FROM MULTI-LABELING TO MULTI-DOMAIN-LABELING: A NOVEL TWO-DIMENSIONAL APPROACH TO MUSIC GENRE CLASSIFICATION

Hanna Lukashevich, Jakob Abeßer, Christian Dittmar, Holger Grossmann
Fraunhofer Institute for Digital Media Technologies, Ilmenau, Germany
{lkh, abr, dmr, grn}@idmt.fraunhofer.de

ABSTRACT

In this publication we describe a novel two-dimensional approach for automatic music genre classification. Although the subject poses a well studied task in Music Information Retrieval, some fundamental issues of genre classification have not been covered so far. Especially many modern genres are influenced by manifold musical styles. Most of all, this holds true for the broad category "World Music", which comprises many different regional styles and a mutual mix up thereof. A common approach to tackle this issue in manual categorization is to assign multiple genre labels to a single recording. However, for commonly used automatic classification algorithms, multi-labeling poses a problem due to its ambiguities. Thus, we propose to break down multi-label genre annotations into single-label annotations within given time segments and musical domains. A corresponding multi-stage evaluation based on a representative set of items from a global music taxonomy is performed and discussed accordingly. Therefore, we conduct 3 different experiments that cover multi-labeling, multi-labeling with time segmentation and the proposed multi-domain labeling.

1. INTRODUCTION

In the field of Music Information Retrieval, automatic genre classification has been covered in numerous publications. Although genre labels as being used in online music stores or music journals mostly represent marketing terms, genre itself embodies both a culturally relevant term and an intuitive concept to categorize music. Single genre labels usually reflect some sort of stylistic elements inherent to a piece of music. Especially nowadays, music is influenced by an increasing amount of different musical styles. This leads to the necessity of describing single recordings with multiple genre labels. At the same time, this increases ambiguity in case a distinct genre classification result is intended. We stumbled across this problem while attempting to train supervised classifiers for a given sub-genre classification taxonomy of global music content. It is obvious that the broad term "World Music" is one of the most ill-defined tags when being used to lump all "exotic genres" together. It lacks justification because this category comprises such a huge variety of different regional styles, influences, and a mutual mix up thereof. On the one hand, retaining the strict classification paradigm for such a high variety of musical styles inevitably limits the precision and expressiveness of a classification system that shall be applied to a world-wide genre taxonomy. On the other hand, multi-labeling is not straight forward to deploy for automatic supervised classification since data sets with multiple class assignments are not well suited as training data due to their inherent ambiguity. To tackle these issues, we considered to break down the multi-label genre classification problem into a set of single-label genre classification tasks, where each classifier can be trained and optimized using well-defined data. The novelty of the proposed 2-dimensional approach for multi-label genre classification consists in the combination of segment-wise and domain-specific genre classifications. The term "domain" refers to the perceived semantic dimensions of music in which the classification is performed, in our case timbre, rhythm and tonality, which represents melody and harmony. We call the introduced approach "multi-domain labeling". In this paper we evaluate and discuss the potential of a more detailed approach directly, compared to multi-label genre classification. The rest of this paper is organized as follows. We give an overview over related work to this topic in the subsequent section. Then, after explaining our novel approach in section 3, we give an overview over the utilized databased as well as the manual genre annotations corresponding to the proposed method in section 4. In the following section, we describe the 3 evaluation experiments that we performed. Details on feature extraction, feature selection, feature space transformation as well as on the applied classification algorithms are presented in section 6. After discussing the results of the experiments in section 7, we conclude our work and provide perspectives for future directions in section 8.

