Plankton Classification

Jeremiah Cloud
Vanessa Lepe
Faraz Niyaghi
Katy Williams

Questions of Interest

- How can we edit the images to make up for the different aspect ratios, orientation, and excess empty space (-1 values in the matrices)?
- What features can we come up with in addition to features calculated by MATLAB's bagoffeatures function?
- Using Random Forest, what model has the best predictive performance for the test set?

bagOfFeatures

Creates a “vocabulary” of SURF features

- Extracts SURF features
- Constructs visual vocabulary by reducing number of features using k-means clustering

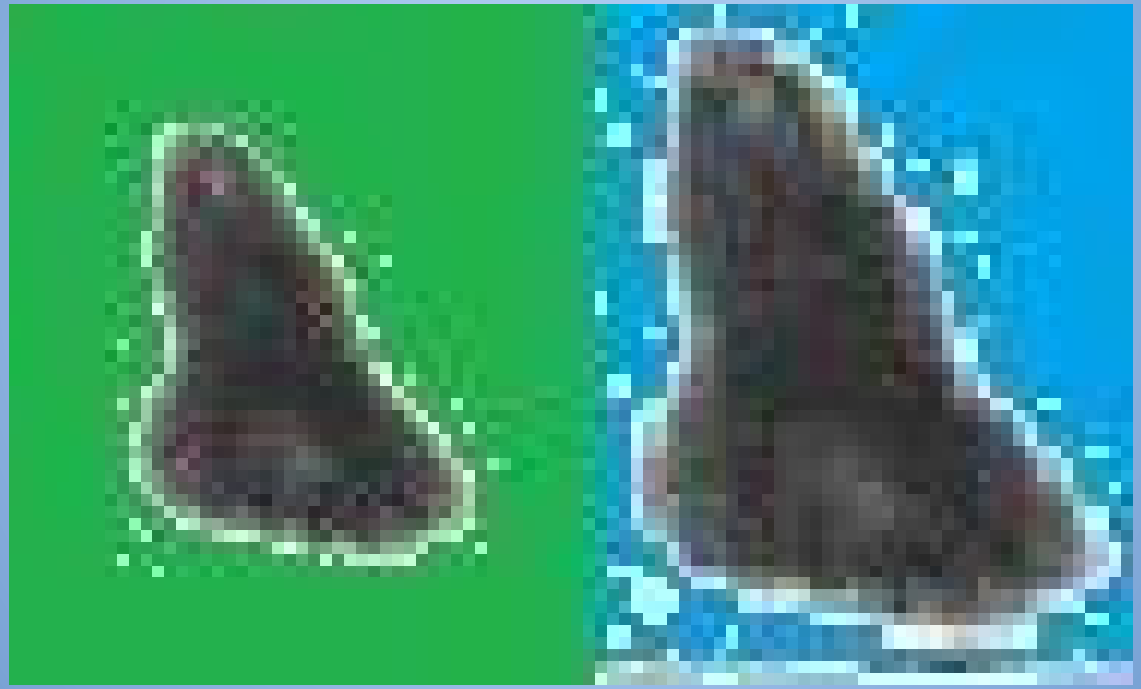
Using MATLAB

- We removed all excess rows and columns of empty white space in both the training data and the test data
- Split the training set into a sub-training set and validation set
- Used bagOfFeatures function to get the same 500 features for the sub-training, validation, and test set
- Manually calculated the mean proportion of white space per image and the length to width ratio of each image
- Converted the matrix of features to .csv for R implementation

