Lab 4 - Cloud detection - Stat 215A, Fall 2017

Olivia Angiuli, Miyabi Ishihara, Yizhou Zhao

November 8, 2017

1   Introduction

In 1999, NASA's Multiangle Imaging SpectroRadiometer (MISR) data generated a massive data set of Arctic images that help provide insight into recent global changes of ice coverage due to increasing surface air temperatures. The dataset is rich, generated by a satellite that collected images along 233 different geographical paths, with 180 blocks in each path, providing images of any given region every 16 days.

One of the biggest statistical challenges of this dataset is to conclude, for each image, which regions are ice-covered, as opposed to snow-covered or simply unknown. Using only a small subset (three images) of this data, each of which produced radiances from 5 different angles, we use expert labels in order to train and evaluate classification models that can help differentiate between icy versus cloudy surfaces.

2   Exploratory Data Anaysis

2.1   Data background

We are provided with three images, each of which provide the following information:

- **Expert labelling.** Experts visually inspected each image, classifying each pixel as cloudy (+1), not cloudy (-1), or unlabelled (0). These are the labels with which we will train our supervised algorithms, and with which we will test our results.

- **Radiances from five angles.**   We were given satellite images for three different 275m x 275m regions, each from five different angles as shown below (where "F" indicates the forward direction of flight and "A" represents the rear direction):

  As will be described in the next section, differences in radiance measurements of the same image from different angles can tell us about the presence or absence of clouds and/or ice.
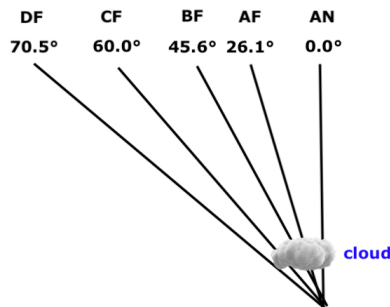

Figure 1: The five angles from which we have radiance measures, for each of the three images.

- **Feature values.**   Shi et. al. developed three features that capture spatial information that helps distinguish between ice and clouds:

  - **CORR -** *an average linear correlation of radiation measurements at different view angles.*
    A cloud may obstruct the view of underlying ice from some angles but not others. Figure 1, above, illustrates a scenario in which DF provides an unobstructed view of the underlying image that the other angles cannot provide.
    This suggests that low CORR images often suggest clouds: however, low-altitude clouds or smooth cloud-free areas may break this trend, requiring the following two features.
  - **SD -** *standard deviation within groups of MISR.*
    SD helps identify smooth surfaces (areas with low standard deviations), as well as to identify a baseline for background measurement error that can be filtered out in the model.
  - **NDAI -** *a proxy for surface roughness, measured by thenormalized difference between measurements in the forward versus backward pointing cameras.*
    The intuition is that a low-altitude cloud has more roughness than an ice- or snow-covered surface. This feature combines with CORR to help differentiate between low CORR images that have low-altitude clouds versus are cloud-free.

2.2   Data visualization

The following plots visualize the three provided images. The first set of images shows the raw data from the perspective of the AN-camera. The second set of images shows the same three images, classified by whether an expert determined that a given area was ice (pink), unknown (green), or cloudy (blue).
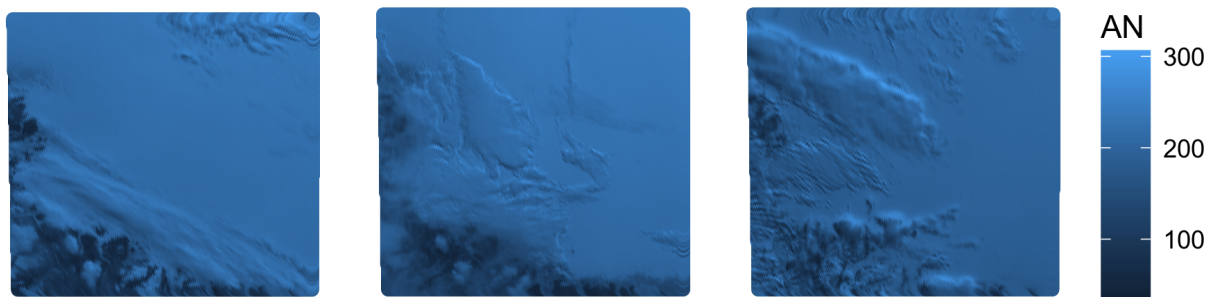

