机器学习辅助肿瘤诊断探索

杨克戎
（遵义医科大学医学信息工程学院，贵州 遵义 563000）

摘要：随着现代信息技术的发展，人工智能已经成为重要的研究热点，并且在各个领域中得以应用，展现出较高的应用价值与意义。机器学习作为人工智能发展的重要板块，在现代医学中已经展现出重要的应用潜力，尤其在辅助肿瘤诊断方面得到了深入研究。本文即以此为研究背景，通过了解机器学习在人工神经网络与深度学习中的发展背景，进而探析机器学习在乳腺癌与结肠肿瘤诊疗中的应用效果。

关键词：机器学习；辅助肿瘤诊断；乳腺癌；结肠肿瘤

一、人工神经网络与深度学习

人工神经网络是20世纪80年代以来提出的一种抽象性概念，其以信息处理的角度为人脑神经网络建立了一种模型，借助不同的连接方式形成了一种独特的大脑网络。神经网络即为以后只能怪运算模型，通过大量节点相互联结，形成一个互联整体，每一个节点可以作为一个输出函数，节点间的连接则称为权重，由此生成一种独特的函数模型。

深度学习则是机器学习的一个研究分支，指的是通过探寻样本数据中的规律，进而在学习过程中对相应的信息有更深的理解，最终目标在于让机器具备像人一样的分析、思考、判断能力。

目前人工神经网络的研究正在快速进展，并且帮助人类在智能机器人、自动控制、医学、经济等领域解决了诸多问题，具有杰出的智能特性。而深度学习则在数据挖掘、机器学习等方面展现出更高级的应用成效。

二、机器学习应用现况

机器学习最早用于语音与图像的识别工作之中，而随着信息技术的发展，智能系统在处理数据方面的能力展现出得天独厚的优势，同时在GPU计算模块的辅助下，矩阵运算能力持续提升，进而将机器学习的技术应用拓展到更多层面。

随着现代科技的发展，机器学习早已成为研究热点中的佼佼者，众多大型公司都在研究与探索，比如Google公司就以机器学习为基础，开发了多国语言识别与翻译系统；苹果公司则开发了Siri语音服务系统；2016年，Alphago成为围棋界的一匹人工智能黑马，这些案例都展现出人工智能在现代社会中的具体价值与功能。

在机器学习发展进程中，深度学习在其中展现出更突出的效果，卷积神经网络作为图像处理能力最强的技术，也成了大型公司用于开发各类图形识别系统的关键，尤其在智能手机、相机以及智能机器人等方面。

而在医学领域，机器学习也有重要的应用空间，自20世纪90年代起，医学领域中就开始应用图像识别技术进行诊疗研究，美国就针对皮肤癌开发了具有人工智能属性的检测系统，在临床验证中发现，其检测结果与专家诊断结构具有较高的重合率，同时还展现了其更高的特异性与灵敏度。而后很多学者都开始深入研究机器学习在医学领域中的应用，尤其在处理肿瘤基因检测数据中具有较高的应用价值，并逐步开始发展到辅助肿瘤诊断方面。

三、机器学习在乳腺癌中的研究应用

（一）数据处理与模型特征

1. 收集数据集

首先要进行数据收集，以乳腺影像学检查设备为基础，通过对患者进行CT拍摄，获取其分层图像集，针对乳腺癌患者则要通过DOT光学成像的方式，以非接触形式采集数据，过程中需要患者俯卧于设备扫描床上，将一侧乳房置于扫描腔内，而后开始扫描，扫描的过程要从胸壁开始，逐层下降高度，一般根据患者的实际情况，每层高度为1–4毫米，直到完成扫描。

2. 预处理数据

根据扫描过程中收集的数据，设定阈值划分，同时通过三维建模的方式将数据分给多个三维字块，根据图像识别的方式，将其中具有异常特征的字块进行特殊标记，对正常字块进行普通标记，同时采取三维分离方法，将所有患者的数据集进行重建，形成若干子VOI，而由于正常实例过多，因此还需要进行采样，对数据进行平衡处理，以此强化分类器的功能性。

