

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

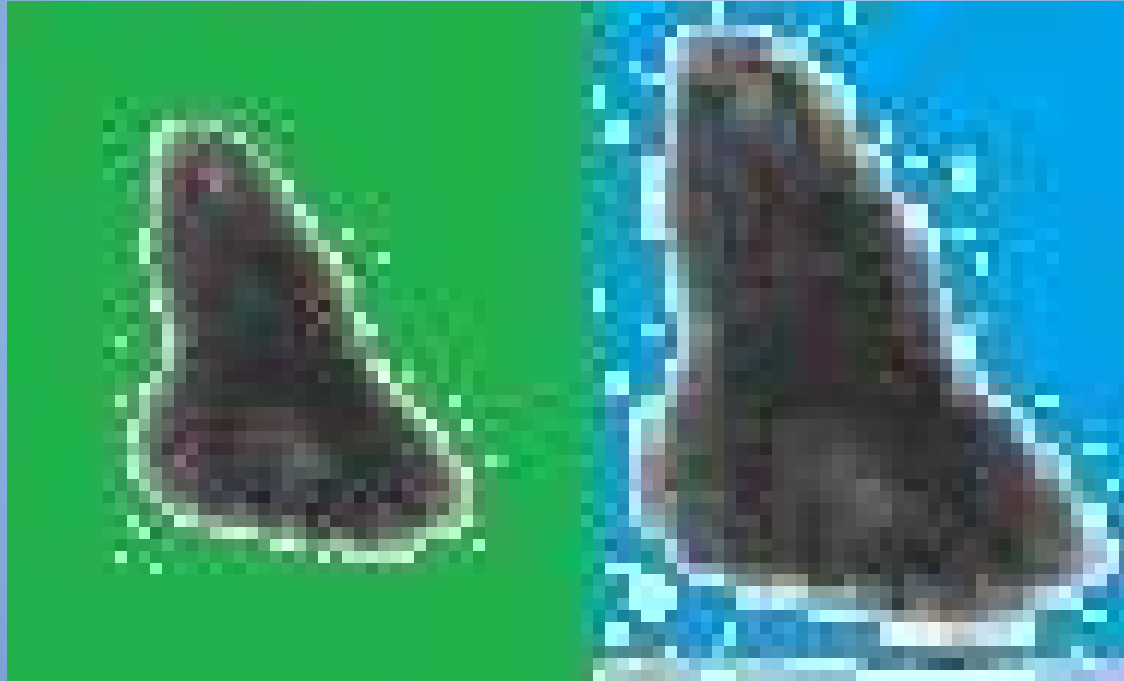
Using MATLAB

- 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

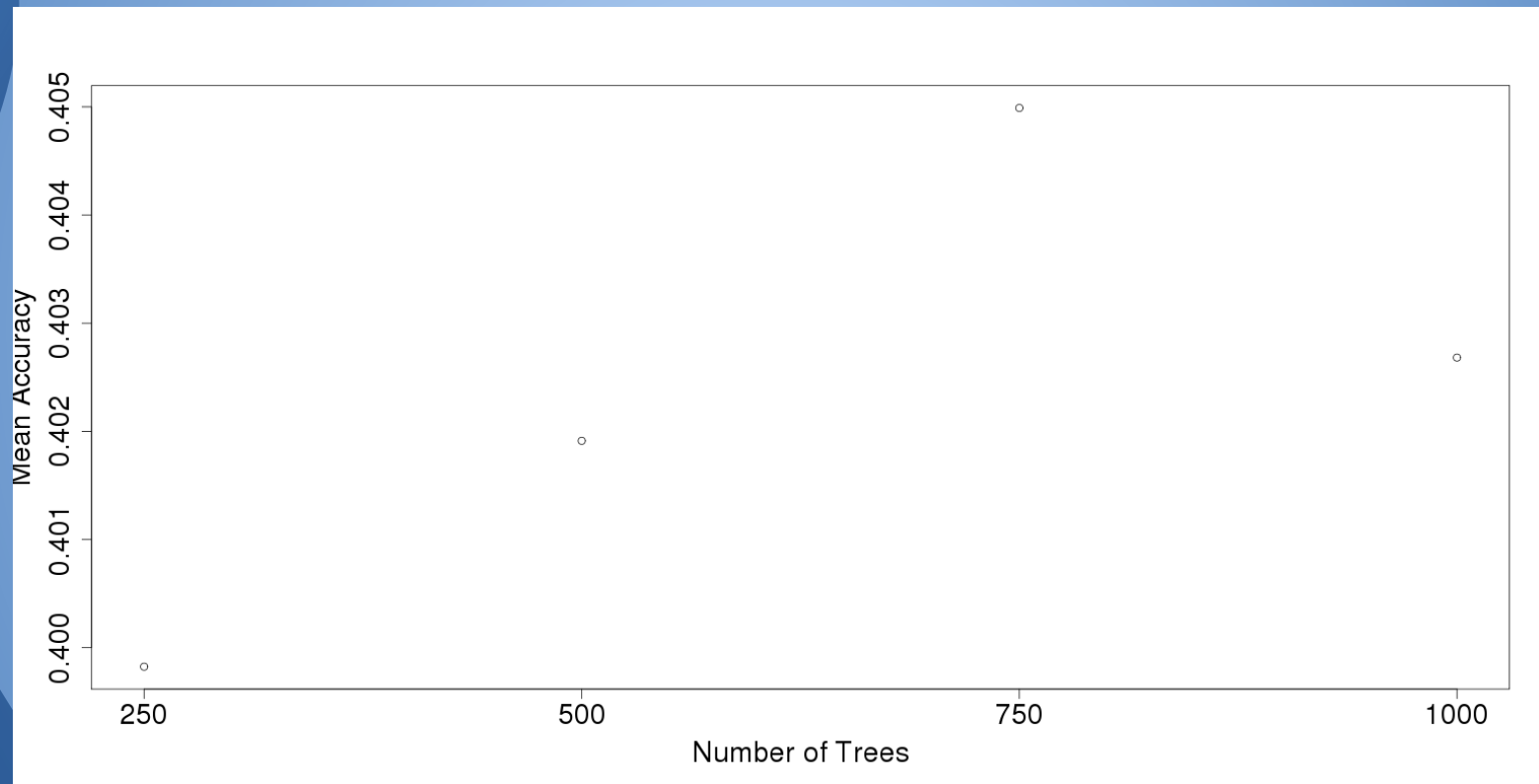
Using R

- 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

662	↑1	Sinbad 	3.326039	3	Wed, 21 Jan 2015 14:27:00
663	↑2	Averroes 	3.341117	3	Wed, 17 Dec 2014 00:05:44 (-8.2h)
-		Faraz	3.341849	-	Thu, 14 May 2015 08:53:35 Post-Deadline
Post-Deadline Entry If you would have submitted this entry during the competition, you would have been around here on the leaderboard.					
664	↓9	Uğur Güney	3.359510	3	Sat, 17 Jan 2015 00:47:20

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.