Nate Kaiser
Adam Pollack
Chainatee Tanakulrungson

Task / Objective

Our project goal is to predict the presence of prostate cancer from medical image data. This project is challenging since it involves machine learning of complex feature sets and also of machine learning of image data, both of keen interest to the authors. Due to the complexity of the problem, we will focus primarily on the machine learning challenges, as opposed to data pre-processing or other auxiliary aspects of the project. To facilitate, we will use freely-available data sets and machine learning libraries for the bulk of the algorithm.

This problem is unique and exciting in that it has impactful and direct implications for the future of healthcare, machine learning applications in personal data and decisions, and computer vision in general. The medical field is a very likely place for machine learning to thrive as medical regulations allow more sharing of anonymized data for the sake of better care. Not only that, but the field is still new enough that our project could even be at the cutting edge of technology.

Data

Medical image databases are readily available, and millions of labeled images are freely available for download (many are even split into training and test sets already). Our initial plan is to use the *PROSTATEx* collection from http://www.cancerimagingarchive.net/. The data consists of 346 distinct patients, 309,251 images, and all the associated metadata. The dataset is very new - completed only as of March 2017. We can also expand to use other datasets if more or different images are desired, but the data set mentioned above should be more than sufficient for constructing a robust learner.

Features

Depending on the type of learner, we will use one of the powerful feature types common to modern image processing (SIFT, SURF, etc). Other learners (ex: Convolutional Neural Nets) may not need feature extraction at all, since the convolving filter *is* the "feature" of interest and is updated during training.

Initial Approach

Pre-processing: The image data comes in a non-standard format, so we will need to use software to convert the data into a format we can process (many variants are freely available for Linux). The images also come with their labeled data separated into a text file, which will need to be mapped to the processed image data as we train our learner. This mapping will require a custom script, but should not be difficult.

Techniques: Since image data is complex, we do not expect simpler methods like k-nearest neighbors and decision trees to work very well. We plan to try a neural network and a non-linear SVM, and will compare the performance of each to A) validate our methods and B) form a comprehensive view of the strengths and weaknesses of the two methods. Other learners may be tried if there is leftover time.

Evaluation: Initial evaluation will be comparing our learner's results to the zeroR baseline. Comparison between learners will also be used to give us a relative success metric, but may not be insightful from an absolute perspective. Finally, while the ultimate measure of success would be to compare to classifications made by medical personnel, we will (probably) not have access to this valuable data. Perhaps we can find a good average error rate from a credible study to compare to. Note the labels don't offer a good comparison, since they will always be 100% accurate when compared to themselves.