MUSIC EMOTION CLASSIFICATION: A REGRESSION APPROACH

Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H. Chen

Graduate Institute of Communication Engineering, National Taiwan University

ABSTRACT

Typical music emotion classification (MEC) approaches categorize emotions and apply pattern recognition methods to train a classifier. However, categorized emotions are too ambiguous for efficient music retrieval. In this paper, we model emotions as continuous variables composed of arousal and valence values (AV values), and formulate MEC as a regression problem. The multiple linear regression, support vector regression, and AdaBoost.RT are adopted to evaluate the prediction accuracy. Since the regression approach is inherently continuous, it is free of the ambiguity problem existing in its categorical counterparts.

1. INTRODUCTION

Music plays an important role in human's history, even more so in the digital age. As the music databases grow, more efficient organization and search methods are needed. Music classification by perceived emotion is a plausible approach, for it is content-centric and functionally powerful.

Though many previous approaches can be found in the literature, music emotion classification (MEC) is still a challenging task. One of the major difficulties lies in the fact that emotions are hard to be described in a universal way. The adjectives used to describe emotions are ambiguous, and the use of adjectives for the same emotion can vary from person to person. A typical solution [1]–[3] is to categorize emotions into a number of emotion classes and apply the standard pattern recognition procedure to train a classifier. The Thayer's emotion plane [4] is commonly adopted to avoid the ambiguity of adjectives. It defines the emotion classes dimensionally in terms of arousal (how exciting/calming) and valence (how positive /negative). For example, the emotion classes can be divided into the four quadrants in the emotion plane as shown in Fig. 1.

However, even with the emotion plane, the categorical taxonomy is still inherently ambiguous. Each emotion class represents an area in the emotion plane, and the emotion states within each area may vary a lot. For example, the first quadrant of the emotion plan contains emotions such as excited, happy, and pleased, which are different in nature. This ambiguity confuses the subjects during the subjective test and confuses the users when retrieving a music piece according to his/her emotion state.

Research supported by a grant from the National Science Council of Taiwan under the contract NSC 95-2752-E-002-006-PAE.

An alternative is to view the emotion plane as a continuous space and recognize each point of the plane as an emotion state. However, one major problem with the continuous approach is that arousal and valence are not necessarily independent and can in fact impact each other. Whether the emotion states should be best modeled as categories or continua has been a subject of debate in psychology, and both perspectives have its pros and cons. Since the continuous approach can resolve the ambiguity problem, we consider it more suitable for MEC.

Specifically, with the continuous approach, we first compute the arousal and valence values (AV values) of each music sample and view the music sample as a point on the emotion plane. Then the user can retrieve music by specifying a point on the emotion plane according to his/her emotion state, and the system would return the closest music pieces. In this way, apparently, the efficiency and accuracy of music retrieval can be much improved.

However, automatic calculation of the AV values is still at its early stage, and the performance of existing methods [5]–[7] is not satisfactory in many aspects (see Section 2). Therefore, the primary purpose of this paper is to develop an effective method for computing the AV values. We show how to formulate the MEC problem as a regression problem and use regression techniques to directly predict the AV values of music samples from extracted features. This approach has sound theoretical basis, assumes no geometric relationship between arousal and valence, and exhibits promising prediction accuracy.

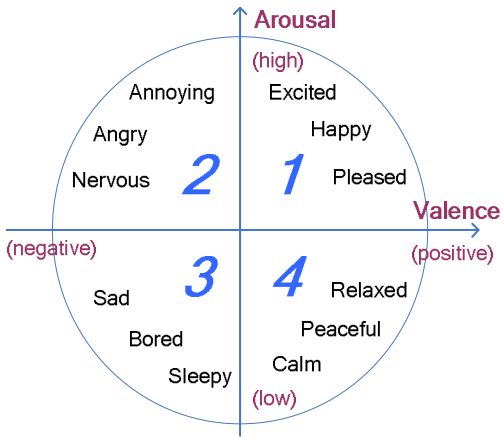

Fig. 1. Thayer's arousal-valence emotion plane.

The paper is organized as following. Section 2 reviews previous work on the computation of AV values. The regression approach and a system overview are described in Sections 3 and 4. Section 5 gives experimental results, Section 6 discussion, and Section 7 conclusion.

2. PREIVOUS WORK

In [5], an AV modeling is proposed to compute the AV values of video sequences. The arousal and valence models are weighted combinations of some component functions, which are computed along the timeline. Three components are used for arousal: the motion vectors between consecutive video frames, the changes in shot lengths of the video, and the energy of sound. For valence, the average pitch of sound is used. Although the AV modeling is based on psychological understandings and the features used are intuitively related to emotion evocation, its theoretical foundation is not mature enough, which makes it difficult to quantitatively evaluate the accuracy of the AV modeling and the effectiveness of the components.

