A HIERARCHICAL GENERATIVE MODEL FOR GENERIC AUDIO DOCUMENT CATEGORIZATION

Zhi Zeng and Shuwu Zhang

Institute of Automation, Chinese Academy of Sciences, Beijing, China

ABSTRACT

In this paper, we call the pattern classification problem that consists in assigning a category label to a long audio signal based on its semantic content as Generic Audio Document Categorization (GADC). A novel generative model is proposed to describe the generic audio document categories and solve the GADC problem. This model is a four-level hierarchical model in which two latent variables "audio topic" and "audio word" are introduced in addition to the two observed variables category and audio feature. We present an iterative learning algorithm including two Expectation-Maximization (EM) cycles to estimate the model parameters and give a discriminative document weighting procedure to make the model more discriminative. Subsequently, the distribution of "audio topic" in the well-trained model is utilized to represent each generic audio document category. This is same with some bag-of-word methods. However, our method is advanced since it does not require quantizing the continuous audio features to a vocabulary of "audio words". Finally, experiment results show the effectiveness of our approach.

Index Terms—Audio content analysis, generic audio document categorization, generative model

1. INTRODUCTION

In audio content analysis, the objective of solving many problems is to automatically categorize a long audio signal into several pre-defined categories. These problems include musical genre classification (MGC) [1-3], audio-based video classification (ABVC) [4] and even spoken language identification (LID) [5]. They share a lot in common: they all focus on one kind of long audio signals which are called as audio documents (AD), such as music and spoken documents. All their goals are to categorize an AD based on its semantic content. Moreover, the category of an AD is almost unrelated to its length. In this paper, we collectively call these problems as Generic Audio Document Categorization (GADC).

GADC is a pattern classification problem that consists in assigning a predefine category label to a generic AD based on its semantic content. We emphasize the use of the word "generic" as the goal is to cope with a wide variety of categories using the same framework.

The state of the art audio classification methods can be roughly divided into two categories. The first category is the acoustic modeling, where acoustic features are modeled by popular models/classifiers [1-2, 4]. The second category is the "audio word" modeling, where an audio signal is transcribed by unsupervised clustering or phoneme recognizers and the scoring is performed on "audio word" strings, e.g., bag-of-word method for MGC [3] and parallel phoneme recognizer followed by vector space modeling (PPR-VSM) for LID [5].

Despite many reported successes, the above methods have some shortcomings when being applied in GADC. It's hard to directly model the acoustic features to categorize generic AD. The "audio word" modeling methods need a vocabulary of "audio words". However, its unsupervised construction may be unable to take AD categories into account, and supervised training of "audio word" recognizers need extra labeling and learning works.

In this paper, we propose a four-level hierarchical generative probabilistic model to categorize generic AD, in which two latent variables "audio topic" (AT) and "audio word" (AW) are used. This model doesn't need construction of a vocabulary of AW. We present an iterative learning algorithm including two Expectation-Maximization (EM) cycles to train the proposed model and a procedure of discriminative document weighting (DDW) to make the model more discriminative.

The paper is organized as follows: Section 2 formulates the GADC problem. Our model 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. PROBLEM FORMULATION

In this section, we formulate the GADC problem. We are given a labeled training set of M ADs $\{(d_1, l_1), (d_2, l_2), \ldots, (d_M, l_M)\}$ in this problem. Here, each d_i is an input AD, which consists of N_i feature vectors x_i^j, and $l_i \in L=\{1, \ldots, C\}$ is the corresponding category label. These ADs can be music signals, audio tracks of videos or spoken utterances, and class labels can be musical/video genres or languages

respectively. In the testing stage, we need assign a category label to an unknown AD q based on its semantic content.

Now we are ready to show the details of our model and how we learn its parameters.

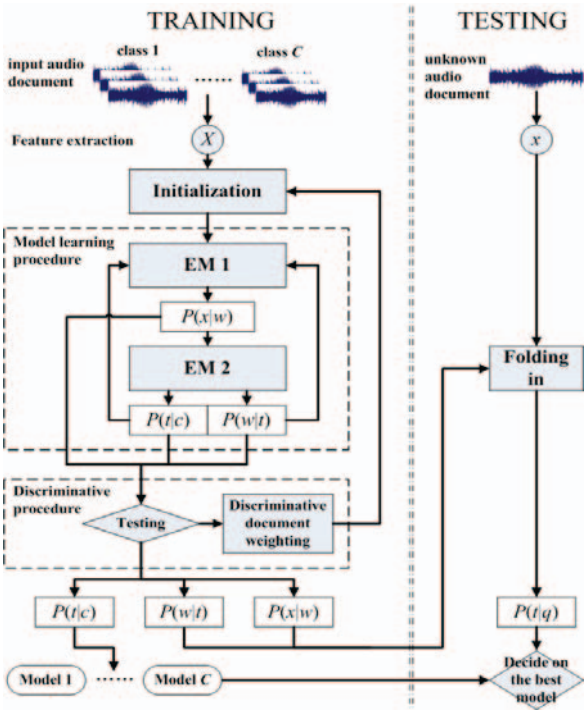

Fig. 1. Flow chart of the algorithm.

