ROC Analysis to Determine Optimal NDVI Threshold

Andrew Pericak, June 2016

Introduction

After meeting with SkyTruth, one of my first tasks was to find a way to choose an optimal classification threshold, ideally using a method with statistical validation. To address this question, I spoke with Dr. Jennifer Swenson, a professor in the Nicholas School at the Environment at Duke, who does a lot of work and research with remote sensing. She proposed using ROC (receiver operating characteristic) analysis to help determine the NDVI threshold. There's a lot written about ROC online, but I'll quickly summarize it here.

ROC, used most commonly for medical studies, helps determine threshold values when doing binary classification (in our case, is a pixel a mine or not.) It does this by taking every possible threshold and finding the true positive rate (sensitivity) and false positive rate (1 – specificity) of using that threshold as the classifier. It finds these rates using provided training data (ground-truthed data, in our case.) The resulting ROC curve plots the true positive rate versus the false positive rate (Figure 1). The goal is to maximize the combination of sensitivity and the specificity (which is to say, maximize the true positive rate and minimize the false positive rate), which on an ROC curve is aiming to have the curve approach the top-left corner. The closer the curve approaches this corner, the better accuracy the model has; in other words, more area under the curve (AUC) is associated with better accuracy.[1] The threshold associated with the point on the curve closest to this corner represents the optimal threshold (if 'optimal' means maximizing true positives and minimizing false positives, although what is optimal will depend on the study.)

This means two things for us. First, by looking at the AUC, the ROC analysis can tell us whether NDVI or other metrics (like SAVI or EVI) are the best to use. Whichever metric has the greatest AUC is the "best" model and thus would have the most accuracy at classifying mines. Second, the ROC analysis can reveal the optimal metric threshold (as defined by maximizing the combination of sensitivity and specificity), and could do this for any year or sensor for which we have ground truth data. This means we can employ ROC to use multiple thresholds, and can justify that decision because we are consistently choosing what combination of sensitivity and specificity to require from our threshold.

Method & Results

Using R and the library "pROC", I first ran the ROC analysis on the combined Landsat 5 (2005) and 8 (2014) training data from Tita's spreadsheet, looking at NDVI, EVI, and SAVI since all

[1] In particular, the AUC represents the probability of a model to correctly "rank" a true positive pixel higher than a true negative pixel. Thus, a high AUC means the model has a high probability of accurately classifying each input location (pixel) relative to each other input location. With a high AUC, most true positives will be labeled as such, but occasionally a true negative may slip in too, resulting in a false positive.

three metrics had values associated with them. In this and subsequent cases, the NDVI "model" (i.e., comparing NDVI thresholds to ground-truthed data) had the greatest AUC (Table 1). This outcome indicates that merely using NDVI is, in fact, the best option in our case for determining mine areas. Notably, the differences in accuracy (AUC) among the three metrics aren't too different (especially for Landsat 8), but because NDVI is always the highest, we can safely say it is our best choice.

Even if NDVI is the best choice, however, we want to ensure that using NDVI will give a statistically significant model result. I used R to run a quick logistic regression, modeling the impact of NDVI on whether a pixel is or isn't a mine. The model revealed using NDVI as a predictor had very good significance ($p < 0.0001$) and a good pseudo-R^2 value (McFadden's $R^2 = 0.7306$). These results indicated that NDVI itself can reasonably predict whether a pixel is a mine. These statistics do not say what threshold is best for that prediction, which instead the ROC analysis can reveal.

Using the NDVI data for the Landsat 5/ 8 model, I looked at the optimal threshold as named by the ROC analysis (Table 1). The derived optimal threshold (0.58705) for the combined Landsat 5 and 8 data greatly exceeded the current 0.51 threshold or the 0.54 threshold that Tita used. Again, this optimal threshold maximizes the combination of true positives and 1 - false positives (Table 2), and we know it can give accuracy around 98.4% (Table 1).

Table 1. Per sensor, AUC values for different models, and the ROC-determined optimal threshold from the NDVI model.

Sensor	NDVI AUC	SAVI AUC	EVI AUC	NDVI Threshold
LS 5 & 8	0.9841	0.9763	0.9711	0.58705
LS 5	0.9691	0.9541	0.947	0.57635
LS 8	0.998	0.9944	0.9914	0.5665
LS 7	0.9849			0.6156

Table 2. Per sensor, the ROC-determined optimal threshold from Table 1, and the associated sensitivity and specificity of those values.

