Music Genre Classification

Abdullah Sahin

Outline

- Problem
- Dataset
- Exploratory Data Analysis
- Features
- Modeling
- Results
- Future Work

Problem

- Many facts make intelligent systems of automatic music genre classification (AMGC) vital these days. The ease of listening to music on devices, the high availability of albums on the Internet, peer-to-peer servers and the fact that artists now actively distribute their songs on their websites make music database management a must. In addition, searching genres and generation of smart playlists to select specific tunes between gigabytes of songs on personal portable audio players are essential tasks that facilitate music mining.

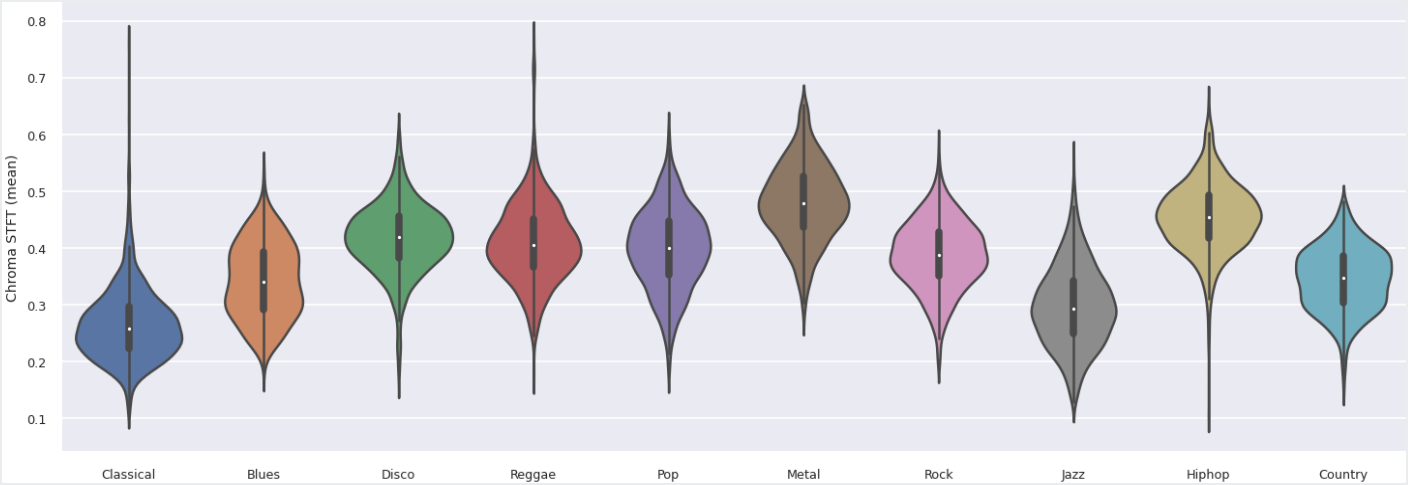
- On the other hand, the classification of music genres is as mentioned above, an ambiguous and subjective activity. It is also a field of research that is being challenged, either because of low classification accuracy or because some say that one is not capable of classifying genres that do not even have clear definitions.

Dataset

The dataset consists of 1000 audio tracks each 30 seconds long. It contains 10 genres, each represented by 100 tracks. The tracks are all 22050Hz Mono 16-bit audio files in .wav format. The files were collected in 2000-2001 from a variety of sources including personal CDs, radio, microphone recordings, in order to represent a variety of recording conditions. From each clip, we sampled a contiguous 3-second window at 10 equally distanced locations, thus augmenting our data to 10000 clips of three seconds each. Since this data was sampled at 22050HZ, this leaves us with 66150 features for the raw audio input. Thus, after pre-processing our input is of shape (10000, 66150), where each feature denotes the amplitude at a certain timestep out of the 66150. The dataset is available on http://marsyas.info/downloads/datasets.html

