MUSIC GENRES CLASSIFICATION USING TEXT CATEGORIZATION METHOD

Kai Chen, Sheng Gao, Yongwei Zhu, Qibin Sun
Institute for Infocomm Research, 21 Heng Mui Keng, Singapore
{kchen,gaosheng,ywzhu,qibin}@i2r.a-star.edu.sg

ABSTRACT

Automatic music genre classification is one of the most challenging problems in music information retrieval and management of digital music database. In this paper, we propose a new framework using text category methods to classify music genres. This framework is different from current methods for Music genre classification. In our framework, we consider music as text-like semantic music document, which is represented by a set of music symbol lexicons with a HMM (Hidden Markov Models) cluster. Music symbols can be seemed as high-level features or semantic features like beats or rhythms. We use latent semantic indexing (LSI) technique that is widely adopted in text categorization for music genre classification. From the experimental results, we could achieve an average recall over 70% for ten musical genres.

1. INTRODUCTION

Music is very popular in modern life, and the amount of digital music increases rapidly nowadays. How to manage the large digital music database has arisen as a crucial problem. Automatic music type classification could be very helpful for managing the music database. Machine learning and pattern recognition techniques, successfully employed in many tasks, can also be applied to music analysis. One of the tasks is to manage the music according the genre. Much work has been reported on music genre classification using the audio acoustic signal or symbolic signal such as MIDI music [2, 3, 4, 5, 9].

Tzanetakis and Cook [3] explore a set of novel Fourier-based feature extraction techniques for musical genre classification using the K-Nearest Neighbors and Gaussian Mixture models. Li [9] achieves superior performance over Tzanetakis using wavelet-based feature and Support Vector Machines (SVM) on the same dataset. Soltau [4] proposes ETM-NN (Explicit Time Modeling with Neural Network) method using the abstraction of acoustical event to the hidden units of a neural network for music types recognition. Mandel and Ellis [5] uses Gaussian Mixtures, KL divergence and SVM to classify music based on MFCC (Mel Frequency Cepstral Coefficients) features.

Currently, the most influential approaches to direct modeling of music signals for automatic genre classification adopt timbre texture, rhythm, and pitch content features. But these features cannot capture the global statistics and long-term dependency among the musical events. We believe they are helpful for distinguishing the genres. In text categorization, many techniques have been exploited to addressing the above issues. It should be helpful to apply the techniques for text categorization to music genre classification? Our goal in the paper is to combine current music genre classification methods and text categorization techniques for automatic music genre classification. The most related work is Soltau [4]. But Soltau only uses 10 music events and simple unigram-counts, bigram-counts, and trigram-counts of the events and other statistics such as the event activation as the features to represent the content of music signal.

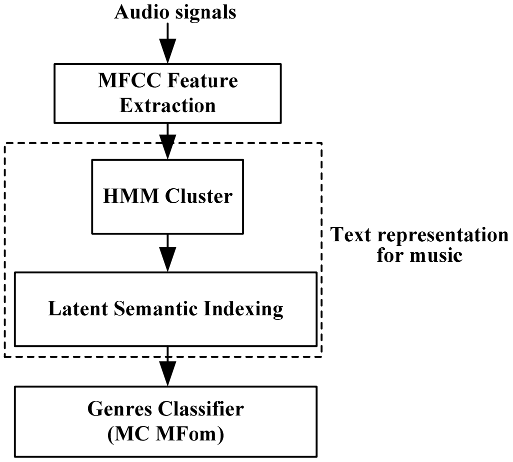

Figure 1 Flowchart of music genre classification

Our framework is presented in Figure 1. The framework includes two main parts: text representation for music content and multi-class classification. In the text representation for music content, firstly, music is represented by a set of music symbols. Music symbols are derived using a HMM based clustering technique. After that, the music symbol sequence is converted into the latent semantic indexing (LSI) based features and then Multi-classes Maximal Figure-of-Merit (MC MFoM) is applied to train multi-class classifier [6]. Our experiments show that our algorithm performs better than traditional GMM schemes for music genre classification.

0-7803-9752-5/06/$20.00 ©2006 IEEE

The paper is organized as follows: Section 2 discusses text representation for music content in detail. Our classification scheme is described in Section 3. In Section 4, the experiments were performed to evaluate our framework.

