Operations Research and Fuzziology 运筹与模糊学, 2024, 14(4), 397-405
Published Online August 2024 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2024.144409

Hans汉斯

基于机器学习算法的乳腺癌诊断分析

张力芝

重庆对外经贸学院数学与计算机学院，重庆

收稿日期：2024年6月28日；录用日期：2024年8月9日；发布日期：2024年8月16日

摘　要

机器学习作为人工智能的重要支撑技术，在多个领域都得到了极为广泛的应用，如图像识别、自然语言处理、医学诊断等。本文针对威斯康星医院乳腺癌数据，采用机器学习中的k近邻、朴素贝叶斯、决策树以及神经网络这四类分类器算法进行分析诊断，即通过收集和预处理乳腺癌数据集，采用机器学习分类方法得到对应的混淆矩阵，计算各分类器的性能评价指标，并用10折交叉验证结果的可靠性，分析最适合的机器学习诊断方法。实验结果表明：k近邻法的评价指标值最高，误判率最低，相比之下更适合该乳腺癌的诊断。

关键词

乳腺癌，k近邻，朴素贝叶斯，决策树，神经网络

Analysis of Breast Cancer Diagnosis Based on Machine Learning Classifier Algorithm

Lizhi Zhang

School of Mathematics and Computer Science, Chongqing College of International Business and Economics, Chongqing

Received: Jun. 28th, 2024; accepted: Aug. 9th, 2024; published: Aug. 16th, 2024

Abstract

Machine learning, as a crucial underpinning technology for artificial intelligence, has found extensive applications in various domains such as image recognition, natural language processing, and medical diagnosis. This study focuses on the analysis and diagnosis of breast cancer using four classification algorithms from machine learning: k-nearest neighbors, naive Bayes, decision trees,

文章引用: 张力芝. 基于机器学习算法的乳腺癌诊断分析[J]. 运筹与模糊学, 2024, 14(4): 397-405.
DOI: 10.12677/orf.2024.144409

and neural networks. Specifically, this paper utilizes these algorithms for data collection and preprocessing of the Wisconsin Hospital Breast Cancer Data Set. Subsequently, employing machine learning classification methods to generate corresponding confusion matrices and calculate performance evaluation metrics for each classifier; furthermore, assessing the reliability of the results through 10-fold cross-validation to identify the most suitable machine learning diagnostic approach. The experimental findings indicate that k-nearest neighbors exhibit the highest evaluation index values and lowest misjudgment rates compared to other classifiers, making it more suitable for breast cancer diagnosis.

Keywords

Breast Cancer, k-Nearest Neighbor, Naive Bayesian, Decision Tree, Neural Network

1. 引言

乳腺癌是全球女性所患恶性肿瘤之一，其发病率和死亡率逐年上升，成为影响女性健康的重要疾病，一直是备受人们关注的重点，现如今早已成为了社会的重大公共卫生问题。乳腺癌的早期诊断和治疗对于提高患者的生存率至关重要。近年来，乳腺癌诊断技术取得显著进展，从传统的影像学检查到分子水平的生物标志物检测，再到人工智能辅助诊断系统的应用，这些技术的发展极大地提高了乳腺癌早期检出率和诊断准确性。

机器学习作为人工智能的核心技术，是一种强大的数据分析工具，能够从大量复杂的数据中进行学习和识别，在疾病诊断领域有着巨大的应用价值，如心血管疾病[1]、肺癌[2]、帕金森[3]、胰腺炎[4]等疾病。文献[5]将机器学习中的决策树、随机森林、支持向量机和 k 近邻算法运用到糖尿病的诊断中，通过建立各自模型来进行优劣性的比较，进而选择最优模型对糖尿病数据进行分析预测。文献[6]运用机器学习中的支持向量机、人工神经网络和深度学习方法对肿瘤诊断与预后预测进行分析，并通过准确率和曲线下面积来评价模型的分类效果。文献[7]通过对心脏病数据集的分析，构建机器学习中决策树和 k 近邻两种算法模型进行分类与预测，并以准确度、精确度、召回率和 F1_score 作为模型的性能评价指标，将两种模型的预测性能进行比较。文献[8]针对乳腺癌数据集，采用决策树中的 C5、CR、QUEST 及 CHAID 这四种算法，通过对乳腺癌腋窝高位淋巴结相关数据的分析构建分类模型，并与 Logistic 模型进行性能比较，结果表明 CHAID 模型的预测性能明显优于 Logistic 模型。文献[9]针对乳腺肿瘤的图像数据提出一种基于 BBO 优化 BP 神经网络算法，对乳腺肿瘤进行分类诊断，结果表明该算法对乳腺癌数据有很好的分类性能，误诊率较低。除此之外，还有很多学者将机器学习中的支持向量机[10] [11]、随机森林[12]等分类器运用到乳腺癌疾病诊断中。