2. RELATED WORK

Various classification schemes for automatic genre classification have been proposed during the last years. [17] provides a comprehensive overview over existing publications in the domain. There different approaches related to expert systems, unsupervised classification, and supervised classification systems have been covered. Considering the general confusion between similar genres, relaxing the strict classification paradigm and allowing for multiple-genre classification seemed to be a reasonable future direction to the authors to implement a more realistic classification system. The earlier work of [3] gave a more pessimistic outlook by considering the term genre to be intrinsically ill-defined and hardly grounded in timbre characteristics. Already in one of the basic works on music genre classification by Tzanetakis [21], separate feature sets representing timbre, rhythm, and tonality were introduced that allowed for different types of similarity measures. However, separate domain-specific genre models have not been proposed there. [17] provides also an overview of different classifier approaches applied in genre classification, such as Support Vector Machines (SVM), Hidden Markov Models (HMM) or Artificial Neural Networks (ANN). Other publications such as [21] utilized Gaussian mixture models (GMM) for this purpose. Among others, the authors of [18] used ensemble-based decision approaches namely a one-against all and a round-robin algorithm to combine binary classifiers. Different feature-space transformation methods such as Linear Discriminant Analysis are applied to increase discrimination between the classes resulting in better classifications scores [17]. Novel musically motivated low- and mid-level features such as the Octave-based Modulation Spectral Contrast [11] or multiscale spectro-temporal modulation features [15] were reported to outperform conventional features such as Mel-Frequency Cepstral Coefficients (MFCCs). Moreover, an increasing amount of publications focused on high-level features that are supposed to better characterize musicological properties as described for instance in [14], [16], and [1]. Further relevant publications regarding feature design are referenced in Section 6.1.

While most research has been conducted using western popular music, only a few works were related to more diverse global music content. A study on the applicability of different classifiers for automatic genre classification of traditional Malaysian music was conducted in [7]. The general issue of multi-label annotations has been addressed only in a few publications so far. In [13], the authors experimented with SVM-based "binary relevance" multi-label genre classification in conjunction with MARSYAS-based features [21]. This approach was continued in [23], where the authors modified a k-Nearest Neighbors classifier in order to handle multi-label data directly. In [20], automatic mood estimation was modeled as a multi-label classification task where every item may belong to more than one class. To the current knowledge of the authors, no publication so far discussed an approach similar to multi-domain labeling, that will be explained in detail in the following section.

3. MULTI-DOMAIN-LABELING

As explained in Section 1, while dealing with musical content from various regional music genres (often referred to as "World music"), the problem frequently arises that songs cannot solely be labeled with one single genre label. Instead, various rhythmic, melodic and harmonic influences conflate into multi-layered mixtures. Common classifier approaches fail because of their immanent assumption that for all song segments, one dominant genre exists and thus is retrievable.

To overcome these problems, we introduce a novel approach called "multi-domain labeling". We aim at breaking down multi-label annotations towards single-label annotations within different musical domains, namely *timbre*, *rhythm*, and *melody / harmony* that are well-known aspects of perceivable music similarity. Furthermore, a separate annotation of each temporal segment of the overall song is enabled. This leads to a more meaningful and realistic two-dimensional description of multi-layered musical content. In addition, the approach facilitates a more precise training of a classifier by avoiding fuzzy multi-labeled data samples.

Figure 1. Structure of the database

4. DATABASE & ANNOTATIONS

The music collection that we used for our investigations consists of 430 full-length tracks from the 16 world music genres. For each genre, the database includes approximately two hours of music on average (see Fig. 1 for details). This music data collection was provided by the content partner of the research project *GlobalMusic2One* [1]. The research project involves educated musicologists working with a world music label and being in regular contact with musicians associated with the applied genres. Annotations were manually made by using an annotation software allowing to label music genres in different domains with regard to an arbitrary amount of time segments. This annotation software includes automatic segmentation algorithm, which makes the first suggestion in order to speed up the annotation process. The experts had a fully freedom to modify borders and assessments of the segments in each of domains.

In this paper, we applied a flat taxonomy with all aforementioned genres considered to be situated at the same hierarchical level. Above all, for our experiments we selected tracks that have been annotated with multi-labels

[1] `http://www.globalmusic2one.net`

in at least one time segment. To evaluate our new annotation approach, the data set was annotated following the principles of multi-domain-labeling as described in Sec. 3. Music experts were allowed to assign up to 4 different genre concepts for each segment - a global genre, a timbre-related genre, a rhythm-related genre, and a genre related to the melodic and harmonic content. The three domain specific annotations were not mandatory. If there were multiple genre influences audible in a single segment, the experts were only allowed to assign one genre label for each domain. This proceeding ensured single-label annotations within each segment and domain. One observation that we made was that these domain-specific genre influences seem to be stable for each segment. The resulting label cardinality (average number of labels per track) of multi-labeled songs per genre was between 1.1 and 2.0 for the selected genres, with 1.0 being a genre that has never been assigned in conjunction with another genre. The label cardinality appeared to be different depending on the music genres.

