Lung Nodule Classification Project

背景介绍

> Lung cancer is the leading cause of cancer-related death worldwide. Lung cancer screening programs using low-dose CT are being implemented in the United States and other countries. Computer-aided detection (CAD) of pulmonary nodules could play an important role when screening is implemented on a large scale.

计算机自动辅助检测系统 (CAD) 的设计目标是根据肺部CT扫描图像，借助计算机系统自动检测出肺结节(pulmonary nodules)。 类似于目标检测 (object detection) 等机器学习领域的过程(pipeline)，检测系统会首先提取出一些甚至大量的候选区域 (candidates or proposasl)，并使用这些候选结节进行下一步的分类或者其他任务。

Lung Nodule Classification Project 的目标是对给定的候选结节区域 (candidate or proposal) 进行分类，判定它们是否是肺结节。

术语

- CT

  CT的全称是计算机层析成像 (computed tomography)，其成像原理主要是使用射线照射人体某一部分具有一定厚度的层面，并用探测器接收透过的射线。不同部位的密度会影响射线穿透的量。探测器接收到的物理信号会经过图像处理的算法进行重建，得到每一层的计算机图像。适度了解一些关于CT的知识有助于你研究这个问题，你可以参考*维基百科*和*XRayPhysics*。

- slice

  原始的CT数据经过重建会形成每一层的CT图像（二维），每一层图像代表肺的一个切面，称作 *slice*。不同设备产生的CT图像的层厚（每个切面的实际厚度）有差异。层厚的具体意义请参考CT部分给出的相关链接。不过关于CT的相关参数信息，包括层厚，都不影响你解决这个问题，你完全可以选择忽略，而使用我们经过选择并统一化过的数据。

- scan

  scan 指一位被检测者的CT图像(*slice*)组成的集合。举例来说，一位被检测者的肺部CT扫描可能有300多张图像(*slice*)，这些图像的集合就是一份CT scan。换言之，*slice* 是一个二维的图像，而若干 *slice* 按序排列形成一个三维图像，即 *scan*。

- HU

  HU值全称为 Hounsfield scale，具体可以参见*维基百科*。简单的说，HU值是CT中各部分衰减程度的度量标准，我们提供的数据都是以HU值为单位。不过你可以将CT图像认为是单通道或者灰度图像，只不过其值域并非是[0,255]。

- candidate

  在图像处理领域，由于各种原因，算法通常只能作用在图像的一个局部。因此，算法设计者会先在图像上选出一些候选区域以训练分类器或者在算法中作为正样本的候选者，这些候选者称为 *candidate*，你的任务就是实现对这些candidate的分类。

关于数据

数据集

我们提供的训练集共包含723个scan。对于所有scan（训练集和测试集），我们已经根据数据提供的信息进行了插值，使得所有图像都是各向同性的，即每个像素都代表一个1立方毫米的立方体。如果你需要有关插值的信息甚至是尚未插值过的原始数据，请联系我们。我们在经过插值的CT图像基础上，提取出 *candidates*（平均每个 *scan* 的正负样本相差几十倍），你的目标就是分类这些 *candidates*。

> 注：本项目不会公开最后的测试数据。

数据格式

我们提供的训练数据以 *scan* 为单位，每个 *scan* 对应一个单独的文件夹，其中仅包含一个文件 caididates.mat，

- caididates.mat(matlab文件)——包含该scan所有 *candidate* 信息的结构体数组(行向量)。每个结构体(struct)的字段有：
  - CANDIDATE_ID: 每个candidate的全局ID;
  - NODULE_ID: 当前scan中该结节的ID。由于可能存在多个正样本对应一个真实结节(nodule)，你可以通过这个ID来选择不同的真实结节。负样本该值为0。
  - VOL: candidate在三维CT图像中对应的40x40x40区域的HU值（已经过插值，保证每个体素代表1立方毫米的立方体，并且HU值通过线性操作保证非负）。存储为single类型。

- LABEL：1表示正样本，0表示负样本。

另外，我们提供一个全局的train_ground_truth.mat，该文件用于你计算自己分类器的AUC（代码已提供）。

正负样本说明

正样本的中心点在真实结节内部或者离真实结节中心2.5毫米内。负样本则不满足上述性质，但可能包含真实结节的一部分。

评价方式

测试数据的格式和训练数据完全一致，但我们不会公布测试数据。

你最后需要提交的有：程序代码、算法报告、最终版本的分类器(classifier)、只需更改路径就能运行分类器的脚本文件以及运行你的分类器所需要的环境配置（不限制语言及框架，但推荐使用C,C++,Matlab,Python,Lua等机器学习框架常用的语言）

