Latent Topic Model Based on Gaussian-LDA for Audio Retrieval

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

National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China

Abstract. In this paper,we introduce a new topic model named Gaussian-LDA, which is more suitable to model continuous data. Topic Model based on latent Dirichlet allocation (LDA) is widely used for the statistical analysis of document collections and other discrete data. The LDA model assumes that the words of each document arise from a mixture of topics, each of which is a multinomial distribution over the vocabulary. To apply the original LDA to process continuous data, discretization based vector quantization must be done beforehand, which usually results in information loss. In the proposed model, we consider continuous emission probability, Gaussian instead of multinomial distribution. This new topic model demonstrates higher performance than standard LDA in the experiments of audio retrieval.

Keywords: Topic model, LDA, Gaussian distribution, Audio retrieval.

1 Introduction

With the development of multimedia and network technology, more and more digital media has been emerging and the interest for content-based information retrieval of multimedia has grown. In case of audio, given the example provided by user, similar audio samples to example are expected from database. Query by example aims at solving this task automatically and many efforts about it have been made. [1,2,3,4,5]

The most intuitive approach for query by example is to view audio clip as a whole and model it as a long term distribution of frame-based features. In [2,3] the 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. Helen *et al.* [3] defined the Euclidean distance between GMMs for retrieval. GMM is a powerful tool capable of representing arbitrary density, but the durations of some audio files in the application of retrieval are very short, which leads to the lack of training data and affects the performance of GMM. 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

C.-L. Liu, C. Zhang, and L. Wang (Eds.): CCPR 2012, CCIS 321, pp. 556–563, 2012.

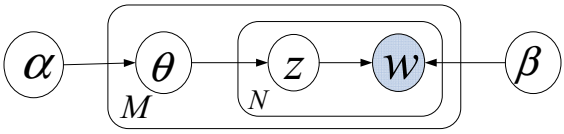

Fig. 1. Graphical model of standard LDA

get the histogram representation. The histogram method has low computation. However, vector quantization(VQ) would result in the loss of information and its main drawback for audio retrieval is that if two observations fall into different cluster bin, they are regarded as different even when they are closely spaced.

In fact, the histogram is similar to bag-of-word model in text processing and each cluster of audio features is like each word in the dictionary. Motivated by this fact, some approaches such as topic model proposed in the context of text information retrieval are successfully considered in the task of audio retrieval. Sundaram *et al.*[5]presented query by example for audio clips in latent perceptual space by using latent semantic index (LSI).Samuel Kim *et al.*[6]model the acoustic latent topics by latent Dirichlet allocation (LDA) and perform audio description classification and retrieval tasks.

The latent topic model assumes that each document consists of hidden topics and each topic in turn can be interpreted as a distribution over words in a dictionary.This assumption is also meaningful for audio because the contents of audio documents are not homogeneous in general. However,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[6,7]. If so, the shortcoming of the histogram method is inherited. In this work, 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.

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 Dirichelt 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 [8]. Fig.1 illustrates the graphical representation for LDA, which is a three-level hierarchical Bayesian model.

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 $d = \{w_1, w_2, .., w_i, ..., w_N\}$, where w_i is the i th word in the document. A

corpus consists of M documents and denoted by $C = \{d_1, d_2, ...d_i, .., d_M\}$.LDA assumes the following generative process for each document d in a corpus C:

1. For each document d ,choose $\theta \sim Dir(\alpha)$
2. for each word w_i in document d ,
   - Choose a topic $z_i \sim Multinomial(\theta)$
   - 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.

After processing of LDA, a document can be mapped onto the latent topic space by θ, of which each dimension indicates the membership probability of this document with respect to the corresponding latent topic. Furthermore, this representation of document can be used for many applications such as document classification and it demonstrates promising performance. However, 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. As mentioned earlier, the discretization has the innate shortcoming. Aiming to this defect of standard LDA for audio analysis, the Gaussian-LDA will be introduced and presented in next section.

3 Gaussian-LDA

The Gaussian-LDA is also built on the basic idea of topic model. As shown in the Figure.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 $C = \{d_1, d_2, ...d_i, .., d_M\}$ 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 $\theta \sim Dir(\alpha)$
2. for each frame-based feature x_i in document d ,
   - Choose a topic $z_i \sim Multinomial(\theta)$
   - Choose a frame-based feature x_i with a probability: $x_i/z_i, u_{1:K}, \Sigma_{1:K} \sim Normal(u_{z_i}, \Sigma_{z_i})$.

