Markus Kangur (1007458038)

STA2453 FINAL REPORT: ZOOPLANKTON CLASSIFICATION

Introduction

Zooplankton are a diverse class of microscopic marine organisms that drift with the tides and currents of the waters they inhabit. They play an important role in the aquatic food chain as a source of food for larger marine organisms. Importantly, zooplankton can be highly sensitive to environmental conditions, meaning zooplankton species composition can be used as an indicator of ecosystem health.

However, manual classification is difficult, time consuming, and can be vulnerable to observation bias if multiple people are contributing to the classification effort. The goal of this project is to use the images and auxiliary information provided in a dataset from the Ontario Ministry of Natural Resources (MNR) to train a discriminative model that can classify zooplankton. In particular, approaches such as XGBoost and logistic regression were applied to estimate a function that maps the inputted data to the appropriate discrete set of plankton species labels.

Data Overview

The MNR has produced a labeled dataset with over two million plankton observations. Environmental features such as depth, water temperature, and sampling time and location were recorded during field sampling in Lake Huron and Lake Simcoe. The FlowCam platform was used to produce the images, followed by manual labeling of the 30 different classes of zooplankton. Along with image dimensions, the dataset includes geometric features computed by the FlowCam such as transparency, symmetry, compactness, and volume.

Exploratory data analysis was conducted to better understand the characteristics of the dataset through preliminary exploration and visualization. Observations tended to cluster according to class labels when plotted on different axes, such as perimeter versus symmetry and image height versus image width. This suggests that even though the images themselves aren't utilized, the geometric features and image characteristics present in the tabular data still provide a significant amount of information.

Many of the predictor variables exhibited high pairwise correlation. This was often due to variables measuring a similar property or being derived from other variables. In a principal component analysis of the 34 numeric variables from the FlowCam, the first eight principal components explained over 90% of the variance, indicating significant opportunities for variable selection and dimension reduction.

Data Preprocessing

Variables that didn't facilitate future prediction or provide meaningful information about the zooplankton were removed, such as processing date, ID numbers, sampling location, and the year of sample collection. Retaining only average depth and removing the maximum and minimum depth values ensured the data matrix was not singular. Variables with over 180,000 entries of missing data were removed before removing any observations with missing data. This ensured that over 96% of observations were retained. Importantly, this step did not significantly alter the distribution of class labels. A data dictionary of the remaining 50 variables is available in Appendix 1.

The MNR is primarily concerned with identifying seven key types of zooplankton: Calanoids, Cyclopoids, Bosmina, Harpacticoids, Chironomids, Chydoridae and Daphnia. Class labels were recoded as factors where all unimportant observations were aggregated into a single class, reducing the total number of classes from 30 to 8. This vastly reduces the computational time associated with performing multiclass classificiation.

Table 1: Distribution of Zooplankton Labels

Label	Coding	Count	Percentage
Other	7	1524279	75.73%
Calanoid_1	1	254139	12.63%
Cyclopoid_1	4	222616	11.06%
Bosmina_1	0	7857	0.39%
Herpacticoida	6	1964	0.10%
Chironomid	2	1335	0.07%
Daphnia	5	594	0.03%
Chydoridae	3	110	<0.01%
TOTAL	**--**	**2012894**	**100.0%**

The counts and percentages for each of the labels in the dataset are provided above in Table 1. Almost a quarter of observations are from important classes, though over 97% of these observations come from only two of the seven important classes. The dataset is extremely unbalanced overall, with over 99% of observations labelled as Other, Calanoid_1, or Cyclopoid_1. Aggregation helped address some of the data sparsity concerns, since most of the extremely sparse classes were aggregated into the Other class.

Methods

The processed data were randomly split into an 80% training set and a 20% test set before two different modeling approaches were applied. In the first approach, extreme gradient boosting (XGBoost) algorithms were applied directly to perform multiclass classification. XGBoost is a modern gradient boosting library that implements parallel tree boosting. It is a very popular machine learning algorithm for solving data science problems that is both flexible and highly optimized. An R package provides a ready-to-run implementation that can be used for multiclass classification.

