第 6 卷◆第 6 期◆版本 1.0◆2024 年
文章类型：论文丨刊号（ISSN）：2705-1102(P) / 2705-1110(O)

医疗大数据挖掘技术在疾病预测中的应用研究

张志超
天津市斯宇克医疗器械销售有限公司
DOI:10.12238/bmtr.v6i6.10981

[摘　要] 目的：本研究旨在探讨医疗大数据挖掘技术在疾病预测中的应用价值,并实证分析不同机器学习算法在疾病预测中的性能。方法：选取电子病历、实验室检验数据等医疗数据集,采用逻辑回归、支持向量机、随机森林和深度神经网络四种典型算法进行对比实验,评估各模型在准确率、召回率、F1值和AUC四个指标上的表现。结果：深度神经网络在所有评估指标上均表现最佳,AUC值达0.94,显示出较强的分类能力和泛化性；随机森林次之,其稳健性和预测精度较高。结论：实验结果表明深度学习算法在疾病预测中的应用潜力较大,提出的算法改进建议包括特征工程、参数调优和集成学习方法,有助于提升疾病预测的准确性与可靠性,为医疗大数据挖掘技术的优化提供了实证支持。
[关键词] 医疗大数据；疾病预测；机器学习；深度神经网络
中图分类号：R363.2+1　文献标识码：A

Research on the Application of Medical Big Data Mining Technology in Disease Prediction

Zhichao Zhang
Tianjin Siyuke Medical Equipment Sales Co., Ltd.

[Abstract] Objective: This study aims to explore the application value of medical big data mining technology in disease prediction, and empirically analyze the performance of different machine learning algorithms in disease prediction. Method: Selecting medical datasets such as electronic medical records and laboratory test data, four typical algorithms including logistic regression, support vector machine, random forest, and deep neural network were used for comparative experiments to evaluate the performance of each model in four indicators: accuracy, recall, F1 score, and AUC. Result: Deep neural networks performed the best in all evaluation metrics, with an AUC value of 0.94, demonstrating strong classification ability and generalization; Random forest comes second, with higher robustness and prediction accuracy. Conclusion: The experimental results indicate that deep learning algorithms have great potential in disease prediction. The proposed algorithm improvement suggestions include feature engineering, parameter tuning, and ensemble learning methods, which can help improve the accuracy and reliability of disease prediction and provide empirical support for the optimization of medical big data mining technology.
[Key words] medical big data; Disease prediction; Machine learning; Deep neural network

借助大数据挖掘技术可以抽取有价值的规律来支持疾病预测与风险评估,促进医疗服务质量的提高。在疾病预测领域中,机器学习算法由于具有优秀的模式识别能力而被广泛使用,但是不同的算法在复杂关系处理中表现出了明显的区别。本项研究针对逻辑回归、支持向量机、随机森林以及深度神经网络进行了深入的比较分析,旨在为算法的进一步优化提供有价值的参考依据。

1 医疗大数据挖掘技术概述

医疗大数据是医疗健康领域产生的各种数据形式以及信息资源的总称,主要包括电子病历,医学影像,实验室检验数据,基因组信息,可穿戴装备数据以及药物使用记录[1]。医疗大数据的挖掘主要涉及数据预处理、特征提取、模式识别和结果分析四个步骤[2]。疾病预测基本过程包括：数据采集,数据预处理,模型训练及测试,模型评估及优化。

2 疾病预测的理论与方法

2.1逻辑回归在疾病预测中的应用。逻辑回归作为一种经典的二分类问题解决手段,在预测疾病时经常被用来确定患者是否受到某种疾病的困扰。通过以若干影响因素为自变量进行逻辑回归,可构造线性模型对二分类结果进行概率的预测。原理是利用Sigmoid函数把特征变量线性组合映射在0~1区间内,以此

来计算发病概率。逻辑回归模型具有模型训练快速,解释性好等优点,适合大规模高维数据的处理[3]。

2.2支持向量机的基本原理及应用。支持向量机是一种监督学习算法,它通过构建超平面来最大化分类间隔,从而进行数据分类。核心思想就是利用非线性映射把原始数据向高维特征空间投影,从而可以在这个高维空间内寻找线性可分超平面。SVM在疾病预测中的应用主要体现在处理非线性数据和高维数据上,通过引入核函数(例如,径向基函数,线性核,多项式核等),SVM可以有效应对复杂的特征关系。

2.3神经网络与深度学习技术的应用。神经网络采用多层结构仿真人脑中神经元连接方式,由输入层至输出层经过多次权重更新逐步拟合出数据复杂规律。深度学习以传统神经网络为基础,进一步提高隐藏层个数,使模型能捕获到数据中更深的特征以提高预测精度。常见的深度学习架构,例如卷积神经网络,循环神经网络等,在医学影像分析,基因数据预测等方面都有着显著优势[4]。

