Pattern Recognition 47 (2014) 1138–1143

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Latent topic model for audio retrieval

Pengfei Hu, Wenju Liu*, Wei Jiang, Zhanlei Yang

National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Intelligence Building 1403, Zhongguancun East Road 95, Haidian District, Beijing 100190, China

ARTICLE INFO

Available online 19 June 2013

Keywords:
Topic model
LDA
Gaussian distribution
Audio retrieval

ABSTRACT

Latent topic model such as Latent Dirichlet Allocation (LDA) has been designed for text processing and has also demonstrated success in the task of audio related processing. The main idea behind LDA assumes that the words of each document arise from a mixture of topics, each of which is a multinomial distribution over the vocabulary. When applying the original LDA to process continuous data, the word-like unit need be first generated by vector quantization (VQ). This data discretization usually results in information loss. To overcome this shortage, this paper introduces a new topic model named Gaussian-LDA for audio retrieval. In the proposed model, we consider continuous emission probability, Gaussian instead of multinomial distribution. This new topic model skips the vector quantization and directly models each topic as a Gaussian distribution over audio features. It avoids discretization by this way and integrates the procedure of clustering. The experiments of audio retrieval demonstrate that Gaussian-LDA achieves better performance than other compared methods.

1. Introduction

With the development of multimedia and network technology, more and more digital media has been emerging and the interest in content-based information retrieval of multimedia has been growing. In case of audio, traditional audio retrieval systems are text-based. However, the human auditory system predominantly relies on perception [1]. Since text captions typically only describe the higher level content, it is not possible to derive perceptual similarity between two acoustic clips. Query by example aims at solving this problem automatically. Given the example provided by user, audio samples which sound similar to example are expected from database.

Obviously, the query-by-example audio retrieval is different from audio classification. Its key issue is how to model each audio clip rather than each audio category. There exist many works for query-by-example. The most intuitive approach is to view audio clip as a whole and model it as a long term distribution of frame-based features. In [2,3] Gaussian mixture model (GMM) was used to model the continuous probability distribution of audio features. Aucouturier et al. [2] built GMM for each audio file and used the Monte-Carlo approximation of the Kulbak–Leibler distance between GMMs for similarity measurement. Besides, Helen and Virtanen [3] defined the Euclidean distance between GMMs for audio retrieval. In fact, GMM is a powerful tool capable of representing arbitrary density, but if the duration of audio file was short, which is very general in practical application, training data of GMM is often inadequate and the performance of the model is declined.

Another way for audio modeling and retrieving is histogram method, in which observation histograms are obtained by quantizing the observation values and calculating their counts within each cluster. For example, Foote [4] constructed a learning tree vector quantizer to get the histogram representation. In essence, the histogram is similar to bag-of-words model in text processing and each cluster of audio features is like each word in the dictionary. Therefore some researchers call it bag-of-words method, too. Besides audio retrieval, this representation has also recently used for copy detection [5] and acoustic event detection [6]. The histogram method is popular because of low computation. But vector quantization (VQ) would result in the loss of information, because it partitions the input space in a hard way, in which each input vector is associated only with one cluster with the nearest center. If two observations fall into different cluster bin, they are regarded as different even when they are closely spaced.

Fortunately, as the equivalent of bag-of-words in text processing, the histogram method enables content-based audio analysis and retrieval following the analogy to the proven text analysis theories and methods. Motivated by this fact, some approaches such as topic model are successfully considered in the task of audio retrieval. Sundaram and Narayanan [7] presented query-by-example for audio clips in latent perceptual space by using latent semantic index (LSI). Kim et al. [8,9] modeled the acoustic latent

* Corresponding author. Tel.: +86 1082614505; fax: +86 1062551993.
E-mail addresses: pfhu@nlpr.ia.ac.cn (P. Hu), lwj@nlpr.ia.ac.cn (W. Liu), wei.jiang@ia.ac.cn (W. Jiang), zhanlei.yang@nlpr.ia.ac.cn (Z. Yang).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.06.010

topics by latent Dirichlet allocation (LDA) and perform audio description classification and retrieval tasks.

The LDA is a generative probabilistic model, which assumes that each document consists of hidden topics and each topic in turn can be interpreted as a distribution over words in a dictionary [10]. It is a powerful method for capturing statistical properties of a collection of conditionally independent and identically distributed random variables. Besides audio retrieval, this framework has been recently used in the context of image processing [11,12] and video analysis [13,14].

It is useful to find the latent topic for audio analysis. But the LDA is designed for text processing, so building topic model for audio by standard LDA must be based on the histogram representation, in which vector quantization provides clusters as word-like units. If so, the shortcoming of the histogram method is inherited and the performance will be affected. In this paper, we propose a modified version of LDA (Gaussian-LDA) to model the latent topic in the task of audio retrieval, in which each topic is directly characterized by Gaussian distribution over audio features and so the insufficiency mentioned above can be avoid.

