From: Dasgupta, Abhijit (NIH/NIAMS) [C]
To: Ward, Michael (NIH/NIAMS) [E]
Subject: updates
Date: Wednesday, June 10, 2020 2:45:00 PM
Attachments: image002.png
image004.png

Hi Mike,

The last two days, I've been down the supervised learning rabbit hole. I labeled subjects into three groups: non-responders, early responders and late responders. All the subjects from the 3 studies who stopped at week 12 were labelled either non-responders or early responders. Late responders were those subjects from the other three studies who did not respond at week 12 but did respond at week 24. This labelling was done for each of the 3 outcomes (ASAS40, CII and ASDAS Major Improvement).

Data cleaning

I omitted smoking, as we had agreed. I also removed ank, ibd, psorias and tnf since they had very little variability in the dataset (tnf was all negative, and the rest were over 90% in one class). I also removed the composite BASDAI score and kept the 6 individual BASDAI scores. I log-transformed the CRP. After this, there were a few (1% or less) missing values in the data, which I imputed using K-Nearest Neighbors (K=5). I standardized age, BMI, CRP, and the joint counts. I ended up with 22 predictors (age, BASDAI 1-6, BASSFI, BMI, CRP (log-transformed), disease duration, MTX, ntp, pga, sjc, ssz, steroid, tbp, tjc, HLA B27, sex and uveitis).

Predictive modeling

The main objective, to my understanding, is to see if baseline metrics can predict response. So I've taken the predictive modeling approach using several machines for classification. I split the data up randomly with 75% for training and 25% for testing. I've been working primarily with the training data, but can quickly score the test data at the end.

In terms of the portfolio of models, I've been playing with 6 models:

1. Decision trees
2. Random forests
3. K-nearest neighbors
4. Gradient boosted trees
5. Linear discriminant analysis
6. Quadratic discriminant analysis

The primary metric of evaluation is accuracy, i.e. what proportion of subjects are classified into their correct classes of response (non-response, early response, late response). The metric is evaluated using 5-fold cross-validation to be able to get an idea of predictive (out of bag) error rates, since it was quickly evident that many of these models were overfitting the training data. For example, decision tree went from 95% accurate using just the training data to around 70% with cross-validation for CII. For the most part, default parameters were used.

Results

It is quite evident that ASAS40 is hard to predict, and is relatively noisy at least with respect to the measured predictors. The ASDAS measures are easier to predict. The following table gives the accuracies of the different machines for the different responses using essentially default parameters

	asas40_response	cii_response	major_response
DT	0.565611	0.693905	0.685663
GB	0.554340	0.689798	0.684643
KNN	0.530533	0.653560	0.655622
LDA	0.583265	0.719769	0.714657
QDA	0.529512	0.672229	0.647423
RF	0.584242	0.711533	0.681523

I also looked at modified responses restricted to week 12 responses, so it becomes binary. With the same process, I can compute accuracy as well as the AUC for the model predictions. That is given below

AUC

	asas40_response	cii_response	major_response
DT	0.605994	0.732859	0.766373
GB	0.628106	0.738373	0.785897
KNN	0.573524	0.638719	0.654351
LDA	0.665246	0.801251	0.816392
QDA	0.613007	0.710148	0.741016
RF	0.650496	0.776952	0.793652

Accuracy

	asas40_response	cii_response	major_response
DT	0.567737	0.717713	0.709449
GB	0.586379	0.718738	0.714647
KNN	0.551167	0.650451	0.671171
LDA	0.610160	0.749730	0.740511
QDA	0.569783	0.699097	0.696004
RF	0.600839	0.728102	0.698082

Precis

Random forests and LDA tend to do the best, and fairly similarly for the classification problem, but LDA does better on the binary problem. We can get some decent performance on the ASDAS outcomes, with better outcomes for the week 12 binary outcomes compared to the week 24 classification problem. We are seeing AUC in the high 70s to low 80s for CII and Major responses at week 12.

Ongoing work

I'm trying to optimize the random forest first to see how good performance could get. This is using a grid search on a set of parameters, using cross-validation and optimizing on the average cross-validation accuracy (for the 3-class problem). This takes a bit of time to run. I'm doing this to see if we can potentially find a combination of predictors and thresholds that can give good separation between the classes. Random forests are easier to tune, since with LDA I'd have to play around with various interactions between the predictors. With LDA, it's a bit easier to find the predictor

partitions since its basically a linear scoring rule and we can figure that out pretty easily.

I had earlier done unsupervised PCA on the predictors to see if any natural separation existed in predictor space… that answer was an emphatic no. So these models are basically contorting the predictor space to find some orientation that can hopefully distinguish reasonably between responders and non-responders.

Abhijit Dasgupta, PhD
Biostatistican (Contractor)
Clinical Trials and Outcomes Branch
National Institute of Arthritis and Musculoskeletal and Skin Diseases
National Institutes of Health
10 Center Drive, Room 4-1350
Bethesda, Maryland 20892-1468
301.496.1622 (W) | 301.385.3067 (M)