Team GAIASAVERS project proposal : plankton identifcation

A. Petit, W. Bouzouita, T. Babinet, M. Chor, E. Wang

1 Background

Among the upcoming challenges of the 21st century, reducing pollution to limit our impact on nature will be very important. In particular, a special care will need to be given to our water bodies as every living being on our planet needs water.
Over the years, several studies have been conducted to find ways to assess the quality of water. As we can see in [NRMA15], a link has been found between the quality of water and the presence of plankton in said water. As water quality increases, the abundance and diversity of plankton does as well. Being an important source of food as well, its prosperity is strongly linked with the ocean ecological balance. In order to facilitate the estimation of water quality, the aim of our project is to build a plankton classifier. This would help to identify them more easily and therefore, be able to better estimate the variety of plankton species found in a given body of water.

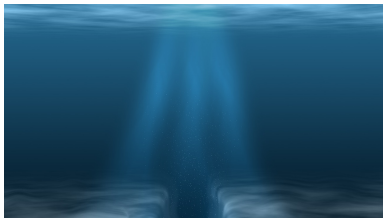

Figure 1: Plankton underwater

2 Material and method

The dataset we chose to use for this problem is the Bering Sea dataset, an in situ plankton dataset published in May 2019 by Kaichang Cheng [Che19]. It was originally used to prove the efficiency of a convolutional neural network.
The original dataset was containing 17920 images split between the train set and the test set as shown in table 1 but its size was almost 4 GB. We reduced the size of all the images to fit in a 300 pixels by 300 pixels square and added white padding when needed. This brought the size of the dataset down to 160 MB.
Due to the even spread of our images in the various categories, the accuracy in each class will have exactly the same weight and to measure the effectiveness of our classifier, we can simply use the accuracy over our whole dataset.

	Chaetognatha	Copepoda	Euphausiids	Fish larvae	Limacina	Medusae	Other
Train set	2048	2048	2048	2048	2048	2048	2048
Test set	512	512	512	512	512	512	512

Table 1: Number of examples from each class in train and test set

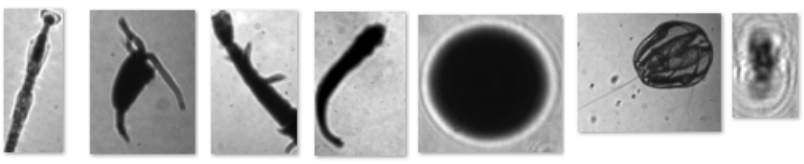

Figure 2: A sample of each class

All of the images are in various shades of gray. As such, we can represent every image by a vector of 90000 (300×300) features where each feature is a float value between 0 and 255 representing the brightness of a pixel (0 being a black pixel and 255 a white one). It is also possible to reduce the size of the images to reduce the number of features as seen in figure 3 where we reduced the size to 100×100 pixels.

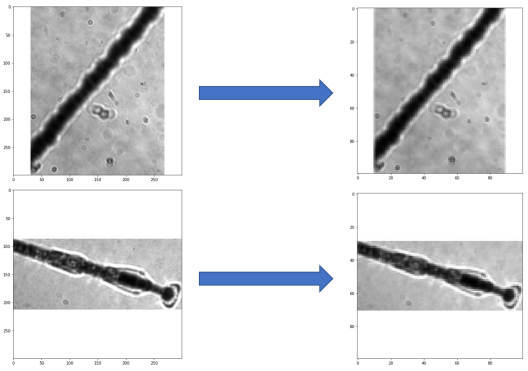

Figure 3: Example of images before and after dimensionality reduction

Then, based on these features, the goal is to predict to which class the image belongs. In order to do so, a multi-class classifier among the wide range available must be used. Simpler models like a Naive Bayes or a Decision tree can be used. They present the advantage of being easier to understand. Some results for this kind of models will be shown in the last section of this proposal. Another more complex option is to use neural networks and in particular convolutional neural networks which are particularly adapted for image analysis. These models often help to reach better performances, however they are often more troublesome to set up and to optimise due to their opaque nature.

