基于 CT 影像的结直肠癌与淋巴结转移诊断

摘要：结直肠癌是胃肠道中常见的一种恶性肿瘤，近年来呈现高发病率，高病死率的特点。因此对结直肠癌的影像学研究具有重要的意义。其中一个重要问题是如何确定肿瘤区域并进行边界分割，过去常用的手段是让影像科医生在 CT 图像的基础上利用软件手动分割，而我们希望设计算法让计算机能够通过 CT 样本的训练和自主学习，对输入的 CT 图像识别肿瘤位置和边界并进行分割。另一个重要问题是如何确定病人淋巴结转移的情况。目前没有明确的手段能够对此进行准确的判断，而我们希望利用病人的 CT 资料，通过数据挖掘技术，提取高通量特征并用于淋巴结转移的预测。本次任务分成四大板块：图像预处理，肿瘤分割，肿瘤特征提取，淋巴结转移预测。

对于图像预处理，我们希望缩小图像的研究区域，凸显图像细节，并减少 CT 成像过程中产生的噪声。为此我们先将 CT 图像每个像素的灰度值转化为 CT 值，而后运用 Windowlevel 算法过滤 CT 值很小和 CT 值很大的区域，并将感兴趣区域的 CT 值均匀拉伸到 0-255 以凸显细节，最后再用中值滤波器对图像进行处理。结果表明一系列处理后的图像具有更高的对比度和细节质量。

对于肿瘤分割，我们希望构建一个具有较高精度和具有较强学习能力的卷积神经网络模型。我们采取两套方案：一套是构建简单的 LeNet-CRF 模型，通过经典的 LeNet-5 网络架构赋予模型学习能力，通过条件随机场对 LeNet 得到的肿瘤粗分割进行更细微的分割。另一套是复杂的 MIGAN 模型，总体框架采用对抗生成网络模型，设置产生肿瘤分割图像的生成器和分辨分割结果真假概率的判断器。生成器使用三维三尺度卷积神经网络模型，根据统计到的肿瘤尺寸信息决定各个尺度的大小。大尺度采用 Inception v1 结构；中，小尺度采用简单的 LeNet-5 结构。判断器采用简化的 LeNet-5 结构。在两个模型的训练过程中均对训练样本图片分割成小尺寸并根据像素中心点的标签设置为正负样本，正负样本交替均衡地输入到训练模型中。最终 LeNet-CRF 模型的总体 Dice 系数达到 73%，MIGAN 模型的总体 Dice 系数达到 75%。两个模型都达到较高的精度，但对肿瘤的细节分割还有待加强。

对于肿瘤特征提取，对于一个病人的两套 CT 图像，分别提取肿瘤截面积最大的一张作为代表用以提取二维特征，分别按扫描顺序集合所有包含肿瘤的切片变成三维图像块用以提取三维特征。本次提取的二，三维特征包括：一阶统计量特征，基于形状和大小的特征，纹理特征，小波特征。先把原来的图像矩阵按照“0-255”，“最小值-最大值”的范围转变成两个跨越 16 个灰阶的灰阶矩阵，用这两种矩阵进一步提取全部四种特征。而后提取小波系数，并用小波系数提取纹理特征。结果我们提取到了包含 Arterial 期和 Venous 期特征，包含二维和三维特征的总计 2922 个特征。

对于淋巴结转移判断，我们尝试了随机森林回归，随机森林分类和二元 Logistic 回归三种方案。对于随机森林回归和分类模型，我们先进行特征降维，降维方案采用三种进行比对：无降维，用 mRMR 降维到 600 个特征，用 relief 降维到 600 个特征。而后分别用三种方法降维后的特征进行预测。对二元 Logistic 回归，先比对双样本-二次交集法，单样本一次交集法，relief-mRMR 法，mRMR-relief 法的降维效果，而后采用前三种降维得到的特征进行预测。最后得到采用双样本-二次交集法和单样本-一次交集法的 Logistic 模型在验证集上表现最好，均有 88.9%的准确率，其次是采用 relief 降维算法的 RF 回归和 RF 分类模型，有 85.2%的准确率。在 F-Score 指标方面采用双样本-二次交集法的 Logistic 模型表现最好，F-Score 为 83.3%。最后我们还适当放大和缩小肿瘤特征提取的窗口用以检验双样本-二次交集法的 Logistic 模型的灵敏度，发现放大或缩小窗口后的模型 F-Score 数值均减小，且窗口缩小的情况更敏感。

关键词：卷积神经网络 条件随机场 特征提取 特征降维 随机森林 Logistic

Colorectal cancer and Lymph gland metastasis diagnosis based on CT

Abstract： Colorectal cancer is a common malignant tumor. In recent years it appeared with high morbidity and high lethality rate, thus radiomics study on colorectal cancer is important. One vital problem is how to determine the region of tumor and segment it out. The means usually used in the past is to let radiomics doctors use software to segment the tumor on CT by hand. But we expect to design a model by which the computer could be trained with sample CT, learn the features of tumor, recognize the region of tumor and finally segment it. Another big problem is how to determine whether the lymph gland has transferred or not. So far there is no clear ways to judge accurately. We desire to make use of the CT data from the patients. By using the technique of data mining, we expect to extract a large quantity of features and use them to make prediction on lymph gland metastasis. The whole work is divided into four parts: Image Preprocessing, Tumor Segmentation, Tumor Feature Extraction and Lymph Gland Metastasis Prediction.

The first part is Image Preprocessing. We expect to narrow the Region Of Interest(ROI), enhance picture details and reduce the impact of image noise. To achieve it we firstly transfer every pixel's gray value to CT value. Then use Windowlevel algorithm to filter regions with too low CT value or too high CT value. Next we uniformly stretch the pixel's CT value in ROI from 0 to 255 to enhance the detail. Finally we use median filter to process the graph. Outcome shows that after a series of processing, the image has higher contrast ratio and high detail quality.

The second part is Tumor Segmentation. We expect to build a Convolution Neural Network(CNN) with high accuracy and strong learning ability. We adopt two methods: one is to build a simple LeNet-CRF model and the other is to build a complex MIGAN model. For LeNet-CRF model, we use classic LeNet-5 structure to give the model a learning ability and use Conditional Random Field(CRF) to do a finer segmentation based on the raw segmentation from LeNet. For MIGAN, the whole structure we use is Generative Adversarial Network(GAN). We respectively set up a generator generating tumor segmentation and a judger judging the possibility of the segmentation. The generator use three dimension-three size CNN. We determine each size according to the tumor size we analyze from the sample. Network in big size use Inception v1 structure, while network in median and small size use LeNet-5 structure. The judger use simplified LeNet-5 structure. In the process of training both the model, sample images are segmented into small images and be classified to positive input and negative input based on the center pixel label(either belongs to the tumor region or not). Positive and negative input are in turn imported into the model. Finally the Dice value of LeNet-CRF reaches 73% and the dice value of MIGAN reaches 75%, 2% more than the former. Both LeNet-CRF and MIGAN have reach a relatively high accuracy but still need to be strengthened in the detail segmentation.

The third part is to extract features from CT. For each set of CT(either Arterial or Venous) in a particular patient, respectively use the CT with the biggest cross section area as a representative to extract two-dimension features. Sort all the CTs containing tumor following the sequence of scan and use them as a whole to extract three-dimension features. The features extracted in this task consist of First Order Statistics, Features based on Shape

and Size, Texture Features and Wavelet Features. First we transfer the original CT value matrix to gray scale matrix in the field of “0-255” and “minimum CT value-maximum CT value”. Then use these two kinds of matrixes to extract all four kinds of features. Then use the wavelet features extracted before to further extract the texture features. Finally we have extracted 2922 features across from Arterial to Venous and from two-dimension to three-dimension.

The last part is to determine lymph gland metastasis. We tried with three models: Random Forest(RF) Regression, Random Forest Classification and Binary Logistic Regression. For RF Regression and RF Classification, we firstly reduce the dimension of features. We adopt three methods to do it: no reduction, reduction to 600 features using Minimum Correlation Maximum Redundancy(mRMR) and reduction to 600 features using relief algorithm. Then respectively use the features filtered by means above to build the prediction model. For Binary Logistic Regression, firstly compare the features filtered by methods of double samples-twice intersection, single sample-once intersection, relief-mRMR and mRMR-relief. As a result we respectively use features filtered by the former three kinds of methods. Results show that Binary Logistic Model using double samples-twice intersection and single sample-once intersection have the best performance on the testing samples, both reach 88.9% accuracy. The second is the RF Regression and RF Classification using relief, both reach 85.2% of accuracy on the testing samples. In F-Score Performance, Binary Logistic model using double samples-twice intersection does the best, the F-Score of which is 83.3%. Lastly we also test the sensitivity of Binary Logistic model using double samples-twice intersection by adjusting the size and shape of feature extracting window. We find that both enlarging the window and narrowing the window cause the F-Score to decrease and it is more sensitive to the case of narrowing the window.

Key words: Convolution Neural Network Conditional Random Field Feature Extraction Feature reduction Random Forest Logistic

目 录

1. 问题背景与任务

- 问题背景

结直肠肿瘤是胃肠道中常见的一种恶性肿瘤，其病死率在消化系统恶性肿瘤中仅次于胃癌，食道癌和肝癌，对患者的生命健康产生严重的威胁。而其发病人数排名第 5 位（在男性中排名第 5 位，在女性中排名第 4 位），同时每年致死人数排名第 5 位（在男性和女性中均排名第 5 位），且发病率和死亡率每年都在逐步上升[1]。鉴于上升的发病率和较高的病死率，开展结直肠癌诊疗及预后的研究显得十分重要。诊疗的一个重要部分是对肿瘤区域的确定和分割，这关系到对肿瘤危险程度的评估以及对手术方案的采取。目前肿瘤分割常用的是由影像科医生用电脑软件进行手动分割，更精确的分割是让不同的医生进行分割，在医学影像上人为添加噪声，利用不同呼吸周期多次分割等[1]。利用数据挖掘技术研究自动分割，实现减少人力资源的调用的同时更准确高效的分割肿瘤，也就成为一个值得研究的课题。而病人预后方面涉及到一个重大问题是对病人淋巴结转移情况的判断。若直肠癌患者在术前能确定没有发生淋巴结转移，则患者无需在术中进行淋巴结清扫，但是目前淋巴结转移的判断没有一个明确的手段。运用数据挖掘技术提取肿瘤影像特征，结合其他可能的相关指标来预测淋巴结转移与否，可以帮助减少患者不必要的痛苦。

- 挖掘任务

①搭建模型，根据已有的肿瘤原始 CT 资料与掩模图像设计自动化算法，对肿瘤区域进行自动分割。该算法需要样本进行训练和自主学习。

②设计指标与相应的算法，从提供的病人 Arterial 期和 Venous 期的 CT 资料挖掘图像矩阵信息，提取包括一阶统计量，基于形状和大小的特征，纹理特征，小波特征等高通量的二维和三维特征。

③利用已提取的特征建立模型预测病人淋巴结转移情况，并对模型预测的表现能力做出评价。

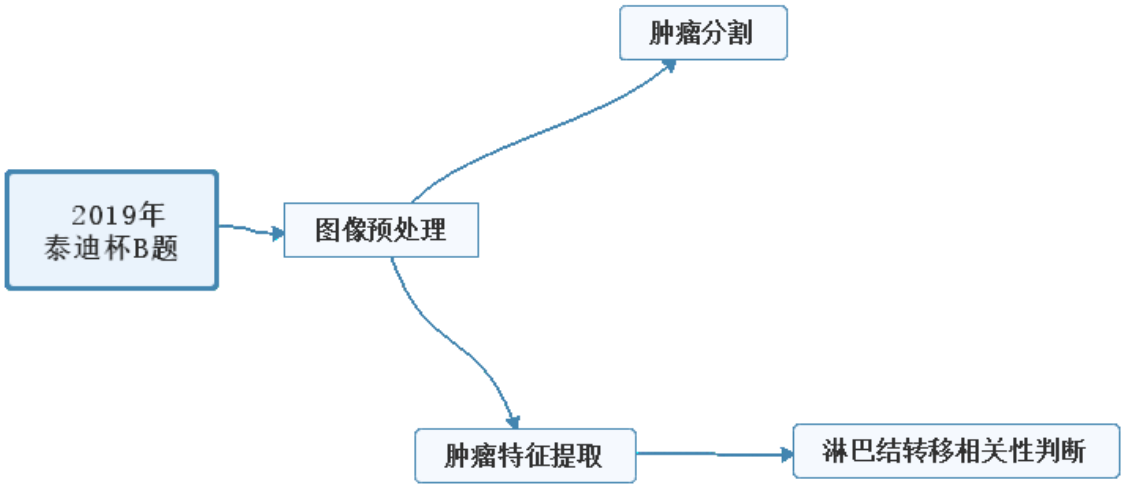


图 1-1 整个工作流程的大概框架

2. 图像的预处理

2.1 任务分析

这次提供的影像资料为病人腹部横断位 Arterial 期和 Venous 期的 CT 拍片这些原始的 CT 图像质量比较粗糙，具体表现在图像中各部分的灰度对比不强烈，导致整张 CT 图像的视觉效果为灰色，模糊；有些图片存在一些无用的纯黑色区域，只有中心圆形里面的部分才是真正的 CT 主成像区。

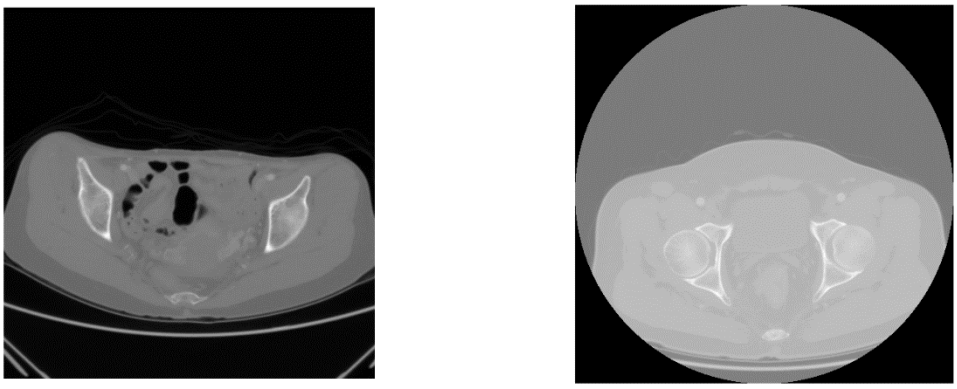


图 2.1-1 原始 CT 影像的视觉效果，
左边为病人 1003 的一张 Arterial 期拍片，右边为病人 1009 的一张 Arterial 期拍片

为了摒弃图像的不相关区域，对有价值的区域进行针对性研究，同时增强对比度以凸显成像细节，并且去除可能存在的噪声干扰，我们拟采用以下步骤：

1）将原 CT 图像矩阵的灰度值转化为 CT 值。CT 值是测定人体某一局部组织或器官密度大小的一种计量单位，通常称为亨氏单位（Hounsfield unit，Hu），空气为-1000，致密骨为+1000。而灰度值与 CT 值之间的变化是灰度线性变化的一种，转化公式为：

$$\mathrm{Hu} = \mathrm{Gray} \times \mathrm{RescaleSlope} + \mathrm{RescaleIntercept}，$$

其中 Hu 是转化后的 CT 值，Gray 是转化前的灰度值，RescaleSlope 是线性转化的斜率，RescaleIntercept 是线性转化的截距。RescaleSlope 和 RescaleIntercept 均用 MATLAB R2017a 自带的 dicominfo( )函数从病人 dicom 文件读取。

2）对转化后的图像矩阵采用影像学 Windowlevel 调窗方法处理。CT 能识别人体内 2000 个不同灰阶密度差别，而人的裸眼只能分辨 16 个灰阶度，每个灰阶度跨度 125（2000/16）个灰阶。而人体软组织 CT 值大多在 20-50Hu 之间变化，人眼无法识别[5]。为此我们需要进行分段观察，而 Windowlevel 方法帮助我们确定观察 CT 值的中心（窗位）和观察 CT 值范围（窗宽），本质上是一种灰度分段线性变换。变换公式如下：

$$\begin{cases} Min = Center - (Width/2) \\ Max = Center + (Width/2) \end{cases}，\quad T = \begin{cases} 0, & Hu < Min \\ 255 \times \dfrac{Hu - Min}{Width}, & Min \le Hu \le Max \\ 255, & Hu > Max \end{cases}，$$

其中 Min，Max 分别是窗宽下限和上限，$Center$，$Width$ 分别是窗中心和窗宽，Hu 是图像矩阵每个像素点的 CT 值。$Center$，$Width$ 均用 MATLAB R2017a 自带的 dicominfo( )函数从病人 dicom 文件读取。

3) 利用 3×3 中值滤波器对调窗后的 CT 片进行噪声过滤。CT 图像在采集过程中容易受到噪声干扰，信噪比比较低，使得肿瘤边界模糊。CT 图像是否含噪声受到多方面的影响。例如 CT 成像医疗设备使用时间过长可能会使 CT 图像含有噪声，质量下降；受检者体型越大，CT 成像噪声也会越大。为了减缓噪声对后续的肿瘤分割，肿瘤特征提取和淋巴结转移判断的干扰，同时避免滤波程度过大使得原图像失真，我们采用 3×3 的小窗口，滑动步长设置为 1，对原图像左右上下各进行一层 paddling 补零使得滤波后的图像保持 512×512 的尺寸。

