Yuqing Xia
Archipelagos, Institution of Marine Conservation

Delphinus Delphis Feeding Behavior Prediction

Introduction

This project goal is to discuss how boat distance, dolphins' reactions and juvenile dolphins' appearance will affect researchers' observations about common dolphins' feeding behaviors.

It is challenging to observe dolphins' feeding behaviors in great Aegean Sea area since their feeding behaviors are mostly under the sea surface. In order to better observe a long time dolphins' feeding behavior without interfere their normal feeding process, we want to build a model to predict possible dolphins' feeding behaviors during boat survey. Due to the randomness of boat survey(like the weather condition and researcher's health condition on the sea) , we would like to increase the efficiency to observe dolphins' feeding behavior by building machine learning models, in a handful and practical way.

Based on our current facilities, we did not include the environmental factors into models, rather I used our boat distance, dolphin's reactions and juvenile dolphin's appearance factors, which are more practical data to collect in boat survey. However, for future study, I suggest we include the environmental factors data that we collected during boat survey into the model(eg,.salinity, PH value), which can be more accurate than data on Copernicus website and also can increase the model accuracy dramatically.

Methodology

I. Dataset Introduction

All data is from Archipelagos, Institute of Marine Conservation, collected from 2016 March to 2017 May, and the sample size is 230.

II. Possible Variables

1. J_A: Juvenile appearance. Factor Data. 0: dolphins accompany with juvenile; 1: adult dolphins only
2. Our.boat.dis : Our Boat Distance to the dolphins. Numerical Data
3. Boat: Numbers of Boat appears around the dolphins. Factor Data(1—5)
4. Reaction: Dolphins reactions to the boats. Factor Data. NEG: negative, NEU: neutral, POS: positive reactions
5. Movement.formation: Factor Data (AL, EC, FR, LI, SP)

III. Data Set Validation

The response in our model is the probability of feeding behaviors. We categorized all observed dolphin behaviors into feeding and non-feeding behaviors. We have 86 feeding behavior and 144

non-feeding behaviors(traveling, socializing, swimming and etc).Their ratio is within 1:2,which is an acceptable proportion even though the positive and negative data is not perfectly 1:1.

In the reaction dataset, we can see NEU has 55% amount, POS has 37.4% amount, while NEG only has 7%. This data reveals the truth that common dolphin's are friendly and hardly show negative reaction to researchers, and thus we don't plan to adjust the proportion of this.

IV. Data pre-processing

- Test & Train data

I used 85% of raw data as train data and 15% as test data, because as a quiet unbalanced data, the major parameter reaction is skewed, thus in order to let the models learn more about this pattern, I increased train data up to 85%. Then I used prop.table to test the response and key parameters' proportion in each train and test data set to ensure they have acceptable distance from each other.

	Feeding	Not Feeding	NEG	NEU	POS	J_A(0)	J_A(1)
Train	0.35	0.645	0.08	0.54	0.38	0.47	0.53
Test	0.5	0.5	0.03	0.62	0.35	0.5	0.5

Table 1. Test and Train data validation

- Cross-Validation

We used 10 fold cross-validation to evaluate and validate our train and test dataset.

IV. Model Selection

Because we have a binary classifier which is feeding or not feeding, I then selected GLM, GLMNET,Naive Byes, SVM and random forest as candidate models.

- Logistic Model

Logistic model is an elementary model for binary classifiers. Similar to linear regression model, logistic model trying to conclude relationship between the mean of dependent variables with independent variables. The difference if that the generalized linear model requires a link function which enables researchers to convert categorical (2-dimensional here) variables into a continuous/ numerical data and then use the same principles of linear regression to fit the model, as indicated below.

$$g(u) = \beta 0 + \beta 1 x1 + \ldots + \beta p x p (x T \beta)$$

The link function, g, describes how the mean response, EY =μ, is linked to the covariates through the linear predictor.

