

文章编号: 2095-4980(2023)08-1014-08

基于 BP_Adaboost 模型的乳腺癌诊断预测方法研究

葛梦飞¹, 李赵旭¹, 刘嘉欣¹, 王宏伟^{1,2}, 王 佳³

(1. 新疆大学 电气工程学院, 新疆 乌鲁木齐 830047; 2. 大连理工大学 控制科学与工程学院, 辽宁 大连 116024;
3. 大连医科大学 基础医学院, 辽宁 大连 116041)

摘 要: 乳腺癌是全球女性发病率位居首位的恶性肿瘤, 研究基于神经网络模型的乳腺癌诊断预测方法的目的是将临床与机器学习相结合, 有助于医疗工作者更加快速准确地判断出患病与否, 同时解决现有模型中存在的过拟合以及漏诊率和误诊率过高的问题, 并提高预测模型的准确率。本文采用加州大学欧文分校(UCI)数据集, 共 669 个样本, 其中包含 357 个良性样本和 212 个恶性肿瘤样本, 共计 10 个特征训练预测模型。将 10 个神经网络模型采用 Adaboost 方法相结合, 即通过 Adaboost 算法组合多个弱分类器从而形成一个强分类器, 最终输出一个具有更高准确率、有较强的自学习能力、自适应能力且泛化性能优良的集成预测模型。结论表明, 该模型的预测准确率达到 98.5507%, 同时准确率(AUC)为 0.9966, 说明所建模型区分度较好, 可以反映模型的诊断价值, 且非常稳定, 具有非常好的验证效果, 为临床应用提供进一步的技术支持和保障。

关键词: 乳腺癌; 早期诊断; 神经网络; 分类器; 预测模型

中图分类号: TP181; R737.9

文献标志码: A

doi: 10.11805/TKYDA2021094

Diagnosis and prediction of breast cancer based on BP_Adaboost model

GE Mengfei¹, LI Zhaoxu¹, LIU Jiaxin¹, WANG Hongwei^{1,2}, WANG Jia³

(1.School of Electrical Engineering, Xinjiang University, Urumqi Xinjiang 830047, China;

2.School of Control Science and Engineering, Dalian University of Technology, Dalian Liaoning 116024, China;

3.College of Basic Medical Sciences, Dalian Medical University, Dalian Liaoning 116041, China)

Abstract: Breast cancer is the first malignant tumor in women worldwide. Studying on breast cancer diagnosis and prediction methods based on neural network models is to combine clinical and machine learning to help medical workers more quickly and accurately determine the disease or not, and solve the problems of over-fitting, missed diagnosis rate and high misdiagnosis rate in existing models, and improve the accuracy of prediction models. The University of California Irvine(UCI) data set contains 669 samples, including 357 benign samples and 212 malignant tumor samples, a total of 10 features to train the prediction model. The 10 neural network models are combined through Adaboost method, that is, multiple weak classifiers are combined by Adaboost algorithm to form a strong classifier. The final output is an integrated prediction model with higher accuracy, stronger self-learning ability, adaptive ability and excellent generalization performance. The conclusion shows that the prediction accuracy of the model is 98.550 7%, and the Accuracy(AUC) is 0.996 6, which indicates that the established model is very stable, and has good discrimination and good verification effects. It provides further technical support and guarantee for clinical application.

Keywords: breast cancer; early diagnosis; neural network; classifier; prediction model

2018 年 2 月, 国家癌症中心发布了最新癌症数据, 乳腺癌逐渐年轻化, 且逐渐成为全球女性发病率最高的癌症之一^[1]。根据统计数据显示, 乳腺癌新发病例每年平均增加 22.9%, 到 2020 年预计达到 1 785 865 例, 中国女性乳腺癌发病率的增长速度已接近全球发病率的 2 倍, 已成为当前社会的重大公共卫生问题。由于乳腺癌易发生

收稿日期: 2021-03-03; 修回日期: 2021-04-30

基金项目: 国家自然科学基金资助项目(61863034)