5. THREE EVALUATION EXPERIMENTS

To evaluate the improvement of the classifier performance, we perform three different experiments as depicted in Fig. 2(a) - 2(c). Therefore, we are moving stepwise from the fuzzy case of multi-labeled songs towards single-labeled segments within different musical domains as described in the previous section.

Multi-labeling (Exp.1)
In the first experiment, all multi-labeled songs are generally used to train multiple classifiers, more precisely all classifier related to the annotated genres.

Multi-labeling with time segmentation (Exp.2)
Bearing the temporal structure of music in mind, we furthermore consider single segments in the second experiment. Multi-labeled segments are repeatedly used as class instances according to their assigned genre labels.

Multi-domain-labeling with time segmentation (Exp.3)
In the third experiment, we are using temporal segments to train three different domain-related classifiers. Therefore, we restricted ourselves to features that can be semantically assigned towards the particular musical domain, as will be detailed in 6.1.

(a) Experiment 1: Multi-labeling

(b) Experiment 2: Multi-labeling with time segmentation

(c) Experiment 3: Multi-domain labeling (**T**imbre, **R**hythm, **M**elody/**H**armony)

Figure 2. Evaluation experiments

6. SYSTEM WORK-FLOW

6.1 Feature extraction

For the experiments conducted in this paper, we utilize a broad palette of features commonly reported in the literature (see Sec. 2). Besides low-level acoustic features, several mid-level representations [4] are extracted. These measures are computed from excerpts of approximately 5 seconds duration by deriving specialized descriptive measures (including musical knowledge) from the observed evolution of low-level features. Besides indifferent usage of all features (in Exp. 1 and Exp. 2), groups of features are assigned to the aforementioned domains in the following manner.

Timbre
In addition to common features, such as Mel-Frequency Cepstral Coefficients (MFCC), Audio Spectrum Centroid (ASC), Spectral Crest Factor (SCF) or Spectral Flatness Measurement (SFM), modulation spectral features [2] have proved to be extremely useful to capture short term dynamics of the low-level features. We applied a cepstral low-pass filtering to the modulation coefficients to reduce their dimensionality and decorrelate them as described in [6].

Rhythm
All rhythmic features used in the current setup are derived from excerpts of the different bands of the Audio Spectrum Envelope (ASE) feature. Part of the measures, such as the Percussiveness [22] and the Envelope Cross-Correlation, are based on the envelope signals. The other part is derived from the Auto Correlation Function (ACF) domain. Besides the measures described in [6], the log-lag ACF and its descriptive statistics are extracted according to [10].

Tonality
Tonality descriptors are computed from a Chromagram based on Enhanced Pitch Class Profiles (EPCP) [12], [19]. The EPCP undergoes a statistical tuning estimation and correction to account for tunings deviating from the equal tempered scale. Most important, the so-called symmetry model, a pitch-space representations as described in [9] are derived from the Chromagram as mid-level features. The model provides an analytic description of aspects of musical consonance and dissonance, as well as functional relationships between probable notes.

6.2 Dimension Reduction

MIR systems usually use a multitude of low-level and mid-level acoustic features. Each feature is designed to correlate with one of the aspects of perceptual similarity, e.g. timbre, tempo, loudness or harmony. The distinct acoustical features are joined together into so called acoustical feature vectors. While temporal changes in one feature often correspond to temporal changes in the other feature (for instance, timbre is changing along with loudness), the individual dimensions of the feature vectors can often be strongly correlated and cause information redundancy. These raw feature vectors could cause various problems on classification stage. One of the usual ways to suppress redundant information in the feature matrix is to utilize dimension reduction techniques. Their purpose is to decrease the feature dimension N while keeping or even revealing the most characteristic data properties. Generally, all dimension reduction methods can be divided into supervised and unsupervised ones. Among the unsupervised approaches the one most often used is *Principal Component Analysis* (PCA). The key idea of PCA [8] is to find a subspace whose basis vectors correspond to the maximum-variance directions in the original feature space. Dimension reduction is obtained then by simply discarding those column vectors with the smallest eigenvalues.

