SEMI-SUPERVISED CLASSIFICATION OF MUSICAL GENRE USING MULTI-VIEW FEATURES

Yunpeng Xu
Tsinghua University
Department of Automation

Changshui Zhang
Tsinghua University
Department of Automation

Jing Yang
Tsinghua University
Department of Automation

ABSTRACT

Musical genre classification is a key problem in multimedia information retrieval. Traditional musical genre classification methods are complete supervised, i.e., large amount of annotations are needed. In addition, if more than one feature sets are used, they are simply concatenated to form a long feature vector, which is sometimes problematic. To solve these problems, we introduce to use multi-view features and Co-Training algorithm. More specifically, we adopt three most popular feature sets to form two feature views for classification, so as to extract as much discriminative information contained by different feature sets as possible. We then use the Co-Training algorithm to classify musical pieces with only a few annotations. Experiments prove the validity and effectiveness of this method.

1. INTRODUCTION

With the rapid developments of various affordable technologies, the amount of multimedia now available on internet remarkably increases. Accordingly, it has become more important to automate the work of querying a database of musical pieces. One of such efforts focuses on automatic classification of musical genre, a fundamental task for music information retrieval.

There have been many studies on musical genre classification, both on feature extraction and selection and on comparison of different classifiers. While extracting musical features, music pieces are generally cut into frames. From each frame, we compute a feature vector of descriptors of timbre, rhythm and pitch, etc. Then the statistics of all frames are computed as the representation of the whole music piece. The most popular feature sets include Short Time Fourier Transform (STFT) based features, Discrete Wavelet Transform (DWT) based features, Mel Frequency Ceptral Coefficients (MFCC), Linear Prediction Coefficients (LPC), rhythm and pitch contents features[3], etc. Deshpande[5] proposes to use spectrogram to classify music in vision domain. Grimaldi[7] discusses musical feature selection and introduces random subspace method. In case of classifiers, the most popular ones include KNN, GMM[3],LDA[8] and SVM[2]. In [7], Grimaldi also introduces bagging, boosting and Round-Rubin techniques. A more detailed review can be found in [6].

However, there are still some problems with the existing methods. First, most of these methods are complete supervised, i.e., large amount of annotations are needed, sometimes 90% of all samples are used for training the classifier. While in a real music querying system, the work of manual annotation is rather time consuming, which means only very few musical pieces can be labelled. Therefore, it makes sense to introduce the method of semi-supervised classification. Second, if more than one feature sets are used, they are simply concatenated to form a long feature vector. However, the concatenated feature would lack its physical meaning and probably result in worse effects. Therefore, it is desirable to study how to effectively combine these different feature sets so that more discriminative information can be extracted from the music pieces.

To properly solve the above two problems, the paper proposes to combine different feature sets into a whole feature set and split it into several feature subsets, which are called views. Co-Training algorithm, a typical algorithm for semi-supervised learning with multi-view data, is then used to classify music pieces with only very few annotations. In this way, discriminative information from different feature sets can be effectively extracted while the number of music pieces required to be labelled reduces.

The paper is organized as follows. Section 2 presents the details of the proposed method for semi-supervised classification of musical genre with multi-view features. In Section 3, experimental results of the proposed method on a dataset of 300 music pieces are given. In the last section, some concluding remarks and the direction of future work are presented.

2. METHOD

In this section, the features used in this paper are first introduced, we then discuss the division of feature views and finally introduce briefly the Co-Training algorithm for semi-supervised classification of multi-view musical data.

2.1. Feature Extraction

The aim of feature extraction is to represent music pieces compactly and efficiently. Different systems may distinctly vary in the features they use. In this paper, we adopt three feature sets that are commonly used in musical genre classification systems, including STFT based features, MFCC and DWT based features. After feature extraction, the classification task can be achieved using standard machine learning methods.