2. TEXT REPRESENTATION FOR MUSIC CONTENT

In information retrieval (IR), a multidimensional feature set is often extracted to represent a text document. Each component of the feature set corresponds to the contribution of a term occurred in the document. According to this, we convert music signals to music symbol sequences and apply music semantic description. Each piece of music can be considered as a text-like semantic music document and latent semantic indexing is adopted for music genre classification. The flowchart of text representation is shown in Figure 2.

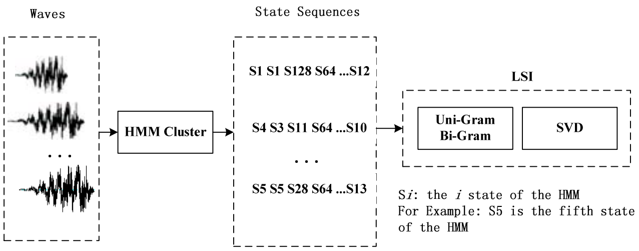

Figure 2 Flowchart of text representation for music

2.1. Music Symbols

Our framework for music genre classification is based on text categorization. The key point is to establish a set of music symbols equivalent to the words in texts, and then the music signal can be regarded as a text-like music document. We assume that each music signal has its special temporal structure composing of different temporal components. Similar temporal components can appear in different music of different genres. Machine learning approaches are possible to uncover such structures. We call these similar temporal components as music symbols which are seemed as high-level features or semantic features such as beats or rhythms. Our experiments and [4] tested that it is a good way to find these music symbols in an unsupervised manner.

HMM is a powerful framework for learning and recognizing the temporal patterns and has been applied in the music summary based on temporal information [7], thus we adopt HMM to explore music symbols in our framework. In our case, one ergodic HMM with each state modeled by Gauss Mixture Model (GMM) is trained on all training dataset. Each state will correspond to a group of similar temporal segments in the audio songs and then each state of the HMM can be seemed as a music symbol. Given the sequence of cepstral features (MFCC) for music, the unsupervised Baum-Welsh algorithm [8] is used to train the HMM. After training, we use Viterbi decoding to determine the most likely state sequence for music. By interpreting the Viterbi alignment results, we can infer temporal structures of songs. In our experiment, the number of states is 128 and each state is modeled by three Gaussian models.

The advantage of using HMM for our purpose is as follows: First, mixture of GMM gives an accurate probabilistic description of music data. Second, the temporal relation between each state and the probability of their occurrence also impose the strong constraints in terms of temporal order of each state and the transitions between them. Comparing with other clustering methods such as k-means, HMM clusters data on the temporal level rather than the frame level.

2.2. Latent Semantic Indexing

When music symbols are obtained, each piece of music can be tokenized using a set of music lexicon for music representation. Like preprocessing of text categorization, the music lexicon can be constructed using each distinctive music symbol, the intra-lexicon statistical association, e.g. unigram and bi-gram which be extracted to describe the co-occurrence of the intra-lexicon music terms, term probabilities and weightings.

Due to the high number of potential n-grams, the dimension of the vector for music document is quite high. It is necessary to reduce the dimension. LSI is a good technique to achieve both feature selection and reduction.

LSI [1] relies on the constituent terms of a document to suggest the document semantic content. It assumes that the variability of word choice partially obscures the semantic structure of the document. By reducing the dimensionality of the term-document space, the underlying, semantic relationships between the documents are revealed, and much of the “noise” (differences in word usage, terms that do not help distinguish documents, etc.) is eliminated. LSI statistically analyses the patterns of word usage across the entire document collection, placing documents with similar word usage patterns near each other in the term-document space, and allowing semantic-related documents to be near each other even though they may not contain common terms.

To build the LSI model, a matrix representation of training music document is created first. For instance we have the following two (short) music documents:

D1 = "S1 S2 S4"

D2 = "S1 S3 S3 S4"

Table 1 shows that documents contain which terms and how many times they appear.

Table 1 A document-term matrix

	S1	S2	S3	S4
D1	1	1	0	1
D2	1	0	2	1

Each entry in the matrix can be a weighted frequency by a function that expresses both the music lexicon's importance in the particular music document and the degree to which the word type carries information in the domain of discourse in general. Next, LSI applies singular value decomposition (SVD) to reduce this large sparse matrix into a compressed matrix.

$$M = USV \qquad (1)$$

The original matrix M is decomposed into a reduced rank term matrix U, a diagonal matrix of singular value S and a document matrix V. The row vector of matrix U and the column vector of matrix V are the projections of word vectors and document vectors into singular value space. Thus the words and documents are represented in a compact way compared to the original. Depending on the different tasks, the number of selected singular value varies. 20 to 400 are typical choices in many regular tasks. 300 is selected in our experiments.

