Music Mood Classification: From a Music Piece to a Computed Mood

MUS-15

Mark Bukowski

July 9, 2015

1 Introduction

Nowadays, music is a prevalent topic in society: anyone listens to music anywhere and anytime [23]. Therefore, it is reasonable that there is the emerging and multidisciplinary research field of *Music Information Retrieval (MIR)* which deals with a wide range of general tasks from music perception, feature extraction and classification to music creation [27, 4].
Due to the "exploding amount of digital music content" and its accessibility to the general public, there is a need of developing solutions to retrieve and manage music of interest [27, 4, 1]. For this "automatic classification techniques play an essential role" [23], which include several tasks like mood classification, genre classification and composer identification [11]. These tasks are also part of the *Music Information Retrieval Evaluation eXchange (MIREX)* which provides a framework to compare MIR algorithms [3].
One important characteristic of music is that it "is created to convey emotions" [27] and express feelings [2, 12]. This leads to the natural process "to categorize music in terms of its emotional associations" [9] and to the need of retrieving emotion or mood information from music pieces [12].
Emotion and mood are used interchangeably in literature [9]. In psychology a mood is a more generalized form of emotional feelings for a longer period of time [9]. Recently, mood tags are an emerging meta-data type and are used for searching specific music pieces [26, 27, 8]. Mood is more subjective and challenging to quantify compared to the traditional terms like genre and artist [26].
How to compute a mood of a music piece is covered by the research field of music mood classification which still is in its early stages [9] and has to deal with multidisciplinary and human related issues [27]. To understand the wide range of this research field, this report gives a compact insight of the state of the art and an exemplary approach to answer the question: How to compute a mood from a music piece?

2 State of the Art

Music mood classification includes several subprocesses which are required to calculate a mood from a music piece [12]. At first an emotion model respectively the representation of music moods, which can be perceived by humans, has to be defined. Additionally, specific music features such as audio-based rhythmic features have to be chosen as characteristics of music pieces.
The overall aim of the classification is to train a classifier by calculating the relation between the terms of the emotion model and the extracted music features of music pieces [27].
For this calculation mostly supervised machine learning techniques are used [12]. To train a classifier with a machine learning technique, a ground truth dataset is needed: a training set with mood labeled music pieces accordingly to the emotion model. If the classifier is trained it can predict the mood classes of unseen music data.

2.1 Emotion Models

The correct representation of moods from music pieces perceived and understood by humans are still an active topic in psychology [9]. This is because of the subjectivity of music perception [4, 18]. There exist two types of emotion models: categorical and dimensional.

A categorical emotion model groups moods in different classes which can be represented by several adjective terms [2]. *Hevner's adjective checklist* [5] was the first model containing eight clusters with overall 67 emotional terms which is still used in modified versions [2, 25]. Another widely used categorical emotion model are the five MIREX clusters [2] derived from *All Music Guide (AMG)*[1] [9].

Mostly two to six mood categories are used [8]. The categorical models are easy to use for machine learning techniques [27], but if there are too less or too much mood classes or terms to represent the perceivable richness of music, there is the problem of oversimplification [2, 27] respectively ambiguity [27].

The dimensional emotion model represents moods in a psychological dimensional real-value space [2, 27, 9]. The most used dimensional emotion model is the *Circumplex Model* by Russel [26, 20], which considers the two dimensions: arousal (level of intensity) and valence (level of pleasantness) [2, 27].

For example, in this 2D-model the emotions *afraid* and *angry* have nearly the same values (high arousal and relatively low valence value), but are from psychological point of view very different [2, 27]. This shows the problem that some important psychological distinctions are blurred [27], wherefore some approaches use further dimensions like *dominance* [27] which can also lead to problems in terms of the visualization [27].

Various approaches with different and individual emotion models [6, 25] exist. This lack of standardization leads to the problem of incomparability. So, there are a lot of challenges which have to be dealt with during the selection of a suitable emotion model.

[1] All Music Guide: `http://www.allmusic.com`

2.2 Music Features

Music features are the representations of specific perceptual acoustical elements of a music piece [24].

