Toward Multiple Emotion Classification from Musical Audio Signals

Jaesung Lee, Jae-Joon Lee, Song Ko, Jin-Hyuk Jo, Hyunki Lim, JongHoon Chae, Jeonghun Yoon, and Dae-Won Kim*

School of Computer Science and Engineering, Chung-Ang University, 221, Heukseok-Dong, Dongjak-Gu, Seoul 156-756, Republic of Korea
dwkim@cau.ac.kr

Abstract. Music emotion recognition is one of the best attractive research in music information retrieval. To develop an efficient music information retrieval system, a music data set related to music emotion is required. In this paper, we released two music emotion data sets, and proposed a feature selection algorithm that can be useful to investigate the relation between musical property and multiple emotion.

Keywords: Music Emotion Recognition, Multi-label classification

1 Introduction

Music is one of the most common ingredients to encourage our daily life. To satisfy music listener's diverse interesting and favorite styles, explosive number of music are published in a day after day. To categorize the large-scale music corpus, researchers focused on emotion, genre, theme and so on [1]. In this paper, we focus on the music emotion recognition (MER). The goal of MER is to recognize the internal emotion of music clip. Early researches on MER were mainly focused on the extraction of emotion-related musical properties. Juslin et al. [2] found that tempo, intensity, and spectrum are highly related to various emotional expressions. After extracting the musical properties, classifiers can be used to recognize the music emotion. Trohidis et al. [3] examined the performance of music emotion recognition in four multi-label music emotion settings.

We summarize our contributions as follows. The first contribution is to create and release two music data sets that consist of musical features and multiple emotions; a music can be assigned maximum four different music emotions. It would be helpful for researchers who have trouble of developing MER system due to the rarity of multi-emotion data. The second contribution is to devise a new multi-label feature selection algorithm that can be applied to MER problem in multi-labeled setting. From the hundreds of musical features, we tried to identify significant features that are highly dependent to emotions, along with minimizing inter-dependency among selected features. The last contribution is to investigate relations between musical audio features and emotions by employing an associative rule mining method to enhance model interpretability.

* Corresponding Author

2 Proposed Method

2.1 Data set

The procedure of constructing a MER data set is given below. First, musical properties were extracted from musical audio signals. Feature extraction were performed by using short-term fourier transform and heuristic musical property detectors [4] and so on. Those signal transformation methods are applied to each music clip, then musical properties were converted from audio signals to numerical values. Second step of forming the data set is to collect emotional responses for each music. To address the emotion, researchers employed the psychological theory related to the emotion in music [3]. In this paper, we exploit the Thayer's emotion model [5] because of its computational effectiveness [1].

We assembled two data sets; `MusicEmo-A` and `MusicEmo-B`. In the case of `MusicEmo-A`, we collected 100 music clips from popular five genres and had them labeled by approximately 500 times through an on-line annotation system. For each music, we used the MIR toolbox that offers integrated set of functions to extract musical audio features [4]. The extracted features fall into six types: dynamics, fluctuation, rhythm, spectral, timbre, and tonal features. At the ends, we obtained `MusicEmo-A` that is composed of 100 patterns and 864 features. For `MusicEmo-B`, we collected 565 music clips from eleven genres, and annotated by approximately 3,600 times. After having applied the MIR toolbox to music clips, we obtained `MusicEmo-B` that is composed of 565 patterns and 346 features; features in `MusicEmo-B` are put forth by the statistical viewpoint (Refer to [4]).

In Thayer's emotion model, a state of human emotion is represented as a point (vector) in a two-dimensional emotion space; valence and arousal. For simple description of multiple emotions for each music, we used the four sub-planes (zones) in Thayer's model as emotion labels. The first plane, positive arousal and positive valence labeled by $l_1 = \{+, +\}$, represent `Excitement` feeling. Other three planes can be labeled as `Distress` $l_2 = \{+, -\}$, `Depression` $l_3 = \{-, -\}$, and `Contentment` $l_4 = \{-, +\}$. We assumed that the perceived emotion may distribute equally all of planes, if a music does not significantly evoke any emotion. Since each music clip is labeled by many participants who may perceive different emotions from it, we represent the multiple emotions of a music clip as a set of aggregated labels over the four emotion planes if an emotion is labeled by more than 33% of participants. Two data sets are publicly available from the Softwares section of `http://ai.cau.ac.kr/`.

