MUSIC EMOTION CLASSIFICATION

Hyeonguk Ryu
School of Computing
hyeonguk@kaist.ac.kr

Jitong Wu
Computer Science
wuatong0721@gmail.com

Keeyeon Park
Business Technology and Management
parkky88@kaist.ac.kr

ABSTRACT

Music emotion classification approaches typically divide into categorical and dimensional methods. However, each method cannot avoid the problems of subjectivity and ambiguity. In this paper, we propose a novel music emotion classification model to combine both methods. In preprocessing, we applied k-means clustering to the dataset and transformed the arousal-valence dimension values into integer values to classify the dataset into 10 classes. Then, we adopt two different methods such as feature extraction and end-to-end method. For feature extraction method, we used 6,669 features to classify music emotion into 10 different classes. In addition, we solve the classification problem using neural networks in an end-to-end manner. Then, we use the centers of each cluster as the final two-dimensional output in the arousal-valence plane in order to evaluate the performance of our approach. The results show that our proposed approach outperforms than the previous approaches used in MediaEval 2013 Emotion in Music.

1. INTRODUCTION

Music is a common element across national borders, human races, and cultures. It has powerful function to arouse emotions and feelings, even greater than language. The technical development such as nearly unlimited computer power has augmented its impact with increased availability and accessibility in daily life of people. Following this music digitalization, the way classifying and retrieving music is getting attention to meet the increasing demand for easier information access [1]. Since music is an integral mechanism that evokes emotion, music classification by perceived emotion is one effective approach.

Even though music emotion classification and retrieval has become popular in both academia and the industry, it is still a challenging assignment due to several reasons [2]. The major difficulty is the subjectivity problem, which describes that people intrinsically perceive different emotions from the same song. The subjectivity nature connects with the fundamental difficulty of performance evaluation of music emotion classification. The results from music emotion classification may not be agreeable to every person. The other issue is emotion taxonomy. Whether emotions are categories or continua, it is hard to describe emotions in a universal way. The adjective expression on emotions is ambiguous because each person has different definitions on the same emotion.

Previous literature has worked on defining music emotion with both dimensional and categorical approaches [3-9]. The dimensional approach defines emotions as numerical values such as valence and arousal and predicts the emotion value as a point in an emotion plane. The major problem in the continuous perspective is that there is no clear answer to independency between arousal and valence. Both can be inter-related.

The categorical taxonomy describes each emotion class as an area in the emotion plane. One basic method is to use four labels of music mood, such as happiness, sadness, anger, and fear and to classify emotions into four categories based on tempo and articulation [6]. Other works segment emotions in terms of arousal and valence to deal with the emotion taxonomy [7-9]. For example, a common emotion plane is Thayer's arousal-valence emotion plane [10] in Fig 1, which uses the four quadrants to divide the emotion classes. The categorical taxonomy, however, has still ambiguity problem. Even the same area in the emotion plane, the emotion states could be different. For example, the fourth quadrant in Thayer's emotion plane contains relaxed, peaceful, and calm, however, these three emotion states are distinguished in different ways in the real world. Therefore, neither continuous nor categorical perspective has the objective answer to classify music emotions.

In this paper, we adopt the essence of continuous and categorical perspectives to reduce subjectivity and ambiguity. We first use the arousal and valence values of each music data, so each sample can be represented as a point in the emotion plane. This attempt will make the ambiguity issue free. Then, we cluster emotion states without using regression techniques to reduce the level of subjectivity. We conducted feature extraction and end-to-end methods to find the best approach. To evaluate the prediction accuracy of the proposed approach, we compare with regression algorithms conducted in previous literature.

This paper presented a preliminary work that combines categorical and continuous perspectives in the emotion classification. We propose a neural approach that solves the emotion regression problem by classification by alleviating some level of subjectivity and ambiguity issues.

Our approach also provides high accuracy and lower mean squared error than previous works.

This paper is organized as follows. Section 2 introduces previous works on the music emotion classification. Section 3 presents our proposed approaches in detail. Section 4 show the experiment results. Section 5 gives discussion and conclusion.

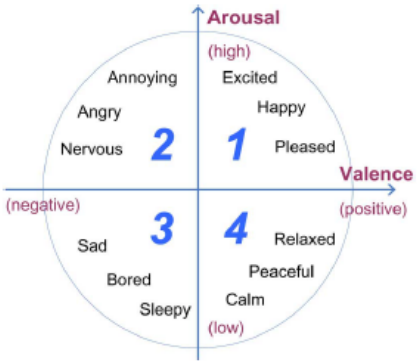

Figure 1. Thayer's arousal-valence emotion plane.

2. RELATED WORKS

Regardless of the variety of emotion models, music emotion classification is still in early stages with many problems such as natural subjectivity and emotion taxonomy. The emotion conceptualization can be divided into two different approaches: the categorical and the dimensional approaches.