The paper is organized as follows. In next section, a brief overview of LDA is given. The proposed topic model is described in Section 3. Experiments and results are provided in Section 4. At last, Section 5 gives the concluding remarks.

2. Latent Dirichlet allocation

In this section, we describe latent Dirichlet allocation, which has served as a springboard for many other topic models. The idea behind LDA is to model document as arising from multiple topics, where each topic is defined as distribution over a fixed vocabulary of terms [10]. Fig. 1 illustrates the graphical representation for LDA, which is a three-level hierarchical Bayesian model [15].

Let K be a specified number of topics, V be the size of vocabulary and w be a V-dimensional vector whose elements are zero except the corresponding word index in the dictionary. A document is a sequence of N words and is represented as $d=\{w_1,w_2,..,w_i,\ldots,w_N\}$, where w_i is the i th word in the document. A corpus consists of M documents and denoted by $C=\{d_1,d_2,\ldots d_i,..,d_M\}$. LDA assumes the following generative process for each document d in a corpus C:

1. For each document d, choose θ~$Dir(\alpha)$.
2. For each word w_i in document d:
   - Choose a topic z_i~$Multinomial(\theta)$
   - Choose a word w_i with a probability $p(w_i/z_i,\beta)$, where β denotes a $K\times V$ matrix whose elements represent the probability of a word with a given topic.

Here θ is a k-dimensional probability variable corresponding to k topics and can take values in the $(k-1)$-simplex. $Dir(\alpha)$ is a Dirichlet distribution with the parameter α, which describes the probability density of θ on the simplex. The more details about Dirichlet distribution can be seen in [10].

In the application of text processing such as document classification, the bag-of-words model can be directly fed into the LDA to train and infer the parameters. In the task of audio retrieval, the application of standard LDA must be based on histogram model, which discretizes the continuous audio features and generates word-like unit by vector quantization. The details are described as follows:

a. The short-term features are first extracted for each audio clip.
b. With a given set of frame-based acoustic features, a dictionary is trained by using a vector quantization algorithm such as k-means. Each word in this dictionary corresponds to each cluster in k-means algorithm.
c. Once the dictionary is built, the extracted acoustic feature vectors from each audio clip can be mapped to acoustic words by choosing the closest word in the dictionary.
d. After extracting acoustic words, the histogram of individual audio clip can be generated by computing the word frequencies.

Based on the histogram model, the LDA parameters will be estimated by using variational inference method. In the end, the posterior Dirichlet parameters are used as final representation of the corresponding audio clips. This modeling method improves the performance of audio classification and retrieval. However, as mentioned earlier, the discretization has the innate shortcoming. Vector quantization would result in the loss of information. Aiming to overcome this defect of standard LDA for audio analysis, Gaussian-LDA will be introduced and presented in next section.

3. Gaussian-LDA

Gaussian-LDA is also built on the basic idea of topic model. As shown in Fig. 2, this topic model shares the most properties of standard LDA and only differs in the last distribution, which defines a Gaussian distribution for each topic over the audio feature data, instead of multinomial distribution over word-like unit. Given the audio document set $D=\{d_1,d_2,\ldots d_i,..,d_M\}$ with frame-based features $x_{1:N}$, the generative process based on Gaussian-LDA model can be described as follows:

1. For each audio document d, choose θ~$Dir(\alpha)$.
2. For each frame-based feature x_i in document d:
   - Choose a topic z_i~$Multinomial(\theta)$
   - Choose a frame-based feature x_i with a probability: $p(x_i/z_i,\mu_{1:K},\sigma_{1:K})$~$Normal(\mu_{z_i},\sigma_{z_i})$.

Given the parameters α, μ and σ, the joint distribution of a topic θ, a set of N topics z,and a set of frame-based features $d=x_1,x_2,..x_n..x_N$ is formulated as

$$p(\theta,z,d|\alpha,\mu,\sigma)=p(\theta|\alpha)\prod_{n=1}^{N}p(z_n|\theta)p(x_n|z_n,\mu,\sigma), \tag{1}$$

where $p(z_n|\theta)$ is simply θ_i for unique i such that $z_n^i=1$. Integrating over θ and summing over z, we obtain the marginal distribution of an audio document

$$p(d|\alpha,\mu,\sigma)=\int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}p(x_n|z_n,\mu,\sigma)\right)d\theta. \tag{2}$$

Finally, taking the product of the marginal probabilities of single audio documents, the probability of an audio corpus can be

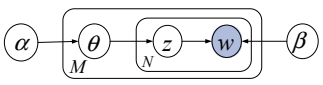

Fig. 1. Graphical model of standard LDA.

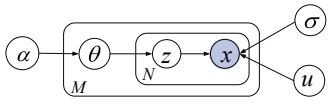

Fig. 2. Graphical model of Gaussian-LDA.