According to the generative process above, we can remodel every audio document. But now, the question is how to estimate the parameters, whose key problem is computing the posterior distribution of hidden variables given an audio document. Unfortunately, it is intractable to compute in general. In standard LDA, this problem is solved by using variational inference, which works by minimizing KL distance between the real distribution and simplified distribution using Jensen's inequality. To utilize this method to estimate the parameters

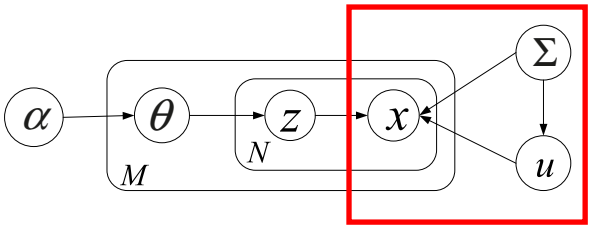

Fig. 2. Graphical model of Gaussian-LDA

of Gaussian-LDA, we firstly introduce an independent Gaussian-Wishart prior governing the mean and variance of each Gaussian, given by

$$\begin{aligned} p(u, \Sigma) &= p(u/\Sigma)p(\Sigma) \\ &= \prod_{k=1}^{K} \mathcal{N}(u_k/m_0, \beta_0^{-1}\Sigma_k)\mathcal{W}(\Sigma_k^{-1}/W_0, \upsilon_0) \end{aligned} \tag{1}$$

This is a conjugate prior distribution when both the mean and variance are unknown. Note that there is a link from u to Σ in Fig 2, since the variance of distribution over u is a function of Σ .In fact, now the problem is transferred as inference of variational Bayesian mixture of Gaussians. Attias[10] has given the solution in detail and we will present the outline of this algorithm.

After introducing the prior, the joint distribution can be written as follows:

$$p(X, Z, \theta, u, \Sigma) = p(X/Z, u, \Sigma)p(Z/\theta)p(\theta)p(u/\Sigma)p(\Sigma) \tag{2}$$

Next we consider a variational distribution which factorizes between the latent variable and the parameters:

$$q(Z, \theta, u, \Sigma) = q(Z)q(\theta, u, \Sigma) \tag{3}$$

There is no assumption on the functional form of $q(Z)$ and $q(\theta, u, \Sigma)$.The functional form of $q(Z)$ and$q(\theta, u, \Sigma)$ is determined automatically by optimization of the variational distribution.The detail of variational inference can be found in [10].

The whole process of parameter estimation can be viewed as variational Bayesian EM algorithm. In the E-step, we compute the optimized variational distributions over latent variables using the current distributions over model parameters. From the optimized $q(Z)$, the responsibilities r_{nk} can be evaluated:

$$r_{nk} \propto \pi_k \Lambda_k^{\frac{1}{2}} \exp(-\frac{D}{2\beta_k} - \frac{\upsilon_k}{2}(x_n - m_k)^T W_k(x_t - m_k)) \tag{4}$$

$$\pi_k = \exp(\psi(\alpha_k) - \psi(\sum \alpha_k)). \tag{5}$$

$$\Lambda_k = 2^D |W_k| \exp(\sum_{i=1}^{D} \psi(\frac{\upsilon_k + 1 - i}{2})). \tag{6}$$

where D is the dimensionality of feature vector.π_k andΛ_k are defined as intermediate variables, and ψ is the digamma function. In M-step, we keep these responsibilities fixed and get three statistics of the observed data given by

$$N_k = \sum_{n=1}^{N} r_{nk}, \overline{x}_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} x_n,$$
$$S_k = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk}(x_n - \overline{x}_k)(x_n - \overline{x}_k)^T. \tag{7}$$

Then the parameters of variational distribution are recomputed according to the following formulas:

$$\alpha_k = \alpha_0 + N_k, \beta_k = \beta_0 + N_k \tag{8}$$
$$\upsilon_k = \upsilon_0 + N_k, m_k = \frac{1}{\beta_k}(\beta_0 m_0 + N_k \overline{x}_k) \tag{9}$$

$$W_k^{-1} = W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k}(\overline{x}_k - m_0)(\overline{x}_k - m_0)^T \tag{10}$$

At last, the parameters m, β, W, υ obtained by EM algorithm describe the posterior distribution of u, Σ and the optimal estimations of u, Σ can be valued as their expectation. Then for a new audio document, it is easy to compute their posterior with respect to each topic. In experiments we use the posterior probability θ as the representation of the corresponding audio clip and cosine metric for retrieval.

4 Experiments and Results

To evaluate the performance of the proposed approach, we collected 1214 audio documents (length: 0.82 seconds to 1 minute). Each audio document is associated with a category and the dataset includes 31 categories: rain, bell, river, laugh, gun, dog and so on. Because these sounds have different formats, we first unify them to 16 kHz sampling and 16 bit per sample. After the silence frames are removed using a threshold of log energy, a set of 26 dimensional feature vectors is extracted for each frame by HTK 3.4. The 26 dimensions include MFCC plus normalized energy and their first order derivatives. The MFCCs provide spectral information considering human auditory characteristics, and they have been widely used in many audio related applications, such as speech recognition and audio classification.

