直肠癌淋巴结转移的智能诊断

摘 要

近年来，随着中国居民生活水平的提升，直肠癌的发病率和死亡率呈逐步上升态势。直肠癌具有向周围组织浸润并发生淋巴结转移的特点，因此，精准判断淋巴结转移情况在医学研究领域意义重大。

针对本次研究的问题一，我们运用边缘检测算法进行处理。该算法基于名目标区域在强度和特征上的差异，借助Python 实现目标边缘的分割。由于边界处会出现两个不同灰度值的相邻区域，我们通过判断结果中灰度值的连续性，运用微分方法进行检测，具体利用一阶导数和二阶导数来检验边缘吻合度。在实际操作中，采用拉普拉斯高斯算子对灰度图像的边缘进行检验，之后运用数学形态学计算方法对直肠肿瘤的边缘进行修正，最终成功获取直肠肿瘤区域的掩模图像。

在解决问题二时，我们依托问题一中得到的分割图像，从肿瘤的颜色特征边缘特征以及纹理特征这三个维度进行特征提取。具体而言，运用直方图统计、计算 Saturation 的值以及分析图像边界周围像素灰度阶跃变化等方法，实现对肿瘤区域特征的有效提取。

对于问题三，我们选择随机森林算法结合特征选择的方式。通过计算 Gini 系数，明确各个特征对淋巴结转移情况判断的精确有效性。这种方法在有效降低拟合现象的同时，提升了预测精度。我们构建分类型，并采用ice 系数和F-Score 评价指标对模型进行全面评估，以确保型的可靠性和准确性。

关键词：微分数学算法　边缘检测 拉普拉斯高斯算子 随机森林

Intelligent diagnosis of lymph node metastasis in rectal cancer

Abstract

In recent years, with the improvement of people's living standards in China, the in cidence and mortality rates of rectal cancer have gradually increased. Rectal cancer has the characteristics of infiltrating surrounding tissues and metastasizing to lymph nodes . Therefore, accurately determining the status of lymph node metastasis is of great sign ificance in the field of medical research.

For the first problem, we used the edge detection algorithm. Based on the differen ces in intensity and features of each target area, we used Python to segment the target e dges. Since adjacent regions with two different gray values appear at the boundary, we detected by judging the continuity of gray values in the results and used the differentia l method. Specifically, we used the first - order and second - order derivatives to test th e edge matching degree. In practical operations, we used the Laplacian of Gaussian op erator to detect the edges of grayscale images. After that, we used the mathematical m orphology calculation method to correct the edges of rectal tumors, and finally obtaine d the mask image of the rectal tumor area.

When solving the second problem, relying on the segmented images obtained in t he first problem, we extracted features from three dimensions: the color features, edge features, and texture features of the tumor. Specifically, we used methods such as histo gram statistics, calculating the value of Saturation, and analyzing the gray - scale step - change of the pixels around the image boundary to effectively extract the features of t he tumor area.

For the third problem, we chose the combination of the random forest algorithm a nd feature selection. By calculating the Gini coefficient, we clarified the precise effecti veness of each feature in judging the lymph node metastasis status. This method can ef fectively reduce the overfitting phenomenon and improve the prediction accuracy at th e same time. We constructed a classification model and comprehensively evaluated the model using the dice coefficient and F - Score evaluation indicators to ensure the relia bility and accuracy of the model.

Key words: Differential Mathematics Algorithms Laplacian Gauss Operator Edge Detection Random Forest

目录

一、挖掘目标

本次挖掘目标是通过直肠癌 CT 影像图，研究出一种准确判断出淋巴结是否转移的情况的方法。利用微分数学算法、直方图统计以及随机森林算法等，达到三个目标。

（1）对 CT 图像进行预处理，运用微分数学算法，Python 软件进行手动分割，找出直肠癌所在的位置，定位肿瘤区域。

（2）从肿瘤的颜色特征、边缘特征、纹理特征三个方面，提取直肠癌肿瘤 CT 影像中与目标结果紧密关联稳定的高维特征，提取得到直肠癌影像的特征。

（3） 通过验证直肠肿瘤区域影像特征和淋巴是否转移的关系，考虑病人的年龄以及实验室指标等等，并建立分类模型来评估模型的有效性。

二、分析方法与过程

2.1 问题一的分析方法与过程

2.1.1 对问题一进行系统性分析

在如下图 1 和图 2 的 CT 动脉期图像上，直肠肿瘤与周围组织的强度上存在差异，这就意味着它们的活性不一样，它们的蛋白表达缺失有明显差异。直肠癌的蛋白表达强度与阳性细胞的表达的数量都明显高于它们在正常组织中的表达。所以，我们利用阳性细胞的表达数量不同，我们可以得到直肠癌与正常直肠的边缘。