2.2 结果

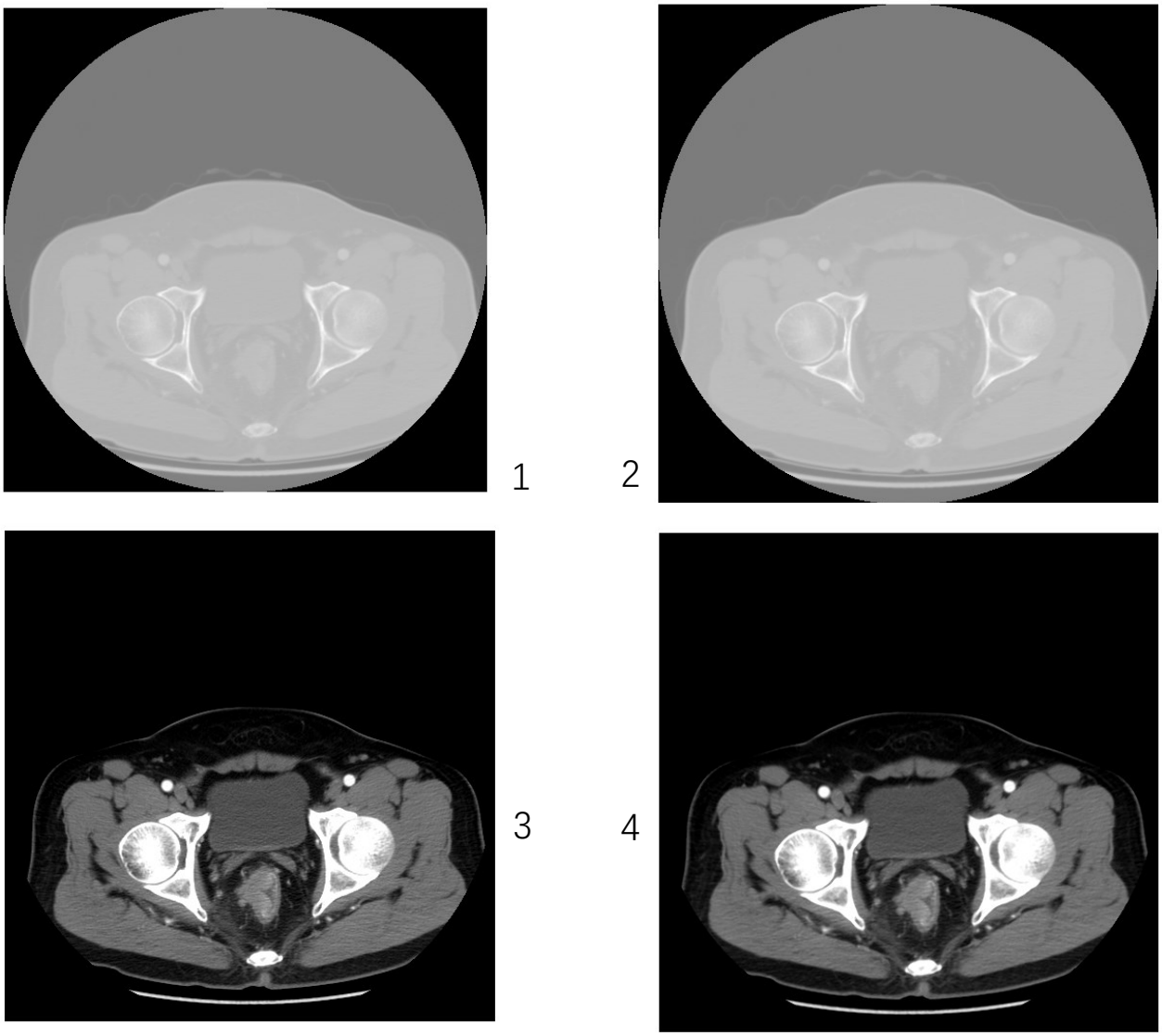


图 2.2-1 病人 1009 Arterial 图像预处理过程的效果展示，按照处理顺序，
左上为原图像，右上为灰度值转化为 CT 值（单位：Hu），左下为 Windowlevel 处理，右下为滤波处理。

我们可以明显地察觉到经过 Windowlevel 处理后的图像（3）质量提升了很多，组织纹理和细节更加清晰。滤波处理后的图像（4）较（3）视觉上几乎没有大的改变。

我们再比对其他窗口下的中值滤波操作。

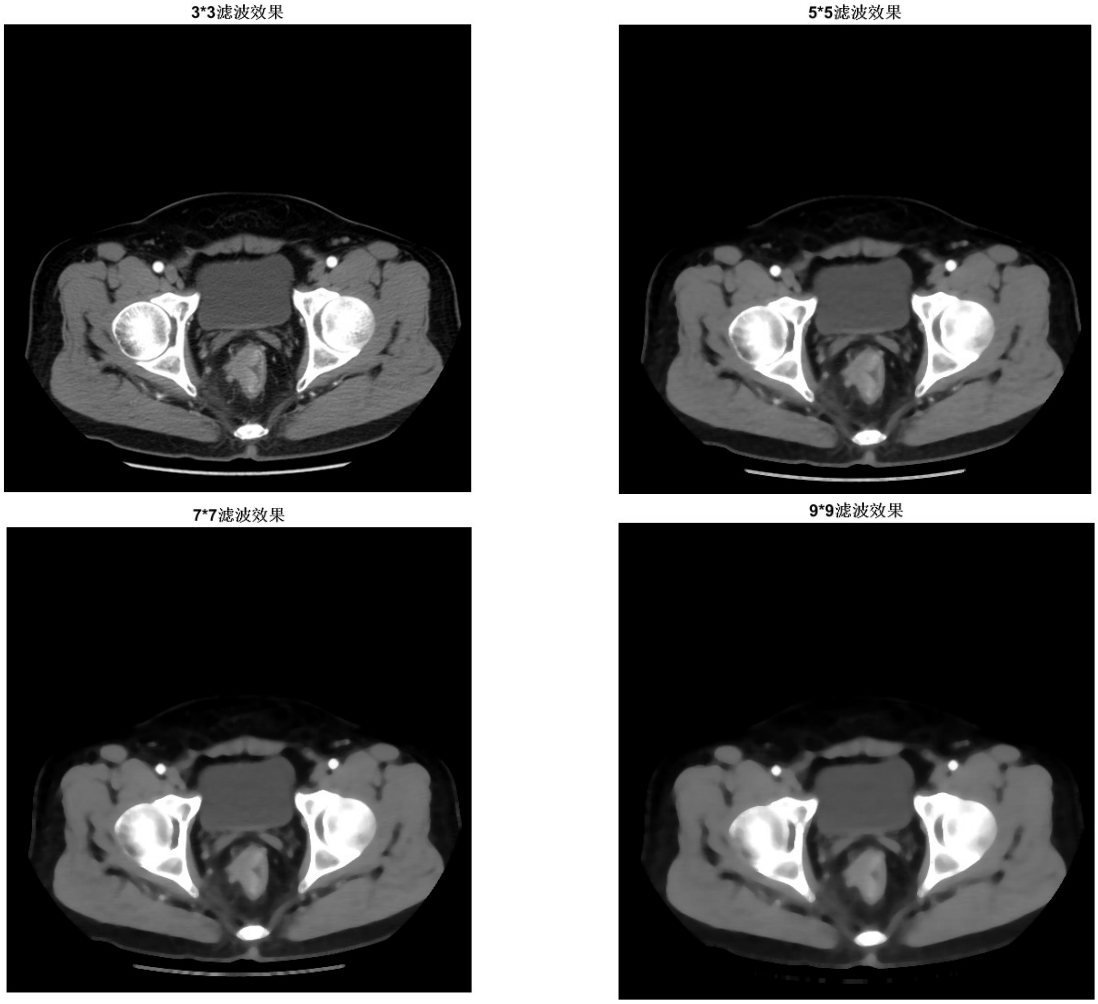


图 2.2-2 不同窗口下的中值滤波器的效果

我们观察到随着窗口的变大，滤波后的图像反而变得越来越模糊，选用9×9窗口时甚至造成严重的失真。因此3×3窗口在本次任务中很适合用来进行滤波。

3. 基于 CT 影像的肿瘤区域分割

3.1 任务的分析

本任务的目的是设计出模型来实现肿瘤边界分割的自动化。已出现并被应用于解决此类问题的算法大体有两种：一种是非机器学习算法，主要代表有基于边缘检测算子的边界分割法，应用 Hough 变换的直线提取法，基于直方图阈值，自动阈值，分水岭或迭代的阈值分割法，包含 K 均值，模糊 C 均值（Fuzzy C-Means），EM（Expectation-Maximization）和分层聚类的聚类算法等。这类算法的一个特点是不需要训练样本，但是需要人为地设定一些分割条件。另一大类是机器学习算法，其中又可细分为无监督机器学习和监督（半监督）机器学习。无监督机器学习的主要代表有区域生长法。这类算法的特点是需要训练样本，但是由于其无监督的性质导致训练后的模型在实际操作中分割精度不太高。监督的机器学习有

传统的极端随机树（ERT），支持向量机（SVM）和从上世纪 90 年代新兴的卷积神经网络（CNN）模型。CNN 与最开始的 BP 神经网络是一脉相承的，运用仿生学通过模拟人众多神经元间刺激的传导来实现一定的学习能力，而 CNN 又新颖在用卷积核实现图像局部感知，参数共享，多个卷积核从多个维度分析特征等优点，因此本任务我们拟搭建 CNN 框架实现对肿瘤边界的分割。

3.2 CNN 的选择和搭建

CNN 搭建的方案有两套。一套是框架比较简单的 LeNet-CRF 模型，另一套是框架比较复杂的 MIGAN（Multi-Scale Inception Generative Adversarial Network）模型。

3.2.1 LeNet-CRF 模型

1998 年 LeCun 等撰写了 CNN 的开山之作。之后 CNN 就被不断发展和开发出更复杂的框架。师冬丽等曾经应用经典的 LeNet-5 网络结构完成对 MRI 成像的脑肿瘤分割[2]，受此启发和鼓励，我们决定也采用简单的 LeNet-5 框架进行肿瘤粗分割。之后再利用 LeNet 输出的初始分割概率输入到条件随机场（Conditional Random Fields，CRF）进行肿瘤强化分割。

- LeNet-5 部分。本次 LeNet-5 的框架参考师冬丽等用到的框架，结构如下：

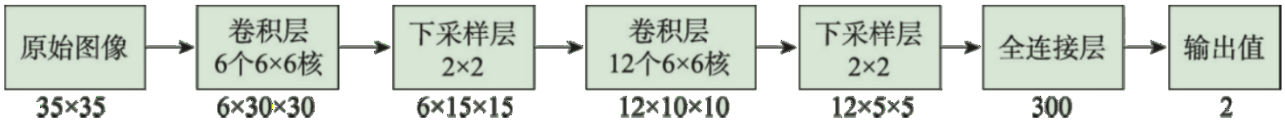


图 3.2.1-1　LeNet-5 的网络结构

原始图像：用随机数抽取将原始的 107 个病人样本按照 2:1 的比例分配为训练集（72 个）和验证集（36 个）。训练集中每个病人每幅 Arterial 期和 Venous 期的 CT 图像以像素点为中心切分为若干个尺寸为 35×35 的小图像，像素中心点属于肿瘤区域对应的小图像为正样本，像素中心点不属于肿瘤区域对应的小图像为负样本。将正，负样本交替作为原始图像输入以保持正，负训练样本的均衡。

卷积层：第一，二层卷积层用到的卷积核滑动步长均设置为 1。卷积层的形式为：

$$x_j^l = f\left(\sum_{i\in M_j} x_i^{l-1}\times k_{ij}^l + b_i^l\right) ,$$

其中 M_j 为神经元 j 对应的局部感受野，k_{ij}^l 是第 l 层的神经元 i 的第 j 个输入对应的权值；b_i^l 为 l 层的第 i 个偏置量，x_i^{l-1} 为 $l-1$ 层神经元 i 的输出，x_j^l 为 l 层神经元 j 的输出。

下采样层：采用 Max-Pooling 法。为了保持输入尺寸的不变适当进行 Paddling 处理。该层形式为：

$$x_j^l = f\left(\beta^l down\left(x_i^{l-1}\right) + b^l\right) ,$$

其中 $down(\ )$ 为下采样，β^l, b^l 分别为可训练参数和可训练偏置。

全连接层：选用 sigmoid 函数。

输出层：输出像素点属于肿瘤区域的概率和不属于肿瘤区域的概率。

- CRF 部分。CRF 的设置参考邢波涛等采用的全连接 CRF 模型[3]，滤波器采用双边高斯滤波器。像素归属为所属标签的能量函数为：

$$E(x)=\sum_i \varphi_u(x_i)+\sum_{i\neq j}\varphi_p(x_i,x_j)\ ,$$

其中 $E(x)$ 为像素归属标签的总能量，$\varphi_u(x_i)$ 为一元能量势函数，$\varphi_p(x_i,x_j)$ 为点对能量势函数。

点对能量势函数 $\varphi_p(x_i,x_j)$ 公式为：

$$\varphi_p(x_i,x_j)=\mu(x_i,x_j)\sum_{m=1}^{M} w^{(m)}k_G^{(m)}(f_i,f_j)\ ,$$

其中 $\mu(x_i,x_j)$ 是标签兼容性矩阵，$k_G^{(m)}(f_i,f_j)$ 是高斯滤波器核，f_i ，f_j 为滤波器特征向量，m 为滤波器数量，$w^{(m)}$ 为滤波器权重。

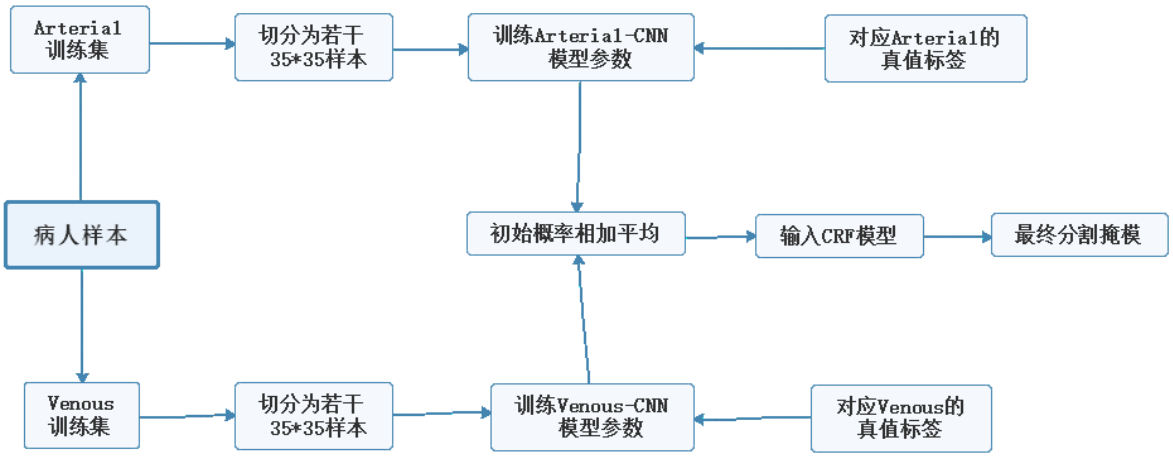


图 3.2.1-2 LeNet-CRF 工作流程

3.2.2 MIGAN 模型

这个模型是我们追求更高精度分割的试验模型。1）3.2.1 中的 LeNet 只是二维平面上的卷积，我们此次想进行三维空间上的卷积。2）3.2.1 中的 LeNet 只基于一个 35×35 的尺度进行，而我们在这一部分设想用多尺度（Multi-Scale）来搭建神经网络。李健等在 MRI 脑肿瘤分割中运用多尺度神经网络在训练层和测试层的分割精度分别达到 95.06%和 83.12%的优良指标[4]。3）Google 团队中 Szegedy 等人在 2014 年提出的 Inception v1 结构能大大减少参数量且提供不同视野的卷积核[7]，我们希望把它融入到神经网络的结构中。4）Luc 等人首次将新兴的对抗生成网络（GAN）使用在图像分割邻域[8]，我们希望将其作为 MIGAN 的总体框架。GAN 是一种特殊的 CNN，设置了分割肿瘤边界的生成器和判断肿瘤掩模真假的判断器，生成器和判断器均采用 CNN 网络结构。通过生成肿瘤掩模的生成器与判断肿瘤掩模真假的判断器的博弈来提升判断器的判断能力，从而更好地分割肿瘤。但原文作者也提到 GAN 网络难训练[8]，难收敛的问题，因此我们还是采用简单的网络架构。

