决策树及支持向量机与深度学习模型在肝癌鉴别诊断中的比较研究

黄辛迪[1],黄 慧[1],刘佳俊[1],丁长松[1,2]

(1.湖南中医药大学信息科学与工程学院,湖南 长沙 410208;
2.湖南省中医药大数据分析实验室,湖南 长沙 410208)

摘要:目的 使用数据挖掘技术研究肝功能检查数据,分析肝功能检查指标与肝癌诊断的关联,探究肝癌早诊断、早治疗的辅助数据分析方法。**方法** 构建决策树C4.5模型并提取决策方法,并以Bagging方法优化;采用网格划分法和粒子群优化算法优化支持向量机模型;构建多层感知机(MLP)和卷积神经网络(CNN)进行性能比较。基于决策树和SVM模型进行特征属性分析和最优特征子集选择。**结果** Bagging决策树模型、SVM、MLP模型的10交叉检验准确率分别为95.18%、95.60%、90.17%,测试准确率分别为94.34%、93.40%、89.78%。在肝功能检查指标中,碱性磷酸酶、谷丙转氨酶、天门冬氨酸转氨酶、年龄、直接胆红素是主要贡献指标,三指标联合诊断对肝癌预测率达86.08%。**结论** 决策树、支持向量机、多层感知机建立的肝癌分类器模型都可用于肝癌辅助诊断,SVM模型略优,预测模型对肝癌早期鉴别有较好的辅助作用。

关键词:肝癌;决策树;支持向量机;深度学习;多层感知机;卷积神经网络

中图分类号:R735.7 **文献标识码:**A **DOI:**10.3969/j.issn.1006-1959.2023.15.012

文章编号:1006-1959(2023)15-0070-05

Comparative Study of Decision Tree, Support Vector Machine and Deep Learning Model in Differential Diagnosis of Liver Cancer

HUANG Xin-di[1],HUANG Hui[1],LIU Jia-jun[1],DING Chang-song[1,2]

(1.School of Information Science and Engineering,Hunan University of Chinese Medicine,Changsha 410208,Hunan,China;
2.Big Data Analysis Laboratory of Traditional Chinese Medicine in Hunan Province,Changsha 410208,Hunan,China)

Abstract:Objective To study liver function test data with data mining technology, analyze the correlation between liver function test indicators and liver cancer diagnosis, and explore methods for early diagnosis and early treatment of liver cancer.**Methods** The decision tree C4.5 model was constructed and the decision method was extracted and optimized by Bagging method. The grid division method and particle swarm optimization algorithm were used to optimize the support vector machine model. Multi-layer perceptron (MLP) and convolutional neural network (CNN) were constructed for performance comparison. Based on decision tree and SVM model, feature attribute analysis and optimal feature subset selection were carried out.**Results** The 10-cross-check accuracy rates of Bagging decision tree model, SVM, and MLP model were 95.18%, 95.60%, and 90.17%, respectively, and the test accuracy were 94.34%, 93.40%, and 89.78%, respectively. Among the liver function test indicators, alkaline phosphatase, alanine aminotransferase, aspartate aminotransferase, age and direct bilirubin were important indexes in the diagnosis of liver cancer with three-index combined diagnosis up to 86.08% in accuracy.**Conclusion** The liver cancer classifier models established by decision tree, support vector machine and multilayer perceptron can all be used in diagnosis of liver cancer, SVM model slightly better. The prediction models are supplementary measures for early identification of liver cancer.

Key words: Liver cancer;Decision tree;Support vector machine;Deep learning;Multi-layer perceptron;Convolutional neural network

原发性肝癌(primary hepatic carcinoma)是全球最常见的恶性肿瘤之一,占癌症发病总数的4.7%和癌症致死总数的8.3%,也是肿瘤致死病因的前3位,严重威胁民众健康[1]。原发性肝癌具有病情进展快、肿瘤转移速度快、术后复发率高等临床特点,早诊断、早干预、早治疗是降低肝癌死亡率的有效方法[2,3]。近年来,现代医学研究在肝癌致病因素、血清学生物标志物、基因组学、代谢组学、病理、生存分析等方面积累了大量相关数据[4,5],而大数据、人工智能等技术的发展为肝癌的危险因素分析[6]、有效诊断[7]、分型[8]、辅助治疗[9]等提供了有效辅助手段。临床常应用血清甲胎蛋白(a-fetoprotein,AFP)、GGT、GGT/ALT[10,11]等标志物筛查判断肝癌,但AFP

