BOOSTING FOR MULTI-MODAL MUSIC EMOTION CLASSIFICATION

Qi Lu, Xiaoou Chen, Deshun Yang, Jun Wang
Peking University
Institute of Computer Science & Technology
{luqi,chenxiaoou,yangdeshun,wangjun}@icst.pku.edu.cn

ABSTRACT

With the explosive growth of music recordings, automatic classification of music emotion becomes one of the hot spots on research and engineering. Typical music emotion classification (MEC) approaches apply machine learning methods to train a classifier based on audio features. In addition to audio features, the MIDI and lyrics features of music also contain useful semantic information for predicting the emotion of music. In this paper we apply AdaBoost algorithm to integrate MIDI, audio and lyrics information and propose a two-layer classifying strategy called Fusion by Subtask Merging for 4-class music emotion classification. We evaluate each modality respectively using SVM, and then combine any two of the three modalities, using AdaBoost algorithm (MIDI+audio, MIDI+lyrics, audio+lyrics). Moreover, integrating this in a multimodal system (MIDI+audio+lyrics) allows an improvement in the overall performance. The experimental results show that MIDI, audio and lyrics information are complementary, and can be combined to improve a classification system.

Key Words: Music Emotion Classification, Multi-Modal, AdaBoost, Fusion by Subtask Merging

1. INTRODUCTION AND RELATED WORKS

Music Information Retrieval is a sub-area of information retrieval. Important research directions include for example similarity retrieval, musical genre classification, or music analysis and knowledge representation. As the music databases grow, classification and retrieval of music by emotion [2]-[7] has recently received increasing attention.

Traditionally music emotion classification (MEC) applies algorithms of machine learning on audio features, such as Mel frequency cepstral coefficient (MFCC), to recognize the emotion embedded in the audio signal. Meanwhile we can also use some mid-level audio features such as chord [5] or rhythmic patterns [8] for this problem, but sometimes it can't get a promising result because of the semantic gap.

Complementary to audio features, lyrics are semantically rich and expressive and have profound impact on human perception of music [17]. It is often easy for us to tell from the lyrics whether a song expresses love, sadness, happiness, or something else. Incorporating lyrics in the analysis of music emotion is feasible because most popular songs sold in the market come with lyrics and because most lyrics are composed in accordance with music signal [18].

Besides music's audio and lyrics features, the MIDI features of music have been ever used in music instrument classification and retrieval. As a popular file format for storing music, MIDI carries more abstract music information than audio. In this paper we firstly apply the music's MIDI file to the music emotion classification.

A multi-modal analysis approach using audio and lyrics features has been proposed and evaluated in music genre classification by Mayer and Neumayer [1]. And promising results have been achieved by combining the audio and lyrics using various types of machine learning algorithms such as SVM and k-NN. Besides, several multi-modal fusion methods using audio and lyrics for music emotion classification are proposed by Yang [2]. However, little has been reported in the literature that applies

AdaBoost to multi-modal automatic music emotion classification. In this paper, we propose a new multi-modal fusing approach that uses features extracted from MIDI files, audio signal and lyrics for 4-class music emotion classification. We focus on how to combine the three modalities: MIDI, audio and lyrics using AdaBoost.

The remainder of the paper is organized as follows. Section 2 describes the MIDI, audio and lyrics features we need respectively. Section 3 describes the details of the proposed multi-modal approach. Section 4 provides the result of a performance study, and Section 5 concludes the paper.

2. FEATURES

In our experiment we use a free program jMIR1.0 with default parameter values to extract MIDI and audio features. jAudio and jSymbolic are two important components of jMIR for extracting audio and MIDI features. jAudio is a software package for extracting features from audio files. These extracted features can then be used in many areas of music information retrieval (MIR) research. jSymbolic is a software package for extracting high-level musical features from symbolic music representations, specifically MIDI files.

2.1 MIDI Features

The MIDI music files are firstly transformed from the corresponding waveform files by a computer tool WIDI Recognition System Professional 4.1 which could be found on the internet [19]. And then we use jSymbolic with default parameter values to extract MIDI features from the MIDI files. The extracted MIDI features, which are listed in **Table 1**, are adopted in our experiments.

