BCUR Statistic Poster Handover Document

1. Introduction

Comment: Should be pretty self-explanatory from the bullet points on the poster

- Our goal is to analyze the types of whistles that dolphins produce which will help with our understanding of Dolphins.
- We aim to research dolphin communication, specifically similarities and differences in whistles across different populations of the rough-toothed dolphin (Steno bredanensis).
- We aim to describe the whistle repertoire of this species by using statistical models to identify key features of the whistles that make them different from each other.

2. Data Gathering

Comment: This is the detailed version of the explanation of data gathering on our poster

The results of the data to be analysed is received from other teams tracing the whistles in a spectrogram and providing the traced data to tools such as ROCCA and ARTwarp. ROCCA is a MATLAB-based tool designed for real time species identification of delphinid whistles during shipboard surveys, and ARTwarp is an automated method for categorizing bioacoustic signals, particularly focusing on dolphin whistles and killer whale calls. Both these tools are used to categorise the whistles into different groups in a variable called Category with an output of NUMBER variables for each whistle analysed.

The issue with all the variables output by ROCCA and ARTwarp from each whistle, is that many of these are collinear, meaning that they provide no tangible benefit to be included in the model, instead complicating the process. Therefore, all the collinear variables were removed reducing the number of variables to 6, being ‘DURATION’, ‘FREQABSSLOPEMEAN’, ‘FREQBEG’, ‘FREQMAX’, ‘FREQQUARTER1’ and ‘FREQPOSSLOPEMEAN’.

To determine the number of data points (whistles) to use for the modelling, a completeness graph was used. From this graph, we can see the discovery curve begins flattening around the mark of 1000 whistles. This means that using more whistles will have negligible additions to the number of categories, therefore the number of 1000 whistles was categories.

- Data is received from teams tracing whistles in a spectrogram and analysed using ROCCA and ARTwarp.
- ROCCA identifies delphinid whistles in real-time, while ARTwarp categorizes dolphin and killer whale calls.
- These tools classify whistles into categories, outputting multiple variables per whistle.
- Many output variables are collinear, so only six key variables were retained: DURATION, FREQABSSLOPEMEAN, FREQBEG, FREQMAX, FREQQUARTER1, and FREQPOSSLOPEMEAN.
- A completeness graph was used to determine the ideal number of whistles for modelling.
- The discovery curve flattens at 1000 whistles, meaning additional data offers minimal benefit.

3. Models and Their Uses

Comment: This is the detailed version of the explanation of the models on our poster

Multinomial GLM

A Multinomial Generalized Linear Model (GLM) is a type of statistical model used when the response variable is categorical with more than two possible outcomes (i.e., a multinomial outcome). It is an extension of logistic regression that can handle multiple categories rather than just binary classification. In this instance, the categories would be considered as multiple possible outcomes, derived from the machine learning grouping techniques used to provide the data to analyse.

- A Multinomial Generalized Linear Model (GLM) is used for categorical response variables with more than two outcomes.
- It extends logistic regression to handle multiple categories instead of just binary classification.
- In this case, categories represent outcomes from machine learning grouping techniques used for data analysis.

Pros:

- It can show the relative importance of each variable with a coefficient of how much each variable impacts the final categorization of the whistle.

- It can help understand the relationship between different key features of whistles (e.g. maximum frequency and duration) in the data.

Cons:

- Model fit is not possible with too many variables. We encountered the problem of having to trim down the variables for our model.
- The model might require a relatively large sample size in order to be accurate.

Random Forest

Random Forest is a machine learning method used for classification and prediction. It works by building many decision trees, where each tree makes a prediction based on different random subsets of the data. Each tree also uses a random selection of variables, making each of the trees diverse. The final prediction is based on the majority vote across all trees. Random Forest helps reduce overfitting and improve accuracy compared to just using a single decision tree.

- Works by building multiple decision trees, each using different combinations of whistle measurements (e.g. max frequency, duration).
- A decision tree is a flowchart-like model where you follow each branch until you get to the prediction in the end.
- The final whistle category prediction is the majority vote across all trees.

Pros:

- Random Forest captures complex, non-linear patterns that could be missed by Multinomial GLM.
- Random Forest ranks feature importance, helping identify which variables have the biggest impact on classification.
- Random Forest is robust to noise and outliers due to using many decision trees, making it more reliable with messy or real-world data.

Cons:

- Random Forest has a limit of 53 categories for classification, meaning data with more categories cannot be fully included in the model and would require exclusion.
- Random Forest makes interpretation harder since it doesn't provide p-values or clear coefficients like Multinomial GLM.

- The model is inconsistent when there are many variables, and it may require a relatively large sample size.

4. Results and Comparison of the Models

Comment: This is the detailed version of the explanation of the model results and comparison on our poster

Multinomial GLM results

We fit a multinomial logistic regression model to predict whistle category from six acoustic features:

We perform stepwise variable selection to reduce the number of predictors variables in the model to six. From this analysis it was found that the most informative variables to predict whistle category are:

- Duration
- FREQBEG (start frequency), FREQMAX (maximum frequency), FREQQUARTER1 (25th percentile)
- FREQABSSLOPEMEAN (absolute slope), FREQPOSSPLOPEMEAN (positive slope)

From model analysis we found that the model explained a substantial proportion of deviance with a Residual Deviance = 5304.35 and AIC = 6144.35 indicating that acoustic features are informative for classifying whistles.

The graph shows the relative contribution of each of the variables with values above 0 contributing more to the model

Random Forest Results

An R Script was made to analyse the previous data that was gathered. This R script builds and evaluates a Random Forest classification model using a dataset of contour statistics, related to audio features. It begins by reading a CSV file and selecting specific columns, including one target variable (SELECTIONNUMBER) and several numerical features such as frequency range and standard deviation. After loading the necessary randomForest library, the script checks the structure of the data and removes any rows with missing values to ensure clean input. The dataset is then split into training (80%) and testing (20%) sets using a fixed seed for reproducibility.

The model is trained using the training data, with SELECTIONNUMBER treated as a categorical variable. A Random Forest is created with 500 trees and 3 variables tried at each split, and variable importance is calculated. After training, the model is evaluated by

predicting the class labels for the test set and comparing them with the actual labels using a confusion matrix. Finally, the accuracy of the model is calculated and displayed, and the importance of each feature is visualized to understand which variables most influenced the predictions.

- Using a set seed of '123' the accuracy of the Random Forest model fitted with this data was found to be 74.32 %
- When another set seed of '456' was used, the accuracy of the Random Forest model fitted with this data was found to be 63.39 %
- After averaging the results of 100 trials with random seeds, the accuracy of this model was found to be 67.68%

Why we decided on Multinomial GLM

- Multinomial GLM allows for hypothesis testing and confidence intervals unlike Random Forest.
- Multinomial GLM is easier to interpret compared to Random Forest as it has clear coefficients and tells you how a variable affects the categorization.
- It is faster to run, making it more time efficient, especially for larger datasets.

5. Final Statements and Conclusion

Comment: Should be pretty self-explanatory from the bullet points on the poster

- From using Multinomial GLM, we found that the maximum and quartile 1 of the frequency of whistles are the most distinguishing factors.
- This helps us understand what the main acoustic features are involved in the categorization and differentiation of whistles.