文章编号：2095-2163(2022)12-0104-06 中图分类号：R737.9;TP181 文献标志码：A

XGBoost 算法在乳腺癌辅助诊断中的应用

李佳思

（上海工程技术大学 数理与统计学院，上海 201620）

摘 要：为了研究 XGBoost 算法在乳腺癌诊断中的应用，将机器学习算法与疾病诊断相结合，提升乳腺癌的诊断效率与准确率。在公开的威斯康星州乳腺癌数据集中建立了 XGBoost 乳腺癌诊断模型，实验过程中通过网格搜索和学习曲线寻找诊断模型的最优参数，并将建立的 XGBoost 模型与决策树、支持向量机、K-近邻和朴素贝叶斯算法的预测效果进行比较，使用准确率、F_1 值、*AUC* 值等评价指标来比较各个模型的预测性能。本文实验结果表明，XGBoost 算法的五折交叉验证准确率达到了 97.89%，优于其他单分类器算法，对提升乳腺癌诊断准确率具有现实意义。

关键词：乳腺癌；疾病诊断；集成学习；XGBoost

Application of XGBoost algorithm in breast cancer diagnosis

LI Jiasi

（School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai 201620, China）

【Abstract】 In order to study the application of XGBoost algorithm in breast cancer diagnosis, machine learning algorithm is combined with disease diagnosis to improve the efficiency and accuracy of breast cancer diagnosis. The XGBoost breast cancer diagnostic model is established in the open Wisconsin Breast Cancer Database. In the experiment, the optimal parameters of the diagnostic model are obtained by grid search and learning curve. The proposed XGBoost model is compared with the prediction results of Decision Tree (DT), Support Vector Machine (SVM), K- nearest neighbor (KNN) and Naive Bayes (NB) algorithm, and *Accuracy*, *Recall*, F_1 value and *AUC* are used to compare the prediction performance of each model. Experimental results show that *Accuracy* of 5-fold cross-validation of XGBoost algorithm is 97.89%, which is better than other single classifier algorithms. The research has practical significance for improving the accuracy of breast cancer diagnosis.

【Key words】 breast cancer; disease diagnosis; ensemble learning; XGBoost

0 引 言

目前，高效率的工作追求改变了人们生活方式和饮食习惯，癌症成为威胁人们健康的常见疾病之一。对女性而言，乳腺癌则已位于癌症发病率的前列，并对国内女性的身体健康造成严重危害。随着医疗技术水平的提升，通过药物、手术等方式可以使早期癌症患者的病情得到有效遏制，而晚期患者的治愈率大大降低，因此，早发现早治疗是提高癌症治愈率的有效手段[1-2]。传统的乳腺癌诊断方式有 X 射线检查、针吸细胞学检查、B 型超声检查和活组织检查等，一般需要经验丰富的医生根据检查结果判定肿瘤恶良性，癌细胞是否已经发生转移和扩散等[3]。由于受到患者健康意识、各地区医疗水平、医疗资源分配以及医生临床经验等因素的影响，现实中乳腺癌误诊、漏诊的情况时有发生。

近年来，医疗器械产品的不断创新推出，能够帮助医生更好地判断患者病情，但是日益繁多的检查指标也使病情与病因的联系更加复杂。同时，计算机硬件和软件性能的高速提升使得海量的医疗数据得以储存，患者病史、生活习惯、家族遗传等因素都可能成为需要考虑的因素，海量的信息加重了医生诊疗的负担。随着人工智能的兴起，机器学习算法已经广泛应用于各个领域并取得了重大突破，这为癌症诊断提供了新的方向，基于机器学习和模式识别的疾病诊断与发病机制研究成为热点研究方向。传统的机器学习模型主要有决策树(DT)、支持向量机(SVM)、K-近邻(KNN)、朴素贝叶斯(NB)等，虽然在很多疾病分类问题上取得了可观的进步，但也具有局限性，如泛化能力不强、在不同数据集中模型表现差距较大等问题。集成学习可以很好地解决上述问题，其主要思想是：训练多个同质或异质的弱分类器对样本进行预测，获得多个预测结果，然后根据某种规则将各个弱分类器的预测结果进行结合得到

