Slotify: an Ensemble Method for Music Genre Classification

Lucas Parzianello
University of Notre Dame
Notre Dame, Indiana
lbarbosa@nd.edu

Eric Tsai
University of Notre Dame
Notre Dame, Indiana
ctsai@nd.edu

1 PROBLEM DEFINITION

In recent years, the increasingly availability of music in several streaming services and personal libraries makes automation a requirement for properly organizing potentially millions of audio tracks. Genre classification is one of the ways we can use to organize such data. In order to automatically categorize the tracks into such categories, we need to first extract audio features from them. This extraction can be hand-crafted (i.e. designed by a specialist) or automated – in the case of the increasingly popular deep neural networks.

This work explores both methods of feature extraction. Our main contributions are:

- Answer which features are more determinant when solving the problem of music genre classification.
- Compare the confusion matrices of handcrafted and automated approaches in order to create an ensemble classifier more accurate than its components in isolation.

2 POSSIBLE SOLUTIONS

Before getting into similar projects and potential solutions, we list below some of the features commonly found to solve tasks involving audio processing.

2.1 Audio Features

From the current literature, we have found the following two lists of handcrafted audio features with potential to aid an automated classifier. They are divided in time and frequency domains:

2.1.1 Time domain.

- Central Moments (CM)
- Zero Crossing Rate (ZCR) [8]
- Root Mean Square Energy (RSME) [12]
- Tempo

2.1.2 Frequency domain.

- Mel-Frequency Cepstrum Coefficients (MFCC) [5, 8–10]
- Spectral Centroid, Bandwidth, Contrast, Roll-Off [7, 8]
- Daubechies Wavelet Coefficient Histogram [8]
- Chroma Features

2.2 Similar Projects

2.2.1 Common classification methods. In Bahuleyan [1], they explore the application of machine learning (ML) algorithms to identify and classify the genre of a given audio file. The conventional ML models that are often seen are Gradient Boosting, Random Forests (RF), Logistic Regression (LR), and Support Vector Machines (SVM). This paper mainly compares the performance of two different classes of methods:

- The first is to make prediction of the genre solely based on its spectrogram as input.
- The second approach is to make prediction of the genre based on features from frequency and time domain.

They train the four conventional ML classifiers mentioned above with these different features and compare their performances. The experiments are conducted on the Audio Set dataset [4] and have an AUC value of 0.894 for an ensemble classifier which combines the two proposed approaches mentioned above.

2.2.2 Hierarchical Taxonomy. In Li and Ogihara [7], they mainly focus on automatic music genre classification based on hierarchical classification with taxonomies. This paper introduce the concept of taxonomy. The hierarchical taxonomy identifies the connection between different genres and provides valuable sources of information for genre classification. This experiment displays different accuracy based on Flat- and Hierarchical-classification, and the Hierarchical-classification has a slightly higher performance in both of their testing Dataset A and B. With this technique, classifiers are able to take care of an easier separable problem and utilize an independently optimized feature set; this leads to improvements in accuracy apart from the gain in training and testing speed. The benefit of applying taxonomy makes the classification errors become more acceptable than in the case of flat classification, which is a type of Divide-and-Conquer approach that makes those errors fall within their level of the hierarchy.

2.2.3 CNN. In Zhang et al. [13], they proposed two ways to improve music genre classification with CNN:

- Method 1: Integrating max- and average-pooling to yield more statistical information to upper level neural networks;
- Method 2: Utilizing "shortcut connections" to bypass one or more hidden layers, a method inspired by residual learning method.

The methodology of their improved CNN is to implement a pile of CNN module, which is used as the feature extractor, for learning mid- and high-level features from the Spectrogram, and followed by a fully connected module, which is utilized as the classifier. CNNs are also used for music genre prediction by Bahuleyan [1].

Data Science, September 2020, University of Notre Dame, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2.3 Our Method

The combination between the mentioned features and models allow us to acquire multiple results. On a first moment, we will train a classifier using traditional techniques such as Support Vector Machines, Logistic Regression, and Random Forests to use them as baseline for comparisons with our ensemble version. In a second moment, we will train neural network using a pre-trained model as our backbone – at the moment we are considering Inception and VGG, in that order. With both models, we propose an ensemble classifier that takes into account the confusion matrices for each model and applies these values as probabilistic weights for a final genre prediction.

3 DATA SOURCES

From our research, we have found a quite few options of datasets containing music tracks (excerpts or integral) with music genre labels. Some popular alternatives are described below:

3.1 GTZAN

GTZAN is one of the most popular public datasets for music genre recognition [11] and is composed of a thousand 30-second audio excerpts labeled across 10 music genres. Despite its popularity, the dataset was not created for music genre classification. Moreover, there are many critics about the dataset quality and whether its size is capable of allowing for accurate or significant results [11].

3.2 SYNAT

The SYNAT database [6] stores over 50 thousand 30-second music tracks in MP3 format, across 22 genres: Alternative Rock, Blues, Broadway & Vocalists, Children's Music, Christian and Gospel, Classic Rock, Classical, Country, Dance and DJ, Folk, Hard Rock and Metal, International, Jazz, Latin Music, Miscellaneous, New Age, Opera & Vocal, Pop, Rap and Hip-Hop, Rock, R&B, and Soundtracks. However, we were unable to find a working download link or request form at the time of writing.