2.1.1. STFT Based Features

Spectral Centroid is the center of gravity of the STFT magnitude spectrum

$$C_t = \frac{\sum_{n=1}^{N} M_t(n) * n}{\sum_{n=1}^{N} M_t(n)} \tag{1}$$

where is the magnitude of the Fourier transform at frame t and frequency bin n.

Spectral Rolloff is another measure of spectral shape. It is defined as the frequency R_t below which 85% of the magnitude distribution is concentrated

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 * \sum_{n=1}^{N} M_t[n] \tag{2}$$

Spectral Flux is a measure of the local spectral change amount, which is defined as

$$F_t = \sum_{n=1}^{N} (N_t(n) - N_{t-1}(n))^2 \tag{3}$$

where $N_t(n)$ is the normalized Fourier transform magnitude at frame t.

Time Domain Zero Crossings is a measure of the signal noisiness, which is defined as

$$Z_t = 0.5 * \sum_{n=1}^{N} |\text{sgn}(x(n)) - \text{sgn}(x(n-1))| \tag{4}$$

where $x(n)$ is the time domain signal for current frame.

Low Energy is defined as the percentage of frames that have less RMS energy than the average RMS energy of all frames.

While extracting STFT based features, music pieces are cut into frames of 512 samples at 22050 Hz sampling rate with a hop size of 256 samples. Then means and variances of the above features (except for the Low Energy feature) are computed . This results in a 9-dimensional feature vector.

2.1.2. MFCC Features

MFCC features are widely used in the field of speech recognition. They are proved to be very effective in modelling the spectrum magnitude of audio signals. The extraction of these features takes into account the human auditory characteristics by adopting filter banks and transforms that are similar to human auditory systems. More Specifically, the extraction of MFCC features involves the following operations.

(1) Cut a music piece into frames. Usually each frame is windowed with a Hamming window to reduce the edge effects.
(2) Perform Discrete Fourier Transform for each frame and take the logarithm of the magnitude.
(3) 40 frequency bins are computed according to Mel scale, then the output is reduced to the desired dimensionality with DCT.

After cutting frames in the same way as above, we compute from each frame the MFCCs, which are typically 13 coefficients. Here, only the first 5 coefficients are taken and used to compute their means and variances. This results in a 10-dimensional feature vector.

2.1.3. DWT Based Features

Wavelet Transform (WT) is developed to overcome the problems of STFT's frequency and time resolution properties. It provides high time resolution and low frequency resolution for high frequencies, and low time resolution and high frequency resolution for low frequencies. The extracted wavelet coefficients is an effective and compact representation of music signals.

The features we extracted from wavelet coefficients are proposed in [4], as follows.

(1) The mean of the absolute value of the coefficients in each subband.
(2) The standard deviation of the coefficients in each subband.
(3) Ratios of the mean absolute values between adjacent subbands.

While extracting DWT based features, music pieces are cut into frames of 3 seconds, with a hop size of 512 milliseconds. 12 subbands coefficients are extracted using DWT, which results in a 35-dimensional feature vector.

2.2. Multi-View Feature Sets

An ideal classification system should be able to extract as much discriminative information contained by different feature sets from the data as possible. As for musical genre classification, people have proposed various feature sets which are proved to be effective to some extent. However, the existing methods also have some problems: if more than one feature sets are used, they are simply concatenated to form a long feature vector. Such an approach is sometimes problematic. As different features have different physical meanings and different classification effects, such a simple concatenation would probably make the resulting feature vector loss its original meaning and even cause worse effects.

In fact, it makes more sense to combine different feature sets into a whole feature set. This feature set can be split into several feature subsets, called views, according to their extraction methods, physical meanings and classification effects. Then learning algorithms using multi-view features, such as Co-Training, Co-EM, Co-Boosting, etc., can be adopted to classify the music. Classifiers designed in this way can both make use of the discriminative information contained by different feature sets and maintain the original physical meaning and effect of each feature set.