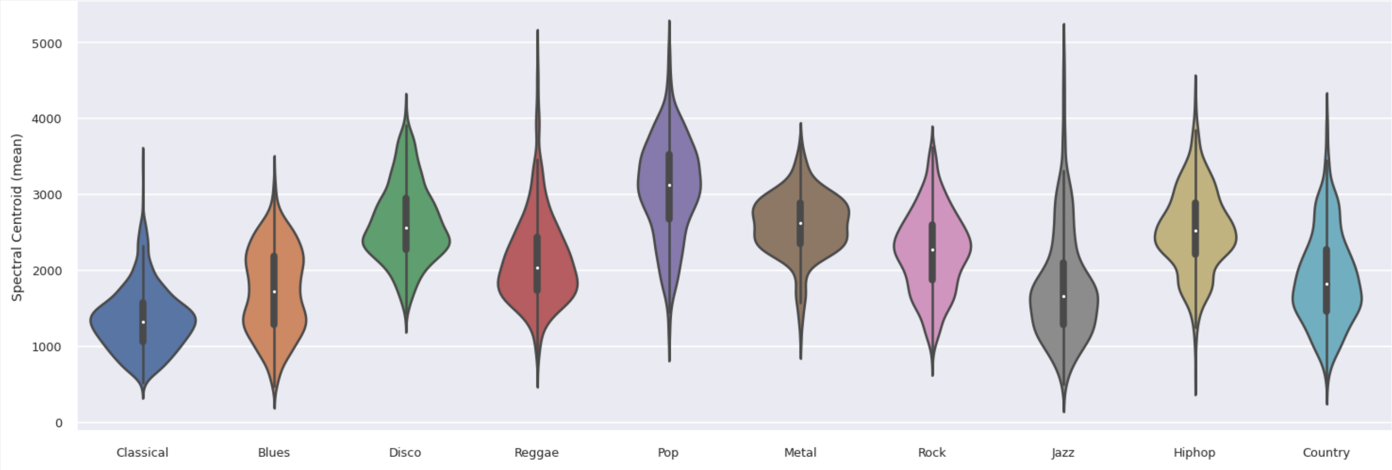
Exploratory Data Analysis -1

Chroma frequency vector discretizes the spectrum into chromatic keys and represents the presence of each key. We take the histogram of present notes on a 12-note scale as a 12-length feature vector. The chroma frequency have a music theory interpretation. The histogram over the 12-note scale actually is sufficient to describe the chord played in that window.



Exploratory Data Analysis - 2

The spectral centroid is commonly associated with the measure of the brightness of a sound. This measure is obtained by evaluating the “center of gravity” using the Fourier transform’s frequency and magnitude information. The individual centroid of a spectral frame is defined as the average frequency weighted by amplitudes, divided by the sum of the amplitudes.



Features

Time Variant

Chroma Features: STFT, CQT, CENS

Spectral: Bandwidth, Rolloff, Centroid

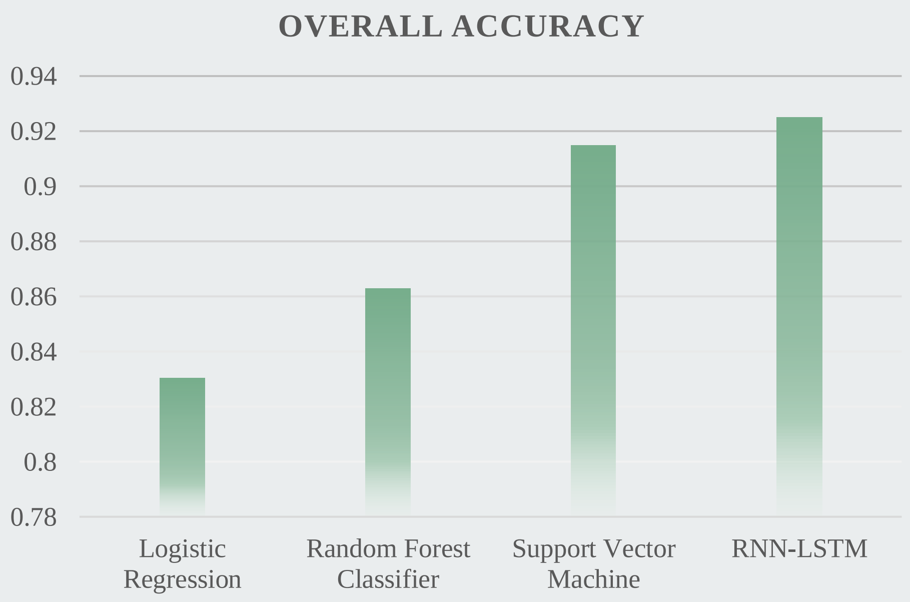
Zero Crossing Rate, Tonnetz

Time Invariant

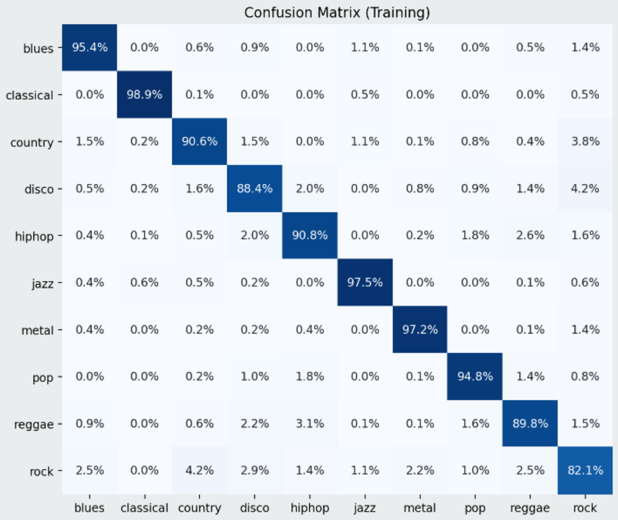
Aggregate statistics over time variant features: Min, Max, Mean, Median, Std, Var, Kurtosis, Skewness etc.