*通信作者: 王 佳 email:jjawang@mail.dlut.edu.cn

转移,随着发病率的升高,对适龄女性进行科学有效的早期筛查是控制乳腺癌发展和死亡率最为有效的方法^[2]。

一直以来,国内都采用人工筛查的方式进行乳腺癌分析,人工筛查的假阳性率可达10%~15%,致使一些未患病的受检者需要再次进行影像诊断,甚至是组织病理检查,对身体造成了不必要的创伤^[3]。因此国内外学者对于早期诊断预测模型进行了大量的研究。根据分析和诊断的手段等不同,建模方法主要可以分为以下2类:a) 基于乳腺癌图像的分析研究。计算机辅助设计(Computer Aided Design, CAD)通过医学图像处理技术对医学图像和检验数据进行分析, CAD技术的出现可有效缓解医生的工作压力,提高医生诊断的敏感性和特异性^[4]。吴辰文等^[5]为优化CAD技术的准确率,提出了一种基于随机森林模型下 Gini 指标特征的支持向量机算法,其准确率达97.7%。b) 基于临床数据的分析研究。赖胜圣等^[6]构建基于序列前向选择算法与支持向量机分类器融合的乳腺癌预测模型,提高了计算机辅助诊断技术对乳腺癌细针穿刺细胞病理的准确率。上述各方法的分类预测性均有不同程度的提高,但均存在泛化性能较差的问题,容易出现用于其他数据集时预测率显著降低的情况。为解决临床上因为各种原因而产生的误诊以及漏诊等问题,需要进一步提高模型准确率与模型的稳定性,同时降低泛化误差,建立临床有效的乳腺癌诊断预测模型。

针对以上问题,本文提出一种基于反向传播(Back Propagation, BP)_Adaboost算法构建的乳腺癌模型预测方法。通过BP神经网络组合强分类器,根据乳腺癌特征基因构建乳腺癌预测模型,实现将病人的情况进行分类判断。最后,采用Wisconsin医院收集的乳腺癌数据样本,分别将本文的方法与其他文献中提出的方法进行比较,证明本文方法的有效性,从而为临床研究提供理论支持。

1 理论与方法

1.1 BP神经网络算法

BP网络是一种多层前馈网络,按误差逆传播算法训练,因此在此基础上提出了BP神经网络算法。通过任意选择一组权值,将给定的目标输出直接作为线性方程的代数和,建立线性方程组,得出待求权值。但传统的BP神经网络容易陷入局部最优且泛化能力不强,而Adaboost通过反复训练不同的弱学习器,通过组合输出使弱学习器互补,从而有效提高学习器的泛化能力。

在神经网络中,连接权与阈值的计算是非常关键的因素,通过连接权的选择与阈值的设定,使样本输出偏差限定在一个极小的范围内,BP神经网络流程如下:

Step1 网络初始化;

Step2 设定输入值与期望输出值;

Step3 训练满足实验要求的网络模型;

Step4 计算输出结果,并将输出结果与期望输出值做比较,若满足误差范围,则继续;若不满足误差范围,则执行Step5;