2.2 Feature Selection

Hundreds of musical features are not typically of equal importance in the emotion classification, moreover irrelevant features may degrade the recognition performance. One straightforward way to select the significant features is to identify emotion-relevant features and their interactions related to multiple emotions.

Let $X \subseteq \mathbb{R}^n$ denote an input space that is constructed from n features and patterns drawn from X are assigned to a certain label subset $\lambda \subseteq L$, where L

$= \{l_1, \cdots, l_m\}$ is a finite set of labels with $|L| = m$. Suppose we already have $S_t = \{f_1, \ldots, f_t\}$; the feature subset calculated from the t-th step. The proposed method selects the next feature f_{t+1} from $F - S_t$ and add it to S_{t+1}; (1) f_{t+1} should have a high correlation to labels, (2) f_{t+1} should be mutually far away from pre-selected feature subset S_t. The first task can be achieved by selecting f_{t+1} that maximizes the mutual information (MI) between f_{t+1} and L:

$$I(f_{t+1}; L) = \underbrace{I(f_{t+1}; l_1)}_{\text{A}} + \underbrace{I(f_{t+1}; l_2|l_1) + \cdots + I(f_{t+1}; l_m|\{l_1, \ldots, l_{m-1}\})}_{\text{B}} \quad (1)$$

where the conditional mutual information (CMI) between f_{t+1} and l_m, given $\{l_1, \ldots, l_{m-1}\}$, is defined as:

$$\begin{aligned} I(f_{t+1}; l_m|\{l_1, \ldots, l_{m-1}\}) =& H(l_1, \ldots, l_{m-1}, f_{t+1}) + H(l_1, \ldots, l_{m-1}, l_m) \\ & - H(l_1, \ldots, l_{m-1}, l_m, f_{t+1}) - H(l_1, \ldots, l_{m-1}) \end{aligned} \quad (2)$$

The entropy $H(S) = -\sum_{s \in S} P(S) \log P(S)$ is the uncertainty measure of given variable set S, and $P(S)$ is a probabilistic mass function of S. Eq. (2) indicates that if one of input variables, f_{t+1} or l_m, is dependent to any of given variables, $l_1, \ldots, l_{m-1}$, then it leads to a lower value of CMI. Thus, we can see that if f_{t+1} is the most relevant feature to l_1 (the part (A) in Eq. (1)), then the part (B) should be minimized because l_1 is given in all terms of (B). From this aspect, we approximate Eq. (1) as:

$$I(f_{t+1}; L|S_t) \approx \max_{l_i \in L} I(f_{t+1}; l_i|S_t) \quad (3)$$

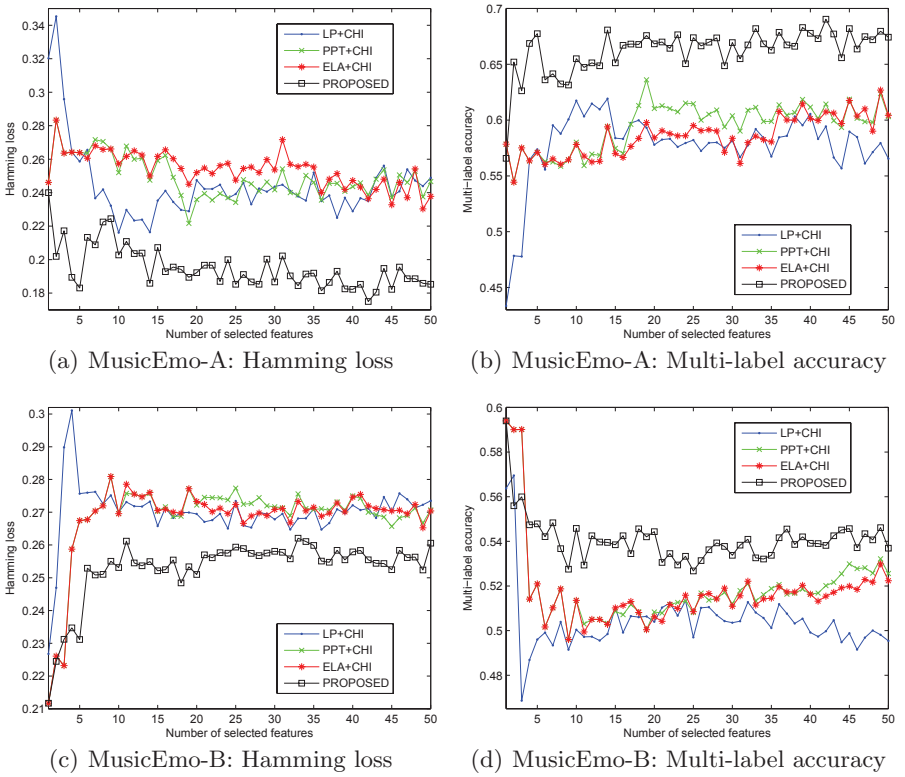
Eq. (3) says that the feature f_{t+1} should be dependent to at least one of labels.