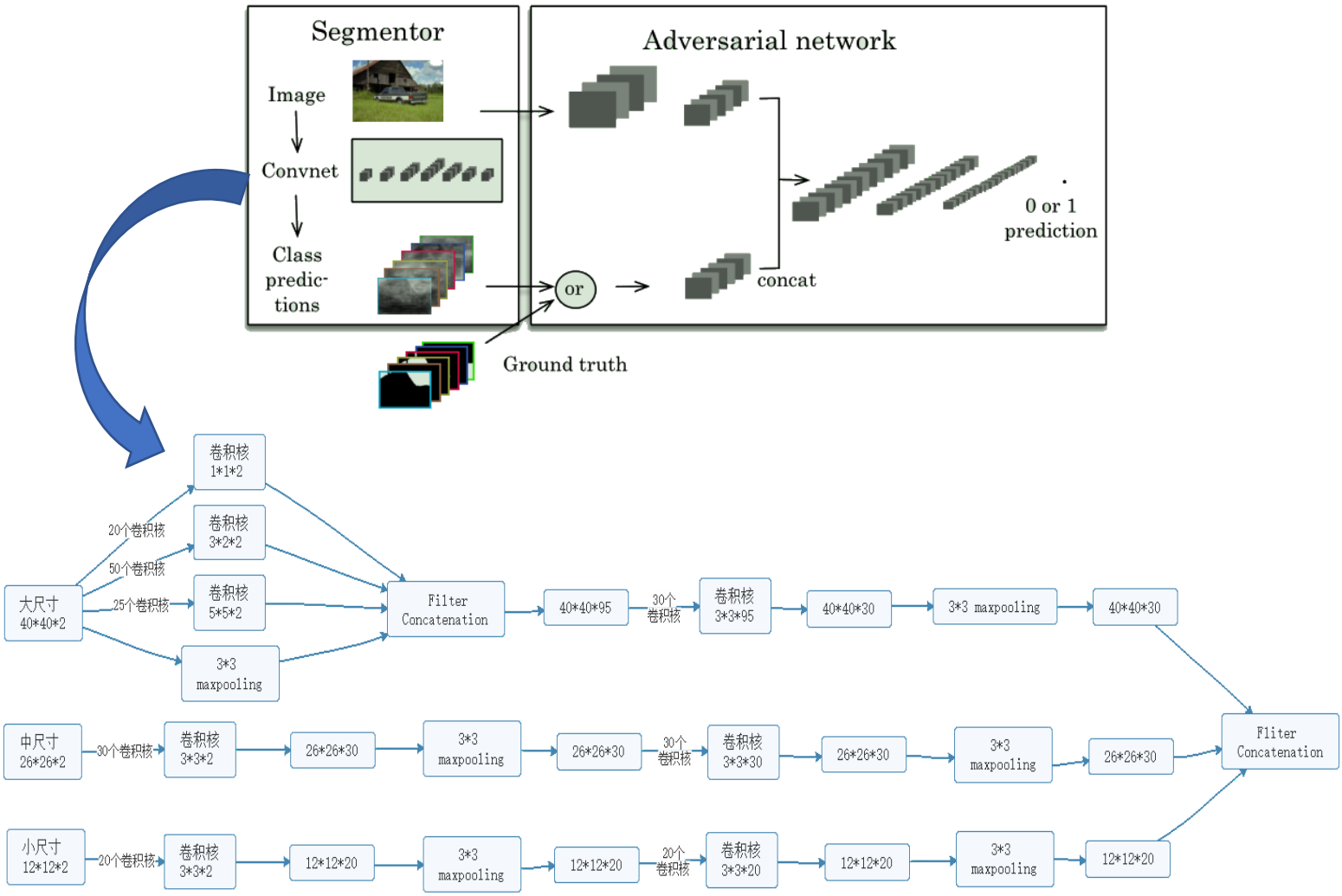


图 3.2.2-1 MIGAN 网络架构图

- 生成器部分采用三维多尺度卷积。

输入图像：先把每个病人 Arterial 期和 Venous 期匹配的 CT 片组合成三维的 $512\times512\times2$ 切片，去掉无法匹配或残缺的 CT 片。用随机数抽取将原始的 107 个病人样本按照 2:1 的比例分配为训练集（72 个）和验证集（36 个）。然后将三维切片以像素为中心点分割成 $40\times40\times2$ 的大尺寸图像，$26\times26\times2$ 的中尺寸图像和 $12\times12\times2$ 的小尺寸图像。大尺寸的选定来自于二维肿瘤掩模面积最大的前 10%的平均值，小尺寸的选定来自于二维肿瘤掩模面积最小的前 10%的平均值，中尺寸的确定按照公式：

$$\text{Median\_Size}=\frac{1}{2}\left(\text{Large\_Size}+\text{Small\_Size}\right),$$

确定。像素中心点属于肿瘤区域的输入图像定为正样本，不属于肿瘤区域的定为负样本。训练时均衡地输入正负样本。

局部结构：大尺寸图像对应的网络中加入 Inception v1 结构，中尺寸和小尺寸的均采用 LeNet-5 结构。

- 代价函数确定参照 Luc 在论文中提出的方法。

全局代价函数：$l(\theta_s,\theta_a)=\sum_{n=1}^{N}l_{mce}\left(s(x_n),y_n\right)-\lambda\left[l_{bce}\left(a(x_n,y_n),1\right)+l_{bce}\left(a(x_n),s(x_n),0\right)\right]$，

生成器代价函数：$\sum_{n=1}^{N} l_{mce}\left(s\left(x_n\right), y_n\right) - \lambda l_{bce}\left(a\left(x_n, s\left(x_n\right)\right), 0\right)$，

判断器代价函数：$\sum_{n=1}^{N} l_{bce}\left(a\left(x_n, y_n\right), 1\right) + l_{bce}\left(a\left(x_n, s\left(x_n\right)\right), 0\right)$，

其中 θ_s，θ_a 分别表示生成器 CNN 和判断器 CNN 的参数设置，

$l_{mce}\left(y, y\right) = -\sum_{i=1}^{H \times W} \sum_{c=1}^{C} y_{ic}$ 表示对于预测的 y 的多级交叉熵损失，

$l_{bce}\left(z_1, z_2\right) = -\left(z_2 \ln z_1 + \left(1 - z_2\right) \ln\left(1 - z_1\right)\right)$ 表示二元交叉熵损失。

3.3 结果与评价

用 LeNet-CRF 和 MIGAN 分割的部分效果图如下：

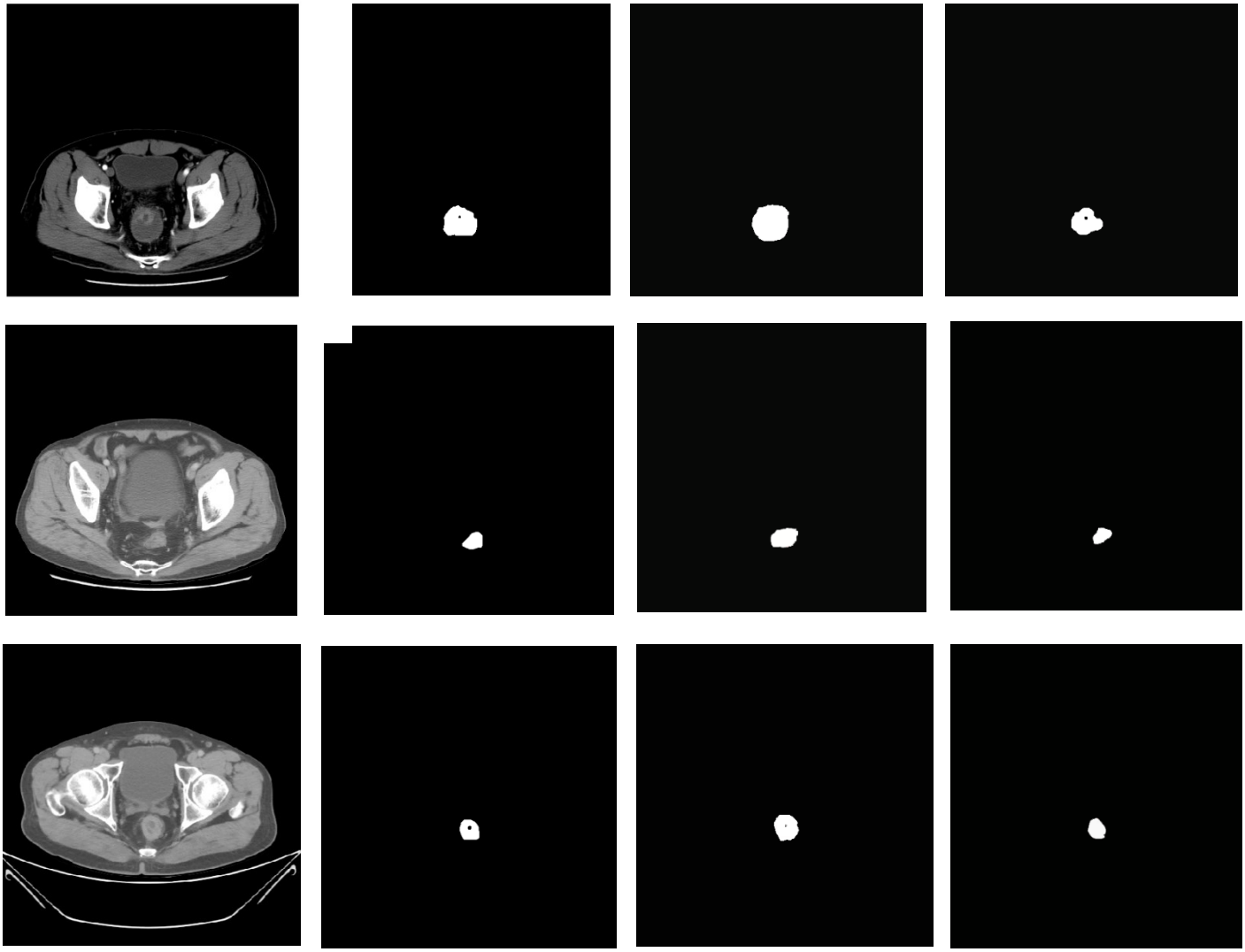


图 3.3-1 CNN 分割肿瘤效果图，第一列为原始 CT 图像，第二列为医生勾画区域，
第三列为 LeNet-CRF 分割结果，第四列为 MIGAN 分割效果

我们观察到：1）用 LeNet-CRF 分割得到的肿瘤掩模一般比金标准要稍大一些，而用 MIGAN 分割得到的肿瘤掩模一般比金标准小一些；2）两种模型分割得到的肿瘤边界或多或少存在一些不平滑的锯齿；3）两种模型对肿瘤细节的分割不稳定。例如图 3.3-1 第一行中金标准的肿瘤区域中有一个没覆盖的

小黑点，LeNet-CRF 模型有体现出来而 MIGAN 没有；第三行中金标准的肿瘤区域内有一个更明显的未覆盖区域，LeNet-CRF 模型没有体现出来而 MIGAN 有。

下面我们各自展开两个模型的精度指标和评价指标。

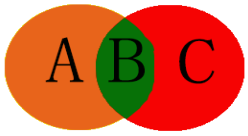


图 3.3-2 精度指标示意图

对于精度指标，定义金标准的像素点集合为 $S_{金}$，模型分割得到的肿瘤区域像素点集合为 $S_{模}$。令 $B = S_{金} \bigcap {}_{模}$，$A = S_{模} \setminus B$，$C = S_{模} \setminus B$。则四个精度指标计算公式如下：

正确分割区域在金标准集合中占比$\left(TS_{金}\right)$：$B/\left(A+B\right)$，

正确分割区域在分割总区域集合中占比$\left(TS_{模}\right)$：$B/\left(B+C\right)$，

金标准区域未被涉及到的部分占金标准区域的比重$\left(FS_{金}\right)$：$A/\left(A+B\right)$，

模型分割区域中错误的部分占分割总区域比重$\left(FS_{模}\right)$：$C/\left(B+C\right)$.

对于评价指标，我们采用 Dice 系数。它是一种集合相似度度量函数，通常用于计算两个样本的相似度，公式为：

$$Dice\left(A,B\right)=\frac{2\left|A\bigcap\right.}{\left|A\right|+\left|B\right|},$$

其中 A 是医生勾画的直肠肿瘤区域， B 是算法分割得到的直肠肿瘤区域。

- LeNet-CRF 模型。

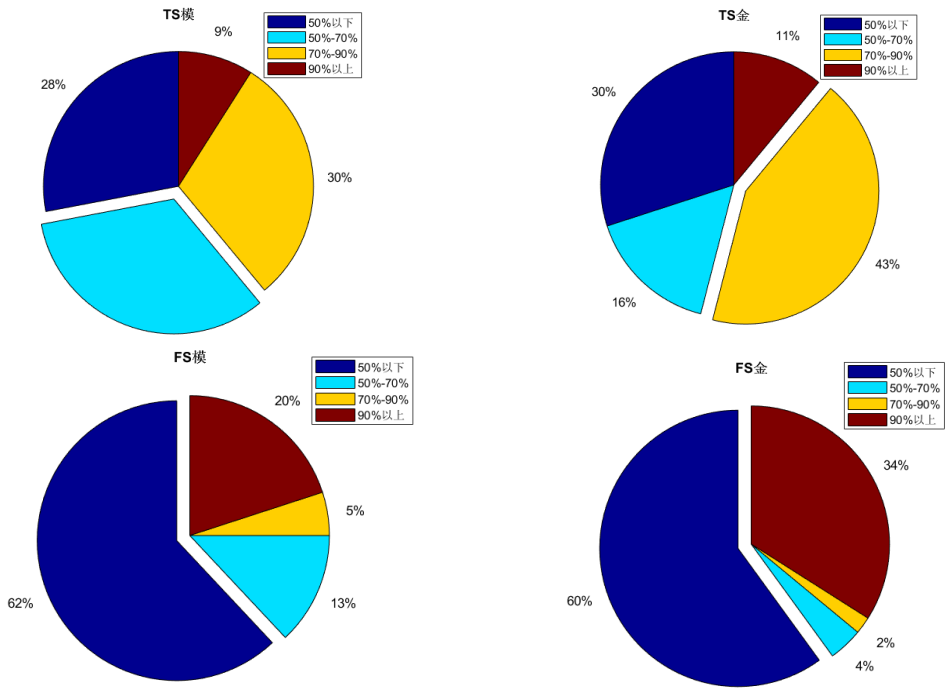


图 3.3-3 LeNet-CRF 模型的四个精度指标

我们观察到对于$TS_{模}$，它在 50%-70%的比例最大，其次是 70%-90%的区间，说明该模型分割出来的区域中大部分是正确区域。然而$TS_{模}$在 50%以下的比重也比较大，这可能是因为在一些医生没标记掩模的 CT 图像上模型也分割出肿瘤区域。对于$TS_{金}$，它在 70%-90%的比例最大，说明金标准中大部分区域被分割区域涵盖到了；比例第二大的是 50%以下的区间，这也是无标记掩模 CT 上依旧分割出肿瘤区域的结果。对于$FS_{模}$和$FS_{金}$，50%以下的区间占超过半数的比例，说明模型分割出的错误区域和金标准未被涵盖到的区域较少，而二者 90%以上的比重较大，这也是因为我们把无标记掩模 CT 上依旧分割出肿瘤区域的情况归类到$FS_{模}=100\%$和$FS_{金}=100\%$的情况。总体而言该模型在这四个指标上的表现可以接受，但精度不高。

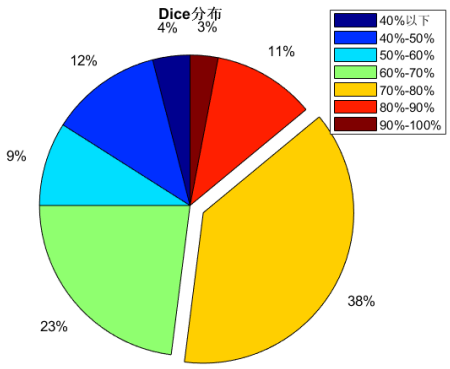


图 3.3-4 LeNet-CRF 模型下 Dice 大小的分布

我们观察到分割结果的 Dice 系数比较集中分布在 70%-80%的区间，其次是 60%-70%的区间。这表明该模型能分割出超过半数的正确区域，但是分割精度仍待进一步的提高。而 Dice 系数分布在 80%以上的比例也有 14%（11%+3%），说明该模型还具有提升精度的潜力。而 Dice 系数分布在 50%以下的比例有 16%（12%+4%），说明该模型还是具有潜在的缺陷，还不足以投入实际生产生活中。

这里还比对了在 LeNet-CRF 模型中同时使用 Arterial 期和 Venous 资料，只使用 Arterial 期资料和只使用 Venous 期资料的平均 Dice 系数，结果如下：

	$A\&V$	A	V
$Dice$	73%	67%	65%

表 3.3-5 不同资料采用下的平均 Dice 系数

我们观察到平均 Dice 系数在同时使用 Arterial 期和 Venous 期资料的时候最大，单独使用 Arterial 期资料其次，单独使用 Venous 期资料最次。于是我们判断 Arterial 期资料对肿瘤分辨有更大的贡献作用且两种类型资料的混合使用能增强模型表现能力。当混合使用两种资料时，平均 Dice 系数可达到七成以上，效果比较可观。