3. 数据增强

当从医院收集的数据较少时，则要通过数据增强的方式进行扩充，通过大量增加样本的数量，保证模型展现出良好的鲁棒性。常见的增强手段有三种，其一为几何变换，即通过对数据的翻转、缩放变形、剪裁等实现；其二为噪声类手段，即在原有的图像上以随机的方式叠加噪声，由此实现数据扩充目的；其三为模糊类手段，通过降低像素点值引起的差异，将图像进行模糊处理，将像素进一步平滑化。

4. 特征选取

纹理特征是分析与解释图像中运用的重要特征，尤其在临床成像中，纹理特征可以展现其表面强度与属性特征，比如平滑度、粗糙度以及规则性等，因此通过对大量样本的识别训练，可以让只能系统具备描述可疑区域的能力，从而了解正常组织与异常组织之间的区别和差异。鉴于纹理特征的特殊价值，在恶性肿瘤判别中成为重要的依据，通过对CTLM图像的检测与识别，即可完成判断与诊疗目的。因此在实际应用中，需要以精确的算法对三维子区域的纹理特征进行描述与提取，从而完成人工智能诊疗的目标。

（二）构建乳腺癌诊断模型

1. 评级标准

适应阈值算法模型是构建评级标准的重要基础，其需要三个指标完善其评价标准，其一为Jaccard，可以有效判别个体之间的相似度，即通过识别个体间相同的特征，通过对比判断其特征细节是否一致；Dice系数则主要用于判断两个集合之间的相似性，而三维数据模型中的像素也可作为不同的集合，因此其同样可以衡量三维模型之间的相似程度。体积重叠误差率则用于判断算法所建立的模型与医生检测中形成的实际模型之间的差异，假设算法与实际两个模型之间的像素重合量为1P，而实际模型为P，则体积重叠误差率u公式如下：

$$u=\frac{P-P_1}{P}$$

模型算法应围绕准确率、敏感度、特异度和 F1 分数确定评价标准，因此在乳腺癌诊疗模型的评估中，就要更加清晰地判别其良性和恶性两种类型，进而可以生成四个识别术语，即真阳性、假阴性、真阴性以及假阳性。

2. 阈值算法选择

在算法选择上，要根据三维子块的特征为基础，通过三种三维自适应阈值建模方法建立识别标准，在临床试验中发现，JACARD 和 DICE 指标分别达到了 96.94% 和 99.24%，明显超出了 FCM 和最大熵，因此为了保证识别成果更为严谨，还需要建立自适应阈值划分算法，并通过建模体重叠误差对分割精度进行量化。

3. 分类预测

分类预测中可以建立四种分类模型，Logistic 回归模型的重点在于确定阈值，针对乳腺癌而言，判定其良性与恶性是核心目标，因此可以设定阈值为 0.5，可以更高效率地保证恶性肿瘤的误诊问题。SVM 模型则需要确定惩罚参数以及核函数参数，C 作为其中具有关键调节作用的指标，需要控制其大小稳定，避免过度拟合；核函数则可以设置为线性、RBF、sigmoid 等不同的形式。在临床试验中，一般选择 C=0.001，并且采用 RBF 核函数。C4.5 决策树模型的关键在于两个方面，其中置信因子会影响算法自动剪枝，而最小化实例数量则要判断节点是否包含噪声，实际应用中一般所设置的惩罚因子为 0.5。BP 算法需要对学习率进行优化设置，其大小会影响模型性能的展现，因此通过 K 折交叉验证方法进一步降低泛化误差后，可以有效保证所有样本都能参与到训练之中。

4. 实验结果

在基于纹理特征和紧致度特征的算法下，四种算法模型展现出不同的准确度与敏感性，而任何一种模型的 F1 分数都高于未组合特征算法，可以证明紧致度与纹理特征在组合后更容易反应数据特征，其中 C4.5 决策树模型的分数结果最高，因此在乳腺癌诊疗应用中具有更好的适用性。

四、机器学习在结肠肿瘤诊疗中的应用

（一）图像采集

首先要完成图像采集的环节，针对结肠肿瘤可以采取超声图像作为检测方式，在遵义医学院附属医院实地采集后，共收集图像 130 张，其中恶性 100 张，良性 30 张，是由该医院专家诊断确诊的真是结果。