上述研究均是通过机器学习的某一个分类器对乳腺癌进行诊断，亦或是通过多个分类器对其它疾病进行诊断，未利用多个分类器分析乳腺癌的肿块特征进行分类诊断。本文根据收集的乳腺癌数据集进行机器学习各类分类器的建模，即通过 k 近邻、朴素贝叶斯、决策树以及神经网络算法得到它们各自的诊断结果和混淆矩阵，选择准确度、Kappa 统计量、灵敏度、特异性、精确度、回溯精确度作为它们的性能评价指标进行比较，并利用 10 折交叉验证分析其结果的可靠性，从而得出最适合该乳腺癌诊断的分类方法。

2. 分类器算法的理论介绍

2.1. 模型选取

机器学习能使计算机在没有明确编程的情况下自主学习。不同于传统方法的分类模型需要对数据有一定的要求或假设，机器学习则是数据驱动，不需任何假设，可通过自主学习来直接对数据进行预测，其结果也可用交叉验证的方法进行验证。机器学习包括监督、半监督、非监督和强化学习，其中监督学习与非监督学习的主要区别在于训练数据有无标签，而由于本文针对的乳腺癌数据具有对应的标签，因此本文采取监督学习的方法进行分类预测。在监督学习中又包含多种分类算法，其中较为常用的算法有k近邻、朴素贝叶斯、决策树、神经网络、支持向量机、随机森林以及逻辑回归[13]等，本文简单总结了这些分类算法的优缺点，列于表1中。

Table 1. Advantages and disadvantages of each classification algorithm
表 1. 各分类算法的优缺点

分类算法	优点	缺点
k近邻	精度高，无假定，无参数模型 对异常值不敏感，可处理高维数据集	复杂度高，k值选取困难
朴素贝叶斯	简单高效，容错性强， 可解释性强，可处理高维数据	性能取决于特征的独立性假设
决策树	效率高，易解释，非参方法， 可处理混合类型特征，可处理高维数据	易过拟合，对缺失值敏感
神经网络	自适应性强，适合各种类型的数据， 可处理高维数据	易过拟合，缺乏理论保证
支持向量机	可处理高维数据，适用于非线性问题	复杂度高，调参困难， 不适用重叠数据
随机森林	准确率较高，可处理高维数据， 适用于非平衡数据	对无关特征不敏感，参数调优， 结果可能存在偏差，不易解释
逻辑回归	简单易懂，计算效率高， 适合线性数据	不适合处理数据特征复杂或非线性数据

由于乳腺癌数据集属于高维数据，且特征较多，可能同时存在线性和非线性数据等特点，通过对比分析表1展示的几种分类算法的优缺点，本文选取无参数模型的k近邻、可解释性和容错性强的朴素贝叶斯、可处理混合型特征和易解释的决策树以及自适应性强且能够处理各种类型数据的神经网络这四种分类算法针对乳腺癌数据集进行分类预测。

2.2. 模型介绍

2.2.1. k近邻(k-NN)

k近邻法(k-Nearest Neighbor, k-NN)是机器学习中最基础的算法之一，是一种既可分类又可回归的算法。它基于实例的学习，不需要显式的训练过程，而是根据已有数据进行训练学习，达到分类或预测的效果。k-NN算法的核心思想是：对于一个新的输入样本，根据其特征空间中k个邻近样本的属性来判断该样本的类别。k近邻法的主要步骤如下：

