Music Emotion Regression based on Multi-modal Features[1]

Di Guan[1] , Xiaoou Chen[2] and Deshun Yang[3],
[1,2,3] Peking University
Institute of Computer Science & Technology
guandiyiyi417@gmail.com
{chenxiaoou, yangdeshun}@pku.edu.cn

Abstract. Music emotion regression is considered more appropriate than classification for music emotion retrieval, since it resolves some of the ambiguities of emotion classes. In this paper, we propose an AdaBoost-based approach for music emotion regression, in which emotion is represented in PAD model and multi-modal features are employed, including audio, MIDI and lyric features. We first demonstrate the effectiveness of our approach, and then focus on exploring the contribution of individual modalities to the regression of each emotion dimension. A series of experiments show that lyric contributes the most to the regression of emotion dimension P, while audio and MIDI contribute more to the regression of dimension A and D. Thinking that the three modalities provide complementary information from different angles, we combine them and show that the best regression performance is obtained when all modalities are used.

Keywords: Music emotion regression, Multi-modal, AdaBoost, PAD.

1 Introduction and Related Works

It is natural for us to organize and search music by emotional contents. Music emotion retrieval has gained increasing attention in the field of music information retrieval during the past few years [1].

Music emotion classification, in which the emotion space is modeled by a given number of classes, is a plausible approach to music emotion retrieval, but the emotional states within each class may vary a lot, and this ambiguity may confuse users when they retrieve music according to emotion. However in music emotion regression, the emotion space is viewed as continuous and each point in the space is considered as a distinctive emotional state [2]. In this way, the ambiguity associated with emotion classes can be successfully avoided, so music emotion regression is considered more appropriate for music emotion retrieval [3]. A regression approach is proposed for music emotion recognition in [3], the best performance evaluated in terms of the R^2 statistics reaches 58.3% for *arousal* and 28.1% for *valence*.

[1] Project supported by the Natural Science Foundation of China (Multi-modal Music Emotion Recognition technology research No.61170167) & Beijing Natural Science Foundation (Multimodal Chinese song emotion recognition)

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

In our work, PAD(Pleasure-Arousal-Dominance) emotion-state model is used to represent music emotion [4].Three nearly independent dimensions, P(pleasure), A(arousal) and D(dominance), are used to represent emotional states in PAD model. P distinguishes the positive-negative quality of emotional states, A refers to the intensity of physical activity and mental alertness, and D is defined in terms of control versus lack of control. In our work, we normalize all dimensions in the range of -4 to 4, in this way each emotional state corresponds to a specific point in PAD model. For example, "anger" corresponds to (-3.51, 2.59, 0.95), which indicates that it is a highly unpleasant, highly aroused, and moderately dominant emotional state.

Audio features have been commonly used in music emotion recognition and audio-based techniques could achieve promising results [5]. As a complementary source, lyric contains rich semantic information of songs and more emotionally relevant information which is not included in audio [6].Additionally, MIDI is used in symbolic music information retrieval [7]. Some previous works applied multi-modal features for music emotion recognition and achieved promising performance [8,9,10].

In this paper, we present an AdaBoost approach for music emotion regression where three-modality features, audio, MIDI and lyric, are employed. We firstly demonstrate the effectiveness of our regression approach by comparing it with several baseline regression algorithms, and secondly use each modality alone to explore the contribution of each modality to the regression of different emotion dimension, and thirdly combine the three modalities to demonstrate the performance improvement of multi-modal feature combination for music emotion regression, lastly use a feature selection technique to reduce feature dimensions and computational complexity.

The paper is organized as follows. Section 2 describes the features and feature processing, Section 3 describes the regression approach, Section 4 provides the analysis of experiment results, and the last Section makes the conclusion and prospect.

2 Dataset and Feature Processing

2.1 Dataset

We download 2500 Chinese songs from network, including audio, MIDI and lyric data for each song, which cover more than 900 singers and more than 1000 albums, and include different genres such as pop, rap and rock. Then 11 volunteers whose ages are from 22 to 50 use Self Assessment Manikins(SAM) [11] to annotate the songs with PAD values ranging from -4 to 4. When a song is annotated by more than 8 volunteers and the emotion values given by different annotators are consistent (all positive or all negative), the song will get a mean value as its emotion label and be retained into our dataset. In this way, the final music dataset includes 1687 songs.

