BST270 Project Proposal

An Implementation of Cluster-adaptive Active Learning with Application to Diabetic Retinopathy Diagnosis

Linglin Huang

1. Data Description

For this project, I will use both simulated datasets as positive and negative controls and a real dataset for diabetic retinopathy diagonisis (DR dataset). The DR dataset is a public dataset on UCI Machine Learning Repository (Link to dataset). This dataset was donated to the repository in 2014, and was preprocessed (Antal and Hajdu 2014) from images in Messidor database (the original data was collected during 2005 to 2006). The Messidor database consists "1200 eye fundus color numerical images of the posterior pole for the Messidor database were acquired by 3 ophthalmologic departments using a color video 3CCD camera on a Topcon TRC NW6 non-mydriatic retinograph with a 45 degree field of view"(Decencière et al. 2014).

The DR dataset contains 1151 samples with 19 covariates and 1 outcome. Specifically, the outcome is a binary class label for diagnosis of diabetic retinopathy (contains signs of DR or no signs of DR). The covarites include(Lichman 2013):

1. A binary result of image quality assessment, where 0 = bad quality 1 = sufficient quality.
2. A binary results of pre-screening, where 1 indicates severe retinal abnormality and 0 its lack.
3. The results of Microaneurysms (MA) detection (7 features). Each feature value stand for the number of MAs found at the confidence levels alpha = 0.5, . . . , 1, respectively.
4. The results of exudates detection (7 features). Similar to features for MA detection.
5. The euclidean distance of the center of the macula and the center of the optic disc, normalized with the diameter of the ROI.
6. The diameter of the optic disc.
7. The binary result of the AM/FM-based classification.

2. Project Goal

In this project, I will implement the cluster-adaptive active learning algorithm proposed by Dasgupta and Hsu (2008), test its performance with simulated datasets (positive and negative controls), and apply it to the DR dataset to construct a classifier for automatic DR screening. Active learning algorithms are particularly useful when it is easy to get unlabeled data but hard to obtain the labels, which is a common scenario in medical science research. And in this project, I will use DR diagnosis as an example to evaluate the efficiency of cluster-adaptive active learning algorithm when applied to biomedical data.

3. Programming Language and Platforms

I plan to use Python for algorithm implementation and R for performance analysis and visualisation. The reports and code will be integrated in R Notebook.

4. Planned methods and analysis

I will implement the cluster-adaptive active learning algorithm with logistic regression as base learner. This model should be appropriate because I have a binary outcome and multiple numerical/categorical covariates.

To evaluate the algorithm, I plan to:

1. Run simulations with positive and negative controls to check correctness.
2. Compare the sample use efficiency of the active learning algorthim and plain logitstic regression with measurements such as gradiants of loss fuctions.

5. Planned visuals

- Summarize the simulation results into tables.
- Plot loss curves for active learning and plain algortihtms in the same figure to compare the efficiencies of learners.
- Summarize and compare the performance (sensitivity, specificiy, etc.) of my DR classifier to other published algorithms.

6. Reference

Antal, Bálint, and András Hajdu. 2014. "An Ensemble-Based System for Automatic Screening of Diabetic Retinopathy." *Knowledge-Based Systems* 60. Elsevier: 20–27.

Dasgupta, Sanjoy, and Daniel Hsu. 2008. "Hierarchical Sampling for Active Learning." In *Proceedings of the 25th International Conference on Machine Learning*, 208–15. ACM.

Decencière, Etienne, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, et al. 2014. "Feedback on a Publicly Distributed Database: The Messidor Database." *Image Analysis & Stereology* 33 (3): 231–34. doi:10.5566/ias.1155.

Lichman, M. 2013. "UCI Machine Learning Repository." University of California, Irvine, School of Information; Computer Sciences. http://archive.ics.uci.edu/ml.