#	Feature	Dimensions
1	Duration	1
2	Acoustic Guitar Fraction	1
3	Average Melodic Interval	1
……		
101	Voice Separation	1
102	Woodwinds Fraction	1

Table 1. MIDI features extracted by jSymbolic.

From **Table 1** we can see there are 102 features extracted by jSmbolic from each MIDI music file. As each feature just has one dimension, a whole MIDI feature vector has 102 dimensions.

2.2 Audio Features

We use jAudio to extract a number of low-level audio features from the waveform files. The extracted features, which are listed in **Table 2**, have been commonly used for MEC in pervious works [3]-[5].

#	Feature	Dimensions
1	Magnitude Spectrum	Variable
2	FFT Bin Frequency Labels	Variable
3	Spectral Centroid	1
……		
25	Zero Crossings	1
26	Beat Sum	1

Table 2. Audio features extracted by jAudio.

From **Table 2** we can see there are 26 features extracted by jAudio from each audio file. Among these 26 features, there are 5 features such as Magnitude Spectrum and MFCC with variable dimensions, other ones with 1 dimension. In our experiment, an audio feature vector has 79 dimensions.

2.3 Lyrics Features

Lyrics are normally available on the web and downloadable with a simple crawler. The acquired lyrics are preprocessed with traditional information retrieval operations such as stopword removal, stemming, and tokenization. In our experiment, two algorithms are adopted to generate textual features.

Uni-gram A standard textual feature representation counts the occurrence of uni-gram terms (words) in each document, and constructs the bag-of-words model [10], which represents a document as a vector of terms weighted by a tf-idf function defined as:

$$tfidf(t_i, d_j) = \#(t_i, d_j) \log \frac{|D|}{\#D(t_i)} \quad (1)$$

where $\#(t_i, d_j)$ denotes the frequency of term t_i oc-

curs in document d_j, $\#D(t_i)$ the number of documents in which t_i occurs, and $|D|$ the size of the corpus. We compute the tf-idf for each term and select the M most frequent terms as our features (M is empirically set to 2000 in this work by a validation set).

Bi-gram N-gram is sequences of N consecutive words [10]. An N-gram of size 1 is a uni-gram (single word), size 2 is a bi-gram (word pairs). N-gram models are widely used to model the dependency of words. Since negation terms often reverse the meaning of the words next to them, it seems reasonable to incorporate word pairs to the bag-of-words model to take the effect of negation terms into account. To this end, we select the M most frequent uni-gram and bi-gram in the bag-of-words model and obtain a new feature representation.

3. TAXONOMY

We adopt Thayer's arousal-valence emotion plane [15] as our taxonomy and define four emotion classes happy, angry, sad, and relaxing, according to the four quadrants of the emotion plane, as shown in **Figure 1**. As arousal (how exciting/calming) and valence (how positive/ negative) are the two basic emotion dimensions found to be most important and universal [16], we can also view the four-class emotion classification problem as the classification of high/low arousal and positive/negative valence. This view will be used in mutli-modal music emotion classification.

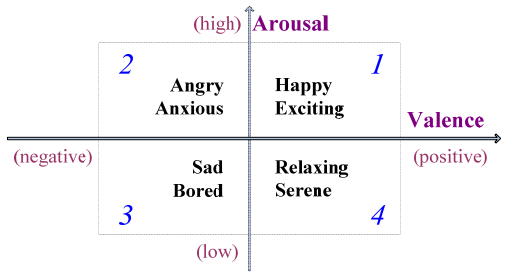

Figure 1. Thayer's arousal-valence emotion plane. We define four emotion classes according to the four quadrants of the emotion plane. We can also subdivide the four-class emotion classification to binary arousal classification and valence classification.

4. PROPOSED APPROACH

In this paper, we use AdaBoost, an ensemble learning algorithm, to train a classifier by integrating MIDI, audio and lyrics features. Boosting is a method to combine a collection of weak classification functions (weak learner) to form a stronger classifier [21]. AdaBoost is an adaptive algorithm to boost a sequence of classifiers, in that the weights are updated dynamically according to the errors in previous learning [22].