In [6], the emotions are divided into the four quadrants in the emotion plane, and the input samples are assigned with a fuzzy vector indicating the relative strength of each class by fuzzy classifiers. The fuzzy vector can then be transformed to AV values by considering the geometric relationship of the four emotion classes in the emotion plane. However, the transformation involves emotion classes that are not necessarily independent of each other. Since the geometric relationship between arousal and valence is inexact, it is improper to apply arithmetic operations on the arousal and valence.

In [7], a more systematic approach is proposed. Emotion is quantified as a time-varying continuous variable, and the system identification technique is utilized to model the music emotion as a function of extracted features. The dependence between arousal and valence is considered, and the average R^2 statistics of the AV values reaches 78.4% and 21.9%. However, this approach aims to estimate the variation of emotion within a specific song, which may be less intractable than the emotion labeling of each song.

3. THE REGRESSION APPROACH

In light of the above observations, we formulate the music emotion classification as a regression problem and train the regressors (regression analysis models) to predict the AV values directly. In the following, we first describe the adopted regressors and present our regression system next.

Given N inputs (x_i, y_i), $1 \le i \le N$, where x_i is the feature vector of the ith input sample, and $y_i \in \mathbb{R}$ ($\mathbb{R}$ denotes a real space) is the real value to be predicted, the regression analysis aims at training a regressor $R(\cdot)$ such that the prediction error is minimized:

$$R \mid \min \sum_{i=1}^{N} \mid y_i - R(x_i) \mid . \quad (1)$$

In other words, given the feature vector, the regressors is applied to predict the AV values directly, whereas the fuzzy classifiers first compute the fuzzy vector and then transform the fuzzy vector to the AV values.

To formulate the music emotion classification as a regression problem, the following factors are considered:

a) Domain of $\mathbb{R}$: The emotion plane is viewed as a coordinate space spanned by the AV values, with each value confined within [-1, 1] ($\mathbb{R}$ =[-1, 1]).
b) Ground truth: Set via a subjective test by averaging the subjects' opinions about the AV values of each input sample. For reliability, the subjects are educated to learn the essence of emotion model and the purpose of the experiment.
c) Number of regressors: Two regressors are trained to predict the AV values respectively.
d) Training fashion: Although it is arguable whether arousal and valence are independent of each other, we train the two regressors separately (with the same x_i) under the assumption that the correlation between arousal and valence is embedded in the ground truth.
e) Feature extraction: The extracted features should be relevant to emotion evocation. The features used in this work are described in Section 4.
f) Type of regressor: Three regression algorithms are adopted and compared in this work: The multiple linear regression (MLR), support vector regression (SVR) [8], and AdaBoost.RT [9].

MLR is a standard regression algorithm which assumes a linear relationship between variables; the linear relationship is estimated by a least squares estimator. Comparatively, SVR nonlinearly maps input vectors to a higher dimensional feature space by the kernel trick, and yields prediction functions that are expanded on a subset of support vectors. Support vectors have been found in many cases competitive with existing machine learning methods. AdaBoost.RT is another nonlinear regression algorithm: A number of regression trees are trained iteratively and weighted according to the prediction accuracy. After the iterative process, the prediction result of each regression tree is combined (weighted mean) to improve the accuracy of prediction.

Note the regression approach has a sound theoretical foundation—the regression theory. It generates more reliable result and allows quantitative performance analysis. Moreover, the regression approach does not assume any geometric relationship between arousal and valence and learns the predicting rules according to the ground truth, so the problem related to inexact relationship between arousal and valence is of no concern. One can also easily convert the regression results to binary or quaternary ones if categorical taxonomy is required. Most importantly, the regression approach is inherently continuous, so it is free of the ambiguity problem commonly existing in its categorical counterparts.

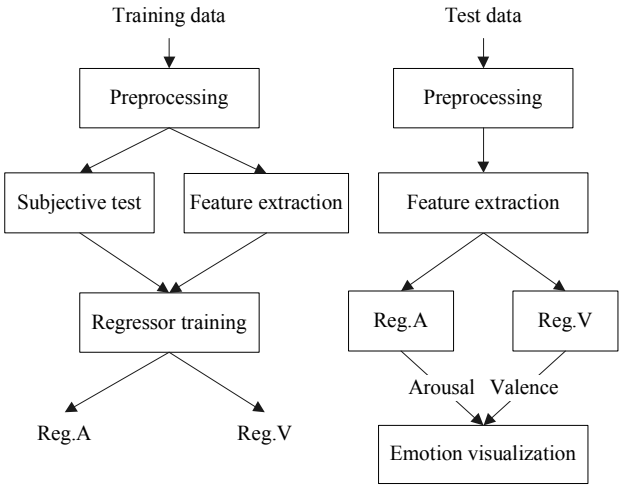

Fig. 2. System diagram of the proposed regression approach. Left: training; right: testing.