Secondly, the feature f_{t+1} should be mutually independent to S_t. This can be done by maximizing the CMI when S_t is given, i.e., $I(f_{t+1}; L|S_t)$. However, it is computationally prohibitive to calculate an accurate value of high dimensional joint entropy owing to the growing size of S_t. To circumvent this situation, we approximate the CMI as:

$$I(f_{t+1}; L|S_t) \approx \min_{f_j \in S_t} I(f_{t+1}; L|f_j) \quad (4)$$

Note that S_t can be considered as the Markov blanket for f_{t+1}, provided that S_t is mutually dependent to f_{t+1} [6]. To find a feature f_{t+1} that is not covered by S_t, in the present work we take the first order Markov blanket for f_{t+1} into account. The conditional dependency of f_{t+1} is reduced when f_j in S_t covers f_{t+1}, so that we can approximate the $I(f_{t+1}; L|S_t)$ by using the minimum value among $I(f_{t+1}; L|f_i)$. By combining Eqs. (3) and (4), we can find the feature f_{t+1} that maximizes the optimization criterion J:

$$J = \max_{f_{t+1} \in F - S_t} \left[\max_{l_i \in L} \min_{f_j \in S_t} I(f_{t+1}; l_i|f_j) \right] \quad (5)$$



(a) MusicEmo-A: Hamming loss (b) MusicEmo-A: Multi-label accuracy

(c) MusicEmo-B: Hamming loss (d) MusicEmo-B: Multi-label accuracy

Fig. 1. Classification performance of the `MusicEmo-A` data set (Top) and `MusicEmo-B` data set (Bottom) according to feature subsets obtained by the proposed method and three conventional feature selection methods: LP+CHI, PPT+CHI, and ELA+CHI.

3 Experimental results

We compared the performance of the proposed method with that of conventional methods (LP+CHI, PPT+CHI and ELA+CHI [3, 7, 8]); LP, PPT, and ELA are the well-known problem transformation methods in multi-label classification, and CHI denotes χ^2 statistics. The classification performances were evaluated using a multi-label naive bayes classifier [9]. The performance was assessed using two measures: hamming loss and multi-label accuracy [9]. Low values of hamming loss and high values of multi-label accuracy indicate good classification performance. To make this study more informative, we employed an associative rule mining algorithm [10] to investigate relations between selected musical features and emotions. We examined the top-ranked features selected by the proposed method in order to preserve model interpretability. Due to the space limitations, we presented top-ranked confident rules obtained from `MusicEmo-B` data set. Through the analysis, we see that human-readable rules can explain the relationship between musical audio features and emotions in music.

Table 1. Top three associative relations between musical features and emotions for `MusicEmo-B` data set. The features employed were TSM (timbre spectralflux mean), TKM (tonal keyclarity mean), TZM (timbre zerocross mean), SCM (spectral centroid mean), and SMS (Spectral Melfrequency cepstral coefficient STD).

Emotion	A/V	Clips	Top-rules	Top Five Features				
				TSM	TKM	TZM	SCM	SMS
`Excitement`	+/+	315	Rule 1		high	**high**		low
			Rule 2		low	**high**		high
			Rule 3			**high**	low	
`Distress`	+/−	85	Rule 1	**low**	low	**high**		
			Rule 2	**low**		**high**	high	
			Rule 3	**low**		**high**		
`Depression`	−/−	135	Rule 1		**high**	low	high	**high**
			Rule 2	low	**high**		high	**high**
			Rule 3		**high**	high		**high**
`Contentment`	−/+	195	Rule 1	**low**	high		**high**	
			Rule 2	**low**			**high**	high
			Rule 3	**low**	low		**high**	high