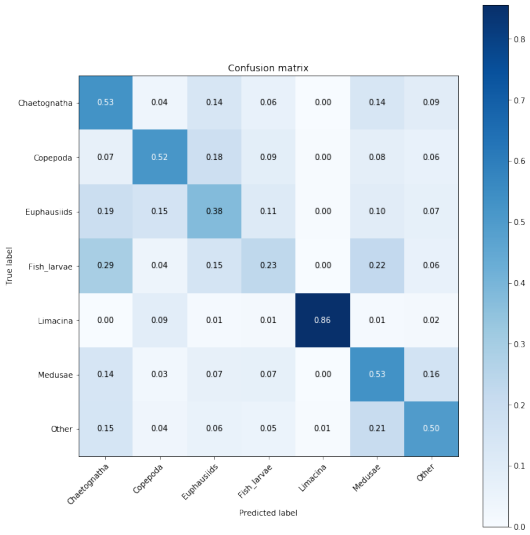

Figure 4: Confusion matrix between the different classes

One more thing we may have to focus on to improve the performance of the classifier is how to separate similar classes. As we can see in figure 4, the Limacina class reaches 86% of good predictions but the Fish larvae class only has 23% of good predictions. It means that focusing on better predicting the latter may result in a substantial improvement in performance.

3 Preliminary results

To see the potential difficulty of our subject, we have used several basic models and measured their performance. For the preprocessing of the data, we reduced the size of our images from 300×300 pixels to 100×100 by aggregating squares of 9 pixels together. This helped reduce the dimensionality of the problem without losing much information. This is only one possible way out of many to preprocess the data.
After this, we shuffled our datasets and fitted several models with our train dataset. We then measured the accuracy of the prediction over the train and the test dataset. The results achieved are reported in table 2. We can clearly see that in our case, the classifier that worked best was the decision tree with an accuracy over the test set of 0.49. The difference with the accuracy over the train set is quite important and shows that the classifier did some overfitting. As a point of comparison, a random prediction over 7 balanced classes would

give only a success rate of 0.14. A result of 0.49 is already a good improvement and it is illustrated by the ROC curves that we plotted. Figure 4 shows good decisions but not great ones.
To improve the results, apart from changing the preprocessing and tweaking the hyper-parameters, we can think of a few potential solutions like using a forest of decision trees or using convolutional neural networks. In fact, the dataset was originally used to demonstrate the effectiveness of an enhanced convolutional neural network in [CCW+19] which achieved an accuracy of 94.52%. This result as well as the disparity in performance between the best and worst classes in table 2 shows that there is a lot of room for improvement compared to our simple decision tree.

Algorithms	Accuracy			
	Train	Test	Best class (test)	Worst class (test)
Gaussian Naive Bayes	0.46	0.41	0.86 (Limacina)	0.21 (Fish larvae)
Decision Tree	0.63	0.49	0.88 (Limacina)	0.13 (Fish larvae)
Multilayer Perceptron	0.28	0.26	1.00 (Chaetognatha)	0.00 (Medusae)

Table 2: Baselines results with different algorithms

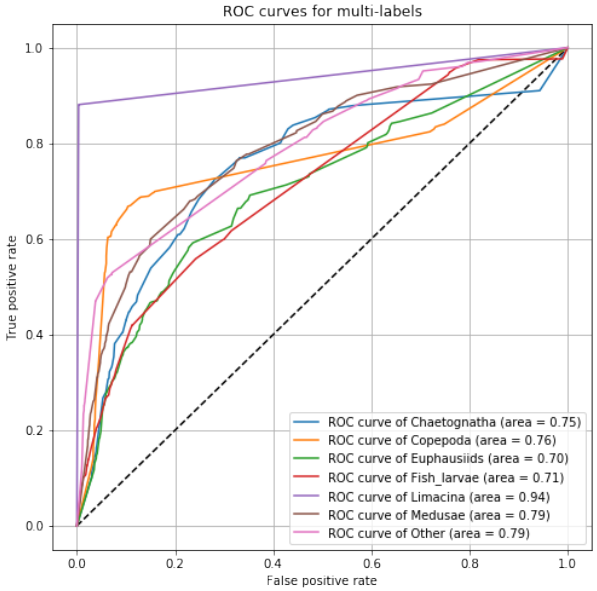

Figure 5: ROC curves

References

[CCW+19] Kaichang Cheng, Xuemin Cheng, Yuqi Wang, Hongsheng Bi, and Mark C. Benfield. Enhanced convolutional neural network for plankton identification and enumeration. *PLOS ONE*, 14:1–17, 07 2019.

[Che19] K. Cheng. Bering sea dataset. https://doi.org/10.6084/m9.figshare.8146283.v3, 2019.

[NRMA15] A. Nair, J.K. Reshma, A. Mathew, and A. Ashok. Effect of water quality on phytoplankton abundance in selected ponds of nedumangad block panchayat, kerala. *Emer Life Sci Res*, 2015.