Markus Kangur (1007458038)

STA2453 EXPLORATORY DATA ANALYSIS: ZOOPLANKTON CLASSIFICATION

Introduction & Project Overview

Zooplankton are a diverse class of microscopic marine organisms that drift with the tides and currents of the waters they inhabit. The goal of this project is to train a discriminative model to classify zooplankton by species using images and auxiliary information provided in a dataset from the Ontario Ministry of Natural Resources. This report details the exploratory data analysis conducted to better understand the characteristics of the dataset through preliminary exploration and visualization. The results will be used to inform future modelling-related decisions.

Distribution of Labels

The dataset is divided by sampling location into two subsets: samples collected from Lake Simcoe and samples collected from Lake Huron. To reduce the computational burden, only plankton collected from Lake Simcoe were analyzed. When processing the data, 1218655 individual images were extracted from the mosaics and matched with their respective auxiliary information. There are 27 different labels for the images, of which 7 are of particular interest to the Ministry. The counts and percentages for each of the labels are provided below in Table 1.

About 33% of observations are from important classes, though 99% of these observations come from only two of the seven classes. The dataset is extremely unbalanced overall, with over 95% of observations coming from the four classes Calanoid_1, Cyclopoid_1, TooSmall, and Floc_1. Additionally, 13 of the 27 classes are extremely sparse, individually comprising less than 0.01% of observations. This data sparsity will likely make it difficult to adequately distinguish between classes with few training examples, especially because fewer observations will be available to train the model after splitting into training and test sets. Undersampling the majority classes, as well as augmenting the minority classes using observations from Lake Huron should be explored reduce the imbalance. Finally, one easily solvable issue is that an extra class was accidentally created when five images were mislabeled as Holopediidae instead of Holopedidae, as seen in the table.

Image Exploration

Exploring the properties of the individual images themselves also yields important information. Firstly, since the plankton images capture a wide variety of spatial orientations and positions, there is significant within-class variation. An example can be seen in Figure 1, where the nine images of Calanoid_1 plankton have very different orientations. The bottom left image also has some spherical particles or artifacts present in the image. Additionally, the coordinates provided for extracting the images from the mosaic are not exact, meaning some of the black pixels surrounding the images were included. This can be fixed by discarding a buffer a few pixels from the perimeter of each image.

Table 1: Distribution of Class Labels

IMPORTANT LABELS			OTHER LABELS		
Label	**Count**	**Percentage**	**Label**	**Count**	**Percentage**
Calanoid_1	203298	16.68%	TooSmall	633937	52.02%
Cyclopoid_1	194176	15.93%	Floc_1	135898	11.15%
Bosmina_1	2814	0.23%	Cyclo_2	20115	1.65%
Herpacticoida	604	0.05%	CountGT500	12277	1.01%
Daphnia	534	0.04%	CopepodSpp	6362	0.52%
Chydoridae	43	<0.01%	LargeZ-1	5743	0.47%
Chironomid	38	<0.01%	Bubbles	633	0.05%
TOTAL	**401507**	**32.95%**	Nauplii	1575	0.13%
			Unknown	235	0.02%
			Eggs	106	<0.01%
			Sididae	91	<0.01%
			CladoceraSpp	86	<0.01%
			Holopedidae	46	<0.01%
			Insecta	19	<0.01%
			Floc_2	10	<0.01%
			Holopediidae	5	<0.01%
			InsectLarvae	4	<0.01%
			Rotifer	3	<0.01%
			Cercopagididae	2	<0.01%
			Polyphemidae	1	<0.01%
			TOTAL	**817148**	**67.95%**

