深圳大学研究生课程论文

题目 利用机器学习从细胞病理学数据中预测乳腺癌 成绩_____

专业 生物医学工程 任课教师 薛武峰

课程名称、代码 2406039-BME5208B-机器学习（01）

学号 2410247047 姓名 余欣震

年级 研一 时间 2424 年 10 月

利用机器学习从细胞病理学数据中预测乳腺癌

摘要:乳腺癌是世界上女性最常见的癌症，传统的诊断方法依赖于医生的经验和主观判断，存在一定的误诊和漏诊风险。机器学习算法能够自动提取和识别病灶特征，提高诊断的准确性和效率。威斯康星诊断数据集包含诊断和一组 30 个特征，描述乳腺肿块细针穿刺(FNA)的数字化图像中细胞核的特征。将使用两种不同的算法创建模型，并使用模型预测诊断。

关键词:机器学习；乳腺癌；

Using machine learning to predict breast cancer from cytopathological data

Abstract: Breast cancer is the most common cancer in women in the world. Traditional diagnostic methods rely on the experience and subjective judgment of doctors, which have certain risks of misdiagnosis and missed diagnosis. Machine learning algorithms can automatically extract and identify the features of lesions to improve the accuracy and efficiency of diagnosis. The Wisconsin Diagnostic dataset contains diagnoses and a set of 30 features that describe the features of nuclei in digital images of fine needle aspiration (FNA) of breast masses. Two different algorithms will be used to create the model and the model will be used to predict the diagnosis.

Keywords: Machine learning; Breast cancer

1 绪论

1.1 研究背景与研究意义

乳腺癌是全球女性中最常见的癌症之一，发病率和死亡率均居高不下。随着医学影像技术的发展，乳腺癌的早期筛查和诊断手段不断丰富。然而，传统的诊断方法依赖于医生的经验和主观判断，存在一定的误诊和漏诊风险。机器学习技术的引入，为乳腺癌的诊断和治疗提供了新的思路和方法。通过对大量医学影像和临床数据的分析，机器学习算法能够自动提取和识别病灶特征，提高诊断的准确性和效率。

1.2 国内外研究现状

1.2.1 国内研究现状

国内研究主要集中在利用机器学习算法对乳腺 X 射线片、MRI 和超声图像进行自动分割和病灶特征提取。例如，研究表明，机器学习可以有效地自动分割病灶并分析其特征，从而提高诊断的准确性[1]。

多模态数据融合：一些研究尝试将乳腺 X 射线、MRI、超声及病理数据进行多模态融合，以提升诊断性能。这种方法可以最大限度地利用不同数据源的信息，提高诊断的全面性和准确性。

影像基因组学：国内也有研究将影像数据与基因组数据结合，利用机器学习技术进行乳腺癌分子亚型的鉴别和癌症风险评估。这种方法在个性化诊断和治疗中显示出巨大的潜力。

1.2.2 国外研究现状

国外在乳腺癌诊断中广泛应用深度学习算法，特别是在影像数据的处理上。深度学习模型如卷积神经网络（CNN）在乳腺癌图像分类和病灶检测中表现出色[2]。

大数据和 AI 技术：国外研究利用大规模公共数据库和先进的 AI 技术，开发出高效的乳腺癌诊断模型。这些模型不仅提高了诊断的准确性，还减轻了临床医生的工作负担。

个性化医疗：国外研究还注重个性化医疗，通过机器学习模型预测患者的治疗反应和预后，从而制定个性化的治疗方案。

1.3 研究方法

本文将采用 r 语言对乳腺癌数据集进行分类，使用 K 最近邻（KNN）和支持向量机（SVM）创建模型，计算其性能指标，并使用模型预测诊断。

2 模型构建与求解

2.1 数据预处理

读取 CSV 格式的乳腺癌数据集。删除 NULL 数据，并将数据保存为 data_file 和 patient 以便备用，删除 data_file 的 id 列，并将 diagnosis 列转换为因子类型，以便进行分类。使用 sample_frac 随机打乱 data_file 数据集，将数据集用 slice_sample 划分为 70%的训练集和 30%的测试集。导入 corrplot 包可视化多个变量之间的关系，并将不显著相关性系数图块上打上 X 符号。可以看出半径均值（radius_mean）和周长均值（perimeter_mean），凹陷点数均值（concave points_mean）和凹陷度均值（concavity_mean）这两对变量之间有很强的正相关性，光滑度均值（smoothness_mean）和凹陷点数均值（concave points_mean），对称性均值（symmetry_mean）和紧致度均值（compactness_mean）这两对变量之间有较强的负相关性。