1) 距离度量：确定一个度量标准来计算样本与样本之间的距离，本文使用欧氏距离来进行度量，即

$$dist(p,q)=\left(\sum_{i=1}^{n}|p_i-q_i|^2\right)^{\frac{1}{2}} \tag{1}$$

其中 $dist(p,q)$ 表示 p 与 q 两点之间的距离，p_i 表示点 p 第 i 个空间向量的值，q_i 表示点 q 第 i 个空间向量的值。

2) 选择 k 值：k 值代表着在进行预测时要考虑的最近邻的数量。K-NN 算法的性能优劣依赖于所选的 k 值，若 k 值较小，则会出现过度拟合的情况，若 k 值较大，则会引入噪声和异常值，出现模型过于平滑的情况，所以选择合适的 k 值是关键。

3) 投票表决：最近邻样本多数样本的属性来判断输入样本的类别。

2.2.2. 朴素贝叶斯

朴素贝叶斯法是一种简单的概率分类方法，它通过计算待分类项属于各类的条件概率，选取最大概率作为分类依据。朴素贝叶斯是基于贝叶斯定理来实现的分类，即

$$P_{post} = P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)} = \frac{P(Y)\prod_i P(x_i|Y)}{P(X)} \tag{2}$$

其中 X 代表样本数据的特征属性，Y 代表类别变量，$P(Y)$表示类别变量的先验概率，P_{post} 表示类别变量的后验概率。

2.2.3. 决策树

决策树是一种常用于分类和回归任务的机器学习方法，它模仿了人类的决策过程，通过一系列规则对数据进行分类。它是一种基于树结构进行决策的模型，包括根节点、内部节点、叶节点和边，其中每个内部节点代表着条件，叶节点代表着类别标签，边则代表着满足条件的结果。

决策树有很多算法可以实现，但本文选取的 C5.0 算法是多种决策树算法之一，因其高效和灵活性，且在多种场景下都是一个非常有力的工具，而成为了现如今生成决策树的行业标准。决策树需要确定根据某一个特征来进行分割，目前有很多不同的度量纯度的方法可以用来确定分割的标准，如熵、基尼系数和误差率，其中 C5.0 算法使用的是熵来度量纯度，即

$$Entropy(S) = -\sum_i p_i \log_2(p_i) \tag{3}$$

其中，S 表示给定的数据集，$Entropy(S)$表示数据集 S 的熵值，p_i 表示类别 i 在数据集 S 中的比例。熵反映了数据集的不确定性，熵越大，不确定性也就越大，数据集也就越随机，纯度越低。

2.2.4. 神经网络

k 神经网络全称为人工神经网络，是一种模拟人脑工作原理的计算模型，通过大量数据训练来识别复杂的模式和特征。这种网络由大量单元或神经元组成，每个神经元接收输入，对其进行处理，并产生输出。神经网络由多个层次构成，包括输入层、隐藏层和输出层，各层都是由多个神经元所组成，每个神经元都有一系列的权重和偏差，如图 1 所示。

图中 $x_1, x_2, \cdots, x_n$ 为输入信号，w_{ij} 表示从神经元 i 和 j 连接的权重值，θ 表示所设置的阈值，神经元 i 的输出与输入的关系表达式为：

