# 实验3 医学图像特征提取与分类

**实验目的**

1．熟悉并掌握数字图像特征计算方法。

2．掌握影像组学特征的计算。

3．理解分类器的作用。

**实验基础**

1. 影像组学特征计算

影像组学或称为放射组学，是传统医学图像纹理特征提取和分析方法的延申和扩充，它的出现基于实体肿瘤在空间与时间上的异质现象，但这是个极富挑战性的问题。

基于有创活检的检测方法难以客观量化描述出随时间和空间的不同而存在的肿瘤的特点，而影像组学方法可以从医学影像学角度去解决这个问题，它可以无创地检测肿瘤的异质性特点。

影像组学方法通过数据表征算法从医学图像重提取高通量的量化特征，这些特征被命名为影像组学特征。

影像组学的假设是由于疾病形式不同，图像特征能反映其特点，因此可用于预测各种病症的预后和治疗反应，从而为个性化治疗提供有价值的信息。2012年3月，Lambin等在《欧洲癌症杂志》发表文章认为，实体肿瘤在空间和时间上是异质的，这限制了基于侵入性活组织检查的分子测定的使用，但基于影像组学的医学图像分析具有巨大潜力，其具有以非侵入方式捕获肿瘤内异质性的能力。同年，他们在《磁共振成像》杂志上发文指出，影像组学数据是可以挖掘肿瘤内部特征的表现形式，它可用于构建一定量图像特征与表型或基因-蛋白特征关联的预测模型。

影像组学特征一般可分为四类：

1. 灰度特征：是对图像区域或是直方图的描述，均值（mean）、方差（variance）、偏度（skewness）、峰度（kurtosis）、均匀性（uniformity）、能量（energy）、熵（entropy）及直方图百分比区间等对图像直方图也就是一阶的衡量。
2. 形态特征：包括体素体积（voxel volume）、表面积（surface area）、表面积体积比率（surface area to volume ratio）、球形度（sphericity）、紧凑性（compactness）、球形比例（spherical disproportion）等对图像区域形态特点的描述。
3. 纹理特征：定量表述出平滑度、粗糙度和规则性等表面属性。影像组学纹理特征通常基于灰度共生矩阵（gray level co-occurrence matrix, GLCM）和灰度游程长矩阵（Gray-level run-length matrix, GLRLM）计算得到。
4. 小波特征：描述灰度和纹理在图像频域表示下的性质。

更详细和完整的影像组学特征定义可参考：

https://pyradiomics.readthedocs.io/en/latest/features.html

1. 分类器

近二十年来，各类机器学习算法层出不穷，取得了极大成功，已顺利应用于社会的各个领域，带领智能科技迅猛发展。从数据中学得模型得过程称为学习或训练，用于学习模型的数据称为训练数据，一个具体数据实体称为样本。利用训练好的模型对未参与训练的数据进行判断，称为测试，这些数据称为测试数据。如测试结果是离散值，此类学习任务称为“分类”（classification），如测试的是连续值，称为“回归”（regression）。模型学习的过程中，如果训练数据有标记（知道其真实类别或真实值），任务为监督学习，最常用的方法是分类和回归模型；如果没有标记，任务为无监督学习，最常用的方法是聚类模型；与前两者均不同的是仅部分数据有标记，称为半监督学习。分类任务中，预测的类别只有两类即正例和反例，称为“二分类”任务，涉及多个类别时，则称为“多分类”任务。机器学习的目标是使学得的模型能很好地适用于未知数据，也就是包括测试数据在内的未见数据，判断训练好的模型适用于新样本的能力，称为“泛化”（generalization）能力，适用于训练数据的能力为模型拟合性。理论上，我们希望模型的拟合和泛化能力都达到最优，但由于数据样本数量较少、数据存在噪声等问题，这两个目标很难同时达到。如果模型在训练时把训练样本学习得很好，这样很可能把训练数据自身的特点当成所有潜在样本都具有的一般性质，这样反而导致对测试样本表现的下降，也就是泛化性的降低，这种现象在机器学习中称为“过拟合”（overfitting），如果对训练样本学习不够，就是与过拟合相对应的“欠拟合”（underfitting）现象。

模型训练后，需要对模型的分类效果开展评估，常用的方法包括留出法、交叉验证等。常用准确率（accuracy）、敏感度（sensitivity）、特异性（specificity）三个指标量化模型在测试数据上的表现。准确率描述预测和标签一致的样本在所有样本中所占的比例，敏感度描述识别出的所有正例占所有正例的比例，特异性描述识别出的负例占所有负例的比例。在模型训练时，希望得到的模型具有良好的泛化性，需要注意以下几点：第一，模型尽量简单，为模型加入正则项是行之有效的方法之一；第二，尽可能增加训练样本，这样样本和总体的差距得以尽量缩小；第三，减少样本的特征维数，特征挑选将有助于缓解样本数小于特征数的问题；第四，利用交叉验证方法观察模型的泛化性能力。

常用的机器学习模型包括决策树（decision tree）、随机森林（random forest）、支持向量机（support vector machine, SVM）、朴素贝叶斯（naive Bayes），Logistic回归、广义线性模型（generalized linear model）、K近邻（K-nearest neighbors, KNN）、最近邻（nearest neighbors, NN）、神经网络（neural network）等。各个模型各有其优缺点，理想状态应是融入研究领域的先验知识到机器学习模型中，搭建适合问题本身的学习模型，寻求其在计算机中的高速求解方法，并在理论上证明其有效性，在尽可能多的样本上训练模型，并进行测试。

在影像组学研究领域，通过前述方法获得图像的特征子集后，采用或搭建机器学习模型（如支持向量机等），在拆分出的训练样本集合中训练模型，在测试样本集合中评价模型效果，经过严格验证后的模型，即可用于临床诊断。

支持向量机是当前机器学习中效果最好的算法之一，其思想由Vapnik在1995年提出，它巧妙地解决了高维样本快速计算的问题和模型过拟合问题，后者的解决主要依据是最优超平面理论的推广和优化。

SVM有很多软件包可以直接使用，其中流传最广的是LIBSVM软件包。它是台湾大学的林智仁博士等开发设计的一套便于开发和集成的SVM模式识别与回归软件包，可以解决分类（包括C-SVC，nu-SVC）、回归（包括epsilon-SVR, nu-SVR）以及分布估计（one-class SVM）等问题，软件可以通过https://www.csie.ntu.edu.tw/~cjlin/libsvm/网站免费获得。

训练好的机器学习模型可以用于临床诊断。

1. 聂生东，邱建峰。医学图像处理，北京：科学出版社，2020.1

**实验内容与步骤**

1. 了解孕囊超声图像数据。本次实验提供了正常和异常两类，各16例早孕期孕囊超声图像。每一例超声图像都包括未标注的原始图像和人工标注孕囊边缘的图像，可以参照人工标注的孕囊边缘坐标信息，到原始图像中去提取孕囊区域数据信息。
2. 搜索网络资源，利用成熟的软件代码计算影像组学特征。例如，https://pyradiomics.readthedocs.io/en/

latest/features.html

1. 实现基于SVM、决策树、随机森林三种模型的分类模型，并对其分类效果进行对比评估，采用十折交叉验证方法。

**实验作业提交要求：**

1. 保存所有代码文件到一个文件夹下，文件夹命名为实验3源代码-班级-学号-姓名，并由班长统一收集。
2. 按照实验报告模板撰写完成实验报告，文件名为实验3-班级-学号-姓名.doc/docx，并由班长统一收集。
3. 实验报告提交时间截止到实验课结束一周，过期按未提交作业处理。