Step5 反向传播:计算误差,更新权值与阈值。

具体算法流程如图1所示。

1.2 BP_Adaboost算法概述

BP神经网络结构简单,运算速度快,研

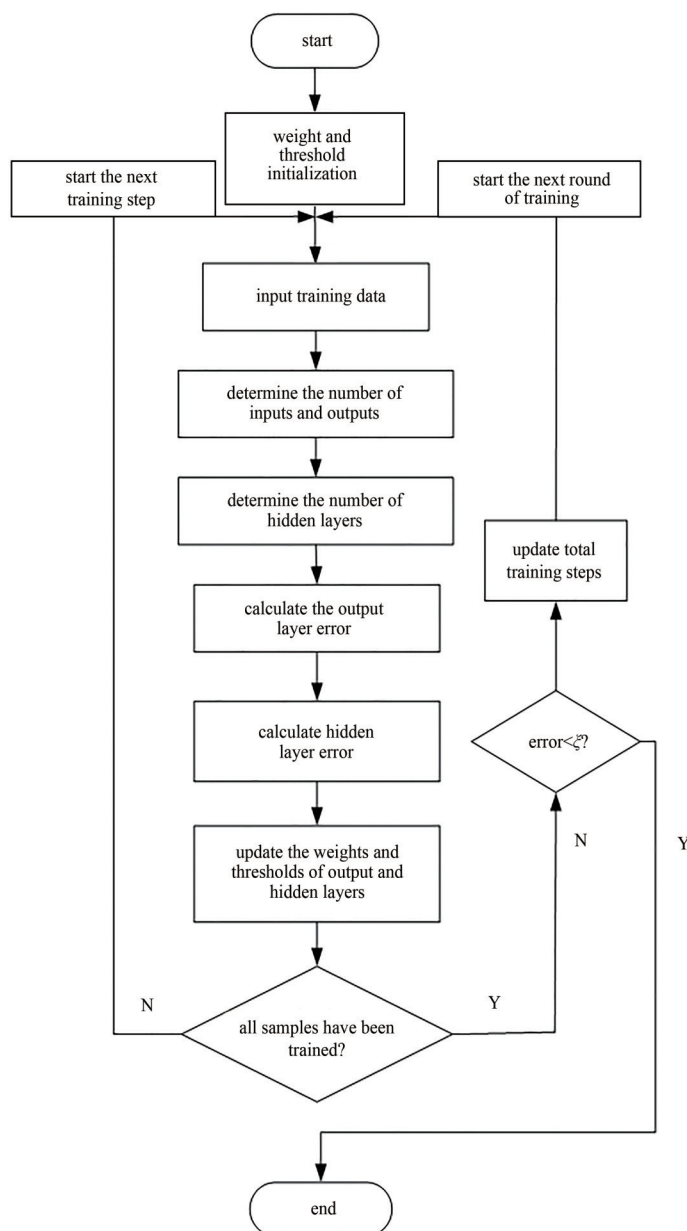


Fig.1 Flow chart of BP algorithm

图1 BP算法流程图

究证明, 3 层神经网络可以逼近任意的非线性关系映射, 适用于工程应用。而通过加权集成的 Adaboost 预测准确率高, 泛化能力强。本文充分总结 BP 神经网络与 Adaboost 算法的优势并将其结合, 既可以改善单个神经网络训练准确率低, 又可以避免传统 Adaboost 算法随训练样本增加, 训练速度突然变慢的问题。换言之, 利用 BP 神经网络的学习能力对训练集进行训练, 在提高验证集的预测准确率的同时, 结合 Adaboost 的高精确度以及高泛化的能力, 进一步提高模型的预测准确率, 从而得到更优秀的模型。Adaboost 算法的核心思想是合并多个“弱”分类器的输出以产生有效分类^[7]。Adaboost 算法通过反复搜索样本特征空间, 获取样本权重, 并在迭代过程中不断更新样本权重, 减小预测准确率高的样本权重, 提高预测准确率低的样本权重, 通过线性组合, 使构造的多个 BP 弱训练模型加权累积到总模型中。每个弱训练器需串行训练且相互依赖, 算法流程图如图 2 所示。

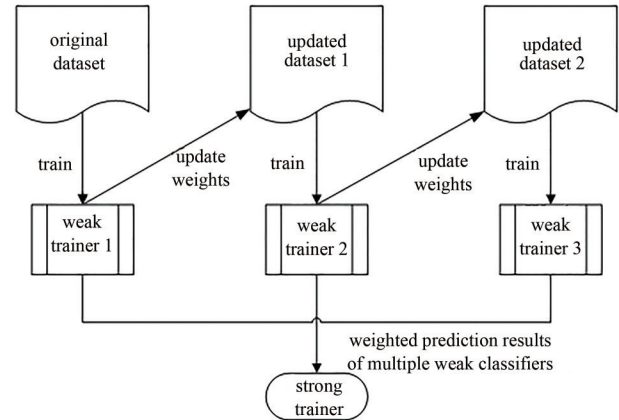


Fig.2 Flowchart of BP_Adaboost algorithm

图2 BP_Adaboost算法流程图

1.3 BP_Adaboost 算法

输入: 数据集 $A = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, q 个基本分类器。 $f_1, f_2, \dots, f_T$ 表示 q 个弱分类器在样本 x 上迭代运行 T 次后的分类函数。

输出: 强分类器。

Step1 初始化训练样本的权重。

for $t \in \{1, 2, \dots, q\}$ do

$$D_t(i) = 1/q$$

end for

Step2 对 q 个弱分类器进行分类预测, 其中 $g(t)$ 为预测分类结果。

for $t \in \{1, 2, \dots, q\}$ do

$g(t)$

end for

Step3 计算弱分类器分类误差率以及预测序列权重。

for $i = 1, 2, \dots, q$, $g(t) \neq y$ do

$$e_t = \sum_i D_t(i)$$

$$a_t = \frac{1}{2} \ln \left(\frac{1 - e_t}{e_t} \right)$$

end for

Step4 更新权重分布。 B_t 为规范化因子, 为了在权重比例不变的情况下使分布权值和为 1,

for $i = 1, 2, \dots, m$, do

$$D_{t+1}(i) = \frac{D_t(i)}{B_t} \times \exp[-a_t y_i g_t(x_i)]$$

end for

Step5 将训练 T 轮后得到 T 组弱分类函数 $f(g_t, a_t)$ 组合得到强分类函数

for $i = 1, 2, \dots, m$, $f(g_t, a_t)$, do

$$h(x) = \text{sign} \left[\sum_{t=1}^T a_t \times f(g_t, a_t) \right]$$

end for

2 基于BP_Adaboost的乳腺癌诊断模型构建

2.1 数据描述

本文采用的数据集是由 Wisconsin 医院 William H Wolberg 博士等收集和整理的基于乳腺肿瘤病灶组织的细胞核图像数据库具体特征信息。该数据集共包含 699 个样本，经过整理与删除存在缺失的样本之后剩余 569 个样本。经过医疗领域的专家诊断结果，将乳腺癌数据的样本标签分为 2 类：-1(良性)与 1(恶性)。采用随机抽样方法，将 569 组样本随机排列后，从中抽取 500 组样本作为训练集用来训练模型，剩下的 69 组样本作为测试集，并采用十折交叉验证，用来测试模型的准确率。数据组成如图 3 所示。