These audio-based features can be grouped among others into rhythm, pitch, harmony, timbre and temporal features [4]. There is a huge amount of different calculations and representations [24]. For example, timbre can be represented by the common used *Mel-frequency Cepstrum Coefficient (MFCC)* or *Spectral Centroid (SC)* [26, 4]. *Marsyas* [24] is a common used framework to extract music features and provides in total 124 different feature extractions [8, 14, 24]. For more details on music features and musicology see recent literature [2, 26].

Emotions are a "complex subjective and conscious experience reflected by complicated psychophysiological expressions" [2]. It is reasonable that there is a gap between the extractable music features and "the human cognitive level of emotion perception" [27]. This gap is called the semantic gap and has to be considered for the classification [27]. Several studies investigate the relevance of specific features for the mood classification [2, 27, 4, 8], whereby rhythm features are the most popular ones [4, 8].

Audio-based music features are the most used features in music mood classification [8], but there are other types and sources of information such as semantically rich lyrics, genre tags or images from album covers which are used to bridge the semantic gap (see Chapter 3). Because "any classification system is only as good as the features that it receives," [14] the selection of suitable music features is very important. The recent best music mood classification systems use a combination of different types of features [9].

Before the feature extraction and analysis of the music pieces, the datasets have to be preprocessed. In the recent research, the music data is converted to the standard format of 22,050 Hz sampling frequency, 16 bit precision, mono channel, normalized sound level and a representative segment extraction of 30 seconds [27, 12].

2.3 Ground Truth

In order to train a classifier, a dataset with assigned mood labels from the emotion model is required. This so-called ground truth is difficult to obtain [4] due to two main issues: emotional perception and emotion annotation [27]. The perception of music pieces and their moods is subjective and influenced by several factors like individual taste, cultural background, age, gender, etc. [27, 4, 9]. The annotation or labeling of moods to music pieces is influenced by the subjectivity problem: humans have an accuracy of 80 % in annotating the correct moods [21, 4].
The emotion annotation is labor-intensive because it demands a heavy cognitive load of the subjects [27], and time-consuming with one minute to annotate a song on average [12]. These problems therefore mostly lead to small datasets with a varying quality in practice [27, 12, 6]. The lack of publicly available ground truth data negatively effects the individual annotation of private datasets by different researchers [27, 4].
Music pieces are conventionally annotated by less than five musical experts which leads to small datasets in practice [3, 27, 9].
Social Tagging is a recent trend to overcome this problem by using mood tags annotated by users on music recommendation websites such as *Last.fm*[2], but additionally inducing a quality weakness [26, 10, 9].
Annotation games try to make the annotation task more playful. *MajorMiner*[3] is a game for categorical emotion models and *MoodSwings*[4] for 2D-emotion models [12, 9].
Obtaining the ground truth is an essential but still challenging problem [9].

2.4 Supervised Machine Learning

With the help of supervised machine learning, a classifier is trained with the ground truth data to provide a mapping from the music feature space to mood labels of the emotion model [4] in order to predict mood labels for unseen data. Often already existing classification algorithms of other research fields are used (e.g. *Data Mining*) [22, 27].
The recent most popular and best performing mood classification model is the *Support Vector Machine (SVM)* [27, 8, 6]. Its main idea is finding the optimal hyperplane with a maximum margin to separate grouped features [4, 19]. *LIBMSVM* is the most used library to apply the SVM algorithm [27].
There are also other common classifiers such as *K-Nearest-Neighbor (KNN)*, *Artificial Neural Network (ANN)* or *Gaussian Mixture Model (GMM)* [2, 4, 8, 6]. For example, for the real-valued dimensional emotion models a regression model is needed, whereby the *Support Vector Regression (SVR)* is used as a modification of SVM [2, 27].

2.5 Types of Labeling

The music mood classification covers different types of labeling: single-/multi-class and single-/multi-modal labeling.
Single-class labeling assigns one mood label as a representative to a music piece, whereby multi-class labeling assigns more labels. There is an increasing interest in using multi-class labeling due to the subjectivity of emotion perception and individual preferences [18]. Because of a "disagreement regarding perception and interpretation of emotions of song or ambiguity within piece itself" [9] the so-called *Music Emotion Variation Detection (MEVD)* exploits temporal information and tracks the changing moods within one music piece [27, 11, 18].
Modal labeling is about the different types of features used for classification. Most music mood classification approaches are using single-modal labeling with only audio-based features [26, 8, 1] which recently leads to an upper limit of performance [9] (the accuracy is bounded by circa 66 % [27]). Other types like lyrics are investigated for single-modal labeling, but they perform not effectively as audio-based features [6, 1].