基金项目:1.湖南省中医药科研计划重点课题(编号:2020002);2.长沙市自然科学基金项目(编号:kq2202265);3.湖南中医药大学校级科研项目(编号:2019XJJJ029)

作者简介:黄辛迪(1987.10-),女,湖南长沙人,硕士,讲师,主要从事中医药信息学、大数据技术的研究

不高者也不能排除患有肝癌的可能性，有 30%~40%的肝癌患者 AFP 呈阴性[12]。加之医学影像检查费用高、肝癌患者早期症状不明显、就诊意识薄弱[13]，因此体检常规肝功能检查对肝癌诊断更具大众化、方便快捷的优势[14]，对增加肝癌早期诊断的灵敏性和普及性具有重要作用。本研究使用数据挖掘技术研究肝功能检查数据，分析肝功能检查指标与肝癌诊断的关联，探究肝癌早诊断、早治疗的辅助数据分析方法。

1 资料与方法

1.1 资料来源 研究数据来自印度 Ramana、Babu 等教授公布的印度肝癌患者数据集，共 583 条，其中肝癌患者数据 416 条，非肝癌患者数据 167 条。数据集中男性患者数据 441 条，女性患者数据 142 条。数据集有 10 个属性特征和 1 个标记位，见表 1。

表 1 数据集结构表

标号	属性名称	取值/单位
1	年龄(age)	4~90(≥90 的都归为 90)
2	性别(sex)	男、女
3	总胆红素(TB)	μmol/L
4	直接胆红素(DB)	μmol/L
5	碱性磷酸酶(ALP)	μ/L(每 L 中酶活力浓度)
6	谷丙转氨酶(ALT)	μ/L
7	天门冬氨酸转氨酶(AST)	μ/L
8	总蛋白(TP)	g/L
9	白蛋白(ALB)	g/L
10	白蛋白/球蛋白(A/G)	/
11	是否患有肝癌(label)	1=患有，2=未患有

1.2 方法

1.2.1 数据预处理 ①处理缺失值：属性 10 A/G 存在 4 个(1%)空缺值，采用均值补全；②属性 age 采用等宽离散化，age 离散化用于决策树算法中计算信息增益；③由于数据集中的正反例数据不平衡，采用随机抽样缓解数据不平衡问题；④采用随机抽样方法划分训练集与测试集；⑤使用 PCA 主成分分析法，计算属性的方差贡献率进行数据降维。

1.2.2 机器学习方法 ①决策树算法：采用决策树 C4.5 算法以最大信息增益率的属性作为分裂条件，对属性设置规则，使得分支节点所包含的样本尽可能属于同一类别，最终自底向上剪枝构建决策树[15]。使用 WEKA3.8.0 软件，运用决策树 C4.5 算法进行模型构建与预测，调试相关参数得到实验结果。基于 sklearn 运用 Bagging 算法构建多棵决策树并行，提高决策准确率；②支持向量机算法：SVM 使用超平面对数据集进行分类，本研究采用最小化经验风险和结构风险的线性组合和随机梯度下降求解最优分割超平面，实现支持向量到决策超平面的距离最大化[16]。SVM 在不同核函数下构建的分类模型性能不同，本研究选用 RBF 径向基函数性能最优。基于 Python 和 sklearn，采用网格划分方法和粒子群优化算法对算法参数惩罚因子 C 和核参数 g 进行优化构建 SVM 模型；③深度学习方法：MLP 输入输出层中包含多个隐层，以全连接神经网络对信息特征进行提取和整合实现数据分类；CNN 中使用权值共享卷积核层级式特征提取的神经网络实现数据分类。深度学习模型经过超参数和参数调试和优化，确定最终模型。使用 Pytorch 完成深度学习模型的构建，深度学习模型采用多层感知机(MLP)和卷积神经网络(CNN)。