Sensor	NDVI Threshold	Sensitivity	Specificity
LS 5 & 8	0.58705	0.950311	0.95119
LS 5	0.57635	0.920792	0.934919
LS 8	0.5665	1.00	0.977381
LS 7	0.6156	0.950311	0.951190

For comparison, I calculated the ROC point of sensitivity and specificity for the 0.51 threshold. This gives a lower true positive rate (sensitivity: 83.2%) but a higher 1 - false positive rate

(specificity: 98.3%) than the values for the ROC-specified threshold (Table 2). In other words, using this threshold will give us fewer true positives than the ROC-specified threshold, but it will also give fewer false positives.

Since Tita's training data was split by sensor (LS5 versus LS8), I also ran the ROC analysis using just the ground truth data for those sensors. Again, the NDVI model gave the greatest AUC, indicating that this is the best metric to use (Table 1). Logistic regressions for these data gave similar, strong outcomes as before. The optimal threshold values were individually both less than the combined threshold value, but were still higher than the 0.51 threshold (Table 1).

I was able to have a student intern in the Bernhardt lab go through the same 900 sample sites Tita used and sort them into mine/unmined for Landsat 7 imagery (for the year 2009). I extracted the NDVI values from the imagery to those points in Earth Engine and again ran the ROC analyses. (Note, in this case I did not extract the SAVI or EVI values, since the other sensors were so strongly indicating that using NDVI would be our best option.) As before, the threshold again exceeded the 0.51 threshold, and in this case exceeded all the other thresholds (Table 1).

Discussion

The ROC analysis revealed two key points about our mine classification. First, using NDVI as the sole classification method is likely the best option, since all sensors had an accuracy of at least 98.4% by using NDVI, and because running a classification solely using NDVI can deliver statistically significant results. Second, the threshold values used for this classification can vary and will ultimately depend on what sort of information we wish to show with the mine dataset. We could use the ROC-defined optimal threshold (maximizing the combination of sensitivity and specificity), or we could seek another threshold that attains some other combination of sensitivity and specificity.

In itself, this latter point is an important consideration about ROC: while the "best" threshold is the one maximizing the combination of sensitivity and specificity, using that threshold could come at a tradeoff if specificity is sacrificed for greater sensitivity, or vice-versa. For the combined LS5 and LS8 dataset, for instance, using the ROC-derived threshold would give us a better true positive rate than using the 0.51 threshold. But in so choosing that optimal threshold, we would also introduce more false positives than if we stuck with the 0.51 threshold. In our case, therefore, the "best" threshold may not necessarily be the one maximizing the combination of sensitivity and specificity. Moreover, note that ROC does not consider true negative or false negative rates. I did not calculate those here but could if necessary, in case we were also concerned about those values.

That each sensor, analyzed individually, had such varying threshold values suggests that we might introduce error into the final data products by using one threshold. For example, the very high "optimal" threshold for LS7 (0.6156; Table 1), illustrates the problem with using one threshold. If we keep the 0.51 threshold, mines classified with LS7 imagery would contain many

false negatives, and would likely contain more false negatives than mines classified with other Landsat imagery. On the other hand, since the thresholds varied so much among each other, we could potentially achieve better accuracy by having unique thresholds for each sensor. The trick to this, though, will be attaining ground truth data for each sensor (although we now have this for Landsat 5, 7, and 8).

Regarding statistical significance, the ROC test itself doesn't have an associated statistical test to "prove" that any threshold is better than another. What it can say is that a model with high AUC (see Table 1) has high accuracy in performing the classification. In our case, since the AUCs for all NDVI-based models are so high, we can assert that using NDVI as a classification measure will give highly accurate results (regardless of where we set the threshold—the high AUC indicates that the model will "rank" a random, true positive ahead of a random, true negative.) Setting the threshold is somewhat an arbitrary exercise because it depends on whether we want to give more worth to sensitivity or specificity (i.e., maximizing true positives or minimizing false positives.) From the logistic regressions, however, we can say that NDVI itself can significantly differentiate between mine and non-mine, so we are "correct" in only using NDVI as our basis of classification.

