Multimodal multi-genre classification using lyrics and audio properties

Allan Misasa Nielsen – QZD426

University of Copenhagen

Table of Contents

Abstract

The topic of Music Recommender Systems has enjoyed rising popularity over the last decade, but studies on genre labelling have historically focused on single-class genre classification. Now, datasets spanning different modalities on music has been released, along with multi-genre labels rendering tests on multimodal multi-class classification on music feasible. With a starting point in cognitive musicology and affective computing, we pick features and test a way to learn algorithms on different modalities and combine them using Restricted Boltzmann Machines to make a joint classification of multiple genres.

Keywords: Multimodal machine learning, multilabel classification

Introduction

Music recommender systems have seen increasing rates of improvement with the richness of public music data that has appeared throughout the last decade. Among the datasets are the Million Song Dataset[1] for audio features, and the musiXmatch[2] dataset for lyrics.

Recent studies (Demetriou, Jansson, Kumar, & Bittner, 2018) have uncovered elements in music that indicates salient attributes to music listeners, of which many exist in the Million Song Dataset. Since genre labelling is a manual task with an element of subjectivity, salient traits influencing human perception should also help classify genres.

Much modern music are fusion genres, e.g. jazz and rock, pop and rock, etc. This means that genres can no longer be treated as mutually exclusive, thus single-genre classification as has historically been the method of classifying genres lacks the ability to capture the genre-spanning music that exists today.

---

[1] http://millionsongdataset.com/
[2] http://millionsongdataset.com/musixmatch/

Related work

Until recently, most genre classifications have been constrained to unimodal representations of audio (Sturm, 2012). Multimodal, multilabel classification has gained traction in later years, where studies managed to reliably classify single genres on albums based on audio and text modalities (Neumayer & Rauber, 2007). Later, deep learning approaches using convolutional neural networks and combinations of audio, visual and textual representations have been tested (Oramas, Nieto, Barbieri, & Serra, 2017), (Orasmas, Barbieri, Nieto, & Serra, 2018).

So we see that multilabel classification has been on the rise, and in 2016, multilabel classification became accessible for less skilled engineers with the release libraries such as scikit-learn for the programming language Python, which also is suitable for music categorization (Szymanski & Kajdanowicz, 2016). These days, datasets of different modalities are released with ever-increasing frequency, enabling multimodal research with ever more detailed features and targets.

Data description

We are using a subset of the Million Song Dataset (MSD) for audio features and genre labels, and the Musixmatch (MXM) dataset for song lyrics.

The MSD dataset contains audio features as processed by the Echo Nest Analyze API (now Spotify API). The MXM dataset supplies us with song lyrics in a bag-of-words format, which serves us well for statistical purposes. It does restrict us from performing more advanced NLP methods however, as any contextual clues, word orders and syntax are missing.

The cognitive aspects

The choice of using language and audio data is not simply a matter of availability. It has been shown that melodic phrases produce neuronal activations common to that of language, including Broca's area, ventral thalamus and posterior cerebellum (Brown, Martinez, & Parsons, 2006).

Emotions play a large role in how we perceive music (Demetriou, Jansson, Kumar, & Bittner, 2018), and there are particular structural features in music that we tend to associate particular emotions to. For instance, the mode (major/minor) tend to be associated with joy and sadness, respectively, tempo can be linked to a scale of emotions from happiness and arousal, to sadness and peace, along with many other factors linked to emotional valence (Gabrielle & Stromboli, 2001). The importance of mode in eliciting emotions is unsurprising, given the breadth and depth of literature surrounding it (K., 1936), (Juslin, 1997), (Vieillard, 2008) .

The choice of features for this project is inspired by the vast papers on the topic of emotion classification incorporate valence and arousal as the two basic dimensions. With outset in Russell's circumplex model (Russell, 1980), we can define a model fitted to music, where key, mode and time signature spans the valence dimension, and duration, loudness and tempo spans the arousal dimension, which falls in line with previous studies that attempt to find music features that correspond to valence and arousal dimensions (Grekow, 2018), (Baume, Fazekas, Barthet, Marston, & Sandler, 2014). Therefore, these features make up the audio classification part of this paper.