- MIGAN 模型。

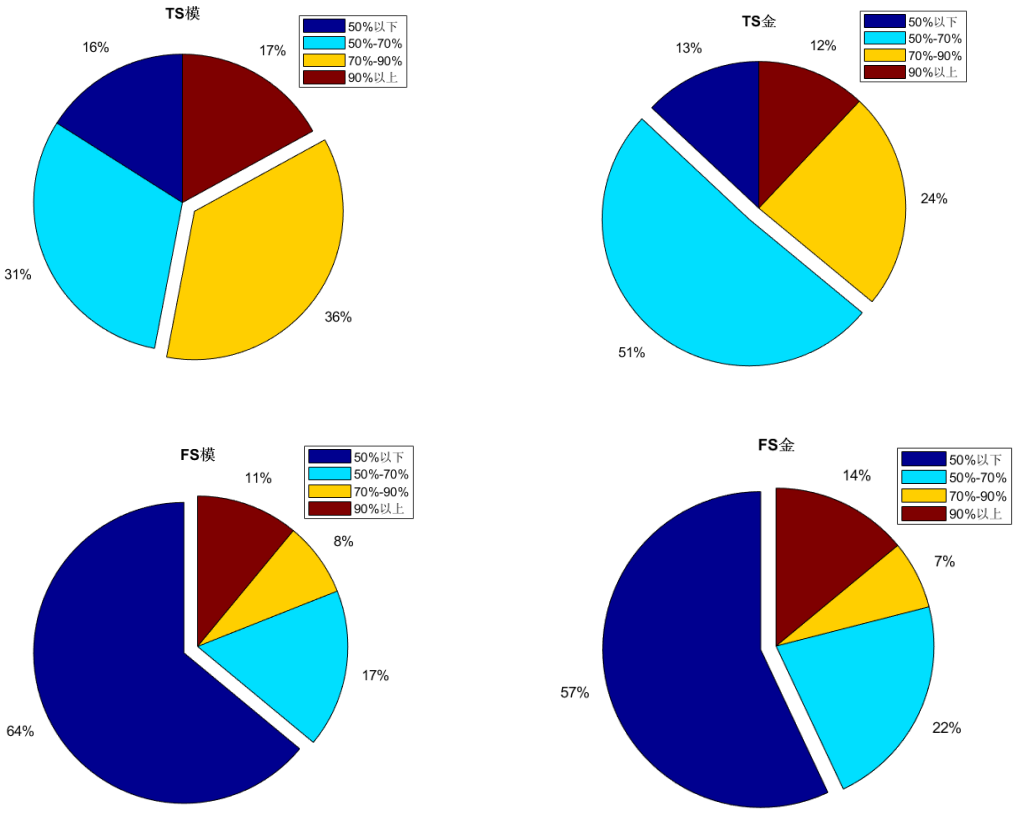


图 3.3-6 MIGAN 模型的四个精度指标

我们观察到对于$TS_{模}$，它在 70%-90%的比例最大，其次是 50%-70%的区间，说明该模型分割出的区域中绝大部分是正确区域。然而$TS_{模}$在 50%以下的比重也不可忽视，这可能是因为在一些医生没标记掩模的 CT 图像上模型也分割出肿瘤区域。对于$TS_{金}$，它在 50%-70%的比例最大，大约一半，说明金标准中超过一半的区域被分割区域涵盖到了；比例第二大的是 70%-90%的区间。50%以下也有 13%的比例，这也是无标记掩模 CT 上依旧分割出肿瘤区域的结果。对于$FS_{模}$和$FS_{金}$，50%以下的区间占超过 60%的比例，说明模型分割出的错误区域和金标准未被涵盖区域较少，而二者 90%以上的比重都是在 15%，这是因为我们把无标记掩模 CT 上依旧分割出肿瘤区域的情况归类到$FS_{模}=100\%$和$FS_{金}=100\%$的情况，但出现这种情况的次数较少。总体而言该模型在这四个指标上的表现可接受，表现效果比 LeNet-CRF 好一点。

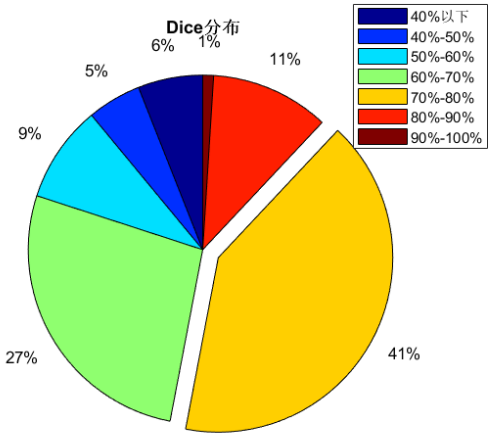


图 3.3-7 MIGAN 模型下的 Dice 系数分布情况

我们观察到分割结果的 Dice 系数比较集中分布在 70%-80%的区间，其次是 60%-70%的区间。这表明该模型能分割出超过半数的正确区域，但是分割精度仍待进一步的提高。而 Dice 系数分布在 80%以上的比例也有 12%（11%+1%），说明该模型还具有提升精度的潜力。该模型最终的平均 Dice 系数为 75%，比 LeNet-CRF 模型的高上 2%。从提高精度的角度来看，这个模型并不是那么成功，原因也是我们这个 MIGAN 的网络架构层次较低，无法进行深层次的学习。而网络层次的加深也会相应地带来网络训练的难题。针对此我们计划，若要相应的加深网络结构，则要配合上卷积层后正则化 BN（Batch Normalizing）技术，全连接层 Dropout 技术等加快网络收敛的技术。

4. 基于 CT 影像的肿瘤特征提取

4.1 任务的分析

我们被提供的原始数据有两类，第一类是病人的 CT 文件，第二类是病人的基本信息表格。第二类数据的表格只提供了每个病人的 ID，性别，年龄，淋巴结转移情况，与肿瘤影像特征无关。因此在该任务中我们仅利用第一类数据。

对于第一类，病人的样本数为 107 个，每个病人有两套 CT 数据，一套是动脉期（Arterial）影像数据，另一套是门静脉期（Venous）影像数据。每套影像下有一个完整的 CT 扫描序列，该序列依据扫描位置进行排序，序列中每一张切片都有对应的肿瘤掩模图像。

对于这项任务，我们想要提取肿瘤的二维特征和三维特征。原始的 CT 资料中的灰度矩阵只提供了整张 CT 每个像素点位置的灰度值，完全不足以直接拿来作为肿瘤特征。高通量特征的提取，即从医学图像单一的信息矩阵中提取成百上千甚至数万以上的特征信息，是影像组学的关键所在[1]。为了做到这一点，我们拟采用以下思路。

1. 进行肿瘤特征提取，首先要进一步选择 ROI（Region of Interest）。通过 MATLAB R2017a 版本自带的 regionprops( ) 函数对每个病人所有切片层肿瘤面积的提取，进行一次比对筛选出每个病人肿瘤的最小和最大截面积，然后分别在所有病人的范围内进行二次比对筛选出所有病人的肿瘤最小和最大截面积。将最小截面积对应的肿瘤掩模和最大截面积对应的肿瘤掩模与各自的原 CT 图像作比对。

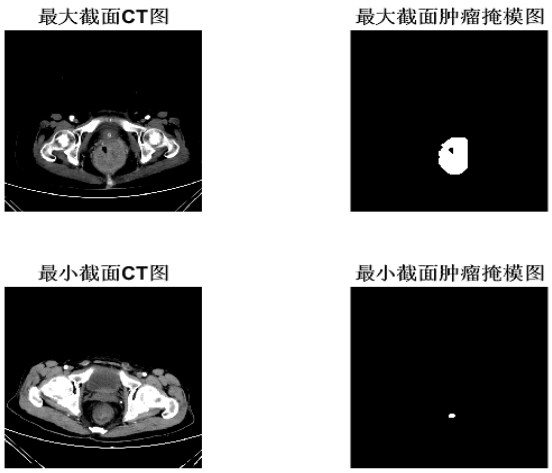


图 4.1-1 最大截面积肿瘤掩模图，最小截面积肿瘤掩模图与各自的 CT 图像的对比，

面积的值以肿瘤截面包含的像素个数来表示。整个 CT 图像的面积是 512*512=262144。最小截面来源于病人“1071”的 Arterial 期中的“10016.Mask.png”，截面面积为 167，占原 CT 图像面积的 0.064%；最大截面来源于病人“1021”的 Arterial 期中的“10019.Mask.png”，截面面积为 6473，占原 CT 图像面积的 2.5%。若定义原 CT 图像中除去所有灰度值为 0 的像素点后剩下的像素点个数为该 CT 的有效面积，则通过 MATLAB 计算得到的最小截面 CT 图的有效面积为 81401，最小的肿瘤截面积占比为 0.21%；最大截面 CT 图的有效面积为 61900，最大的肿瘤截面积占比 10.5%。因此在原 CT 灰度图像中寻找包含肿瘤一部分或整体的，合适大小的窗口来提取肿瘤特征是很有必要的。为了实现：1）从多个视角提取肿瘤特征；2）模拟实际情况中不同医生对同一 CT 资料手工分割得到的肿瘤区域不一致；3）评判特征提取窗口的形状和大小对淋巴结转移验证模型的影响程度，我们决定采用 5 个类型的窗口。

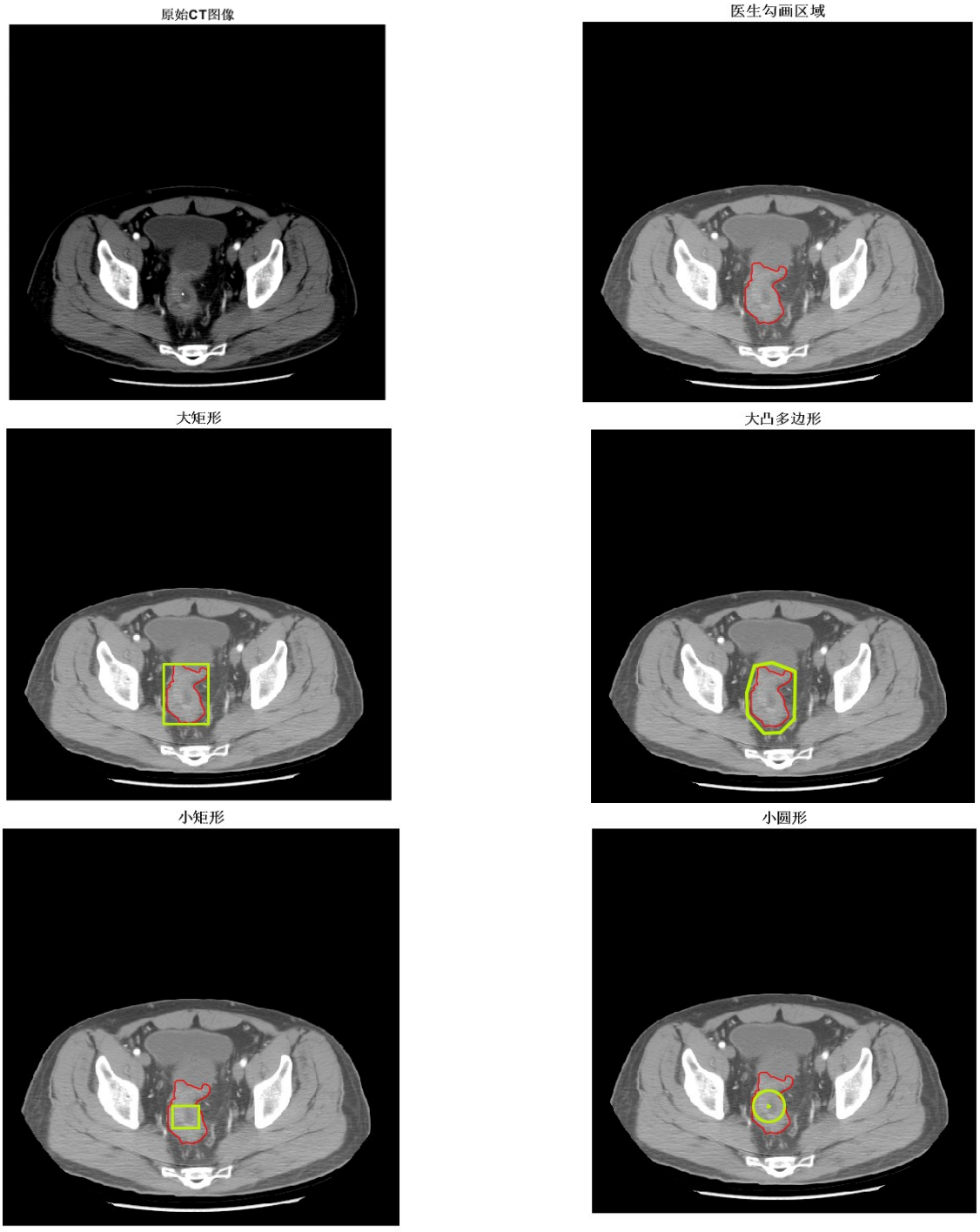


图 4.1-2 以病人“1005”的某一 Arterial 期切片“10070”为例子的
五种特征提取窗口，黄色框线为提取窗口

第一个类型的窗口正好是医生勾画的不规则区域。第二个类型的窗口是一个矩形，是包含原不规则区域的最小矩形，简单地模拟了所选区域适当扩大。第三个类型的窗口是一个凸多边形，边数视具体的肿瘤形状而定，这个多边形也是包含原不规则区域的最小多边形，比较真实地模拟了适当扩大选区后得到的边界较为光滑的区域。第四个类型的窗口是矩形，其对角线的交点是肿瘤像素位置分布的质心，对角线的长度的一半取自于质心到肿瘤边界八个极值点的距离的最小值的一半，公式表示为：

$$\frac{l}{2}=\min\left\{\sqrt{(X-X_i)^2+(Y-Y_i)^2}\right\},i=1,2,...,8\ ,$$

其中 l 是矩形对角线长度，X,Y 分别是质心的横，纵坐标，X_i,Y_i 分别是第 i 个极值点的横，纵坐标。第四个类型的窗口简单地模拟了所选区域适当缩小的情况。

第五个类型的窗口是圆形，记为圆 A。其圆心位置即为肿瘤像素位置分布的质心位置。圆 A 的半径来源于面积与原肿瘤截面积相等的圆 B 的半径长度的一半。公式表示为：

$$\begin{cases} r=\dfrac{R}{2} \\ \pi R^2=S_{\text{肿瘤}} \end{cases},$$

其中 r 是圆 A 的半径，R 是圆 B 的半径，$S_{\text{肿瘤}}$ 是肿瘤的截面积。第五个类型的窗口比较粗糙地模拟了适当缩小选区后得到的边界较为光滑的区域。

2. 选定窗口后进行肿瘤特征的第一次粗提取。为了提炼原 CT 灰度矩阵中的灰度值大小，灰度分布空间信息，灰度等级，灰度个数统计，灰度变化等信息，减小噪声的干扰性，我们决定进行以下操作。

对二维切片中的 ROI，把所有像素位置的灰度值转化为对应的灰阶。对于以什么灰度范围分割灰阶，我们采用两个范围：1）灰度最小值—灰度最大值范围；2）0—255 的灰度范围。两个范围下得到的灰阶矩阵分别记为 X_1, X_2。对于总灰阶数的确定，由于人的裸眼能识别的灰阶数一般为 16，我们把两种范围的总灰阶数都确定为 16。由此得到两种灰阶矩阵。此外，对原 ROI 区域进行二维小波分解得到四种小波系数，系数为二维矩阵形式。这四个小波系数分别为：

$$X_{LL}, X_{LH}, X_{HL}, X_{HH}\ ,$$

其中 X 是二维 ROI，L, H 分别对应低通滤波器，高通滤波器；X_{AB} 表示对 x 轴进行 A 型滤波，对 y 轴进行 B 型滤波得到的小波系数，$A,B\in\{L,H\}$。

对三维切片中的 ROI，也采用上述的两个灰度范围进行 16 阶灰度的分割，将原 ROI 三维灰度矩阵转化为两种三维灰阶矩阵 Y_1, Y_2。此外，对原 ROI 区域也进行三维小波分解得到八种小波系数，系数均为三维矩阵形式。这八个小波系数分别为：

$$Y_{LLL}, Y_{LLH}, Y_{LHL}, Y_{LHH}, Y_{HLL}, Y_{HLH}, Y_{HHL}, Y_{HHH}\ ,$$

其中 Y 是三维 ROI，L, H 分别对应低通滤波器，高通滤波器；Y_{ABC} 表示对 x 轴进行 A 型滤波，对 y 轴进行 B 型滤波，对 z 轴进行 C 型滤波，$A,B,C\in\{L,H\}$。

总结得到的第一次粗提取特征形式如下：

	X1	X2	X(L,L)	X(L,H)	X(H,L)	X(H,H)	Y1	Y2	Y(L,L,L)	Y(L,L,H)	Y(L,H,L)	Y(L,H,H)	Y(H,L,L)	Y(H,L,H)	Y(H,H,L)	Y(H,H,H)
病人1																
病人2																
病人3																
.....																

