ST599 Group Project #2

Spring 2015

Jeremiah Cloud

Vanessa Lepe

Faraz Niyaghi

Kate Williams

**Background**

There are two datasets of plankton images collected from Oregon State University’s Hatfield Marine Science Center. Nearly 50 million images were collected over an 18-day period, between May-June 2014, in the Straits of Florida using the In Situ Ichthyoplankton Imaging System (ISIIS). The training dataset contains approximately 30,000 images from 121 different classes of plankton. The images were of varying size and rotations. The test dataset contains plankton images, not categorized into any classes.

Using the training dataset, we will create an algorithm that assigns class probabilities to any given image in test dataset. The algorithm includes using random forests and approximately 500 features.

**Feature Extraction**

Before extracting any features, the white margins were removed from each image to reduce the noise. Due to R’s lack of image processing tools, we used a combination of MATLAB and R. Many features invariant to scaling and rotation (e.g. sift, surf, gabor features) could be extracted using MATLAB but, we decided to use the bag of features approach (also known as bag of words)[[1]](#footnote-1). This approach creates a “vocabulary” of SURF features which involves extracting the SURF features from all the images in the training dataset and then constructing a visual vocabulary by reducing the number of features using k-means clustering. This approach extracted 500 features. In actuality, 70% of the training dataset was used as the sub-training data and the remaining 30% as the validation dataset (the same features were extracted from the validation set as the sub-training dataset). This was done in order to find the optimal Random Forest to use before applying the classification model to the test dataset. In addition, two more features were included. The first feature was the length-to-width ratio of each image. The second feature was the proportion of white space per image. These same features were also extracted for the test data.

**Classification**

After obtaining the features in MATLAB, Random Forests was applied in R on the sub-training dataset and then the validation dataset was used to find the optimal random forest model. A few parameters were changed, such as the number of variables at each node (fifteen) and the number of trees (250, 500, 750, 1000). Using 750 trees produced the most accurate results. Using the MeanDecreaseGini criterion obtained from random forests, the first 100, 200, 300, 400 and 500 most important variables were used and it was found the model with the first 100 was the best. After a few more rounds of grid search, it was found that the first 68 most important variables resulted in the highest accuracy. Random forest will then be applied one last time on all the training dataset using the first 68 significant variables and 750 trees. The number of variables at each node was fixed to fifteen variables to reduce computation time.

**Results**

We will use the classification model found via Random Forests to classify each image from the test data and expect results by Friday.

**Complications**

Initially we considered scaling the images to the same size so it would be easier to extract features from. That proved to be unnecessary because it was possible to extract features that were invariant to size, as well as image rotation. We were only able to extract a few features in R because it lacks functions dealing with image processing. This then required transitioning into MATLAB, which we all only had little experience with. We were also all unfamiliar with image processing techniques so it was difficult to extract features, even in MATLAB. The majority of complications were from extracting features for the images. As with the last project, it was computationally expensive to run classification methods on the test data. Thus, the test dataset was split into subsets to run on multiple computers.

**Future Work**

Given extra time, we could first scale the images (as we had initially thought to do) and then extract features. This could allow for more features to be extracted that are not invariant to scaling and therefore potentially different results. It would be interesting to compare these results to our initial results to see if scaling the images is necessary. In addition, only Random Forests was used for classification. It would be of interest to compare this classification method to another, such as k-nearest neighbors.

1. http://www.mathworks.com/help/vision/examples/image-category-classification-using-bag-of-features.html [↑](#footnote-ref-1)