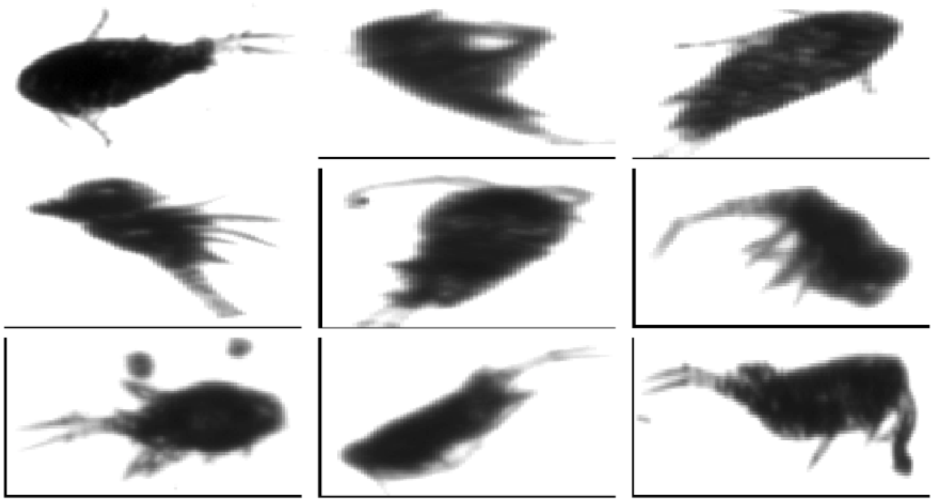

Figure 1: Example Cyclopoid_1 Images

Additionally, the individual images have very different sizes, even between classes. Figure 2 below shows a scatterplot of the image sizes for the four most prominent classes. While the maximum height and width were 1809 and 981 pixels respectively, the axes were limited to 400 by 400 pixels for visualization purposes. Importantly, many of the classes tend to cluster according to their dimensions, suggesting that image height and width could be important predictors.

However, the differences in image size also present a modelling challenge. Many convolutional neural network (CNN) architectures and dimension reduction techniques require that the input images have the same dimensions. After choosing a sufficient size, it may be necessary to pad the smaller images with whitespace or downscale the larger ones to reach a reasonable middle ground.