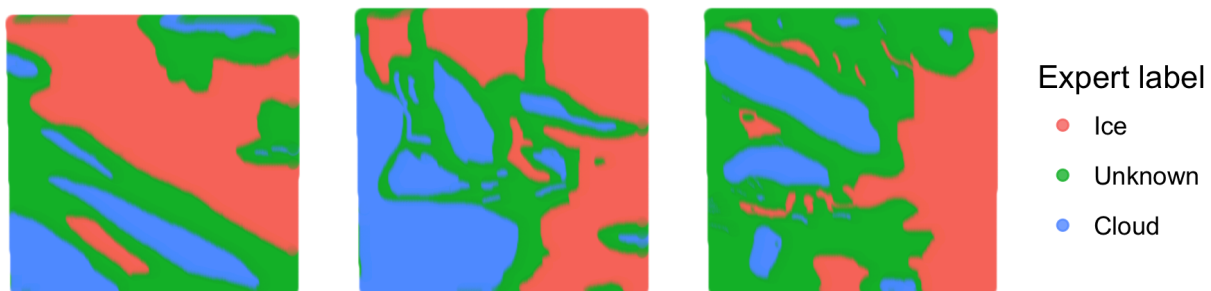Figure 2: Raw (from AN-camera) image data for the three provided images.

Figure 3: Expert-labelled image data for the three provided images.

2.3   Do different angles provide different information?

Yes! Below, we show image 1 from all 5 different angles that we are provided in the data set: from left to right, 70.5° (DF), 60.0° (CF), 45.6° (BF), 26.1° (AF) in the forward directions, and 0° (AN).

One can spot higher resolution in the bottom left corner (which corresponds to mostly cloudy and/or unknown regions) as the angles becoming increasingly acute. This supports the reasoning for Shi et. al.'s construction of the "CORR" and "NDAI" features: cloudy regions might have more variation in radiances between different angles.
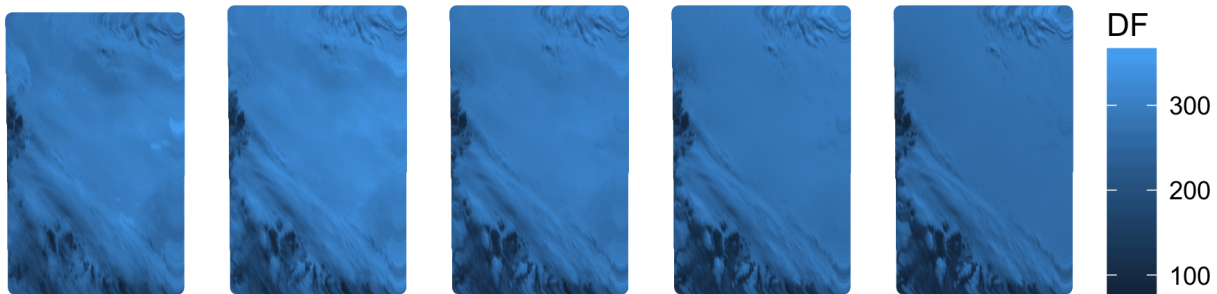
Image 1, taken from increasingly acute angles

Figure 4: Image 1, by increasingly acute angles.