XGBoost computes a normalized importance metric that describes how important each predictor variable is in the model and then clusters them. This importance metric provides a convenient criterion to perform variable selection. While XGBoost is quite robust to correlated predictors, excluding unimportant predictors can reduce overfitting and improve model generalizability.

Three different XGBoost models were fit: one full model with all predictors, and two reduced models selected based on importance. The learning rate was set to 0.5 to balance protection against overfitting with computational time so that fewer training rounds were needed. The maximum tree depth was kept at the default of 6 over 60 initial training rounds with the objective function set to the multiclass softmax function.

In the second approach, logistic regression was used to perform binary classification and predict whether the input is from an important class. The input proceeds to XGBoost multiclass classification only if it is predicted to be from an important class, helping address the imbalance in the dataset. A receiver operating characteristic (ROC) curve was used to determine the optimal threshold values for prediction purposes.

However, the IRLS algorithm for fitting the logistic regression failed to converge. The deviance was not uniformly decreasing and often increased significantly before getting “stuck” at a local minimum. Examining correlations between predictors and computing their VIFs revealed extremely high multicollinearity in the dataset. While multicollinearity isn’t often a concern in prediction problems, in this case it was preventing the logistic regression from being fit. Unfortunately, it was also unclear which variables to remove due to the complex correlation structure. Deleting variables in decreasing order of VIF values was also unsuccessful.

To remedy this issue, principal components regression was applied. Since the primary focus is to reduce multicollinearity, not dimension reduction, enough principal components were retained to account for 99% of the variation in the dataset. This approach allowed IRLS to converge successfully. Finally, the performance of the models was evaluated using a combination of training accuracy, test accuracy, and a confusion matrix as discussed below.

Results

Table 2: Preliminary Model Performance

Model Structure	Total Time	Time Per Round	Training Accuracy	Testing Accuracy
Full Model	1:23:28	83 seconds	96.02%	95.61%
Reduced Model 1 (18 variables)	0:38:14	38 seconds	95.82%	95.45%
Reduced Model 2 (26 variables)	0:46:01	46 seconds	95.94%	95.54%
PCA + Logistic + XGBoost	0:17:02	17 seconds	90.14%	89.82%

The results of the preliminary model training with 60 boosting rounds are shown above in Table 2. Since cross validation was extremely computationally intensive, as an alternative the training and testing accuracy rates were compared to guard against overfitting. Overall, there was little difference between these rates, indicating good model generalizability.

