Comparison of four machine learning algorithms for audio classification

Serenay Doganca Cetin
u966915
Group 04

Abstract

The use of machine learning for audio classification could be beneficial in tracking animal diversity and population. For that purpose, four machine learning algorithms were compared in their performance in classifying audio data. The results showed very similar scores for all algorithms which were not as high as expected. However, SGDC (Stochastic Gradient Descent Classifier) performed relatively better than the other three. The results suggest considering other algorithms for a similar comparison.

Keywords: machine learning; animal diversity; audio classification; decision trees; support vector machines; stochastic gradient descent classifier; logistic regression; audio features

Introduction

Animals are suffering from the increase of urbanization, the detriment of forests and rural areas for farming and building residential areas. The areas once used to belong to only them, are now shared with humans. According to Buerkert & Schlecht (2019), the worldwide urban population could reach 68 percent by 2050, up from 55 percent presently. This situation leads to high levels of heavy metals and plastics in the soil and, plastics consumed by ruminants in grazing fields or on municipal dumpsites build in their rumen and can cause animal mortality. Other than mortality, they might have trouble finding food, and a safe place for their off-springs and themselves. Some of the animals may adjust to the change in their habitat and some may not. For animal preservation and conservation, keeping track of the numbers of animals to assess population sizes is very important. (Keeping & Pelletier, 2014) One way to do that is by using machine learning algorithms to classify animal sounds, which can lead to tracking the number of animals in various habitats.

Currently, predictive analytics and structured data are at the center of the general usage of big data, which is dominated and affected by the marketing activities of large software and hardware developers where the majority of big data, which is unstructured and can be found in the form of audio, photos, video, and unstructured text is ignored. (Gandomi & Haider, 2015) The development of algorithms that lead to the correct segmentation and classification of audio data would aid intelligent agents in gaining a better understanding of their surroundings. The purpose of this paper is to answer the question “Which algorithms given the most highly correlated audio features, produce the most accurate classification model of audio tags on the freefield1010 data set.”

Methods

Dataset

The data to be used for this research consists of 7690 standardized excerpted audio clips from the Freesound field recordings which were recorded in different locations and regions. Each audio clip lasts about 10 seconds and is tagged with a variety of keywords that indicate the qualities and biodiversity of these locations. The data set cannot be guaranteed to contain solely ”field-recordings” because the original tagging (as well as the audio contributions) was crowdsourced. The data set was divided into ten equal-sized sections at random. This enables machine learning to employ fixed data subsets or do 10-fold cross-validation. There are around 128 minutes of audio in each of the ten subgroups, totaling nearly 21 hours.

Each.wav file contains a numeric ID as a filename that corresponds to the ID of the original audio file in the Freesound archive. Each .wav file is 10 seconds long and extracted from the middle of the audio recording, WAV, mono, sample rate 44.1 kHz, 16-bit PCM, amplitude normalized to -2 dB for each file. Some of the files contained complete silence and since this may cause normalization issues, they were removed from the data set. Each.wav file contains a corresponding .json file, which is merely metadata reported by the Freesound server while retrieving the original audio. User, license, tags, id, and geotag are among the important keys in the.json files. Fields like ”channels,” ”duration,” and ”samplerate” in the.json data correspond to the actual audio contribution, not the clip contained in the data set.

Preprocessing steps

At first, we have loaded the audio files using librosa (Version 0.9.1; McFee et al., 2022) package. Librosa is a Python package that analyses music and audio. It contains the components required to construct music information retrieval systems. (10.5281/zenodo.6097378) Then we needed to extract useful information from the audio files that are not visible in the raw data. According to Bhattacharjee et al. (2018), in feature extraction for audio classification, Zero-Crossing Rate, Spectral Centroid, Spectral Roll-off, and Spectral Flux are some of the most commonly used spectral features whereas the most popular features from the temporal category are energy, entropy and root mean square (RMS). Apart from these, in a few studies, spectrograms have also been used as features and processed as images. For each audio file, we have calculated amplitude envelope, root-mean-square energy, zero crossing rate, band energy ratio, spectral centroid, bandwidth, spectrogram, mel-spectogram, mel-frequency cepstral coefficients as characteristics that can help in categorizing. All of these features were calculated by averaging the values in each frame for each file. **”water”, ”train”, ”voice”, ”people”, ”nature”, ”city” and ”birdsong”** tags were selected as the target variables since their appearance on the data set were slightly greater than the others. (Stowell & Plumbley, 2014)

With a total of 9 features, Spec_mean, Spec_std, Mel_Spec_mean, Mel_Spec_std, Mfcc_mean, Mfcc_std, Amp_Env, RMSE, Zero_Cross, BER, Band_Width, a correlation matrix was created and according to the results, **Spec_mean, Spec_std, Mel_Spec_mean, Mel_Spec_std, Mfcc_mean, Amp_Env, RMSE** are chosen to be used as predictor variables.