Plotting the conditional densities of cloud- versus ice-covered regions confirms that different angles do indeed help differentiate between cloud and ice. We notice two trends in particular. First, icy regions tend to occur at a resolution of around 275. This can effectively set a high "prior" for iciness for pixels with a similar resolution. Second, we notice that cloudy regions have a much wider range of resolutions that change with the angle of measurement. Again, this confirms Shi et. al.'s intuition that differences between measurements from different angles often correspond to cloudiness.

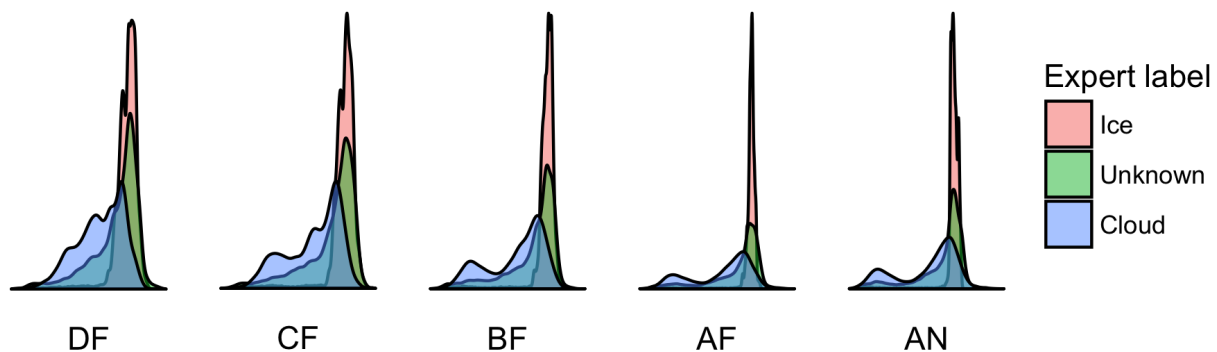Image 1 conditional densities, by measurement angles

Figure 5: Conditional densities of ice vs. cloud vs. unknown by different angles. Cloud resolutions become increasingly bimodal as measurement angle decreases.

2.4   How do features help distinguish between cloud and ice?

Recall that the three features, as described in detail above, are:

- CORR (correlation of images from different view angles)
- SD (the standard deviation within groups of MISR)

- NDAI (a proxy for surface roughness, measured by the normalized difference between measurements in the forward vs. backward pointing cameras)

By itself, NDAI has the most discriminatory power between the presence of ice and the absence of ice (either cloudy or unknown). This can be seen by the distinct peaks in the leftmost graph below.
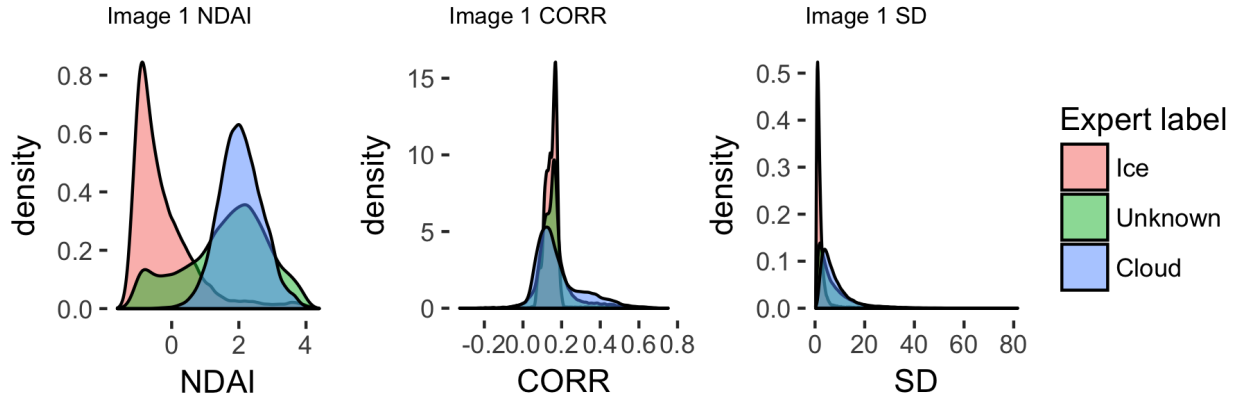

Figure 6: Conditional densities of ice vs. cloud vs. unknown by different angles. Cloud resolutions become increasingly bimodal as measurement angle decreases.

