Student: Edwin Meyers
Date: 25-Jul-20

Machine Learning Engineering Nanodegree
Udacity

Capstone Proposal:
Classifying the severity of prostate cancer from microscopy scans of prostate biopsy samples using different machine learning techniques

Dear reader,

This Capstone Project is based off of a challenge from Kaggle.com named “Prostate cANcer graDe Assessment (PANDA) Challenge” (link).

Please find below details regarding how I plan to approach solving the challenge:

1. Domain background[1]:

The project is derived from the field of pancreatic cancer research. Radboud University Medical Center and Karolinska Institute teamed up to organize a Kaggle competition in collaboration with their colleagues from Tampere University. The Computational Pathology Group (CPG) of the Radboud University Medical Center is a research group that develops computer algorithms to aid clinicians. Karolinska Institute’s Department of Medical Epidemiology and Biostatistics (MEB) includes an interdisciplinary research group to improve the diagnostics and treatment of prostate cancer. Together, they hope to further their existing research to make a significant impact on the healthcare of prostate cancer patients.

Kaggle challenge organizer team: Wouter Bulten, Geert Litjens, Hans Pinckaers, Peter Ström, Martin Eklund, Lars Egevad, Henrik Grönberg, Kimmo Kartasalo, Pekka Ruusuvuori, Tomi Häkkinen, Sohier Dane, Maggie Demkin.

2. Problem statement[2]:

With more than 1 million new diagnoses reported every year, prostate cancer (PCa) is the second most common cancer among males worldwide that results in more than 350,000 deaths annually. The key to decreasing mortality is developing more precise diagnostics. Diagnosis of PCa is based on the grading of prostate tissue biopsies. These tissue samples are examined by a pathologist and scored according to the Gleason grading system. There are times however, when two pathologists may diagnose two sample tissues differently, which could have potentially life-impacting effects on patient treatment and outcome.

3. Datasets and inputs[3]:

File: [train/test].csv

- image_id: ID code for the image.
- data_provider: The name of the institution that provided the data. Both the Karolinska Institute and Radboud University Medical Center contributed data. They used different scanners with slightly different maximum microscope resolutions and worked with different pathologists for labeling their images.
- isup_grade: Train only. The target variable. The severity of the cancer on a 0-5 scale.
- gleason_score: Train only. An alternate cancer severity rating system with more levels than the ISUP scale. For details on how the gleason and ISUP systems compare, see the Additional Resources tab.

---

[1] From the Kaggle challenge description
[2] From the Kaggle challenge description
[3] From the Kaggle challenge description

Folder: [train/test]_images
- These are the images. Each is a large multi-level tiff file. There are roughly 1,000 images in the hidden test set. Slightly different procedures were in place for the images used in the test set than the training set. Some of the training set images have stray pen marks on them, but the test set slides are free of pen marks. Each individual image is quite large. It should be interesting to explore strategies that can efficiently locate areas of concern to zoom in on.
- The labels are imperfect. This is a challenging area of pathology and even experts in the field with years of experience do not always agree on how to interpret a slide. This will make training models more difficult, but increases the potential medical value of having a strong model to provide consistent ratings. All of the private test set images and most of the public test set images were graded by multiple pathologists, but this was not feasible for the training set. You can find additional details about how consistently the pathologist's labels matched here.

Folder: train_label_masks:
- Segmentation masks showing which parts of the image led to the ISUP grade. Not all training images have label masks, and there may be false positives or false negatives in the label masks for a variety of reasons. These masks are provided to assist with the development of strategies for selecting the most useful subsamples of the images. The mask values depend on the data provider:
  - Radboud: Prostate glands are individually labelled. Valid values are:
    0: background (non tissue) or unknown
    1: stroma (connective tissue, non-epithelium tissue)
    2: healthy (benign) epithelium
    3: cancerous epithelium (Gleason 3)
    4: cancerous epithelium (Gleason 4)
    5: cancerous epithelium (Gleason 5)
  - Karolinska: Regions are labelled. Valid values are:
    0: background (non tissue) or unknown
    1: benign tissue (stroma and epithelium combined)
    2: cancerous tissue (stroma and epithelium combined)

File: sample_submission.csv: A valid submission file.