The argument of making classifications partially based on emotive correspondences, would be moot if there did not exist a unified perception of emotional impacts of music. Indeed, an argument could pe posited, that due to a combination of different personalities, experiences and

cultural backgrounds, people *should* perceive music differently in terms of emotions. This does not seem to be the case, instead, those personal factors determine their preferences in music emotion (Bonneville-Roussy, Rentfrow, Xu, & Potter, 2013), not so much their perception of said emotion (Limer, 2017). And just as people generally agree on broad genre definitions, they tend to agree on the emotions elicited by music (Krumhansl, 2002). Therefore, we can consider the perception of emotion in music and of genre labels stable across individuals, which serves as an argument for the approaches in this paper.

Data pre-processing

We found 2350 matching Track IDs in the two datasets (MXM & MSD). Of these, 1095 have non-existent genre labels which we need as targets, which leaves us with 1255 tracks after removing the unlabeled tracks.

For the labels, we want to use genres of the tracks, but since they are user-submitted they need cleaning. After some data exploration, we find that most genre tags are amongst, or include keywords that correspond to major, distinct genres, listed below:

['jazz', 'pop', 'rock', 'metal', 'indie', 'classical', 'punk', 'electronic', 'blues', 'ambient', 'folk', 'hip hop', 'country']. This reduces the number of distinct labels from more than a 100, to just 11, while preserving the multiple tags that many songs have, and without losing too much data due to incongruences in genre types.

Because we still have 59 unique mixtures of classes, ratio of data-to-class is too low for efficient training. Therefore, we algorithmically leave one genre out at a time, attempting to lower the number of unique class mixtures, while preserving enough data. This leaves us with data on 738 tracks, with 11 unique class mixtures, comprising the genres [‘Jazz’, ‘Rock’, ‘Classical’, ‘Electronic’, ‘Folk’]. These genres seem to make intuitive sense, insofar as they seem different

enough that confusion between two genres is minimized, while still capturing a broad spectrum of music – most genres are sub-genres of the chosen genres[3]; in other words, the genres chosen coincides with the proposed major genre clusters that were identified by the creators of Musicmap[4]. Given that the genres chosen are subjectively considered unique clusters, they may aid the differentiation between them – hence the genre "pop" is not included, due to it spanning a wide range of temperaments, often crossing other genre boundaries.

Preparing for classification

A Pearson correlation matrix was made on all the MSD features of the chosen subset.

Index	Duration	KeySignature	Loudness	Tempo	TimeSignature	Key	Mode
Duration	1	-0.0770772	0.0554797	-0.0837418	0.0339228	-0.0770772	-0.0594382
KeySignature	-0.0770772	1	0.0236055	-0.000212886	-0.0266979	1	-0.156336
Loudness	0.0554797	0.0236055	1	0.172955	0.112011	0.0236055	-0.0424305
Tempo	-0.0837418	-0.000212886	0.172955	1	0.129737	-0.000212886	0.0154879
TimeSignature	0.0339228	-0.0266979	0.112011	0.129737	1	-0.0266979	-0.0871958
Key	-0.0770772	1	0.0236055	-0.000212886	-0.0266979	1	-0.156336
Mode	-0.0594382	-0.156336	-0.0424305	0.0154879	-0.0871958	-0.156336	1

Notice that the features KeySignature and Key are fully correlated; they are likely duplicates, therefore we drop the feature KeySignature. The rest of the data are uncorrelated; thus, we can safely leave the rest of the features as is.

Lyrics were supplied in a bag of words format, which in itself is a poor feature for classification (Elena Rudkowsky, 2018). It is possible to unpack the lyrics and proceed to get more detailed features, and we managed to create TF-IDF features from the unraveled lyrics, which becomes the basis for the textual modality.

---

[3] https://www.musicgenreslist.com/
[4] https://www.musicmap.info/

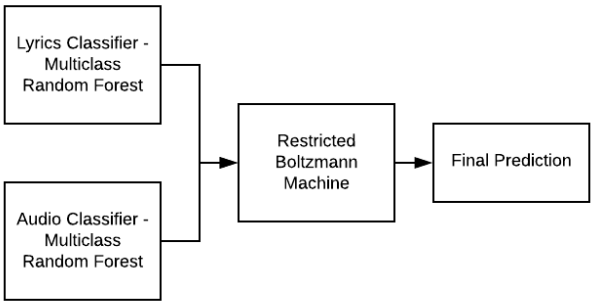
We are allowing fusion genres that can be any combination of the five chosen genres. In order to ensure that different combinations of genres are mutually exclusive, we employ the MultiLabelBinarizer algorithm from the Python library SKlearn[5], which forms a binary matrix indicating the presence of a class label out of the labels. For example, in a four-genre classification task, with genres [Rock, pop, jazz, classical], a pop-rock track is labelled 1100, and a purely classical track is labelled 0001.