The discriminatory power of NDAI can also be seen from the following side-by-side plot of NDAI resolution values versus the true classes of image 1. Note that higher values of NDAI tend to correspond to "no ice" regions.
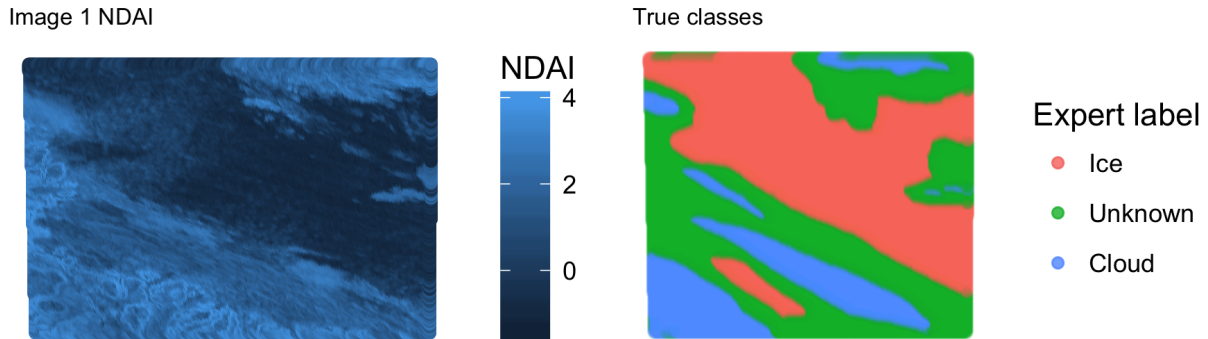

Figure 7: High NDAI values tend to correspond to unknown or cloudy regions; low NDAI values tend to correspond to icy regions.

Shi et. al. also discuss how NDAI and CORR, together, provide even more information than the two features alone: they hypothesize that high CORR + low NDAI images would more strongly suggest ice than high CORR alone. This relationship, though, is not readily seen in the following pairwise feature plots, largely because there are few high CORR pixels that correspond to ice-covered regions.
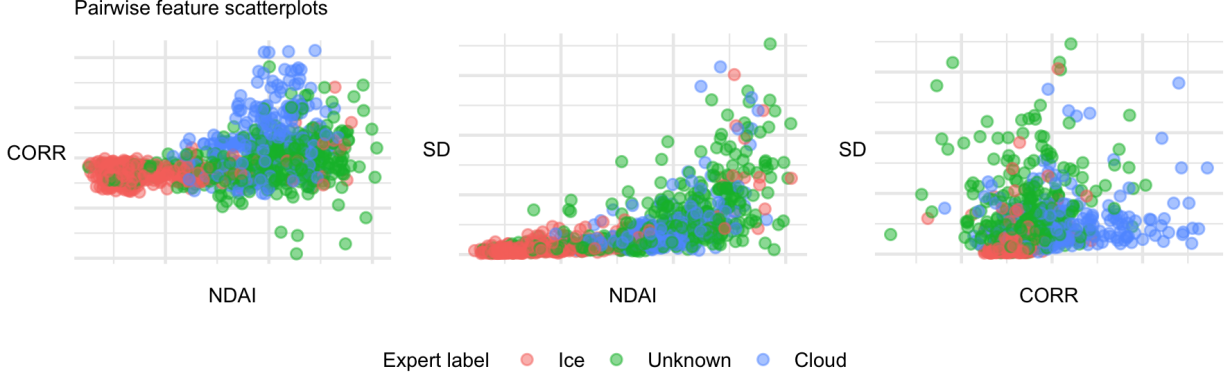

Figure 8: High NDAI values tend to correspond to unknown or cloudy regions; low NDAI values tend to correspond to icy regions.