Conclusion & Recommendations

With all this information, we now have some choices to make. First, we should decide whether we want to have one, uniform threshold, or whether we would want to have separate thresholds per sensor or even per year. Second, we can decide whether we want to use the "optimal", ROC-derived threshold, or whether to use a different threshold that attains some different combination of sensitivity and specificity. To help with these goals, I used pROC to export a table of all points along the ROC curves for the four variations I ran here (5&8, 5, 7, and 8) (Figures 2 – 5). I graphed the sensitivity and specificity curves by threshold; where these curves intersect is the ROC-defined optimal threshold. These curves show, for instance, that we could likely sacrifice some specificity in favor of sensitivity (or vice-versa) without having a significant impact on the accuracy of the specificity.

Having looked at this data for the last few days, I recommend using multiple thresholds, at least by sensor. This choice might take more time to explain in a final, public write-up, but given the observed variation in thresholds for each sensor, I think that trying to simplify those thresholds with one uniform threshold would lead to much error in our output datasets.

For consistency's sake, I would also recommend using the ROC-defined optimal threshold, but I could also support an argument for choosing a different, varying threshold that maximizes specificity (i.e., minimizes false positives.) By maximizing specificity, we are ensuring that we are not overstating the true extent of mines, while acknowledging that we are likely not describing the full extent of the mines. I hope this document is useful for our discussions in deciding where to set the threshold(s).

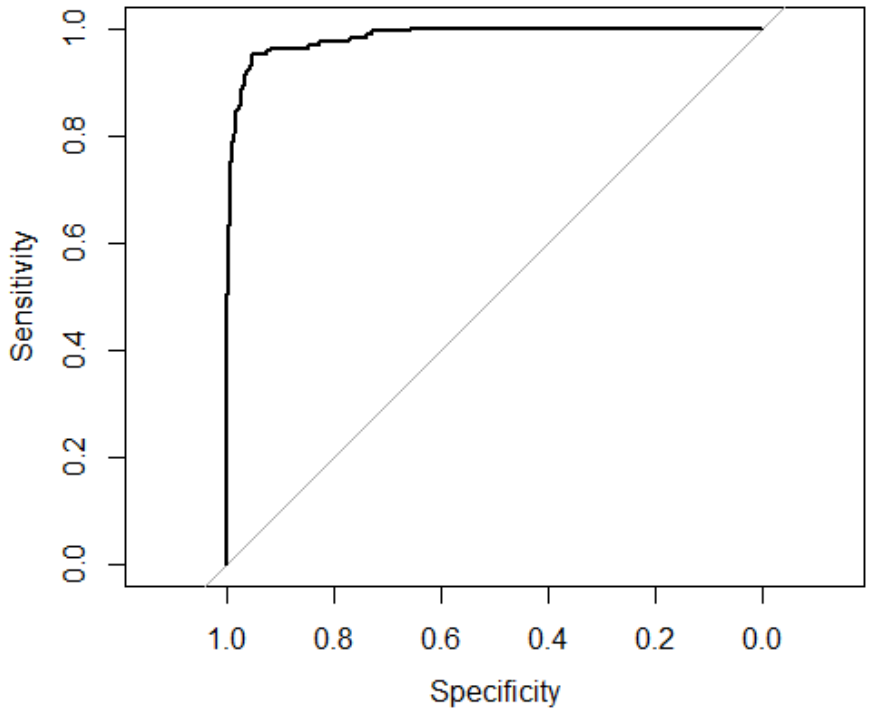

Figure 1. A sample ROC plot, depicting the sensitivity (true positive rate, *y*-axis) and specificity (1 – false positive rate, *x*-axis). This happens to be the ROC plot for the combined LS5 and LS8 ground-truth data. The light gray line ($y = x$) is a reference: an ROC curve falling on this line depicts a binary classification model that is no better than guessing (i.e., sensitivity and specificity are exactly inversed). Most ROC curves fall above the reference line, suggesting the model does better at classifying than a random guess. If the curve were below the reference line, we would be better off guessing a classification than using a model at all.