图 1　CT 动脉图像

图 2　CT 动脉图像

2.1.2 对 CT 图像做预处理

我们先对图像进行预处理。经过大量的实际调查发现，性别对淋巴结的转没有多大影响，所以当 CT 图像中除了性别这一区别以外没有其他区别的，我们只对一种展开研究，即只是研究性别。不考虑不具有直肠癌的图片。

2.1.3 基于数学算法的分割

用微分数学算法对 CT 图像中直肠癌的特征进行研究。人体直肠癌与其他正常部位的活性细胞数量有关，直肠癌处的细胞活性较弱，细胞数量较少，由此大致分割出直肠癌区域。

先对一张图片的直肠癌区域进行分割，其他图片用类似的方法也将直肠癌区域分割出来。

下面以 8 张图片为例。根据直肠癌活性细胞数量的不同，利用微分算法可以将其与正常区域分割。如某一个存在生命现象的细胞数量相邻的区域，其一阶导数是常数，那么二阶导数也是常数，则可以把这两个相邻区域合并，以此类推，将 CT 图像从小一小部分慢慢提取直肠癌区域，最终就可以将图像进行边缘分割。

根据边缘检测算法，从各个目标区域的特征，强度的差异，是分离直肠肿瘤区域的重要算法。只有将目标的边缘用算法提取出来，才能将肿瘤的区域和非肿瘤的区域分开；两个不相同边界灰度值的相邻区域灰度值是不连续的，通常可以利用微分的方法进行检测这种不连续，所以一阶导数和二阶导数常用来检测边界。

检测边缘首先利用边缘增强算子，为了突出 CT 图像中局部的边缘，然后那个点的边缘用某些特定符号去定义，提取边缘的点集，通过设值阈值的方法来进行。但是尽量避免由影像模糊不清的噪音因素，最终检验到的边界处有可能会有间断的情形出现。所以边缘检测有以下两个内容:

(1) 边缘点的集合用边缘算子来提取。

(2)从边缘点的集合中取出某些边缘点集，增强这些边界点，再将得到的边

缘点顺序连接成为一个轮廓。

利用拉普拉斯高斯算子（缩写log算子）来检验灰度图像的边缘。拉普拉斯—高斯算子，是一种二阶微分算子，用这个的算法的结果会在图像的边缘处产生一个陡峭的零交叉。高斯算子是一个线性的的算子，检测边缘点是通过寻找图像灰度值中二阶微分是 0 的点。

log是常用的边缘检测算子，它是各向同性的二阶导数，对连续函数$f(x,y)$，有

$$\nabla^2 f(x,y)=\frac{\partial^2 f}{\partial^2 x}+\frac{\partial^2 f}{\partial^2 y}$$

经边缘检测后的图像$g(x,y)$为

$$\varphi(x,y)=\ f(x,y)-\ r\nabla^2 f(x,y)$$

式中，扩散效应与系数r有关。检测出边缘图像$\varphi(x,y)$是图像$f\ (x,y)$经过拉普拉斯算法之后得到。特别需要注意的是，对系数r的选择太大，会使图像中的轮廓边缘相冲，太小就会使得到边缘图像不明显，所以我们一定要对系数r进行合理的选择。

对数字图像来讲，$f(x,y)$的二阶偏导数可以表示为

$$\begin{cases}\dfrac{\partial^2 f(x,y)}{\partial^2 x}=[f(x+1,y)-f(x,y)]-[f(x+1,y)-f(x-1,y)]\\ \qquad\qquad=f(x+1,y)f(x-1,y)-2f(x,y)\\ \dfrac{\partial^2 f(x,y)}{\partial^2 y}=f(x,y+1)f(x,y-1)-2f(x,y)\end{cases}$$