Classification

The classification problem is of the multimodal, multiclass type, which requires one classifier for each modality[6], after which there are many methods of unifying them, such as ensemble averaging, winner-takes-all, and the method used in this paper – a Boltzmann machine, as it has seen an increase in popularity among multimodal learning circles (Srivastava & Salakhutdinov, 2014).

During our first classification, One-versus-rest classification was performed by training two random forest classifiers separately on the two modalities. The outputs of these were fed to a restricted Boltzmann machine (RBM) to create a joint classification. The main idea, is that the RBM learns from the co-occurrences of the outputs of the classifiers. An example of this usage case of RBMs is clarified by Chris Nicholson[7] .

The classification pipeline is shown to the right:



[5] https://scikitlearn.org/stable/modules/generated/sklearn.preprocessing.MultiLabelBinarizer.html

[6] Except for joint representations, such as one produced by a restricted Boltzmann machine if used in stage 1.

[7] https://pathmind.com/wiki/restricted-boltzmann-machine

The idea in a real-world scenario comes from advances in recommender systems (voting systems), where outputs from each modality's classifier can be regarded a vote, and the RBM serves as a sophisticated mechanism in making a joint representation of the 'votes' in the classifiers. It assigns the probability to a vector **v** (the proposed output) given by the Boltzmann distribution:

$$P(\mathbf{v};\theta) = \frac{1}{z(\theta)}\sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}^1, \mathbf{h}^2, \mathbf{h}^3; \theta)).$$

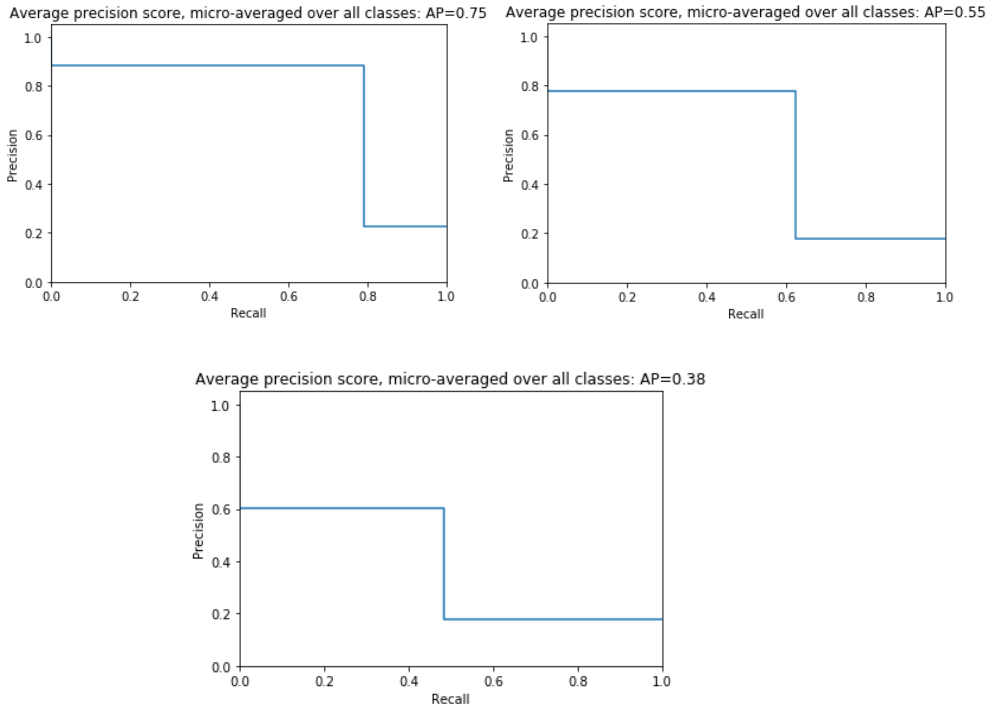
The choice of RBM over a conventional Boltzmann machine stems from the fact that RBM's neurons must form a bipartite graph (Larochelle & Bengio, 2008), far simplifying the workload, and maintaining a simplicity that is appropriate for the task. In addition, it is also common to use RBM's when mixing modalities (Srivastava & Salakhutdinov, 2014).