3   Classifying ice- versus snow-covered surfaces

The classification approach of Shi et. al. hinges upon an enhanced linear correlation matching (ELCM) algorithm. However, we noted that this approach depended on having data images available from multiple orbits and/or consecutive MISR blocks. Since we do not have access to this full data set, reproducing this ELCM approach is not possible.

For that reason, we instead resort to more classic classification methods: QDA/LDA, logistic regression, and random forests. First, we train baseline versions of these methods. We note that random forests produces the best AUC values of the three baseline models, so we proceed to do additional parameter tuning and analysis for that model.

We decided before training these methods on a few approaches:

- **Three-fold cross validation.**   Since labels of adjacent pixels in an image are not truly independent, we decided to perform three-fold cross validation: that is, we would train three different versions of each model, each time setting aside an entire image for testing purposes.
  This is much preferred over randomly sampling points, since regions of ice and cloud often span many adjacent pixels, meaning that training on one of these pixels would enhance the test performance of the pixels nearby to it.

- **Training only on positive- and negative- expert labels.**   We do not train our model on data points whose expert classification was a 0 (unknown): we would only like our model to learn classification based on points whose true value is known.

3.1   LDA/QDA

Quadradic discriminant analysis(LDA) for classification aims to separate the feature space using lines(hyperplanes in higher dimenstion). It assumes that each class i has a Gaussian distribution with mean μ_i and covariance matrix Σ_i, that is

$$p(y = i|x) = \frac{\pi_i |2\pi\Sigma_i|^{-\frac{1}{2}} \exp[-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i)]}{\sum_k \pi_k |2\pi\Sigma_k|^{-\frac{1}{2}} \exp[-\frac{1}{2}(x - \mu_k)^T \Sigma_i^{-1}(x - \mu_k)]}$$
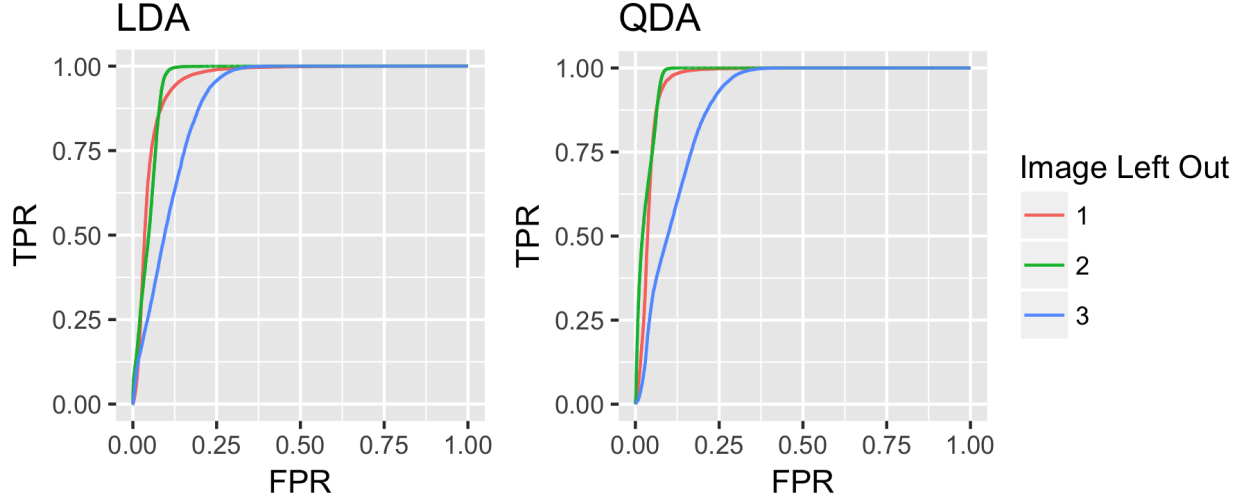
5

Figure 9: ROC curve for LDA/QDA three-fold cross-validation.

