Music Mood Classification using Reduced Audio Features

Babu Kaji Baniya and Choong Seon Hong
Department of Computer Engineering
Kyung Hee University
{ babu775, cshong}@khu.ac.kr

Abstract

Music emotion is a crucial component in the field of multimedia database retrieval and computational musicology. Nowadays, the large online musical datasets are major challenges for searching, retrieving, and organizing the music content. Therefore, there is a need for robust automatic music emotion classifier system for organizing various music pieces into different classes according to the specific viable information. Two fundamental components are to be considered for music emotion classification: audio feature extraction and classifier design. In this paper, we propose diverse audio features to precisely characterize the music content. The feature sets belong to four groups: dynamic, rhythmic, spectral, and harmonic. From the features, five statistical parameters are considered as representatives, including the fourth-order central moments of each feature as well as covariance components. The large number of insignificant parameters is controlled by minimum redundancy maximum relevance (MRMR) algorithm and principal component analysis (PCA). Support Vector Machine (SVM) is used as a classifier to classify the music mood.

1. Introduction

The automatic music emotion classification has gained increasing attention in the field of music information retrieval. The research activities in this field are not only highly diversified, but also consistently growing. The diversity comes from the fact that emotions classification establishes certain relationships between music and its effect in human emotional state like happy, anger, sad, tender, etc. In addition, the growth is inevitable nowadays, to increase the accessibility to music databases. As the amount of music content continues to explode, the searching time is unexpectedly increasing. The solution of widespread music grouped under different emotions could lead to a reduction in the information retrieval search time on the online system.

Some of the major difficulties in MER are related to the fact that the perception of emotions evoked by the song is inherently subjective: different listeners often receive distinct emotions while listening the same song. Besides, even when listeners agree in the perceived emotion, there is still much ambiguity regarding its description. Furthermore, it is not yet well-understood how and why music elements create specific emotional response in listeners [1]. In general, there are two models to describe emotions i.e. a categorical and dimensional one [2]. The categorical model focuses on the characteristics that distinguish emotions from one another. The essentiality of this model is the concepts of basic emotions, such as happiness, sadness, anger, fear, disgust, and surprise. On the other hand, the dimensional model focuses on identifying emotions based on their position in a continuous dimensional emotion spaces with a small number of axes. In the emotion space, each bipolar axis usually has its own meaning i.e. valance and arousal.

The music mood classification with reduced feature sets using support vector machine (SVM) classifier is shown in the Fig. 1. It represents the overview of proposed method of mood classification. Basically, there are certain problems need to be addressed in music mood classification i.e. audio feature extraction and classifier design. Besides these, feature analysis also plays a significant role in mood classification. The feature analysis means finding out the most discriminative feature or set of features from feature pool. In this scheme, minimum redundancy maximum relevance (MRMR) [3] and PCA [4] approaches implement for feature reduction. MRMR gives the

maximum relevance value as a score of each feature statistics in descending order. Based on the requirement, number of audio feature statistics has been selected to achieve maximum classification accuracy. Furthermore, the unique aspect of MRMR is to keep the original features as it is. However, PCA transforms the original features and order them based on variance difference. Later, desire number of transform feature statistics has been selected for classification.

2. Overview of Proposed Method

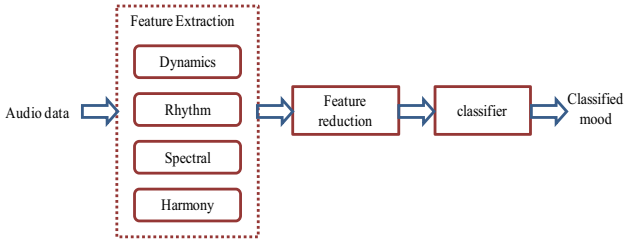

Figure 1: Block diagram of overview of proposed method

2.1. Audio Feature Extraction

Audio feature extraction is the extraction of meaningful information from an audio sample in order to obtain a compact and concise description that is machine-readable. Features are usually selected in the context of specific tasks and domains, and are divided into two categories, i.e., low- and high-level, according to the frame size. The frame lengths for the low-level and high-level features were 46 ms and 2 s, respectively, with both having 50% overlap. Each frame-based feature provides a sequence of frame-wise values for audio data, which allows a set of a small number of meaningful scalar values to be determined. Subsequently, frames of each song have been integrated by means of the different statistical values [5] like mean (M), standard deviation (S_{td}), skewness (S_k), kurtosis (K_t), and covariance (Cov) over a music piece. The mean (μ) and standard deviation (σ) for frame-wise feature values (X_n) in a N-frame music.

3. Feature Reduction Methodologies

3.1. Audio Feature Extraction

The MRMR criterion was proposed in [6], [7] in combination with forward selection search strategy. Given a set X_s of selected variables, the criterion consists of updating X_s with the variable $X_i \in X_t$ (it is difference between the original set of input and set of variables X_s selected so far) that maximizes $u_i - z_i$, where u_i is a relevance term and z_i is a redundancy term. Moreover, I is the mutual information of two variables, u_i is the relevance of X_i to the output Y alone, and z_i is the average redundancy of X_i to the selected variables $X_j \in X_s$.

