Affective Feature Extraction for Music Emotion Prediction

Yang Liu
Department of Computer Science
Hong Kong Baptist University
Kowloon Tong, Hong Kong SAR, P. R. China
csygliu@hkbu.edu.hk

Yan Liu
Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Hong Kong SAR, P. R. China
csyliu@comp.polyu.edu.hk

Zhonglei Gu
AAOO Tech Limited
Shatin, Hong Kong SAR, P. R. China
allen.koo@aaoo-tech.com

ABSTRACT

In this paper, we describe the methods designed for extracting the affective features from the given music and predicting the dynamic emotion ratings along the arousal and valence dimensions. The algorithm called Arousal-Valence Similarity Preserving Embedding (AV-SPE) is presented to extract the intrinsic features embedded in music signal that essentially evoke human emotions. A standard support vector regressor is then employed to predict the emotion ratings of the music along the arousal and valence dimensions. The experimental results demonstrate that the performance of the proposed method along the arousal dimension is significantly better than the baseline.

1. INTRODUCTION

The *Emotion in Music* task in MediaEval 2015 Workshop aims to detect the emotional dynamics of music using its content. Specifically, given a set of songs, participants are asked to automatically generate continuous emotional representations in arousal and valence. More details of the task as well as the dataset can be found in [1].

Feature extraction, which aims to discover the intrinsic factors while capturing essentials of original data according to some criteria, plays an important role in music emotion analysis. Some algorithms have been proposed to learn the genuine correlates of music signal evoking emotional responses. You *et al.* presented a multi-label embedded feature selection (MEFS) method for the task of music emotion classification [12]. Liu *et al.* introduced an algorithm called multi-emotion similarity preserving embedding (ME-SPE) by considering the correlation between different music emotions, and then analyzed the relationship between the low-dimensional features and the music emotions [6]. In this paper, we propose a feature extraction algorithm, arousal-valence similarity preserving embedding (AV-SPE), which inherits the basic idea from ME-SPE. The difference is that the emotion labels in ME-SPE are in the binary form, i.e., 0 or 1, while those in AV-SPE could be any real number between $[-1, 1]$.

In order to learn the relationship between the feature space and the dimensional emotion space, which is composed of the arousal dimension and the valence dimension, many popular machine learning approaches have already been employed to train the model, such as k-Nearest Neighbor [11], Support Vector Regression [9], Boosting [7], Conditional Random Fields [2], and Gaussian Process [8]. In this task, we employ the ν-Support Vector Regression (ν-SVR) [10] to predict the arousal and valence labels of the music.

In the remaining part, we first introduce the methods used for feature extraction and label prediction in Section 2. In Section 3, we report the evaluation results. Finally, we conclude the paper in Section 4.

Copyright is held by the author/owner(s).
MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

2. METHOD

2.1 Feature Extraction via Arousal-Valence Similarity Preserving Embedding

In order to discover the intrinsic factors in music signals that convey or evoke emotions along the arousal and valence dimensions, we propose a supervised feature extraction algorithm dubbed AV-SPE to map the original high-dimensional representations into a low-dimensional feature subspace, in which we hope that a clearer linkage between the features and emotions could be discovered.

Let $\mathbf{x} \in \mathbb{R}^D$ be the high-dimensional feature vector of the music at a certain time point (in our specific task, $D = 260$), and $\mathbf{y} = [y^{(1)}, y^{(2)}]$ be the corresponding emotion label vector, where y_1 and y_2 denote the arousal value and valence value, respectively. The idea behind AV-SPE is simple: if two pieces of music can convey similar emotions, they should possess some hidden features in common. Specifically, given the training set $\{(\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_n, \mathbf{y}_n)\}$, AV-SPE aims to learn a transformation matrix $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_d] \in \mathbb{R}^{D \times d}$ which is able to project the original D-dimensional data to an intrinsically low-dimensional subspace $\mathbb{Z} = \mathbb{R}^d$, where the data with similar emotion labels are close to each other. The objective function of AV-SPE is formulated as follows:

$$\begin{aligned} \mathbf{U} &= \arg\min_{\mathbf{U}} J(\mathbf{U}) \\ &= \arg\min_{\mathbf{U}} \sum_{i=1}^{n} \sum_{j=1}^{n} \|\mathbf{U}^T\mathbf{x}_i - \mathbf{U}^T\mathbf{x}_j\|^2 \cdot S_{ij}, \end{aligned} \tag{1}$$

where $S_{ij} = \langle \hat{\mathbf{y}}_i, \hat{\mathbf{y}}_j \rangle = \langle \mathbf{y}_i/||\mathbf{y}_i||, \mathbf{y}_j/||\mathbf{y}_j|| \rangle$ denotes emotional similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$ $(i, j = 1, ..., n)$.

