Ensemble Learning Applied to Classify GPS Trajectories of Birds into Male or Female

Dewan Fayzur
fayzur20@gmail.com

DevScope S.A.
Porto, Portugal

Introduction

The goal of this work is to predict gender of shearwater based on how they navigate themselves across a big ocean from its GPS trajectory. The trajectories are collected from GPS loggers attached on shearwaters' body, and represented as a variable-length sequence of GPS points (latitude and longitude), and associated meta-information, such as the sun azimuth, the sun elevation, the daytime, the elapsed time on each GPS location after starting the trip, the local time (date is trimmed), and the indicator of the day starting the from the trip. This work achieved the first-place to the Animal Behavior Challenge (ABC 2018). The Animal Behavior Challenge was organized by the 2018 Symposium on Systems Science of Bio-Navigation, sponsored by Technosmart and proposed as a CodaLab competition. The source code of our first-place solution can be found online `https://github.com/dfayzur/Animal-Behavior-Challenge-ABC2018`.

Motivation

The idea is to have a prediction model, that could help to understand shearwater more efficiently and how they navigate themselves, like male and female shearwater could use different trajectories along the way of trip. While predicting gender of birds can be straightforward from fixed length of sequence of trajectories (latitude and longitude), predicting from variable-length of sequence of GPS points present a challenge to the feature representation to the modeling.

Data and competition setup

The training dataset is composed of all the GPS trajectories of 631 streaked shearwaters (326 male and 305 female) breeding on Awashima Island, Japan. Each datapoints in the training dataset representing a complete bird trip and being composed of the following attributes:

- longitude
- latitude
- sun azimuth: clockwise from the North
- sun elevation: upward from the horizon
- daytime: 1 being day, or 0 being night
- elapsed time: after starting the trip
- local time: only time with a format (hh:mm:ss)
- days: days after the trip starts

In the competition setup, the testing dataset is composed of all the GPS trajectories of 275 streaked shearwaters. In the Development Phase of the competition 10% of the submission labels are randomly modified to report score.

Feature Engineering

- We first created velocity, acceleration, distance features for each of the GPS points from the gives dataset. At this time we have 7 key features for each GPS points to work with, such features are: velocity, acceleration, distance, longitude, latitude, azimuth, and elevation.
- We also created the differences of above features at time t to the next point at time $t+1$. We call these features as *delta* of velocity, longitude, latitude, azimuth, and elevation.
- From these 12 features, we took quintiles at 0%, 5%, 10%, 20%, 25%, 30%, 40%, 50%, 60%, 70%, 75%, 80%, 90%, 95%, and 100% of them.
- In addition, we also calculated average, minimum, and maximum of these 12 features.
- We also calculated the number of times a velocity exceeds to values (average and quintiles at 5%, 10%, 15%, 25%, 50%, 75%, 80%, 85%, 90%, 95%, and 99%) calculated over all GPS trajectories combined.
- Finally we took first 5 longitude and latitude values of each birds' trajectories.
- We also included Principal component analysis (PCA) on longitude, latitude, azimus, elevation, and velocity of individual birds' and added them to the final features list.

Preparing Dataset for modeling

We have created two training dataset based on the generated features.

- In the first dataset, we splitted the original dataset to *day* and *night* trajectories and apply features generation. So, this operation doubles the number of features, and we call this dataset as *split*.
- In the second dataset, we consider all trajectories together and apply features generation process. We call this dataset as *together*.
- The test dataset were created on similar ways.

Approach

- The winning model was ensemble of several variants of Gradient Boosting Classifier along with Gaussian Process Classifier and Support Vector Classifier after extensive feature engineering.
- All together we trained 20 models on newly created *split* and *together* dataset with 5 fold cross validations, and predicted on the test dataset.
- The trained models are:
  - Gradient boosted decision trees: We modeled variants of Gradient boosted decision trees:
    - XGBoost with binary logistic objective.
    - XGBoost with pairwise rank objective.
    - LightGBM as traditional Gradient Boosting Decision Tree.
    - LightGBM as Random Forest.
    - CatBoost from Yandex
    - GradientBoostingClassifier from scikit-learn.
    - RandomForestClassifier from scikit-learn.
    - ExtraTreesClassifier from scikit-learn.
  - SVC, a libsvm based Support Vector Machines estimator from scikit-learn.
  - GPC, a probabilistic predictions with Gaussian process classification estimator from scikit-learn.

Results

- A 5 fold cross validation strategies was used to find hyperparameters of each models.
- The F1 score was used as evaluation criteria at cross validation.
- The classification decision boundary is 0/1, thus every model predicts 0/1 decision on the test dataset.
- Finally, we applied simple majority vote ensemble to the output of every models' collected from previous step.

Table 1: Models accuracy on test dataset (Final standing)

Model	Accuracy
Ours (First-place team)	0.7200
Second-place team	0.7018
Third-place team	0.6909
Median competition scores	0.6436
Average competition scores	0.6181

Conclusion

We introduced an ensemble learning approach to predict gender of shearwater based on trajectory and associated metadata. The ensemble learning approach was able to model trajectories to output the predictions. As a future direction, we want to make 3D (latitude, longitude, and elevation) clustering of grids from the trajectories and analyze the problem.