Tieu and Viola [12] adapted AdaBoost algorithm for natural image retrieval. They made the weak learner work in a single feature each time. So after T rounds of boosting, T features are selected together with the T weak classifiers. We adapted AdaBoost algorithm of Tieu and Viola's version for music emotion classification and retrieval. In each iteration, we made the weak learner work on each modality independently. So we can get three classifiers which are trained according to MIDI, audio and lyrics features respectively each time. And then we select the classifier of the minimum learning error as the representative of this iteration. After T rounds of boosting, T weak classifiers are produced in the end.

The classic AdaBoost algorithm is only used for binary classification. In a 4-class scenario, we propose a two-layer classifying strategy called Fusion by Subtask Merging.

•**Fusion by Subtask Merging (FSM):** Use AdaBoost to classify arousal and valence separately and then merge the result. To enhance readability, we denote the classification model trained by AdaBoost for classifying arousal and valence as M_A and M_V, respectively. For example, a negative arousal (predicted by M_A) and negative valence (predicted by M_V) would be merged to class 3. We make the three modalities focus on different emotion classification subtasks because empirical test reveals MIDI, audio and text clues are complementary and useful for different subtasks. In addition, training models for arousal and valence separately has been shown adequate.

4.1 AdaBoost

The AdaBoost algorithm We adapted in our experiment as follows:

Input: 1) n training examples

$(x_1, y_1), \ldots, (x_n, y_n)$ with $y_i = 1$ or 0;

2) the number of iterations T.

Initialize weights $w_{1,i} = \frac{1}{2l}$ or $\frac{1}{2m}$ for $y_i = 1$ or 0, with $l + m = n$.

Do for $t = 1, \ldots, T$ && $\varepsilon_t \leq 0.5$:

1. Train one hypothesis h_j for each modality j with w_t, and error $\varepsilon_j = \sum_{i=1}^{n} (h_j(x_i) - y_i) * w_{t,i}$.

2. Choose $h_t(\cdot) = h_k(\cdot)$ such that $\forall j \neq k, \varepsilon_k \prec \varepsilon_j$. Let $\varepsilon_t = \varepsilon_k$.

3. Update: $w_{t+1,i} = w_{t,i} \beta_t^{e_i}$, where $e_i = 1$ or 0 for example x_i classified correctly or incorrectly respectively, and $\beta_t = \frac{\varepsilon_t}{1 - \varepsilon_t}$.

4. Normalize the weights so that they are a distribution, $w_{t+1,i} \leftarrow \frac{w_{t+1,i}}{\sum_{j=1}^{n} w_{t+1,j}}$.

Output the final hypothesis,

$$h_f(x) = \begin{cases} 1 \ \ if \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 \qquad\qquad otherwise \end{cases} \quad (2)$$

where $\alpha_t = \log \frac{1}{\beta_t}$.

4.2 Support Vector Machine

Support vector machine (SVM) learns an optimal separating hyperplane (OSH) given a set of positive and negative examples. Kernel functions are used for SVM to learn a non-linear boundary if necessary. See Vapnik [14] for a detailed introduction of SVM. Li and Guo [13] tried to use the SVM for audio classification and retrieval. In this paper, SVM is selected as our weak learner. In our experiment we use the SMO which is a fast implementation of SVM algorithm provided by WEKA3.6.1 [20].

5. EXPERIMENTS

The music database is made up of 500 Chinese pop songs, whose emotions are labeled through a subjective test conducted by 8 participants. The corresponding lyrics are downloaded from the Internet by a web crawler. Classification accuracy is evaluated by randomly selecting 400 songs as training data and 100 songs as testing data. We conducted 2 experiments. To assure the confidence, we performed the experiments based on a five-fold cross validation. We use the features extracted by jSymbolic for MIDI feature representation, the features extracted by jAudio for audio feature representation and the uni-gram and bi-gram based bag-of-words model for lyrics feature representation.

5.1 Single Feature Sets

In our first experiment, we apply SVM to mono-modal based music emotion 4-class classification (MEC) using MIDI, audio and lyrics information respectively. Therefore, we got three SVM classifiers which are trained on each mono-modality. Our SVM implementation is the SMO algorithm provided by WEKA3.6.1 and the kernel function is Polynomial. To enhance readability, we denote the classification model trained by MIDI, audio and textual features as **MO**, **AO** and **LO** respectively.