（二）图像归一化预处理

在预处理环节中，要采取归一化自相关系数和均方差系数展开研究，通过对结肠肿瘤图像中的纹理特征进行分析，分别判断图像粗糙程度的度量以及像素之间的相关性，进而建立相应的数学模型：

$$\gamma(\Delta m, \Delta n) = \frac{A(\Delta m, \Delta n)}{A(0,0)}$$

$$A(\Delta m, \Delta n) = \frac{1}{(M-\Delta m)(N-\Delta n)}$$

$$\sum_{i=0}^{M-\Delta m}\sum_{j=0}^{N-\Delta n}(f(i,j)-\bar{f})(f(i+\Delta m, j+\Delta n)-\bar{f})$$

其中 M、N 指的是诊断过程中形成的感兴趣区域大小；f（i，j）指图像中某一坐标位置的亮度，可以表示感兴趣区域图像的亮度均值。

（三）提取结肠肿瘤轮廓

良性或恶性的结肠肿瘤与正常的组织都表现出不同的物理特性，因此在图像纹理、亮度等特征上表现出较大的差异性，而由此即可通过模型在图像上提取出肿瘤的轮廓，一般提取的方法如下：

1. 针对感兴趣区域图像范围内实施小波变换操作，进而得出图像子带的图像；

2. 在子带图像中找到结肠肿瘤中心或相近之处，并由此向四周作出大量射线；

3. 根据形成的射线，从中点开始向图像边界移动，并针对每一个坐标设置（m，n）大小的评估窗口，在窗口内进行特征识别，找出方差最大的窗口即为结肠肿瘤和正常组织的临界点；

4. 通过对每一条射线作上述操作，进而找到所有临界点后进行连接，由此便完成了结肠肿瘤的轮廓线，确定了其轮廓形状。

（四）提取结肠肿瘤的特征

在临床应用中发现，结肠肿瘤图像在良性与恶性之间的特征差异较为明显，尤其在纹理、亮度、几何等方面都展现出不同的形态特征，因此需要以此为依据，通过构建模型对纹理、对比度、结肠肿瘤形状等特征进行统计整理，进行形成区别良、恶性的判别标准。

（五）建立数学模型

通过数学模型构建生成结肠肿瘤自动判定算法，则是实现机器学习辅助结肠肿瘤诊断的关键，可以采取支持向量机方法，通过监督式机器学习对结肠肿瘤诊断中的两个重要参数进行确定，分别为调节参数 C 和核函数参数，可以采取交叉项参数验证的方法进行计算，主要步骤如下：

1. 把 130 份样本随机分成 k 等分；

2. 取其中一组用来训练诊断系统，其余 k–1 份用来测试诊断系统；

3. 再从步骤 2，余下的 k–1 份中取一组数集用来训练诊断系统，其余 k–1 数集用来测试；

4. 重复步骤 3，训练和测试诊断系统共 k 次，得出最佳参数。

（六）仿真实验

根据上述提出的算法对第二组中 60 份结肠肿瘤图像进行实验，通过支持向量机算法完成仿真，并根据数据成果与专家诊断结果进行对比分析，发现其中存在的问题与不足，并直到算法改进到结果满意。

五、结语

综上所述，机器学习在肿瘤诊断中的应用成为当前医学发展的重要趋势，通过图像识别技术与算法模型的应用，可以在无人工干预的情况下自动对患者进行肿瘤诊断，由此不仅提高了医疗效率，而且大大减轻了医生的工作量，为现代医学发展提供了重要的支持作用。

参考文献：

[1] 李俊衡 . 基于机器学习的 CTLM 乳腺癌辅助诊断的研究应用 [D]. 西北大学，2020.

[2] 张婷婷，渠宁，郑璞 . 机器学习在甲状腺肿瘤诊疗中的应用 [J]. 中国癌症杂志，2017，27（12）：992-995.

[3] 王祥旭，潘伟，张琼，黄陆光，张红梅 . 人工智能辅助恶性肿瘤诊断的应用进展 [J]. 肿瘤防治研究，2020，47（10）：788-792.