$$u_i = I(X_i; Y) \tag{1}$$

$$z_i = \frac{1}{d} \sum_{X_j \in X_s} I(X_i; Y) \tag{2}$$

$$X_i^{MRMR} = \arg \max_{X_i \in X_t} \{u_i - z_i\} \tag{3}$$

3.2. Principal Component Analysis

PCA is a statistical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of variables called principal components. The number of principal components is less than or equal to the number of original variables. It is called a discrete version of Karhunen-Loeve transform. For a set of M-dimensional data $X = [x_1, x_2, ..., x_M]^T$, let $\{\lambda_1, \lambda_2, .., \lambda_R, .., \lambda_M\}$ and $[w_1, w_2, ..., w_R, ..., w_M]$ be the eigenvalues in a descending order and the corresponding orthonormal eigenvectors of $E[XX^T]$. To reduce the M-dimensional data to R-dimensional space, the reduced number of eigenvectors $W = [w_1, w_2, ..., w_R]$ is applied as given below

$$Y = W^T X \tag{9}$$

Table 1: Coimbra dataset having several adjectives and corresponding songs

Cluster	Adjective in class	No. of songs
C1	passionate, rousing, confident, boisterous, rowdy	170
C2	rollicking, cheerful, fun, sweet, and amiable/good natured	164
C3	literate, poignant, wistful, bittersweet, autumnal, brooding	215
C4	humorous, silly, campy, quirky, whimsical, witty, wry	191
C5	aggressive, fiery, tense/anxious, intense, volatile, visceral	163

4. Experimental Setup and Data Preparation

The dataset is taken from Informatics and System of University of Coimbra Portugal. There are 903 mp3 audio in total. It contains five clusters with several emotional categories each given Table 1.

Fig. 2 shows that classification accuracy including lower order moments, higher order moments, and

covariance components. The feature dimension increases sharply in this scheme. There are 538 feature (statistic) dimensions in total. Therefore, two different feature reduction methodologies have been implemented before the classification. Based on reduced feature sets, 25 number of features interval is considered as shown in Fig. 2 i.e. multiple of 25. The classification accuracy is continuously increased up to 150 features in both cases. Thereafter, the classification accuracy has not been improved. The classification accuracy using MRMR is 57.42% while considering 150 feature statistics. At the same time, the classification accuracy using PCA is only 47.27%. From the experiment, it is known that only 27.88% feature statistics are enough to achieve above classification accuracy. It means that these 27.88% feature statistics have sufficient discriminative ability to maximize the classification accuracy. Ultimately, minimize the computational complexity of a classifier.

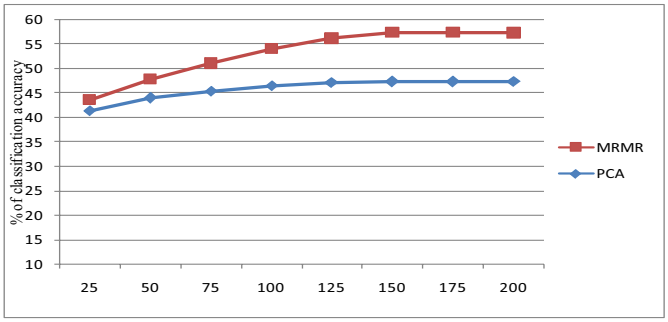

Figure 2: Number of features Vs classification accuracy considering all statistics

Table 2: Confusion matrix of Coimbra dataset

	C1	C2	C3	C4	C5
C1	**49.12**	11.56	6.65	18.21	14.46
C2	17.65	**50.23**	6.37	14.26	11.73
C3	7.26	12.43	**67.02**	9.08	4.21
C4	20.52	9.03	12.71	**51.44**	6.30
C5	11.78	2.41	7.26	14.23	**64.32**

5. Conclusion

In this paper, diverse audio feature such as dynamics, rhythm, spectral, and harmony are selected for music emotion classification. In the next stage, these features are integrated using lower (mean and standard deviation) and higher order moments (skewness and kurtosis). Similarly, covariance components are also calculated to improve the classification. Based on obtained feature statistics, the experiments were performed an experiment. The direct consequence of considering all statistics was to rapid increment feature dimension. Therefore, they were controlled by two feature reduction methodologies i.e. PCA and MRMR. The classification accuracies of lower order statistics achieved 57.42% and 47.27% using MRMR and PCA reduction feature sets respectively. Similarly, 72.12% of features were insignificant for mood classification. Only 27.88% feature statistics contributed for mood classification. The MRMR based feature reduction method using SVM comparatively gave better accuracy than other contemporary methodologies. Moreover, MRMR algorithm calculated maximum relevance features from the feature pool.

Acknowledgement

This research was supported by Basic Science Research Program through National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2014R1A2A2A01005900). Dr. CS Hong is the corresponding author.

References

[1] R. Panda, R. Malheiro, B. Rocha, A. Oliveira, and R. P. Paiva, “Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis,” *International Symposium on Computer Music Multidisciplinary Research* 2013

[2] J. A. Russell, A.Weiss, and G. A. Mendelsohn, “Affect grid: A single item scale of pleasure and arousal,” *J. Personality Social Psychol.*, vol. 57, no. 3, pp. 493-502, 1989

[3] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp.1226-1238, Aug. 2005.

[4] L. Smith, “A Tutorial on Principal Components Analysis,” Available:www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf, 2000

[5] Kyoungro Yoon, Jonghyung Lee, Min-Uk Kim, “Music recommendation system using emotion low-level features,” *Transactions on Consumer Electronics,* vol. 58-2, pp. 612-618, 2012

[6] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp.1226-1238, Aug. 2005.

[7] H. Peng and F. Long, “An efficient max-dependency algorithm for gene selection,” *in 36th Symp. Interface*: Computational Biology and Bioinformatics May 2004