3. PROPOSED MODEL

Fig. 1 is a summary of our algorithm in both learning and categorization. The goal of learning is to achieve a model that best represents the distribution of feature vectors in each category. An iterative model learning procedure including two EM cycles and a DDW procedure is used to train the model. In categorization stage, a folding-in procedure is utilized to represent the unknown AD with the trained model, and the obtained mixing coefficients are then used to classify the test AD.

3.1. Model Structure

Now we turn to a description of the proposed generative model, whose graphical representation is shown in Fig. 2. This model is a latent variable model for co-occurrence data which associates an unobserved AT variable $t \in T = t_1, \ldots, t_T$ and an unobserved AW variable $w \in W = w_1, \ldots, w_K$ with each observation, where $T << K$. An observation is the occurrence of a feature vector in a particular category.

It is easier to understand the model by going through the generative process for obtaining a feature vector in a specific category. First, let us introduce the following probabilities: $P(c)$ is used to denote the probability of observing a particular category c, $P(t|c)$ denotes the conditional probability of a specific AT t conditioned on the category variable c, $P(w|t)$ denotes the conditional probability of a specific AW w conditioned on the unobserved AT variable t, and finally $P(x|w)$ denotes the conditional probability of a specific feature vector x conditioned on the unobserved AW variable w. Using these definitions, one may define a generative model by the following scheme:

- Select a category c with probability $P(c)$,
- Pick a latent AT t with probability $P(t|c)$,
- Pick a latent AW w with probability $P(w|t)$,
- Generate a feature vector x with probability $P(x|w)$.

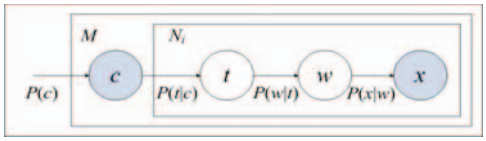

Fig. 2. Graphical representation of the proposed model.

As a result, one obtains an observation pair (c, x), while the latent AT variable t and latent AW variable w is discarded. Translating the data generation process into a joint probability model results in the expression:

$$P(c,x) = P(c) \cdot \sum_t P(t \mid c) \cdot \sum_w P(w \mid t) \cdot P(x \mid w) \qquad (1)$$

It's important to note that the conditional probabilities $P(t|c)$ and $P(w|t)$ obeys multinomial distribution, and they are discrete probability distributions. But the conditional probability $P(x|w)$ is continuous probability distributions and can be chosen to various distributions based on the specific problem. For example, we can directly model it by Gaussian distribution, or more complex, the ASM n-grams [5]. In this paper, we choose Gaussian distribution with diagonal covariance for the reason of simplicity and low computational cost. Let $\mu(x|w)$ and $\Sigma(x|w)$ denote the mean vector and diagonal covariance matrix of Gaussian $P(x|w)$ respectively.

In general we do not need to learn the prior $P(c)$ from the training data since the prior is more a property of the way the training data was collected than of the real world frequencies.

3.2. Model Learning

At the beginning of model learning, for getting rid of the influence of the different number of ADs in each category and the different number of feature vectors in each AD, we need to weight the feature vectors of training ADs. The weight of a feature vector x_i^j is defined as follow:

$$n(x_i^j) = \frac{\lambda \cdot weight(d_i)}{length(d_i) \cdot \sum_{l_k = l_i} weight(d_k)} \qquad (2)$$

where λ is a constant, $length(d_i)$ is the number of feature vectors belonged to d_i, and $weight(d_i)$ is the weight of d_i, which is set to 1 in this stage.

After feature weighting, we use the EM algorithm to maximize likelihood in the presence of latent variables. Firstly, we fix the parameters $P(w|t)$ and $P(t|c)$ and consider the optimization with respect to the parameters $\mu(x|w)$ and $\Sigma(x|w)$ governing the Gaussian distribution $P(x|w)$. This is easily maximized using EM algorithm (EM 1 in Fig. 1):