Linear discriminant analysis(LDA) is a special case of QDA, which makes further assumption that the covariance matrix are the same across all classes. Our problem deals with binary data, and we can classify each point by selecting a probability threshold and assign a hard hyperline to determine whether a pixel belongs to clear or cloud.

Figure ??? shows the ROC curves for the leave-one-image-out cross validation results from LDA and QDA models, in which QDA performs a little better than LDA in the first and second fold, which is in line with our exploratory analysis: the covariances for NDAI/CORR/SD may be difference between cloud and clear pixels.
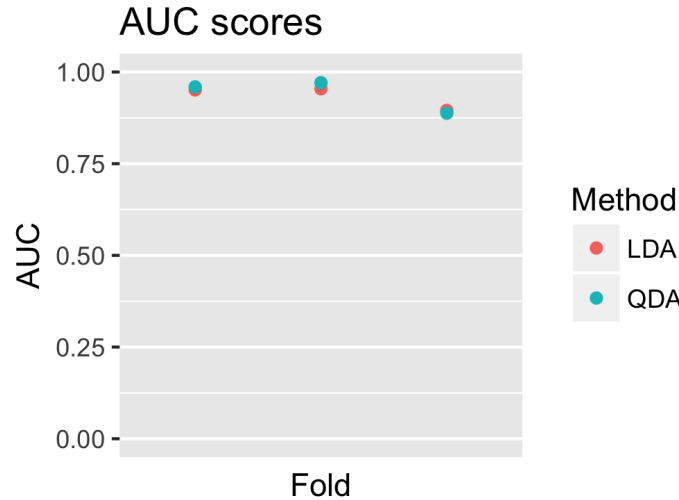

Figure 10: AUC curve for LDA/QDA three-fold cross-validation

Cross-validation for QDA and LDA tells that while they work really well in the first and second fold, finding that the AUC scores almost 1, it falls to perform as well in the third fold, where we regard image1 and image2 as training data and image3 as testing data. More specifically, the AUC scores for qda in the three folds are $0.960, 0.971, 0.888$ respectively, and for lda $0.951, 0.954, 0.895$.

Figure ??? shows that confusion plot after we selected the threshold to be 0.11, which is the average of the thresholds in the three-fold validation to make the point (TRR, FPR) farest away from the line $y = x$. As the above plot shows, the classification suffers more from false positive and false negetive with some regions completely misclassified. Actually, those regions are hard to be classified from our bare eyes. Calculation shows that the false positive rates of the three folds are $0.121, 0.122, 0.195$ and false negative rates are
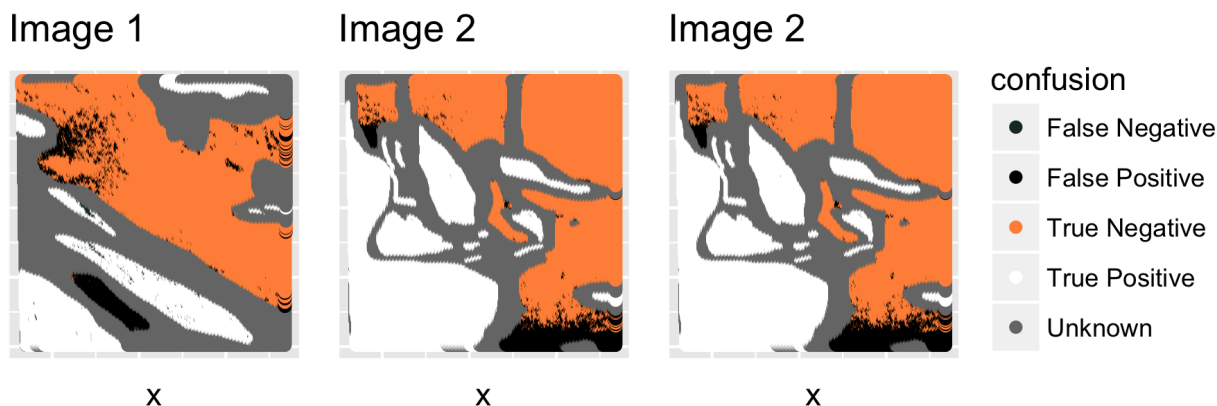
Figure 11: Confusioin plots for qda