2.4模型评估指标与算法性能优化。为对疾病预测模型性能进行综合评价,通常采用的评价指标有准确率,召回率,F1值,AUC等。准确率度量了模型对全部样本的正确分类率并适用于样本类别平衡的情景；召回率主要关注于识别正类样本的能力,并能有效地评价模型在漏诊率方面的性能。F1值是精确度与召回率之间调和平均的结果,比较适合样本类别失衡的场景。并且AUC体现了模型总体分类性能,ROC曲线下方区域的计算度量了其区分能力[5]。

3 模拟仿真实验设计与实施

3.1实验数据集的选择与预处理。为了证实医疗大数据挖掘技术在疾病预测方面的实用性,我们选择了多个公开的医疗数据集进行研究,包括UCI Machine Learning Repository中关于糖尿病和心血管疾病的数据集。这些数据集涉及不同的特征变量,例如年龄,性别,血压,血糖和体重指数。由于原始数据集含有噪声,缺失值以及异常值等特征,为了改善数据质量,该研究先对数据预处理。具体步骤包括数据清洗(删去缺失值大于某一比例的样本及特征)、缺失值填补(使用均值填补或者插值的方法)、特征归一化(如果把所有的特征都缩放在[0,1]的范围内)和特征选择(利用卡方检验或者主成分分析,筛选出对预测结果有显著影响的性状)。

3.2不同算法的实验方案设计。在实验方案设计中,选择了逻辑回归、支持向量机、随机森林和深度神经网络四种典型的机器学习算法进行对比分析。这些算法分别代表了线性模型、核方法、集成学习方法和深度学习方法,能够较好地涵盖疾病预测中常见的建模方法。实验过程中,逻辑回归用于建立基线模型,其损失函数定义为对数似然损失：

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log h_\theta\left(x^{(i)}\right) + \left(1-y^{(i)}\right)\log\left(1-h_\theta\left(x^{(i)}\right)\right)\right]$$

其中, $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$ 是逻辑回归模型的输出概率,是模型参数,是样本数量。

对于SVM,其目标是通过最大化分类间隔来提高分类精度,损失函数定义为：

$$L(w,b) = \frac{1}{2}|\quad |w|\quad |^2 + C\sum_{i=1}^{m}\max\left(0,1-y_i(w\cdot x_i+b)\right)$$

其中,w为权重向量,b为偏置,C为惩罚参数,用于控制分类间隔与误分类率之间的平衡。

随机森林则通过集成多棵决策树来增强模型的鲁棒性和泛化能力,其决策函数为多个树模型输出的加权平均：

$$f(x) = \frac{1}{N}\sum_{i=1}^{N}T_i(x)$$

其中,$\mathrm{T_i}(x)$为第i棵树对样本 x 的预测结果,N为树的数量。

深度神经网络通过多层非线性变换来学习数据的复杂模式,其损失函数采用交叉熵损失函数：

$$L = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\hat{y}^{(i)} + \left(1-y^{(i)}\right)\log\left(1-\hat{y}^{(i)}\right)\right]$$

其中, $\hat{y}^{(i)}$ 是模型对第i个样本的预测值。

3.3模型训练及参数调优过程。在模型训练时,采用参数调优的方法增强预测性能。在逻辑回归中,调优主要针对正则化项,而SVM对核函数参数进行调优；随机森林所涉及的树数,最大深度和节点分裂最小样本数等；深度神经网络对学习率,隐藏层数目,神经元数目以及激活函数进行了调优。利用网格搜索与随机搜索相结合的调优方式,通过网格搜索来枚举参数组合并通过随机搜索对参数空间内的参数进行随机抽样以提高工作效率。为了避免过拟合的发生,将L2正则化以及Dropout技术引入深度神经网络。

3.4交叉验证及实验数据的划分方法。为了验证该模型的泛化能力,本文使用K折交叉验证的方法对数据集进行K次子集划分,每使用1次子集作为验证集,剩余部分作为训练集,重复K次,取其均值,通常K值5或者10来权衡计算成本和评估精度。将数据集按照8:2比例分为训练集与测试集对未知数据进行性能评价。

4 实验结果与分析

4.1实验结果的性能比较。在本次实验中,使用逻辑回归、支持向量机、随机森林和深度神经网络四种算法对医疗数据集进行了训练和预测。表1展示了各模型在测试集上的四个主要评估指标(准确率、召回率、F1值、AUC)的实验结果。

从表格1可以明显观察到,深度神经网络在四个评价标准上都超越了其他模型,特别是在AUC值(0.94)上,这表明它在处理复杂的非线性特性时展现出了更为出色的分类性能。随机森林在准确性(0.90)和F1值(0.89)方面表现出色,这进一步证明了集成学习方法在增强模型稳定性方面具有明显的优越性。尽管支持向量机在整体性能上略逊于随机森林和深度神经网络,但

其表现仍然优于逻辑回归，这证明了它在处理非线性问题时的高效性。

表1 不同模型在测试集上的性能比较

模型	准确率	召回率	F1值	AUC值
逻辑回归	0.85	0.82	0.83	0.87
支持向量机	0.87	0.85	0.86	0.89
随机森林	0.9	0.88	0.89	0.92
深度神经网络	0.92	0.9	0.91	0.94