1.2.3 特征属性分析 ①属性重要性分析：采用基于 Bagging 的决策树模型，对每棵决策树计算特征不纯度再取平均，根据平均不纯度大小对特征重要性排序，不纯度使用 Gini 值；对每棵决策树加入随机噪声，计算前后的平均袋外数据误差，即对样本的准确率的影响，根据准确率对特征的重要性进行排序；②特征子集选择：使用 SVM 模型作为学习器和 Wrapper 包装法，采用循环删除单个重要性低的特征和循环加入重要性高的特征相结合的混合搜索方式。

2 结果

2.1 数据预处理结果 缺失值处理：A/G 有 4 个空缺值，以该项均值 0.947 补全。age 属性离散化用于决策树判别，age 分成 3 组采用等宽离散化，取值范围分别为[min，32.7]、[32.7，61.3]、[61.3，max]。缓解正反例数据不平衡：采用随机抽样得到正反两类类别分别为 402 例和 181 例，使正反比例接近于 2:1。建立训练集和测试机：采用随机抽样得到 477 例为训练集，106 例为测试集，用于后续建立判别模型。数据降维：使用 PCA 降维，经计算得前 9 个属性的方差贡献率达到 99.438%，故删去方差贡献率最低的属性 ALB，见图 1。

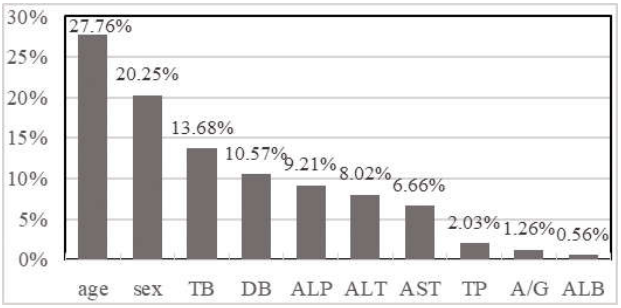

图 1 属性方差贡献率

2.2 决策树算法预测 采用 C4.5 算法建立决策树，对置信度、叶子节点最小实例数、子树上升 3 个参数进行调节，确定对应取值为 0.25、2 和 True 时，随机抽样下决策树的总准确率为 92.45%。提取的决策规则见表 2，基于此规则采用 Bagging 算法构建决策树模型，训练后模型过拟合，在训练集上准确率达到 100.00%，且采用 Bagging 算法的决策树预测结果高于单一决策树模型，见表 3。

2.3 支持向量机 SVM 预测 由于支持向量机选择只有 2 个参数，使用网格划分方法准确率十分接近，综合考虑最终选择结果稳定的网格划分方法得出的参数，即(C,g)=(1,222.8609)。10 折交叉检验和测试集的预测结果见表 4、图 2。

表 2 决策树规则表

规则	类别	准确率(%,总实例数/错误预测)
R1:(ALT≤32)(ALP≤187)(TP≥7.1)	no	81.25(32/6)
R2:(TB≤1.6)(ALP≤210)(AST≤23)(AST≥16)	no	72.34(47/13)
R3:(DB≤1)(ALP≤165)(TP≤6.8)(ALP≥128)	no	90.48(21/2)
R4:(TB≤1.6)(TB≥0.9)(DB≤0.2)(AST≥32)	no	86.96(23/3)
R5:(DB≤1)(ALP≥194)(ALP≤271)(A/G≤1.2)(AST≤74)(TB≥1.1)	no	94.44(18/1)
R6:(sex=Female)(TP≤7.2)(ALT≤34)(ALP≥194)(AST≥12)	no	93.75(16/1)
R7:(age=youth)(TB≤0.5)	no	100.00(3/0)
R8: -R1-R2-R3-R4-R5-R6-R7	yes	96.85(317/10)

表 3 决策树与 Bagging 算法的优化结果

数据集	方法	灵敏度	特异度	F1 值	均方根误差	ACC(%)
测试集	决策树	0.896	0.825	0.896	0.2976	89.62
10 折交叉检验	决策树	0.891	0.854	0.891	0.3139	89.10
测试集	Bagging+决策树	0.943	0.890	0.943	0.2111	94.34
10 折交叉检验	Bagging+决策树	0.952	0.928	0.952	0.2144	95.18