Fig. 1 shows the classification performance of each feature selection method for `MusicEmo-A` data set and `MusicEmo-B` data set. The horizontal axis represents size of the selected feature subset according to each feature selection method, and the vertical axis indicates that hamming loss value (Left) and multi-label accuracy (Right). We see from Fig. 1(a) that the hamming loss of the proposed method improved with size of the selected feature subset. The best classification performance is achieved by the proposed method with 0.175 of hamming loss value when 42 features are selected. Fig. 1(b) shows that multi-label accuracy was improved to a great extent by using the proposed method. Two bottom figures represent the classification performance for `MusicEmo-B` data set. From Fig. 1(c), it is observed that the proposed method demonstrated superior performance to conventional methods. The hamming loss by LP+CHI, PPT+CHI, ELA+CHI, and the proposed method were 0.270, 0.270, 0.270, and 0.253 respectively for 10 selected features. In Fig. 1(d), the multi-label accuracy by the proposed method was 0.546 when 10 features were selected, whereas LP+CHI, PPT+CHI, and ELA+CHI achieved multi-label accuracies of 0.500, 0.514, and 0.514, respectively. In summary, it is evident from test results that the proposed method is superior to three conventional methods.

We analyzed relations between selected features and emotions by using the association rule mining method [10], after discretizing each feature according to its average; values of a feature over its average are assigned to `high`, otherwise values are assigned to `low`. By using the top five audio features selected by the proposed method from `MusicEmo-B`, the top three rules with high confidence were identified for each emotion. Table 1 shows the emotion, A/V status, the number of music clips, and the body of rules. Each rule is composed of five features and their assigned values, for example, Rule 1 in `Excitement` indicates that if TKM

and TZM are given high values and SMS is assigned a low value, then emotion of a music clip is assigned to `Excitement`. Common features in the three rules are marked in bold text. Table 1 indicates that a music clip is likely to be assigned to `Excitement` or `Distress` when TZM takes high values. In contrast, `Depression` is dominantly influenced by TKM(high) and SMS(high), while `Contentment` is affected by TSM(low) and SCM(high). It is interesting to note that `Excitement` and `Distress` belong to sub-planes with positive arousal in Thayer's emotion model. Therefore, we can say that a high value of TZM in a music clip makes users feel a high value of arousal; TZM indicates the number of time domain zero crossings in the signal. Rock and Dance music often contain a high value of TZM. In the `Contentment`, two features TSM and SCM are reported to be dominant; TSM measures the amount of local spectral change and SCM is a measure of spectral brightness. Thus, a low amount of local change and a high brightness determines the softness of music, which is agreed on by human experts.

4 Conclusion

In this study, we have released new, publicly available two music data sets that consist of musical features and corresponding multiple emotions. The superiority of the proposed method over other conventional method was clearly demonstrated through several experiments.

Acknowledgments. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0001772)

References

1. Huron, D.: Perceptual and Cognitive Applications in Music Information Retrieval, Perception. 10, 83–92 (2000)
2. Juslin, P.: Cue Utilization in Communication of Emotion in Music Performance, J. Exp. Psychology: Hum. Perception Perform. 26, 1797–1813 (2000)
3. Tsoumakas, K., Kalliris, G., Vlahavas, I.: Multi-label Classification of Music into Emotions. In: Int. Conf. Music Inform. Retr., Philadelphia (2008)
4. Lartillot, O., Toiviainen, P.: MIR in Matlab (II): A Toolbox for Musical Feature Extraction from Audio. In: Int. Conf. Music Inform. Retr., Vienna (2007)
5. Thayer, R.: The Biopsychology of Mood and Arousal, Oxford Univ. Press (1989)
6. Yu, L. and Liu, H.: Efficient Feature Selection via Analysis of Relevance and Redundancy, J. Mach. Learn. Res., 5, 1205–1224 (2004)
7. Read, J.: A Pruned Problem Transformation Method for Multi-label Classification. In: New Zealand Comput. Sci. Res. Student Conf., Venue (2008)
8. Chen, W. et al.: Document Transformation for Multi-label Feature Selection in Text Categorization. In: IEEE Int. Conf. Data Min., Omaha (2007)
9. Zhang, M.L., Peña, J.M., Robles, V.: Feature selection for Multi-label Naive Bayes Classification, Inform. Sci. 179, 3218–3229 (2009)
10. Liu, B., Hsu, W., Ma, Y.: Integrating Classification and Association Rule Mining. In: Int. Conf. Knowl. Discov. Data Min., New York (1998)