A NOVEL APPROACH TO MUSICAL GENRE CLASSIFICATION USING PROBABILISTIC LATENT SEMANTIC ANALYSIS MODEL

Zhi Zeng, Shuwu Zhang, Heping Li, Wei Liang and Haibo Zheng

Institute of Automation, Chinese Academy of Sciences, Beijing, China
{zzeng, swzhang, hpli, wliang, hbzheng}@hitic.ia.ac.cn

ABSTRACT

A novel approach based on the probabilistic latent semantic analysis model (pLSA) for automatic musical genre classification is proposed in this paper. Unlike traditional usage, the pLSA is used to model musical genre instead of single music signal in the proposed approach. First, an unsupervised clustering algorithm is utilized to group temporal segments in music signals into several natural clusters. By this means, each music signal is decomposed into a bag of "audio words". Subsequently, the pLSA model of each musical genre is trained through a new iterative training procedure and well-known EM algorithm. This training procedure can iteratively update the pLSA model parameters by discriminatively computing weight of each training music signal and evidently improve the model's discriminative performance. Finally, these models can be used to classify new unseen music signals. Experiments on two commonly utilized databases show that our pLSA based approach can give promising results and the iterative learning procedure is effective.

Index Terms—Musical genre classification, pLSA, MFCC

1. INTRODUCTION

In recent years, due to the increasing growth of network bandwidth and computer storage, there has been a rapid proliferation of digital music database. How to effectively manage, classify and retrieve the large digital music database has arisen as a crucial problem.

Musical genres are labels created and used by humans for categorizing and describing the vast universe of music. Automatic musical genre classification could be very helpful for managing the music database. However, musical genre is sometimes ambiguous. In general, it is agreed that audio signals of music belonging to the same genre should contain some common characteristics. These common characteristics have motivated recent research activities to classify music into genres automatically.

Automatic musical genre classification can be divided into two stages: feature extraction and classifier design. Several feature extraction methods have been developed. In [1], features of timbral texture, rhythmic content and pitch content are thoroughly investigated by Tzanetakis and Cook. E. Pampalk [2] presented three music similarity measures: MFCC-EMD, Fluctuation Pattern and Spectrum Histogram. These measures have also been used by Y. Song *et al.* [3]. Another novel feature extraction method is proposed in [4], in which local and global information of music signals are captured by computation of histograms on their Daubechies wavelets coefficients. In recent, A. Holzapfel *et al.* [5] present a novel feature set to classify musical genres by computing a Non-negative Matrix Factorization (NMF) on spectrograms of music signals.

While there have been various feature extraction methods, many classifiers are also employed for automatic musical genre classification, such as K-nearest neighbor (KNN) and Gaussian mixture model (GMM) classifier [1], [4], AdaBoost [6], radial basis function (RBF) [7], support vector machine (SVM) [8] and semi-supervised method [3].

A comprehensive survey about this area can be found in [16]. As mentioned above, several classifiers have been used to classify musical genre. However, none of them can give perfect results, seeking of other models to classify musical genre is still required. The Probabilistic Latent Semantic Analysis (pLSA) model proposed by Hofmann [9], a generative model from the statistical text literature, uses a probabilistic model and the expectation-maximization (EM) method for text classification and other applications. In recent years, this model has been applied to several other fields, such as scene classification [10], human action category [11] and video classification [12].

In this paper, we apply pLSA model to classify musical genres. Our method is a type of bag-of-word method and includes two parts: unsupervised clustering and multi-class classification. In clustering, music signal is decomposed into a bag of "audio words" by k-mean clustering algorithm. Then, unlike other bag-of-word classification methods, we combine all the music signals with the same genre in the training set to form an "audio document". PLSA model is used to model these "audio documents" which are on behalf of genres. A new music signal is classified by considering its similarity to those "audio documents".

The paper is organized as follows: Section 2 describes the audio feature calculation frontend. Our proposed approach is represented in Section 3. In Section 4, the experiments were performed to evaluate our approach. Finally, we summarize our work in Section 5

2. FEATURE EXTRACTION

The music signal is first down-sampled at 16000 Hz and 16 bits/sample, and divided into frames of 25 ms with 50% overlap. Short-time analysis is performed over these frames and Hamming window is applied to remove edge effects. In this study, the first 20 MFCC (Mel Frequency Cepstrm Coefficients) [13] coefficients (expect the 0th coefficient) and its first-order derivative are extracted from these frames. All these features are collected into a 38-dimensional feature vector per audio frame.