$0.018, 0.001, 0.165$. The third fold cross-validation has such a high false negative rate because image3 has a small area of cloud and some parts of it on the upper left of the image are misclassified. However, the main problem is that the qda/lda classifier is hard to judge whether the "dark" regions in the images belong to cloud to clear, for example, the middle lower left region in image one, the lower right region in image two and lower middle and lower right regions on image three. All of that informs us of the irregularity of image three, for its large fraction of unknown area and complicated cloud and clear regions.

3.2   Logistic regression

3.3   Random forests

Random forests are an ensemble learning method that combine the prediction of multiple component decision trees in order to make a probabilistic determination of the class of an input.

Random forests are "random" in a few different steps:

- Tree bagging. Each component tree randomly selects, via bootstrapping, a training subset from the total training set.

- Feature bagging. At each candidate split of a component tree, a random subset of the features are used.

This is illustrated visually in the below picture.

In our case, we fit random forest with 100 trees three times (using the three-fold cross validation approach described above). The resulting ROC-curve is shown below, and reveals very strong performance, shown by the model's ability to achieve a high true positive rate with quite a low corresponding false negative rate.

3.3.1   Parameter tuning

The three main parameters to tune are the number of trees in the model, the maximum number of leaves (nodes) in each tree, and the number of features to use in each tree.

First, we examine the marginal behavior between AUC and each parameter. We see from the above graph that AUC is maximized with:

- **A high number of trees**. This is intuitive because more trees lead to a more stable prediction. Since each tree uses different baseline data and different features, overfitting is usually not a big concern.