4. SYSTEM OVERVIEW

The system diagram of the proposed regression approach is shown in Fig. 2, and the details are described below.

In *preprocessing*, music clips are trimmed to 25 seconds and converted to a uniform format (22,050 Hz, 16 bits, and mono channel PCM WAV). The same music database as described in [6] is used, which contains 195 popular songs from Western, Chinese, and Japanese albums. Popular songs are chosen rather than classical ones because we think it is the pop-music that dominates the everyday music-listening for most people.

After preprocessing, the following algorithms are applied in *feature extraction*: PsySound [10], Marsyas [11], spectral contrast [2], and DWCH [1]. These feature extraction algorithms are listed and described in Table I. A total of 114 features are extracted for each music sample. Besides, since 15 of the PsySound features have been found particular useful for MEC in [6], we run the following experiments using both the whole feature sets (denoted as All) and the 15 PsySound features (Psy15).

The *subjective test* sets the ground truth of the AV values. Subjects (most college students) are asked to listen to a subset of music dataset and to choose two values, each ranges from –1.0 to 1.0 in 11 levels, to indicate their feeling about the AV values of the music sample. The ground truth is set as the mean of the AV values of all subjects tested. On the average, more than ten pairs of AV values are collected from the subjective test for each music sample. We have made the dataset available on our website [12].

Given the 195 inputs (x_i, y_i) from feature extraction and subjective test, the regression algorithms are applied to train the regressors. MLR can be easily implemented using Matlab. The SVR implementation is based on the library LIBSVM [13]. A grid parameter search is also applied to find out the best parameters for SVR.

TABLE I
The feature extraction algorithms adopted in this work

Method	Number of features	Description
PsySound	44	Extracts features including loudness, level, pitch multiplicity, and dissonance based on psychoacoustic models.
Marsyas	30	Extracts timbral texture, rhythmic content and pitch content features.
Spectral contrast	12	Represents the relative characteristics of each spectral sub-band, and reflects the distribution of harmonic components.
DWCH	28	The Daubechies wavelets coefficient histogram which uses histogram moments to estimate the probabilistic distribution.

For AdaBoost.RT [9], the threshold φ for demarcating correct and incorrect predictions are empirically determined as 0.1, and the number of iterations is set to 30. We further introduce a new threshold θ to control the tree pruning process: A tree node cannot be pruned if its depth is lower than θ. The threshold θ is empirically determined as 5.

5. EXPERIMENTAL RESULTS

We evaluate the performance of the proposed regression approach in terms of R^2 statistics, mean absolute error (MAE), and classification accuracy A_c. R^2 statistics is often interpreted as the proportion of response variation explained by the regressors in the model. A negative value of R^2 means the prediction model is worse than simply taking the sample mean. MAE is measured to provide some geometric insight to the prediction accuracy. A_c is measured by treating positive values and negative values as two different classes. The performance of the approach is evaluated by the 10–fold cross validation technique. The sample variance for arousal and valence of our database are 0.118 and 0.063.

Experimental results are shown in Table II, based on which the following observations can be made:

a) The best R^2 statistics, obtained by using SVR, reaches 60% and 19%, respectively, for arousal and valence, similar to that obtained by [7]. The A_c of SVR reaches 84% and 68%, which is as competitive as existing MEC works.
b) Using Psy15 as the feature set not only largely reduces the computational effort but also improves the results without any serious overfitting problem.
c) On the average, arousal is much easier to predict than valence. This finding is consistent with existing MEC systems. To further improve the performance, other emotion models or feature sets may be required; the regression approach can then be applied as well.

Fig. 3 shows the distributions of the AV values for the ground truth (blue crosses) and the predicted AV values by SVR using Psy15 features (red stars).