Evaluating the classification

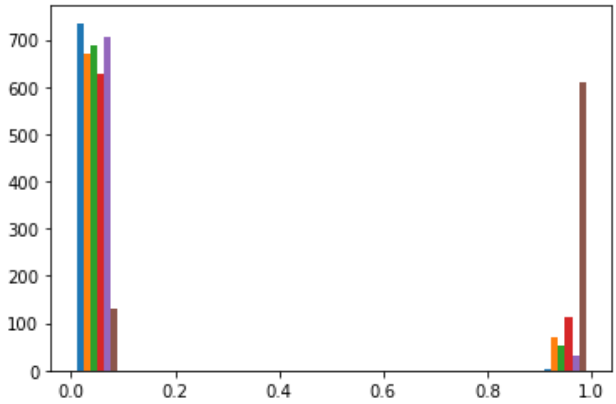
We have opted for precision-recall scores over the simpler, and more widely used accuracy scores. Given the imbalanced nature of the dataset, precision and recall serves as preferable estimates to accuracy, as the number of false positives may be high, which accuracy will not necessarily capture. Additionally, it has been found that precision-recall curves yield more information in imbalanced scenarios than receiver operating characteristic curves, which might be deceptive in evaluating this type of task (Saito & Rehmsmeier, 2015). Since this is a multiclass setting, we opt for the average precision scores as a benchmark, which relates the precision and recall scores of all targets, as given by the formula

$$AP = \sum_n (R_n - R_{n-1}) P_n$$

, where P_n and R_n are the precision and recall scores at the nth threshold[8].

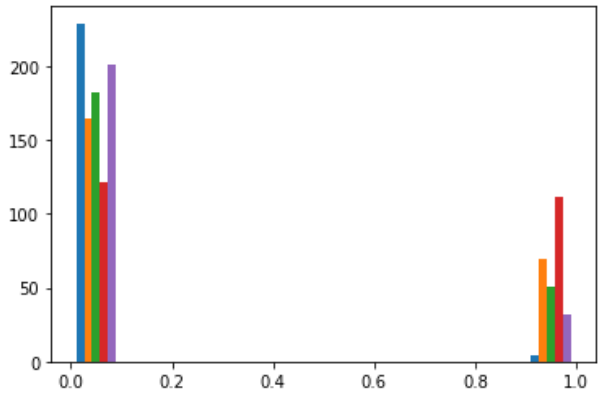


The above three plots show the average precision scores of 4, 5, and 6 genre classification schemes. They show that the classification scheme does not respond well to classifications of genre combinations comprised of more than 4 distinct genres; it is likely that they get increasingly confused with each other, which is unsurprising due to the slight overlap between valence and arousal between genres (Eerola, 2011). Another factor could be the imbalance in the dataset. In fact, in the six-genre classification, rock music (brown) created a heavy imbalance as evident in the following histogram:



[8] https://towardsdatascience.com/breaking-down-mean-average-precision-map-ae462f623a52

By removing rock music from the label set, we already see a much higher balance:



It is likely that the weak recall at higher number of labels is the consequence of a higher false positive rate, by the introduction of heavy bias toward the majority label - an effect that is widely observed in real-world datasets (López, Fernandez, Garcia, Palade, & Herrera, 2013).

Discussion

While the results are satisfactory given the novelty and complexity of the problem, it should also be discussed that the approach is admittedly rudimentary, partially due to the main focus being on the cognitive aspects of modelling our perception of music genres. The possible quality degrading effects are mostly seen in the use of pre-engineered audio features from a third party (MSD), and the lack of full lyrics to generate deeper features from, such as a particular emotion feature since emotion seems intrinsically linked to genre perception (Eerola, 2011). It could also be interesting to add a third modality, such as album covers, or music videos. However, with limited computing resources and the lack of a unified, multimodal music dataset, the approaches used are satisficing. With enough data to balance out the class imbalance, the classification method should scale to class permutations $n > 4$ with fewer false positives.

We used restricted Boltzmann machines partially due to the lower computing power requirements. It would be interesting to see how an unrestricted Boltzmann machine would perform in this environment. Using Boltzmann machines before the classification step could be of particular interest, as this will result in a joint representation of the chosen modalities, which would then allow a broader array of classification methods to be used – in our example, we were restricted to *k*-neighbors classification and tree models due to the multi-output, multi-label situation, originating in the maintained separation of input modalities.