表 4.1-3 第一次特征粗提取的规模

3. 进行完第一次特征粗提取后进行第二次的特征细提取。第二次的特征细提取得到的特征数据大致可分为四组：第一组是一阶统计量特征（First order statistics），第二组是基于形状和大小的特征（Shape and size based features），第三组是纹理特征（Textural features），第四组是小波特征，直接取自粗提取特征中的小波系数。第一至三组特征的提取方法参考 Aerts 发表在期刊 Nature Communications 的方法和影像组学软件 pyradiomics 官方网站 Radiomic Features 一栏介绍的方法[10]。

至此我们绘制该项任务的工作流程如下：

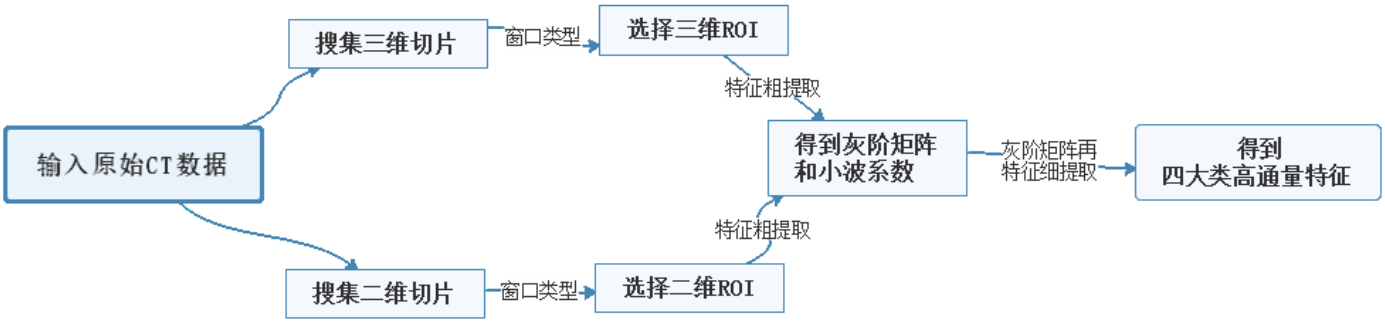


图 4.1-4 肿瘤特征提取任务的流程图

4.2 问题的发现与解决

1. 对于二维切片的选择问题。有的病人一套 CT 序列里面就只有 3 张是有勾画出肿瘤掩模的（如病人 1003 的 Arterial 期 CT 序列），而有的病人一套 CT 序列里面有多达 14 张具备肿瘤掩模（如病人 1043 的 Arterial 期 CT 序列）。如若把所有病人的所有含肿瘤二维切片都提取出来，分别进行二维特征提取，那么计算量将是无比庞大的，且出自同一病人的肿瘤 CT 片提取出来的特征可能有重复相似的地方，造成信息冗余。为了减小计算量的同时最大程度地提取每个病人的二维肿瘤特征，我们决定对每个病人，先计算每张含肿瘤切片中肿瘤截面积的大小，选出肿瘤截面积最大对应的 CT 切片代表全部含肿瘤切片来进行计算。

2. 对于三维特征的计算问题。首先是三维 ROI 的选择，三维图像矩阵的每个切片的 ROI 选择几乎都是各不相同的。各层切片的 ROI 的选择不统一会带来计算的麻烦与不规范。为了统一各层切片的 ROI，我们决定把各层 ROI 区域的并集作为最终每一层提取特征的窗口。设对于某一个病人，含肿瘤 CT 切片有 L 张，第 i 张对应的肿瘤掩模的二值图像的 ROI 像素集合记为 $P(i)=\{(x_k,y_k)|(x_k,y_k)\text{处的灰度值为}255\}$，那么最终每层的区域集合为 $P=P(1)\cup P(2)\cup P(3)\cup\ldots\cup P(L)$.

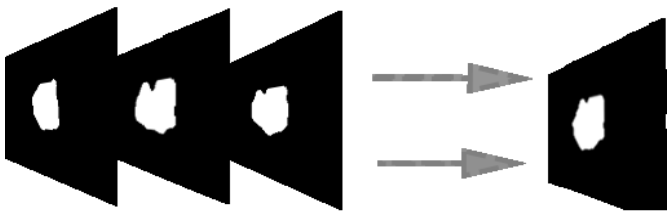


图 4.2-1 三维 ROI 并集选择法的投影理解

此外，肿瘤掩模图像还可能出现一些极端情况。比如同一切片上有两个或两个以上的不连通肿瘤区域，或者一段连续的 CT 序列中出现“第一张有肿瘤掩模→第二张无肿瘤掩模→第三张再次出现肿瘤掩

模”。对于肿瘤区域不连通情况，有两种方式进行处理：其一是保留这些切片，但要更注重 ROI 选择涵盖到所有区域的肿瘤面；其二是对这些切片进行剔除（这些切片数量占比很少），免除编程和计算上的麻烦。而对于 CT 序列中肿瘤掩模的“有-无-有”出现模式，如果出现，则要意识到该病人的直肠癌大概率不止一个，一般计算还是可以把这多个肿瘤视为一个整体，但诸如体积，表面积这类三维特征的计算就要注意多个肿瘤分开计算再求和。

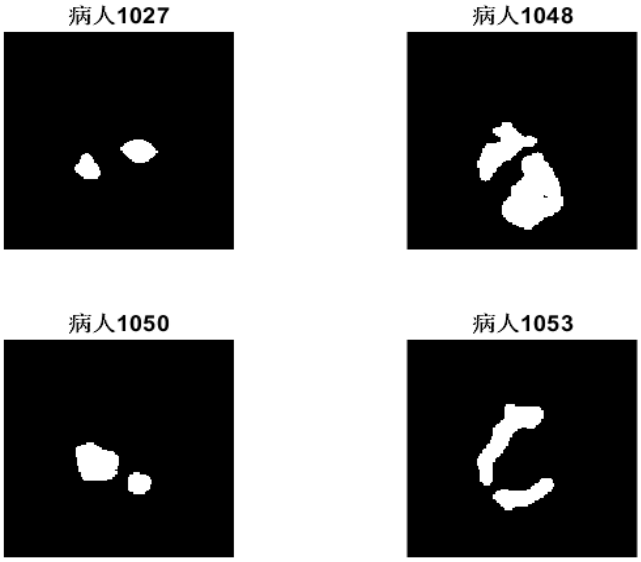


图 4.2-2 同一张切片出现肿瘤区域不连通的情况

对于任务推荐的三维特征：体积，表面积，如何通过一系列没有厚度的肿瘤切片的堆叠来计算我们想要的体积和面积？我们重点也对这个问题提出自己的算法。其他特征的算法详见附录。

对于一个系列的肿瘤切片，我们视任意相邻两个肿瘤面之间构成一个台体（两个相邻肿瘤面一模一样则构成柱体）。台体的高度用 MATLAB R2017a 的 dicominfo( ) 函数读取病人 dcm 文件信息中的“Spacing between slices”（单位：mm）来表示。设上截面面积为 S_{UP}，下截面面积为 S_{DOWN}。假定上截面到下截面的变化是连续的，将每个台体等间隔地用平行于切片方向的平面分割成充分多的 N 个部分（N 充分大，取 10000 以上），则每个部分可近似视为上下截面相同的几何体。从上往下数，第 i 个部分的上下截面面积均为 $S_{UP}+(i-1)\dfrac{S_{DOWN}-S_{UP}}{N}$，高度为 $\dfrac{spacing\ \ between\ \ slices}{N}$。

对于体积计算，一种方法是将每个上下截面相同的小几何体视为台体，利用台体体积公式计算。本次任务我们采取另一种计算方便的方法——根据祖暅原理的思想，将每个上下截面相同的小几何体体积化为小圆柱体的体积来计算。公式表示为如下：

$$Volumn=\sum_{j=1}^{n}\sum_{i=1}^{N_j}\left(S_{j,UP}+\frac{S_{j,DOWN}-S_{j,UP}}{N_j}\bullet(i-1)\right)\bullet\frac{spacing\ \ between\ \ slices}{N_j},$$

其中 n 表示肿瘤切片之间的间隔数，j 是切片间隔的索引，N_j 是第 j 个间隔对应的台体被分割成小圆柱体的数目，i 是第 j 个间隔里面的圆柱体索引，$S_{j,UP}$，$S_{j,DOWN}$ 分别表示第 j 个间隔台体的上，下截面积。

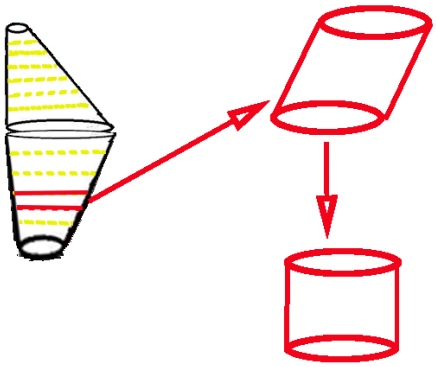


图 4.2-3 体积计算中祖暅思想运用简单示意图

对于表面积计算，将每个上下截面相同的小几何体视为圆柱体。利用圆柱体表面积公式计算。考虑最简单的“病人只有一个直肠肿瘤”的情况。计算公式如下：

$$Surface\ Area = S_{1,UP} + S_{n,DOWN} + \sum_{j=1}^{n}\sum_{i=1}^{N_j}\left(C_{j,UP} + \frac{C_{j,DOWN} - C_{j,UP}}{N_j}\bullet(i-1)\right)\bullet\frac{spacing\ \ between\ slices}{N_j},$$

其中 $S_{1,UP}, S_{n,DOWN}$ 分别为第 1 张肿瘤切片截面积和最后一张肿瘤切片截面积，n 表示肿瘤切片之间的间隔数，j 是切片间隔的索引，N_j 是第 j 个间隔对应的台体被分割成小圆柱体的数目，i 是第 j 个间隔里面的圆柱体索引，$C_{j,UP}$, $C_{j,DOWN}$ 分别表示第 j 个间隔台体的上，下截面周长。

4.3 特征提取结果

4.3.1 二维特征提取

对于肿瘤二维特征的提取。在对每个病人选出代表性的二维切片后，我们分别用了“db2”，“sym2”，“coif1”，“coif2”和“bior2.4”这五个小波基对二维切片进行二维小波分解。每张切片在应用某一小波基分解后都得到四个小波系数。将小波系数可视化如下：

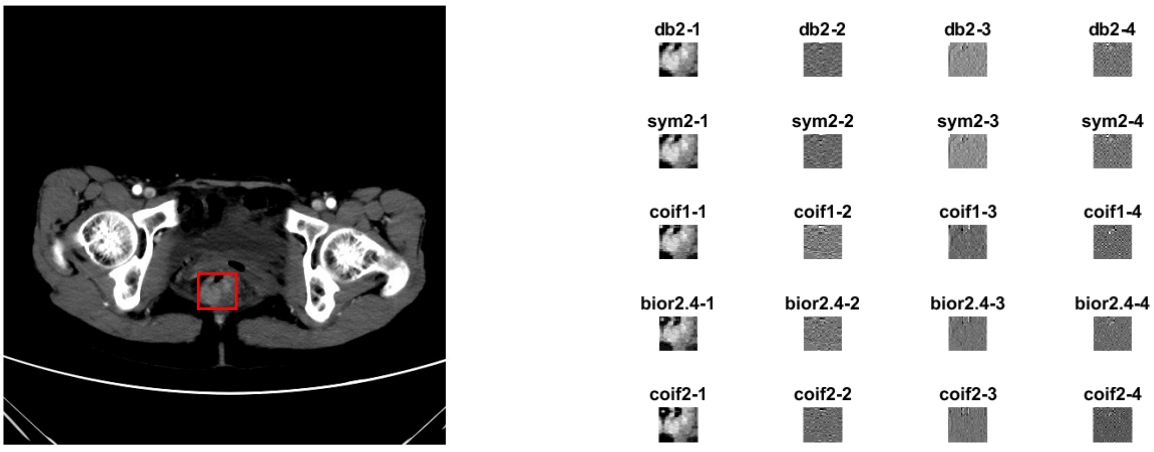


图 4.3-1 病人 1003 动脉期一切片小波分解

之后选出小波基“coif1”分解得到的全部结果，并与经 Windowlevel 预处理得到的 CT 灰度矩阵和基

于“0-255”和“矩阵元素最小值-矩阵元素最大值范围”分割得到的灰阶矩阵一起提取其他高通量特征如下：

1）一阶统计量特征: Energy, Entropy, Kurtosis, Maximum Intensity, Mean Intensity, Median Intensity, Minimum Intensity, Mean Absolute Deviation, Range of Intensity, Root Mean Square (RMS), Skewness, Standard Deviation, Uniformity, Variance。样本结果如下：

	1	2	3	4	5	6	7	8	9	10
1	'Energy_...	'Energy_2...	'Energy_itf'	'Entropy_...	'Entropy_i...	'Mean_WL'	'Mean_255'	'Mean_itf'	'Maximu...	'Minimun...
2	3.3612e+...	229595	255050	2.8766	3.5041	117.1039	1.5092	10.2503	191.9211	0
3	1.7482e+...	233035	243198	2.9928	2.8733	101.6806	0.0781	12.0049	141.3087	0

	11	12	13	14	15	16	17	18	19	20
1	'Median_...	'Range_WL'	'Kurtosis_...	'Kurtosis_...	'Kurtosis_...	'MAD_WL'	'MAD_255'	'MAD_itf'	'RMS_WL'	'RMS_255'
2	123.4737	181.6707	4.4864e+...	152.2188	29.4249	29.1520	6.9919	2.4151	122.5793	10.1309
3	105.4061	129.3038	2.4107e+...	155.4563	30.1579	13.0970	11.6260	1.4682	103.2782	11.9240

	21	22	23	24	25	26	27	28	29	30
1	'RMS_itf'	'Skewnes...	'Skewnes...	'Skewnes...	'SD_WL'	'SD_255'	'SD_itf'	'Uniformi...	'Uniformi...	'Variance...
2	10.6777	-0.7836	12.1147	-2.2489	36.2345	8.9466	2.9915	0.0442	0.1026	1.3129e+...
3	12.1812	-1.6235	12.3572	-3.2049	18.1008	11.8509	2.0658	0.1586	0.1750	327.6396

	31	32
1	'Variance...	'Variance...
2	80.0414	8.9489
3	140.4449	4.2674

图 4.3-2 病人 1001 和 1002 Arterial 期二维一阶统计量提取的特征（共计 32 个）

2）基于形状和大小的特征：Section Area, Perimeter。样本结果如下：

	33	34
1	'Surface_Area'	'Perimeter'
2	2237	167.6208
3	1639	143.4777

图 4.3-3 病人 1001 和 1002 Arterial 期二维形状特征的提取结果（共计 2 个）

3）纹理特征：

采用 Gray level co-occurrence matrix（GLCM），二维 GLCM 产生方案有 8 种如下表所示。

方向 / 间距	x轴正向	y轴正向	(1,1)	(-1,1)
1	Y	Y	Y	Y
5	Y	Y	Y	Y

表 4.3-4 二维 GLCM 产生方案，Y 表示采纳，N 表示不采纳

提取特征：Autocorrelation, Cluster Prominence, Cluster Shade, Cluster Tendency, Contrast, Correlation, Difference Entropy, Dissimilarity, Energy, Entropy (H), Homogeneity 1, Homogeneity 2, Informational Measure of Correlation 1 (IMC1), Informational Measure of Correlation 2 (IMC2), Inverse Difference Moment Normalized (IDMN), Inverse Difference Normalized (IDN), Inverse Variance, Maximum Probability, Sum Average, Sum entropy, Sum Variance, Variance。计算方法参考附录。样本结果如下：

	1	2	3	4	5	6	7	8	9	10
1	'Autocorr...	'Cluster_P...	'Cluster_S...	'Cluster_T...	'Contrast'	'Correlati...	'Differenc...	'Dissimilil...	'Energy'	'Entropy'
2	69894	5.5193e+...	-3.7833e...	26034599	4727	-1.5521e...	1.0162e+...	1453	142512	-5.9379e...
3	67707	1.8575e+...	-2522163...	3484449	2196	-4.1900e...	5.1474e+...	810	19342	-2.6331e...

	11	12	13	14	15	16	17	18	19	20
1	'Homoge...	'Homoge...	'IMC1'	'IMC2'	'IMDN'	'IDN'	'Inverse_v...	'Maximu...	'Sum Ave...	'Sum entr...
2	820.3202	770.7811	-0.6996	1	1.2480e+...	1.2478e+...	380.9119	362	14811	7.5885e+...
3	438.9905	412.9234	-0.7526	1	700.0000	699.8963	275.6600	94	12274	3.5251e+...

	21	22
1	'Sum vari...	'Variance'
2	7.1642e+...	29716
3	8.6122e+...	4.0815e+...

