Impact of Skewed Distributions on an Automated Plankton Classifier

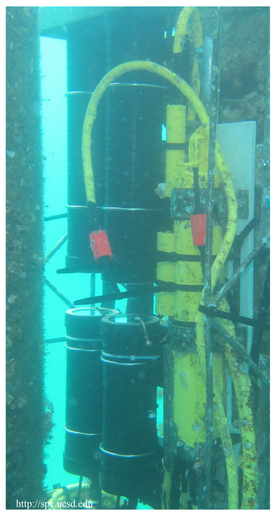
Will Chapman, Emal Fatima, William Jenkins, Steven Tien, Shawheen Tosifian

SCRIPPS INSTITUTION OF OCEANOGRAPHY

UC San Diego

Overview

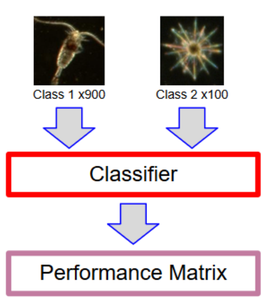
As the most abundant form of life in our oceans, plankton - microscopic organisms which include bacteria, photosynthesizing organisms (phytoplankton), and the tiny organisms that eat them (zooplankton) - are the foundation of life in the sea. Their abundance directly impacts the ecology of the oceans, as their position at the lowest trophic levels of the food chain ensure their population dynamics can be felt by even the largest of apex predators. Additionally, the combined respiratory and photosynthetic activity of plankton are major contributors to our planet's atmospheric constituents, serving as a biological "pump" and regulating both oxygen production and carbon sequestration. Yet for all their importance, much remains to be learned about the ecology and population dynamics of these ubiquitous organisms. In an effort to get a sense of their population dynamics, in 2014-2015 the Jaffe Lab for Underwater Imaging at Scripps Institution of Oceanography (SIO) developed the Scripps Plankton Camera (SPC, pictured below), which has been successfully imaging tens of thousands of these small organisms each day. Near real-time images are visible at http://spc.ucsd.edu.



The Scripps Plankton Camera (left) is mounted at the end of the pier at Scripps Institution of Oceanography (above).

With millions of images collected, accurate classification of each image is impossible to do by hand. To assist in the classification of these large data sets, machine learning algorithms have been implemented which rely on sets of manually labeled data to train the classifier. Of particular interest in this project was clarifying how the underlying statistics of the plankton population affect the performance of a classifier algorithm, in our case, the random forest. In nature, a classifier will often encounter situations where the environment of interest is largely dominated by instances of a singular class, for example, during a red tide offshore La Jolla. We were interested in seeing how the performance of a pre-trained classifier may be affected by an input with a skewed distribution. To do this, we trained a simple classifier that uses random forests and tested how its accuracy was affected by differing distributions. The results suggest that distributions that are largely dominated by one class tend to cause the classifier to underestimate the amount of correct classifications.

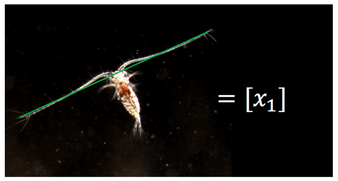
Data

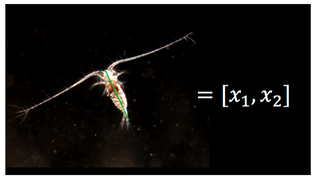


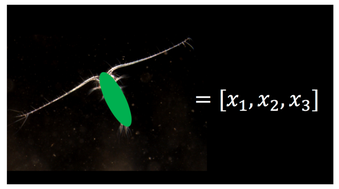
For this project, we utilized approximately 12,000 labeled images from 12 different classes composed of various plankton types. To simulate different orientations for the plankton and increase our image count, we created transformed (rotated/ sheared) copies of the images for a total of 24,000 images for 12 classes. The images varied in size from about 100-200 pixels square, and were represented in color.

Features

Images from the camera were filtered, cleaned, and segmented such that the image field of each plankton was obtained and all extraneous image information discarded. From this reduced image field, we extracted 72 features, each of which corresponded to a geometric property of the plankton image, for example, major axis length, minor axis length, aspect, elliptical properties, etc. This was accomplished algorithmically by finding contours in the image and converting them into weighted values using Fourier methods. By using geometric properties of the plankton, we are able to take into account distinguishable aspects of the data while reducing the dimensionality of the data (versus feeding the entire image into the model).