In order to reduce the computational complexity of the proposed approach, we choose to group audio frames into longer temporal audio segments, and to use these longer segments as the basis for the subsequent processing steps. A sliding window of 1 s with 0.5-s overlap is used to segment the frame sequence. At each window position, the mean and standard deviation of the frame-based features (76-dimensional) are computed and used to represent the corresponding one-second-long audio segment.

3. PROPOSED APPROACH

In our framework for musical genre classification, each music signal is represented as a bag of "audio words" firstly by using clustering algorithm. Subsequently, pLSA model is utilized to model each genre. Classification of unseen music similarly proceeds in two stages: First, the test music is also decomposed into "audio words". Then, its most similar genre is gained by using pLSA model. The detail of the proposed approach will be given in the following subsection.

3.1. Audio Words

After audio feature extraction, we need to group the feature vectors extracted from the training set into V clusters. V is the total number of cluster. In order to group similar feature vectors into V clusters, which are adopted as "audio words", we use the traditional k-means clustering algorithm, meanwhile, the Euclidean distance is adopted to measure the distance between two feature vectors.

By using clustering algorithm, every music signal in the training set could be represented as a bag of "audio words". Suppose we have a collection of music signals $S = s_1, ..., s_N$ with words from "audio vocabulary" $W = w_1, ..., w_V$, where N is total number of music signals in training set and V is total number of "audio words" in the "audio vocabulary". The training set could be denoted by a $V \times N$ co-occurrence table of counts $T'_{ij} = n(w_i, s_j)$, where $n(w_i, s_j)$ denotes how often the word w_i occurred in a music signal s_j. And a new unseen test music signal could also be represented as a word index by choosing the nearest cluster center. This means that a test music signal could be represented by a V×1 vector of counts $T''_i = n(w_i, s_{test})$.

3.2. PLSA Model

The probabilistic Latent Semantic Analysis (pLSA) model [9] is a latent variable model for co-occurrence data which associates an unobserved topic variable $z \in Z = z_1, ..., z_K$ with each observation, an observation being the occurrence of a word in a particular document.

Let us introduce the following probabilities: $P(d_j)$ is used to denote the probability of observing a particular document d_j. $P(w_i|z_k)$ denotes the conditional probability of a specific word conditioned on the unobserved topic variable z_k, and finally $P(z_k|d_j)$ denotes a document specific probability distribution over the latent variable space. Using these definitions, one may define a generative model by the following scheme:

- Select a document d_j with probability $P(d_j)$,
- Pick a latent class z_k with probability $P(z_k|d_j)$,
- Generate a word w_i with probability $P(w_i|z_k)$.

As a result, one obtains an observation pair (d_j, w_i), while the latent topic variable z_k is discarded. A representation of the pLSA model in terms of a graphical model is depicted in Figure 1.

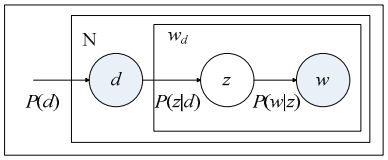

Fig. 1. Graphical model representation of the pLSA model.

Translating the data generation process into a joint probability model results in the expression:

$$P(w_i, d_j) = P(d_j) \sum_{k=1}^{K} P(w_i \mid z_k) P(z_k \mid d_j) \qquad (1)$$

As described in [9], the pLSA model is fitted by using the Expectation Maximization (EM) algorithm. The goal is to determine the model that gives high probability to the distribution of words that appear in the corpus.

3.3. Musical Genre Classification

Conventionally, we can interpret a music signal as a document in the definition of pLSA model, and use the training music set to train this model. However, the pLSA model has a problem that the number of parameters which must be estimated grows linearly with the number of

training documents. The parameters for a K-topic pLSA model are K multinomial distributions of size V and N mixtures over the K hidden topics. This gives $KV+KN$ parameters and therefore linear growth in N. The linear growth in parameters suggests that the model is prone to overfitting, and it's a serious problem.

