Country Classification with feature selection and network construction for folk tunes

Cornelia Metzig[1], Roshani Abbey[2], Mark Sandler[1], Caroline Colijn[3]

[1]Centre for digital music, Queen Mary University of London, UK
[2]Royal Academy of Music, London, UK
[3]Simon Fraser University, Burnaby, Canada

June 28, 2019

Abstract

We explore two approaches to quantify folk song similarity. In the first part of this paper, we investigate to what extent the a folk tunes differ by country of origin. If it is possible for a human with a limited set of tunes in mind to guess the country of origin correctly, then there must be a signal that can be detected with automatic classification. This question has been addressed in the literature, where classifiers were trained both on global and local features [6, 5, 3].

In this paper, we aim at predicting country of origin based on extracted features We use a large number local features (n-grams of note successions and rhythm successions), for a MIDI dataset of songs from England, Scotland, Ireland, Germany, USA, Spiritual songs and African songs. Each group contains 80 songs, although larger groups have been used for comparison. These extracted features contain indirectly information of global features such as mode and time signature. Since we use a high number of initially extracted features ($\approx$ 7000 that are not 0) we reduce this number with statistical filtering methods, before performing random forest classification on them. In addition we use feature selection, based on the importances of the random forest classifier, which strongly increases accuracy. Our method allows us to predict country of origin with up to 91% accuracy, depending also on the degree of similarity of the two countries that were used for training. We use 10-fold cross validation to reduce overfitting, and compare results to a classifier trained on the same data with randomly assigned country labels. The importances of the features are interesting since they reveal typical patterns of tunes for each country. Interestingly, rhythm grams and melodic grams of different lengths are all relevant. The results depend strongly on the country pair. They confirm established results from musicology but also have the potential to complement them, due to the systematic way they are found.

However, country of origin is not the only relevant aspect when studying song similarity. We construct similarity networks of songs, based on these extracted features. This approach can reveal song families that follow other patterns than countries, and may contain interesting information regarding the history of songs, for example many links between Scottish and American songs, or between African songs and Spirituals. To construct a network, we first calculated Hamming and Euclidean distances between the song vectors of extracted features. Whenever the distance between two songs falls below a certain threshold, a link between those songs is created. This simple method bears the problem of so-called "hubness" [1, 2], which is an aggregation artefact due to the fact that the distances are calculated in very high-dimensional spaces. Songs without apparent similarity will appear close (and therefore connected in our network), and highly connected hub songs will emerge. We explore two methods correcting for this: (i) to use fewer dimensions, which are selected based on random forest importances in the classification. This reduces hubness and creates plausible networks, however some hghly connected songs remain, unless the number of dimensions is reduced to a number where the interesting similarities disappear as well. We compare our feature selection approach with a method of reducing hubness in audio similarity introduced by [4], the so-called mutual proximity (MP). Based on the idea of using only the closest songs and constructing a symmetric distance from it, MP reduces hubness, however loses some of the interesting information as well. We propose a combination of MP and selecting important dimensions. Although it is ultimately a matter of choice which songs to consider to be similar, the proposed method yields the most plausible results. We find strong clusters that belong within one country, and also that songs from certain countries

of origin (e.g. Germany) are much more distant than others. Also heterogeneity within one country group differs strongly. Genre seems to be much less informative than country, e.g. drinking songs and Christmas carols of one country are often connected.

The classification method has the potential to be applied to smaller geographic regions than countries, as a tool to help location, or to melodies from different composers, to identify authorship.

References

[1] Jean-Julien Aucouturier and Francois Pachet. A scale-free distribution of false positives for a large class of audio similarity measures. *Pattern recognition*, 41(1):272–284, 2008.

[2] Arthur Flexer, Dominik Schnitzer, and Jan Schlüter. A mirex meta-analysis of hubness in audio music similarity. In *ISMIR*, pages 175–180, 2012.

[3] Ruben Hillewaere, Bernard Manderick, and Darrell Conklin. Global feature versus event models for folk song classification. In *ISMIR*, volume 2009, page 10th. Citeseer, 2009.

[4] Dominik Schnitzer, Arthur Flexer, Markus Schedl, and Gerhard Widmer. Using mutual proximity to improve content-based audio similarity. In *ISMIR*, volume 11, pages 79–84, 2011.

[5] Peter van Kranenburg, Anja Volk, and Frans Wiering. A comparison between global and local features for computational classification of folk song melodies. *Journal of New Music Research*, 42(1):1–18, 2013.

[6] Anja Volk and Peter Van Kranenburg. Melodic similarity among folk songs: An annotation study on similarity-based categorization in music. *Musicae Scientiae*, 16(3):317–339, 2012.