表 4 实验结果对比表

数据集	灵敏度	特异度	F1 值	AUC	均方根误差	ACC(%)
测试集	0.934	0.840	0.931	0.938 06	0.2642	93.40
10 折交叉检验	0.956	0.895	0.955	0.927 14	0.2098	95.60

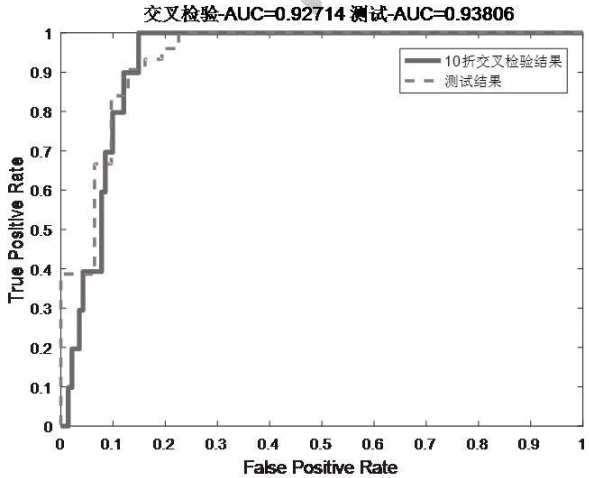

图 2 SVM 模型 ROC 曲线

2.4 深度学习模型预测 MLP 模型包含 3 个隐层，隐层节点数分别为 20、40、20，学习率设置为 0.003。CNN 模型包含 2 个隐层，隐层节点数为 30，卷积核大小为 5，步长为 1，padding 为 3，学习率为 0.001。模型的训练和测试采用 10 折交叉验证，结果显示 3 层 MLP 和 2 层 CNN 模型基本已达到预测功能，见表 5。

2.5 基于 Bagging 的决策树模型特征属性分析 根据不纯度和准确率 2 种方式，建立模型来对特征的重要性排序，结果显示 age 特征比较重要，而特征 TP、ALB、sex 和 ALB 重要性较低。此外，特征 DB 在基于不纯度的方法中重要性排名靠后，而在基于准确度的方法中排名靠前，见图 3、图 4。

表 5 深度学习模型的实验结果

数据集	方法	灵敏度	特异性	F1 值	均方根误差	ACC(%)
测试集	MLP	0.859	0.897	0.894	0.103	89.78
10 折交叉检验	MLP	0.876	0.928	0.899	0.098	90.17
测试集	CNN	0.787	0.826	0.778	0.213	80.23
10 折交叉检验	CNN	0.791	0.839	0.809	0.185	81.50

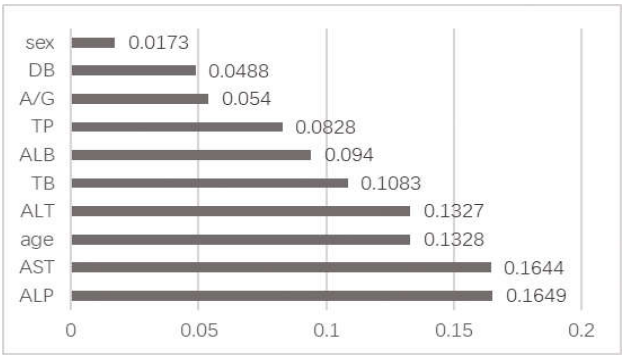

图 3 基于不纯度的特征排序

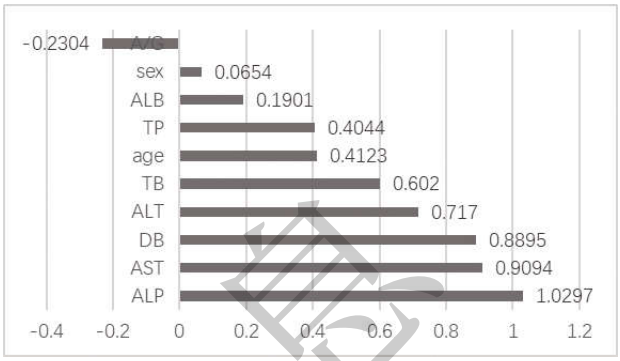

图 4 基于准确率的特征排序

2.6 支持向量机 SVM 特征子集选择 通过循环搜索策略，结果显示 ALP、ALT 和 AST 三个指标的联合诊断和 age、ALP 和 ALT 三指标的联合诊断的性能都达到最优，阳性检出率为 86.08%，约登指数为 0.641，见表 6。