所以拉普拉斯高斯算子$\nabla^2 f(x,y)$为

$$\begin{aligned}\nabla^2 f(x,y)&=\frac{\partial^2 f}{\partial^2 x}+\frac{\partial^2 f}{\partial^2 y}\\ &=\ f(x+1,y)+f(x-1,y)+f(x,y-1)-4(x,y)\end{aligned}$$

得到的图像用拉普拉斯高斯算子来检测边缘值，由所得到点的灰度值减去这个点邻域的平均灰度值得到。

如果在图像中出现了一些亮点在某个较暗区域中，那么采用拉普拉斯算法就可以使这些亮点变得更亮，这是因为图像中的边缘就是那些灰度发生跳变的区域组成，所以利用拉普拉斯算子去边缘检测，有着很大的作用。

检测的微分算法具体过程如下流程图 3.

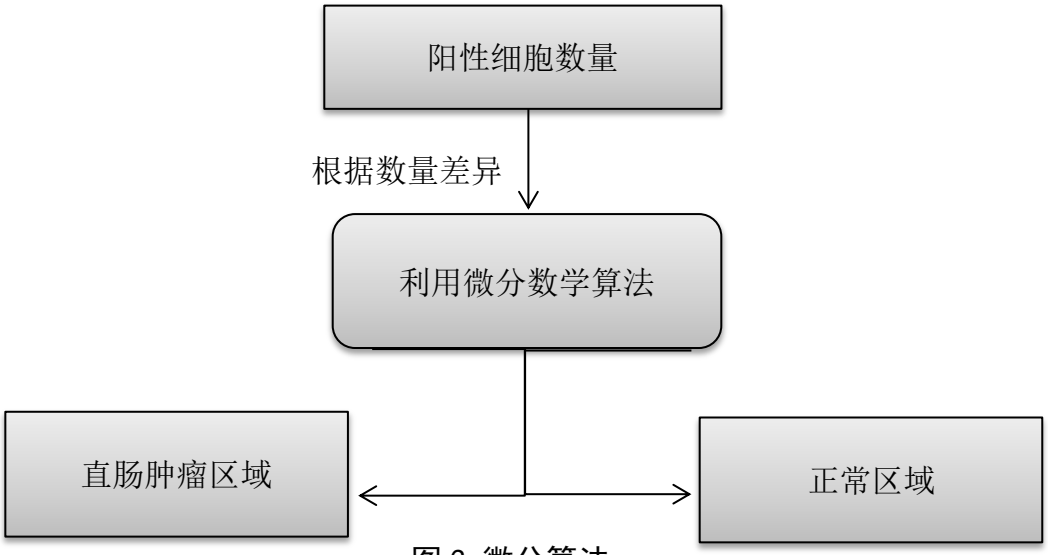
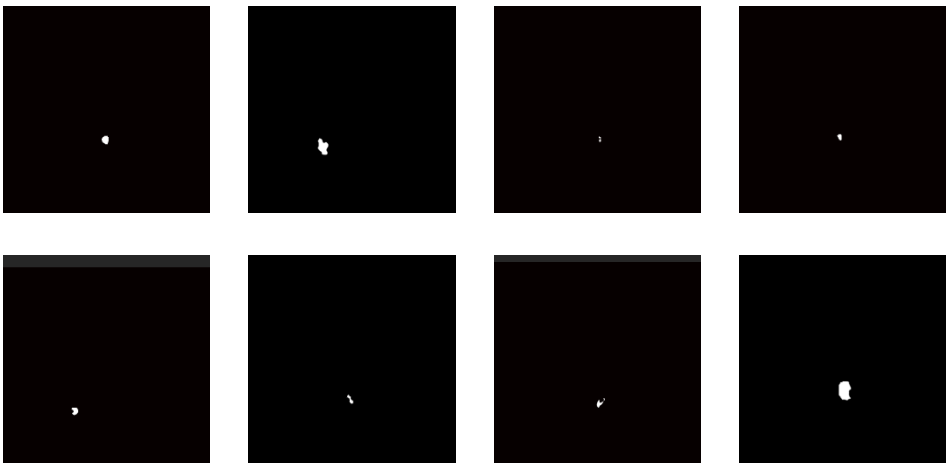