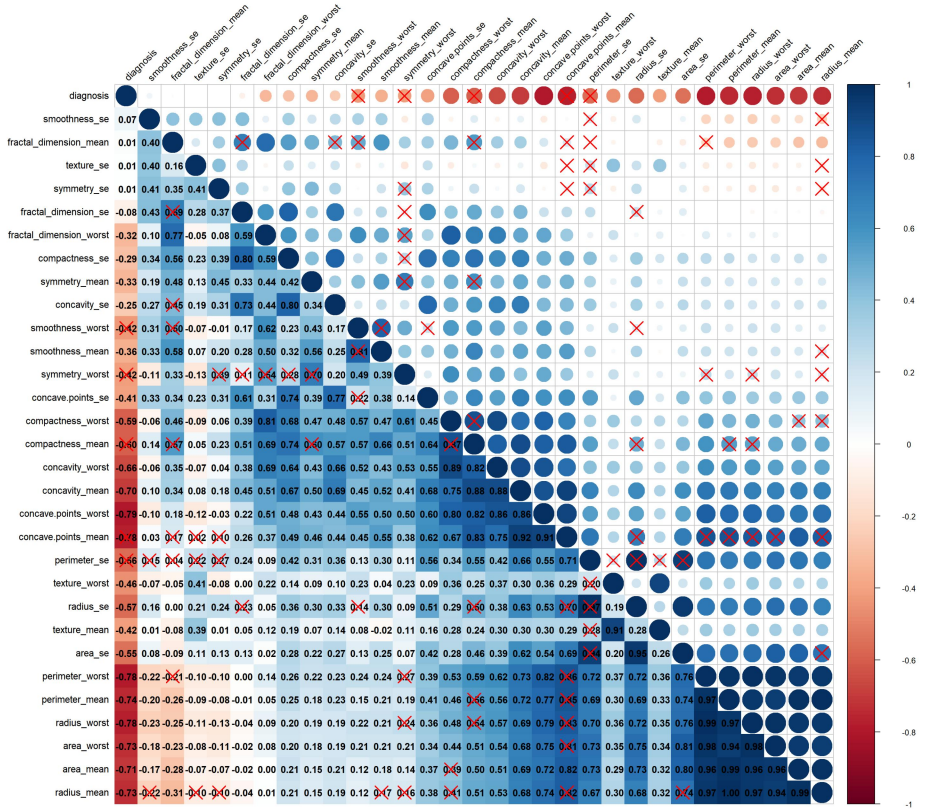

图 1 变量相关性

2.2 混淆矩阵

混淆矩阵（confusion matrix）是可视化工具，特别用于监督学习，在无监督学习一般叫做匹配矩阵。矩阵的每一列代表一个类的实例预测，而每一行表示一个实际的类的实例。TP：True Positive，也就是正确的正向预测，FP：称为 False Negative，也就是错误的负向预测，TN：True Negative，也就是正确的负向预测，FP：False Positive，也就是错误的正向预测，通常情况下 TP 和 TN 越多越好。把所有正确的情况，也就是 TP 和 TN，把它加总起来除上所有情形个数，那就是 Accuracy，这也是最常用的指标，但是在某些情形下这个指标会失效，Precision（准确率）和 Recall（召回率）这个时候就派上用场了，Precision 和 Recall 同时关注的都是 TP，但是角度不一样，Precision 看的是在预测正向的情形下，实际的精准度是多少，而 Recall 则是看在实际情形为正向的状况下，预测能召回多少实际正向的答案。如果 Precision 和 Recall 都同等重要，要用一个指标来统合标志它，这就是 F1 Score 或称 F1 Measure，这是真实阳性率(召回率)和正确率的加权平均值。Sensitivity 和 Specificity，Sensitivity 就是 Recall，它代表的是诊断方法是否够灵敏可以将真正得病的人诊断出来，其实就是真正有病症的患者有多少可以被侦测出来，而 Specificity 则代表实际没病症的人有多少被检验正确的。两种指标都是越高越好。Kappa 系数（Cohen’s Kappa）在实际应用中，Kappa 系数通常用于评估分类模型的性能，特别是在类别不平衡的情况下，它比简单的准确率（Accuracy）更能反映模型的真实表现，1 表示完全一致，0 表示与随机猜测一致，负值表示一致性低于随机猜测。