$$net_i = \sum_{j=1}^{n} w_{ij} x_j - \theta \tag{4}$$

$$y_i = f(net_i) \tag{5}$$

其中，y_i 表示神经元 i 的输出，函数 f 称为激活函数，若 net_i 为正，则称该神经元处于激活状态，否则称为抑制状态。

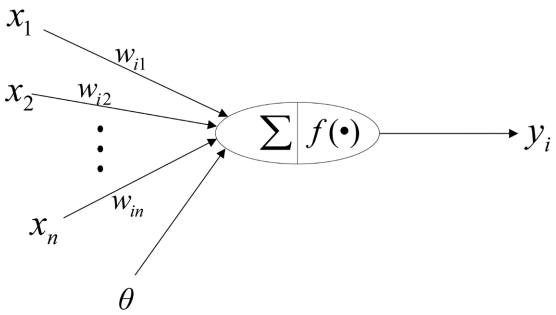

Figure 1. Schematic diagram of an artificial neuron model
图 1. 人工神经元模型示意图

3. 各分类器的乳腺癌诊断分析

3.1. 数据收集与预处理

本文选取来自 UC Irvine Machine Learning Repository 网站上威斯康星大学医院的乳腺癌数据集，该乳腺癌数据集总共涉及到 699 个乳腺癌患者的相关检测样本，每个样本包含 10 项特征数据以及一个表明该患者所患乳腺癌为良性或恶性的类标签，具体特征名称和类标签如表 2 所示。表中除了患者的 ID 编号不用于机器学习建模以外，其余特征为有序变量，均可用于机器学习建模。

Table 2. Feature name and feature explanation
表 2. 特征名称与特征解释

编号	名称	解释
1	Sample code number	ID 编号
2	Clump Thickness	肿块厚度(1~10)
3	Uniformity of Cell Size	细胞大小的均匀度(1~10)
4	Uniformity of Cell Shape	细胞形状的均匀性(1~10)
5	Marginal Adhesion	边际粘合力(1~10)
6	Single Epithelial Cell Size	单个上皮细胞大小(1~10)
7	Bare Nuclei	裸核(1~10)
8	Bland Chromatin	布兰德染色质(1~10)
9	Normal Nucleoli	正常核仁(1~10)
10	Mitoses	有丝分裂(1~10)
11	Class	类别(2 (良性)，4 (恶性))

因本文所采用的乳腺癌数据集存在缺失值情况，需要对该数据集进行数据预处理，即将样本中存在缺失值的 16 名患者信息进行删除，对剩余的 683 个乳腺癌患者样本进行建模分析。通过统计类型变量 Class，分析乳腺癌的良性肿块和恶性肿块的比例，得知 683 个样本内有 444 个良性肿块和 239 个恶性肿块，所占比例分别为 65%和 35%。本文将乳腺癌数据集的前 583 个乳腺癌样本数据作为训练集用于学习建模，后 100 个乳腺癌样本数据作为测试集用于预测结果的验证。

3.2. 分类器结果展示

本文采用了机器学习中的 k 近邻、朴素贝叶斯、决策树以及神经网络这四类分类器算法，并且根据

各类分类器的混淆矩阵，分别计算相应的性能评价指标，以此来判别最适合该乳腺癌数据集的分类方法。由于本文所感兴趣的类别是阳性和阴性，即乳腺癌的恶性肿块和良性肿块，则各类分类器的混淆矩阵中包含有真阳性、假阳性、假阴性和真阴性，落在其中的次数分别表示为 *TP*、*FP*、*FN* 和 *TN*。*TP* 表示将恶性肿块诊断为恶性的数量，*FP* 表示将良性肿块诊断为恶性的数量，*FN* 表示将恶性肿块诊断为良性的数量，*TN* 表示将良性肿块诊断为良性的数量，表 3 展示了各分类器混淆矩阵结果。

Table 3. Confusion matrix table for each classifier
表 3. 各分类器的混淆矩阵表

混淆矩阵 / 分类器	*TN*	*FP*	*FN*	*TP*
k 近邻	79	0	0	21
朴素贝叶斯	77	2	0	21
决策树	59	4	1	36
神经网络	78	1	0	21