表 6 联合诊断的性能

指标	灵敏度	特异度	约登指数	AUC	阳性检出率(%)
ALP+ALT+AST	0.889	0.752	0.641	0.820	86.08
age+ALP+ALT	0.889	0.752	0.641	0.820	86.08
sex+DB+ALB	0.796	0.668	0.464	0.732	82.09
sex+TP+ALB	0.830	0.662	0.492	0.746	81.89
sex+DB+TB	0.755	0.665	0.420	0.710	81.82

3 讨论

近年来，基于医学影像、生化指标的疾病预测研究显示，运用机器学习、深度学习、迁移学习等建立诊断模型具有较高的诊断价值[17,18]。在小数据样本上，机器学习预测模型预测准确率相较于深度学习模型有一定差异，在医疗诊断逐渐电子智能化的背景下对于疾病诊断的效率及准确率具有辅助作用。

本研究中单一决策树的 10 折交叉检验准确率为 89.10%，决策树结合 Bagging 算法的 10 折交叉检验准确率为 95.18%。因为决策树结合 Bagging 算法强强制约构建多棵决策树，通过投票机制可以解决决策树容易过拟合的问题，提高模型的准确率，可并行化提高建模速度。但是由于创建了多棵决策树，使用的多数投票规则会使得模型变得更加复杂，剥夺了单棵决策树直观可解释性，且易过拟合，其泛化性能需进一步研究。

SVM 模型的 10 折交叉检验准确率为 95.60%，且所有评估指标都优于单一决策树模型，与 Bagging+决策树模型的判断效果差异不大。以上结果提示，参数选择时使用交叉检验的准确率作为评优标准有利于提高泛化能力。本研究结果还发现，不均衡类问题对算法的性能优劣有较大的影响，提示预测模型只能作为辅助作用而不能完全依赖，对于诊断结果还要依据实际情况分析。

另外，深度学习模型一般需要大量数据进行特征学习，而本研究的数据集有限，模型极易产生过拟合和欠拟合。在本研究数据集特征较为有限的情况下，结果显示 3 层 MLP 和 2 层 CNN 模型基本已达到预测功能，准确率在 90%左右，与本研究中的单一决策树模型基本相当，略低于基于 Bagging 的决

策树模型和SVM模型。可见,多层感知机MLP在本研究小数据集和小模型规模下已达到较高的准确率,另外可能由于样本中每个特征本身较独立,因此MLP在本研究中优于CNN模型。王钰涵等[19]在基于决策树和神经网络的高血压病危险因素研究中也发现,决策树和MLP神经网络的准确率较优。

从本研究中决策树的判别条件来看,提取的反例判别方法主要为ALP、ALT、TB、DB、age的值域判别组合。在R5、R6规则中,ALP≥194时,联合AST、TB、A/G等进行多重判别,也可排除其肝癌可能。ALP、ALT和AST在决策树和SVM模型中都是排名较前的重要特征。相关研究也显示[20],AFP联合血清酶可提高原发性肝癌的诊断阳性率。多指标的联合诊断效果优于单指标的效果,本研究中ALP+ALT+AST、age+ALP+ALT、sex+DB+ALB三个指标的联合诊断的性能在所有指标集合中诊断效果较好,也就是说肝癌的判别指标中ALP、ALT、AST、age是主要影响因素,阳性检出率达到86.08%,对于肝癌预测和诊断具有一定的辅助作用。基于准确率的Bagging决策树模型中,age特征比较重要,而特征TP、ALB、sex和A/G重要性较低,特征DB在基于不纯度的方法中重要性排名靠后,而在基于准确度的方法中排名靠前,可能因为TB和DB存在较大相关性。

综上所述,决策树、支持向量机、多层感知机建立的肝癌分类器模型都可用于肝癌辅助诊断,SVM模型略优,预测模型对肝癌早期鉴别有较好的辅助作用。

参考文献:

[1]刘宗超,李哲轩,张阳,等.2020全球癌症统计报告解读[J].肿瘤综合治疗电子杂志,2021,7(2):1-14.