So here, our response is the probability of feeding behavior appearance. And the model is listed below. We selected J_A, Our.boat.dist and Reaction as independent variables and drop the boat number because is not significant. I used quasi binomial function because it deals more properly for data with high variance, compared to binomial method.

model_glm<- glm(Behavior~ J_A+Our.boat.dist+Reaction,train_data,family = quasibinomial)

- GLMNET model

Compare to GLM model, the GLMNet model works very well on small sample size and it's unique lasso and ridge functions helps to constrain on your coefficient and prevent overfitting scenario.

Because the logistic model performs pretty bad, thus I used the glmnet model, which is a bagging version of glm model. Bagging is one of the ensemble method that can use bootstrap sampling on our current training dataset. We realized that when we changed the train data ratio with test data, the proportion of each levels in Reaction predictor is not stable and the performance of very model will change according. For a small sample and as a unstable learning model, we think GLMNET can be a better candidate for classification prediction.

- Naive Byes

Naive Byes is a common machine learning model to calculate the posterior probability based on the class densities Pk(x) and the prior probabilities. In our case, we wanted to discover how other predictors contributed to the the probability of dolphins feeding behaviors. I included boat, J_a, reaction and movement.formation into the model to explore how all those features can lead to best prediction of feeding behavior probability.

P(Feeding| boat ∩ J_A ∩ Reaction ∩ Movement.formation) =P(boat|Feeding) P(J_A | Feeding)P(Reaction|Feeding) P(Movement.formation|Feeding) P(Feeding)/P(boat) P(J_A)P(Reaction) P(Movement.formation)*

- SVM Model

SVM model is popular in machine learning because it's very accurate and highly applicable. One advantage of SVM model is that, in classification process, it creates a hyperplane that can separate different classes with maximum margin with each class for separable variables. In

linear regression models, where we separate classes using one line to separate them, we can use a hyperplane in SVM to do that.

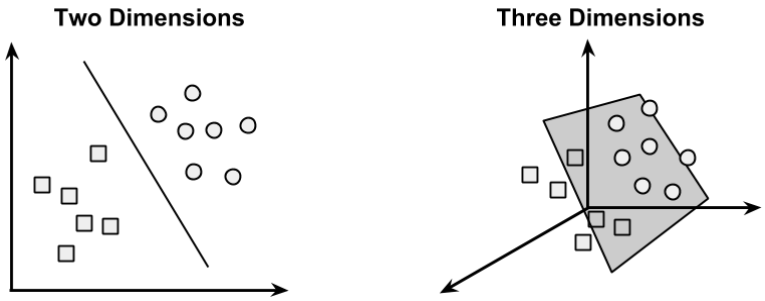

Figure 1: SVM Hyperplane Explanation

However the most powerful skill set of SVM is that it use a slack variables to tolerate for points that are are not exactly corrected classified. This slack variable is called kernel and it act as a transform function that can map problem into higher dimensions and then separate those points properly. In other words, if all data we all have is locked in a black box, and all of the types are mixed together. In this case, we mix the feeding and non-feeding behaviors. Then if we shake the box and then a virtual hyperplane can can be created to separate those points. The power to uplift those points in order to separate them is kernel.

In order to let SVM model performs best, we need to choose propitiate kernel type and the optimal cost that minimizes the mean of cross-validation errors.[1]

$$\min \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

- Random Forest

Compare to decision trees, random forest can improve models' accuracy by fitting many trees and each split with fit each one to a bootstrap sample of our data. In addition, with advantage of using cross validation, random forest itself can be resistant to overfitting scenario.

Instead of using common random forest model, I use caret package because it is faster and provide opportunity for us to tune the model by adding tune length, tune grid and separated cross validation method. We used tune length equal to 10 because is more accurate than smaller tune length, and I used 10 fold cross validation method. After defining the tune grid and customizing

[1]A optimal cost value is applied to all points that violate the constraints, and rather than finding the maximum margin, the algorithm attempts to minimize the total cost.

tune grid, I found when mtry equal to 2, the model reaches best performance. Mtry is important in Random forest because it is the number of randomly select parameters at each split.

