Project Documentation
Anuran Calls

Kadir Hizarci

2022

Introduction

Aim

The aim of this project is to determine the species of the frog by using the frequency of their calls from the dataset. There are various models implemented (gain ratio, gini index, naïve bayes, neural network, support vector machine) and their results have been measured with different methods (runtime, confusion matrix, accuracy etc.). On top of that, used models are enchanced with cross-validation and bagging ensemble methods and re-evaluated.

Technology

Libraries used are: numpy, pandas, timeit and sklearn.

Numpy: Numpy is be used to perform a wide variety of mathematical operations on arrays. It adds powerful data structures to Python that guarantee efficient calculations with arrays and matrices and it supplies an enormous library of high-level mathematical functions that operate on these arrays and matrices.

Pandas: Pandas is one of the most useful libraries for data science/machine learning. It provides a wide range of pre-defined operations for use in best practices. In this project, it's used for filtering data according to certain conditions, or segmenting and segregating the data according to preference.

Timeit: Timeit is used to measure the runtime of a code segment accurately.

Sklearn: Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

Method

ML models

Algorithms used are: Naïve Bayes, Decision Tree with Gain Ratio, Decision Tree with Gini Index, Neural Network, Support Vector Machine

Naïve Bayes yields low accuracy at the cost of very fast runtime.
Fastest Variation: Default Naïve Bayes (0.012 secs)
Highest Accuracy: Default Naïve Bayes (91.76%)

Gain Ratio gives a rather balanced result in terms of runtime and accuracy.
Fastest Variation: Default Gain Ratio (12.84 secs)
Highest Accuracy: Bagging Ensemble Applied Gain Ratio (96.85%)

Gini Index corresponds quite well with the given dataset with fast and accurate results.
Fastest Variation: Default Gini Index (5.93 secs)
Highest accuracy: Bagging Ensemble Applied Gain Ratio (96.66%)

Neural Networks (1 Hidden Layer) yields high accuracy at the cost of runtime.
Fastest Variation: Default Neural Network (20.39 secs)
Highest Accuracy: Bagging Ensemble Applied Neural Network (98.66%)

Neural Networks (2 Hidden Layers) yields slightly better results than 1 hidden layer variation with slower runtime.
Fastest Variation: Default Neural Network (26.98 secs)
Highest Accuracy: Default Neural Network (98.84%)

Support Vector Machine yields the best accuracy with moderate runtime. Arguably the best algorithm for this dataset.
Fastest Variation: Support Vector Machine (5.93 secs)
Highest Accuracy: Support Vector Machine (98.85%)

Future work for this dataset could include prediction of the poisonous species with no physical contact neededed from their calls.

Functions

Each of the used machine learning models has its advantages and disadvantages. The evaluation in this report is mainly done by going over the model runtimes and accuracies, however, results folder includes a more detailed report with confusion matrix and other criteria included.

Links

Link to the repository (eg GitHub, but if you do not want to publish your work, you can use google drive). The folder should contain the code used to train the model.

Link to GitHub
Link to Dataset