6.3 Classification

In this section we shortly describe the applied classifier and bring the architecture details regarding all three experiments.

Gaussian Mixture Models

Gaussian Mixture Models (GMM) is a commonly used generative classifier. Single data samples of the class are thought of as generated from various sources and each source is modeled by a single multivariate Gaussian. The probability density function (PDF) of the feature frames is estimated as a weighted sum of the multivariate normal distributions. Each single i-th mixture is characterized by its mean vector μ_i and covariance matrix Σ_i. Thus, a GMM is parametrized in $\Theta = \{\omega_i, \mu_i, \Sigma_i\}$, $i = \overline{1, M}$, where ω_i is the weight of the i-th mixtures and $\sum_i \omega_i = 1$. The generalization properties of the model can be adjusted by choosing the number of Gaussian mixtures M. The parameters of the GMM can be estimated using the Expectation-Maximization algorithm [5].

Classifier architecture for three experiments

On the classification stage for each data frame the likelihoods of all class models are calculated. We do not use prior distribution information. The classification decision is therefore made using maximum likelihood rule. In a case of Exp. 1 and Exp. 2 the same data samples may belong to multiple data classes. To tackle the problem, here the classification task is reduced to a set a binary classification decisions, where every binary classifier H_c is trained to make a binary decision (if the data sample belong to a class c or not). These decisions of binary classifiers are joined together to form the multi-label classification. In a case of Exp. 3 as described above only single labels are used within each domain and time segment. Thus for each domain we train one GMM classifier. On the classification stage firstly each domain is classified and post-processed (see Sec. 6.4 for details) independently, and later the results for all domains are joint together.

6.4 Post-processing

Classification with GMM results in class decision for each frame of the feature vector. Thus we apply the following post-processing procedure to reduce frame-level classification to the full-track multi-labels. For Exp. 1 and Exp. 2 the procedure is identical. For all frames of the track for each of the genres we sum up the number of frames associated to these genres. Then be build the normalized histogram of these data. The maximum of this histogram is pointing out the most probable genre for this track. As we are expecting more then single label per track, probably, we also have to accept the second maximum of the normalized histogram. This decision is made by a simple thresholding of the normalized histogram. The track is considered to be associated to those genres, where the values of the normalized histogram are above the threshold. As the histogram is normalized, the threshold is set to $(0, \ldots, 1)$. The choice of the threshold crucially influences the performance of the system. For instance, too low threshold causes high recall values, but might lead to poor precision. Thus the threshold values have to be optimized for each of the experiments. In a case of Exp. 3 we first perform the thresholding for each of domains independently, and then joint the results.

6.5 Evaluation Measures

In multi-label classification each data sample (in our case each song or song segment) is associated with a set of labels $Y \subseteq L$, where L is a full set of labels. Let D be a multi-label dataset, consisting of $|D|$ multi-label examples (X_i, Y_i), $i = 1 \ldots |D|$, $Y_i \subseteq L$, where X_i is a feature matrix of the data example i and Y_i is a set of (ground-truth) labels associated to the data example i. The label cardinality of D is defined as follows:

$$LC(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} |Y_i| \, . \tag{1}$$

Given the multi-label classifier H, the estimated set of labels for sample i is $Z_i = H(X_i)$. The traditional information retrieval evaluation measures for multi-label case are written as:

$$Precision(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap Z_i|}{|Z_i|}, \tag{2}$$

$$Recall(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \cap Z_i|}{|Y_i|}, \tag{3}$$

$$F\_measure(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{2 * |Y_i \cap Z_i|}{|Y_i| + |Z_i|}. \tag{4}$$