图 3 微分算法

2.1.4 掩模图像的提取

通过微分数学算法，运用 **Python 软件**手动分割，再用数学形态学计算方法修正直肠肿瘤边缘，最终得到的掩模图如下列图片。



2.2 问题二的分析方法与过程

2.2.1 对问题二进行系统性分析

问题一中已经将直肠癌与正常区域分割出来，并得到相应的影像。如果已经知道直肠癌肿瘤影像特征，在医学上就可以很好地指导临床工作。一般来说，癌的低分化组织更容易浸润地生长，浸润到周围的其他组织，因此存在淋巴结转移。下面，我们从影像的影像图的形状，纹理，边缘，肿瘤浸润深度、亮度一些方面对影像图进行研究，提取出直肠癌肿瘤区域的影像特征。

2.2.2 影像特征的提取

（1）纹理特征

利用直方图统计来提取直肠癌肿瘤区域影像的纹理特征。不同的直方图对应着纹理的不同特点，反之，特征相似的直方图会对应着相似的纹理。这就说明纹理与直方图存在着一些关系。因此，我们利用直方图获取统计特征作为图像的纹理特征。

从以下三个方面来统计纹理特征：

① 均值是度量纹理平均亮度。

$$m = \sum_{i=0}^{L-1} z_i p(z_i)$$

其中L是灰度级总计数，z_i代表第i个灰度级，$p(z_i)$是归一化直方图灰度级分布中灰度为z_i的概率。

② 标准方差是度量纹理平均对比度。

$$\sigma = \sqrt{\sum_{i=0}^{L-1} (z_i - m)^2 p(z_i)}$$

其中根号中的内容是均值的二阶矩μ_2。一般均值m的n阶矩表示为:

$$\mu_n(z) = \sum_{i=0}^{L-1} (z_i - m)^2 p(z_i)$$

③ 度量纹理亮度的相对平滑度是平滑度。

对于灰度级的区域一致，平滑度$R=1$；对于灰度级区域的值，有着较大差异的，$R=0$。

$$\mathrm{R}=\frac{1}{(1+\sigma^2)}$$

（2）颜色特征

直肠癌影像的颜色有差别，直肠癌的颜色比正常的部位更加深。对图像中的某一个像素，可以根据 Saturation 的值来判断函数阈值，并将它们划分到相应颜色的区间。采用下面的有效线性函数作为阈值：

$$\frac{1}{1+\alpha*Vaue},\ (1\leq\alpha\leq 4)$$

定义阈值：

$$\mathrm{T}=\frac{1}{1+\alpha*Vaue},(1\leq\alpha\leq 4)$$

对于影像中的任意像素 X(H,S,V)，判断它属于哪一个颜色的区间，就可以提取出颜色的特征。

（3）边缘特征

根据部分区域边缘的直线特征，能找到一些邻域内直线部分的高精度位置，再根据内部边缘的直线特性，用线段的中点来拟合整个直线边界，可以得到亚像素精度的影像边界。拟合过程中，根据直线段的转角变化排除掉噪声点，提高算法在图中定位的精度。利用边缘检测，分离出各个区域的边界，图像的边界是由于周围像素灰度阶跃变化或屋顶变化，他是图像分割的重要根据，因为边界直方图有尺度不变的性质，所以能够较好的描述出直肠肿瘤的大致形状。