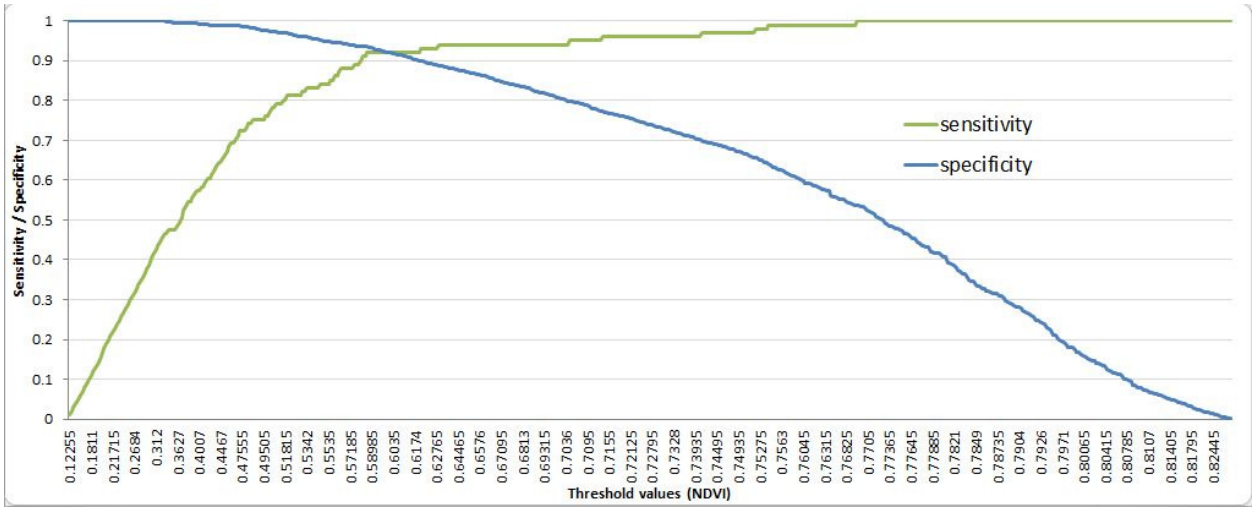

Figure 2. Sensitivity and specificity per threshold for combined LS5 and LS8.

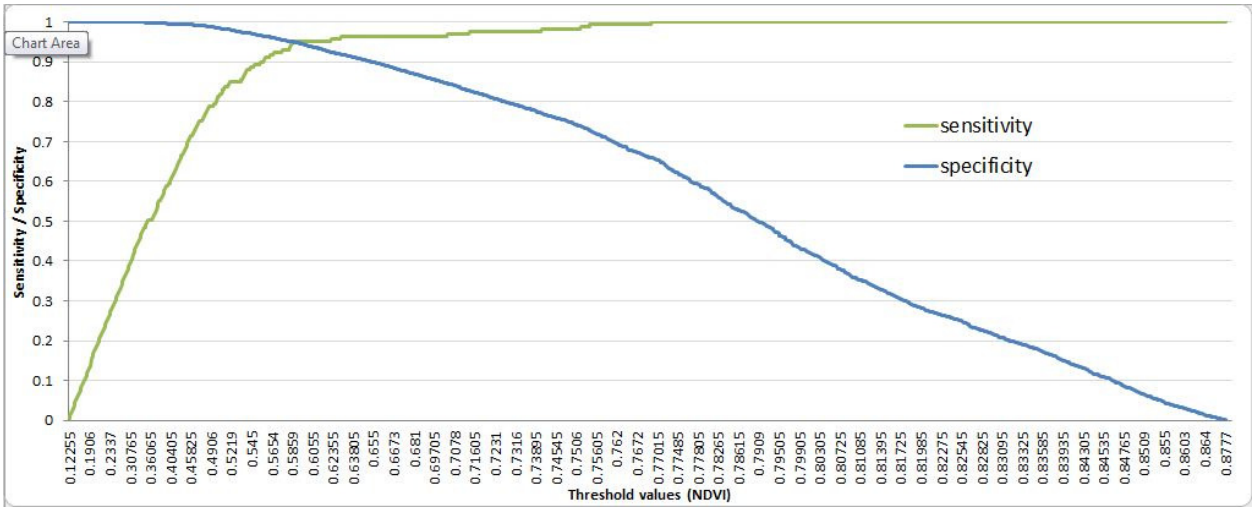

Figure 3. Sensitivity and specificity per threshold for LS5.