作者简介：李佳思(1996-)，女，硕士研究生，主要研究方向：机器学习、生物统计。

通讯作者：李佳思 Email：m130319107@sues.edu.cn

收稿日期：2022-03-09

最终预测[4-5]。这种方法提升了模型的泛化能力，具有鲁棒性，通常能够得到比单个分类器更好的预测准确率，在人脸识别、文本分类、疾病辅助诊断等方面得到了广泛的应用。

目前许多学者将机器学习算法用于乳腺癌的诊断中，取得了良好的分类效果。邓卓等人[6]将集成学习算法用于乳腺癌的诊断研究，在美国威斯康星州乳腺癌数据集上分别建立了随机森林和 XGBoost 诊断模型，并与决策树算法的分类性能进行比较，结果表明，集成学习算法相比传统的决策树分类模型具有更高的分类准确率。张红斌等人[7]提出了一种改进的自适应提升算法用于乳腺癌图像识别，从不同角度提取图像特征并进行特征融合，在 CBIS-DDSM 数据集上分别建立了逻辑回归、随机森林、AdaBoost、梯度提升决策树等诊断模型，并将改进的 ERGS 算法与传统算法进行结合用于乳腺癌诊断。实验结果显示，所提出的 ERGS-Ada 算法的预测准确率最高，达到了 86.24%，对阳性图像的识别精度达到了 99.18%。Khuriwal 等人[8]提出了一种自适应集成投票方法进行乳腺癌检测，在威斯康星乳腺癌数据集上验证了算法的有效性，使用卡方检验和递归特征消除方法选择出了影响模型识别准确率的 16 个特征，建立了逻辑回归和神经网络模型，采用投票法获得了最终预测结果，实现了 98.50%的分类准确率。Sun 等人[9]提出了基于深度学习的集成 CNN 模型，对 266 例乳腺癌患者的术前磁共振成像和分子信息进行训练来区分 2 种乳腺癌亚型（管腔型和非管腔型），结果显示，所提出的模型在测试集中的识别准确率达到了 85.2%。岳鹏等人[10]提出了一种 XLC-Stacking 方法进行乳腺癌诊断，该方法基于 XGBoost 算法选择出影响乳腺癌分类的最佳特征，并使用威斯康星州乳腺癌的诊断数据来验证模型有效性。结果表明，所提出的 XLC-Stacking 方法与单一的 XGBoost 和一般的 Stacking 方法相比准确率得到提升，分类准确率为 97.73%。Hou 等人[11]使用四川大学华西医院的数据分别建立了深度神经网络、XGBoost 和随机森林模型来预测中国居民患乳腺癌的风险，实验结果显示，XGBoost、深度神经网络和随机森林模型的 *AUC* 值分别为 0.742、0.728 和 0.728。

为了弥补传统的乳腺癌诊断方式的不足，提升诊断效率和精度，本文将 XGBoost 算法用于乳腺癌诊断，将数据的 70%作为训练集、30%用于测试，同时建立了决策树、支持向量机、K-近邻和朴素贝叶斯分类模型，使用五折交叉验证准确率、F_1 值、*AUC* 值等评估指标来评价模型的优劣。

1　相关理论介绍

1.1　XGBoost 算法

XGBoost（Extreme Gradient Boosting）算法在 2014 年由陈天奇博士提出，这是对梯度提升算法的一种改进，近年来在各大机器学习和数据挖掘类的比赛中取得了优异的成绩。提升算法的思想是从某个弱学习器开始，不断学习得到不同的弱分类器，然后将这些弱分类器通过某种策略进行融合得到更加强大的分类器。Boosting 方法中各个弱分类器是相互关联的，每次迭代根据前一个分类器的结果来调整样本的权重，通过加大容易被错误分类的样本的权重使得下阶段生成的弱分类器更加关注被错误分类的样本。Boosting 方法的工作原理如图 1 所示。