2.2 Features

Audio Features. We extract audio features from wave files of the dataset by jAudio [12], which is a system to extract the basic features from audio signal. We set the window size to 512ms (the signal sampling rate to 22KHz) to extract audio features,

including one-dimension (e.g. RMS) and multi-dimension vectors (e.g. MFCC's). 27 kinds of audio features that have been commonly used in MIR are extracted to compose an audio feature vector of 112 dimensions for a song. Table 1 shows part of the audio features.

MIDI Features. We extract MIDI features from MIDI files of the dataset by jSymbolic [13], which is a feature extraction system for extracting high-level musical features from symbolic music representations, specifically from MIDI files. Unlike audio data, MIDI data contains the information reflecting music concepts directly. 102 kinds of MIDI features are extracted to compose a MIDI feature vector of 1022 dimensions for a song. Table 2 shows part of the MIDI features.

Table 1. The partial list of audiofeatures.

Audio features	
Feature	**Dimensions**
MFCC's	13
LPC	10
Spectral Rolloff	1
Spectral Flux	1
RMS	1
Compactness	1
Zero Crossings	1
…	…
Power Spectrum	variable
All	**112**

Table 2. The partial list of MIDI features.

MIDI features	
Feature	**Dimensions**
Basic Pitch Histogram	128
Beat Histogram	161
Melodic Interval Histogram	128
Pitch Class Distribution	12
Acoustic Guitar Fraction	1
Amount of Arpeggiation	1
Note Density	1
…	…
Duration	1
All	**1022**

Lyric Features. We firstly download the lyrics of all the songs from Internet, and then do some pre-processing to them with traditional NLP tools, including stop-words filtering and word segmentation etc. Finally Unigram, Bigram and Trigram features are extracted from the lyrics.

Unigram. Unigram refers to the sequences of single word appeared in documents.

Bigram. Bigram refers to a distinctive term containing 2 consecutive words appeared in documents. Because negation words often reverse emotion of the words next to them, it seems reasonable to incorporate word-pairs to take effect of negation words into account in emotion analysis.

Trigram. Trigram refers to a distinctive term containing 3 consecutive words appeared in documents. Because bigrams only reflect parts of useful multi-word patterns for emotion expression, we take trigrams into account additionally.

Finally, in order to reduce the lyric feature space, we select the 3000 most frequently appeared N-grams (*n=1, 2, 3*) as lyric features. In our work, the feature vector of a lyric can be expressed as (v_1, v_2, v_3, ... v_{3000}). Here $v_i \in \{0, 1\}$: if N-gram *i* appeared in the lyric, v_i=1; otherwise v_i=0.

2.3 Feature Processing

In our work, seven different feature sets are employed for the regression of emotion dimension P, A and D, including the set of audio features(A), the set of MIDI

features(M), the set of lyric features(L), the set of audio and MIDI features(A+M), the set of audio and lyric features(A+L), the set of MIDI and lyric features(M+L), and the set of audio, MIDI and lyric features(A+M+L). A simple concatenation scheme is employed to combine the multi-modal features. For example, a concatenated feature vector of the three modalities can be expressed as (A_1,A_2,...A_x , M_1,M_2,...M_y , L_1,L_2,...L_z). Where A_1~A_x are audio features, and $x=112$; M_1~M_y are MIDI features, and $y=1022$; L_1~L_z are lyric features, and $z=3000$. The number of dimensions of a concatenated feature vector is $x+y+z=4134$.

The space formed by the raw concatenated features has a huge number of dimensions. To reduce the computational complexity of learning and regression, increase the efficiency and generalization capability of the regression model, we do feature selection on each of the original feature sets, to find a subset of the original set which could maximize the performance of regression model. Correlation-based Feature Subset Selection [14] with BestFirst as its search method is employed in our work, which evaluates the worth of a feature subset by considering the individual predictive ability of each feature along with the degree of redundancy between them, subsets of features that are highly correlated with the class while having low inter-correlation are preferred [15].