[2]Last.fm: `http://www.lastfm.de`
[3]MajorMiner: `http://majorminer.org/info/intro`
[4]MoodSwings: `http://moodswings.ece.drexel.edu`

Recent studies show that multi-modal labeling systems have a general improvement in accuracy [26, 27, 8, 6, 9]. Using additionally semantically rich feature types like lyrics, genre or images seem to be a recent trend of the field of music mood classification [26, 27, 9, 7].
For the recent state of the art in music mood classification, there is a variety of multidisciplinary approaches which use different combinations of the types of labeling, also different emotion models, music features, ground truth data and classification techniques [26, 2, 27, 11, 9, 8]. This individuality leads to new unique solutions but also to a lack of standardization and comparability.

3 Exemplary Approach

In order to compactly explain the whole process of computing moods from a music piece and the related issues one exemplary approach is selected: *Exploiting Online Music Tags for Music Emotion Classification* by Lin et al. from 2011 [12].

3.1 Idea

The main idea of this exemplary approach is to exploit genre tags to bridge the semantic gap in order to improve the performance of music mood classification. The assumption is that the genre of a music piece is closely related to the mood: if music pieces have the same genre they will contain the same performance techniques perceived by humans, and will express the same moods. For example, a sad emotion can be expressed by different performance techniques: *Rock* is loud and rough, however, *Country* is more smooth and tender. This reasons the assumption of a genre-specific characteristic [9].
This approach assigns several mood labels to one music piece (multi-class labeling) and uses genre tags and audio-based features (multi-modal labeling). It has a two-layer structure: first grouping the music pieces by their genre, then building classifiers for each group.

3.2 Methods

This exemplary approach uses all 183 mood classes respectively labels from AMG as its categorical emotion model and AMG's six different genre tags: *Blues*, *Country*, *Jazz*, *R&B*, *Rap* and *Rock*.
It uses all 124 audio-based features which are provided by Marsyas: 68 timbral, 48 pitch and eight rhythmic features [24].
In order to obtain the ground truth, the mood and genre labels, which are only assigned at album-level at AMG, are propagated to all music pieces. The audio files are web crawled from *YouTube.com*[5] (7,922 music pieces out of 1,036 albums). Finally, all music features are extracted by using Marsyas.
The music pieces are separated into the genre groups and to each group supervised machine learning is applied. For each mood, a binary SVM classifier is calculated to predict if the specific mood will be expressed by the music piece. Therefore, 183 binary SVM classifiers composed to one classifier in total are computed for each genre group.
Lin et al. figure out that there is an imbalance in the mood classes of the online taxonomy: many mood classes are underrepresented in the ground truth data (about 75 % of all mood classes are in less than 10 % of the music pieces). As this imbalance would negatively affect the SVM classification, there is a need of data sampling before building the classifiers. A common approach is undersampling: for each mood class only a subset of the majority set (music pieces without the specific mood) with the size of the minority set (music pieces with the specific mood) is sampled. To reduce the risk of discarding potentially useful data, the classification process is adapted by using ensemble learning: each classifier is trained with different undersampled data and finally the best performing set and classifier is chosen.
The moods of an unseen music piece can then be predicted by web crawling the genre tag, extracting the music features and finally applying it to the genre-specific classifier. See the system

[5] YouTube.com: `https://www.youtube.com`

overview in Fig. 1 in Appendix A.
Furthermore, Lin et al. investigated the genre-specific characteristic of music pieces by computing pairwise similarities according to the music features.