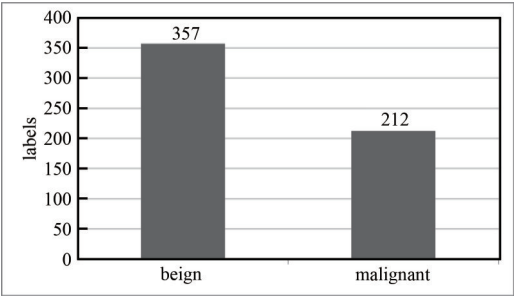


Fig.3 Data composition
图3 数据组成

2.2 网络层数设计

本文采用改进的BP神经网络方法，对数据集进行训练测试。BP神经网络的拓扑结构和节点数对网络模型预测精确度非常重要，网络层数过少，会导致训练结果准确率较低；网络层数过多，则会导致计算量过大。本研究筛选 10 个变量作为输入神经元，是否会发生乳腺癌作为输出神经元，不断调整神经网络使模型更优化，最终确定为 3 层模型结构，即传统的输入层、隐含层、输出层。隐含层节点选取如表 1 所示。

表 1 隐含层节点选取
Table1 Hidden layer node selection

neural network	hidden layer	training steps	error
Net1	4	15	1.13×10
Net2	5	17	1.76×10 ⁻⁵
Net3	6	5	5.89×10 ⁻⁶
Net4	7	19	3.08×10 ⁻⁵
Net5	8	26	2.76×10 ⁻⁵
Net6	9	15	5.79×10 ⁻⁶
Net7	10	32	5.21×10 ⁻⁶

隐含层与神经元的个数取决于多种因素，如输入层和输出层神经元数目、训练样本大小、样本噪音大小、所面对问题的复杂程度等^[8]，本文采用经验公式 $n_1 = \sqrt{n + m} + a$ (n_1 为隐含层个数， a 为 1~10 之间的常数， m 为输出神经元个数， n 为输入神经元个数)，由表 1 可知， n_1 为 6 时最合适。

本文采用的 BP 神经网络如图 4 所示。

2.3 模型评估

为了评估模型的预测性能，在实验过程中，对测试集的预测结果采用准确率(Accuracy)作为模型评估的指标，准确率计算公式为：

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

式中： TP 代表在测试集样本中真实标签为 1(恶性)，模型预测结果也为 1 的样本个数； FP 代表在测试集样本中真实标签为 -1(良性)，模型预测结果却为 1(恶性)的样本个数； TN 代表在测试集样本中真实标签为 -1(良性)，模型预测结果也为 -1 的样本个数； FN 代表在测试集样本中真实标签为 1(恶性)，模型预测结果却为 -1(良性)的样本个数。

基于 BP_Adaboost 的乳腺癌预测模型的受试者工作特征曲线(Receiver Operating Characteristic curve，ROC)如图 5 所示。图 6 为 BP_Adaboost 方法下 10 个弱分类器的分类错误数分别为 23、20、29、20、21、21、21、20、21 时的分类效果图，其分类误差率为 0.314 5，而强分类器的分类误差率平均为 0.014 5。

图 7 为强预测器预测误差绝对值与 10 个弱预测器的平均误差绝对值对比图，进一步表明，将多个 BP 神经网络即多个弱分类器加权可以得到更精准的分类效果。

为进一步验证模型的性能，在多组随机试验的基础上采用十折交叉验证，得到基于 BP-Adaboost 的乳腺癌预测模型的准确率平均值为 98.59%，其基于 BP-Adaboost 预测模型的 AUC 为 0.996 6。说明所建模型区分度较好，可以反映模型的诊断价值，且非常稳定，验证效果非常好。