3.3 MSD

The Million Song Dataset (MSD) [2] is a collection of one million songs for which over 190 thousand tracks have consistent genre annotations. Due to the large size of the dataset (around 300GB), MSD is publicly available for research purposes as an AWS EC2 snapshot, rather than a direct download.

3.4 FMA

The Free Music Archive (FMA) dataset [3] is a publicly available alternative containing over 100 thousand audio tracks with four dataset versions of varying track number, lengths, and genres, ranging from 8 thousand tracks of 30 seconds of 7.2GB in total size; to over 106 thousand untrimmed tracks across 161 genres summing 879GB. The audio tracks are under a Creative Commons license and it appears to be the best documented alternative.

Due to the public availability, ease of access, and good documentation, we are inclined to start experimenting with a subset of the FMA dataset.

4 EVALUATION

In order to compare the classification models and fulfill our contributions of (i) which features are most relevant in the handcrafted method, and (ii) build an ensemble model for music genre classification; we plan to extract a list of metrics from our classifiers. Firstly, our dataset will be split into training, validation, and testing sets with disjoint audio tracks and uniform representation across music genres, when possible. Then, once the classifiers models are built, we will extract the metrics below in isolation, and lastly, from our ensemble version.

4.1 Metrics

Some of the metrics we are considering using are:

- Mean accuracy – for a simplified overall idea of a model's performance;
- F1 score – takes into account precision and recall;
- Confusion matrix across genres – to identify which pairs are most challenging for each model and use this information to improve a collective decision;
- Area under the ROC Curve (AUC) – to attest for the model's performance independently of thresholding decisions;

5 TIMELINE

With synchronization checkpoints every Monday, we planned the following execution timeline:

Date	Week	Tasks
Sep. 14	1	Dataset selection, download, and cleanup.
Sep. 21	2	Handcrafted feature extraction.
Sep. 28	3	Implementation of baseline classifier; milestone report.
Oct. 5	4	Improvements on baseline; metrics extraction.
Oct. 12	5	Alternative model implementation.
Oct. 19	6	New metrics and comparison.
Oct. 27	7	Final report writing and presentation.
Nov. 2	8	Further experiments / improvements.
Nov. 9	8.5	Revision and delivery.

REFERENCES

[1] Hareesh Bahuleyan. 2018. Music Genre Classification with Machine Learning Techniques. (2018), 1–4. https://doi.org/10.1109/siu.2017.7960694 arXiv:arXiv:1804.01149v1

[2] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. 2011. The Million Song Dataset. *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011* (2011), 591–596.

[3] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. 2017. FMA: A Dataset for Music Analysis. In *18th International Society for Music Information Retrieval Conference (ISMIR).* arXiv:arXiv:1612.01840v3

[4] Jort F. Gemmeke, Daniel P.W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (2017), 776–780. https://doi.org/10.1109/ICASSP.2017.7952261

[5] Piotr Hoffmann, Andrzej Kaczmarek, Paweł Spaleniak, and Bożena Kostek. 2016. Music Recommendation System. *Asian Journal of Information Technology* 15, 21 (2016), 4250–4254. https://doi.org/10.3923/ajit.2016.4250.4254

[6] Bozena Kostek, Adam Kupryjanow, Pawel Zwan, Wenxin Jiang, Zbigniew W Raś, Marcin Wojnarski, and Joanna Swietlicka. 2011. Report of the ISMIS 2011 Contest: Music Information Retrieval. In *Foundations of Intelligent Systems*, Marzena Kryszkiewicz, Henryk Rybinski, Andrzej Skowron, and Zbigniew W Raś (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 715–724.

[7] Tao Li and Mitsunori Ogihara. 2005. Music Genre Classification with Taxonomy. *In Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, PA, USA* (2005), 197–200.

[8] Tao Li and Mitsunori Ogihara. 2006. Toward Intelligent Music Information Retrieval. *IEEE Transactions on Multimedia* 8, 3 (2006), 564–574. https://doi.org/10.1109/TMM.2006.870730

[9] Shin Cheol Lim, Jong Seol Lee, Sei Jin Jang, Soek Pil Lee, and Moo Young Kim. 2012. Music-Genre Classification System Based on Spectro-Temporal Features and Feature Selection. *IEEE Transactions on Consumer Electronics* 58, 4 (2012), 1262–1268. https://doi.org/10.1109/TCE.2012.6414994

[10] Loris Nanni, Yandre M.G. Costa, Alessandra Lumini, Moo Young Kim, and Seung Ryul Baek. 2016. Combining Visual and Acoustic Features for Music Genre Classification. *Expert Systems with Applications* 45 (2016), 108–117. https://doi.org/10.1016/j.eswa.2015.09.018

[11] Bob L. Sturm. 2013. The GTZAN Dataset: Its Contents, Its Faults, Their Effects on Evaluation, and Its Future Use. 11 (2013), 1–29. https://doi.org/10.1080/09298215.2014.894533 arXiv:1306.1461

[12] Ran Tao, Zhenyang Li, and Ye Ji. 2010. Music Genre Classification Using Temporal Information and Support Vector Machine. In *ASCI Conference, 2010.*

[13] Weibin Zhang, Wenkang Lei, Xiangmin Xu, and Xiaofeng Xing. 2016. Improved Music Genre Classification with Convolutional Neural Networks. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 08-12-Sept (2016), 3304–3308. https://doi.org/10.21437/Interspeech.2016-1236