表 4.3-5 基于 GLCM 矩阵提取的 Arterial 期二维纹理特征的其中一种（每种包含 22 个特征）

采用 Gray-Level Run-Length matrix（GLRLM），二维 GLRLM 产生方案有 4 种如下表所示。

方向	x轴正向	y轴正向	(1,1)	(-1,1)

表 4.3-6 二维 GLRLM 产生方案

提取特征：Short Run Emphasis（SRE），Long Run Emphasis（LRE），Gray Level Non-Uniformity（GLN），Run Length Non-Uniformity（RLN），Run Percentage（RP），Low Gray Level Run Emphasis（LGLRE），High Gray Level Run Emphasis（HGLRE），Short Run Low Gray Level Emphasis（SRLGLE），Short Run High Gray Level Emphasis（SRHGLE），Long Run Low Gray Level Emphasis（LRLGLE），Long Run High Gray Level Emphasis（LRHGLE）。计算方法参考附录。样本结果如下：

	1	2	3	4	5	6	7	8	9	10	11
1	'SRE'	'LRE'	'GLN'	'RLN'	'RP'	'LGLRE'	'HGLRE'	'SRLGLE'	'SRHGLE'	'SRHGLE'	'LRLGLE'
2	0.7581	2.6329	203.3161	840.6456	0.7014	0.0351	65.8445	0.0260	49.8535	0.1125	175.8305
3	0.5275	8.9391	170.3217	195.4000	0.4210	0.0349	45.8319	0.0234	22.5088	0.2017	455.8333

表 4.3-7 基于 GLRLM 矩阵提取的 Arterial 期二维纹理特征的其中一种（每种包含 11 个特征）

4.3.2 三维特征提取

对于肿瘤三维特征的提取。在提取每个病人肿瘤三维切片后，我们分别用了“db2”，“sym2”，“coif1”，“coif2”和“bior2.4”这五个小波基对三维切片进行三维小波分解。每个三维肿瘤灰度矩阵在应用某一小波基分解后都得到八个小波系数（三维矩阵形式）。以“coif1”分解结果为例，将分解得到的八个三维矩阵的每一层可视化如下：

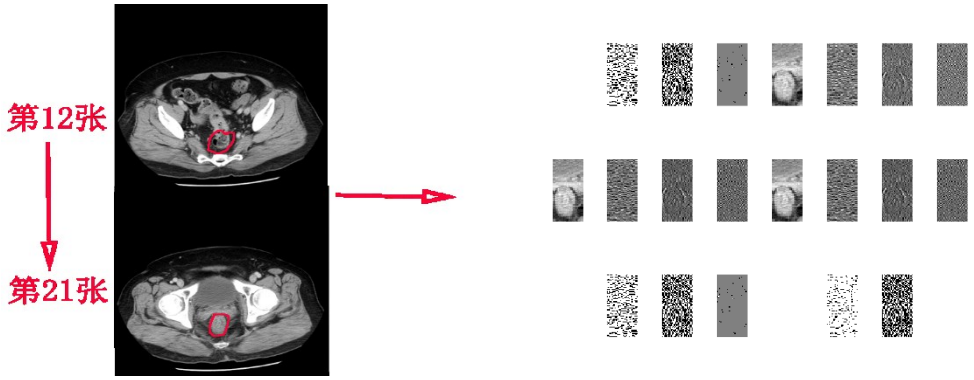


图 4.3-8 病人 1020 静脉期三维肿瘤小波分解，同一列的属于分解后的同一三维矩阵不同层次

之后选出小波基“coif1”分解得到的全部结果，并与经 Windowlevel 预处理得到的 CT 灰度矩阵和基于“0-255”，“矩阵元素最小值-矩阵元素最大值”范围分割得到的灰阶矩阵一起提取其他高通量特征如下：

1）一阶统计量特征: Energy, Entropy, Kurtosis, Maximum Intensity, Mean Intensity, Median Intensity, Minimum Intensity, Mean Absolute Deviation, Range of Intensity, Root Mean Square (RMS), Skewness, Standard Deviation, Uniformity, Variance。结果以一个病人为例子表示如下：

	1	2	3	4	5	6	7	8	9	10
1	'Energy_...	'Energy_2...	'Energy_itf'	'Entropy_...	'Entropy_i...	'Mean_WL'	'Mean_255'	'Mean_itf'	'Maximu...	'Minimun...
2	1.2791e+...	564307	856498	3.1432	3.4571	115.7238	0	9.5133	205.3421	0
3	8.2591e+...	374783	1109080	1.9939	2.7951	101.1456	0	11.7339	144.0704	0

	11	12	13	14	15	16	17	18	19	20
1	'Median_...	'Range_WL'	'Kurtosis_...	'Kurtosis_...	'Kurtosis_...	'MAD_WL'	'MAD_255'	'MAD_itf'	'RMS_WL'	'RMS_255'
2	119.4474	195.8288	4.4208e+...	80.8943	25.5808	28.5230	7.7312	2.2350	121.1440	8.0464
3	103.5650	132.3365	3.5734e+...	51.9312	40.9800	12.3882	6.8268	1.4275	102.8353	6.9273

	21	22	23	24	25	26	27	28	29	30
1	'RMS_itf'	'Skewnes...	'Skewnes...	'Skewnes...	'SD_WL'	'SD_255'	'SD_itf'	'Uniformi...	'Uniformi...	'Variance...
2	9.9130	-0.6436	8.8195	-1.6576	35.8332	8.0468	2.7866	0.1311	0.1054	1.2840e+...
3	11.9167	-2.1323	7.1563	-4.1243	18.5661	6.9278	2.0792	0.3316	0.1844	344.7012

	31	32
1	'Variance...	'Variance...
2	64.7512	7.7652
3	47.9937	4.3231

图 4.3-9 病人 1001 和和 1002 Arterial 期三维一阶统计量特征提取（共 32 个）

2）基于形状和大小的特征：Compactness 1, Compactness 2, Spherical diameter, Spherical disproportion, Sphericity, Surface Area, Surface to Volume Ratio, Volumn。样本结果表示如下：

	33	34	35	36	37	38	39	40
1	'Volumn'	'Surface_...	'Compact...	'Compact...	'Surface_t...	'Spericity'	'Spherical...	'Spherical...
2	3.7195e+04	5.9207e+...	63.9510	0.7535	0.1592	0.9134	41.2971	1.1056
3	3.2812e+04	5.6933e+...	57.9086	0.6595	0.1735	0.8736	39.6085	1.1557

图 4.3-10 病人 1001 和 1002 Arterial 期三维形状特征的提取（共 8 个）

3）纹理特征：

采用 Gray level co-occurrence matrix（GLCM），三维 GLCM 的产生方案有 17 种如下表所示。

方向 / 间距	x轴正向	y轴正向	z轴负向	(1,1,-1)	(1,1,1)	(-1,-1,-1)	(-1,-1,1)	(0,1,-1)	(0,1,1)	(1,0,-1)	(1,0,1)	(1,1,0)	(1,-1,0)
1	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
5	Y	Y	N	N	N	N	N	N	N	N	N	Y	Y

表 4.3-11 三维 GLCM 产生方案，Y 表示采纳，N 表示不采纳

提取特征：Autocorrelation, Cluster Prominence, Cluster Shade, Cluster Tendency, Contrast, Correlation, Difference Entropy, Dissimilarity, Energy, Entropy (H), Homogeneity 1, Homogeneity 2, Informational Measure of Correlation 1 (IMC1), Informational Measure of Correlation 2 (IMC2), Inverse Difference Moment Normalized (IDMN), Inverse Difference Normalized (IDN), Inverse Variance, Maximum Probability, Sum Average, Sum entropy, Sum Variance, Variance。计算方法参考附录。样

本结果如下：

	1	2	3	4	5	6	7	8	9	10
1	'Autocorr...	'Cluster_P...	'Cluster_S...	'Cluster_T...	'Contrast'	'Correlati...	'Differenc...	'Dissimila...	'Energy'	'Entropy(...
2	471908	4.1268e+...	-4.7280e...	5.4170e+...	37868	-4.6926e...	7.3763e+...	12100	1187957	-4.9010e...
3	315419	2.5559e+...	-3.2223e...	4.0624e+...	10580	-3.7049e...	7.1060e+...	5100	5878239	-5.8595e...

	11	12	13	14	15	16	17	18	19	20
1	'Homoge...	'Homoge...	'IMC1'	'IMC2'	'IDMN'	'IDN'	'Inverse_v...	'Maximu...	'Sum_ave...	'Sum_ent...
2	3.5578e+...	3.1225e+...	-0.6844	1	7.1110e+...	7.1096e+...	2.9751e+...	365	114388	6.4037e+...
3	4.4750e+...	4.3620e+...	-0.5560	1	6.4570e+...	6.4563e+...	2.9237e+...	1904	89992	6.4780e+...

	21	22
1	'Sum_vari...	'Variance'
2	2.9146e+...	2.7617e+...
3	2.7085e+...	2.1226e+...

图 4.3-12 基于 GLCM 矩阵提取的 Arterial 期三维纹理特征的其中一种（每种包含 22 个特征）

采用 Gray-Level Run-Length matrix（GLRLM），三维 GLRLM 产生方案有 13 种如下表所示。

方向	x轴正向	y轴正向	z轴负向	(1,1,-1)	(1,1,1)	(-1,-1,-1)	(-1,-1,1)	(0,1,-1)	(0,1,1)	(1,0,-1)	(1,0,1)	(1,1,0)	(1,-1,0)

表 4.3-13 三维 GLRLM 产生方案

提取特征：Short Run Emphasis（SRE），Long Run Emphasis（LRE），Gray Level Non-Uniformity（GLN），Run Length Non-Uniformity（RLN），Run Percentage（RP），Low Gray Level Run Emphasis（LGLRE），High Gray Level Run Emphasis（HGLRE），Short Run Low Gray Level Emphasis（SRLGLE），Short Run High Gray Level Emphasis（SRHGLE），Long Run Low Gray Level Emphasis（LRLGLE），Long Run High Gray Level Emphasis（LRHGLE）。计算方法参考附录。样本结果如下：

	1	2	3	4	5	6	7	8	9	10	11
1	'SRE'	'LRE'	'GLN'	'RLN'	'RP'	'LGLRE'	'HGLRE'	'SRLGLE'	'SRHGLE'	'LRLGLE'	'LRHGLE'
2	0.9174	1.4015	990.0773	6.3141e+...	0.8995	0.0389	63.4746	0.0371	57.3403	0.0473	93.7892
3	0.7422	2.9736	1.4844e...	2.7087e+...	0.6754	0.0368	46.6389	0.0310	33.6313	0.0787	147.3314

表 4.3-14 基于 GLRLM 矩阵提取的 Arterial 期二维纹理特征的其中一种（每种包含 11 个特征）

5. 基于肿瘤特征的淋巴结相关性验证

5.1 任务分析

经过上一部分的肿瘤特征提取，在“医生勾画的肿瘤区域”窗口下，我们提取了 2920 个特征如下：

	一阶统计	形状	纹理*GLCM*	纹理*GLRLM*
二维特征	32×2	2×2	22×6×2×2	11×6×2×2
三维特征	32×2	8×2	22×10×3×2	11×10×3×2

表 5.1-1 特征构成情况，红色数字表示特征种类数，蓝色数字表示 CT 种类数（动脉期和门脉期），紫色数字表示纹理矩阵来源的矩阵数（2 个原灰阶矩阵+6 或 8 个小波系数），黑色数字表示纹理矩阵灰阶方向检测的取向数（x，y 或 z）

Arterial期		Venous期	
二维	三维	二维	三维
1-430	431-1460	1461-1890	1891-2929

表 5.1-2 特征在 Arterial 期，Venous，二维，三维的分布

现阶段我们要利用提取的特征结合每个病人淋巴结转移的情况来构建淋巴结转移模型并进行预测。为此我们需要把拥有的 107 个病人样本划分为模型构建用途样本（训练样本）和模型验证用途样本（检验样本）。然后对模型的预测性能做出评价。最后再调整一些指标的值来评价模型的稳定性。

总结得到本任务流程如下：

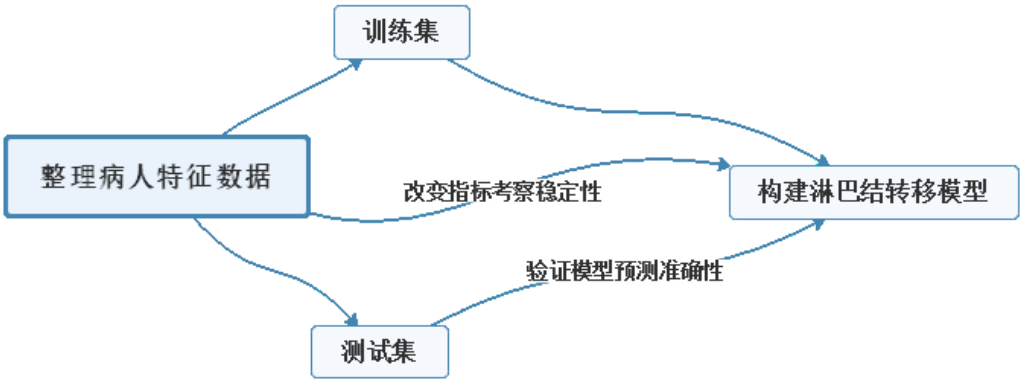


图 5.1-3 淋巴结相关性验证任务的工作流程

5.2 解决任务的方法

1. 我们提取到的病人特征维度接近 3000，但是病人样本的数量最多也只有 107。明显地，特征维度过高。但这不仅会增加计算的复杂程度，给后续的问题带来负担，还会对模型分辨性能造成负影响，引发“维度灾难”[6]。为了将如此庞大的特征维度控制在一个合适的数量下，我们有必要进行特征降维操作。

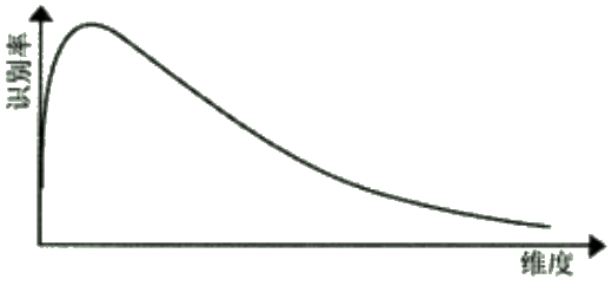


图 5.2-1 特征维度与模型识别率的关系

目前特征降维的思路主要有两大类：一类是基于原特征集通过映射产生新的特征集，即特征映射，代表技术有 PCA（Principal Components Analysis）。PCA 是通过原特征的线性组合产生新的综合指标，新特征按包含原特征信息程度进行排序依次为第一主成分，第二主成分，…… 另一类是按照一定法则选取原特征集的一个子集作为新特征集，即特征抽取，代表技术有 mRMR（Minimum Redundancy Maximum Correlation），Relief 和 LASSO 回归等。

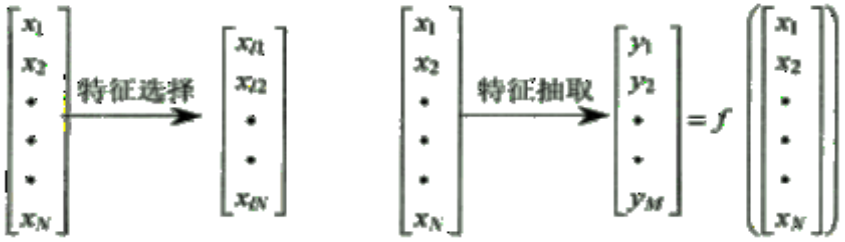


图 5.2-2 特征降维的不同思路

以 PCA 为代表的特征映射技术适用于样本数大于特征维度数的情况。因此我们主要考虑特征抽取思路。为了根据淋巴结转移结果来剔除大多数的无关或低相关因素，我们拟采用双样本-二层交集法降维，提取 100 个病人样本分成两组 A 和 B，病人 1001 至病人 1051 为 A 组，病人 1052 至病人 1101 为 B 组。在 A 组中应用 mRMR 算法得到相关性最强的前 300 个特征集 S_{A1}（约占原特征总数的 10%），应用 Relief 算法得到相关性最强的前 300 个特征集 S_{A2}，取交集特征集 $S_A = S_{A1} \bigcap$ 。同样方法得到 B 组交集特征集 S_B。最后再取 $S = S_A \bigcap$ 作为最终择取的特征。表现良好的特征应该在多种分类器上均有体现[1]，因此二层交集法抽取出的特征是比较有代表性的。但是这种方法采集到的特征数可能过少，因此我们还考虑双样本-二层交集法的缩减版本——单样本的一层交集法。将 A，B 组合并成 C 组，使用 mRMR 和 reliefF 各自得到 300 个特征集后进行交集运算。此外再比对单样本中先后使用两种抽取方法：1）先在 C 组使用 mRMR 抽取 1000 个特征（约占比 35%），再在 1000 个特征里运用 relief 选出 50 个特征；2）先在 C 组使用 relief 抽取 1000 个特征，再在 1000 个特征里运用 mRMR 选出 50 个特征。过程中涉及到的 mRMR 算法和 Relief 算法简介如下：

- mRMR：这是一种滤波式的特征选择方法，由 Peng et.al 提出。核心思想是最大化特征与分类变量之间的相关性，最小化特征与特征之间的相关性。

  最大相关性（连续变量）：$\max D$，$D = \frac{1}{|S|}\sum_{x_i \in S} F(x_i, c)$ ，

  最小冗余度（连续变量）：$\min R$，$R = \frac{1}{|S|^2}\sum_{x_i, x_j \in S} C(x_i, x_j)$ ，

  目标函数（加法整合）：$\max \Phi(D, R)$，$\Phi = D - R$，

  S 为特征子集，x_i 为第 i 个特征，c 为类别变量，$F(x_i, c)$ 为 F 统计量，$C(x_i, x_j)$ 是相关函数。

- Relief：该算法最早由 Kira 提出，基本内容是初始化各个特征的权重为 0，然后从训练集 D 中随机选择一个样本 R，从与 R 同类的样本中寻找 k 最近邻样本 H，从和 R 不同类样本寻找 k 最近邻样本 M，最后按公式更新每个特征的权重。

$$W(A) = W(A) - \sum_{j=1}^{k} diff(A, R, H_j)/(mk) + \sum_{C \in class(R)} [1 - \frac{p(C)}{1 - p(Class(R))}] \sum_{j=1}^{k} diff(A, R, M_j(C))]/(mk),$$

$$diff(A, R_1, R_2) = \begin{cases} |R_1[A] - R_2[A]| / (\max(A) - \min(A)), & A\text{是连续变量} \\ 0, & A\text{是离散的且}R_1[A] = R_2[A] \\ 1, & A\text{是离散的且}R_1[A] \neq R_2[A] \end{cases},$$