In feature selection process, we have found that some features are effective to all the 3 emotion dimensions, such as LPC, beat histogram, basic pitch histogram, melodic interval histogram, etc. But some features only effective to some of the 3 dimensions, such as staccato incidence, spectral rolloff point, etc, which only effective to dimension A and D. After feature processing, we get seven final feature sets for emotion dimension P, A and D.Table 3 shows the number of selected features of each feature set.

Table 3. The number of selected features in each feature set.

	P	**A**	**D**
A	17	16	16
M	44	43	54
L	262	410	474
A+M	43	51	54
A+L	349	208	331
M+L	115	67	203
A+M+L	116	69	86

3 Regression Algorithm

AdaBoost is a commonly used boosting method, which works by iteratively running weak learners on different distributions of training data, so as to get an integrated regression model more powerful than weak learners.

We present an AdaBoost regression approach in this paper, which follows most of the steps of AdaBoost.R2 [16] and uses MultiLayerPerceptron(MLP) [17] as the weak learner. MLP is a typical feed forward neural network connecting several perceptrons by a hierarchy, and uses error back propagation to adjust connection weights

continuously. We called our approach AdaBoost.RM(R refers to Regression and M-MultiLayerPerceptron).

Given a set of m training instances: $(x_1, y_1),\ldots,(x_m, y_m)$, where $x_1 \ldots x_m$ are the features, and $y_1 \ldots y_m \in [-4, +4]$ are the P, A or D emotion values of the instances. Initially, we set the weights of the training instances as $D_t(i) = 1/m$, then iteratively running MLP on the instances to train a regression model for P, A, or D and modify the weights of the instances accordingly. We set the number of iterations to 10, because the performance of the algorithm no longer improves when the number of iterations is greater than 10. The instance weight modification method is as follows:

$$\overline{L_t} = \sum_{i=1}^{m} \left(\frac{f_t(x_i) - y_i}{\max_{i=1,2\ldots m}(f_t(x_i) - y_i)} \right)^2 D_t(i) \tag{1}$$

$$D_{t+1}(i) = \frac{D_t(i) \left(\frac{\overline{L_t}}{1 - \overline{L_t}} \right)^{(1 - L_t(i))}}{Z_t} \tag{2}$$

Where f_t is the regression model learned in iteration t, $f_t(x_i)$ is the regression result of x_i from model f_t. $D_t(i)$ is the weight of instance i in iteration t, $\overline{L_t}$ is the average loss of f_t, z_t is a normalization factor that makes $\sum_i D_{t+1}(i) = 1$.

This reweighting procedure makes the poorly predicted instances get higher weights but well predicted ones get lower weights. Finally, an average formula is used to calculate the final regression result instead of the "*INF*" formula of AdaBoost.R2:

$$f_{final}(x) = AVE \left[y \in Y: \sum_{t: f_t(x) \le y} \log \frac{1 - \overline{L_t}}{\overline{L_t}} \ge \frac{1}{2} \sum_t \log \frac{1 - \overline{L_t}}{\overline{L_t}} \right] \tag{3}$$

Where $Y = \{f_1(x), f_2(x), \ldots, f_T(x)\}$, *AVE* is the average function.

4 Experiments and Results Analysis

4.1 Evaluation Criteria

We conduct a series of experiments to evaluate the performance of our regression approach. Different regression algorithms and different feature sets are tried to build a regression model for each emotion dimension(P, A and D), and the performances are measured in terms of correlation coefficient (CF) and R^2 statistics both developed by Karl Pearson. They are defined as follows:

$$CF_{XY} = \frac{\sum_{i=1}^{N} (R(X_i) - \overline{R(X_i)})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N} (R(X_i) - \overline{R(X_i)})^2} \sqrt{\sum_{i=1}^{N} (Y_i - \overline{Y})^2}} \tag{4}$$

$$R^2{}_{XY} = 1 - \frac{\sum_{i=1}^{N} (Y_i - R(X_i))^2}{\sum_{i=1}^{N} (Y_i - \overline{Y})^2} \tag{5}$$

Where Y_i is the emotion label, $R(X_i)$ is the regression value of feature vector X_i.