Following some standard operations in linear algebra, above optimization problem could be reduced to a trace minimization problem:

$$\mathbf{U} = \arg\min_{\mathbf{U}} tr(\mathbf{U}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{U}), \tag{2}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \in \mathbb{R}^{D \times n}$ is the data matrix, $\mathbf{L} = \mathbf{D} - \hat{\mathbf{S}}$ is the $n \times n$ Laplacian matrix [3], $\mathbf{D}$ is a diagonal

Table 1: Averaged RMSE of The Baseline and Proposed Methods

	Arousal	Valence
Random Baseline ($D = 260$)	0.28 ± 0.13	0.29 ± 0.14
Multilinear Regression ($D = 260$)	0.27 ± 0.11	0.366 ± 0.18
ν-SVR ($D = 260$)	0.2377 ± 0.1089	0.3834 ± 0.1943
AV-SPE + ν-SVR ($d = 10$)	0.2414 ± 0.1081	0.3689 ± 0.1863

Table 2: Averaged Correlation of the Prediction and Ground Truth

	Arousal	Valence
Multilinear Regression ($D = 260$)	0.36 ± 0.26	0.01 ± 0.38
ν-SVR ($D = 260$)	0.5610 ± 0.2705	-0.0217 ± 0.4494
AV-SPE + ν-SVR ($d = 10$)	0.5806 ± 0.2290	0.0133 ± 0.4811

matrix defined as $D_{ii} = \sum_{j=1}^{n} \hat{S}_{ij}$ $(i = 1, ..., n)$, and $tr(\cdot)$ denotes the matrix trace operator. Obviously, $\mathbf{L}$ is positive semi-definite and $\mathbf{D}$ is positive definite. By transforming (1) to (2), the optimal solution can be easily obtained by employing standard eigendecomposition. Additionally, we introduce the constraint $\mathbf{U}^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{U} = \mathbf{I}_d$ to remove the scaling factor in the learning process, where $\mathbf{I}_d$ denotes the d-dimensional identity matrix. So for the first transformation vector $\mathbf{u}_1$, the problem becomes

$$\mathbf{u}_1 = \underset{\substack{\mathbf{u}_1 \\ \mathbf{u}_1^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{u}_1=1}}{\arg\min} \mathbf{u}_1^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{u}_1. \quad (3)$$

Then we obtain the Lagrangian equation of (3)

$$L(\mathbf{u}_1, \lambda) = \mathbf{u}_1^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{u}_1 - \lambda(\mathbf{u}_1^T\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{u}_1 - 1). \quad (4)$$

Letting $\partial L(\mathbf{u}_1, \lambda)/\partial \mathbf{u}_1 = 0$, the optimal $\mathbf{u}_1$ is the eigenvector corresponding to the smallest non-zero eigenvalue of the generalized eigendecomposition problem

$$\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{u} = \lambda\mathbf{X}\mathbf{D}\mathbf{X}^T\mathbf{u}. \quad (5)$$

Similarly, $\mathbf{u}_2, ..., \mathbf{u}_d$ are the eigenvectors corresponding to the 2-nd, ..., d-th smallest non-zero eigenvalues of (5), respectively.

2.2 Music Emotion Prediction via Support Vector Regression

After feature extraction, we can obtain the reduced features by $\mathbf{z}_i = \mathbf{U}^T\mathbf{x}_i$. We then use the reduced features as the input to predict the emotion labels of the music via the ν-Support Vector Regression (ν-SVR) [10]. Given the training set $\{(\mathbf{z}_1, \mathbf{y}_1), ..., (\mathbf{z}_n, \mathbf{y}_n)\}$, where $\mathbf{z}_i$ is the extracted feature vector and $\mathbf{y}_i = [y_i^{(1)}, y_i^{(2)}]$ is the corresponding label vector including arousal and valence values. For predicting the arousal and valence values, i.e., $y_i^{(1)}$ and $y_i^{(2)}$, we train two regressor separately. The final optimization problem, i.e., the dual problem that ν-SVR aims to solve is:

$$\begin{aligned} \min_{\alpha, \alpha^*} \quad & \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T\mathbf{K}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) + (\mathbf{y}^{(m)})^T(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \\ s.t. \quad & \mathbf{e}^T(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) = 0, \ \mathbf{e}^T(\boldsymbol{\alpha} + \boldsymbol{\alpha}^*) \leq C\nu \\ & 0 \leq \alpha_i, \alpha_i^* \leq C/n, \quad i = 1, ..., n, \end{aligned} \quad (6)$$

where α_i, α_i^* are the Lagrange multipliers, $\mathbf{K}$ is an $n \times n$ positive semidefinite matrix, in which $K_{ij} = K(\mathbf{z}_i, \mathbf{z}_j) = \phi(\mathbf{z}_i)^T\phi(\mathbf{z}_j)$ is the kernel function, $m = 1$ or 2, $\mathbf{e} = [1, ..., 1]^T$ is the n-dimensional vector of all ones, and $C > 0$ is the regularization parameter. The prediction label of a new coming vector $\mathbf{z}$ is:

$$y = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)K(\mathbf{z}_i, \mathbf{z}) + b. \quad (7)$$

3. EVALUATION RESULTS

In this section, we report the experimental settings and the evaluation results. The features used in our experiments are extracted via the openSMILE toolbox [5]. The original dimension of the feature space, i.e., D, is 260. We set the reduced dimension $d = 10$ for AV-SPE. The SVR toolbox we have used is the LIBSVM [4]. In the training process, we use the radial basis function (RBF) as the kernel function. The ten-fold cross-validation is employed to select the best parameters γ and C. We finally select $\gamma = 0.125$, $C = 2$, and $\nu = 0.5$ for our model.

Table 1 and Table 2 list the averaged Root-Mean-Square Error (RMSE) and the averaged correlation, respectively. From the tables, we can observe that the arousal result of ν-SVR and that of AV-SPE + ν-SVR are significantly better than the baseline. Moreover, the results on the reduced feature space ($d = 10$), i.e., the results of AV-SPE + ν-SVR, are comparable to the results on the original feature space ($D = 260$), which indicates that the extracted features play an important role in representing the music emotions.

4. CONCLUSIONS

In this working notes paper, we have introduced our system for music emotional dynamics prediction. The system is composed of a feature extraction algorithm and a support vector regressor. The evaluation results shown that the features extracted by the proposed AV-SPE are informative, and the system worked well in predicting the arousal values. Our future work will focus on extending the proposed algorithm by considering the dynamic nature of the music data.

5. ACKNOWLEDGMENTS

The authors would like to thank the reviewer for the helpful comments. This work was supported in part by the National Natural Science Foundation of China under Grants 61373122.

6. REFERENCES

[1] A. Aljanaki, Y.-H. Yang, and M. Soleymani. Emotion in music task at mediaeval 2015. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, September 2015.

[2] T. Baltrušaitis, N. Banda, and P. Robinson. Dimensional affect recognition using continuous conditional random fields. In *IEEE Conference on Automatic Face and Gesture Recognition*, 2013.

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003.

[4] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[5] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proc. 21st ACM International Conference on Multimedia*, pages 835–838, 2013.

[6] Y. Liu, Y. Liu, Y. Zhao, and K. Hua. What strikes the strings of your heart? – feature mining for music emotion analysis. *IEEE Transactions on Affective Computing*, PP(99):1–1, 2015.

[7] Q. Lu, X. Chen, D. Yang, and J. Wang. Boosting for multi-modal music emotion classification. In *Proc. 11th ISMIR*, pages 105–110, 2010.

[8] K. Markov and T. Matsui. Music genre and emotion recognition using gaussian processes. *IEEE Access*, 2:688–697, 2014.

[9] S. Rho, B.-j. Han, and E. Hwang. Svr-based music mood classification and context-based music recommendation. In *Proc. 17th ACM Multimedia*, pages 713–716, 2009.

[10] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Comput.*, 12(5):1207–1245, 2000.

[11] Y.-H. Yang, C.-C. Liu, and H. H. Chen. Music emotion classification: A fuzzy approach. In *Proc. 14th ACM Multimedia*, pages 81–84, 2006.

[12] M. You, J. Liu, G.-Z. Li, and Y. Chen. Embedded feature selection for multi-label classification of music emotions. *International Journal of Computational Intelligence Systems*, 5(4):668–678, 2012.