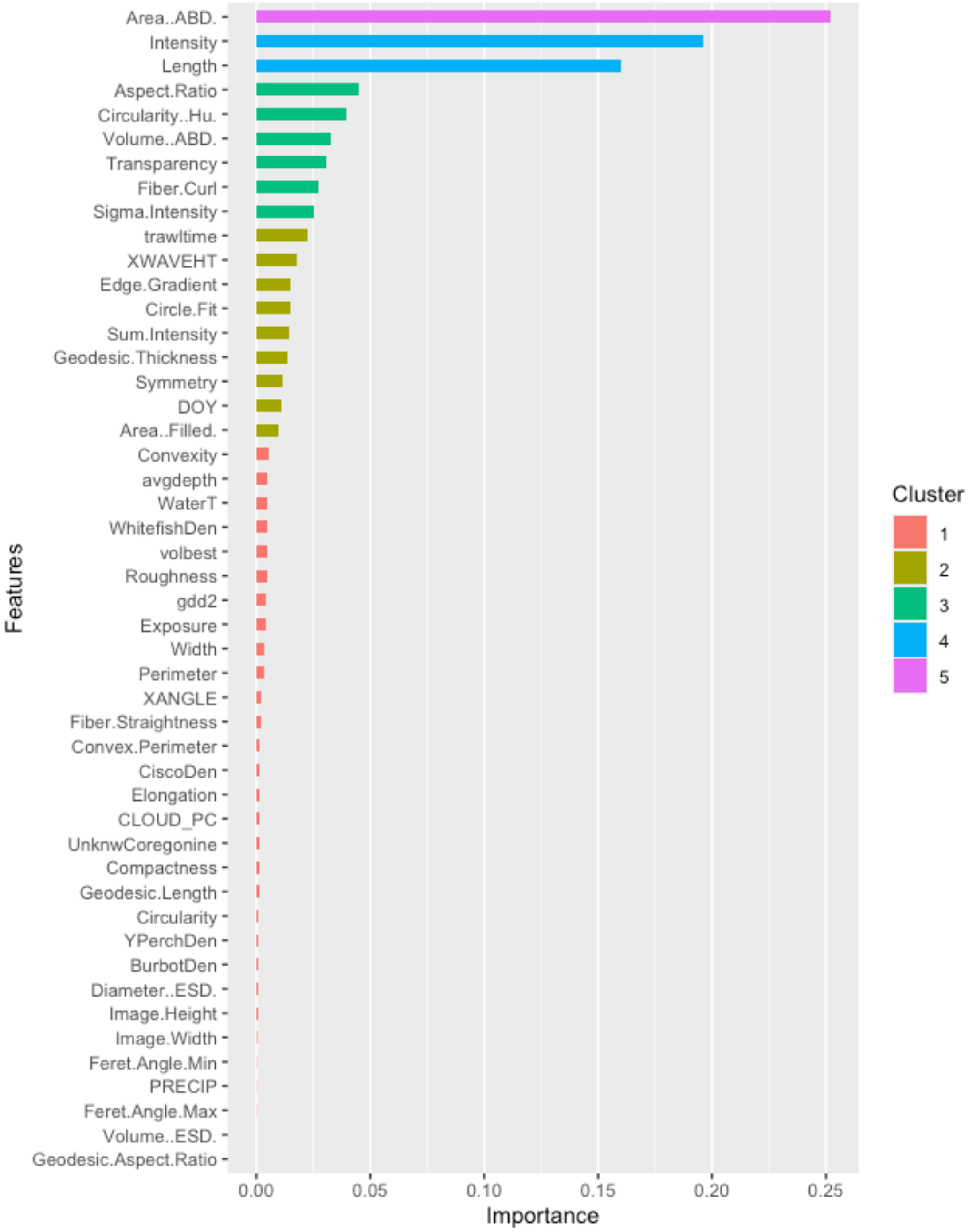


Figure 1: Importance plot for the full XGBoost model.

Figure 1 above the importance plot for the full model, where the predictors are clustered into five classes by their relative importance. In the first reduced model, we excluded cluster 1, the predictors with the lowest importance. This model retained 18 variables and 93.8% of the importance in the full model. In the second reduced model, we excluded variables less important than the Exposure variable. elbow point on the importance graph, retaining 26 variables explaining 97.6% of the importance. The full and reduced models had very similar accuracy rates, differing mainly in XGBoost computation time.

In the PCA + Logistic + XGBoost model, we needed to include 32 principal components to retain 99% of variance explained, all of which were statistically significant in the logistic regression. The ROC curve and AUC value is shown above in Figure 2. The logistic model exhibited strong discriminative power with an AUC of 0.97. To balance sensitivity and specificity, the optimal cutoff vale of 0.223 was chosen. This model had much lower training accuracy than the other models, but much faster XGBoost training time since it was trained only on classes 0-6.

