Timbral and Semantic Features for Music Playlists

Andreu Vall, Hamid Eghbal-zadeh, Matthias Dorfer, Markus Schedl ANDREU.VALL@JKU.AT
Department of Computational Perception, Johannes Kepler University, Linz, Austria

Abstract

Music recommendation methods based on song-level features (either derived from audio content or metadata), suffer from not being able to identify clear relations between music items and listeners, whose perception of the quality of a received recommendation is affected by a wider range of factors. This problem is particularly severe for the task of generating music playlists. The analysis of song characteristics in hand-curated playlists exhibits large within-playlist variability, indicating that generating music playlists or creating suitable continuations may be an ill-defined problem. In this paper we analyze two different features, based on either timbral or semantic descriptors of songs, for the task of predicting whether a song is suitable or not for a playlist. Our empirical results on a dataset of hand-curated playlists indicate that features extracted from semantic descriptors are better suited for this task.

1. Suitable Features for Playlist Generation

Music recommender systems try to assist listeners to navigate large music collections. Automated music playlist generation is a special music recommendation task, where the recommender system needs to pay special attention to the whole sequence, regarding aspects such as intent, flow or coherence. However, this is only a general target that may be fulfilled in many different ways. The analysis of interviews with practitioners and postings to a dedicated web site in (Cunningham et al., 2006) gives a hint of the complexity of preparing playlists. Also, the distribution of content-based features within hand-curated examples was not discriminative of possible different groups of playlists in (Choi et al., 2015).

Models based on collaborative filtering (Aizenberg et al., 2012) are able to uncover complex latent relations. However, they are not suited for tasks where the song characteristics may be important, e.g. finding songs with similar instrumentation, or dealing with new songs in the system.

In this paper we propose to examine further which may be suitable song-level features for automated music playlist generation. To do so, we tackle a simplified version of the problem, set as a classification task. We consider playlists as classes and study the predictive power of the song characteristics to classify songs to the right playlists. We experiment with two different types of features, namely the so-called *i-vectors* extracted from timbre descriptors of the songs, and vector representations of words providing semantic descriptions of the songs. For convenience, we will refer to the latter as *word-vectors* from now on.

2. Methodology

2.1. I-Vectors from Timbre Descriptors

I-vector features were first introduced in the field of speaker verification (Dehak et al., 2011) and recently have been successfully utilized for music similarity and music artist recognition tasks (Eghbal-zadeh et al., 2015a;b).

Given a collection of songs, a so-called *universal background model* is built on the basis of frame-level features (e.g. MFCCs), in order to capture similarities among them. For the process to be successful, there should be a sufficient number of songs, representative of the classes (e.g. playlists) we may be interested in. The captured similarities are then discarded using factor analysis, by projecting the songs into a new space where the remaining factors are in strong correlation with the eventual class variability. The new song components in this final space are expected to contain rich discriminative information.

2.2. Word-Vectors from Semantic Terms

The embedding of words into continuous vector spaces, as e.g. in (Mikolov et al., 2013), allows us to exploit any

© Andreu Vall, Hamid Eghbal-zadeh, Matthias Dorfer, Markus Schedl, Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Andreu Vall, Hamid Eghbal-zadeh, Matthias Dorfer, Markus Schedl, "Timbral and Semantic Features for Music Playlists" *Machine Learning for Music Discovery Workshop* at the 33^{rd} *International Conference on Machine Learning*, New York, NY, 2016.

words related to songs as semantic features.

We utilize song titles, artist names and artist-related terms (mostly genre tags), which we assume to carry relevant semantic information. We first build a representative and large enough text corpus by querying Wikipedia[1] with the aforementioned terms and gathering the returned content. We use the implementation of the continuous bag-of-words algorithm available in *word2vec*[2] on top of the text corpus to obtain word-vectors for the most relevant terms in the corpus. Finally, each song is represented by the average of the word-vectors corresponding to its terms.

2.3. Classification Task

Given a song and a collection of playlists, we want to decide if the song is suited or not for each of the playlists. We propose to use a feed-forward neural network that takes song features as inputs. The output is passed through sigmoidal activation functions, so that we make as many independent decisions as playlists in the collection. Although the final decisions are independent, the network is trained considering all the predictions jointly, eventually learning representations useful for all the decisions.

We use features of 200 dimensions for both i-vectors and word-vectors. They are input to a neural network with 2 fully connected layers and 500 hidden units each. The hidden layers have hyperbolic tangents as activation functions and the output layer has logistic functions. The network is optimized to minimize the average binary cross-entropy between predictions and targets.