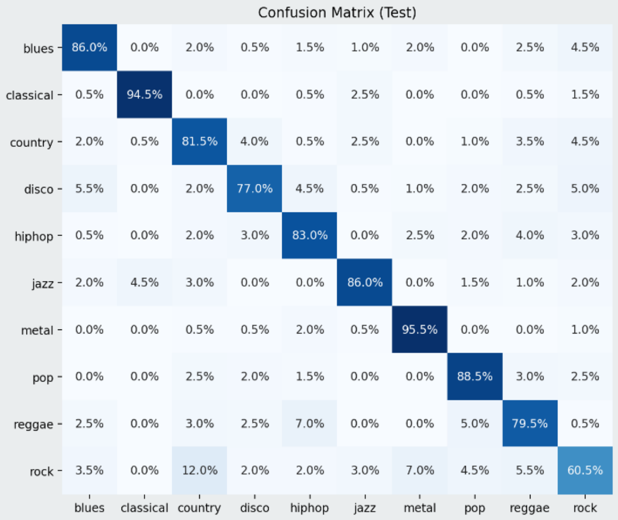
Modeling

- Logistic Regression (Baseline Model)
- Random Forest Classifier
- RBF Kernel Support Vector Machine
- Recurrent Neural Network with LSTM Layer



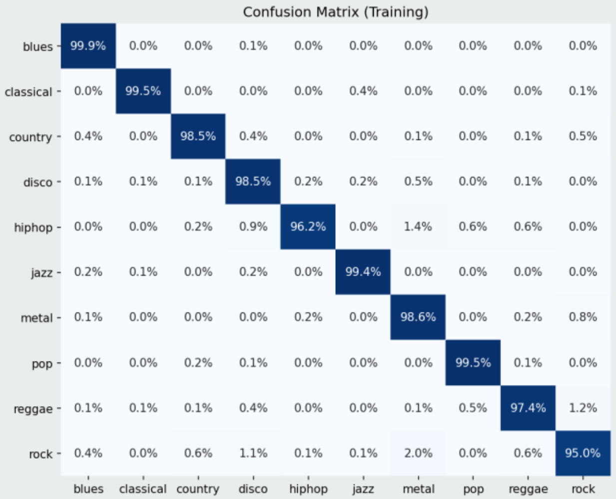
Results (Baseline Model)

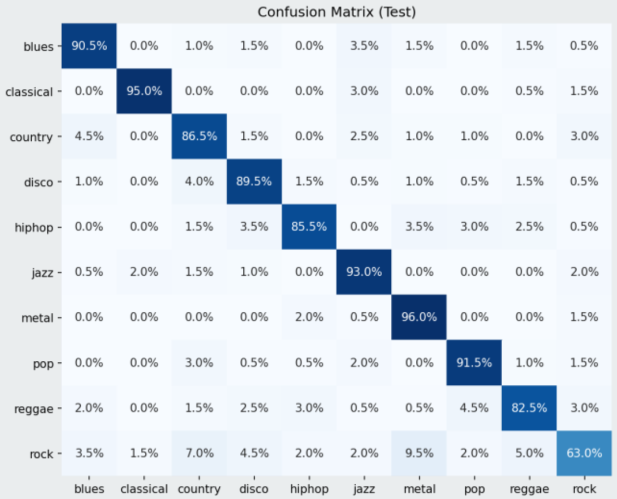




- Lowest score for Rock Genre
- Good achievement for few genres (metal, classical)

Results (Random Forest)





- Little improvement for Rock Genre
- Catching nonlinearity features improved accuracy for other genres

Results (Support Vector Machine)

Confusion Matrix (Training)