There is also the question of the genre effects of multimodality. It can be argued that classifications of genres are ambiguous across modalities; that lyrics may tell one story, and audio features may tell another, resulting in rising ambiguity due to the directions of the different vectors representing each feature.

Bibliography

Baume, C., Fazekas, G., Barthet, M., Marston, D., & Sandler, M. (2014). Selection of audio features for music emotion recognition using production music. *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio.* Retrieved from http://www.aes.org/e-lib/browse.cfm?elib=17110

Bonneville-Roussy, A., Rentfrow, P. J., Xu, M. K., & Potter, J. (2013). Music through the ages: Trends in musical engagement and preferences from adolescence through middle adulthood. *Journal of Personality and Social Psychology, 105*(4), 703-717. doi:10.1037/a0033770

Brown, S., Martinez, M. J., & Parsons, L. M. (2006). Music and language side by side in the brain: A PET study of the generation of melodies and sentences. *European Journal of Neuroscience, 23*(10), 791-803. doi:https://doi.org/10.1111%2Fj.1460-9568.2006.04785.x

Demetriou, A., Jansson, A., Kumar, A., & Bittner, R. M. (2018). *Vocals in music matter: The relevance of vocals in the minds of listeners.*

Eerola, T. (2011). Are the Emotions Expressed in Music Genre-specific? An Audio-based Evaluation of DatasetsSpanning Classical, Film, Pop and Mixed Genres. *Journal of New Music Research, 40*(4), 349-366. doi:10.1080/09298215.2011.602195

Elena Rudkowsky, M. H. (2018). More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures*, 140-157.

Gabrielle, A., & Stromboli, E. (2001). The influence of musical structure on emotional expression. *Music and Emotion: Theory and Research*, 223-242.

Grekow, J. (2018). Audio features dedicated to the detection and tracking of arousal and valence in musical compositions. *Journal of Information and Telecommunication, 2*(3), 322-333. doi:10.1080/24751839.2018.1463749

Juslin, P. N. (1997). Emotional communication in music performance: a functionalist perspective and some data. *Music Perception, 14*(4), 383-418.

K., H. (1936). Experimental studies of the elements of expression in music. *The American Journal of Psychology, 48*(2), 246-268.

Krumhansl, C. L. (2002). Music: A Link Between Cognitionand Emotion. *Current Directions in Psychological Science, 11*(2), 45-50.

Larochelle, H., & Bengio, Y. (2008). Classification using Discriminative Restricted Boltzmann Machines. *Proceedings of the 25th international conference on Machine learning (ICML '08)*, (pp. 536-543). doi:https://doi.org/10.1145/1390156.1390224

Limer, C. (2017). Perceiving Emotions Through Music. *FODL Library Awards*.

López, V., Fernandez, A., Garcia, S., Palade, V., & Herrera, F. (2013). An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics. *Information Sciences*, 113-141.

Neumayer, R., & Rauber, A. (2007). Integration oftext and audio features for genre classification in mu-sic information retrieva. *European Conference onInformation Retrieval* (pp. 724-727). Springer.

Oramas, S., Nieto, O., Barbieri, F., & Serra, X. (2017). Multi-Label Music Genre Classification from Audio, Text, and Images using Deep Features.

Orasmas, S., Barbieri, F., Nieto, O., & Serra, X. (2018). Multimodal Deep Learning for Music Genre Classification. *Transactions of the International Society forMusic Information Retrieval.*

Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161-1178. doi:10.1037/h0077714

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. (G. Brock, Ed.) *PLOS One, 10*(3). doi:10.1371%2Fjournal.pone.0118432

Srivastava, N., & Salakhutdinov, R. (2014). Multimodal Learning with Deep Boltzmann Machines. *Journal of Machine Learning Research 15*, 2949-2980.

Sturm, B. L. (2012). A survey of evaluation in musicgenre recognition. *International Workshop onAdaptive Multimedia Retrieval*, (pp. 29-66).

Szymanski, P., & Kajdanowicz, T. (2016). scikit-multilearn: A scikit-based Python environment forperforming multi-label classification. *Journal of Machine Learning Research 1*, 1-15.

Vieillard, S. P. (2008). Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognition & Emotion, 22*(4), 720-752.