To ensure the validity of the results, we use 5-fold cross validation to evaluate the performance of regression models. The dataset is randomly broken into five subsets of the same size, with four being used for training and one for testing, and this process is repeated 5 times and finally the mean CF and R^2 value is taken.

4.2 Comparison of AdaBoost.RM with Baseline Algorithms

Because of the different used datasets, it is not reasonable to compare our approach with existing ones. However we compare our approach with three baseline algorithms. The first is LinearRegression [18] which uses linear regression for prediction, the second is SMOreg [19] which implements support vector machine for regression, and the last is original AdaBoost.R2 algorithm. The experiments are based on the three-modality feature set(A+M+L) introduced in Section 2.3, and the results are showed in Table 4.

Table 4. Performance of our approach compared with that of baseline ones.

	P		A		D	
	CF	R^2	**CF**	R^2	**CF**	R^2
LinearRegression	0.688	0.476	0.823	0.693	0.715	0.529
SMOreg	0.692	0.48	0.828	0.696	0.72	0.542
AdaBoost.R2	0.536	0.284	0.778	0.61	0.672	0.435
AdaBoost.RM	**0.702**	**0.488**	**0.843**	**0.708**	**0.755**	**0.558**

Table 4 shows that among all the regression algorithms our approach achieve the best performance for the regression of all the emotion dimensions(P, A and D), this indicates the effectiveness of our approach. It's to be noted that our approach performs better than AdaBoost.R2, which demonstrates the effectiveness of our modification to AdaBoost.R2. On the other hand we can see that all the regression algorithms have achieved promising performance on our three-modality feature set, which indicates that our feature processing technique and the selected features are really effective to music emotion regression.

4.3 Contributions of Different Modality and Effectiveness of Multi-modal Feature Combination

We conduct a series of experiments to explore the contribution of different modality to the regression of each emotion dimension, and demonstrate the effectiveness of multi-modal feature combination.

The seven feature sets introduced in Section 2.3 are employed for the regression of all the emotion dimensions(P, A and D). The results are showed in Table 5.

Table 5. Contributions of individual modality and effectiveness of multi-modal features.

		P		A		D	
#	**Feature Set**	**CF**	R^2	**CF**	R^2	**CF**	R^2
1	**A**	0.473	0.166	0.724	0.516	0.667	0.38
2	**M**	0.541	0.305	0.823	0.685	0.73	0.508
3	**L**	0.623	0.383	0.461	0.202	0.575	0.328
4	**A+M**	0.571	0.285	0.812	0.663	0.719	0.474
5	**A+L**	0.681	0.465	0.745	0.554	0.728	0.53
6	**M+L**	0.68	0.469	0.828	0.681	0.738	0.542
7	**A+M+L**	**0.702**	**0.488**	**0.843**	**0.708**	**0.755**	**0.558**

In Table 5, the 1st to 3rd rows show that:

1. Among the three modalities, lyric has the biggest contribution to the regression of emotion dimension P.
2. To the regression of dimension A and D, audio and MIDI contribute more than lyric, and MIDI has the biggest contribution among the three modalities. This indicates that MIDI features contain more useful information related to emotion dimension A and D compared with audio and lyric features.
3. Audio and lyric are complementary on the regression of dimension P and A, the reasons maybe that audio signal contains a large amount of energy related information such that the extracted audio features reflect emotional intensity more directly, while lyric contains more semantic information so as to express emotion more directly.

The 4th to 7th rows show that:

1. The regression performance has been enhanced on all the emotion dimensions when any two modalities are combined.
2. The best regression performance has been achieved on all the emotion dimensions when all the three modalities are combined. This indicates that the three modalities provide useful and complementary information for music emotion regression, and the greatest improvement of performance can be achieved when all the three modalities are used.