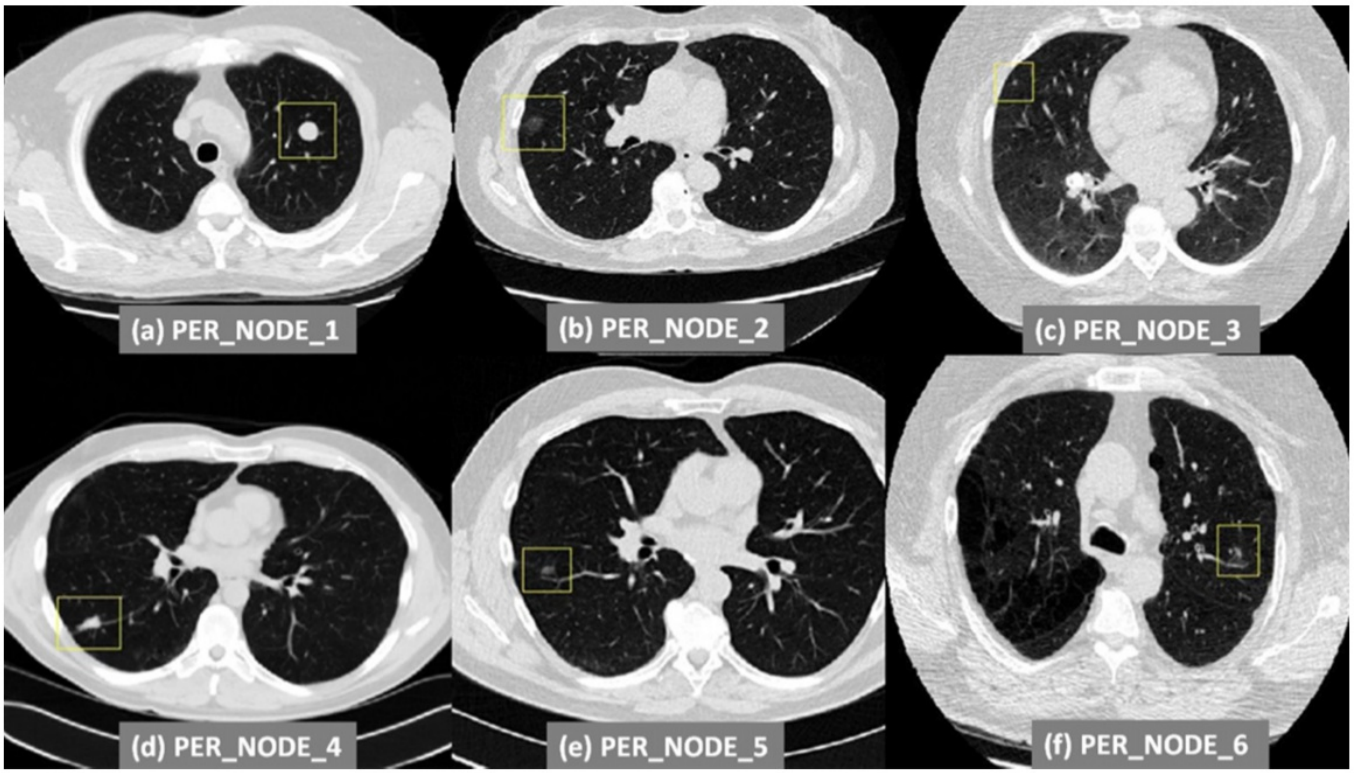
我们目前支持的机器学习框架有：Caffe,Tensorflow,Keras,Torch,Theano。建议你使用这些框架，如果你使用的是除这些以外的框架请记录好你安装框架的流程，我们可能需要请你帮助安装你用的框架。

你提供的运行脚本需要是只用更改数据目录和你的结果的输出目录就可以运行（你训练的时候是train_data文件夹，给我们测试时只需要将这个文件夹路径改为test_data)。你提交的分类器应该输出每个 *candidate* 是否是结节的置信度(confidence)，取值为[0,1]，存储格式为mat文件。该mat文件需要包含一个N*2的数组，N代表测试数据的个数，第一列记录CANDIDATE_ID（升序），第二列记录置信度，具体格式可以参看我们给出的AUC计算程序。该程序会根据置信度计算全局所有结节对应的AUC来进行评估。

文件列表

目录名	描述
train_data	训练数据目录，测试数据的格式完全一致
raw_data	10份完整的CT图像和相应的matlab看图工具
AUC.m	计算AUC的matlab脚本
readme.pdf	说明文档

结节知识补充



图片表示六种不同的结节类型

a. 孤立大结节，呈孤立球状，大部分结节为类似的球状但大小可以有变化。

b. 贴肺壁的磨玻璃结节，亮度较低

c: 肺壁磨玻璃小结节，结节直径较小且亮度较低。

d: 肺膜处结节，结节生长在肺膜上（肺膜为结节下面那一条线），形状不近球形。

e: 血管处磨玻璃结节，结节生长在血管附近（但可能有血管会穿过结节）。

f: 血管处结节。

我们提供的raw_data文件夹里面包含了10个scan 的完整三维CT图像 *rawVol* 和逐像素标记的结节位置 *rawLabel* 以及对应的看图程序。如果你觉得有必要，可以依照看图程序的输入要求和操作方法查看scan里的结节。

联系人

- steve_gao@pku.edu.cn 高俊
- gujiayuan@pku.edu.cn 顾家远
- 1400012710@pku.edu.cn 张梦晓