3. Experimental Study

3.1. Dataset

We use the Art of the Mix 2011 dataset, presented in (McFee & Lanckriet, 2012). The dataset is a collection of hand-curated playlists from the Art of the Mix[3] database. Each playlist is represented by the song titles and artist names and also has a category label providing a general description of its content. Some of the songs within the playlists are linked to the Million Song Dataset (MSD) (Bertin-Mahieux et al., 2011), allowing us to use the timbre features and the artist-related terms from the Echonest[4] to compute timbre i-vectors and semantic word-vectors.

3.1.1. Subset for Feature Extraction

In order to build the universal background model and the text corpus required to extract the timbral and semantic features, we subset the dataset seeking to obtain a representative sample of the different types of playlists available. We select playlist segments with at least 6 contiguous songs linked to the MSD. We discard the playlist segments belonging to rare categories (i.e., those tagged with category labels used in 30 or less playlist segments), and for the highly populated categories (50 or more playlist segments) we keep only 50 random playlist segments. There remain 1,085 playlist segments containing 6,088 unique songs.

3.1.2. Subset for the Classification Task

For the classification task we subset playlists with at least 19 songs linked to the MSD, so that all playlists have a minimum of examples. Since the word-vectors are informed about the song artist, we further filter out the playlists with less than 5 unique artists, to make the comparison fairer. This results in 226 playlists with 4,123 unique songs.

3.2. Results and Discussion

We conduct 5-fold cross-validation, balancing the number of observations of each playlist between folds. The network is trained using AdaGrad (Duchi et al., 2011) with Nesterov momentum for a maximum of 100 epochs, although the process may stop before if no significant progress is made. We use mini-batches of 50 observations and a learning rate of 0.1.

Table 1 reports the classifier's binary cross-entropy loss and average precision score (i.e., area under the precision-recall curve) for the epoch in which minimum validation loss is achieved.

The i-vectors could not classify the songs correctly, as it is clear from the validation average precision score under 1%. They also showed a very fast and severe overfit if trained longer. The word-vectors exhibited difficulties as well, but demonstrated some extent of classification power achieving a validation average precision score around 8%.

The presented findings suggest interesting next steps. The timbre-based features appeared to be too specific for this task, while the semantic-based features could encode part of the variability of songs within playlists. Further analysis is required, as well as studying whether both features can complement one another if used together.

Table 1. Performance of the classifier.

	Loss		AveP	
feature	train	valid	train	valid
i-vectors	0.0306	0.0302	2.98%	0.64%
word-vectors	0.0232	0.0250	16.72%	8.13%

[1] en.wikipedia.org
[2] code.google.com/p/word2vec
[3] www.artofthemix.org
[4] the.echonest.com

References

Aizenberg, Natalie, Koren, Yehuda, and Somekh, Oren. Build your own music recommender by modeling internet radio streams. In *Proc. WWW*, pp. 1–10. ACM, 2012.

Bertin-Mahieux, Thierry, Ellis, Daniel PW, Whitman, Brian, and Lamere, Paul. The million song dataset. In *Proc. ISMIR*, pp. 591–596. University of Miami, 2011.

Choi, Keunwoo, Fazekas, Gyrgy, and Sandler, Mark. Understanding music playlists. In *Machine Learning for Music Discovery Workshop*, Lille, France, 2015.

Cunningham, Sally Jo, Bainbridge, David, and Falconer, Annette. "More of an art than a science": Supporting the creation of playlists and mixes. In *Proc. ISMIR*, Fairmont Empress Hotel, Victoria, BC, Canada, 2006.

Dehak, Najim, Kenny, Patrick J, Dehak, Rda, Dumouchel, Pierre, and Ouellet, Pierre. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, May 2011. ISSN 1558-7916, 1558-7924. doi: 10.1109/TASL.2010.2064307.

Duchi, John, Hazan, Elad, and Singer, Yoram. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.

Eghbal-zadeh, Hamid, Lehner, Bernhard, Schedl, Markus, and Widmer, Gerhard. I-vectors for timbre-based music similarity and music artist classification. In *Proc. ISMIR*, October 2015a.

Eghbal-zadeh, Hamid, Schedl, Markus, and Widmer, Gerhard. Timbral modeling for music artist recognition using i-vectors. In *Proc. EUSIPCO*, pp. 1286–1290. IEEE, August 2015b.

McFee, Brian and Lanckriet, Gert. Hypergraph models of playlist dialects. In *Proc. ISMIR*, pp. 343–348, 2012.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S., and Dean, Jeff. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.