$diff(A, R_1, R_2)$ 表示样本 R_1 和 R_2 在特征 A 上的差，$M_j(C)$ 表示类 C 中的第 j 个最近邻样本，

$W(A)$是特征A的权重。

2. 抽取完主要特征后进行的是淋巴结转移模型的构建。首先为了充分利用患者本身的信息，把患者的年龄，性别并入降维后的特征，同时作为影响淋巴结转移的自变量，淋巴结转移情况作为因变量。对于性别，淋巴结转移阴阳性这类非连续的标签变量，我们将其转变为数值变量。性别男对应数值“1”，女对应数值“0”；淋巴结转移阳性对应数值“1”，阴性对应数值“0”。而基于特征分析的预测模型一般有多元 Logistic 回归，Lasso 回归，SVM 分类器，RF（Random Forest）分类器等。对此我们希望考虑回归算法中的 Logistic 算法和 RF 回归以及分类模型中的 RF 分类器。

- Logistic Regression：这是一种概率型非线性回归模型，是研究二分类观察观察结果y与一些影响因素$(x_1,x_2,...,x_n)$之间关系的多变量分析方法。

  已知θ求x的分类预测函数形式：$h_\theta(x)=g(\theta^T x)=\frac{1}{1+e^{-\theta^T x}}$，

  转化为已知x求θ的函数：$P(y|x;\theta)=(h_\theta(x))^y(1-h_\theta(x))^{1-y}$，$y=0$或$1$，

  对数似然函数：$l(\theta)=\sum_{i=1}^{m}\left(y^{(i)}\log h_\theta(x^{(i)})+(1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right)$，

  求解向量θ：$\theta_j=\theta_j-\alpha(h_\theta(x^{(i)})-y^{(i)})x_j^{(i)}$，$j=0,1,...,n$，

  θ是线性分类函数的参数向量，x是特征向量，y是分类向量，α是一个常量。

- Random Forest：随机森林是通过集成学习的思想将多棵树集成的一种算法。基本决策单元是决策树，每个决策树均进行迭代学习，最后由训练过的决策树群体一起针对某个问题做出决策。决策树是分类树则是 RF_Class 模型，决策树是回归树则是 RF_Reg 模型。随机森林的一大优点就是能处理特征数量较多的数据并且通过对策略树学习设置正则化等条件避免过拟合问题，而迭代次数的设置一般与特征数呈正相关，特征数越大，迭代次数相应地越多。因此可适当增加输入 RF 的特征数。
  - ✧ RF_Class：计算原则是基尼指数（Gini），越小表示集合中被选中样本被分错概率越小，即集合纯度越高。Gini 公式为

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

    其中p_k表示选中样本属于 k 类别的概率，K是总类别数。
  - ✧ RF_Reg：计算原则是最小均方差。对于任意划分特征 A，对应的任意划分点 s 两边划分成的数据集 D1 和 D2，求出特征划分点满足：1）D1 和 D2 各自均方差最小；2）D1 和 D2 均方差之和最小。表达式为

    $$\min_{A,s}\left[\min_{c_1}\sum_{x_i\in D_1(A,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in D_2(A,s)}(y_i-c_2)^2\right],$$

    其中c_1是 D1 数据集的样本输出均值，c_2是 D2 数据集的样本输出均值。

5.3 结果

5.3.1 特征降维

使用医生勾画区域窗口下提取的特征。使用 MATLAB R2017a, 将 relief 算法的迭代次数设置为 10000, 将 relief 算法和 mRMR 算法抽取特征的规模均设定为 300。特征散点图如下：

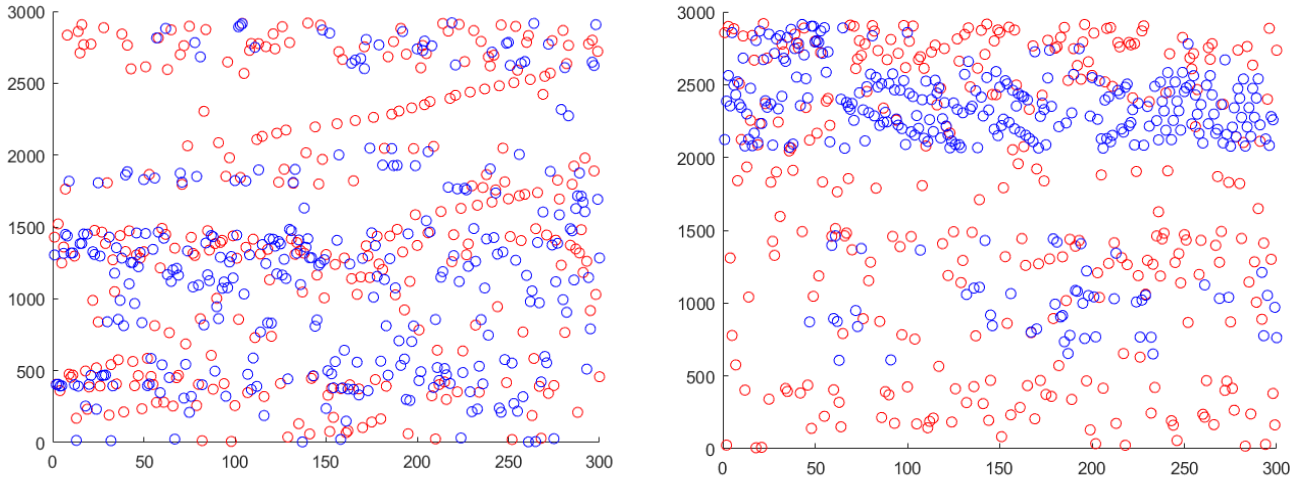


图 5.3.1-1 左图为 A 组使用两种方法各自得到的特征排序（x 轴）和特征分布（y 轴）情况，右图为 B 组。红色对应 mRMR 算法，蓝色对应 relief 算法。精准上讲，作直线 y=k（k>0 且 k 为整数），若直线过两点，则第 k 项特征是一次交集特征里的元素之一。

A，B 组选出的一次交集特征和最终的二次交集特征制表如下：

A组一次特征交集：									
1305	404	406	393	395	1316	1448	1437	1319	403
392	1384	1385	254	405	1327	1308	232	1807	1297
466	468	470	1429	13	857	989	1283	813	585
1344	2815	1464	347	2881	1121	1811	210	2782	2683
402	1440	945	1306	391	769	188	1417	1415	1382
1165	1483	1474	1317	1358	1234	1407	642	1242	361
2048	1459	1371	2045	300	411	1758	353	1261	408
1326	2784	2650	2916	1392	422	1488	1473	1148	2907
1284									
B组一次特征交集：									
1305	404	406	395	1316	1448	1437	1818	1318	1319
403	1384	1451	1450	405	467	1327	1807	468	470
1355	813	1884	1252	1458	1223	834	1454	407	2815
347	541	1421	2881	452	1121	1164	1811	1333	1822
1286	2601	1459	1371	2766	1134	1927	581	300	531
最终二次特征交集									
1305	404	406	395	1316	1448	1437	1319	403	1384
405	1327	1807	468	470	813	2815	347	2881	1121
1811	1459	1371	300						

表 5.3.1-2 医生勾画区域窗口下的特征提取的一次交集与二次交集

结果我们得到 24 个筛选后的特征。将之前特征集元素规模由 300 改为 150,600（分别变为原来的一半和两倍），再进行测试，二次交集结果如下：

规模为600情况下的二次交集
1408 403 392 405 188
规模为150情况下的二次交集
无

表 5.3.1-3 不同规模的二次交集情况

300 和 600 规模下特征比例随着交集次数的变化如下：

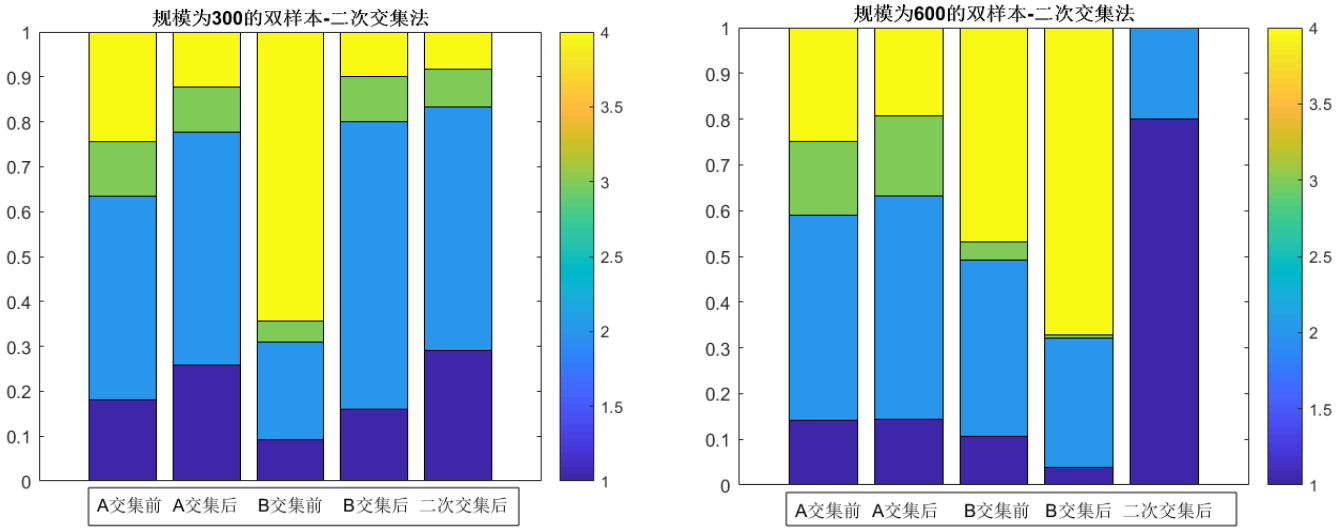


图 5.3.1-4 特征比例变化趋势图，

1 为动脉期二维，2 为动脉期三维，3 为静脉期二维，4 为静脉期三维

以上的结果表明： 规模为 300 时，二次交集有 24 个特征，特征最大编号为 2881，Arterial 期三维特征的占比最大，其次是 Arterial 期二维特征，Venous 期特征占比较小。规模为 600 时，二次交集只有 5 个特征，4 个 Arterial 期二维特征和 1 个 Arterial 期三维特征，无 Venous 特征。规模为 150 时，二次交集元素为 0。在双样本-二层交集法采集下，Arterial 期特征成为判断淋巴结转移相关性的主要特征。

接着我们再考虑单样本-一层交集法，“单样本：mRMR→relief”，“单样本：relief→mRMR”方法，“单样本 mRMR”方法和“单样本-relief”方法。

单样本_一层交集法									
1811	2857	2868	210	188	874	2736	610	2211	2725
470	1807	607	1429	1464	2903	585	1474	13	2331
1371	1862	2067	1380	949	2233	468	822	2916	1086
651	1242	2683	866						
先mRMR后relief									
2541	2921	1126	862	2585	818	439	442	1117	853
809	856	2803	1311	1076	1150	1075	1033	1300	1443
1820	1120	32	1355	1119	1223	2397	1469	1098	2133
1432	1074	1897	1118	854	776	18	1472	1097	1809
2760	1096	541	820	2771	964	15	1062	864	939
先relief后mRMR									
25	2772	87	2834	425	1790	78	1841	454	1422
100	2789	7	1821	2701	791	122	1763	813	1584
2600	466	2212	1606	1768	1327	1628	2807	2698	36
1152	2884	1271	1844	1650	2710	725	2919	56	1801
1854	284	1672	446	23	58	10	1716	1481	1843

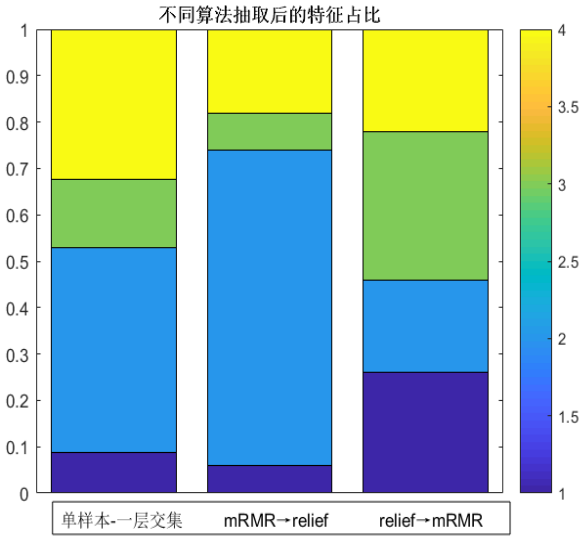


图 5.3.1-5 三种算法下的特征组成情况

1 为动脉期二维，2 为动脉期三维，3 为静脉期二维，4 为静脉期三维

单样本-一层交集法抽取得到特征 34 个，Arterial 期和 Venous 期的特征占比非常接近，其中以 Arterial 期三维特征占比最大，“单样本先 mRMR 后 relief”法抽取得到 50 个特征中，Arterial 期特征占比大于 70%，为主要特征；“单样本先 relief 后 mRMR”法抽取得到 50 个特征中，Arterial 期和 Venous 期特征占比比较接近，Venous 相对比重大一些，其中 Venous 期二维特征占比最大。

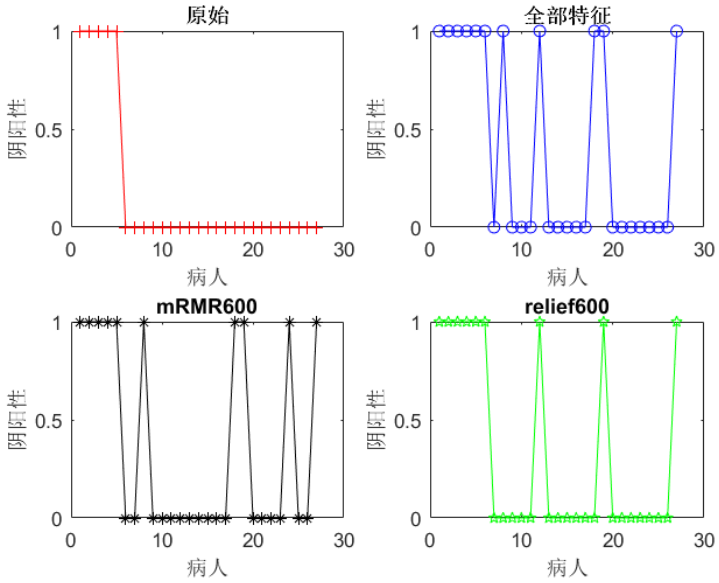
5.3.2 淋巴结转移判断结果

最终我们针对不同的预测模型选取不同的特征抽取原则。

- 选取出 5.3.1 中的三套规模较小的特征抽取结果：双样本二层交集降维得到的以 Arterial 期为主要特征的特征集，单样本一层交集得到的 Arterial 期和 Venous 期特征分布均衡，Arterial 期特征偏多的特征集和单样本先 relief 后 mRMR 得到的 Arterial 期和 Venous 期特征分布均衡，Venous 期特征偏多的特征集来训练 Logistic 回归模型。
- 选取“单样本 mRMR”法和“单样本 relief”法得到的规模较大（600 个特征）的特征集来训练 RF 回归/分类模型。

107 个病人中有 62 个检测出阳性，45 个检测出阴性。抽取阴阳性患者各 40 例作为训练用样本，训练前先将数据标准化。剩下的 27 例作为验证集。Logistic 回归的实现使用 SPSS 17.0，RF 模型的实现使用 MATLAB R2017a。

1）用 RF 回归模型预测。回归模型的预测值一般不会恰好是 0 或 1 的整数，由于定义“阳性”为 1，阴性为“0”，因此我们对预测结果值 $k(0<k<1)$ 进行以下处理：$k<0.5 \Rightarrow$ 阴性$(-)$，$k \geq 0.5 \Rightarrow$ 阳性$(+)$。对于全部特征法，设置决策树 1000 棵，迭代次数为 1000 ;对于 mRMR 和 relief 各自选出的 600 个特征，