Conclusions

I. Key factors analysis

Among all 5 variables we have, we found Reaction and J_A are most significant factors influencing dolphin's feeding behavior. Thus we dropped other variables to achieve best model performance.

II. Model Evaluation

To evaluate the models' performance, we used Kappa, sensitivity and specificity as evaluation standards.

As a statistic outcome in confusion matrix, Kappa give us a way to think beyond only accuracy. The rate of accuracy is equal to sum of true positive and false positive divided by all the test data. While Kappa trying to consider how accuracy different from probability that is created by chance alone would lead the predicted and actual values to match.

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

The ROC is a visualization method that show how well the model predicts the positive data. The diagonal line in ROC means no predicted value, while the left corner(the perfect classifier) means there is 100% true positive rate and 0% of negative false rate. Since our research purpose is to predict the dolphin appearance data, we would like to use ROC curve as an implication. A quantitive version of ROC is AUC, which is the area under the ROC. As we mentioned the principles of ROC, if the AUC is 50%, then means ROC is basically the diagonal line, which means the classier cannot differentiate negative and positive data. If AUC is 100%, which means the model can predict true positive rate 100% accurate. In this research, we combine both ROC and AUC to evaluation model.

However, ROC and AUC here is not very suggestive. For example, in GLM model, where we have 0 kappa value, and we get 76% in ROC and 85% of AUC.

Model / Evaluation	Kappa	Sensitivity	Specificity
GLM	0	0.3529	0.1176
GLMNET	0.3529	0.8750	0.6154
Naive Byes	0.5294	0.8462	0.7143
SVM	0.6471	0.8462	0.7143
Random Forest	0.5693	0.6176	0.8621

Table 2. Model Selection Criterions Comparison

From the table listed above, we chose SVM model with dependent variables J_A, Reaction and Boat as parameters. We then tested optimal cost from 1e-1 to 1e+3, and gamma value from 0.5-100. We finally choose cost=1, gamma=100 on a 10 fold cross validation basis based on it's final classification error. Then from all the kernels, ranging from rbfdot, polydot, tanhdot, vanilladot, laplacedot, besseldot, anovadot, we chose basseldot kernel because we can get highest kappa value from it.

```
svm_test <-ksvm(Behavior ~ J_A+Boat+Reaction,data = train_data,kernel =
"besseldot",cost=1,gamma=0.5)
```

Discussion

In this chapter, I will discuss why I chose certain variables into our models and discuss their drawbacks. Further more, I will listed other variables to be collected in the future in order to improve model accuracy.

J_A and reactions are two significant factors in our models based on all 13 variables Archipelagos have been collected so far. Researchers supposed that dolphins will be more readily seeking food if they have children accompany and if they were in neutral mood status. More specifically, dolphins may seek social activities if they shown positive attitudes.

Our boat distance and boat number help some models to achieve minor performance but not as significant as former two did. Those two factors are substitute of "noise" measurement to dolphins. Since dolphins will be less likely to have feeding behaviors if there are noises around, so this is a very valuable factor to consider. However those two factors doesn't work on well together. The boat distance only measures our boat distance to the dolphins without considering other boats distance. And the boat number is a categorical variable, which obscures the significant differences of real "noises". An alternative to measure "noise" could be counting each boat's average decibels and add all of them together. Other than just boat numbers, adding

industry standard decibels for each boats may help models predict feeding behaviors more accurately.

Other variables to consider will be environmental data(like salinity and PH value) and sea surface biomass values. It reasonable to assume that those factors will influence prays of dolphins' distribution and thus are reasonable to be included. However, it will be unreasonable if we have a very small sample size(230), and the validated GPSs of dolphins' appearance are only 23 points. In other words, we observed multiple behaviors in one spot. Thus, it will be statistically insignificant if we would extract those environmental variables by their latitude and longitude. Hopefully in future research, we can include salinity,PH value and see surface biomass values into the models until we have a large dataset.