Machine Learning End Of Year Project Repport

Arnaud Martin

Dataset overview

I choosed to work with the image (colorectal-histology) dataset. The dataset was released with a paper (Kather, 2016) [1]. It is composed of 5,000 images of colon divided in eight classes. The classes are:

1. tumor epithelium,
2. simple stroma
3. complex stroma
4. immune cells
5. debris and mucus
6. mucosal gland
7. adipose tissue
8. background

Author's work overview

Authors converted images to grey scale. And developed a set of problem specific features divided in 6 broad categories that we won't detail here. Then they used four classifiers 1-NN, ensmble tree, linSVM and rbfSVM. And they defined two types of of problems. The first is multiclass problem where they try to predict the label of the image. The second is the conventional problem where they try to predictif if the patient has a cancer or not.

Multiplying feature sets, classifiers and problems there are 48 models authors compare using error rates. Error rates varies from 4% to 50%.

Problem Definition

The point of the project is to showcase how to improve (& build) a predictive model using these images.

In my non-expert understanding, stroma (simple or complex) [2], immune cells [3], debris and mucus [4], mucosal glands [5], adipose tissue [6] and background are not conditions. The only class that require special attention is the first one [7].

I want to build (a little bit) on the authors work. I assume in a real scenario with such an application a doctor will confirm any positive. So I think error rate is not the most important metric for the patient. So I will try to make a model that prioritize not missing tumor prediction even at the expanse of predicting other classes as tumor.

My plan is to first build a model that the authors did not explore a CNN. Then to compare it's error rate to the authors models. Then depending on the time I have I will either try to improve it and compare it's recall to my model or try a model made by the authors and compare it's recall to my model.

Bibliography & Sources

[1] - Kather, J., Weis, CA., Bianconi, F. *et al.* Multi-class texture analysis in colorectal cancer histology. Sci Rep **6**, 27988 (2016). https://doi.org/10.1038/srep27988

[2] - https://en.wikipedia.org/wiki/Stroma_(tissue)

[3] - https://en.wikipedia.org/wiki/Nonspecific_immune_cell

[4] - https://www.frontiersin.org/articles/10.3389/fcimb.2020.00248/full

[5] - https://en.wikipedia.org/wiki/Mucous_gland

[6] - https://en.wikipedia.org/wiki/Adipose_tissue

[7] - https://en.wikipedia.org/wiki/Epithelioma