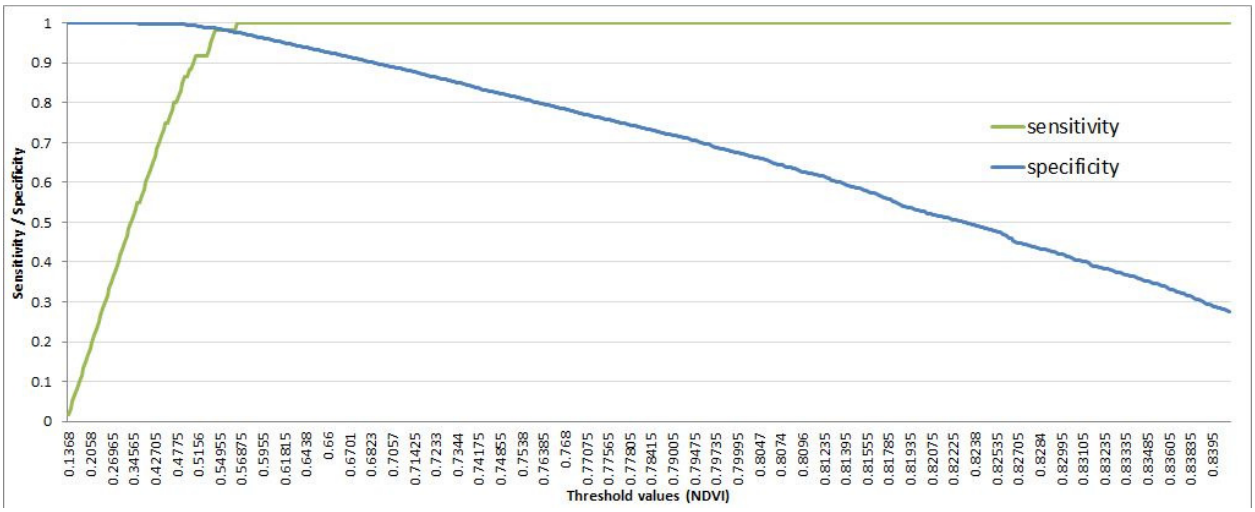

Figure 4. Sensitivity and specificity per threshold for LS8.

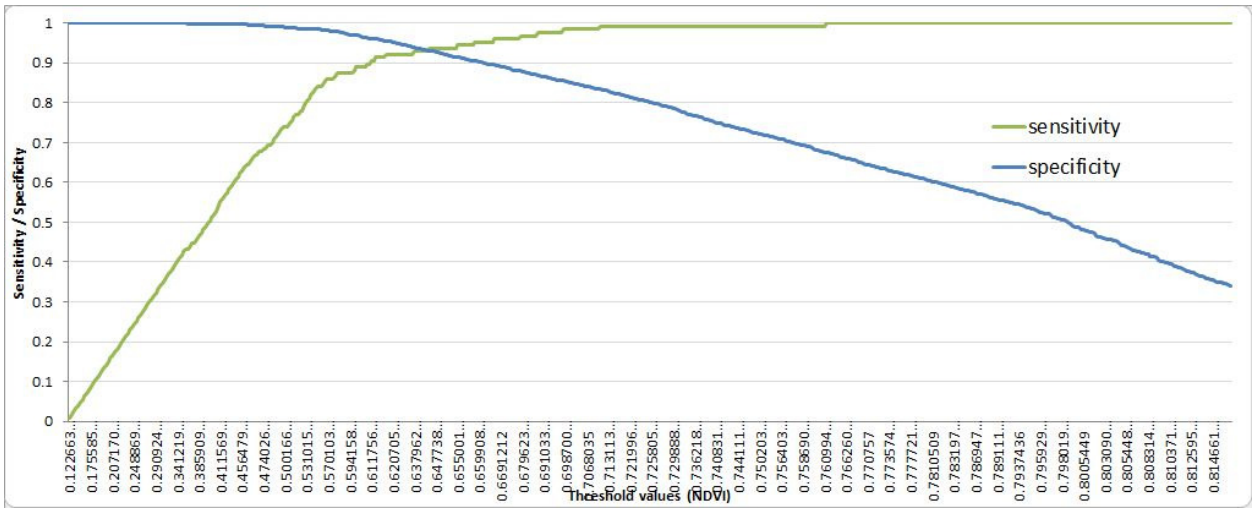

Figure 5. Sensitivity and specificity per threshold for LS7.