2.3 问题三的分析方法与过程

2.3.1 对问题三进行系统性分析

病人的年龄，实验室指标等可能与淋巴结有关，分析不同年龄，不同实验室指标的病人的直肠癌影像特征，检验这些特征是否能说明与淋巴结转移有关。

2.3.2 建立分类模型

为了防止单棵的决策树易发生拟合的现象，提高它的预测精度，第三问我

们采用随机森林算法和特征选择，提高各个特征分别对淋巴结是否转移情况的精确性，通过构遗多个弱分类器合成一个强分类器，在有效减少拟合现象的同时，提高它的预测精度。

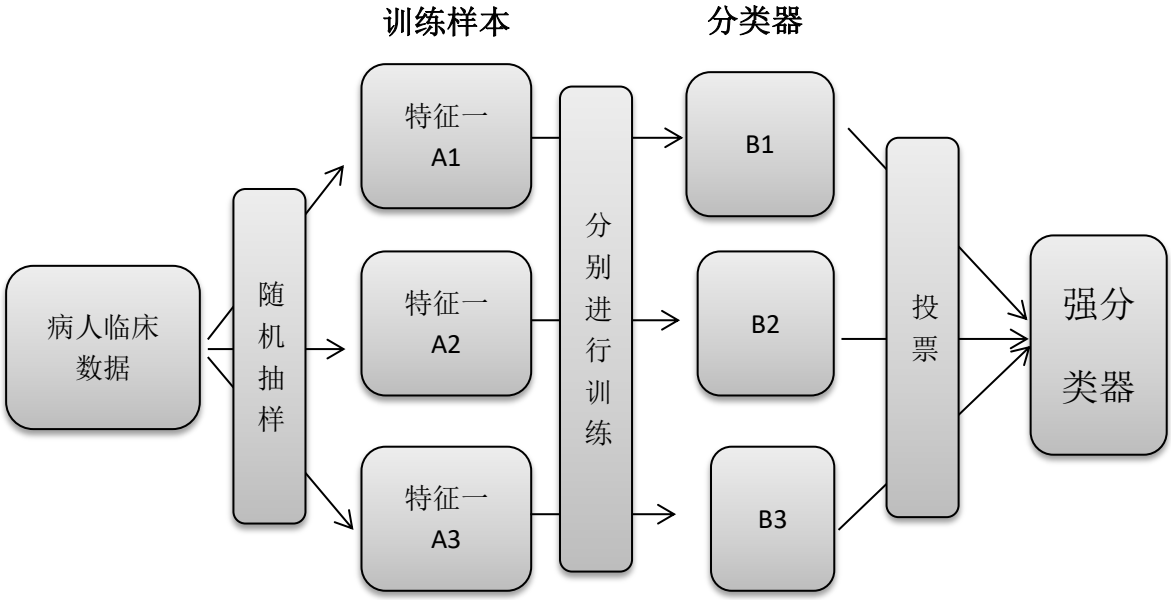
在影响直肠癌淋巴结转移特征中，找出权重较大的特征因素。首先构建预测模型，具体方法是：设置K个弱分类器，使用$Gini$系数计算，其中系数越大，特征的不确定性越大，反之，不确定性就越小，分割得更干净，将这些相似的样本放在同一个弱分类器中，采用 K-means 聚类算法进行训练，并使用均值组合方式，进行模型训练。

$$\mathrm{Gini}(p)=\sum_{k=1}^{K}p_k(1-p_k)=1-\sum_{k=1}^{K}{p_k}^2$$

在训练阶段，主要是根据计算特征的样本，划分好K个弱分类样本后，再进行随机森林训练。随机森林的算法是多棵决策树对样本进行训练之后，再预测出结果的一种分类器。分类决策的公式为:

$$\mathrm{H}(X)=\arg max\sum_{i=1}^{k}I\left(h_i(x)=Y\right)$$

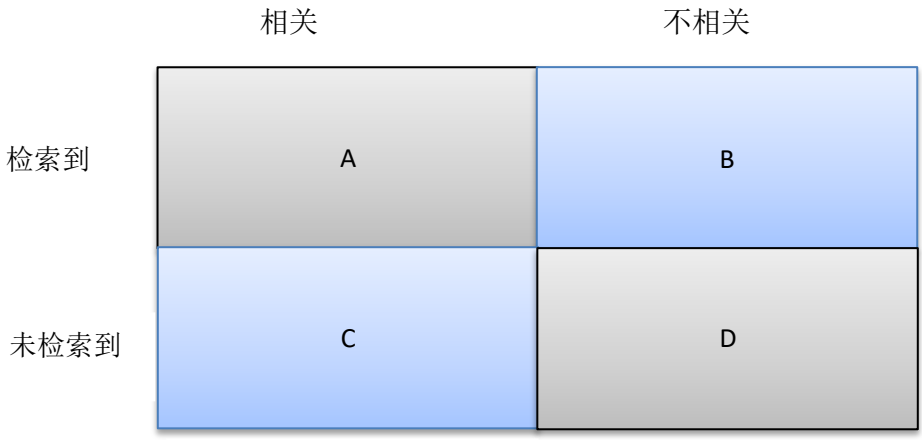
其中，$H(x)$表示随机森林的分类结果，$h_i(x)$表示单个的分类结果，Y表示分类的目标。该式为随机森林的分类问题，最终结果取各个决策树结果占多数的。而对于随机森林的回归问题，则可以选择每个决策树的结果期望值作为最后的结果。