4. Solution statement

The objective is to develop models for detecting PCa on images of prostate tissue samples, and estimate severity of the disease using the most extensive multi-center dataset on Gleason grading yet available. Having a strong model to provide consistent ratings holds important medical value in assisting doctors in making their final diagnosis of the condition of a patient's pancreas.

5. A benchmark model

This Capstone project is based on a Kaggle competition. Kaggle has a leaderboard that displays the precision of other participants' models. Please find the leaderboard in the following url: *https://www.kaggle.com/c/prostate-cancer-grade-assessment/leaderboard*

6. Evaluation metrics[4]

The submissions must include predictions of the `isup_grade` for each `image_id`. Submissions are scored based on the quadratic weighted kappa, which measures the agreement between two outcomes. This metric typically varies from 0 (random agreement) to 1 (complete agreement). In the event that there is less agreement than expected by chance, the metric may go below 0.

The quadratic weighted kappa is calculated as follows. First, an N x N histogram matrix *O* is constructed, such that *Oi,j* corresponds to the number of `isup_grades` *i* (actual) that received a predicted value *j*.

An *N-by-N* matrix of weights, *w*, is calculated based on the difference between actual and predicted values:

$$w_{i,j} = \frac{\left(i-j\right)^2}{\left(N-1\right)^2}$$

An *N-by-N* histogram matrix of expected outcomes, *E*, is calculated assuming that there is no correlation between values. This is calculated as the outer product between the actual histogram vector of outcomes and the predicted histogram vector, normalized such that *E* and *O* have the same sum.

From these three matrices, the quadratic weighted kappa is calculated as:

$$\kappa=1-\frac{\sum_{i,j}w_{i,j}O_{i,j}}{\sum_{i,j}w_{i,j}E_{i,j}}.$$

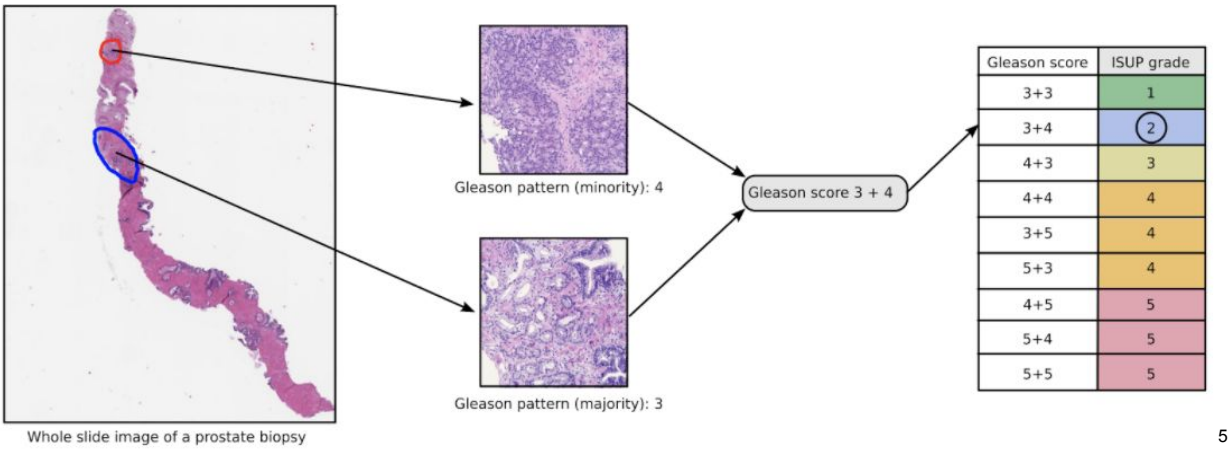
7. Project design outline:

Preliminarily speaking, the solution is expected to be reached by implementing the following steps:

Step 1: Defining areas of interest in the biopsies:

Segmentation masks show which parts of an image led to an ISUP grade. The initial objective will be to use these masks to train a model that will help identify areas of interest that merit additional study. Hopefully, in this stage we will be able to preliminarily discard benign from non-benign tissue sections so as to take a first step in narrowing possible diagnoses.