As for the feature sets used in this paper, STFT based features and MFCC features share more similarities because both are computed using FFT and are measures of FFT spectral structure and shape of music signals. In addition, they are complementary in classification. Therefore, we group them to be one feature view and concatenate them to be a long feature vector. We then take DWT based features to be another feature view. In fact, the two feature views correspond to the two most popular feature sets used in musical genre classification.

Note that to make the concatenation more reasonable, it is better for different feature sets to have the same scales in the resulting feature vector[9]. To achieve this, PCA is first applied to each feature sets respectively to compute the eigenvectors U^j and eigenvalues λ_i. Then for each music pieces, project each kind of feature vector V^j to the eigenvectors and normalize them by the sum of eigenvectors

$$\overline{V^j} = U^j V^j / \sqrt{\sum \lambda_i^j}. \qquad (5)$$

2.3. Co-Training Algorithm

In a real system of musical genre classification, due to the restriction of large amount of music pieces, it is unrealistic for extensive manual annotations. Therefore, an algorithm of training with only a few labelled samples is required. In addition, the division of feature sets inspires us to use semi-supervised learning algorithm using multi-view features. In this paper, we adopt the Co-Training algorithm. This algorithm is originally used in semi-supervised classification of web-page data[1]. It is designed to improve the performance of a learning machine with a few labelled samples aided by large amount of cheap unlabelled samples. Let V_1 and V_2 be two different feature views, L be the labelled samples, T be their labels, and U be the unlabelled samples. The algorithm works as Table 1.

Table 1. Flow chart of Co-Training Algorithm

```
1. Iterate the following step until all samples are labelled.
(1)Train Classifiers $h_1$ and $h_2$ with $L(V_1), T$ and $L(V_2), T$, respectively.
(2)For each classifier $C_i$,
   ·Classify all unlabelled data by $h_1$ and $h_2$.
   ·Let $E_1$ and $E_2$ be the samples that most likely belong to $C_i$. Remove $E_1$ and $E_2$ from $U$ and put them in $L$.
2. Classify according to the decision by $h_1$ and $h_2$.
```

It has been proven that if the feature subsets satisfy the assumptions of compatibility and conditional independence, the number of labelled data used by learning machine can be reduced. In addition, in case of PAC learnable problems, any weak classifier can improve its accuracy to any high level with Co-Training algorithm and unlabelled data. Although the assumptions for analysis are strong, the algorithm still performs well on real data.

3. EXPERIMENTS AND RESULTS

In this section, experiments are conducted to evaluate and validate the proposed musical genre classification method.

The music database used in this experiments contains 300 music pieces which cover 3 different genres, including classical, pop and metal. There are 100 pieces in each class and each piece is 30 seconds in length. All data are 22050 Hz sample rate. After features are extracted from each music piece, they are split into two feature views, which are 19 dimensional (STFT based + MFCC) and 35 dimensional (DWT based), respectively. Then Co-Training algorithm using the classifier of nearest neighborhood is adopted. We label randomly only a few pieces in each class. The following results are the average of 50 runs. In Figure 1, classification accuracy of Co-Training algorithm using multi-view feature sets with different number of annotations is presented. It can be seen that our method can achieve a satisfactory result in musical genre classification, even with only a few annotations.

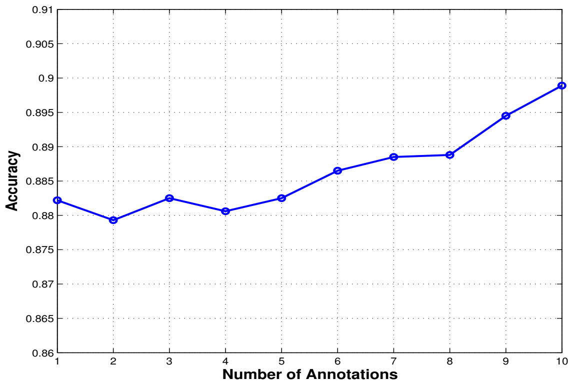

Figure 1. Classification accuracy of Co-Training algorithm using multi-view feature sets with different number of annotations.

