Notes on Analysis of Recursion COVID Data

Randall Reese

March 2, 2021

0 Premise

This document is to supply background and context to some of the thought processes behind my analysis of the RxRx19a data set from Recursion Pharmaceuticals. The code and relevant notes directly relating to the code and analysis are found in an accompanying Jupyter notebook entitled `RRWorkbook.ipynb`.

Below I will address the questions given in the project description.

1 Model Selection

Question 1.a) What is your chosen approach/model(s), and why did you choose it?

My overall approach here was to use Random Forest (RF) models. I chose this type of model because RFs are usually decently easy to train out-of-the-box and they produce favorable results. RFs are non-parametric and do not require many assumptions about the data. (Thus eliminating bias). I also reduced the feature space down to 63 features using PCA. I don't know if this is what the embedding data already was, but just for expediency, I reduced the embedding space further. The choice of 63 features was the minimal amount necessary for 90% explanation of variance.

Question 1.b) What assumptions does your method make about the features?

Random Forests makes almost no assumption about the data. There is no bias towards certain geometries or data relations (such as LDA, QDA, linear regression, etc.). We do assume that sampling is representative and that our data has no underlying time correlation component. RFs do tend to bias towards multi-level categorical features, but all of our features here are continuous, so this is not an issue. We also assumed that all treatment/disease condition classes were well represented. Examining the data, this seemed to be the case.

Question 1.c) How well does your approach discriminate between healthy and disease states of the provided data?

I fit three initial models: Using only HRCE data, using only VERO data, using the full (after cleanup, see the Jupyter notebook) data set. The accuracy of each model is summarized below in Table 1:

Model	OOB Accuracy %	Test Accuracy %
RF-HRCE	95.98%	96.01%
RF-VERO	97.35%	97.40%
RF-Full	96.09%	96.12%

Table 1: Model performance

The Out-of-Bag (OOB) accuracy is a cross-validated measure of model fit during the training phase. The test accuracy is the accuracy of the model on the randomly selected test data. These accuracies suggest that our models discriminate very successfully between healthy and diseased states.

2 Model Generalization

Question 2.a) How well does your model(s) generalize to different plates, experiments, or cell types?

My models were tested across a test set including a wide selection of plates, experiments, and cell types and achieved very high test accuracy. None of my models had any contingencies or reliance on plate number or well location. Adding these variables to the model could have resulted in overfitting the model and would have likely compromised the ability of the models to generalize to different plates or wells.

Question 2.b) If you were provided new data from new experiments (i.e. HRCE-3 and VERO-3), how would you expect your model(s) to perform?

I would expect the models to perform at least decently well, assuming that all other aspects of the experiment were maintained. Some of this would come down to understanding how the imaging works, how much variability occurs in the images between plates, experiments, etc. I included randomly sampled data from HRCE-1, HRCE-2, VERO-1, and VERO-2 in my final full RF model. This should at least give us some confidence in the generalizability of the model.

3 Treatment Scoring

My scoring system first fits a RF model to data of both healthy and diseased untreated sites. (I.e. those sites for which no treatment was given). The untreated sites with Active COVID were considered to be the canonical examples of "diseased"; the untreated sites with no virus or UV killed virus were considered "healthy." Once this RF model is fit, we use the model to determine class probabilities (diseased, healthy) for each treated site. Using a naive cutoff of 0.5, any treated site with a class probability of being healthy was considered "rescued." This cutoff could be re-visited in a more extensive analysis. (For example, is 55% probability of being healthy strong enough to say a site was rescued?) Overall, this approach is suggestive of a semi-supervised learning method.

Once each site is labeled as rescued or not, we calculate the percentage of rescued sites for each treatment. This percentage of rescues is our final score.

Question 3.a) What are the top 20 compounds, as found by your measure of "rescue"?

I found the 20 (actually 24) compounds with the highest rescue scores. (There was a six-way tie for 18th). There may be ways of implementing a tie breaker for those last spots if we knew more about the treatments. (E.g. is one treatment much less costly? Easier to perform? etc.) This would require some content knowledge of the treatments.

The "best" rescuers are listed below in Table 2, including the rescue scores:

Question 3.b) What are the strengths and weaknesses to your selected approach to scoring treatment conditions?

I think that one strength of this method is that it relies on the known (labeled) data to train a model that will provide our scores. It also allows for a weighted scoring that gives weight to percentage of rescues, not just total rescues. Some drawbacks of this approach could be that it is not guaranteed that rescued sites will present visually in the same way as a site with no COVID to begin with. This comes down to what the actual dye testing does. This is not my area of expertise, so I may be overlooking something that could lead to a more robust score. I think that it also may be worth considering what the cutoff for "rescue" should be. (Should it be 0.5? 0.7? 0.3?). A degree of uncertainty may be introduced into the scoring there. This might be mitigated by bootstrapping up the model more slowly by retraining using treated sites that scored as highly probable to be healthy. (This would be a sort of heuristic self-training).

Treatment	Frequency	Percent Rescues %
etazolate	25	17.36%
gepefrine	24	16.67%
GS-441524	138	15.97%
MLN2238	23	15.97%
D-Mannitol	23	15.97%
Ramosetron	22	15.28%
Ulipristal	22	15.28%
oxiconazole	21	14.58%
Losartan Carboxylic Acid	20	13.89%
telotristat	20	13.89%
gemifloxacin	20	13.89%
secnidazole	20	13.89%
phenolsulfonphthalein	20	13.89%
Primidone	19	13.19%
Domperidone	19	13.19%
4-Biphenylacetic acid	19	13.19%
Dichlorophen	19	13.19%
Mycophenolic acid	19	13.19%
nalfurafine	18	12.50%
dehydroepiandrosterone-sulfate	18	12.50%
norethindrone-acetate	18	12.50%
Tafluprost	18	12.50%
Dequalinium	18	12.50%
Glipizide	18	12.50%

Table 2: Top rescue treatments

Question 3.c) Would you judge the treatment predictions/scores to be necessary and/or sufficient to conclude the treatment effectively rescues the disease effect (makes disease look like healthy)? Why or why not?

This scoring system is likely not sufficient. This is certainly not something I would put out as a final decision on rescue. I would need to know quite a bit more about how the dye testing presents under various conditions and mediums. And, after all, we are not looking at the actual images here. Just from a philosophical stand-point, it is rather risky to ever say that a preliminary scoring system is "sufficient." Suggestive would be a good word to use, in that these scores might suggest treatments to further consider. But this is certainly not sufficient.

I tend to think that treatments scoring highly here are necessarily more likely to be prone to rescue. If a treatment scores very low, it is less likely that the treatment in question is highly viable. This would mean that a very large percentage of sites do not look the same as healthy sites, which necessitates our asking if the treatment did any rescuing.

Overall, this is something that would need to be done in consult with people more familiar with the actual dye/imaging process. How do healthy cells present? Is there a difference between mock and UV-inactivated visually? Does having dead cells in the site/well make a difference? I would think that the scoring would ultimately be a first step towards further testing and analysis, not a final conclusive measure.