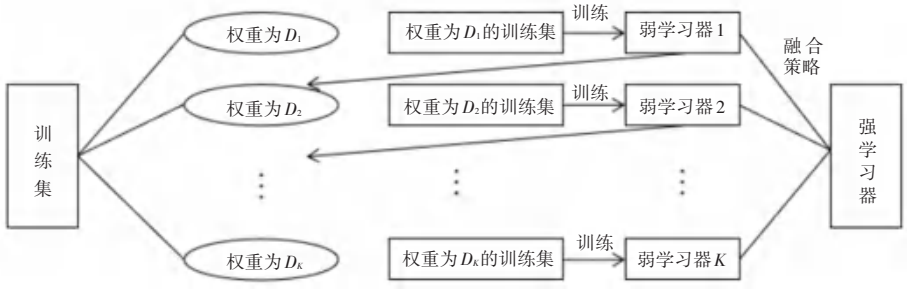

图 1　Boosting 算法原理图

Fig. 1　Schematic diagram of Boosting algorithm

XGBoost 算法由多棵决策树组成，模型以前 $k-1$ 棵树生成的模型所产生的残差来建立第 k 棵树，从而不断提升模型的预测精度。XGBoost 算法的流程如下：

给定二分类数据集，每个样本由特征 X 和标签 Y 构成：

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\} \quad (1)$$

其中，$x_i \in R^n$，$y_i \in \{+1,-1\}$，+1 和-1 分别

表示 2 个不同的类别。

预测模型可表示为：

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{2}$$

其中，K 为树的个数；$f_k(x_i)$ 表示第 k 棵树；$\hat{y}_i$ 表示样本 x_i 的预测值。

XGBoost 的目标函数为：

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3}$$

其中，$l(y_i, \hat{y}_i)$ 表示第 i 个样本的训练误差；n 表示样本个数；$\Omega(f_k)$ 为第 k 棵树的正则项。正则项可以减小模型的过拟合风险，降低模型复杂度。其定义如下：

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w^2{}_j \tag{4}$$

其中，T 为叶子节点个数，γ 和 λ 为控制参数。

可将上述目标函数写为：

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i{}^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) = \sum_{i=1}^{n} l(y_i, \hat{y}_i{}^{(t-1)} + f_t(x_i)) + \Omega(f_t) + C \tag{5}$$

其中，C 表示常数。

使用泰勒展开可将上述目标函数近似为：

$$Obj^{(t)} \simeq \sum_{i=1}^{n} [l(y_i, \hat{y}_i{}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) + C \tag{6}$$

其中，

$$g_i = \frac{\partial l(y_i, \hat{y}_i{}^{(t-1)})}{\partial \hat{y}_i{}^{(t-1)}} \tag{7}$$

$$h_i = \frac{\partial^2 l(y_i, \hat{y}_i{}^{(t-1)})}{\partial (\hat{y}_i{}^{(t-1)})^2} \tag{8}$$

去掉常数项可得：

$$Obj^{(t)} \simeq \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{9}$$

上述目标函数可改写为：

$$Obj^{(t)} \simeq \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j{}^2 \tag{10}$$

公式(10)可以写为：

$$Obj^{(t)} = \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2}(H_j + \lambda) w_j{}^2 \right] + \gamma T \tag{11}$$

其中，$G_j = \sum_{i \in I_j} g_i, H_j = \sum_{i \in I_j} h_i$。

对式(11)求导，令一阶导数为零可得：

$$w_j{}^* = -\frac{G_j}{H_j + \lambda} \tag{12}$$

将式(12)代入目标函数可得：

$$Obj^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j{}^2}{H_j + \lambda} + \gamma T \tag{13}$$

在此，将 Obj^* 作为评价决策树结构好坏的函数，取值越低表示模型越好。

1.2 模型评价方法

混淆矩阵[12-13]是分类任务中常用的评价方法，可以将数据的真实情况和预测结果清晰地显示出来，矩阵中 *Positive* 表示正例，*Negative* 表示负例。以二分类模型为例，混淆矩阵见表 1。