7

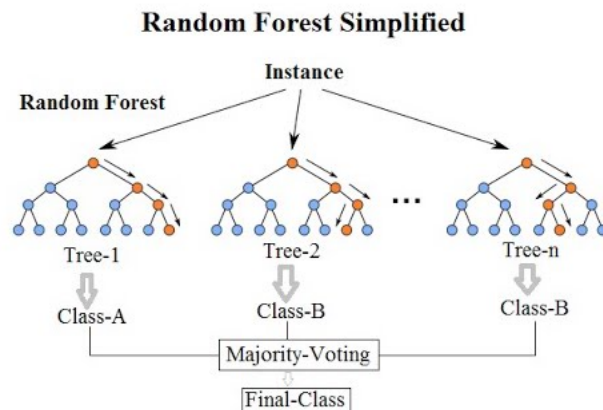**Random Forest Simplified**

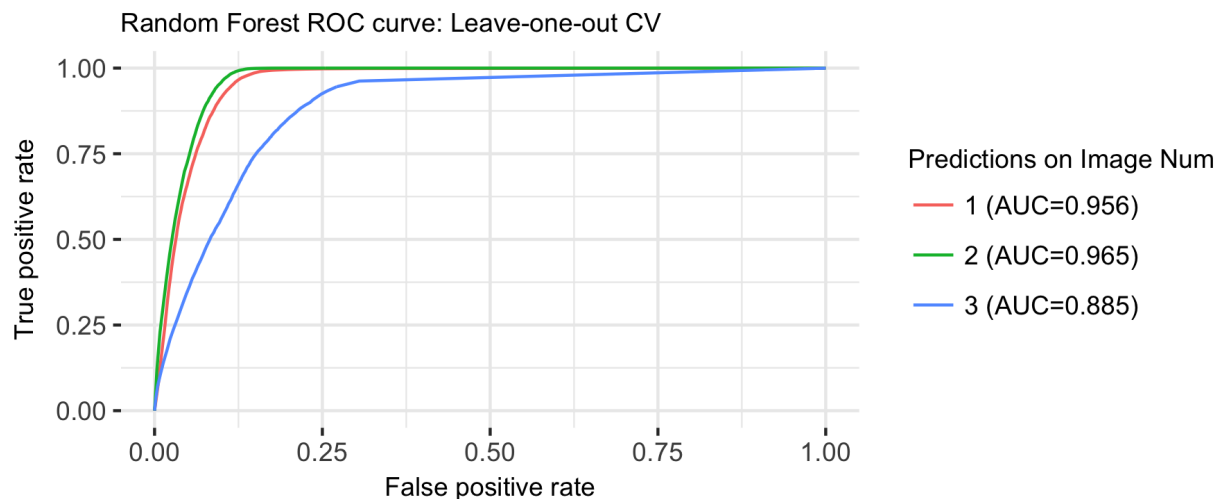Figure 12: Graphical introduction to random forests [2].

Figure 13: ROC curve for random forest three-fold cross-validation.

- **A moderate number of maxnodes**. This suggests that having too many leaves in any given component tree may lead to overfitting.

- **A small number of features tried in each tree.** At first, this may be counterintuitive: you would think if more features are considered within each tree, this would lead to a strictly better result. However, since our features may be correlated, then a small number of features may be better for prediction since it may decouple the correlation between features and avoid overfitting.

The below table shows the three best and three worst parameter combinations, in terms of AUC performance. The best parameter combination coresponds to 500 trees with a maximum of 10 nodes per tree and 1 feature tried at each point, which outputs a model with an AUC of 0.900.

Num Trees	Max Nodes	Num Features	AUC
500	10	1	0.900
100	10	1	0.898
50	10	1	0.896
...	...	...	...
100	10	3	0.810
50	10	3	0.809
10	10	3	0.803
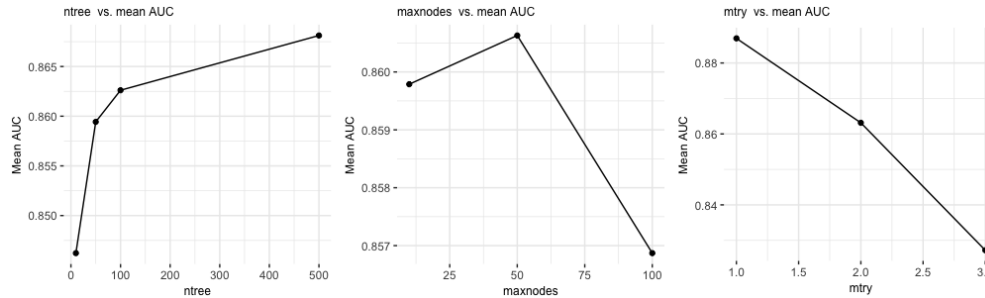
3.3.2   Systematic misclassification errors

3.3.3   Generalizability of the model

Ultimately, a model learns best about the subspaces that it has trained on, and a random forest model is no different. Decision criteria within every component decision tree of a random forest are chosen to be optimal for data whose ranges are similar to those of the training examples.

Although we don't have information about the variability of other images within the feature space of NDAI, CORR, and SD, we can reasonably assume that **the model will perform fairly well in images whose feature combinations are similar to those seen in Images 1, 2, and 3**, and will perform less accurately for images whose feature combinations are not.

4   Concluding thoughts

5   References

(1) Shi, Tao, et al. "Daytime arctic cloud detection based on multi-angle satellite data with case studies." Journal of the American Statistical Association 103.482 (2008): 584-593. (2) Random Forest Interactive Discussion with Example. `https://www.youtube.com/watch?v=ajTc5y3OqSQ`