3.3 Results and Review

A genre-specific characteristic is shown: the similarity within each genre is higher than the average similarity between all music pieces. So, the genre-grouping seems reasonable to bridge the semantic gap and can also be exploited in future approaches.
The usage of the two-layer structure improves the overall performance of the mood classification about 50 %: average f-score from 0.23 to 0.36. The f-score is the harmonic average of precision and recall. Precision measures the proportion of truly relevant mood tags among all the predicted ones. Recall measures the proportion of truly relevant mood tags in the ground truth that are correctly detected. So, the f-score measures the quality of mood tags assigned to a music piece. With a 36 % precision and recall on average, the accuracy seems unsatisfying but outperforms the random guess (f-score: 0.12). Yet, there is no comparison of the accuracy possible, because there are no similar approaches performed under the same conditions. Only an overall improvement can be achieved by using multi-modal labeling, which also supports other recent research results, however, no statement can be made about the real-world sufficiency.
The propagation of mood labels from the album-level to the music pieces is the biggest point of criticism. This assumes that all music pieces of an album express the same moods, which is not necessarily granted. This false assumption probably leads to a low quality of the ground truth and influences the music mood classification.
Another point of criticism is that all features provided by Marsyas are taken without considering recent research about specific features with higher impact [4, 8].
Finally, this is an exemplary approach out of music mood classification which can compute the moods from a music piece, however, with the lack of comparability due to the individual emotion model, features, ground truth and the application of the SVM classification. Major take aways of this approach are the investigations of genre specificity, the data sampling procedure to overcome mood class imbalance and the general improvement of using multi-modal labeling.

4 Conclusion & Outlook

Nowadays, there are many application areas like music recommendation systems where it is needed to compute the expressed mood from a music piece. Within the research field of music mood classification, there exists a variety of solutions. Best performing systems use knowledge based on multiple domains and a combination of different music feature types [9]. Within the MIREX community there seems to be an accuracy boundary at about 68 % also for multi-modal approaches [17, 16]. This accuracy is quite satisfactory relative to the human accuracy of 80 % [21] considering that this field of research is quite young and has to deal with open issues regarding its multidisciplinarity and human related nature. There are attempts such as MIREX or particular systems like the *Autonomous Classification Engine (ACE)* to provide a standardized framework [15, 13, 14], but in general the recent variety of approaches has a lack of standardization and comparability [27, 4, 9, 18, 25, 14].
Based on these issues, there are still open challenges and future work for music mood classification. To overcome the problem of individual subjectivity, one possibility is to use personalized learning systems with individual profiling [2, 9]. A future-looking approach could also consider the environment of music listening such as waking up, eating or driving [27], or take into account physiological changes (e.g. heart rate) [2].
Finally, music mood classification shows a great potential and a basis for further multidisciplinary research to compute a mood from a music piece.

References

[1] T.-T. Dang and K. Shirai. Machine learning approaches for mood classification of songs toward music search engine. *2009 International Conference on Knowledge and Systems Engineering*, pages 144–149, 2009.

[2] J. J. Deng, C. H. C. Leung, A. Milani, and L. Chen. Emotional states associated with music. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(1):1–36, 2015.

[3] J. S. Downie, X. Hu, J. H. Lee, K. Choi, S. J. Cunningham, and Y. Hao. Ten years of MIREX (music information retrieval evaluation exchange): Reflections, challenges and opportunities. In *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, pages 657–662, 2014.

[4] Z. Fu, G. Lu, K. M. Ting, and D. Zhang. A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, 13(2):303–319, 2011.

[5] K. Hevner. Experimental studies of the elements of expression in music. *The American Journal of Psychology*, 48(2):246–268, 1936.

[6] X. Hu. *Improving Music Mood Classification using Lyrics, Audio and Social Tags*. PhD thesis, University of Illinois at Urbana-Champaign, 2010.

[7] X. Hu and J. S. Downie. Exploring mood metadata: Relationships with genre, artist and usage metadata. In *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 67–72, Vienna, Austria, September 23-27 2007. `http://ismir2007.ismir.net/proceedings/ISMIR2007_p067_hu.pdf`.

[8] X. Hu and J. S. Downie. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, pages 159–168, New York, NY, USA, 2010. ACM.

[9] Y. E. Kim, E. M. Schmidt, R. Migneco, O. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull. Music emotion recognition: A state of the art review. In *11th International Society for Music Information and Retrieval Conference*, pages 255–266, 2010.

[10] J. H. Lee and X. Hu. Generating ground truth for music mood classification using mechanical turk. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '12, pages 129–138, New York, NY, USA, 2012. ACM.

[11] T. Li, M. Ogihara, and G. Tzanetakis. *Music Data Mining (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series)*. CRC Press, 2011.