3. MC MFOM CLASSIFIER

After music is represented by a set of music lexicons, we can extract a feature vector to describe its content. Then we will discuss the MC MFoM learning [6] for classifier design based on this representation. Different from the conventional techniques, MC MFoM method attempts to integrate any performance metric of interest (e.g. accuracy, recall, precision, or F1 measure) into the design of any classifier. The corresponding classifier parameters are learned by optimizing an overall objective function of interest.

Given N genres, $c = \{c_j, 1 \le j \le N\}$, and a training set, $T = \{(X,Y) \mid x \in R^D, Y \subset C\}$, where C_j is the j-th genre, with X being a sample in a D-dimensional space, Y being a subset of C, representing the set of labels of X. N-category classifier with the parameter set, $\Lambda = \{\Lambda_j, 1 \le j \le N\}$, is estimated form T. If the discriminant function for the j-th genre is $g_j(X;\Lambda_j)$, then the decision rule is as

$$\begin{cases} \text{Accept } X \in C_j \text{ if } g_j\left(X;\Lambda_j\right) - g_j^-\left(X;\Lambda^-\right) > 0 \\ \text{Reject } X \in C_j, \textit{Otherwise}, \end{cases} \quad 1 \le j \le N (2)$$

Where $g_j^-\left(X;\Lambda^-\right)$ is the competitive model for the j-th genre, named class anti-discriminant function, defined as,

$$g_j^-\left(X;\Lambda^-\right) = \log\left[\frac{1}{\left|C_j^-\right|}\sum_{i \in C_j^-} \exp\left(g_i\left(X;\Lambda_i\right)\right)^\eta\right]^{1/\eta} \qquad (3)$$

Where C_j^- is a subset containing the most competitive genres against C_j, $|C_j^-|$ its cardinality, Λ^- the parameter set for the competitive genres, and η a positive constant.

$g_j(X;\Lambda_j)$ describes the similarity between the sample X and j-th genre, while Eq. (3) measures the score from the competing categories which functions as a negative model. The main idea for the MC MFoM learning is to design a smoothing objective function for optimization, which function will embed any preferred performance metric and any type of classifier.

4. EXPERIMENTAL RESULTS

4.1. Datasets

We used the same dataset used in [3] to evaluate our algorithm. The dataset contains 1000 songs over ten genres with 100 songs per genre. The ten genres are *Blues, Classical*, *Country*, *Disco*, *Hiphop*, *Jazz*, *Metal*, *Pop*, *Reggae* and *Rock*. To ensure variety of different recording qualities the excerpts were taken from radio, compact disks, and MP3 compressed audio files. The length of each song is 30s and each song is stored as 16k Hz, 16-bit, mono.

4.2. Experimental Results

The classification results are calculated using a 3-fold cross-validation evaluation where the dataset to be evaluated is randomly partitioned so that 70% is used for training and 30% for testing. GMM classifier [3] is simple and successful in audio genre classification. We use it to compare with our proposed algorithm. For each class, we assume the existence of a probability density function expressible as a mixture of multi-dimensional Gaussian distributions. GMM with 32 mixture components are applied in our experiments and Expectation Maximization algorithm is used to estimate the parameters of GMM model with MFCC features. Each clip in the testing set is classified into the class that has the highest probability density according to Bayesian criterion.

Table 2 and Table 3 show results of our method and GMM in the form of a confusion matrix. In the confusion matrix, the rows correspond to the actual genre and the columns to the predicted genre. For example, in Table 2, the cell of row 1, column 3 with value 13.3 means that 13.3% of the *Blues* music was wrongly classified as *Country*. The percentages of correct classification lie in the diagonal of the confusion matrix. From these two tables, it is clearly seen that the proposed methods outperform GMM method. The average of recall, 70.1%, is better than GMM, 62.7% and 61% of [3] which used KNN, GMM as classifiers.