The E step is given by

$$P(w \mid x, c(x)) = \frac{P(w \mid c(x)) \cdot P(x \mid w)}{\sum_w P(w \mid c(x)) \cdot P(x \mid w)} \tag{3}$$

where $c(x)$ represents the category of the AD containing feature x, and

$$P(w \mid c(x)) = \sum_t P(w \mid t) \cdot P(t \mid c(x)) \tag{4}$$

The M step is given by

$$\mu(x \mid w) = \frac{1}{N(x \mid w)} \sum_{i=1}^{M} \sum_{j=1}^{N_i} n(x_i^j) \cdot P(w \mid x_i^j, c(x_i^j)) \cdot x_i^j \tag{5}$$

$$\Sigma(x \mid w) = \frac{1}{N(x \mid w)} \sum_{i=1}^{M} \sum_{j=1}^{N_i} n(x_i^j) \cdot P(w \mid x_i^j, c(x_i^j)) \cdot (x_i^j)^2 - \mu^2(x \mid w) \tag{6}$$

where $$N(x \mid w) = \sum_{i=1}^{M} \sum_{j=1}^{N_i} n(x_i^j) \cdot P(w \mid x_i^j, c(x_i^j)) \tag{7}$$

Next we fix the parameters of $P(x|w)$ and consider the optimization with respect to the parameters $P(w|t)$ and $P(t|c)$. At the beginning, we can count up the co-occurrence of category and AW by

$$n(c, w) = \sum_{l_i = c} \sum_{j=1}^{N_i} n(x_i^j) \cdot P(w \mid x_i^j, c) \tag{8}$$

where $P(w \mid x_i^j, c)$ is computed by equation (3). Then we can use EM algorithm to learn the parameters $P(w|t)$ and $P(t|c)$ and maximize the likelihood (EM 2 in Fig. 2).

The E step is given by

$$P(t \mid c, w) = \frac{P(w \mid t) \cdot P(t \mid c)}{\sum_t P(w \mid t) \cdot P(t \mid c)} \tag{9}$$

The M step is given by

$$P(w \mid t) = \frac{\sum_c n(c, w) \cdot P(t \mid c, w)}{\sum_w \sum_c n(c, w) \cdot P(t \mid c, w)} \tag{10}$$

$$P(t \mid c) = \frac{\sum_w n(c, w) \cdot P(t \mid c, w)}{\sum_w n(c, w)} \tag{11}$$

These two EM cycles are implemented alternately and they yield valid EM algorithm in which the likelihood never decreases. This procedure is presented in the up dashed frame in Fig. 1.

3.3. Categorization decision

After model parameters estimation, we can use the AT distribution conditioned on a category to represent this category, which means the parameter $P(t|c)$ is feature vector of category c. Given a unseen AD q, $P(t|q)$ is computed by running EM algorithm in a similar manner to that used in learning, but only $P(t|q)$ are updated and other parameters are kept fixed. Then, the intersection of $P(t|c)$ and $P(t|q)$ is computed to determine the similarity of AD q and category c, which is given by

$$sim(q, c) = \sum_t \min(P(t \mid q), P(t \mid c)) \tag{12}$$

The categorization decision is made by choosing a category which has the biggest similarity with test AD.

3.4. Discriminative document weighting

As mentioned above, we assume the weights of all training ADs equal to 1.0, which means each AD has the same influence to the model training. However, since the training ADs with same category can not equally represent their category's character, it's not proper to set their weights equally. To improve the model's discriminative power, we use the following steps to obtain better weights:

1) Set $weight(d_i) = 1.0$ for $i = 1, \ldots, M$.
2) Perform the above training process and categorization task over all the training ADs. If the error rate is less than a predefined value or some stopping condition is satisfied, the training is completed, else proceed.
3) If d_i can not be right classified, then compute the posterior probability:
$$P(c \mid d_i) = \frac{\prod_{j=1}^{N_i} P(c, x_i^j)}{\sum_c \prod_{j=1}^{N_i} P(c, x_i^j)} \tag{13}$$
4) Set $weight^{new}(d_i) = weight^{old}(d_i) + \delta \cdot (1 - P(c \mid d_i))$, where d_i can not be right classified. Then go to 2).

It's important to note that smart choice of parameter δ is important. We find that good results can be given when the parameter δ is set to 1 with about 5 iterations.

4. EXPERIMENTS

In order to evaluate the performance of our model on GADC problem, we make experiments on MGC and ABVC in this paper, and Gaussian mixture model (GMM) with diagonal covariance matrices are applied to be the baseline method. In these experiments, the GMM has 100 or 200 components and the number of AWs in our model is set to 1000 or 2000. This setting makes the evaluation fair, because there are almost 10 classes in each experiment and the total number of the GMM's components is between 1000 and 2000.

In our experiment, all the ADs are first down-sampled at 16 kHz and 16 bits/sample, and then divided into frames of 25 ms with 50% overlap. A 28-dimensional feature vector, which includes short-time energy (STE), zero-crossing rate (ZCR), the first 14 (except the 0th) MFCC and its first-order derivatives, are extracted from each audio frame. In order to reduce the computational complexity, we choose to group audio frames into longer temporal audio segments by a sliding window of 1-s with 0.5-s overlap. At each window position, the mean and standard deviation of the frame-

based features (56-dim) are obtained and used to represent the corresponding one-second-long audio segment.