[4] From the Kaggle challenge description



5

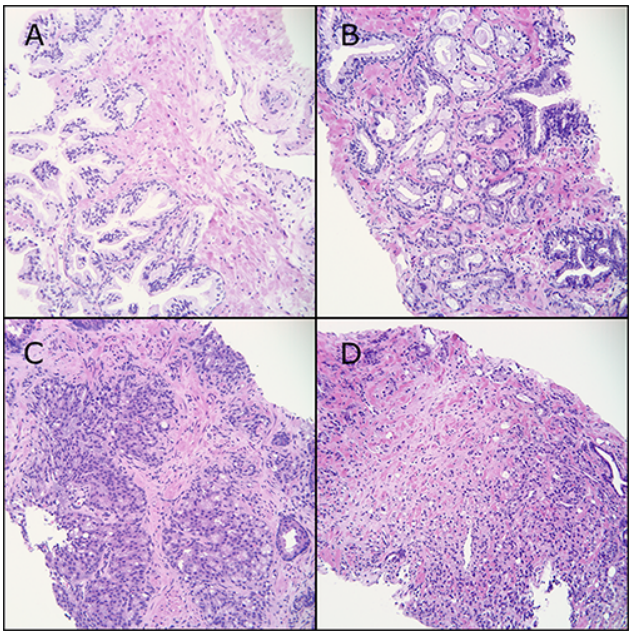
Given that only a portion of the images have their corresponding segmentation masks, we will likely have to rely on unsupervised learning methods to train a model to recognize areas of interest. Preliminarily speaking, it seems that using clustering methods would be a good way to identify areas of interest. Color and formational features of the biopsies will serve as good input for a clustering model to correctly identify areas of interest. Looking forward, the model that will be used to identify areas of interest could be replaced with a potentially better model by using the method described in Step B (iii) (further below)

Step 2: Zooming in on areas of interest to train models:
Once the areas of interest are defined, we will train a new set of models, using both masked and unmasked pictures, dedicated to providing a Gleason Score for the areas of interest. Several models should be evaluated, preliminarily a neural network model (e.g., vgg16) and a Random Forest classifying model.

Detailed examples of biopsy close-ups and their respective Gleason indices (aka ISUP) (in reference to picture of 4 biopsies below, obtained from Kaggle competition's website):

A. Benign prostate glands with folded epithelium. The cytoplasm is pale and the nuclei small and regular. The glands are grouped together.
B. Prostatic adenocarcinoma - Gleason Pattern 3 has no loss of glandular differentiation. Small glands infiltrate between benign glands. The cytoplasm is often dark and the nuclei enlarged with dark chromatin and some prominent nucleoli. Each epithelial unit is separate and has a lumen.
C. Prostatic adenocarcinoma - Gleason Pattern 4 has partial loss of glandular differentiation. There is an attempt to form lumina but the tumor fails to form complete, well-developed glands. This microphotograph shows irregular cribriform cancer, i.e. epithelial sheets with multiple lumina. There are also some poorly formed small glands and some fused glands. All of these are included in Gleason Pattern 4.
D. Prostatic adenocarcinoma - Gleason Pattern 5 has an almost complete loss of glandular differentiation. Dispersed single cancer cells are seen in the stroma. Gleason Pattern 5 may also contain solid sheets or strands of cancer cells. All microphotographs show hematoxylin and eosin stains at 20x lens magnification.

[5] Picture from the Kaggle challenge description



Step 3: Hyperparameter tuning to improve desired prediction outcomes:
Self explanatory

Additional pre-/post-processing steps should be taken in parallel when working with this data:

Step A: Given that there are two data sources (ie. Radbound and Karolinska) with slightly differing image qualities, special consideration should be taken. Mixing both sources when training the data may potentially be such a consideration that should be taken to make sure there are no biases implicit to image qualities that could potentially affect the quality of a prediction. Additionally, it may be beneficial to rotate and flip the images to help the model, especially if using neural networks, so as not to overfit.

Step B: Given that there two doctors can give different diagnoses, special consideration should be taken. Training a model with existing doubt on what is ground truth will be challenging and could require a combination of (i) eliminating pictures with doubtful ground truths before training models, (ii) creating average Gleason scores for conflicting diagnoses before training models, (iii) once a first model is created, use it to predict label on both test and train data, and only keep pictures that were correctly predicted by the model, with a given error margin. Then use those pictures to train a new model that will theoretically be trained on pictures that are known to generate greater consensus between pathologists and the predictor.[6]

Step C: If time permits, it would be useful to build the model in the cloud and to create an application that would allow users to upload the images directly through an internet portal.

[6] This idea is borrowed from the Kaggle competition's winning team's strategy (link)