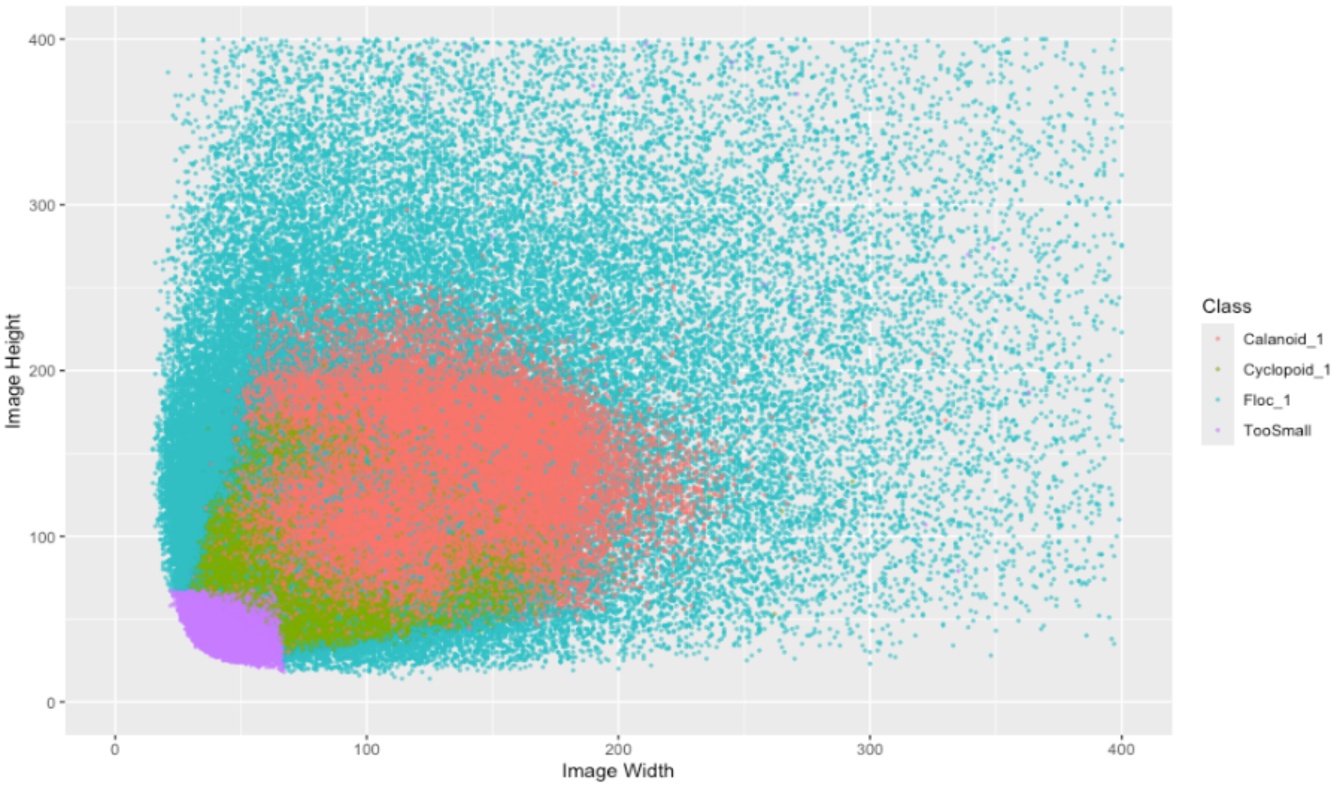

Figure 2: Scatterplot of Image Sizes

Computational Considerations

The computer did not have sufficient RAM to hold the images in memory, even after optimizing image storage by converting the 3-channel greyscale to single-channel greyscale and storing only the cropped images vectors. While there were no issues dealing with the tabular data for all the observations at once, only about 12% of images can be held in memory at once. This would leave very little computing power available for working with the data and training the model, especially for intensive methods like CNNs. As a result, major modifications must be made to the modelling process to utilize the information contained in the images. We propose using principal component analysis (PCA) to reduce the dimensionality of the images. By retaining only a few transformed features that explain a large amount of the variance in the images, we can vastly reduce the computational power required while still making use of the available information. Incorporating these features along with the geometric data would create a dataset well suited to gradient boosting algorithms like XGBoost to perform classification.

Auxiliary Information Exploration

The master table contains environmental features recorded during sampling for each mosaic, but these variables were not explored because they are constant for each mosaic, and each mosaic contains many different species. This means we expect these features to be poorly separated by class. However, the geometric features computed by the FlowCam provide a useful set of predictors. 34 meaningful numeric variables were explored including geometric features such as area, perimeter, compactness, transparency, and symmetry, as well as other image characteristics like height and width. As an example, Figure 3 below shows the relationship between indices of convexity and circularity for the four most common classes.