表 1 混淆矩阵

Tab. 1 Confusion matrix

真实情况	预测结果	
	正例	负例
正例	*TP*(真正例)	*FN*(假负例)
负例	*FP*(假正例)	*TN*(真负例)

由混淆矩阵可以计算出准确率、精准率、召回率、F_1 值和 *AUC* 值，多角度评估模型性能。对此拟做探讨分述如下。

(1)准确率 (*Accuracy*)：表示分类器预测正确的样本个数与样本总数之比。通常，模型的准确率越高，表示模型分类效果越好。准确率计算公式如下：

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{14}$$

(2)精准率 (*Precision*)：阳性病例中正确分类的样本数量与所有预测为阳性的样本数量之比。计算公式如下：

$$P = \frac{TP}{TP + FP} \tag{15}$$

(3)召回率 (*Recall*)：阳性病例中正确分类为阳性的样本数量与真实为阳性的样本数量之比。计算公式如下：

$$R = \frac{TP}{TP + FN} \tag{16}$$

(4) F_1 – *Score*：表示精准率与召回率的调和平

均。计算公式如下：

$$F_1 = \frac{2P \cdot R}{P + R} \tag{17}$$

（5）*AUC*：表示 *ROC* 曲线下面积，值越接近于 1 表示模型性能越好。

2　数据来源及预处理

2.1　数据来源

本文实验数据来自 UCI 数据库中的威斯康星州乳腺癌诊断数据集。数据集共包含 569 条数据样本，32 个特征，分别含有 212 例恶性样本和 357 例良性样本。其中，ID 为患者编号，*diagnosis* 为样本标签，*M* 表示恶性病例，*B* 表示良性病例。除去患者编号和 *diagnosis* 两列后剩余的 30 个特征是从患者乳房肿块的细针抽吸数字影像中提取的数据，表示了细胞的 10 个细胞核信息（平均值、标准误差和最大值），分别为半径（radius）、纹理（texture）、周长（perimeter）、面积（area）、平滑度（smoothness）、密实度（compactness）、凹度（concavity）、凹点（concave points）、对称性（symmetry）和分形维数（fractal_demension）。

2.2　数据预处理

探索性数据分析是建立模型前的重要步骤，有助于更好地了解数据结构，发现数据中的异常值及特征之间的相关关系，帮助选择合适的数据处理方法和预测模型。实验前需要对数据进行清洗来提升模型的拟合能力，图 2 展示了数据集中前 9 个特征的描述性统计，可以看出，不同特征之间存在较大的量纲差异，因此，在建立模型之前对数据进行标准化处理。图 3 绘制了前 10 个特征的箱线图，描述了数据的分布信息，从箱线图中可以分析数据的对称性、分散程度、偏态及异常值等信息。由图 3 可以看出，周长均值和面积均值两个特征具有相似的分布，由此可以推测二者具有一定相关关系。

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
mean	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919	0.181162
std	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803	0.027414
min	6.981000	9.710000	43.790000	143.500000	0.052630	0.019300	0.000000	0.000000	0.106000
25%	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310	0.161900
50%	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500	0.179200
75%	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	0.195700
max	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	0.304000