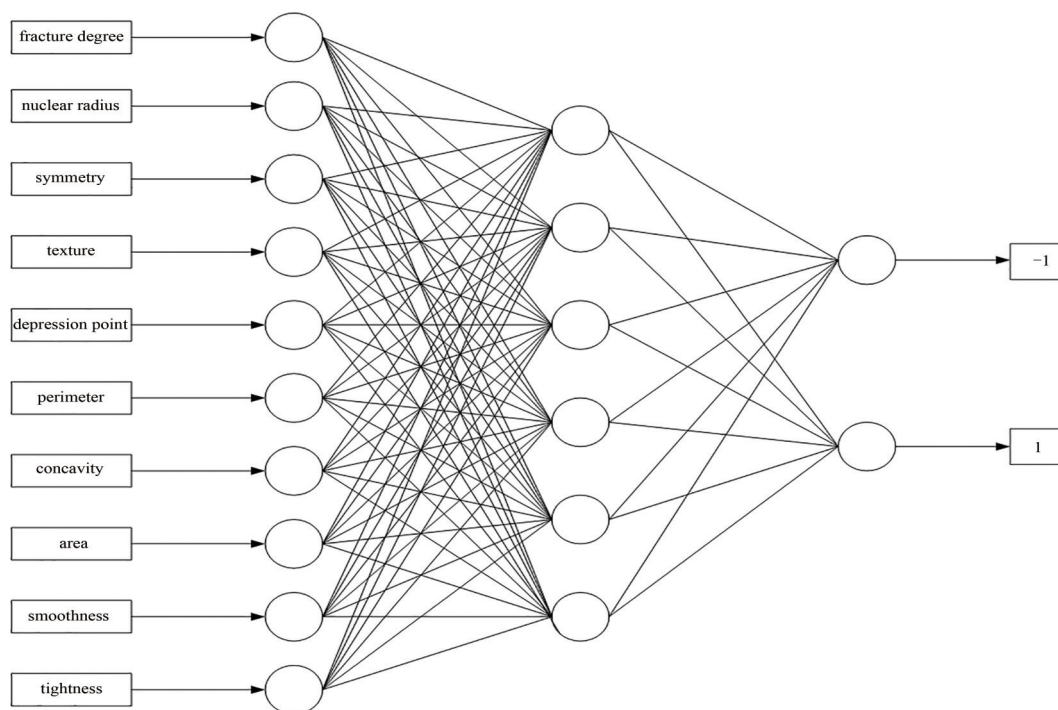


Fig.4 BP neural network
图4 BP神经网络图

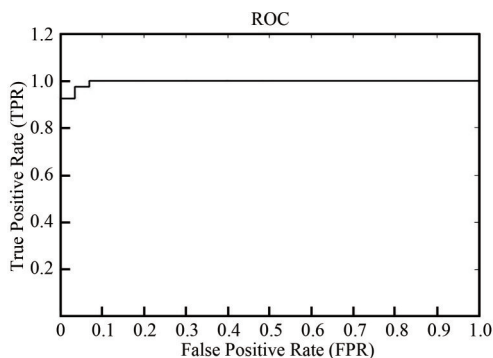


Fig.5 ROC curve based on BP-Adaboost breast cancer prediction model

图5 基于BP-Adaboost乳腺癌预测模型的ROC曲线

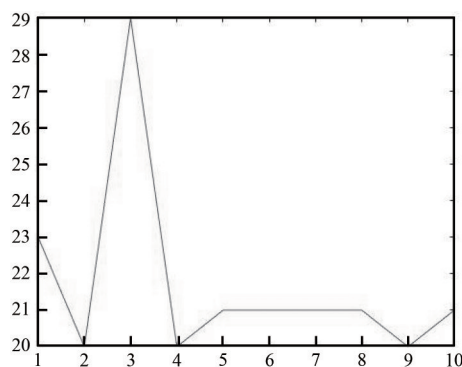


Fig.6 Classification effect of weak classifier

图6 弱分类器分类效果

3 实验结果与分析

3.1 模型仿真验证

在 Matlab 环境下搭建仿真模型, 训练集样本 500 例, 测试集样本 69 例。由表 2 仿真实验结果可知, 其中良性肿瘤确诊 47 例、误诊 1 例, 良性肿瘤检测准确率为 97.916 7%; 恶性肿瘤确诊 21 例、误诊 0 例, 恶性肿瘤检测率为 100%; 总检测样本 69 例, 确诊 68 例、误诊 1 例, 检测准确率为 98.550 7%。仿真实验预测结果如图 8 所示。

同时, 在相同的 Matlab 环境下采用其他方法进行仿真实验。将 BP_Adaboost 方法分别与 BP 神经网络模型、学习向量量化(Learning Vector Quantization, LVQ)模型^[9]、决策树模型^[10]、详尽可能性模型(Elaboration Likelihood Model, ELM)^[11]、随机森林^[12]等方法进行仿真实验对比,