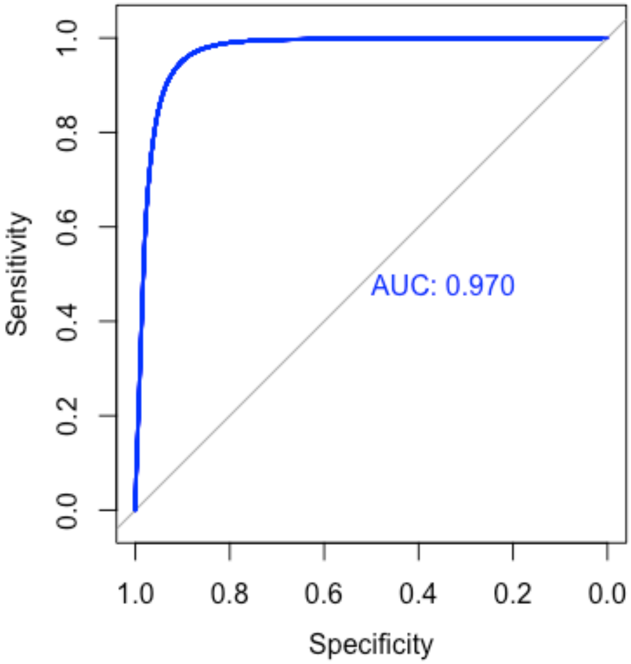
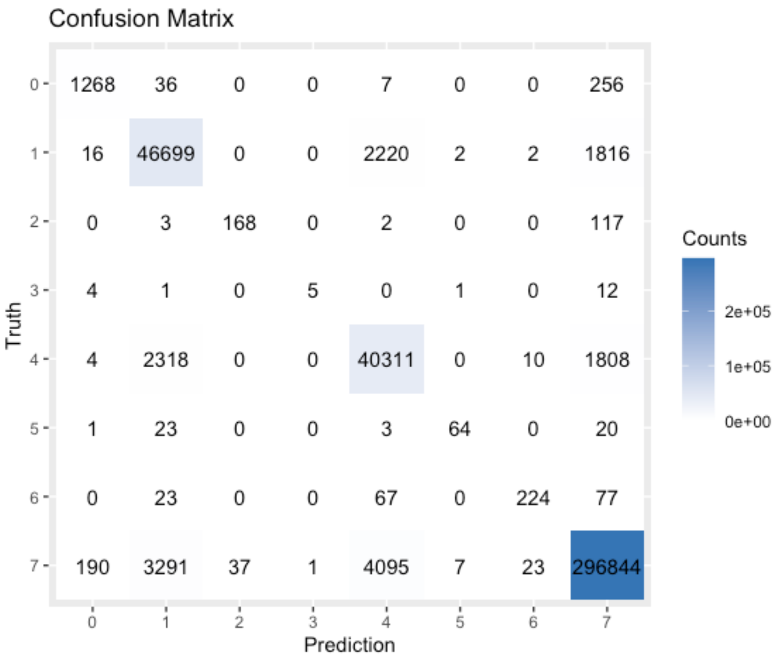


Figure 2: ROC curve for the logistic regression.

Reduced model 2 was chosen for the final model since its training accuracy was nearly the same as the full model, yet it required significantly less training time and relied on fewer variables. After 100 boosting rounds, the final model achieved 96.53% training accuracy and 95.90% testing accuracy. The confusion matrix for the final model when predicting on the test set is provided below in Figure 3. From the confusion matrix we can see that the model performs well overall, with most observations falling on the diagonal. Recall that Class 3 (Chydoridae) has the fewest observations, making up less than 0.01% of the data. While only 5 of 23 true Chydoridae in the test set were correctly classified, 5 of 6 predicted Chydoridae are correct. Class 5 (Daphnia), the next smallest class, also has similar behaviour. Otherwise, while some large values do occur off the diagonal, these are from classes with a lot of observations, meaning they make up a small percentage of observations in the class.



Truth \ Prediction	0	1	2	3	4	5	6	7
0	1268	36	0	0	7	0	0	256
1	16	46699	0	0	2220	2	2	1816
2	0	3	168	0	2	0	0	117
3	4	1	0	5	0	1	0	12
4	4	2318	0	0	40311	0	10	1808
5	1	23	0	0	3	64	0	20
6	0	23	0	0	67	0	224	77
7	190	3291	37	1	4095	7	23	296844

Figure 3: Confusion matrix for the final model.

Discussion & Limitations

Overall, the final model achieved very high testing accuracy even without directly using the zooplankton images. As seen in the importance plots, the most important predictors in the XGBoost algorithms were the geometric features, which were originally extracted from the images. This suggests we can view the geometric features as a type of feature extraction, somewhat analogous to how a CNN classifies images. The PCA + Logistic + XGBoost model was not as effective at predicting class labels as the other models. One possible reason for this is that the importance values don't follow the order of the principal components, meaning parts of the data with low variance are still often important for prediction.