The categorical approach consists of the limited number of primary emotion classes, such as happiness, sadness, anger, fear, disgust, and surprise [3]. However, the number of primary emotion classes are relatively small to human perceptions on music emotion. In addition, the adjectives and languages for emotions are ambiguous, because each researcher brings with different emotion classes.

The dimensional conceptualization of emotion emphasizes on classifying emotions based on the placement on any defined dimensions, which match to human perception on emotion. There are many attempts to define dimensions of emotions for interpretation. Russell [11] proposed the *circumplex model* of emotion, which consists of two-dimensional structure, pleasant-unpleasant and arousal-sleep. This structure was adopted by Thayer [10] to music emotion model [10]. Thayer corresponds energy to arousal and stress to pleasure. Even though the basic meaning of each dimension is similar, the names are various in the literature [13]. The stress and energy structure of Thayer was adopted to classify music emotions into four quadrants. The four emotion classes represent the basic perceptions to music. Then, it became the important structure in emotion model because it is relatively unambiguous and discriminable. The dimension names of emotion were changed to valence (positive and negative states) and arousal (energy and stimulation). However, the dimensional approaches cannot avoid the criticism of psychological distinctions of emotion classes.

In our system, we also adopt arousal-valence plane (AV plane) with four primary emotion classes. But we divide emotion classes into larger number of segments and define new class names to increase variety of emotion responses to songs.

3. PROPOSED APPROACHES

3.1 Dataset

We used EmoMusic 1000 songs database [14], which provides 744 songs, 6,669 features, continuous annotation and annotation for the whole song. The 744 songs are selected from Free Music Archive where 45 seconds excerpts are extracted from random starting point. Each song is annotated for arousal and valence separately. The dataset is treated in two different ways, feature extraction and end-to-end method.

In preprocessing, instead of implementing two times regression training, we applied k-means clustering directly to the dataset and transformed the AV dimension values into integer values which represent the cluster ID. We divide the dataset into 10 classes as a result. Figure 2 shows the clustering in the AV plane.

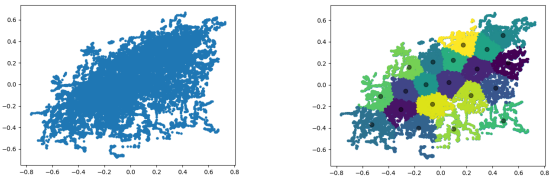

Figure 2. Preprocessing with k-means.

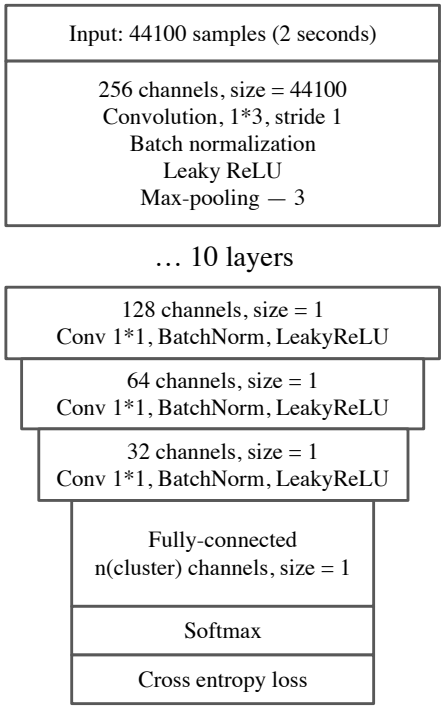

Figure 3. Network structure of the end-to-end method.

3.2 Feature Extraction Method

The training data contains 500 songs which are picked randomly from the dataset. In order to consider data augmentation, we chose to have features from a smaller segmentation, rather than taking all feature values. From the 15th and 45th seconds, the songs are separated into five segments, thus, the training dataset becomes five times larger than the original. The data is split into 60% training set, 20% validation set, and 20% test set.

The convolutional neural network (CNN) is used for the classifier. The model consists of three convolution layers which corresponds to batch normalization, ReLU, and Conv. The input is 6,669 features while the output is the cluster ID of 10 classes.

3.3 End-to-End Method

The songs were split into the development set and evaluation set, as presented in the database. Consequently, we obtained the development set of 619 songs and the evaluation set of 125 songs. We used 10% of the development set as a valid set. To filter out the outliers, we exploit the correlation between the arousal and valence. We performed a linear regression in the AV plane, then excluded points with large prediction error, before performing the k-means clustering. As a result, approximately 5% of the points were removed from the train set. Note that the filtering process was performed only in the train set.

The peak-to-peak values of the arousal and valence data were scaled to 1.0 and 1.5, respectively. Our model tends to overfit to arousal, and the unbalanced scales make the accuracy higher in the valid set. However, when calculating RMSE, we rescaled the peak-to-peak values to one. Also, we resampled the clips from 44,100 Hz to 22,050 Hz and scaled all the sound clips in order that the mean RMS power of each clip was the same. We also augmented the dataset in two ways, reducing the volume and adding Gaussian noise.

