第七届“泰迪杯”数据挖掘挑战赛——

B 题：直肠癌淋巴结转移的智能诊断

一、问题的背景

直肠癌是指从齿状线至直肠乙状结肠交界处之间的恶性肿瘤，是消化道最常见的恶性肿瘤之一。近几年在中国，直肠癌的发病率越来越高，特别在一些大城市，它已经跃居至恶性肿瘤发病率排行榜前三位。直肠癌易向肠外浸润并发生淋巴结及远处转移，一旦发生转移病人常常需先进行辅助放化疗才能获得手术机会，患者预后较早期直肠癌患者的预后差。直肠癌患者是否有淋巴结转移对治疗方案的决策以及病人预后有重要的影响，因此对是否有淋巴结转移的准确判断是直肠癌治疗的重要步骤，但目前尚无一种方法能在术前准确地判断淋巴结转移情况。直肠癌肿瘤本身的特性和周围淋巴结转移存在一定的关联性，所以本问题期待参赛者能够设计出有效的算法通过对直肠癌 CT 影像特征的判断来对淋巴结转移情况进行评估，提高影像学对淋巴结转移判断的准确性。

二、任务

本问题提供直肠肿瘤病人的动脉期和门脉期两类影像数据，参赛者可选择其中的一类或两类影像开展研究。

1. 直肠肿瘤分割

在 CT 动脉期和门脉期增强图像上，肿瘤区域和周围组织在强度上存在差异。以数据集 1 中提供的 CT 影像和医生标记出的直肠肿瘤掩模为训练样本，设计图像分割算法，分割出直肠肿瘤所在的区域。由于 CT 图像是一个连续的扫描序列，参赛者需要给出每个病例每幅图像中标记肿瘤区域的掩模图像（掩模图像的解释见第三部分数据集说明）。

2. 直肠肿瘤部分特征的提取

提取 CT 图像中直肠癌肿瘤区域的影像特征（参见[1]），可以是平面的二维特征如形状、面积、强度、纹理、小波系数等，也可以是肿瘤部位的三维特征如体积、表面积等，也欢迎提出新的影像特征。

3. 肿瘤影像特征与淋巴结转移的相关性验证

所给数据中包含淋巴结转移阴性和阳性的两类病例，请分析直肠肿瘤区域影像特征与是否淋巴结转移之间的关系，基于 CT 影像数据建立分类模型并评估模型的有效性。可适当引入病人临床数据（如性别、年龄、实验室指标等）。

直肠肿瘤的影像学研究中，精确分割肿瘤区域一直是实践中的一个难点。本问题拓展的一个方向是通过实验来验证当所选取区域适当扩大（例如覆盖一部分肿瘤的周边组织）或适当缩小（例如只选取肿瘤的部分核心区域）时，是否仍然能构建有效的模型来判断肿瘤是否已发生淋巴结转移。

三、数据集说明

赛题数据包括临床数据和 CT 影像数据。临床数据保存在 EXCEL 表中，包含病人 ID、性别、年龄、是否淋巴结转移等字段。CT 影像方面，提供每个病人的 DCM 格式的腹部横断位动脉期和门脉期两种增强 CT 影像数据，分别保存在以病人 ID 命名的文件夹下的“arterial phase”和“venous phase”子文件夹中。

用于模型训练的数据集 1 包含多位患直肠癌病人的临床数据和相应的 CT 图像文件（格式为 dcm），以及与每幅 CT 图像对应的由医生标记出的直肠肿瘤的掩模图像文件（在 CT 图像文件名后加“_mask”，格式为 png，例如：如果 CT 图像文件为“10066.dcm”，对应的掩模图像文件为“10066_mask.png”）。掩模图像是一个与 CT 图像具有同样分辨率的二值图像，像素取值为 1 表示该像素属于目标（肿瘤区域），像素取值为 0 表示该像素属于背景（非肿瘤区域）。如果某幅 CT 图像中不存在直肠肿瘤，则对应的掩模图像为全黑。数据集 1 将于 2019 年 4 月 13 日在竞赛官网发布。

用于模型测试的数据集 2 的临床数据中没有标识病人是否淋巴结转移的字段，也不提供直肠肿瘤的掩模图像。参赛者需提交每个病人每幅 CT 图像对应的肿瘤掩模图像以及该病人是否发生淋巴结转移的预测结果。数据集 2 将于 2019 年 4 月 27 日上午 9:00:00 在竞赛官网发布。

四、评价方案

1. 分割的评价指标

图像分割采用 Dice 系数进行评价, 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：

$$\mathrm{Dice}(A,B)=\frac{2|A\cap B|}{|A|+|B|}$$

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。Dice 系数的取值范围是[0,1]，取值越接近 1 表明直肠肿瘤分割的结果与医生给出的结果越接近。

2. 淋巴结转移预测的评价指标

使用 F-Score 对分类结果进行评价：

$$F=\frac{2PR}{P+R}$$

其中P为查准率（Precision），R为查全率（Recall）。

五、DCM 格式

DCM 文件是遵循 DICOM（DICOM：Digital Imaging and Communications in Medicine，医学数字成像和通信）标准的一种文件。DICOM 标准支持的设备包括心电图、核磁共振成像、血管镜、超声心动图等多种医疗设备，因而 DCM 文件被广泛应用于医疗行业。DCM 格式文件可以用 MedImaView、Millensys DICOM MiniViewer、DICOM Viewer 等软件打开。MATLAB 和 Python 中都有专门的函数用于读取 DCM 文件。

参考文献

[1] 魏炜、刘振宇、王硕、田捷，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，2018，37:513-520.

附录：

请仔细阅读以下说明：

1、 关于赛题数据

① 示例数据：2019 年 3 月 16 日随赛题公布。

② 全部数据：2019 年 4 月 13 日公布。

③ 测试数据：2019 年 4 月 27 日 9:00:00 公布。

2、 提交作品

① 命名方式：论文命名为“B 题”，附件命名为“作品附件”，测试结果命名为“作品测试结果”。

② 论文及附件内请勿出现队号、学校、学院、队员以及指导老师相关任何信息，否则该作品视为无效作品。

③ 请参赛队于 2019 年 4 月 26 日 16:00 之前在竞赛官网【提交作品】处提交论文（PDF 版，大小不超过 50M）及附件（论文正文（Word 版）、源数据（组委会提供的源数据除外）、过程数据、程序的压缩包，大小不超过 200M）。

3、公布测试数据，提交测试结果

2019 年 4 月 27 日 9:00:00 准时放出测试数据，请在"赛题与数据"页面对应的题目右下方下载测试数据，并于 2019 年 4 月 28 日 9:00:00 前请在“提交测试结果”页面提交测试结果。