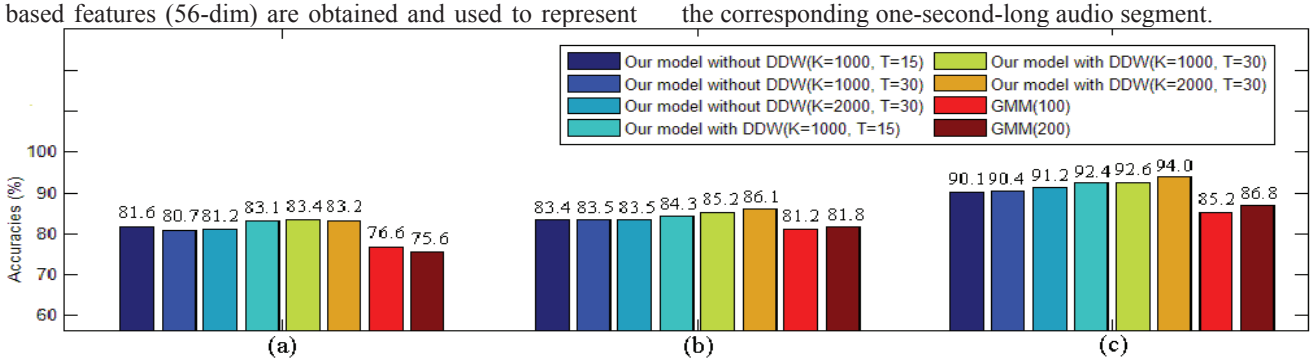

Fig. 3. Categorization Performance: (a) MGC for D1; (b) MGC for D2; (c) ABVC.

4.1. Music Genre Classification

For the experiments of MGC, two different data sets have been used. The first dataset (D1) is collected by G. Tzanetakis [1] and consists of ten classes[1]. The second dataset (D2) was downloaded from the website of the ISMIR contest in 2004 [6]. It is classified into six genres[2]. For D1, a 5-fold cross validation has been used. For D2, we use its training songs to train our model, and its development songs are used for evaluation. The results are given in Fig. 3 (a) and (b) respectively.

The results show evidently that the proposed method outperforms the baseline method and the DDW procedure improves the performance. Our method also performs well in comparison with other published methods. On D1, Holzapfel *et al.* [2] reported an accuracy of 74%, while our former work [3] reported 81.5%. On D2, the winner of the ISMIR'04 Audio Description contest reached an accuracy of 84.07% and Holzapfel *et al.* [2] reported an accuracy of 83.5%, while our former work [3] reported 84.4%.

4.2. Audio-based Video Classification

For the experiments of ABVC, we collected video sequences of TV programs containing six classes[3], each class includes 46, 14, 64, 30, 40, 106 video sequences respectively. They vary in duration from 5 minutes to 1 hour and are collected from different Chinese TV channels on different dates to ensure the variety. To evaluate the proposed method, audio tracks are extracted from these video sequences and a 2-fold cross validation is used. The results are given in Fig. 3 (c).

The experiment results show that our method works well on ABVC and outperforms the baseline method.

5. CONCLUSION

In this paper, we present a four-level hierarchical generative model to solve the GADC problem. This model doesn't require obtaining a vocabulary of AWs, while it is trained by an iterative learning algorithm and a DDW procedure.

Owing to limited space, this paper don't investigate how categorization performance is affected by the various parameters comprehensively, there is a need to perform more experiments in future work.

6. ACKNOWLEDGMENTS

This work has been supported by the National Key Technology R&D Program of China under Grant No. 2009BAH48B02, 2009BAH43B04, 2011BAH16B01 and 2011BAH16B02.

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

[3] Z. Zeng, S. Zhang, H. Li, W. Liang and H. Zheng, "A novel approach to musical genre classification using probabilistic latent semantic analysis model," in *Proc. IEEE ICME'09*, New York, USA, Jun. 2009, pp. 486-489.

[4] D. Brezeale and D. J. Cook, "Automatic Video Classification: A Survey of the Literature," *IEEE Trans. Systems, Man, and Cybernetics-Part C: Applications and Reviews*, vol. 38, no. 3, pp. 416-430, May 2008.

[5] H. Li, B. Ma and C.-H Lee, "A vector space modeling approach to spoken language identification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 1, pp. 271-284, Jan. 2007.

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

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

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

[3] News, Sports (basketball matches), TV plays, Variety shows, Talk shows, Music videos