TABLE II
Experimental results of the regression approach

Regression algorithm	Feature subset	R^2 statistics of arousal (training)	R^2 statistics of valence (training)	R^2 statistics of arousal (test)	R^2 statistics of valence (test)	MAE of arousal (test)	MAE of valence (test)	A_c of arousal (test)	A_c of valence (test)
MLR	All	88.00%	80.91%	-8.49%	-130.03%	0.266	0.2716	76.84%	62.11%
MLR	Psy15	64.88%	27.90%	56.84%	10.94%	0.1805	0.1927	84.74%	67.37%
SVR	All	67.65%	42.88%	56.15%	9.32%	0.1897	0.2132	77.37%	61.58%
SVR	Psy15	62.78%	32.76%	59.80%	19.18%	0.1752	0.1798	84.21%	67.89%
AdaBoost.RT	All	81.14%	56.65%	54.26%	-7.96%	0.1731	0.2026	86.84%	61.58%
AdaBoost.RT	Psy15	77.78%	51.97%	58.10%	10.86%	0.1831	0.1889	83.68%	66.84%

6. DISCUSSION

To maintain clarity, the paper has focused on the ability of the regression approach in solving the ambiguity problem. However, the regression approach can also help reduce subjectivity, which is the other critical issue of MEC.

The subjectivity problem stems from the fact that music perception is intrinsically subjective and is under the influence of many factors such as personality, age, listening mood, and cultural background. Therefore, as discussed in [6], typical categorical approaches that simply assign one emotion class to each song in a deterministic manner will not perform well in practice. Since the regression approach represents each song as a point on the emotion plane and thus offers more freedom in describing the song, the subjectivity problem is reduced by the regression approach.

Being helpful in alleviating both the ambiguity problem and subjectivity problem, the continuous view of emotion, as well as the regression approach, is suggested by the authors to applications involved emotion recognition, such as speech emotion recognition and video affective analysis.

7. CONCLUSION

In this paper, we have presented a preliminary work that formulates music emotion classification as a regression problem and utilizes MLR, SVR, and AdaBoost.RT for direct prediction of the AV values. Since emotions are modeled continuously in the regression approach, the ambiguity problem commonly existing in categorical approaches is successfully avoided.

The R^2 statistics of the AV values reach 60% and 19% by SVR, similar to that obtained by [7]. The classification accuracy reaches 84% and 68%, competitive to existing categorical approaches. The regression model could be further improved with more appropriate emotion models or mid-level features such as the lyrics.

Besides, the regression approach does not assume any geometric relationship between arousal and valence, so the problem related to inexact relationship between arousal and valence is of no concern. With AV values, each song can be visualized as a point on the emotion plane, allowing for more efficient music retrieval and management.

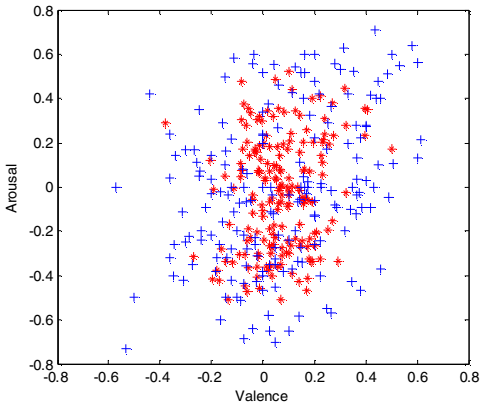

Fig. 3. Distribution of the AV values of the ground truth (blue crosses) and the predicted AV values by SVR+Psy15 (red stars).

8. REFERENCES

[1] T. Li and M. Ogihara, "Content-based music similarity search and emotion detection," *ICASSP*, pp. 17–21, 2004.

[2] L. Lu, D. Liu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 5–18, 2006.

[3] D. Yang and W. Lee, "Disambiguating music emotion using software agents," *ISMIR*, pp. 52–58, 2004.

[4] R. E. Thayer, *The Biopsychology of Mood and Arousal*, 1989.

[5] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Trans. Multimedia*, pp. 143–154, 2005.

[6] Y.-H. Yang, C.-C Liu, and H.-H. Chen, "Music emotion classification: A fuzzy approach," *ACM MM*, pp. 81–84, 2006.

[7] M. D. Korhonen, D. A.Clausi, and M. E. Jernigan, "Modeling emotional content of music using system identification," *IEEE Trans. Sys. Man. and Cyber.*, vol. 36, no. 3, pp. 588–599, 2006.

[8] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, 2004.

[9] D. P. Solomatine and D. L. Shrestha, "AdaBoost.RT: a boosting algorithm for regression problems," *IJCNN*, 2004.

[10] D. Cabrera, "PsySound: A computer program for psychoacoustical analysis," *Austra. Acous. Conf.*, pp. 47–54, 1999.

[11] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech and Audio Processing*, vol. 10, pp. 293–302, 2002.

[12] http://mpac.ee.ntu.edu.tw/~yihsuan/icme07/.

[13] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm.