IEEE J Biomed Health Inform，2020，24（7）：1952－1967.

28 Shenoi S J，Vi L，Sarvesh S，et al. Developing a Search Engine for Precision Medicine［J］. AMIA Summits on Translational Science Proceedings，2020（2020）：579－588.

29 Struck A，Walsh B，Buchanan A，et al. Exploring Integrative Analysis Using the BioMedical Evidence Graph［J］. JCO Clin Cancer Inform，2020（4）：147－159.