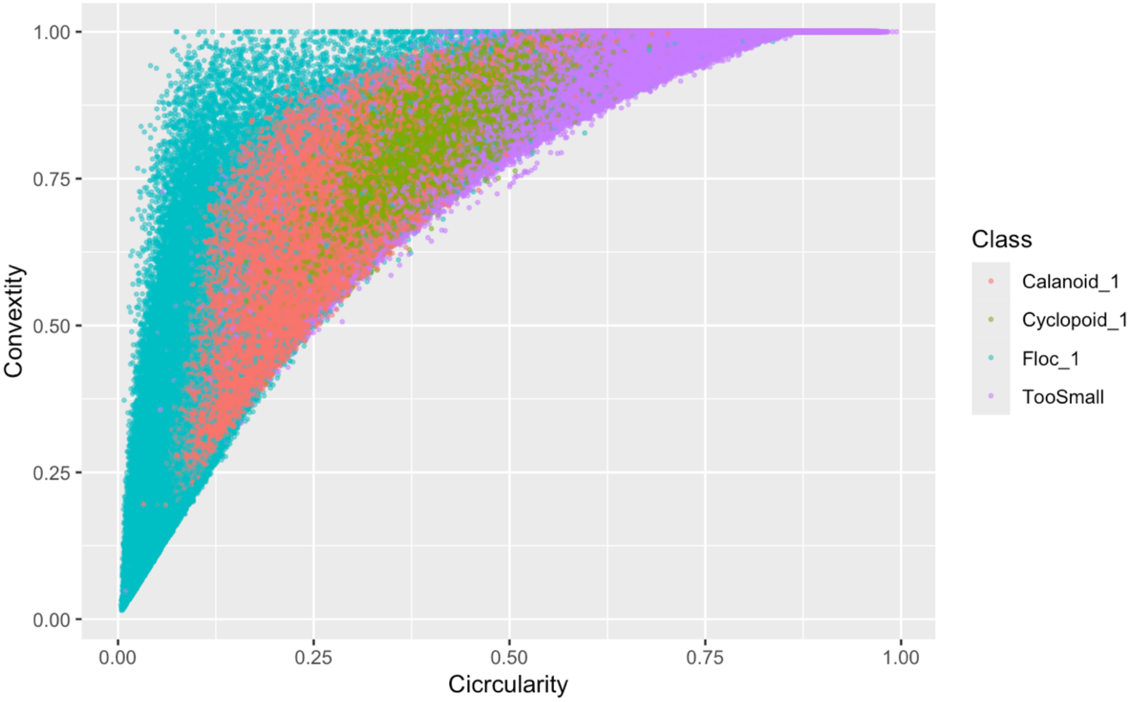

Figure 3: Convexity vs. Circularity

While there is not sufficient space to examine all variables in detail here, we present the results of PCA conducted on the 34 variables considered below in Figure 4. Many outliers can be seen in the graph. The first principal component explains about 44% of the variance, while the second explains about 17%, and the first nine principal components explain 90% of variance. Some clustering of classes can be observed, indicating these variables (or the principal components themselves) could be used alongside the image data during modelling.

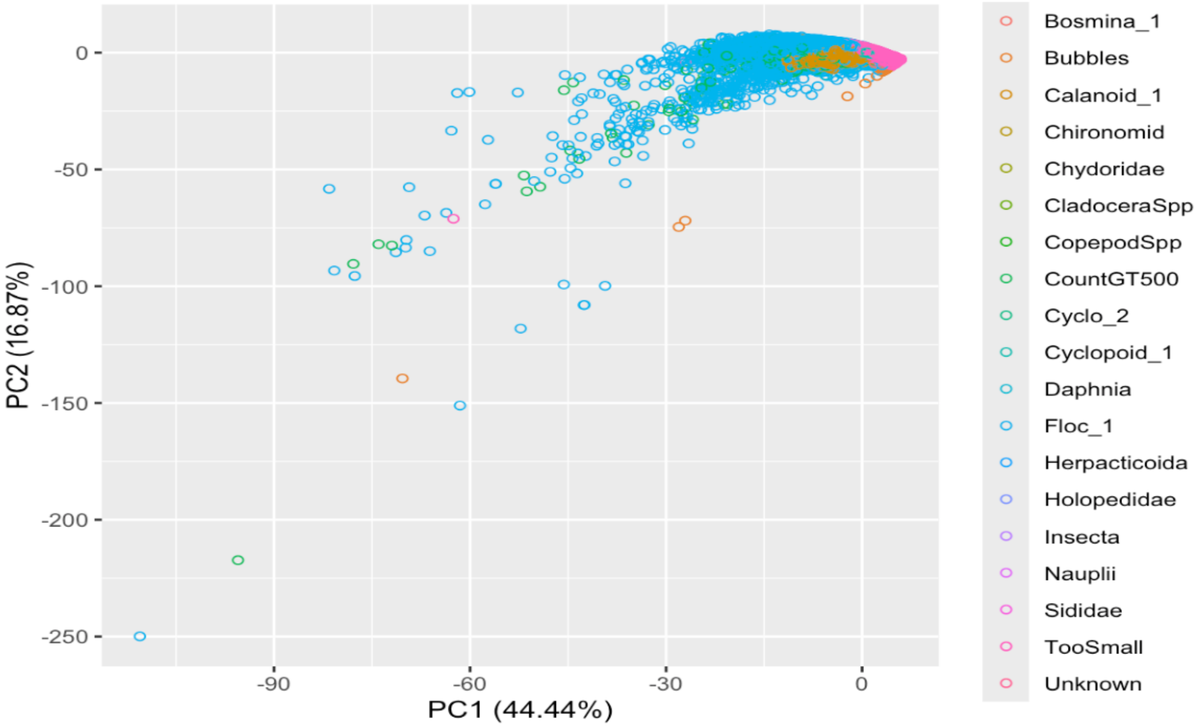

Figure 4: Principal Component Analysis of Geometric Variables