图 2　部分特征的统计性描述

Fig. 2　Statistical description of some features

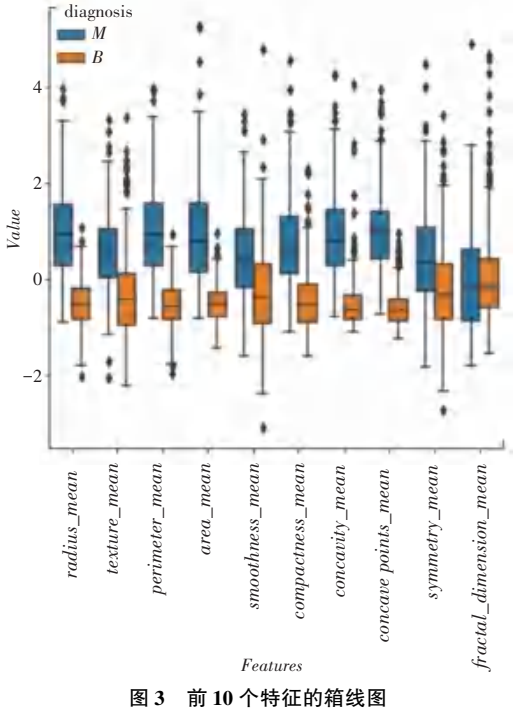

图 3　前 10 个特征的箱线图

Fig. 3　Box diagram of the first 10 features

3　实验过程与结果

3.1　实验过程

本文将 XGBoost 模型用于乳腺癌患者的诊断数据对乳腺肿块的恶良性进行预测。实验过程主要包括数据预处理、模型建立和结果分析。其中，数据预处理主要通过统计学方法了解数据特性，对数据进行清洗；模型建立即是对数据的拟合过程，根据研究目标，选择合适的预测模型和参数，本文使用网格搜索和交叉验证寻找模型的最优参数；结果分析主要是通过模型评估指标对模型性能进行评价和分析。本文实验步骤如图 4 所示。交叉验证是建立模型时常用的一种方法，可以获得更加稳定的模型。*K* 折交叉验证即是将数据平均分为 *K* 份，每次建模使用其中一份作为测试集，剩余 *K* − 1 份作为训练集，共进行 *K* 次训练和测试，最后返回 *K* 次测试结果的平均值来增加结果的可靠性[14]。实验过程中，将 70% 的样本作为训练集，30% 的样本作为测试集，通过网格

搜索获得模型的最佳参数组合,以五折交叉验证准确率、F_1 值和 *AUC* 值来对模型性能进行评价。

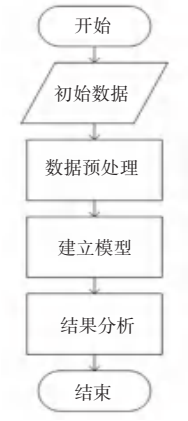

图 4 实验流程

Fig. 4 Experimental process

3.2 XGBoost 诊断模型的建立

实验使用 70%的数据作为训练集,30%的数据作为测试集,在经过预处理后的数据上建立 XGBoost 模型。合理调整模型参数有利于降低模型过拟合风险,提升模型预测能力,实验过程中使用了五折交叉验证准确率作为评价标准确定模型的最优参数。通过学习曲线可以直观地看到模型性能随着参数的变化情况,设置 *n_estimators* 的迭代范围为 10 到 200,以 10 为步长,得到五折交叉验证准确率随弱分类器个数的变化关系如图 5 所示。由图 5 可以看出,当弱分类器数量小于 50 时,模型五折交叉验证准确率随着弱分类器数量的增加而升高,在弱分类器数量达到 100 时,模型的分类准确率最高、为 97.72%,随后再增加分类器数量,模型的分类准确率先下降、后趋于平稳。

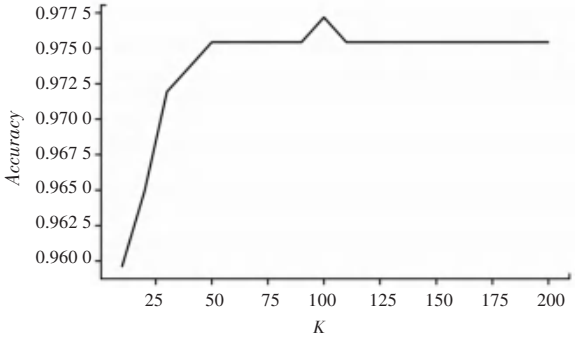

图 5 分类准确率与基学习器个数的关系

Fig. 5 Relationship between classification accuracy and the number of base learners

经过实验验证,在 *random_state* = 10 的情况下,*n_estimators* = 100,*learning_rate* = 0.65 时模型的五折交叉验证准确率达到最高、为 97.89%。XGBoost 模型在测试集上的混淆矩阵如图 6 所示。