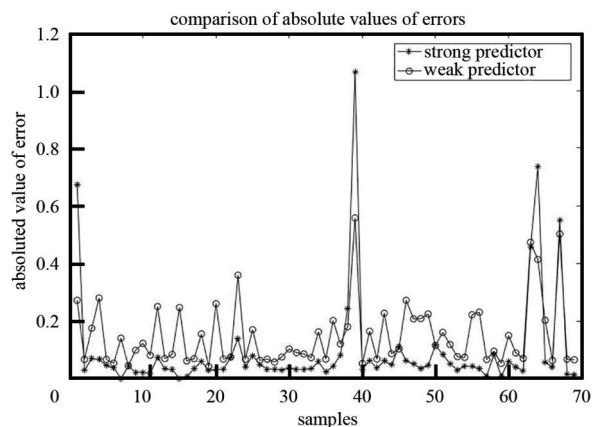


Fig.7 Prediction error comparison

图7 预测误差对比图

每组实验进行 100 次，将所得模型的准确率求均值，仿真结果如表 2 所示。多组随机实验下，BP 神经网络模型的准确率为 88.405 8%，BP_Adaboost 模型的准确率为 98.550 7%，LVQ 模型的准确率为 94.202 9%，决策树模型的准确率为 89.855 1%，ELM 模型的准确率为 86.956 5%，随机森林模型的准确率为 94.202 9%。对比图如图 9 所示，本文提出的 BP_Adaboost 方法在模型预测准确率上较其他方法具有显著优势。