From the tables, *Classical, Metal* and *Reggae* can be easily classified by our method and GMM. Comparing to the result with GMM, most of genres have an improvement, 10% is improved for *Hiphop* and *Pop*. It seems that *Rock*, *Country* and *Blues* are more difficult to discriminate. According to WIKIPEDIA [10], *County* has much different music structure and the temporal structures of music in this genre have many kinds, which will be confused with other genres. *Rock* has the worst classification accuracy and is easily confused with other genres because of its broad nature [3]. *Blues* is a vocal and instrumental form of music based on a pentatonic scale as well as a characteristic twelve-bar chord progression. The scale and chord may be most important for *Blues*, so we will consider combining other features in future.

5. CONCLUSION

In this paper, we investigate that text categorization technologies for music genre classification on Wave format not MIDI format has been tested. The scheme provides a flexible framework based on LSI and MC MFoM. Using the proposed scheme classification of 70.1% has been achieved in a dataset consisting of ten musical genres, better than the results obtained using the popular music genre classification such as GMM. In future, we will try to use other methods and low-level features to extract music lexicon as high-level features to evaluate our scheme.

6. REFERENCES

[1] C. H. Papadimitriou, P. Raghavan, H. Tamaki, S. Vempala "Latent semantic indexing: A probabilistic analysis", PODS, 1998

[2] Carlos Pérez-Sancho, José Manuel Iñesta Quereda, Jorge Calera-Rubio, "A Text Categorization Approach for Music Style Recognition", IbPRIA (2), pp 649-657, 2005

[3] George Tzanetakis and Perry Cook, "Musical genre classification of audio signals", IEEE Transactions on Speech and Audio Processing, 10(5):pp293–302, 2002.

[4] H. Soltau, T. Schultz, M. Westphal and A. Waibel., "Recognition of Music Types", ICASSP, 1998

[5] M. Mandel, D. Ellis, "Song-Level Features and Support Vector Machines for Music Classification", ISMIR-05, London, 2005

[6] Sheng Gao, Wen Wu, Chin-Hui Lee, Tat-Seng Chua, " A MFoM learning approach to robust multiclass multi-label text categorization", ICML, 2004

[7] Stephen Chu Beth Logan, "Music summary using key phrases", http://www.hpl.hp.com/techreports/Compaq-DEC/CRL-2000-1.pdf

[8] Steve Young, Phil Woodland and Gunnar Evermann, HTK Book, http://htk.eng.cm.ac.uk, 2005

[9] Tao Li, Mitsunori Ogihara, and Qi Li, "A comparative study on content-based music genre classification", SIGIR '03, pp. 282–289, 2003,

[10] WIKIPEDI, www.wikipedia.org

Table2 Genres Confusion Matrix of GMM

	Blues	Classical	Country	Disco	Hiphop	Jazz	Metal	Pop	Reggae	Rock
Blues	56.7	0	13.3	3.3	0	3.3	16.7	0	0	6.7
Classical	0	86.7	6.7	3.3	0	3.3	0	0	0	0
Country	6.7	0	40.1	13.3	0	3.3	13.3	13.3	0	10
Disco	6.7	0	0	60	0	0	0	20	6.7	6.6
Hiphop	0	0	0	3.3	76.8	0	3.3	3.3	13.3	0
Jazz	3.3	13.3	16.7	3.3	0	53.4	3.3	0	0	6.7
Metal	0	0	0	0	0	0	86.7	3.3	0	10
Pop	3.3	0	3.3	3.3	16.7	3.3	10	60.1	0	0
Reggae	0	0	3.3	3.3	6.7	0	0	0	86.7	0
Rock	3.3	0	40	13.3	3.3	0	16.7	3.3	0	20.1

Table3 Genres Confusion Matrix of Text Category Method

	Blues	Classical	Country	Disco	Hiphop	Jazz	Metal	Pop	Reggae	Rock
Blues	53.4	3.3	20	0	0	3.3	13.3	0	6.7	0
Classical	0	93.3	6.7	0	0	0	0	0	0	0
Country	10	0	50.1	6.7	0	3.3	23.3	3.3	0	3.3
Disco	6.7	0	0	60	0	0	0	20	6.7	6.6
Hip-hop	0	0	3.3	0	86.8	0	3.3	3.3	3.3	0
Jazz	0	13.3	0	3.3	0	76.7	6.7	0	0	0
Metal	0	0	0	0	0	3.3	93.4	3.3	0	0
Pop	0	0	3.3	3.3	13.3	3.3	0	70.1	0	6.7
Reggae	0	0	3.3	3.3	6.7	0	0	3.3	83.4	0
Rock	6.7	6.7	16.7	10	0	0	13.3	3.3	10	33.3