	27个病人预测准确率
全部特征	77.80%
mRMR600	81.50%
relief600	85.20%

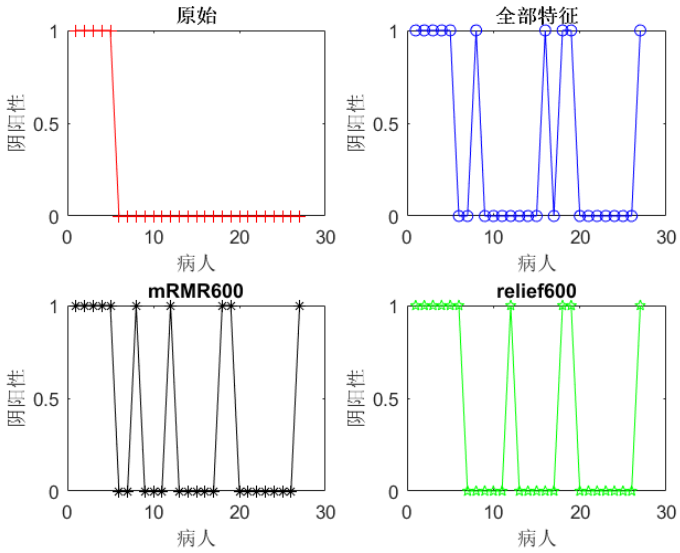
设置决策树 1000 棵，迭代次数为 200 次。迭代次数的确定参照公式 $times = \max\left\{[\frac{feature\_quantity}{3}], 1\right\}$。

RF 回归结果如下：

图 5.3.2-1　RF 回归模型在验证集上的预测效果

结果表明 relief 算法提取的特征在回归预测模型的验证集上有着更好的表现。三种算法在淋巴结转移呈阳性的 5 个病人上都能准确预测，准确率达到 100%；而在淋巴结转移呈阴性的 22 个病人上的预测表现则较差，但总体还是能保持 70%以上的准确率。显然地，三种特征提取法都在预测阳性病人方面具有更大的优势和准确度。

2）用 RF 分类模型预测，预测值为 0 或 1。决策树设置，迭代次数设置同 1)。结果如下：

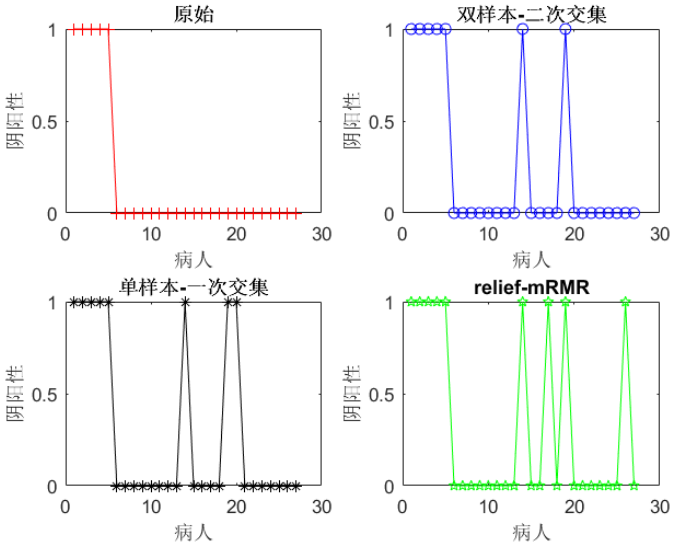


	27个病人预测准确率
全部特征	81.50%
mRMR600	81.50%
relief	85.20%

图 5.3.2-2　RF 分类模型在验证集上的预测效果

结果表明 relief 算法提取的特征在分类预测模型的验证集上有着更好的表现。三种算法在淋巴结转移呈阳性的 5 个病人上都能准确预测，准确率达到 100%；而在淋巴结转移呈阴性的 22 个病人上的预测表现则较差，但总体还是能保持 70%以上的准确率。显然地，三种特征提取法都在预测阳性病人方面具有更大的优势和准确度。这与用 RF 回归模型得到的结果很相似。

3）用 Logistic 回归模型。



	27个病人预测准确率
双样本-二次交集	88.90%
单样本-一次交集	88.90%
relief-mRMR	85.20%

图 5.3.2-3　二元 Logistic 回归模型在验证集上的预测效果

结果表明双样本-二次交集和单样本-一次交集算法提取的特征在验证集上表现优良。三种算法在淋巴结转移呈阳性的 5 个病人上依然能准确预测，准确率达到 100%；而在淋巴结转移呈阴性的 22 个病人上的预测表现则较差，但比 RF 回归和 RF 分类模型的情况好，总体还是能保持 80%以上的准确率。与前面的 RF 回归模型和 RF 分类模型不同的是，这个模型在预测淋巴结转移阳性和阴性问题上有着相近的能力，预测阳性能力稍优。

5.4 评价

5.4.1 模型准确度评价

从外部验证和内部验证来评价，RF 回归，RF 分类和二元 Logistic 回归的表现如下：

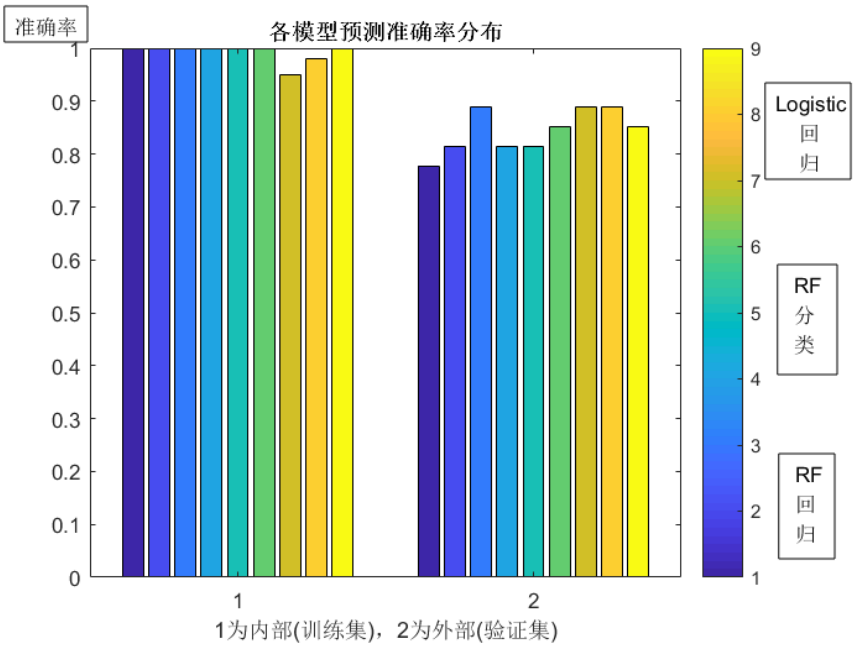


图 5.4.1-1 淋巴结转移预测模型的内外部验证情况

我们观察到使用 RF 算法在训练集内部进行验证时准确率维持着一个很高的水平，但在训练集外部的验证集准确率总体就略微低一些；相比之下，Logistic 算法在训练集内部进行验证时准确率不能达到100%的匹配率，但在训练集外部的验证集准确率总体略高一些。

用 F-Score 指标进行评价。F-Score 的计算公式是 $F\text{-}Score=\frac{2PR}{P+R}$，其中 P 是查准率，R 是查全率。结果如下：

	内部(80)	外部(27)	整体(107)
*RF*回归法1	(1,1)	(0.455,1)	(0.882,1)
*RF*回归法2	(1,1)	(0.5,1)	(0.9,1)
*RF*回归法3	(1,1)	(0.556,1)	(0.918,1)
*RF*分类法1	(1,1)	(0.5,1)	(0.9,1)
*RF*分类法2	(1,1)	(0.5,1)	(0.9,1)
*RF*分类法3	(1,1)	(0.556,1)	(0.918,1)
*Logistic*法1	(0.974,0.925)	(0.714,1)	(0.933,0.933)
*Logistic*法2	(0.952,1)	(0.625,1)	(0.9,1)
*Logistic*法3	(1,1)	(0.556,1)	(0.918,1)

表 5.4.1-2 查准率与查全率表，格式为(P,R)

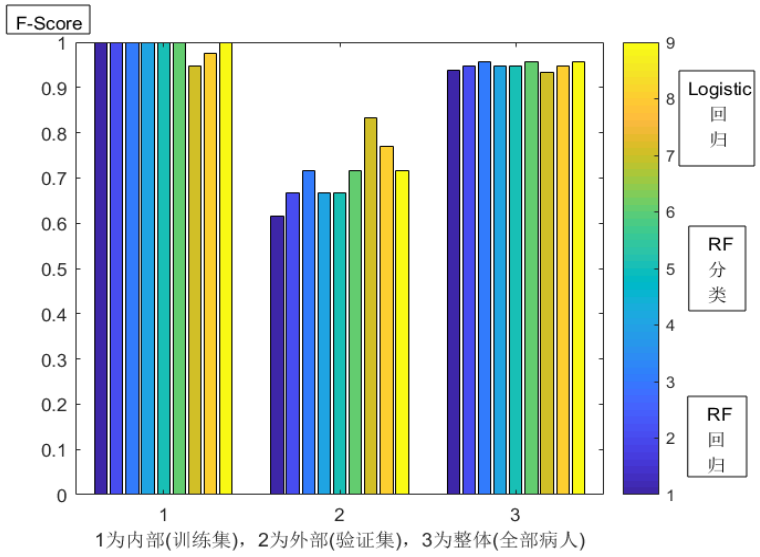


图 5.4.1-3 各模型 F-Score 分布情况

我们观察到训练集内 F-Score 的值普遍很高，这正是用训练集训练模型所产生的结果；其次 F-Score 在整体中表现较好，均有 90%以上；外部验证集表现最差，最低值为 0.616，最大值也才 0.833。外部中的 F-Score 大小对预测精度有着较好的代表性，因此我们更倾向于使用 F-Score 为 0.833 对应的双样本-二次交集 Logistic 回归预测模型。

5.4.2 模型灵敏度评价

模型灵敏度的评价可通过以下几个方面考察：1）肿瘤特征提取窗口的形状和大小变化后模型预测情况的变化，具体可考虑窗口适当增大和缩小的问题；2）肿瘤纹理特征数量的适当减少或增加对模型预测准确度的影响，具体可考虑产生 GLCM（Gray-Level Co-Occurrence Matrix）和 GLRLM（Gray-Level Run-Length Matrix）过程中灰阶变化取向的适当增减和矩阵元素间距方案的适当增减，也可考虑其他用以提取高通量的矩阵 GLSZM(Gray-Level Size Zone Matrix)，NGTDM(Neighbouring Gray Tone Difference Matrix)，GLDM（Gray-Level Dependence Matrix）[9]。本次由于时间限制，我们只从第一个方面进行测试。

除了采用医生勾画的窗口 R_1，我们再采用包围 R_1 的最小矩形窗口 R_2，包围 R_1 的最小凸多边形窗口 R_3，含于 R_1 内的小矩形窗口 R_4，含于 R_1 内的球形窗口 R_5。针对双样本-二次交集 Logistic 回归预测模型和 Relief-RF 分类预测模型比对不同窗口下的预测情况。结果如下：

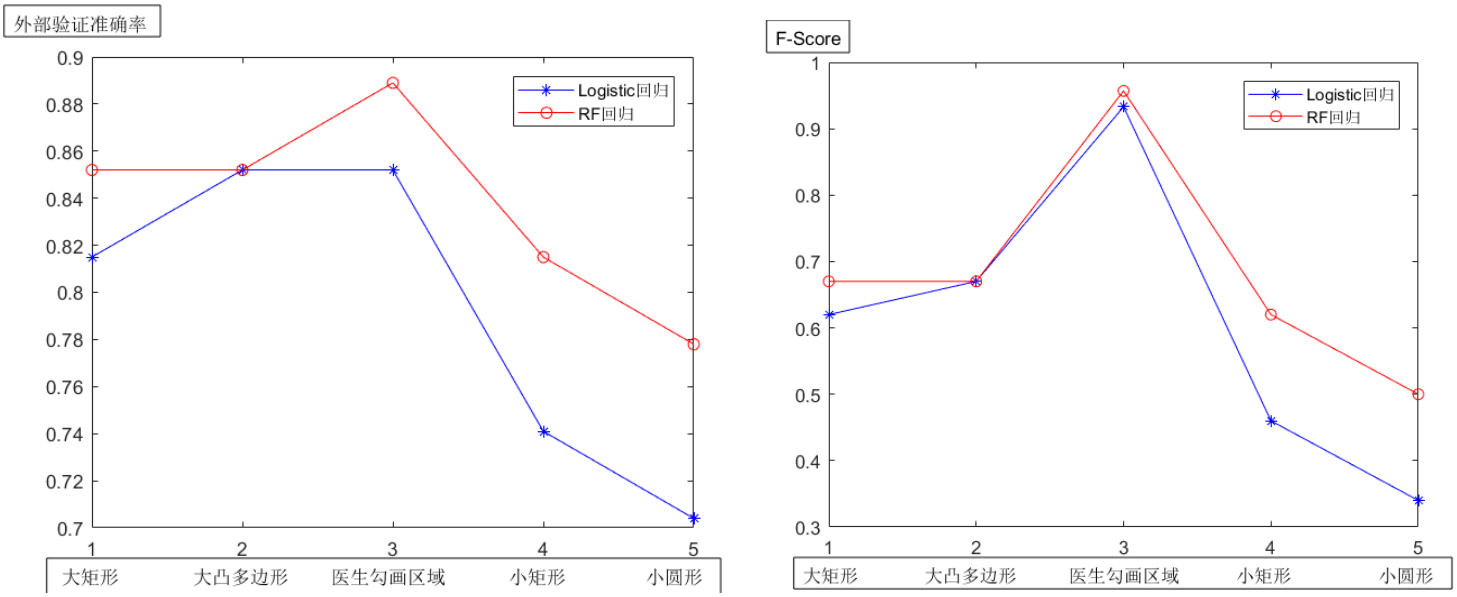


图 5.4.2-1 不同窗口特征提取下的模型外部验证准确率和 F-Score 表现

	大矩形	大凸多边形	医生勾画区域	小矩形	圆形
Logistic	(0.5,0.8)	(0.57,0.8)	(0.933,0.933)	(0.375,0.6)	(0.29,0.4)
*RF*分类	(0.57,0.8)	(0.57,0.8)	(0.918,1)	(0.5,0.8)	(0.43,0.6)

表 5.4.2-2 不同窗口特征提取下的模型查准率和查全率情况，格式为(查准率，查全率)

我们观察到使用医生勾画的窗口效果最佳，两个模型在这个窗口下的外部验证准确率和 F-Score 均有着最好的表现。而当所选窗口适当扩大时，Logistic 模型下的外部验证准确率有略微的下降，F-Score 有着明显的下降；RF 模型下的外部验证准确率在窗口为大凸多边形时不变，窗口为大矩形时有略微下

降，而 F-Score 在这两个窗口也有着明显的下降。当所选窗口适当缩小时，则两个模型的外部验证准确率和 F-Score 都有着明显的下降。由此我们做出判断：当窗口由原来医生勾画的区域发生或大或小的变化时，两个预测模型的表现都会变差，且模型对窗口变小更敏感。预测模型对窗口大小和形状变化的稳定性较差。

6. 展望

- 肿瘤分割部分

我们构造的 LeNet-CRF 和 MIGAN 模型虽然能够达到 70%以上的 Dice 系数，但是也存在一些问题。比如肿瘤细节部分的勾勒未能表现或表现不稳定，一个典型的代表就是肿瘤外轮廓内有个别小区域是不属于肿瘤本身的，但是模型仍旧把那部分归为肿瘤区域。此外，分割得到的肿瘤边界多为不平滑的锯齿状，这一方面是受限于肿瘤边界的体现来自于一个个离散的像素块，但也有模型不细腻的因素。而两个模型 Dice 系数分布在 80%以上的区间的比例都有 10%以上，这也说明模型还有提升精度的空间。针对以上问题我们有一个初步的想法。这个想法是适当加深 CNN 的网络结构，使之具备更强的学习能力，从而能更好地识别和分割肿瘤细节，并且提高分割准确率和 Dice 系数。Google 团队研发的深层次 GoogLeNet 的网络结构可用以借鉴解决深层次网络的一些诸如参数量巨大，网络难以训练的问题。

7. 参考文献

[1]. 魏炜，刘振宇，王硕，等. 影像组学技术研究进展及其在结直肠癌中的临床应用. 中国生物医学工程学报, 2018, 37(5).

[2]. 师冬丽，李锵，关欣. 结合卷积神经网络和模糊系统的脑肿瘤分割. 计算机科学与探索, 2018, 12(4).

[3]. 邢波涛，李锵，关欣. 改进的全卷积神经网络的脑肿瘤图像分割. 信号处理, 2018, 34(8).

[4]. 李健，罗蔓，罗晓，等. 基于多尺度卷积神经网络的磁共振成像脑肿瘤分割研究. 中国医学装备, 2016, 13(2).

[5]. 唐陶富. CT 诊断学. 北京：人民卫生出版社发行部, 2005.

[6]. 张铮，倪红霞，苑春苗，等. 数字图像处理与识别. 北京：人民邮电出版社, 2013.

[7]. Szegedy, Liu, Jia, et al. Going Deeper with Convolutions. Computer Science, 2014.

[8]. Luc, Couprie, Chintala, et al. Semantic Segmentation using Adversarial Networks. Computer Science, 2016.

[9]. Pyradiomics(n.d.).Radiomic Features. From https://pyradiomics.readthedocs.io/en/latest/features.html.

[10]. Aerts, Velazquez, Leijenaar, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach [J]. Nature Communications, 2014, 5:4006.