3.3. 性能评价指标

为了检验模型的有效性，度量模型的性能优劣，本文将准确度、*Kappa* 统计量、灵敏度、特异性、精确度和回溯精确度作为性能评价指标。

1) 准确度：正确诊断肿块类别的数量除以诊断总数，即

$$\text{准确度} = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

2) *Kappa* 统计量：通过解释完全因为巧合而诊断正确的概率，*Kappa* 统计量对准确度进行了调整，*Kappa* 值越大，则一致性越好，即

$$Kappa = \frac{\Pr(a)-\Pr(e)}{1-\Pr(e)} \tag{7}$$

其中 Pr 指的是分类器和真实值之间的真实一致性(a)和期望一致性(e)的比例，且 Pr(e)是完全偶然性导致的肿块诊断结果和真实结果相同的概率。

3) 灵敏度和特异性：灵敏度指的是度量恶性肿块样本被正确诊断的比例；特异性指的是度量良性肿块被正确诊断的比例，即

$$\text{灵敏度} = \frac{TP}{TP+FN} \tag{8}$$

$$\text{特异性} = \frac{TN}{TN+FP} \tag{9}$$

4) 精确度和回溯精确度：精确度指的是恶性肿块诊断正确的数量占所有预测为恶性肿块的数量之比；回溯精确度指的是恶性肿块诊断正确的数量占所有恶性肿块的诊断数量之比，即

$$\text{精确度} = \frac{TP}{TP+FP} \tag{10}$$

$$\text{回溯精确度} = \frac{TP}{TP+FN} \tag{11}$$

表 4 列出了各类分类器的性能评价指标，从表中可看出，在这四类分类器中 k 近邻的各个评价指标值最高，性能评价指标均为 1，且均值最大，其次是神经网络，性能评价指标的均值为 0.9782，而后是朴素贝叶斯，其性能评价指标的均值为 0.958，最后是决策树，其性能评价指标的均值为 0.9335。由此可见，针对威斯康星医院的乳腺癌数据集而言，这四类分类器诊断最为准确的是 k 近邻法。

Table 4. Performance evaluation metrics for each classifier
表 4. 各分类器的性能评价指标

分类器 / 评价指标	k 近邻	朴素贝叶斯	决策树	神经网络
准确度	1	0.9800	0.9500	0.9899
Kappa 统计量	1	0.9418	0.8945	0.9703
灵敏度	1	0.9130	0.9000	0.9545
特异性	1	1	0.9833	1
精确度	1	1	0.9730	1
回溯精确度	1	0.9130	0.9000	0.9545
指标平均值	1	0.9580	0.9335	0.9782

3.4. 10 折交叉验证

本文针对上述几种分类方法都用 10 折交叉验证的方法来判断其结果的可靠性。通过采用 10 折交叉验证的方法来计算 k 近邻分类、朴素贝叶斯分类、决策树分类以及神经网络分类方法的误判率，将各种分类器的 10 折交叉验证结果展示在图 2 中(kknn 表示 k 近邻分类，bayes 表示朴素贝叶斯分类，tree 表示决策树分类，nnet 表示神经网络分类)，图中对应于每一种分类方法的条高指的是测试集的平均误判率，而表 5 展示的是各分类器进行 10 折交叉验证的各折误判率以及平均误判率(均值)。

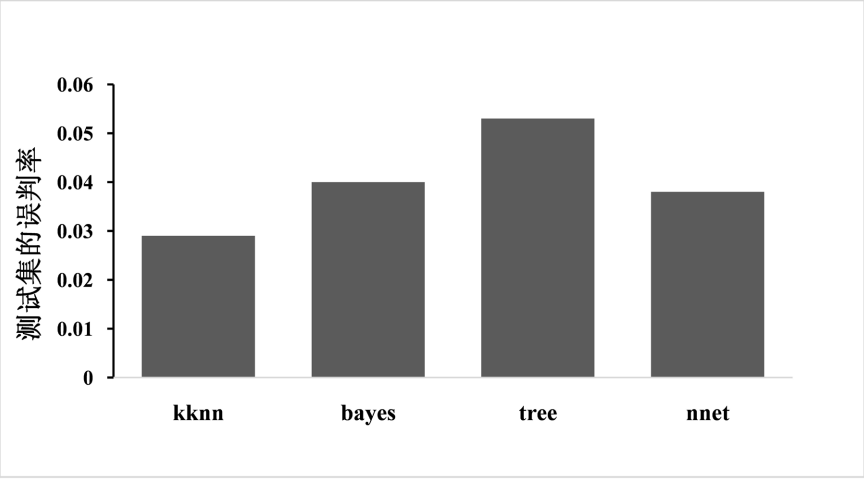

Figure 2. Average misclassification rate for 10-fold cross-validation
图 2. 10 折交叉验证的平均误判率