7. RESULTS

First of all we detail the settings of the system. The feature extraction procedure results in 233 dimensions for timbre features, 768 dimension for rhythmic features, and 187 dimensions for tonal features. All in all it leads to 1188 dimensions of the feature vector. It is well known that GMMs are sensitive to the curse of dimensionality. As the available annotated database is relatively small, we applied PCA to reduce the dimensionality of feature vectors within each of domains to 100 dimensions. The PCA algorithm has been trained on the randomly chosen training set (70% of the database) and then applied to the test set (30% of the database). This PCA-transformed data have been used in all 3 experiments. The GMMs have been trained with 1, 5, 20, and 50 mixtures, only diagonal covariance matrices have been used. The threshold for the post-processing (as described in Sec. 6.4) has been varied within a range of 0 and 1 for Exp. 1, Exp. 2, and for each of three domains in Exp. 3. Figure 3 depicts the dependency of the F-measure on the thresholding for all above mentioned cases. It is interesting to note, that for Exp. 1 and Exp. 2 achieved F-measure significantly differs depending on the amount of mixtures in the GMM, while in all domains for Exp. 3 the values of F-measure become comparable. Using 5 mixtures results into highest F-measure values for all experiments. The optimal thresholds values are within a range of 0.15 and 0.25.

Within Exp. 3 we found out, that the optimal thresholds for each of domains separately do not form the optimal combination of the thresholds leading to the best F-measure performance when the domains are joined together. Thus, in a case of GMM1 (using only one gaussian to model the class) the optimal thresholds in all domains are found within a range of 0.20 and 0.25, while in a case of GMM20 the optimal thresholds lies within a range of 0.30 and 0.35. Figure 4 depicts the F-measure performance for all three experiments. The F-measure values for each number of mixtures in GMM are increased for Exp. 3 in comparison to Exp. 1 and Exp. 2. The best performance is achived in Exp. 3 for GMM with 5 mixtures reaching the F-measure of 0.61. The significant performance raise of about 10% is observed for the case of using only one gaussian to model the class information. It can be explain with a fact, that in a case of Exp. 3 the classes are less overlapped and easier to model then in a case of the set of binary classifiers (as in Exp. 1 and Exp. 2).

Note that for Exp. 3 the involved GMMs include about two times less free parameters than in a case of Exp. 1 and Exp. 2. As we used only diagonal covariance matrices, the number of free parameters for each GMM can be approximated as $m \cdot (2d + 1)$, where m is a number of mixtures and d is the dimensionality of the feature vector. Thus for Exp. 3 the number of all free parameters comprises $3 \cdot k \cdot m\,(2d' + 1)$, where k is a number of classes and d' is the features dimensionality within one domain; GMMs are trained within each of three domains. Whilst in a case of Exp. 1 and Exp. 2 the amount of free parameters for the set of binary classifiers reaches $2k \cdot m\,(2 \cdot 3d' + 1)$.

Figure 3. Dependency of F-measure on the thresholding while post-processing as described in Sec. 6.4. For Exp. 1 and Exp. 2 the F-measure performance strongly depends on the number if mixtures in GMM.

8. CONCLUSIONS & FUTURE WORK

The paper presented a novel two dimensional approach to music genre classification. It allows to decompose the multi-label classification problem into multiple single-class classification problems by breaking it down in two dimensions. First results demonstrate high potential of the proposed approach. Future work will be directed towards applying Support Vector Machines as alternative classification technique, as it has been proved to perform better than GMM for binary classification. In a case of multi-domain classification we shall make use of supervised feature selection and feature space transformation methods, which can not be utilized in a case of multi-label classification. Furthermore, in the context of the research project *GlobalMusic2One*, we are going to use *vocals* and *instrumentation* as additional domains. We believe the presented approach to be extensible to other music genres as the semantic partitioning of music into different musical domains is universal for most of the world's regional music styles.

Figure 4. F-measures for all three experiments

9. ACKNOWLEDGMENTS

This work is part of the German research project *GlobalMusic2One* funded by the German Ministry of Education and Research (BMBF-FKZ: 01/S08039).

10. REFERENCES

[1] J. Abeßer, C. Dittmar, and H. Großmann. Automatic genre and artist classification by analyzing improvised solo parts from musical recordings. In *Proc. of the Audio Mostly Conference (AMC)*, Piteå, Sweden, 2008.