To overcome this problem, we treat the entire music signals with the same genre in the training set as a document in the pLSA model. Let M be the total number of musical genres, and we have a collection of musical genres $D = d_1, ..., d_M$. Then, the training set could be denoted by a $V \times M$ co-occurrence table of counts $T_{ij} = n(w_i, d_j)$, where

$$n(w_i, d_j) = \sum_{s_k \in genre\ j} a_k n(w_i, s_k) \quad (2)$$

where a_k is the parameter which control the importance of music signal s_k in genre j. It is initially set to 1 and its update rule will be introduced later.

Then the co-occurrence table of counts T will be used to train the pLSA model by utilizing EM algorithm. After the training process, each musical genre will be represented by a V-dimensional vector ($P(w_1|d), ..., P(w_V|d)$), where

$$P(w \mid d) = \sum_{z \in Z} P(w \mid z) P(z \mid d) \quad (3)$$

After the training process, the estimated $P(w|z)$ parameters are used to estimate new parameters $P(z|s)$ and $P(z|w, s)$ for a new test music signal s_{test}. This is called the folding-in process [9]. The new test music signal s_{test} is also represented by a V-dimensional vector ($P(w_1|s), ..., P(w_V|s)$), where

$$P(w \mid s) = \sum_{z \in Z} P(w \mid z) P(z \mid s) \quad (4)$$

Then, the similarity between the new test music signal and any musical genre can be calculated by the following cosine function:

$$sim(d, s_{test}) = \frac{P(w \mid d) \bullet P(w \mid s_{test})}{\left|P(w \mid d)\right| \times \left|P(w \mid s_{test})\right|} \quad (5)$$

The most similar genre can be assigned to be the new test music's genre.

Since the training music signals with same genre can not equally represent their genre's character, it's not proper to set the parameters a_k equally. To improve the model's discriminative performance, we use the following steps to update the parameters a_k and train the pLSA model:

1) Set $a_k = 1$ for $k = 1, \ldots, N$.
2) Perform the above training process.
3) Perform the above genre identification task over the training data. If the error rate is less than a predefined value or some stopping condition is satisfied, the training is completed, else proceed.
4) Set $a_k = a_k + \delta$, where training music signal s_k can not be right classified. Then goto 2).

Increasing the weights for songs that could not be classified can make the model more effective on the training set, but this also increases the weights for outliers. Therefore, smart choice of parameter δ to reach a compromise is important. In our experiments, we find that good results can be given when the parameter δ is set to 1.

Following these steps, the pLSA model for musical genre classification is well trained. Supportive experimental results that show the improvements in the classification are presented in next section.

4. EXPERIMENTS

4.1. Databases

For the experiments, two different data sets have been used. The first database (D1) consists of ten classes[1], each containing 100 subsections of musical pieces of 30 seconds length. This database was collected by G. Tzanetakis [1] and has been used for performance evaluation by other researchers [6], [7]. The second database (D2) was downloaded from the website of the ISMIR contest in 2004 [14]. There are 729 training songs and 729 development songs, which are classified into six genres[2]. And they are not equally distributed among the classes as they are in D1. For D1, a five-fold cross validation has been used. For D2, we can use the training songs to train our model, and the development songs are used for evaluation. However, the second database (D2) is well known to be heavily biased due to the artist effect [15]. Especially when using features derived from MFCCs, this should be taken into account. To avoid this problem, we also run a five-fold cross validation on the training set of D2 and split the original training set into a new test and training set where the same artist does not occur in both.

4.2. Classification Results

In order to evaluate the performance of our classification approach it is necessary to compare with some types of standard procedure used in many publications. For this, we use the GMM classifier to compare with our proposed method. This model is widely used in the field of musical genre classification [1], [4], [5]. In our experiments, we applied GMMs with 20 and 30 mixture components and diagonal covariance matrices.

In our classification approach, there are several parameters needed adjust: the number of "audio words" (V) and the number of latent topics (K). In the experiments, we set (V, K) to (3000, 30), (2000, 20) and (1200, 20) respectively. To show the improvements of iterative training, we also test our algorithm without iterative training.

Table 1 shows the classification results on the two databases. The values in parentheses denotes the parameters setting of our method and the number of mixture components in GMM.

