Project Milestone

Priyank Shelat & Peng Cheng

Abstract

My team plans on using the Amphibians Dataset for the UCI ML Repository (https://archive.ics.uci.edu/ml/datasets/Amphibians) to test multiple unsupervised methods for classification. The problem this dataset poses is to predict the presence of amphibian species near water reservoirs based on features obtained from GIS systems and satellite images. The unique aspect, and challenge, of this dataset is that not only does it have multiple labels to predict but that each reservoir can have multiple labels. We plan on applying Bayesian classifiers, clustering methods, and other possible methods we learn further on in the class. We will test these against neural networks, KNNs, and/or other supervised classification methods. Outside of this goal, we may try to build visualization tools and/or apply some sort of association rule mining or predicter system if we see fit. From our preliminary findings we see that human impact and location tend to play a fairly significant role

Introduction

Our team is working on predicting which species of frogs and toads will appear in a water reservoir based on the geological and ecological features of the area. The data we will be using is from UCI ML Repository (https://archive.ics.uci.edu/ml/datasets/Amphibians). It was collected using GIS systems and satellite images as part of an environmental impact assessment report for two planned road projects – the planned A1 motorway section in Pyrzowice and the planned Beskidy Integration Way on the Bielsko BiaÅ‚a-Wadowice-GÅ‚ogoczÃ³w section of the S52 motorway – in Poland.

Our motivation for working on this problem is two-fold. It gives us a great chance to test the prediction accuracy of unsupervised methods such as Bayesian classification, unsupervised clustering, and association rule mining against more robust supervised methods such as KNNs, logistic classification, ensemble methods, and neural networks. This dataset will be very good for this comparison since it has multiple class labels that overlap – meaning each reservoir may have more than one type of amphibian resident. This will not only allow for us to test binary and multi-class classification but also multi-class overlapping classification (probabilistic), if this exists otherwise, we will just test multiple multi-class classifications. The dataset also has 16 features with both continuous and categorical feature types. It will truly test the all-around classification capabilities of the unsupervised methods. Aside from this comparison, we are also hoping to learn about what amphibians are more adaptive to human environmental impact and what human impacts affect these little critters the most.

Background

Preliminary Results

We have begun our exploration of the data. We started out by ingesting the data and splitting it into a usable array. It was immediately clear that we will have to be careful with our data types as well as what we use as inputs for our different models. The class labels are 7 one-hot vectors merged into a column-wise array. This gave us several challenges with visualization as well as figuring out what the best way to interpret and model the data is.

Aside from the actual usability of the dataset, which we had known might be an issue, we continued forward to analyze it. We saw right away that it was split into 2 by location. For any environmental and ecological data set, location can play a huge role as it will change the landscape, human impact, weather, and seasonality that many organisms depend on. We thought

this would be an important factor and investigated further. The locations were in the same country and were described by road-ways. This is because the data was originally part of environmental assessment studies for road projects in Poland. Since both locations were in the same country, we assume that weather and seasonality do not vary too much but landscape and human impact will. We have other features that deal with these particular issues. We made the decision to ignore seasonality because of the small size of the country. As for ignoring possible confounding from weather, it is still possible however when looking at how some of the species of frogs were split between the two locations, we were able to find that the area of the reservoirs in the two locations might play a bigger role and that it is likely that one of the locations, the S52 motorway, may be more rural than the other, the A1 motorway.

We have begun exploring and visualizing the interaction between the environmental and human impact features as well as their relationship to the species of amphibians. All our progress can be found here: https://github.ccs.neu.edu/prs14004/DS5320-Project. This includes our visuals, as well as our current working jupyter notebook and any other working documents.

Discussion

We will continue our work analyzing the data and begin working towards model selection, fit, optimization early next week once we have a better understanding of our predictor variables and their biases. We will begin by implementing k-nearest neighbor algorithm to get a baseline prediction accuracy since it comes with a minimum error guarantee. We will play around with a few other supervised methods such as logistic classification, ensemble methods, and neural networks to quickly get a high baseline target for our unsupervised learners. After that, our time will be focused on clustering methods, Bayesian classification, and any other methods we may be able to implement and

test. Hopefully, we will be able identify which factors affect amphibian habitat selection and compare unsupervised learners to their supervise counterparts.