Measuring the Accuracy of Music Genre Classifier Models Using Cross-collection Evaluation

Borna Salarian • B.salarian@student.tudelft.nl •
Supervisors: Cynthia Liem & Jaehun Kim

TU Delft

1. Background

- Music Information Retrieval (MIR) attempts to analyze and process music and is used for tasks such as recommendation, identification among other tasks.
- There are libraries and tools developed by the MIR community to further advance the field, one of the main ones being Essentia, an open-source library for audio analysis and music based information retrieval.
- Evaluation experiments and interpretation of their results is key to advancing the field. The output of experiments would be evaluated against the ground truth.
- Performance of the ML pipelines commonly used for music processing tasks are better in cross-validation evaluations than in reality.
- Cross-collection evaluation is evaluation of a classifier model with a validation set that has been annotated with an independent source of ground-truth.

2. Objective

- How accurate are the three genre classifiers that are used in Essentia?
  - AcousticBrainz Rosamerica Collection (ROS)
  - GTZAN Genre Collection (GTZAN)
  - Music Audio Benchmark Data Set (MABDS)

3. Methodolgy

To perform cross-collection evaluation, three components are needed:

- **An independent collection of music:**
  We partnered with Muziekweb to gather a collection of music.
- **A set of classifiers:**
  The three genre classifiers that are in Essentia.
- **An independent set of ground truth:**
  To measure the output of classifiers against. We used the collection of tag annotations made possible by the collaboration of users of Last.Fm.

4. Experiments

- From Muziekweb, 1050 songs were randomly chosen. There are roughly 50 songs evenly distributed across 20 genres (Blues, Country, Classical, Disco,Dance, Electronic, Funk, Hip-Hop, Latin-Metal,Rock, Reggae, Pop, Jazz plus 6 regional world songs).
- To gather ground truth, the user generated tags of each of the songs were sought after. Of the 1050, only 385 songs had user generated annotations. The genre distribution heavily changed as more popular songs had more tags. (53 Caribbean/Latin,38 hip-hop,30 country songs, etc.)
- Each song was processed with the models and if the model correctly guessed at least one of the genres of the song, they'd gain more accuracy.

5. Results

The f1-score of each of the models were calculated and can be seen below. The models performed worse than in cross-validation, in particular GTZAN which classified most songs as only jazz. A confusion matrix of MABDS can be seen below.

GTZAN	MABDS	ROS
0.07	0.37	0.29

Accuracy of models

ground truth \ MABDS	blues	country	electronic	hip-hop	jazz	rock
african	0	0	3	0	0	1
asian	1	0	1	0	0	1
blues	0	0	2	0	0	2
caribbean_latin	6	1	28	0	0	12
country	4	4	6	0	0	3
easy_listening	0	0	0	0	1	0
electronic	0	0	106	0	0	1
folk	2	0	19	0	0	6
hip_hop	0	0	20	3	0	2
jazz	10	3	14	0	0	5
pop	0	2	24	0	0	9
rnb	1	1	18	0	0	2
rock	0	1	26	0	0	34

Lorem ipsum

6. Conclusions

The genre classifiers in Essentia did not perform well, with the best one working around only less than 40% accuracy.

A larger database with a better distribution could give a less biased insight into the true performance of the pipelines.