Generally Speaking, audio signal contains a large number of energy relevant information which reflects emotional intensity more directly; MIDI data contains more information which reflects the concept of music more directly; lyric includes more semantic information which describes emotional inclinations more directly. Audio and MIDI have big contribution to emotion dimension A and D, while lyric has big contribution to emotion dimension P. The three modalities provide complementary information for music emotion regression, and the greatest improvement of regression performance can be achieved when all the three modalities are combined.

5 Conclusion and Future Work

In this paper, we present three main parts of our research work on music emotion regression. First we demonstrate the effectiveness of our regression approach, and then we expound the contribution of each modality to the regression of each dimension of PAD model, and last we verify the performance improvement of emotion regression models brought about by the combination of multi-modal features.

There are two focuses in the future, one is to find more informative features for music emotion recognition, and the other is to build a music emotion retrieval system based on our regression model, in which songs can be retrieved by specifying an emotional state.

References

1. Y. Feng, Y. Zhuang, and Y. Pan: Popular Music Retrieval by Detecting Mood. In: Proc. ACM SIGIR, pp. 375--376(2003)
2. Russell and James A: A Circumflex Model of Affect. Journal of Personality and Social Psychology, vol.39, no.6, pp.1161--1178(1980)
3. Yi-Hsuan Yang, Yu-Ching Lin et al: A Regression Approach to Music Emotion Recognition. IEEE Transactions on Audio, Speech, and Language Processing, vol.16, no. 2(2008)
4. Mehrabian, A.: Framework for A Comprehensive Description and Measurement of Motional States. Genetic, Social, and General Psychology Monographs, vol. 121, pp. 33--361(1995)
5. C.Laurier and J.Grivolla and P.Herrera: Multimodal Music Mood Classification using Audio and Lyrics. In: Proceedings of the 7th International Conference on Machine Learning and Applications(2008)
6. Xiao Hu.et al: Lyric Text Mining in Music Mood Classification. In: ISMIR 2009 Conference Proceedings, pp.411--416(2009)
7. Tzanetakis G., Ermolinskyi, A., and Cook, P: Pitch Histograms in Audio and Symbolic Music Information Retrieval. In: ISMIR 2002 Conference Proceedings, pp.31--38, Paris: IRCAM(2002)
8. Yi-Hsuan Yang et al: Toward Multi-modal Music Emotion Classification. In: Proc. PCM, pp. 70--79(2008)
9. Xiao Hu.et al: Improving Mood Classification in Music Digital Libraries by Combining Lyrics and Audio. In: Proc. of the 10th Annual Joint Conference on Digital Libraries, New York, USA(2010)
10. Q.Lu et al: Boosting for Multi-modal Music Emotion Classification. In: ISMIR 2010 Conference Proceedings, pp.105--110(2010)
11. Lang P. J.: Behavioral Treatment and Bio-behavioral Assessment: Computer Applications. In: J.Sidowski, J. Johnson, & T. Williams (Eds.), Technology in Mental Health Care Delivery Systems. pp.119--137. Norwood, NJ: Ablex(1980)
12. McEnnis, D., C. McKay, and I. Fujinaga: jAudio: A Feature Extraction Library. In: Proc. of the International Conference on Music Information Retrieval(2005)
13. McKay, C., and I. Fujinaga: jSymbolic: A Feature Extractor for MIDI Files. In: Proc. of the International Computer Music Conference(2006)
14. M. A. Hall: Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand(1998)
15. Weka: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/.
16. Drucker H.: Improving Regressors using Boosting Techniques. In: Proceedings of the Fourteenth International Conference on Machine Learning, pp. 107--115, Morgan Kaufmann, Burlington, Mass(1997)
17. Rumelhart.D et al: Learning Internal Representations by Error Propagation. Parallel Distributed Processing Explorations in the Microstructure of Cognition, vol.1, MIT(1986)
18. Douglas C. Montgomery et al: Introduction to Linear Regression Analysis. 4th Edition, Wiley(2008)
19. S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, K.R.K. Murthy: Improvements to the SMO Algorithm for SVM Regression. IEEE Transactions on Neural Networks(1999)