- **MIDI-Only (MO):** Use MIDI features only and apply SVM to classify emotion. This serves as a baseline. MO is used to assess the importance of the MIDI modality.
- **Audio-Only (AO):** Use audio features only and apply SVM to classify emotion. This serves as a baseline because many existing MEC work adopts it [1-2]. AO is used to assess the importance of the audio modality.
- **Lyrics-Only (LO):** Use lyrics features only and apply SVM to classify emotion. This serves as a baseline because many existing MEC work adopts it [1-2]. LO is used to assess the importance of the text modality.

The Results of experiment 1 are shown in **Table 3**:

Classifier Name	Features	Accuracy(4-class)
MO	MIDI	0.586
AO	audio	0.598
LO	lyrics	0.491

Table 3. Results of mono-modal method using SVM for 4-class emotion classification.

5.2 Multi-Modal Feature Set Combinations

In our second experiment, we apply AdaBoost to multi-modal based music emotion classification. And we select SVM as the weak learner in AdaBoost. We develop and evaluate the following method for fusing MIDI, audio and lyrics. To enhance readability, we denote the classification model trained by MIDI and audio features set, MIDI and lyrics features set, audio and lyrics features set, MIDI, audio and lyrics features set as **MA**, **ML**, **AL** and **MAL** respectively.

• **MIDI+Audio (MA):** Use MIDI and audio features and apply AdaBoost to classify emotion. The weak learner is SVM.

• **MIDI+Lyrics (ML):** Use MIDI and lyrics features and apply AdaBoost to classify emotion. The weak learner is SVM.

• **Audio+Lyrics (AL):** Use audio and lyrics features and apply AdaBoost to classify emotion. The weak learner is SVM.

• **MIDI+Audio+Lyrics (MAL):** Use MIDI, audio and lyrics features and apply AdaBoost to classify emotion. The weak learner is SVM.

The Results of experiment 2 are shown in **Table 4**:

Classifier Name	Features	Accuracy(4-class)
MA	MIDI+audio	0.616
ML	MIDI+lyrics	0.712
AL	audio+lyrics	0.72
MAL	MIDI+audio+lyrics	0.724

Table 4. Results of multi-modal fusion method using AdaBoost for 4-class emotion classification.

4.3 Comparison and Analysis of Experimental Results

Because of the different database, it is difficult to quantitatively compare the proposed approach with existing ones. Alternatively, we treat MO, AO and LO as the three baselines, and compare the classification accuracy of mono-modal and multi-modal approaches.

It can be observed from row 2 to 4 of **Table 3** that MIDI features, audio features and textual features performs very poor on 4-class emotion classification, with MO's accuracy 58.6%, AO's accuracy 59.8%, LO's accuracy 49.1%. But from row 2 to 4 of **Table 4**, we can see MIDI features, audio features and lyrics features are fairly complementary, because the combination of any two of them outperforms the mono-modal approach, with MA's accuracy 61.6%, ML's accuracy 71.2%, AL's accuracy 72.0%. **Table 4** also indicates that the 4-class emotion classification accuracy can be significantly improved by fusing all the three modalities. Among the fusion methods (rows 2-5 of **Table 4**), MAL achieves the best classification accuracy (72.4%) and contributes a 23.3% relative improvement over the lyrics-only (LO) baseline. This seems to imply the individual strength of the three modalities should be emphasized separately.