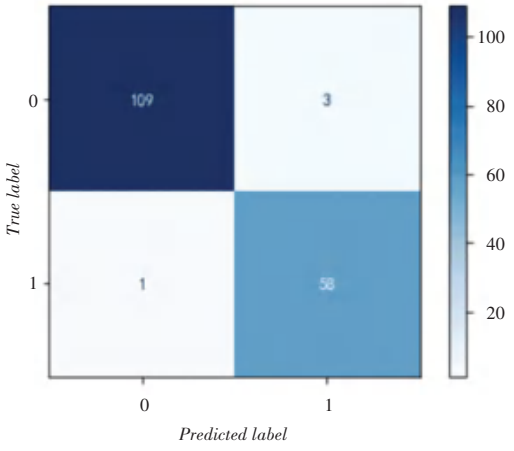

图 6 XGBoost 模型的混淆矩阵

Fig. 6 Confusion matrix of XGBoost model

3.3 实验结果分析

为了将 XGBoost 模型的诊断效果与其他模型进行比较,在威斯康星州乳腺癌数据中分别建立了决策树、支持向量机、K-近邻和朴素贝叶斯分类模型,得到不同模型的精准率、召回率、F_1 值、*AUC* 值以及五折交叉验证准确率见表 2。

表 2 不同模型的分类结果

Tab. 2 Classification results of different models

模型	*Precision*	*Recall*	F_1 – *Score*	*AUC*	*Accuracy*/%
DT	0.892	0.983	0.935	0.960	94.55
SVM	0.919	0.966	0.942	0.996	97.36
KNN	0.983	0.983	0.983	0.997	96.84
NB	0.903	0.949	0.926	0.984	92.79
XGBoost	0.950	0.966	0.958	0.998	97.89

由表 2 可知,XGBoost 诊断模型的分类准确率为 97.89%,高于其他 4 种单分类器的准确率,支持向量机的准确率排名次之,达到了 97.36%;紧随其后是 KNN 和决策树模型,准确率分别为 96.84%和 94.55%,朴素贝叶斯的预测效果最差,准确率为 92.79%。5 个模型中 XGBoost 模型的 *AUC* 值达到最高、为 0.998,KNN 排名次之、为 0.997,其次为 SVM 和 NB,分别为 0.996 和 0.984,决策树的 *AUC* 值最低、为 0.960。由结果可知,整体来看 XGBoost 模型的分类效果优于其他 4 种单分类器,XGBoost 算法在乳腺癌诊断中可以提升模型的泛化能力,将机器学习算法应用于医疗数据与疾病诊断中,建立疾病诊断辅助系统,能够有效提升诊断效率和准确率,具有一定现实意义。

建立 XGBoost 模型后得到特征重要性排序如图

7 所示。由图 7 可知，*area_se*、*texture_worst*、*texture_mean*、*symmetry_se*、*symmetry_worst*、*concave points_worst*、*perimeter_worst* 是影响乳腺癌恶良性的较重要特征。这对于找出影响乳腺癌恶良性的因素十分重要，医生在为患者进行诊断时可以优先关注这些特征，有利于发现更加有效的治疗方式，对降低乳腺癌死亡率具有一定现实意义。