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

2.3 KNN

K 最近邻（K-Nearest-Neighbor，KNN）是一种非线性的分类算法，KNN 处理分类问题的方法是：找 K 个距离待遇测样本最近的点，然后根据这几个点的类别来确定新样本的类别。首先创建一个空向量，用于存储测试集的准确率，使用 for 循环 30 次，每次改变 k 的值进行超参数调优，将 k 和准确率存储在一个数据框中，找到准确率最高的 k 值，用最优的 k 值进行预测并得出 confusionMatrix，可以看出 knn 的 Accuracy : 0.936，, Kappa : 0.855 结果还是比较好的。但是我们希望拥有更高的 Accuracy。

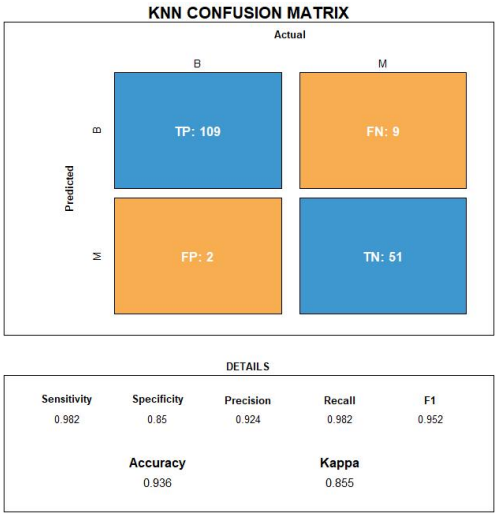

图 2 KNN CONFUSION MATRIX

2.4 SVM

支持向量机（support vector machine，SVM）是一个非常强大的算法，它既可以用于回归也可以用于分类，而且还可以用于生存分析中，在神经网络中也很热门。使用 svm（kernel = "radial"）进行训练，probability 设为 T，对 train_data 进行预测，并将结果存储在 svm_pred_df 中，并得出 confusionMatrix，可以看出 svm 的 Accuracy : 0.982，, Kappa : 0.962 结果比 knn 好了很多。

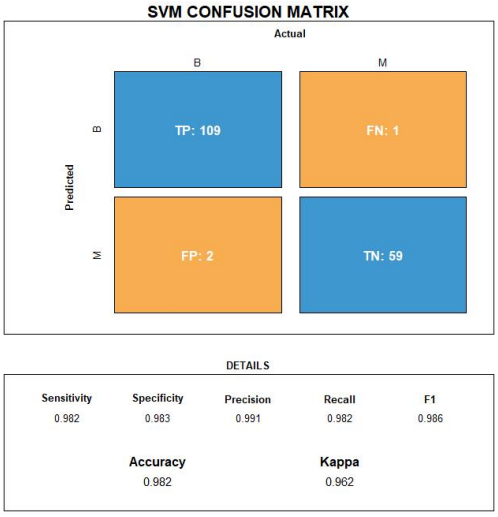

图 3 SVM CONFUSION MATRIX

3 结果分析

因为 SVM 的 predict_model 结果更好因此使用 SVM，核函数选择 radial（高斯径向基核），将之前的 patient 的 diagnosis 列取出保存为 orgin_diagnosis_res，再将 patient 的 diagnosis 列删除，使用 patient 作为测试数据集进行预测得出预测结果，预测正确的概率达到了 98.2%。

表 1 预测结果统计

CORRECT	N	PERCENTAGE
TRUE	559	0.982
FALSE	10	0.018

结论

随着机器学习的发展,利用机器学习算法挖掘乳腺癌数据和潜在信息,可以促进乳腺癌医学诊断与治疗的进步,提高乳腺癌临床诊断的准确率和效率,减少医生的工作量。在可预见的未来仍有望提高乳腺癌临床诊断、预后评估和治疗的准确率,实现乳腺癌的精准治疗。机器学习和大数据技术的应用在促进乳腺癌学进步的同时也带来了一些局限性,因此,在未来工作中还须对其进行深入研究。

参考文献

[1] 卢峰,张开炯,吴立春,等.机器学习对乳腺癌诊断应用现状研究进展[J].社区医学杂志,2023,21(24):1315-1322.

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