Implementation

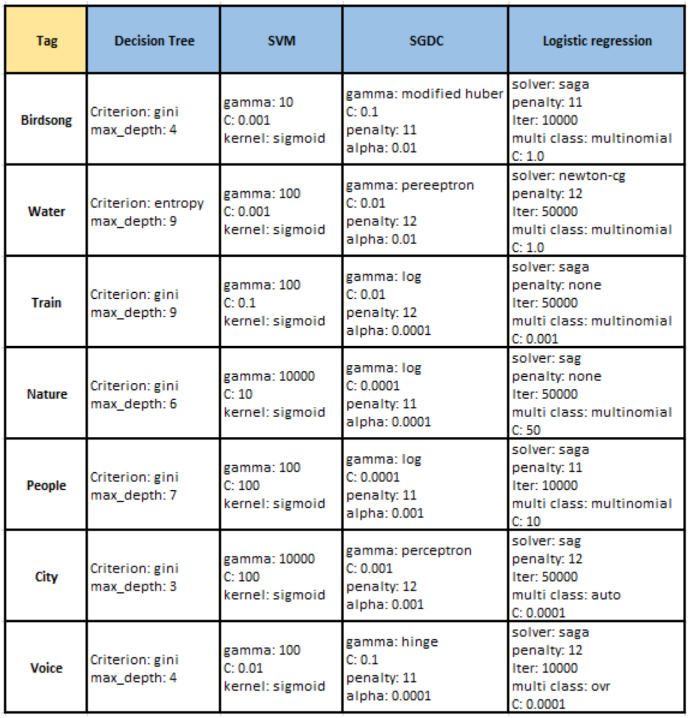
For each tag among water, train, voice, people, nature, city and birdsong, which are the target variables, the new data set containing the features indicated in the above section, which are the predictor variables, is split into training, validation and test sets with a %20 test set, %20 validation set, and %60 training set ratio. Decision trees, SVM (Support Vector Machines), SGDC (Stochastic Gradient Descent Classifier) and Logistic Regression models are compared in accuracy of predicting the target variable, for each variable. A pipeline containing Standard-Scaler, PCA (Principal Component Analysis) and the estimator to be used, is fit to the training data. The reason to use a scaler is to standardize the range of features, and PCA is used to transform correlated variables into a smaller number of uncorrelated variables, reducing dimensionality and, as a result, reducing the quantity of the data and improving the performance of the machine learning algorithms. GridSearchCV with 5 fold stratified cross-validation and roc_auc as scoring is used to find the optimum hyperparameters for the algorithms. The ranges of tried values for hyperparameters for each algorithm is shown in Table 1.

Table 1: Grid of hyperparameter values.

Algorithm	Hyperparameter	Range
PCA	n_components	range(1,X.shape[1]+1,1)
DecisionTree	criterion	gini, entropy
DecisionTree	max_depth	2,3,4,5,6,7,8,9,20,50
SVC	gamma	10,100,10000,20000
SVC	kernel	sigmoid
SVC	C	0.001, 0.01, 0.1,10,100
SGDC	loss	hinge,log,modified_huber,perceptron
SGDC	epsilon	.01, .001, .1, .0001, .05
SGDC	penalty	l1, l2, elasticnet
SGDC	alpha	.00001, .0001, .001, .01, .1
LogisticR	solver	newton-cg, sag, saga
LogisticR	penalty	l1, l2, elasticnet, none
LogisticR	C	100, 10, 1.0, 50, 20, .001, .0001
LogisticR	max_iter	10000, 30000, 50000
LogisticR	multi_class	auto, ovr, multinomial

Performace Evaluation

The best hyperparameters for each algorithm and each tag are shown in Figure 1.