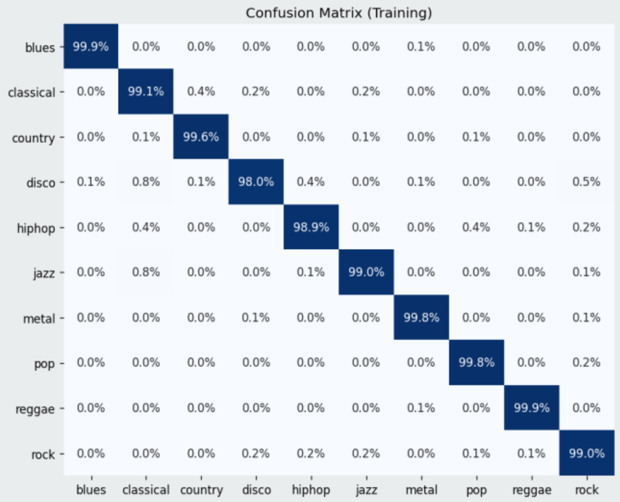
	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	98.9%	0.0%	0.0%	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.8%
classical	0.0%	99.9%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%
country	0.1%	0.1%	98.5%	0.2%	0.0%	0.0%	0.0%	0.4%	0.0%	0.6%
disco	0.1%	0.1%	0.4%	96.1%	0.6%	0.1%	0.5%	0.5%	0.4%	1.1%
hiphop	0.0%	0.0%	0.2%	0.9%	96.4%	0.1%	0.8%	0.4%	0.1%	1.1%
jazz	0.0%	0.1%	0.1%	0.0%	0.0%	99.6%	0.0%	0.0%	0.0%	0.1%
metal	0.2%	0.0%	0.0%	0.0%	0.2%	0.0%	98.2%	0.0%	0.0%	1.2%
pop	0.0%	0.0%	0.2%	0.2%	0.0%	0.1%	0.0%	99.0%	0.2%	0.1%
reggae	0.0%	0.0%	0.2%	0.4%	0.1%	0.0%	0.1%	0.2%	98.5%	0.4%
rock	0.5%	0.0%	0.1%	0.9%	0.2%	0.0%	2.1%	0.1%	0.2%	95.8%

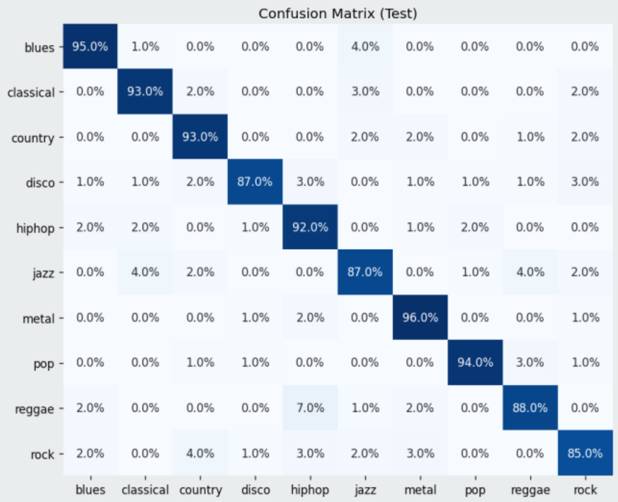
Confusion Matrix (Test)

	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	95.0%	2.5%	0.5%	0.0%	0.0%	0.5%	0.0%	0.0%	0.0%	1.5%
classical	0.0%	98.0%	0.0%	0.0%	0.0%	1.0%	0.0%	0.0%	0.0%	1.0%
country	1.0%	1.5%	93.0%	2.0%	0.0%	1.0%	0.0%	0.0%	0.5%	1.0%
disco	1.5%	0.5%	3.5%	87.5%	1.5%	0.5%	1.0%	0.5%	0.0%	3.5%
hiphop	0.0%	0.0%	2.0%	1.5%	92.5%	0.0%	0.5%	1.5%	1.0%	1.0%
jazz	0.5%	3.5%	3.0%	0.0%	0.0%	92.0%	0.0%	0.0%	0.0%	1.0%
metal	0.0%	1.0%	0.0%	0.0%	2.5%	0.5%	95.0%	0.0%	0.0%	1.0%
pop	0.0%	0.5%	2.0%	1.0%	0.5%	0.5%	0.0%	94.0%	0.5%	1.0%
reggae	1.5%	0.0%	0.0%	0.5%	1.5%	0.0%	0.0%	3.5%	92.0%	1.0%
rock	2.0%	1.5%	4.5%	4.0%	1.5%	2.0%	6.0%	1.5%	0.5%	76.5%

- Remarkable improvement for Rock Genre
- Best results with time invariant features

Results (RNN-LSTM)





- Classified some genres very good while sacrificing accuracy for other genre types
- Best results for classifying rock music

Future Work

- Since there are limited number of music samples, 1000, we augment our data to 10,000 by resampling. One drawback of this is that music mostly has chorus that repeat itself again and again. Therefore, our models might have seen some of the training data at testing stage. To overcome this issue, either number of resampling needs to be decreased or models need to be trained with a larger data.

- We realized that each model performs better for certain genres. For example, RNN model has its best score for rock music, while reggae is classified best at support vector machine. This problem arises the need of ensemble models that averages out different models. We believe that this might make classifier more robust and accurate.