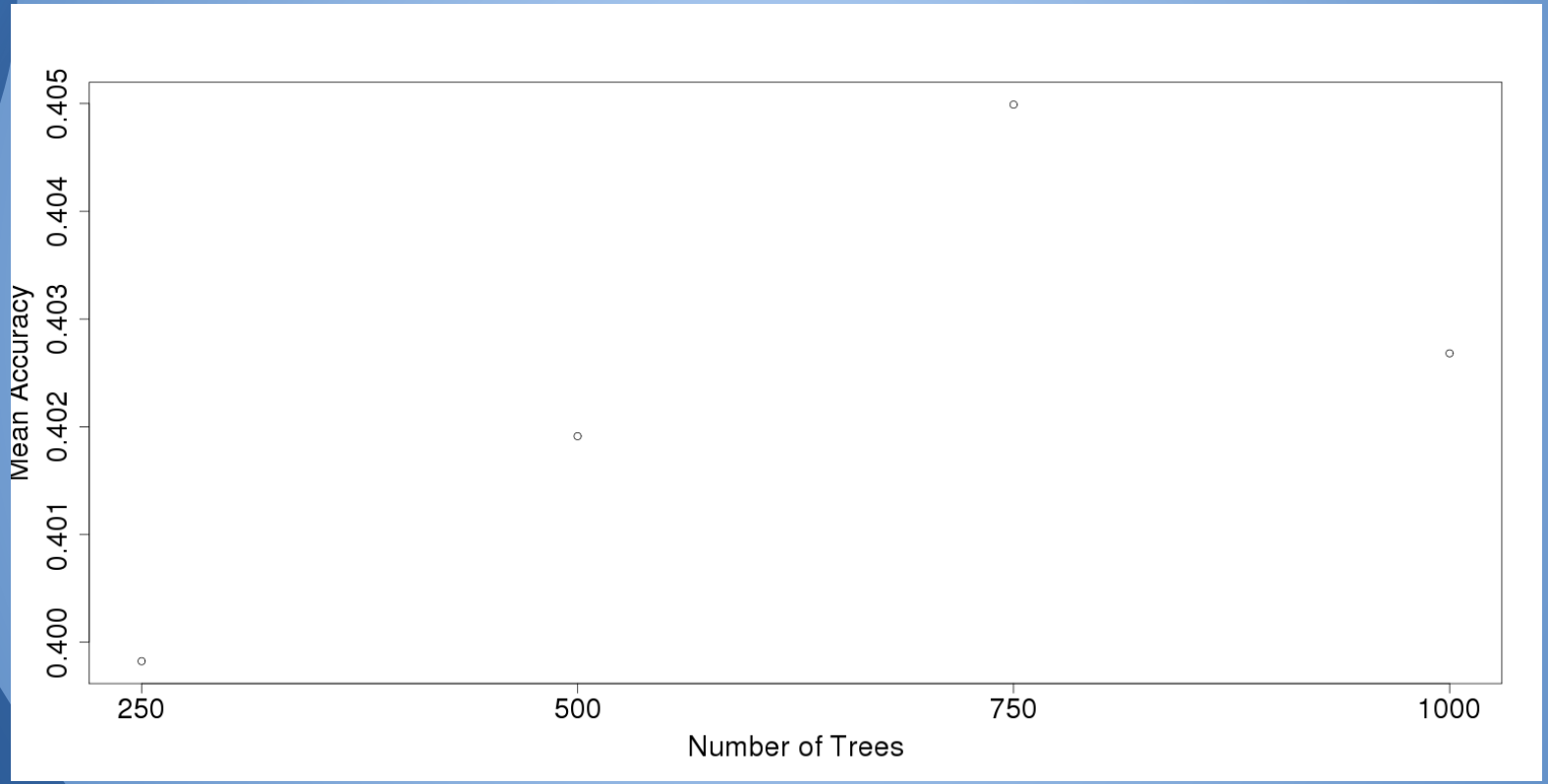
Using R

- Used Random Forest with 15 features at each node to sequentially test the number of trees up to 1000 and then each sub interval until achieving an optimal number of trees (68 in our case)
- Determined the important features from the Random Forest results for which included both of the features we manually created as two of the top features
- Fit the test data using the results from the training data
- Extracted the prediction matrix and scaled the probabilities to avoid having log(0) when uploading to Kaggle

Sample of edited image



Mean Accuracy of N Random Forest Trees



Results

629	↓5	talasag	2.926808	16	Tue, 27 Jan 2015 22:11:26 (-46.7h)
630	↓2	Mickey	2.928097	26	Mon, 23 Feb 2015 21:58:28 (-11.4h)
-		**Faraz**	**2.933300**	-	**Fri, 15 May 2015 03:33:35** **Post-Deadline**

Post-Deadline Entry
If you would have submitted this entry during the competition, you would have been around here on the leaderboard.

631	—	Yosuke Katada	2.963693	3	Thu, 25 Dec 2014 15:53:01

Complications

- Scaling Images
- Lack of image processing functions in R
- Computationally expensive

Future Work

- Implement a technique that find features for scaled images
- Compare other classification methods to Random Forest
- Use Snow and/or Rpython to break up the computations in order to train with a larger number of variables per node.