[2]中华人民共和国国家卫生健康委员会.原发性肝癌诊疗指南(2022年版)[J].肿瘤防治研究,2022,49(3):251-276.

[3]Singal AG,Pillai A,Tiro J.Early detection, curative treatment, and survival rates for hepatocellular carcinoma surveillance in patients with cirrhosis: a meta-analysis[J].PLoS Med,2014,11(4):e1001624.

[4]王敏,陈泽峰,韩志伟,等.基于TCGA数据库分析肝细胞肝癌组织中HMMR表达与患者临床特征及预后的相关性[J].现代肿瘤医学,2021,29(7):1173-1178.

[5]王玉,杨雪,靳晓杰,等.基于中医药整合药理学平台、GEO数据库芯片及分子对接探讨大黄抗肝癌的作用机制[J].中草药,2020,51(20):5207-5219.

[6]Ibragimov B,Toesca DAS,Chang DT,et al.Deep learning for identification of critical regions associated with toxicities after liver stereotactic body radiation therapy[J].Med Phys,2020,47(8):3721-3731.

[7]Nishida N,Yamakawa M,Shiina T,et al.Current status and perspectives for computer-aided ultrasonic diagnosis of liver lesions using deep learning technology [J].Hepatol Int,2019,13(4):416-421.

[8]Owens AR,McInerney CE,Prise KM,et al.Novel deep learning-based solution for identification of prognostic subgroups in liver cancer (Hepatocellular carcinoma)[J].BMC Bioinformatics,2021,22(1):563.

[9]Zheng C,Chen L,Jian J,et al.Efficacy evaluation of interventional therapy for primary liver cancer using magnetic resonance imaging and CT scanning under deep learning and treatment of vasovagal reflex[J].Journal of Supercomputing,2021,77(7):7535-7548.

[10]胡仁智,赵世巧,申波,等.血清甲胎蛋白及其异质体和异常凝血酶原对原发性肝癌的诊断价值 [J]. 中华肝脏病杂志,2019,27(8):634-637.

[11]赵晓玲,王晶晶,赵巧玉,等.甲胎蛋白异质体比率在原发性肝癌鉴别诊断中的应用 [J]. 国际检验医学杂志,2016,37(9):1228-1229,1231.

[12]中华人民共和国卫生部.原发性肝癌诊疗规范(2011年版)[J].临床肝胆病杂志,2011,27(11):1141-1159.

[13]张子梅.利用机器学习方法识别肝癌早期诊断标志[D].成都:电子科技大学,2021.

[14]钟锐,张俊.肝功能指标在原发性肝癌中的诊断价值[J].中国现代医学杂志,2015,25(8):102-105.

[15]聂斌,李欢,罗计根,等.融合GINI指数的C4.5算法的分类研究[J].江西师范大学学报(自然科学版),2019,43(5):469-472.

[16]蔺轲,谢俊卿,胡永华,等.支持向量机在ICU急性肾损伤患者住院死亡风险预测中的应用[J].北京大学学报(医学版),2018,50(2):239-244.

[17]余美慧,袁泉,曾书娥,等.基于超声图像的迁移学习模型在乳腺肿块良恶性鉴别诊断中的价值 [J]. 临床超声医学杂志,2022,24(9):652-656.

[18]吴树才,王新举,纪俊雨,等.基于深度学习卷积神经网络的肺结核CT诊断模型效能初探 [J]. 中华结核和呼吸杂志,2021,44(5):450-455.

[19]王钰涵,段鹏喆,张鑫,等.基于决策树和神经网络的高血压病危险因素研究[J].世界科学技术-中医药现代化,2021,23(8):2784-2794.

[20]赵斌,赵经川,李昭宇.AFP联合三项血清酶对原发性肝癌的诊断评价[J].宁夏医学杂志,2010,32(1):20-21.

收稿日期:2022-08-29;修回日期:2022-10-31

编辑/杜帆