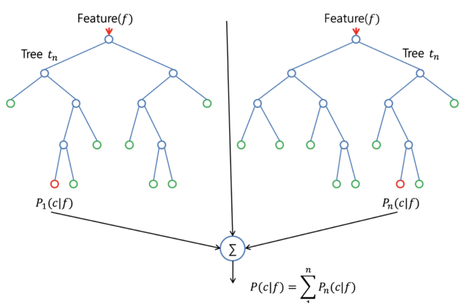




An example of feature vector generation is shown above for a copepod.

Model

The random forest algorithm relies on devising multiple decision trees during training and outputting the class that is the mode of the classes of the individual trees. Training random forests relies on the technique of bootstrap aggregating, otherwise known as bagging. In this process, samples are selected randomly with replacement from the training set. The decision trees are then fitted to these samples. What random forest does specifically is that when devising the decision trees, the split is based on a random selection of K predictors. Following training, predictions can then be made on an unknown sample x by averaging prediction from the individual trees on x. For our experiment, our distributions were generated by testing 2000 times, randomly sampling from the test population.



$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

Results

Figure 1 shows the PDFs representing the bootstrapped samples, in which the spread of the accuracy for the dominant and non-dominant classes can be observed. In Figure 2, the biases for the Chain class (top performing) dominant distribution and Agg class (worst performing) dominant distribution are shown.

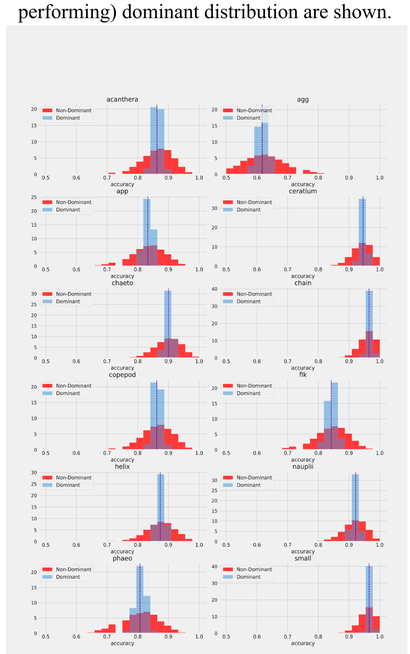


Figure 1.

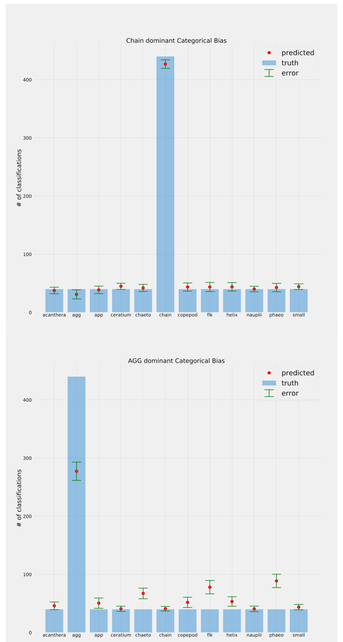


Figure 2.

Discussion

There is no meaningful shift in the means of the distribution (Figure 1), though the distribution widens, a result which can be attributed to the classifier seeing more and more images and converging to an accurate class skill. The categorical biases depicted in Figure 2 show how the Agg class dominant distribution led to poorer performance of the model, as opposed to the Chain class dominant distribution; other dominant distributions performed similarly to the Chain class. This can be attributed to the physical nature of the Agg class, which is composed of dead pieces of various plankton and those particles which cannot be classified as one of the other classes. As a result, the classifier performs poorly off as there is a varied distribution within the dominant class, leading to the difficulty of the classifier to correctly identify the class itself along with the other non-dominant classes. In addition, it appears that a general trend in the classifier's bias toward dominant classes is that the bias is generally underspecified, as seen in Figure 2.

Future

We suggest extending this experiment to another classification model, such as a convolutional neural network. By comparing our results with other models, strengths and weaknesses of the models can be better quantified. Afterwards, the models could be tested against larger and more varied distributions that are more representative of real-time conditions so their performances could be measured and evaluated against other metrics such as computational resources. Eventually, a classifier would be implemented which is capable of accurately identifying plankton images near real time.

References

1.E. C. Orenstein and O. Beijbom, "Transfer Learning and Deep Feature Extraction for Planktonic Image Data Sets," *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*

2. Orenstein, Eric., Automated Analysis of Planktonic Image Data; Lecture UCSD, 2018