从图 2 或表 5 中可看出，测试集误判率最低的是 k 近邻分类器，最高的是决策树分类器，神经网络和朴素贝叶斯从低到高分别排第二和第三，这与上述各种分类器的性能评价指标的结果一致，说明相比之下，k 近邻分类器的诊断结果是最为准确的，也是最适合对该乳腺癌数据集进行诊断的方法。

Table 5. Misclassification rate for each classifier at each fold
表 5. 各分类器各折误判率标

分类器 / 折数	k 近邻	朴素贝叶斯	决策树	神经网络
1	0.043	0.058	0.101	0.072
2	0	0.029	0.043	0.014
3	0.043	0.043	0.072	0.014
4	0.029	0.014	0.014	0.043
5	0.029	0.044	0.044	0.015
6	0.044	0.029	0.059	0.044
7	0.059	0.044	0.074	0.074
8	0	0.044	0.044	0.029
9	0.015	0.029	0.029	0.029
10	0.030	0.060	0.045	0.045
均值	0.029	0.040	0.053	0.038

4. 结论

随着人工智能的普及，机器学习在乳腺癌诊断中的应用迅速发展，不仅提高了诊断的准确性和效率，还有望改善患者的治疗与预后。本文利用机器学习中的 k 近邻、朴素贝叶斯、决策树以及神经网络分类器，通过对乳腺癌数据集的分析进行分类诊断。首先，针对乳腺癌数据集预处理后，利用不同分类器计算得到相应的混淆矩阵；然后，根据混淆矩阵计算性能评价指标，即准确度、*Kappa* 统计量、灵敏度、特异性、精确度、回溯精确度，且用 10 折交叉验证结果的可靠性，得到测试集误判率；最后，通过比较各自的性能评价指标及误判率选择最优分类器。研究结果表明，四种分类器中 k 近邻法的性能最好，误判率最低，最适合该乳腺癌数据集的分析与诊断。

参考文献

[1] 张耀祖. 基于机器学习的心血管疾病预测研究[D]: [硕士学位论文]. 大连: 大连交通大学, 2023.

[2] 朱勇, 晏峻峰. 机器学习在肺癌诊断中的研究和应用[J]. 计算机与数字工程, 2024, 52(3): 751-756.

[3] 李西, 姜孟. 机器学习在帕金森病诊断中的应用研究[J]. 电子科技大学学报, 2024, 53(2): 315-320.

[4] 李龙, 尹梁宇, 种菲菲, 等. 基于改进的机器学习模型对重症急性胰腺炎诊断的早期预测[J]. 陆军军医大学学报, 2024, 46(7): 753-759.

[5] 吴兴惠, 周玉萍, 邢海花, 等. 机器学习分类算法在糖尿病诊断中的应用研究[J]. 电脑知识与技术, 2018, 14(35): 177-178+195.

[6] 施维, 薛均, 潘璀然, 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.

[7] 梁靖涵, 许亚杰. 基于机器学习算法的心脏病预测诊断模型研究[J]. 现代信息科技, 2022, 6(19): 67-70.

[8] 易静, 苏新良, 王润华. 决策树在乳腺癌高位淋巴结转移判别诊断中的应用[J]. 重庆医科大学学报, 2009, 34(5): 606-609.

[9] 李卉. 基于 BBO 优化 BP 神经网络的乳腺癌诊断[J]. 山西电子技术, 2018(5): 35-36+44.

[10] 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.

[11] 刘兴华, 蔡从中, 袁前飞, 等. 基于支持向量机的乳腺癌辅助诊断[J]. 重庆大学学报(自然科学版), 2007, 30(6): 140-144.

[12] 全雪峰. 基于随机森林的乳腺癌计算机辅助诊断[J]. 软件, 2017, 38(3): 57-59.

[13] 李洪成, 许金炜, 李舰. 机器学习与 R 语言[M]. 北京: 机械工业出版社, 2015.