To further illustrate the advantage of the proposed method, especially in case of very few number of annotations, we compare our method with some other methods, including nearest neighborhood (NN), SVM, LDA and Hastie LDA, using different feature sets. Here, we label randomly only 1 musical piece in each class. This is in accord with the situation in real systems: the musical database is so large that only very few musical pieces can be labelled. Table 2 gives the comparison result among these methods. It can be seen that in case of very few annotations, our method achieves higher accuracy than other methods. Table 3 shows the confusion matrix of the proposed method in this case.

Table 2. Comparison results

	KNN	SVM	LDA	Hastie LDA	Co-Training
STFT+MFCC+DWT	81.38	81.38	83.75	83.84	**88.05**
STFT+MFCC	83.04	83.04	83.74	83.50	
STFT	82.80	82.80	82.21	82.90	
MFCC	65.37	65.37	63.66	66.89	
DWT	71.88	71.88	71.16	70.02	

Table 3. Confusion matrix of the proposed method

	Classical	Metal	Pop
Classical	0.9192	0.0505	0.0303
Metal	0.0404	0.8586	0.1010
Pop	0.0404	0.0909	0.8687

4. CONCLUSION AND FUTURE WORK

In this paper, we have presented an automatic semi-supervised classification method for musical genres using multi-view features. In order to make full use of the discriminative information contained by music pieces, we combine different feature sets, including STFT based features, MFCC features, DWT based features. These features are split into two feature views. We then adopt Co-Training algorithm to classify musical genres with only a few annotations. Experimental results demonstrate the validity of this method.

There are several problems need to be investigated in the future. The first one is to test this method on larger musical database. The second is try to combine more features to form more feature views and the third is to test other semi-supervised algorithms, especially algorithms using multi-view features, such as Co-EM, Co-Boosting, etc.

Acknowledgements
The authors would like to thank Dr. George Tzanetakis
for his kind help in sharing his musical database.

5. REFERENCES

[1] A. Blum, T. Mitchell. ”Combining Labeled and Unlabeled Data with Co-Training”, *COLT98:Proceedings of the Workshop on Computational Learning Theory*, 1998.

[2] C. Xu, N.C. Maddage, X. Shao, F. Cao, Q. Tian. ”Musical Genre Classification using Support Vector Machines”, *ICASSP03:IEEE International Conference on Acoustics, Speech and Signal Processing*, 2003.

[3] G. Tzanetakis, P. Cook. ”Musical Genre Classification of Audio Signals”, *IEEE Transaction on Speech and Audio Processing*, Vol. 10, No. 5, Jul, 2002.

[4] G. Tzanetakis. ”Manipulation, Analysis and Retrieval System for Audio Signals”, *Ph. D Thesis*, Jun, 2002.

[5] H. Deshpande, R. Singh, U. Nam. ”Classification of Music Signals in the Visual Domain”, *Proceedings of the COST G-6 Conference on Digital Audio Effects(DAFX-01)*, Dec, 2001.

[6] J. Aucouturier, F. Pachet. ”Representing Musical Genre: a State of the Art” *Journal of New Music Research*. Vol. 32, No. 1, 2003.

[7] M. Grimaldi, P. Cunningham, A. Kokaram. ”An Evaluation of Alternative Feature Selection Strategies and Ensamble Techniques for Classification Music”, *The 14th European Conference on Machine Learning (ECML) and the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, 2003.

[8] T. Li, G. Tzanetakis. ”Factors in Automatic Musical Genre Classification of Audio Signals”, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct, 2003.

[9] X. Wang, X. Tang ”Using Random Subspace to Combine Multiple Features For Face Recognition”, *the Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 2004.