[2] L. Atlas and S. A. Shamma. Joint acoustic and modulation frequency. *EURASIP Journal on Applied Signal Processing*, 2003:668–675, 2003.

[3] J.-J. Aucouturier and F. Pachet. Representing musical genre: A state of the art. *Journal of New Music Research*, 32:83–93, 2003.

[4] J. P. Bello and J. Pickens. A robust mid-level representation for harmonic content in music signals. In *Proc. of Intl. Conf. on Music Information Retrieval (ISMIR)*, London, UK, 2005.

[5] A.P. Dempster, N. M. Laird, and D. B. Rdin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[6] C. Dittmar, C. Bastuck, and M. Gruhne. Novel mid-level audio features for music similarity. In *Proc. of the Intl. Conf. on Music Communication Science (ICOMCS)*, Sydney, Australia, 2007.

[7] S. Doraisamy, S. Golzari, N. M. Norowi, M. N. B. Sulaiman, and N. I. Udzir. A study on feature selection and classification techniques for automatic genre classification of traditional Malay music. In *Proc. of Intl. Conf. on Music Information Retrieval (ISMIR)*, Philadelphia, Pennsylvania, USA, 2008.

[8] K. Fukunaga. *Introduction to Statistical Pattern Recognition, Second Edition (Computer Science and Scientific Computing Series)*. Academic Press, September 1990.

[9] G. Gatzsche, M. Mehnert, D. Gatzsche, and K. Brandenburg. A symmetry based approach for musical tonality analysis. In *Proc. of Intl. Conf. on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007.

[10] M. Gruhne and C. Dittmar. Improving rhythmic pattern features based on logarithmic preprocessing. In *Proc. of the 126th AES Convention, Munich, Germany*, 2009.

[11] C.-H. Lee, J.-L. Shih, K.-M. Yu, and J.-M. Su. Automatic music genre classification using modulation spectral contrast feature. In *Proc. of the IEEE Intl. Conf. on Multimedia and Expo (ICME)*, 2007.

[12] K. Lee. Automatic chord recognition from audio using enhanced pitch class profile. In *Proc. Intl. Computer Music Conf. (ICMC), New Orleans, USA*, 2006.

[13] T. Li and M. Ogihara. Detecting emotion in music. *Proceedings of the Fifth International Symposium on Music Information Retrieval*, pages 239–240, 2003.

[14] C. McKay and I. Fujinaga. Automatic genre classification using large high-level musical feature sets. In *Proc. of Intl. Conf. on Music Information Retrieval (ISMIR)*, 2004.

[15] I. Panagakis, E. Benetos, and C. Kotropoulos. Music genre classification: A multilinear approach. In *Proc. of Intl. Conf. on Music Information Retrieval (ISMIR)*, 2008.

[16] C. Pérez-Sancho, P. J. Ponce de León, and J. M. Iñesta. A comparison of statistical approaches to symbolic genre recognition. In *Proc. of the Intl. Computer Music Conf. (ICMC)*, pages 545–550, 2006.

[17] N. Scaringella, G. Zoia, and D. Mlynek. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine*, 23:133–141, 2006.

[18] C. N. Silla, C. A. A. Kaestnerlso, and A. L. Koerich. Automatic music genre classification using ensemble of classifiers. In *Proc. of IEEE Intl. Conf. on Systems, Man, and Cybernetics*, pages 1687–1692, October 2007.

[19] M. Stein, B. M. Schubert, M. Gruhne, G. Gatzsche, and M. Mehnert. Evaluation and comparison of audio chroma feature extraction methods. In *Proc. of the 126th AES Convention, Munich, Germany*, 2009.

[20] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas. Multilabel classification of music into emotions. In *Proc. of Intl. Conf. on Music Information Retrieval (ISMIR)*, Philadelphia, Pennsylvania, USA, 2008.

[21] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[22] C. Uhle, C. Dittmar, and T. Sporer. Extraction of drum tracks from polyphonic music using independent subspace analysis. In *Proc. of the 4th Intl. Symposium on Independent Component Analysis*, 2003.

[23] A. Wieczorkowska, P. Synak, and Z. W. Ras. *Multi-Label Classification of Emotions in Music*. Springer Berlin Heidelberg, 2006.