Tag	Decision Tree	SVM	SGDC	Logistic regression
Birdsong	Criterion: gini max_depth: 4	gamma: 10 C: 0.001 kernel: sigmoid	gamma: modified huber C: 0.1 penalty: 11 alpha: 0.01	solver: saga penalty: 11 Iter: 10000 multi class: multinomial C: 1.0
Water	Criterion: entropy max_depth: 9	gamma: 100 C: 0.001 kernel: sigmoid	gamma: pereeptron C: 0.01 penalty: 12 alpha: 0.01	solver: newton-cg penalty: 12 Iter: 50000 multi class: multinomial C: 1.0
Train	Criterion: gini max_depth: 9	gamma: 100 C: 0.1 kernel: sigmoid	gamma: log C: 0.01 penalty: 12 alpha: 0.0001	solver: saga penalty: none Iter: 50000 multi class: multinomial C: 0.001
Nature	Criterion: gini max_depth: 6	gamma: 10000 C: 10 kernel: sigmoid	gamma: log C: 0.0001 penalty: 11 alpha: 0.0001	solver: sag penalty: none Iter: 50000 multi class: multinomial C: 50
People	Criterion: gini max_depth: 7	gamma: 100 C: 100 kernel: sigmoid	gamma: log C: 0.0001 penalty: 11 alpha: 0.001	solver: saga penalty: 11 Iter: 10000 multi class: multinomial C: 10
City	Criterion: gini max_depth: 3	gamma: 10000 C: 100 kernel: sigmoid	gamma: perceptron C: 0.001 penalty: 12 alpha: 0.001	solver: sag penalty: 12 Iter: 50000 multi class: auto C: 0.0001
Voice	Criterion: gini max_depth: 4	gamma: 100 C: 0.01 kernel: sigmoid	gamma: hinge C: 0.1 penalty: 11 alpha: 0.0001	solver: saga penalty: 12 Iter: 10000 multi class: ovr C: 0.0001

Figure 1: Best performing hyperparameters.

Results

Decision Trees, SVM, SGDC and Logistic regression models were performed on the data set which consists of seven audio features as predictor variables, and their performance on predicting seven tags were measured. The training and prediction results were obtained for different machine learning algorithms and for each tag separately. The best performing hyperparameters found at preprocessing stage were used in the algorithms. On table 2, the "Training" column represents the accuracy scores of different algorithms on the training set and the "Prediction" column represents the accuracy scores on the test data set for each tag. As can be seen in Table 2, the accuracy on the test data set for algorithms is usually 0.5. For the accuracy scores on the training data set, the minimum is 0.508, the maximum is 0.673, and the mean is 0.580.

Table 2: Accuracy scores on training and test data sets.

Algorithm	Training	Prediction
Bird_Tag		
Decision Tree	0,673	0,5
SVM	0,635	0,5
SGDC	0,634	0,5
Logistic Regression	0,579	0,5
Water_Tag		
Decision Tree	0,552	0,51
SVM	0,572	0,5
SGDC	0,59	0,5
Logistic Regression	0,523	0,5
Train_Tag		
Decision Tree	0,567	0,5
SVM	0,58	0,5
SGDC	0,632	0,5
Logistic Regression	0,587	0,5
People_Tag		
Decision Tree	0,588	0,51
SVM	0,637	0,5
SGDC	0,588	0,5
Logistic Regression	0,584	0,5
City_Tag		
Decision Tree	0,543	0,5
SVM	0,56	0,5
SGDC	0,578	0,5
Logistic Regression	0,508	0,5
Voice_Tag		
Decision Tree	0,532	0,5
SVM	0,56	0,5
SGDC	0,566	0,5
Logistic Regression	0,531	0,5

The mean accuracy scores on training data set for each of the algorithms are presented on Table 3.

Table 3: Mean accuracy scores on training data set.

Algorithm	Mean Training Score
Decision Tree	0.577
SVM	0.589
SGDC	0.596
Logistic Regression	0.557

Discussion & Conclusion

This study has tried to find the best performing machine learning algorithm for predicting certain tags on the freefield1010 data set. Even though the accuracy scores on the test data set are mostly 0.5, it does not mean that the models performed poorly. However, the results were not significant either. The accuracy scores on the training data set were slightly higher than on the test data set, but the increase is also not noteworthy. To answer the research question, Table 3 can be observed. Since the highest average score was obtained by SGDC algorithm, it can be said that it is the best performing machine learning algorithm in classifying sounds, out of the four algorithms that were considered in this study. However, since the accuracy was not significantly high, considering other machine learning algorithms would be a logical step for the future.

References

Bhattacharjee, M., Prasanna, S. R. M., & Guha, P. (2018). *Time-frequency audio features for speech-music classification.* arXiv. doi: 10.48550/ARXIV.1811.01222

Buerkert, A., & Schlecht, E. (2019, 09). Rural–urban transformation: a key challenge of the 21st century. *Nutrient Cycling in Agroecosystems*, *115*, 137-142. doi: 10.1007/s10705-019-10008-1

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144. doi: https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Keeping, D., & Pelletier, R. (2014, 05). Animal density and track counts: Understanding the nature of observations based on animal movements. *PLoS ONE*, *9*, e96598. doi: 10.1371/journal.pone.0096598

McFee, B., Metsai, A., McVicar, M., Balke, S., Thomé, C., Raffel, C., ... Thassilo (2022, February). *librosa/librosa: 0.9.1.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.6097378` doi: 10.5281/zenodo.6097378

Stowell, S., & Plumbley, M. (2014, January). An open dataset for research on audio field recording archives: freefield1010. *Journal of the Audio Engineering Society*.