While the plankton images would be a valuable source of information, 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. Only about 10% of images could be held in memory at once, leaving very little computing power available for working with the data. The variation in image sizes also presents a modelling challenge, since many feature extraction techniques require the image dimensions to be the same. As a result, only the tabular data were analyzed.

Finally, importance weighting was explored in an attempt to account for the class imbalances. Weights were computed on a per observation basis such that each class was weighted equally during XGBoost training. However, this approach yielded much poorer performance overall, especially on sparse classes.

Conclusion

Using a labeled dataset from the MNR, a discriminative model based on the XGBoost algorithm achieved 96% accuracy when classifying important zooplankton species. While there were some misclassification issues stemming from data sparsity and class imbalances, the model performed strongly overall. This automated approach to zooplankton classification can facilitate cost-effective and efficient ecological monitoring for government and conservation authorities. Future work could focus on reducing classification error in sparse classes by either collecting more observations during additional field sampling or by using methods such as over or under sampling.

Appendix 1: Data Dictionary

Variable Name	**Description**
Area..ABD.	Number of pixels in the thresholded greyscale image converted to an area measure.
Area..Filled.	The area represented by the particle edge and all the pixels inside the edge.
Aspect.Ratio	The ratio of the lengths of the axes of the Legendre ellipse of inertia of the particle.
Circle.Fit	Normalized deviation of the particle edge from a best-fit circle.
Circularity	A shape parameter computer from the perimeter and the filled area.
Circularity..Hu.	Alternative measure of circularity.
Compactness	A shape parameter derived from the perimeter and the filled area. Inverse of Circularity.
Convex.Perimeter	No description provided.
Convexity	A shape parameter computed as the ratio of the filled area to the area of the convex hull.
Diameter..ABD.	Diameter based on a circle with an area equal to the ABD area.
Diameter..ESD.	The mean value of 36 feret measurements.
Diameter..FD.	No description provided.
Edge.Gradient	Average intensity of the pixels on the outside border after applying a Sobel Edge Detect filter.
Elongation	The inverse of Geodesic Aspect Ratio.
Feret.Angle.Max	Angle of the largest feret measurement.
Feret.Angle.Min	Angle of the smallest feret measurement.
Fiber.Curl	A shape parameter computed from Geodesic Length and Length.
Fiber.Straightness	A shape parameter computed from Geodesic Length and Length.
Geodesic.Aspect.Ratio	The ratio of Geodesic Thickness to Geodesic Length. Inverse of Elongation.
Geodesic.Length	Length obtained by modeling the particle as a rectangle.
Geodesic.Thickness	Thickness obtained by modeling the particle as a rectangle.
Image.Height	Image height in pixels.
Image.Width	Image width in pixels.
Intensity	The average grayscale value of the pixels in the particle.
Length	The maximum value of 36 feret measurements.
Perimeter	The length of the particle edge not including edges of holes in the particle.
Roughness	Measures irregularity of the particle’s surface as the ratio of perimeter to convex perimeter.
Sigma.Intensity	Standard deviation of greyscale values.
Sum.Intensity	Sum of greyscale pixel values.
Symmetry	A measure of the symmetry of the particle about its center.
Transparency	A measure of transparency computed from the ABD Diameter and ESD Diameter.
Volume..ABD.	Sphere volume calculated from ABD Diameter.
Volume..ESD.	Sphere volume calculated from ESD Diameter.
Width	The minimum value of 36 feret measurements.
DOY	Julian day of the year.
gdd2	Growing degree days, a measure of heat accumulation.
WaterT	Surface water temperature
Avgdepth	Mean water depth.
Trawltime	How long the trawl was trawled.
XANGLE	A measure of accumulated wind forcing.
PRECIP	Precipitation codes as a ranked factor.
XWAVEHT	Wave height in metres.
CLOUD_PC	Percent cloud cover.
volbest	The volume of water that was sampled.
WhitefishDen	Density of larval lake whitefish.
UnknwCoregonine	Density of larval coregonines of unknown species.
CiscoDen	Density of larval cisco.
Exposure	A measure of accumulated wind forcing.
YPerchDen	Density of larval yellow perch.
BurbotDen	Density of larval burbot.