6. CONCLUSION

In this paper we have described a preliminary multi-modal approach to music emotion classification that exploits features extracted from the MIDI, audio and the lyrics of a song. We apply AdaBoost algorithm to ensemble the three modalities. A new approach of multi-modal fusion method called Fusion by Subtask Merging (FSM) is developed and evaluated. Experiments on a moderately large-scale database show that MIDI, audio and lyrics indeed carry semantic information complementary to each other. By the proposed fusion by subtask merging strategy, we can improve the classification accuracy from 49.1% to 72.4%. Using lyrics features also significantly improves the accuracy of valence classification from 61.6% to 72.4%. Meanwhile, we find that MIDI and audio features contribute fairly to the music emotion classification. From the result, we can see that the accuracy of MO is 58.6%, while that of AO is 59.8%. Besides, the accuracy of ML is 71.2%, while that of AL is 72.0%. An explanation for this phenomenon is that there exists some redundancy between MIDI and audio information. As well, an exploration of more natural language processing algorithms and more effective features for modeling the characteristics of lyrics is underway. Besides, we're trying to verifying more ensemble learning algorithms on multi-modal music emotion classification.

7. REFERENCES

[1] Rudolf Mayer et al: "Multi-modal Analysis of Music: A large-scale Evaluation," *In Proceedings of the Workshop on Exploring Musical Information Spaces* , 2009.

[2] Yang, Y.-H. et al: "Toward multi-modal music emotion classification," *Proc. PCM*, pp. 70-79, 2008.

[3] Yang, Y.-H. et al: "A regression approach to music emotion recognition," *IEEE Trans. Audio, Speech and Language Processing*, Vol. 16, No. 2, pp. 448-457, 2008.

[4] Lu, L. et al: "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio, Speech and Language Processing*, Vol. 14, No. 1, pp. 5-18, 2006.

[5] Cheng, H.-T. et al: "Automatic chord recognition for music classification and retrieval," *Proc. ICME*, pp. 1505-1508, 2008.

[6] Yang, D. et al: "Disambiguating music emotion using software agents," *Proc. ISMIR*, pp. 52-58, 2004.

[7] Chuang, Z.-J. et al: "Emotion recognition using audio features and textual contents," *Proc. ICME*, pp. 53-56, 2004.

[8] Chua, B.-Y. et al: "Perceptual rhythm determination of music signal for emotion-based classification," *Proc. MMM*, pp. 4-11, 2006.

[9] Yo-Ping Huang and Guan-Long Guo et al: "Using Back Propagation Model to Design a MIDI Music Classification System," *Int. Computer Symposium*, Dec. 15-17, 2004, Taipei, Taiwan.

[10] Sebastiani F.: "Machine learning in automated text categorization," *ACM CSUR*, Vol. 34, No. 1, pp. 1-47, 2002.

[11] G. Guo, H. Zhang, and S. Z. Li: "Boosting for content-based audio classification and retrieval: an evaluation," *Microsoft Research Tech. Rep.* MSR-TR-2001-15.

[12] K.. Tieu and P. Viola: "Boosting image retrieval," in *Proc. of Computer Vision and Pattern Recognition*, v. 1, pp. 228-235, 2000.

[13] S. Z. Li and G. Guo: "Content-based audio classification and retrieval using svm learning," *(invited talk), PCM*, 2000.

[14] V. N. Vapnik: "Statistical learning theory," John Wiley& Sons, New York, 1998.

[15] Thayer, R. E. et al: "The Biopsychology of Mood and Arousal", Oxford University Press, New York, 1989.

[16] Russel, A.: "A circumplex model of affect", *Journal of Personality & Social Science*, Vol. 39, No. 6, pp. 1161-1178, 1980.

[17] Omar Ali, S. et al: "Songs and emotions: are lyrics and melodies equal partners", *Psychology of Music*, Vol. 34, No. 4, pp. 511-534, 2006.

[18] Formas, J.: "The words of music", *Popular Music and Society*, Vol. 26, No. 1, 2003.

[19] http://www.widisoft.com/english/download.html

[20] http://www.cs.waikato.ac.nz/ml/weka/

[21] Yoav Freund and Robert E. Schapire: "A short introduction to boosting," Journal of Japanese Society for Artificial Intelligence, Vol. 14, No. 3, pp. 771–780, 1999.

[22] Y. Freund and R. E. Schapire: "A decision-theoretic generalization of online learning and an application to boosting," *Journal* of Computer and *System* Sciences, Vol. 55, No. 1, pp. 119-139, 1997.

[23] C. Cortes and V. Vapnik: "Support-Vector Networks," Machine Learning, Vol. 20, No. 3, pp. 273-297, 1995.