[12] Y.-C. Lin, Y.-H. Yang, and H. H. Chen. Exploiting online music tags for music emotion classification. *ACM Trans. Multimedia Comput. Commun. Appl.*, 7S(1):26:1–26:16, 2011.

[13] C. McKay, J. A. Burgoyne, J. Thompson, and I. Fujinaga. Using ACE XML 2.0 to store and share feature, instance and class data for musical classification. In *Proceedings of the 10th International Society for Music Information Retrieval Conference, ISMIR 2009, Kobe International Conference Center, Kobe, Japan, October 26-30, 2009*, pages 303–308, 2009.

[14] C. Mckay, R. Fiebrink, D. Mcennis, B. Li, and I. Fujinaga. Ace: A framework for optimizing music classification. In *Proceedings of the International Conference on Music Information Retrieval*, 2005.

[15] C. McKay and I. Fujinaga. Improving automatic music classification performance by extracting features from different types of data. *Proceedings of the international conference on Multimedia information retrieval - MIR '10*, pages 257–266, 2010.

[16] Music Information Retrieval Evaluation eXchange (MIREX). MIREX 2013 results, 2013. `http://www.music-ir.org/mirex/wiki/2013:MIREX2013_Results` (Accessed on 2015-07-08).

[17] Music Information Retrieval Evaluation eXchange (MIREX). MIREX 2014 results, 2014. `http://www.music-ir.org/mirex/wiki/2014:MIREX2014_Results` (Accessed on 2015-07-08).

[18] E. E. P. Myint and M. Pwint. An approach for mulit-label music mood classification. *2010 2nd International Conference on Signal Processing Systems*, pages V1–290–V1–294, 2010.

[19] S. Rho, B.-j. Han, and E. Hwang. Svr-based music mood classification and context-based music recommendation. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, pages 713–716, New York, NY, USA, 2009. ACM.

[20] J. A. Russel. A circumplex model of affect. *Journal of Personality and Social Psychology*, 93(6):1161–1178, 1980.

[21] P. Saari, T. Eerola, and O. Lartillot. Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. *IEEE Trans. Audio Speech Lang. Process.*, 19(6):1802–1812, 2011.

[22] J. Sander and M. Ester. Data mining algorithms: Chapter 5. *Lecture Notes, RWTH Aachen University, Data Management And Exploration Group, Prof. Dr. rer. nat. Thomas Seidl*, 2013.

[23] M. Schedl, E. Gmez, and J. Urbano. Music information retrieval: Recent developments and applications. *FNT in Information Retrieval*, 8(2-3):127–261, 2014.

[24] G. Tzanetakis and P. Cook. Marsyas: a framework for audio analysis. *Organised Sound*, 4(3):169–175, 2000.

[25] B. Van De Laar. Emotion detection in music, a survey. *4th Twente Student Conference on IT*, 1, 2006.

[26] H. Xue, L. Xue, and F. Su. Multimodal music mood classification by fusion of audio and lyrics. *Lecture Notes in Computer Science*, pages 26–37, 2015.

[27] Y.-H. Yang and H. H. Chen. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):40:1–40:30, 2012.

A Appendix

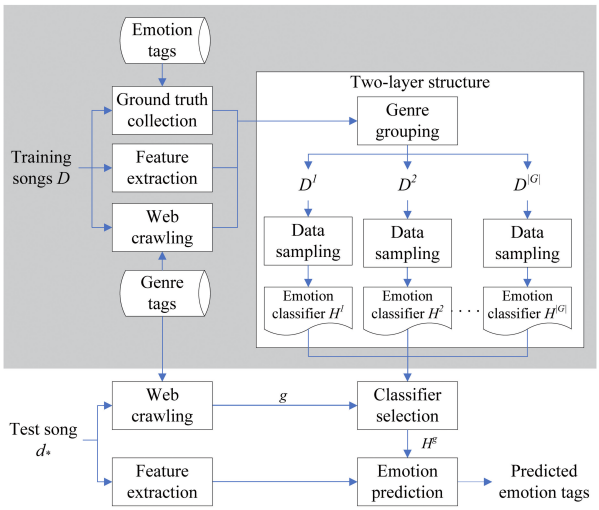

Figure 1: System Overview of the Exemplary Approach [12]