Markus Kangur (1007458038)

STA2453 PROGRESS REPORT: ZOOPLANKTON CLASSIFICATION

Introduction

Zooplankton (singular zooplankter) 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, we are interested in applying machine learning methods 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 to be used for this purpose, where they are primarily concerned with identifying 7 key types of zooplankton: Calanoids, Cyclopoids, Bosmina, Harpacticoids, Chironomids, Chydoridae and Daphnia. After completion of field sampling in Lake Huron and Lake Simcoe, they used a FlowCam platform to image the plankton, followed by manual labeling of the 30 different classes of zooplankton. When processing the data, information corresponding to 209384 individual images were extracted and compiled into a single dataset.

Along with class labels and image dimensions, the dataset includes two main classes of variables; geometric features computed by the FlowCam (transparency, symmetry, compactness, volume, etc.) and environmental features recorded during sampling (latitude, longitude, depth, date, etc.). Every image from the same mosaic has the same values for the environmental features because they were recorded at time of sampling. They are therefore expected to have poor separability by class, since each mosaic contains many different zooplankton species. However, the geometric features computed by the FlowCam for each individual image contains over 30 meaningful numeric variables. During exploratory data analysis, many of these variables showed clustering by labels when plotted, such as convexity versus circularity and perimeter versus symmetry. There are also a significant number of outliers, likely due to a combination of measurement error from the FlowCam and high natural variability.

Table 1: Distribution of Class Labels

IMPORTANT LABELS			OTHER LABELS		
Label	**Count**	**Percentage**	**Label**	**Count**	**Percentage**
Calanoid_1	274579	13.11%	TooSmall	1053399	50.31%
Cyclopoid_1	231692	11.07%	Floc_1	378806	18.09%
Bosmina_1	8053	0.38%	LargeZ-1	80554	3.85%
Herpacticoida	1984	0.09%	Cyclo_2	23622	1.13%
Chironomid	1336	0.06%	CountGT500	16905	0.81%
Daphnia	669	0.03%	CopepodSpp	10758	0.51%
Chydoridae	118	0.01%	Bubbles	5473	0.26%
TOTAL	**518431**	**24.76%**	Nauplii	2585	0.12%
			Unknown	1256	0.06%
			CladoceraSpp	907	0.04%
			Sididae	492	0.02%
			InsectLarvae	148	0.01%
			Holopediidae	131	0.01%
			Floc_2	111	0.01%
			Eggs	109	0.01%
			Insecta	69	<0.01%
			Naididae	36	<0.01%
			Nematode	30	<0.01%
			Rotifer	8	<0.01%
			Cercopagididae	6	<0.01%
			Polyphemidae	3	<0.01%
			Leptodoridae	1	<0.01%
			Trombidiforme	1	<0.01%
			TOTAL	**1575410**	**75.24%**

The counts and percentages for each of the labels in the dataset are provided above in Table 1. About 25% of observations are from important classes, though 98% of these observations come from only two of the seven classes. The dataset is extremely unbalanced overall, with over 93% of observations coming from the four classes Calanoid_1, Cyclopoid_1, TooSmall, and Floc_1. Additionally, 13 of the 30 classes are extremely sparse, individually comprising at most 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.

While the images themselves aren't utilized, the geometric features and image characteristics present in the tabular data still provide a significant amount of information. For example, when examining the relationship between image height and image width, it was found that many of the classes tend to cluster according to their dimensions, suggesting that they could be important predictors.

Additionally, many of the predictor variables are highly correlated. For example, there were many variables measuring the same property, such as three for diameter and three for circularity. In a principal component analysis of the 34 numeric variables from the FlowCam, the first principal component explained about 44% of the variance, while the second explained about 19%, and the first eight explained over 90%. This method could be used to reduce the dimensionality of the data while ensuring the predictors aren't highly correlated, but this will sacrifice interpretability. Importantly, many of the methods discussed below are robust to correlated predictors, so this will only be explored if modelling issues occur.

Methods

We first conducted exploratory data analysis to better understand the characteristics of the dataset through preliminary exploration and visualization. Variables that were constant or did not provide meaningful information about the plankton were removed, such as date of sample processing and ID numbers. The prepared dataset initially had 96 columns, of which 27 were removed at this time. Next, classes with under 10 instances were excluded due to sparsity, especially after splitting into training and testing sets. Several typos in class labels were also corrected. Categorical variables, including the class labels, were re-coded as factors.

After splitting the prepared data into a 75% training set and 25% test set, two different modeling approaches were applied. In the first approach, we conducted multiclass classification immediately to assign each of the inputs to one of the 30 classes. In the second approach, we applied logistic regression to conduct binary classification and predict whether the input is from an important class. A receiver operating characteristic (ROC) curve was used to determine the most appropriate threshold values for prediction purposes. The input proceeds to the multiclass classification algorithm only if it is predicted to be from an important class. This reduces the multiclass step from 30 to 7 classes, potentially simplifying the model and improving performance.

We applied two different machine learning algorithms to perform multiclass classification: extreme gradient boosting (XGBoost) and support vector machines (SVMs). XGBoost is a modern gradient boosting library that implements parallel tree boosting. It is one of the most popular machine learning libraries for regression and classification, and it is highly optimized and computationally efficient. SVMs are a classical machine learning approach finds optimal hyperplanes that maximize data separability in high dimensions.
While well suited for classification, they can be computationally intensive. Readily available R packages provide ready-to-run implementations of both approaches.

After training the models on the training set, the labels of the test set were predicted. Model performance was evaluated using a confusion matrix. Extra emphasis was placed on classification accuracy within the class of important zooplankton species. This helped mitigate the effects of the unbalanced dataset, since high accuracy rates can be achieved simply by predicting the most common classes while still misclassifying relatively uncommon classes.

Results

Preliminary results indicate that logistic regression is quite effective at differentiating between important and unimportant plankton classes, achieving a prediction accuracy of approximately 87%. This accuracy can likely be improved by tuning the cutoff probability. The next modeling work will include implementing the multiclass classification algorithms and evaluating their performance.

Limitations

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. While there were no issues dealing with the tabular data for all the observations at once, only about 10% 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 convolutional neural networks. The variation in image sizes also presents a modelling challenge, since many dimensional reduction techniques such as principal component analysis require the images to be stored as vectors of common size. As a result, only the tabular data was analyzed.

Conclusion

In progress.