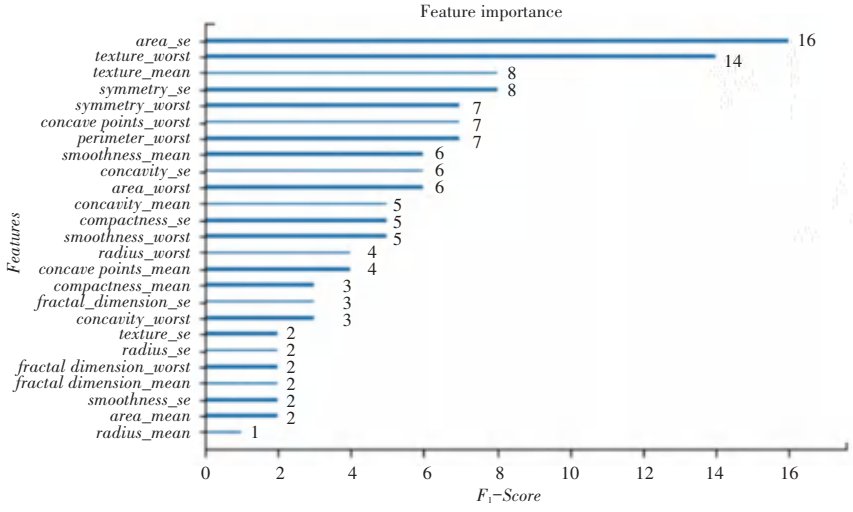

图 7　特征重要性排序

Fig. 7　Feature importance ranking

4　结束语

本文将 XGBoost 模型应用于乳腺癌诊断，使用 UCI 数据库中的威斯康星州乳腺癌数据集验证了模型的有效性，同时建立了决策树、支持向量机、K-近邻和朴素贝叶斯分类模型，通过五折交叉验证准确率、F_1 值、*AUC* 值等评估指标对比了不同模型的分类性能。由实验结果可知，XGBoost 模型的预测效果优于其他 4 种分类模型，取得了 97.89%的分类准确率，一定程度上说明集成学习相比单分类器具有更好的泛化能力，在今后可以将集成学习应用于乳腺癌的诊断来提升诊断准确率。此外，XGBoost 模型给出了特征的重要性排名，可以发现不同特征对分类结果的影响，进而为医生提供决策指导。此研究对识别早期乳腺癌患者并进行针对性治疗具有现实意义，后续研究中可以将其他集成学习方法用于乳腺癌辅助诊断，也可将 XGBoost 算法应用于其他类型的疾病进一步探究模型的表现及其适用性。

参考文献

[1] CAPLAN L. Delay in breast cancer: Implications for stage at diagnosis and survival[J]. Frontiers in Public Health, 2014, 2: 87.

[2] ABDAR M, ZOMORODI-MOGHADAM M, ZHOU X, et al. A new nested ensemble technique for automated diagnosis of breast cancer[J]. Pattern Recognition Letters, 2020, 132: 123-131.

[3] 鲍琦莉. 基于分类监督学习算法的乳腺癌预测诊断研究[D]. 海口：海南大学，2020.

[4] 陈思萱. 基于机器学习的乳腺癌导诊和诊断预测研究[D]. 兰州：西北师范大学，2021.

[5] 石胜源，朱磊，叶琳，等. 基于随机森林算法的心血管疾病预测研究[J]. 智能计算机与应用，2021，11(04)：176-178，181.

[6] 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.

[7] 张红斌，邬任重，蒋子良，等.基于改进的自适应提升算法的乳腺癌图像识别研究[J]. 华中师范大学学报(自然科学版)，2020，54(06)：935-943.

[8] KHURIWAL N, MISHRA N. Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm[C]//2018 IEEMA Engineer Infinite Conference (eTechNxT). India: IEEE, 2018: 1-5.

[9] SUN Rong, MENG Zijun, HOU Xuewen, et al. Prediction of breast cancer molecular subtypes using DCE-MRI based on CNNs combined with ensemble learning[J]. Physics in Medicine & Biology, 2021, 66(17): 175009.

[10] 岳鹏，侯凌燕，杨大利，等. 基于 XGBoost 特征选择的疾病诊断 XLC-Stacking 方法[J]. 计算机工程与应用，2020，56(17)：136-141.

[11] HOU Can, ZHONG Xiaorong, HE Ping, et al. Predicting breast cancer in Chinese women using machine learning techniques: Algorithm development[J]. JMIR Medical Informatics, 2020, 8(6): e17364.

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

[13] 郭海湘，黄媛玥，顾明赟，等. 基于自适应多分类器系统的甲状腺疾病诊断方法研究[J]. 系统工程理论与实践，2018，38(08)：2123-2134.

[14] 郑惠文. 机器学习算法在内科疾病诊断中的应用[D]. 哈尔滨：哈尔滨工业大学，2020.