Proposal for 6.862 project

Shaoling Han | shaoling@mit.edu

Project background

Dermatosis is a major health burden for people nowadays. For example, according to wikipedia, skin cancer is the most common form of cancer in United State(1). Early detection, distinguishing and diagnosis of skin cancers and benign dermatosis are of high importance. The main method of dermatosis diagnosis is based on observation of skin. In this project, I will explore some machine learning methods that can help efficient diagnosis/detection of skin disease based on medical skin images.

Data description

The data used in this project is called **HAM10000** ("Human Against Machine with 10000 training images"), which consists of 10015 dermatoscopic images and labels of these images, including patients' ID, image ID, ground truth disease classification within one of these 7 categories: Actinic keratoses and intraepithelial carcinoma / Bowen's disease, basal cell carcinoma, benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses), dermatofibroma, melanoma, melanocytic nevi and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage), method of obtaining ground truth (histopathology/follow-up/expert consensus/confocal microscopy), age, gender and location of dermatosis(2). There are 58 records with missing data of age. Each image is of about 200 kb size. To process the data, images will be represented as coordinate of pixel and its color vector in RGB. The ground truth will be used as label, and the method of obtaining ground truth could indicate how difficult this image is for human experts to judge.

Problem description

The problem here is a supervised classification problem. The main input is the image and corresponding label, and the output is the prediction of the disease category. To evaluate the prediction models, classification accuracy could be used.
I also plan to perform clustering with these images: use images and ignore the labels, and see whether the model can find the pattern of the diseases without supervision. Silhouette score and gap statistics will be used for evaluation of clustering.

Methods and models

For classification, a simple baseline model will be multinomial logistic regression, using the image information, age, gender, and even location of dermatosis to predict category of dermatosis. The loss function will be cross entropy. Candidate models for classification include k-nearest-neighbor (a non-parametric model) and Convolutional Neural Network (a parametric model), and the loss function is also cross entropy.
KNN basically calculates distance between a new image and all image in training dataset, select the k nearest neighbors and gives the most frequent category. CNN

basically takes the spatial information between pixels into account and extracts features through several rounds of convolution and pooling. After reduction of the high dimensions, the “flattened” features could be used to train a neural network. For clustering, we can use the features obtained from CNN as datapoints and perform k-means/ Hierarchical clustering/DBScan, and compare the clustering to the expected 7 categories.

- **Computational resources**

I will use my laptop for EDA and data processing, and use AWS cloud for complicated modeling. I will use Tensorflow and sklearn in python for modeling.

- **Relevant work**

This dataset has been released for over one year and many people have explored it before(3-5). It was also used as a part of ISIC 2018 challenge(6). These works mainly focus on classification problem using deep learning, and rarely pay attention to other methods. I think exploring different methods can help me better understanding the pros and cons of the methods.

- **Timeline**

Mar 20: run models for classification (k-means and CNN) and compare performance.
Apr 2: tune classification models and summary for intermediate report.
Apr 15: run models for clustering.
Apr 25: tune models for clustering.
May 5: compare all methods, search for literatures and finish discussion.
May 7: final report due.

- **Potential risk**

One potential risk is that it may take longer time than expected when training models, due to complexity and high-dimention of imaging data. To mitigate this, I need to setup cloud service early, run less intensive tasks locally and more intensive tasks online simutanously.

- **Reference**

1. . Available from: https://en.wikipedia.org/wiki/Skin_cancer#Epidemiology.
2. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Available from: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T.
3. Tschandl P, Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data. 2018;5(180161).
4. Tschandl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. The Lancet Oncology. 2019;20(7):938-47.

5. Skin Cancer MNIST: HAM10000. Available from: https://www.kaggle.com/kmader/skin-cancer-mnist-ham10000.

6. ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection. Available from: https://challenge2018.isic-archive.com/task3/test/.