We solve the classification problem using neural networks in an end-to-end manner. We split the 45-seconds long clips into two-seconds long segments, with 1-second overlap. The last few samples of each clip were dropped. As a final step, a majority of the predictions were used as the final prediction of each clip.

The structure of the neural network is shown in Figure 3. The input of the model is a two-seconds of raw sound samples, and the output is an index of the predicted cluster. Our model consists of 14 layers of convolution and a fully-connected layer. Each convolutional layer consists of a one-by-three or one-by-one convolution, a batch normalization, a leaky ReLU activation, and a max pooling layer. The cross-entropy loss and Adam optimizer were used to train the model. 744 clips are too few for end-to-end learning and we used dropout and L2 regularization to prevent overfitting, with a dropout rate of 0.2 and a weight decay parameter of 0.003.

4. EXPERIMENTAL RESULTS

To evaluate the performance of our approach against the previous regression approaches, we add a slight modification in the final stage. We first solve the classification problem and use the centers of each cluster as the final two-dimensional output in the arousal-valence space.

First, the approach using feature extraction has 24.1% average test accuracy. The result of the end-to-end approach is shown in Table 1. Our method outperforms the baseline method and shows a similar performance than the methods submitted to MediaEval 2013 Emotion in Music. We can see that our methods showed relatively poor performance in valence, suggesting that predicting valence requires more complex features and large amounts of data than in arousal. While it cannot be compared directly with other previous methods, the classification accuracy was 0.24 in the feature-based method and 0.34 in the end-to-end method.

Run	Arousal		Valence	
	RMSE	R^2	RMSE	R^2
Baseline [14]	0.12	0.48	0.15	0
TUM [14]	0.10	0.59	0.11	0.42
ToA [14]	0.10	0.63	0.12	0.35
UU [14]	0.10	0.59	0.12	0.31
Ours	0.11	0.58	0.13	0.18

Table 1. RMSE and R^2 results on the test set.

5. CONCLUSION

In this paper, we have presented a preliminary work that combines categorical and continuous approaches in music emotion classification. However, both methods still have subjectivity and ambiguity problems. Our proposed neural approach alleviates the levels of these problems by classification. We proposed a new method solves the emotion regression problem by classification, featuring rescaling and k-means, and showed small errors in both arousal and valence. Further research could use neural networks that pretrained on larger datasets, to improve performance in predicting valence.

6. AUTHOR CONTRIBUTIONS

"Keeyeon Park" did the writing of the introduction and related work sections and organized all works. "Jitong Wu" collected datasets and conducted experiments for the feature extraction method. "Hyeonguk Ryu" proposed a new idea, implemented it, and conducted experiments for the end-to-end method.

7. REFERENCES

[1] Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. Proceedings of the IEEE, 96(4), 668-696.

[2] Yang, Y. H., & Chen, H. H. (2012). Machine recognition of music emotion: A review. ACM Transactions on Intelligent Systems and Technology (TIST), 3(3), 40.

[3] Yang, Y. H., Lin, Y. C., Su, Y. F., & Chen, H. H. (2007, July). Music emotion classification: A regression approach. In 2007 IEEE International Conference on Multimedia and Expo (pp. 208-211). IEEE.

[4] Yang, Y. H., Lin, Y. C., Su, Y. F., & Chen, H. H. (2008). A regression approach to music emotion recognition. IEEE Transactions on audio, speech, and language processing, 16(2), 448-457.

[5] Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Transfer learning for music classification and regression tasks. arXiv preprint arXiv:1703.09179.

[6] Feng, Y., Zhuang, Y., & Pan, Y. (2003, July). Popular music retrieval by detecting mood. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 375-376). ACM.

[7] Wu, T. L., & Jeng, S. K. (2006, March). Extraction of segments of significant emotional expressions in music. In Workshop on Computer Music and Audio Technology (pp. 76-80).

[8] Yang, Y. H., Liu, C. C., & Chen, H. H. (2006, October). Music emotion classification: A fuzzy approach. In Proceedings of the 14th ACM international conference on Multimedia (pp. 81-84). ACM.

[9] Lu, L., Liu, D., & Zhang, H. J. (2005). Automatic mood detection and tracking of music audio signals. IEEE Transactions on audio, speech, and language processing, 14(1), 5-18.

[10] Thayer, R. E. (1990). The biopsychology of mood and arousal. Oxford University Press.

[11] Russell, J. A. (1980). A circumplex model of affect. Journal of personality and social psychology, 39(6), 1161.

[12] Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Transactions on Pattern Analysis & Machine Intelligence, (10), 1175-1191.

[13] Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. Journal of Experimental Psychology: Human perception and performance, 26(6), 1797.

[14] Soleymani, M., Caro, M. N., Schmidt, E. M., Sha, C. Y., & Yang, Y. H. (2013, October). 1000 songs for emotional analysis of music. In Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia (pp. 1-6). ACM.