[1] Blues, Classical, Country, Disco, Hip Hop, Jazz, Metal, Pop, Reggae, Rock

[2] Classical, Electronic, Jazz, Metal/Punk, Rock/Pop, World

Table 1. Classification Accuracies (%)

	D1 (5 cross-validation)	D2 (testing by using development set)	D2's training set (5 cross-validation)
Our method (V=3000, K=30)	81.5	84.4	80.1
Our method (V=2000, K=20)	80.4	82.7	79.1
Our method (V=1200, K=20)	79.3	82.3	76.9
Our method without iterative training (V=2000, K=20)	79.1	81.3	75.5
Our method without iterative training (V=1200, K=20)	77.8	80.9	73.2
GMM(20)	65.3	64.5	58.8
GMM(30)	70.8	68.3	60.2

The results show that the proposed method outperforms the baseline method. This is evident for both two databases. The results also show that the iterative training certainly can improve the performance of our algorithm. Even though our approach only uses the MFCC features of music signal it also performs well in comparison with other published methods. On D1, Holzapfel and Stylianou [5] reported an accuracy of 74% while Bergstra *et al*. [6] reported 83%. On D2, the winner of the ISMIR'04 Audio Description contest reached an accuracy of 84.07% while Holzapfel and Stylianou [5] reported an accuracy of 83.5%.

5. CONCLUSION

This paper presented a novel approach to musical genre classification by using pLSA model. Unlike conventional approach, we don't treat a music signal as a document in pLSA model. By combining the entire music signal with same genre to form an "audio document" and using a proposed iterative algorithm, we have trained a pLSA model of musical genre. A new unseen test music signal can be classified by considering the similarity of smoothed version of the "audio word" frequency between it and each genre. Experiments show that our approach is effective. Future work includes the task to utilize other effective audio features.

6. ACKNOWLEDGMENTS

This work has been supported by the National Science and Technology Supporting Program of China under Grant No. 2008BAH21B03-04.

7. REFERENCES

[1] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293-302, May 2002.

[2] E. Pampalk, "A matlab toolbox to compute similarity from audio," in *Proc. Int. Conf. Music Information Retrieval (ISMIR)*, Barcelona, Spain, 2004.

[3] Y. Song and C. Zhang, "Content-based Information Fusion for Semi-Supervised Music Genre Classification," *IEEE Trans. Multimedia*, vol. 10, no. 1, pp. 145-152, January 2008.

[4] T. Li and M. Ogihara, "Toward intelligent music information retrieval," *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 564-574, June 2006.

[5] A. Holzapfel and Y. Stylianou, "Musical genre classification using nonnegative matrix factorization-based features," *IEEE Trans. Speech Aud. Process.*, vol. 16, no. 2, pp. 424-434, February 2008.

[6] J. Bergstra, N. Casagrande, D. Erhan, D. Eck and B. Kegl, "Aggregate features and adaboost for music classification," *Mach. Learn.*, vol. 65, no. 2-3, pp. 473-484, June 2006.

[7] D. Turnbull and C. Elkan, "Fast recognition of musical genres using rbf networks," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 580-584, April 2005.

[8] C. Xu, N. C. Maddage and X. Shao, "Automatic music classification and summarization," *IEEE Trans. Speech Aud. Process.*, vol. 13, no. 3, pp. 441-450, May 2005.

[9] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning*, vol. 42, no. 2, pp. 177-196, February 2001.

[10] A. Bosch, A. Zisserman and X. Munoz, "Scene classification using a hybrid generative/discriminative approach," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 712-727, April 2008.

[11] J. C. Niebles, H. Wang and F. Li, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Computer Vision.*, vol. 79, no. 3, pp. 299-318, September 2008.

[12] Z. Zeng, W. Liang, H. Li and S. Zhang, "A Novel Video Classification Method Based on Hybrid Generative/Discriminative Models," *Joint Int. Workshops on Structural and Syntactic Pattern Recognition and Statistical Techniques in Pattern Recognition (S+SSPR)*, Orlando, FL, 2008.

[13] J. Junqua, J. Haton, *Robustness in automatic speech recognition*, Kluwer Academic, Boston, 1996.

[14] *ISMIR Audio Description Contest*, 2004 [online]. Available: http://ismir2004.ismir.net/genre_contest/index.htm

[15] A. Flexer, "A closer look on artist filters for musical genre classification," in *Proc. Int. Conf. Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007.

[16] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: a survey", *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133-141, 2006.