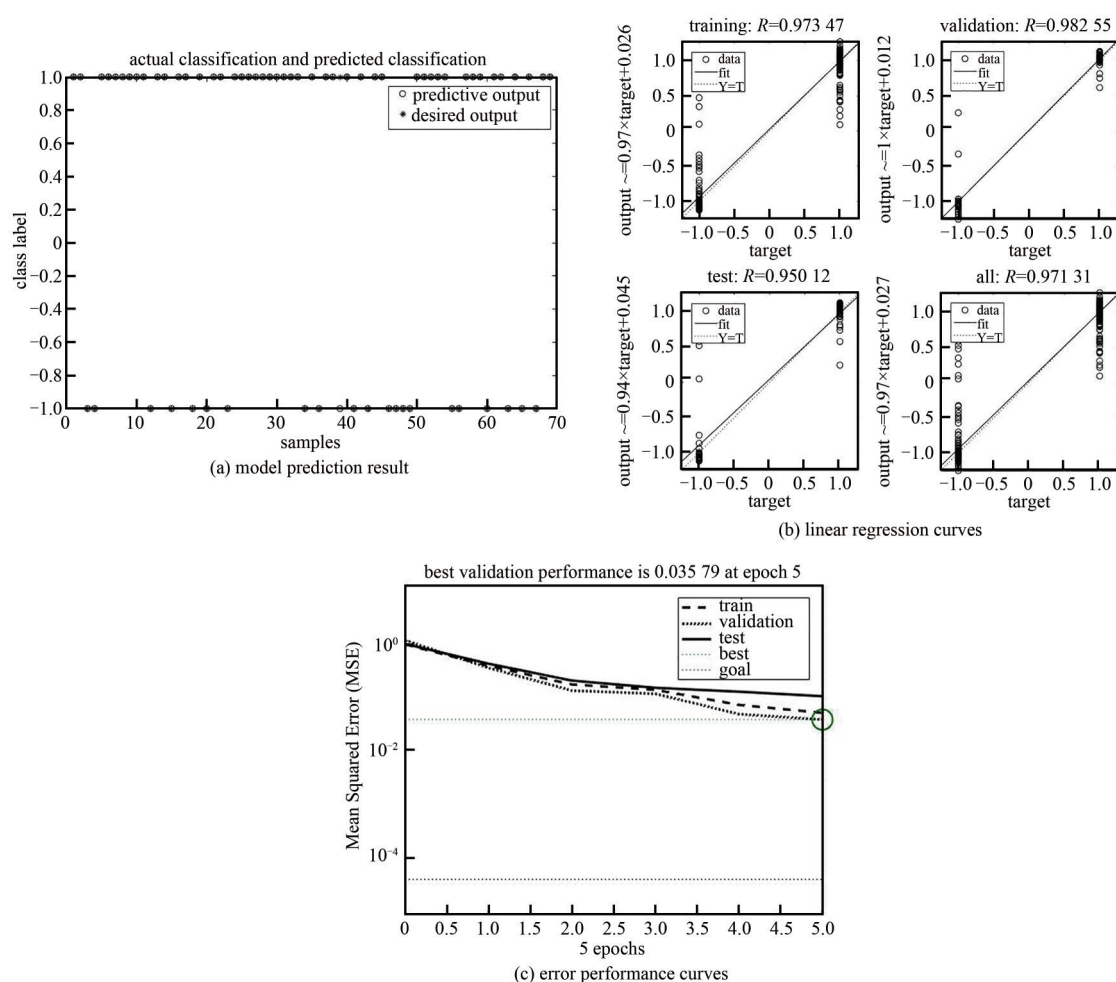


Fig.8 Classification results of BP_Adaboost model

图 8 BP_Adaboost模型分类结果

表 2 不同方法下模型准确率对比

Table2 Comparison of model accuracy by different methods

		confirmed	misdiagnosis	accuracy/%
BP	benign	32	4	88.888 9
	malignant	29	4	87.878 8
BP_Adaboost	benign	47	1	97.916 7
	malignant	21	0	100.000 0
LVQ	benign	38	1	97.435 9
	malignant	27	3	90.000 0
DecisionTree	benign	40	3	93.023 3
	malignant	22	4	84.615 4
ELM	benign	39	1	97.500 0
	malignant	21	8	72.413 8
random forest	benign	42	1	97.674 4
	malignant	23	3	88.461 5

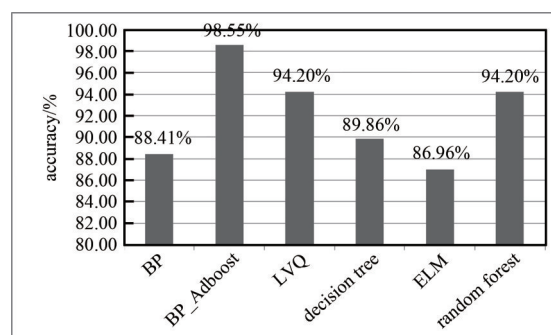


Fig.9 Comparison of model accuracy

图 9 模型准确率对比

3.2 不同研究结果对比

对比其他模型建立方法，如表 3 所示。Huiling Chen 等基于粗糙集的特征选取方法，采用分层采样的验证方

法构建模型^[13]。Yen Chen 等采用人工神经网络(Artificial Neural Network, ANN)对非小细胞性肺癌术后患者的生存情况进行预测,但此模型的泛化能力较差,只能对非小细胞型肺癌的患者数据进行预测^[14]。Tomczak 等运用分类受限玻尔兹曼机对乳腺癌术后患者 10 年内的复发情况进行预测,进一步实验结果发现肿瘤分期超过“50 mm”,对乳腺癌复发的影响力就较弱^[15-16]。对比结果表明,本模型分类性能稳定,误差小,且具有更好的泛化能力,同时预测模型的准确性显著提高,有助于医疗工作者更加快速、方便地诊断,为进一步的临床实现提供强有力的理论依据。

表3 不同研究结果对比

Table3 Comparison of different studies

author	tumor type	cases	data type	AUC
Chen.et al ^[13]	breast cancer	683	clinical, nuclear characteristic	0.994
Chen.et al ^[14]	lung cancer(cross dataset)	440	clinical, genetic	0.835
Tomczak	breast cancer	949	clinical	0.738
this study	breast cancer	569	nuclear characteristic	0.997

4 结论

针对现有乳腺癌预测模型中存在的过拟合现象以及漏诊率和误诊率过高的问题,提出了基于BP_Adaboost的乳腺癌预测模型。该模型兼具BP神经网络和Adaboost方法的优点,实现了分类精确度高且收敛效果好的特点。通过随机实验验证表明该模型分类性能稳定,误差小且具有更好的泛化能力,同时显著提高预测模型的准确性,有助于医疗工作者更加快速、方便地诊断。乳腺癌预测模型准确率的提高对于后续将临床与机器学习相结合有重要的意义。目前临床数据的积累已十分充足,将这十分庞大的数据加以利用与研究,创造出更简便易操作且准确率高的机器,可以减轻医疗人员的负担。综合性医学信息支撑平台的搭建,利用网络信息平台以及数据库信息突破某些局限,用户与资源平台共建共享,可进一步推动医学事业的发展。对医疗数据进行分析运用并从中取得规律也将具有十分重大的意义,在后续的研究中,将继续深入结合临床医学知识,对已建立的乳腺癌预测模型进行不断更新,找到更加实用的设计方法。

参考文献:

- [1] 宋争放. 建立我国癌症筛查制度的战略思考[J]. 肿瘤预防与治疗, 2019,32(3):199-206. (SONG Zhengfang. Strategic thinking on the establishment of cancer screening system in China[J]. Journal of Cancer Control and Treatment, 2019,32(3):199-206.)
- [2] BISWAS S K, MUKHERJEE D P. Recognizing architectural distortion in mammogram: a multiscale texture modeling approach with GMM[J]. IEEE Transactions on Biomedical Engineering, 2011,58(7):2023-2030.
- [3] HAMIDINEKOO A, DENTON E, RAMPUN A, et al. Deep learning in mammography and breast histology, an overview and future trends[J]. Medical Image Analysis, 2018(47):45-67.
- [4] DHUNGEL N, CARNEIRO G, BRADLEY A P. Fully automated classification of mammograms using deep residual neural networks[C]// 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). Melbourne, VIC, Australia: IEEE, 2017:310-314.
- [5] 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. (WU Chenwen, LI Changsheng, WANG Wei, et al. Application of an improved support vector machine algorithm in the diagnosis of breast cancer[J]. Computer Engineering & Science, 2017,39(3):562-566.)
- [6] 赖胜圣, 刘虔铖, 余丽玲, 等. 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(7):826-829. (LAI Shengsheng, LIU Qiancheng, YU Liling, et al. Construction of breast cancer prediction model based on SFS-SVM[J]. Chinese Journal of Medical Physics, 2019,36(7):826-829.)
- [7] 曹莹, 苗启广, 刘家辰, 等. AdaBoost算法研究进展与展望[J]. 自动化学报, 2013,39(6):745-758. (CAO Ying, MIAO Qiguang, LIU Jiachen, et al. Advance and prospects of AdaBoost algorithm[J]. Acta Automatica Sinica, 2013,39(6):745-758.)
- [8] WOLBERG W H, MANGASARIAN O L. Multisurface method of pattern separation for medical diagnosis applied to breast cytology[J]. Proceedings of the National Academy of Sciences of the United States of America, 1990,87(23):9193-9196.
- [9] SI J R, YAN Z, HU S J, et al. Comparison of LVQ and BP neural network in the diagnosis of diabetes and retinopathy[C]// The 4th International Conference of Pioneering Computer Scientists, Engineers and Educators(ICPCSEE 2018). Zhengzhou, China: Springer, 2018:455-466.
- [10] 王宏, 张强, 王颖, 等. 基于ELM的改进CART决策树回归算法[J]. 计算机系统应用, 2021,30(2):201-206. (WANG Hong, ZHANG Qiang, WANG Ying, et al. Improved CART decision tree regression algorithm based on ELM[J]. Computer Systems &

- Applications, 2021,30(2):201–206.)
- [11] 王成凯,杨晓敏,严斌宇. 基于随机森林的红外图像超分辨力算法[J]. 太赫兹科学与电子信息学报, 2020,18(4):665–671, 686. (WANG Chengkai, YANG Xiaomin, YAN Binyu. Infrared image super-resolution algorithm based on random forest[J]. Journal of Terahertz Science and Electronic Information Technology, 2020,18(4):665–671,686.)
- [12] 李琳,杨日东,王哲,等. 基于机器学习方法的原发性肝癌患者预后预测研究[J]. 中国数字医学, 2019,14(3):34–37. (LI Lin, YANG Ridong, WANG Zhe, et al. Prognosis of primary liver cancer patients based on machine learning method[J]. China Digital Medicine, 2019,14(3):34–37.)
- [13] CHEN H L, YANG B, LIU J, et al. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis[J]. Expert Systems with Applications, 2011,38(7):9014–9022.
- [14] CHEN Y C, KE W C, CHIU H W. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories[J]. Computers in Biology and Medicine, 2014(48):1–7.
- [15] TOMCZAK J M. Prediction of breast cancer recurrence using classification restricted Boltzmann machine with dropping[EB/OL]. (2013–10–30)[2015–08–28]. <https://arxiv.org/abs/1308.6324>.
- [16] 施维,薛均,潘隼然,等. 机器学习在肿瘤早期诊断与预后预测中的应用[J]. 医学信息学杂志, 2016,37(11):10–14,22. (SHI Wei, XUE Jun, PAN Cuiran, et al. Applications of the machine learning in early diagnosis and prognostic prediction of tumors[J]. Journal of Medical Informatics, 2016,37(11):10–14,22.)

作者简介：

葛梦飞(1995–)，女，在读硕士研究生，主要研究方向为生物信息数据挖掘。

刘嘉欣(1999–)，女，在读硕士研究生，主要研究方向为生物信息数据挖掘。

李赵旭(1992–)，男，硕士，主要研究方向为图像处理。

王宏伟(1969–)，男，博士，教授，主要研究方向为模式识别。

王 佳(1984–)，女，博士，讲师，主要研究方向为生物信息数据挖掘。