4.2四个评估指标的结果对比。该研究对4种评估指标进行了比较，并揭示了各种算法的优劣性。深度神经网络与随机森林预测精度都达到了0.90以上，对疾病预测具有突出的性能。深度神经网络的召回率高达0.90，有效地减少了漏诊率，其中F1的值是最高的(0.91)，这表明它的综合性能是最好的。支持向量机的F1值最低(0.86)，略逊于其他非线性模型。在AUC值方面，深度神经网络的表现是最为出色的(0.94)，而随机森林和支持向量机的表现则紧随其后，两者的表现均超过0.89，这表明这三种算法在样本区分方面具有较高的能力。逻辑回归AUC值达到0.87，说明线性模型存在局限性。

4.3结果的可视化展示与讨论。为了更清晰地呈现模型在准确率、召回率、F1值和AUC这四个评价标准上的性能，图1展示了这四种算法在这些标准下的性能比较。深度神经网络的各项指标性能最好，特别是AUC值，为0.94，说明该网络对复杂特征及大规模数据的分类能力及稳健性较强。其次是随机森林的准确率及F1值表现突出。

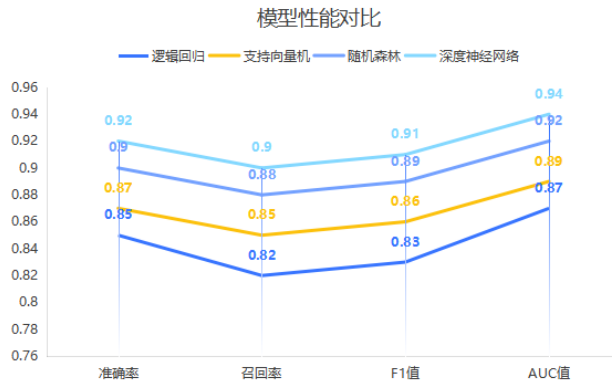

图1 不同机器学习模型在各项评估指标上的性能对比

4.4模型预测性能的改进建议。根据实验结果可对每个模型的表现提出改进意见。对逻辑回归而言，可通过引入多项式特征或者交互特征来提高模型非线性能力进而增强预测效果。对于SVM，可以通过调节核函数参数(如径向基核的宽度)或采用更高级的核函数来提升模型的泛化能力。对随机森林及深度神经网络可通过增大数据集规模或者使用数据增强技术来增强模型鲁棒性。也可通过采用集成学习方法综合多个模型预测结果来进一步提高疾病预测性能。表2列出了使用以上改进策略的4种模型在性能上的变化，以供后续分析对比。

表2 不同模型改进前后性能比较

模型	改进前准确率	改进后准确率	改进前召回率	改进后召回率	改进前F1值	改进后F1值	改进前AUC	改进后AUC
逻辑回归	0.85	0.88	0.82	0.85	0.83	0.86	0.87	0.89
支持向量机	0.87	0.89	0.85	0.87	0.86	0.88	0.89	0.91
随机森林	0.9	0.92	0.88	0.9	0.89	0.91	0.92	0.94
深度神经网络	0.92	0.93	0.9	0.92	0.91	0.92	0.94	0.95

通过上述改进，所有模型的评估指标均有不同程度的提升，尤其是逻辑回归和支持向量机的改进效果最为显著。这表明在模型优化过程中，合理的特征工程和参数调优能够有效提升预测的准确性和稳健性。

5 结论

研究实证分析4种机器学习算法用于医疗大数据疾病预测。结果表明：深度神经网络对非线性特征及大规模数据的处理效果最好，AUC值达到了0.94，分类能力强，泛化性能好。在准确性和F1值方面，随机森林表现得比支持向量机和逻辑回归更为出色，这证明了集成学习方法的稳健性。传统算法(SVM和逻辑回归)虽不如深度学习，但通过特征工程和参数调优可提升性能。研究给出了高级特征工程，参数调优和集成学习策略等优化建议，并对算法的优化进行了定位。

[参考文献]

[1]Partha S.Understanding Differences in Behaviour Patterns of Healthcare Service Elements Among Regions Applying Data Mining[J].Journal of Health Management,2023,25(2):369-381.

[2]于国庆,沈飞.数据挖掘技术在医疗大数据分析中的应用——评《医疗大数据分析与数据挖掘处理研究》[J].中国科技论文,2022,17(07):847.

[3]秦泽宁,崔雨萌.大数据助力现代医疗体系建设研究[J].网络安全技术与应用,2022,(04):106-107.

[4]童俊.数据挖掘技术在医疗大数据中的应用[J].电子技术与软件工程,2021,(12):156-157.

[5]郑秀娟.数据挖掘技术在医疗大数据中的应用研究[J].电脑知识与技术,2020,16(32):26-27+35.

作者简介：

张志超(1986--),男,汉族,天津人,硕士研究生,任职于天津市斯宇克医疗器械销售有限公司,研究方向：医疗大数据分析与实践应用。