2.3.3 评估模型

我们在做医学图像分割，并对分割结果进行评价时，常用$Dice$系数来评价，在 python 中用$SimpleITK$函数来进行并实现图像分割。 $Dice$系数是一种集合相似度函数，我们用它来计算直肠癌和其他正常区域的样本相似度。Dice $= \frac{2|A \cap B|}{|A|+|B|}$. A 代表题目中所给勾画出的直肠肿瘤区域，B 是我们分割得到的直肠肿瘤区域。

我们通过分类模型的精确率($Precision$)和召回率($Recall$)来进行评估指标。对于精确率和召回率，虽然从计算公式来看，并没有什么必然的相关性关系，但是，在大规模数据集合中，这 2 个指标往往是有某种特定的联系。所以，综合权衡这两个指标，这是我们很多时候需要考虑的，综合$Precision$和$Recall$的调和值，然后找出了一个新的评估指标$F-score$。



P(准确率）$= \frac{A}{A+B}$

R（召回率）$= \frac{A}{A+C}$

三、结论

在中国，直肠癌的发病率和死亡率均呈逐渐上升趋势。直肠癌极易向肠外组织浸润，进而引发淋巴结转移。因此，精准判断淋巴结是否转移以及转移淋巴结的具体区域，对制定患者的治疗方案、保障手术成功起着关键作用。为此我们采用了一些有效的微分数学算法和边缘检测等手段。依据边界细胞活性和灰度值的差异与相似性，先将直肠癌肿瘤区域和其他正常区域分割开来，借助高斯算子以及其他用于修整区域轮廓的数学方法，获取肿瘤区域的掩模图。接下来，研究直肠肿瘤区域的影像，提取形状、纹理、颜色、边缘等影像特征。最后，通过计算基尼系数，分析这些影像特征与淋巴结转移之间的关系，并建立随机森林分类模型，以此检验结果的有效性和准确性。在患者的临床数据中，阳性代表患病，阴性代表未患病。通过上述算法，能够有效判断出淋巴结的转移情况。

准确判断直肠癌淋巴结是否转移，对患者的治疗方案和手术成功意义重大这能为患者争取更多生存机会，助力更多人恢复健康，在医学影像研究领域也有着重要价值。

四、参考文献

[1] 王飞,董国礼. 结直肠癌的影像学研究进展[J]. 川北医学院报,2014,29(1):107-112.doi:10.3969/j.issn.1005-3697.2014.01.25.

[2] [17]肖辉, 郝元涛, 徐晓, 等.基于随机森林算法和Logistic回归模型的糖尿病风险因素研究[J].中国数字医学, 2018, 13 (1) :33-35, 40.DOI:10.3969/j.

[3] 蒯玉娴(综述),左长京(审校). 直肠癌淋巴结转移磁共振诊断研究进展[J].功能与分子医学影像学杂志(电子版), 2016, (2):949-954.doi:10.3969/j.issn.2095-2252.2016.02.015.

[4] 赵泽亮, 杨新辉, 孙振强.直肠癌淋巴结转移与临床病理特征分析.中国实用外科杂志, 2011, 31(8):696-699.

[5] 李德福, 张敏.直肠癌淋巴结转移与临床病理特征分析(附79例报道).中国普外基础与临床杂志, 2008, 15(9):652-655.

[6] 汪晓东, 冯硕, 游小林.等.结直肠肿瘤多学科协作诊治模式下的随访体系建设.中国普外基础与临床杂志, 2007, 14(6): 709.

[7] 屠世良, 叶再元, 邓高里, 等.结直肠癌淋巴结转移的规律及其影响因素.中华胃肠外科杂志, 2007, 10(3):257-260.

[8] 郭骏，潘申，胡小建基于灰度形态学的烟叶图像边缘检测[J].计算机工程，2007, 33 (21) :163-165.