4.1 Evaluation Procedure

In our experiments, we randomly select 100 documents as query files and the rest of the dataset are kept as training set. For each query, all of the audio documents from training set are ranked based on the similarity measure. The

documents are considered similar when they belong to the same category as the query. The number of correctly retrieved samples n_c is calculated for each query, the precision and recall rates can be got by using the formula below

$$precision = \frac{n_c}{R_r}, \tag{11}$$

$$recall = \frac{n_c}{N_c}, \tag{12}$$

where R_r is the number of all the samples retrieved and N_c is the total number of samples in the same class as the query. The performance of the experiments is measured by using average precision and recall rates of the whole database.

4.2 Results and Discussion

For comparison, we select histogram method, GMM and standard LDA as the baselines. All of these methods use the same frame features. In the histogram method, LBG-VQ is used for vector quantization and the number of cluster is set to 530, which is got by Bayesian information criterion (BIC). For GMM method, we train 8, 16 mixtures GMM for each audio and compute the likelihood of query data on different GMMs. The standard LDA is implemented based on the histograms by using the code in [11].

Since both standard LDA and Gaussian LDA are latent topic model, we first investigate their performance variation according to number of latent topics. Then a proper number of topics is set for these two models and we compare those with the histogram and GMM method. Table 1 gives the average precision at top 5 ranks with respect to different number of latent topics and the comparison between topic model with 200 topics and conventional methods is given in Fig 3.

The results in Table 1 clearly show that the Gaussian-LDA topic model outperforms the standard LDA topic model. This significant improvement is evident regardless of the number of latent topics. We argue that discretization of audio features when applying standard LDA to audio analysis affects the performance because vector quantization leads to information loss. Gaussian-LDA directly models latent topic as Gaussian distribution over audio features, which avoids discretization and hence gets better performance.

In the Fig 3, the precision and recall for tested methods are given and we can see that the proposed Gaussian-LDA produces higher precision and recall than the others. For GMM method, the 8-mixtures GMM gets better performance than 16-mixtures GMM. As mentioned earlier, this is because more mixtures

Table 1. The average precision at top 5 ranks with respect to different topic numbers

Topic numbers	50	100	150	200
standard LDA	53.7 %	55.4 %	57 %	57.4 %
Gaussian-LDA	55.4 %	58.9 %	59.7 %	59.9%

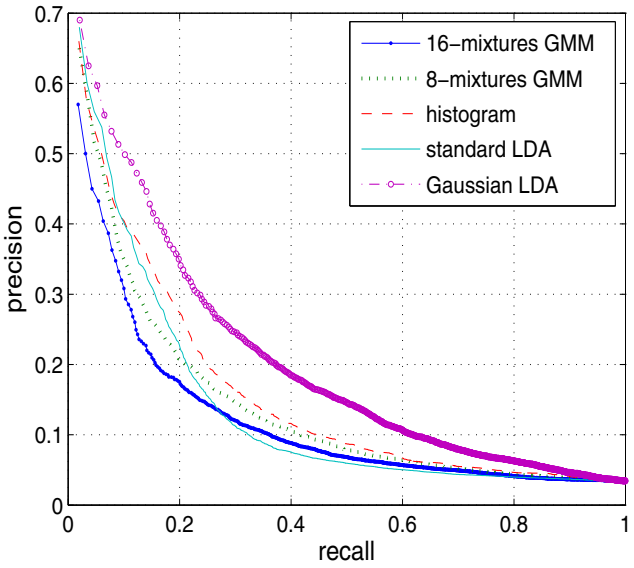

Fig. 3. the precision and recall for tested methods

usually need more training data and most audio clips in the task of query by example haven't enough long duration. The result of histogram method is obviously better than GMM method and the application of standard LDA enhances the performance of histogram when less than 10 most similar samples are retrieved. The experiment results verify that the introduction of latent topic model is helpful for audio retrieval and the Gaussian-LDA is more suitable for modeling continuous data.

5 Conclusion

In this paper, we proposed a latent topic model suited to process continuous data, which assumes that each audio document consists several latent topics and each topic is a Gaussian distribution over the continuous features rather than a discrete distribution over word-like unite. Compared with the standard LDA, this method models each topic in a continuous way,which skips vector quantization and avoids the loss of information. We also adopted the variational inference method to train Gaussian-LDA topic model and apply it to audio retrieval by using posterior topic probability as audio's representation. The results of experiment showed that the proposed topic model outperforms the standard LDA topic model in the task of audio retrieval.

Since the Gaussian-LDA has showed the superiority on processing continuous data, we will consider other application of this topic model in the future work, such as audio annotation and acoustic event detection.

Acknowledgments. This work was supported in part by the China National Nature Science Foundation (No.91120303, No.90820011 and No.90820303), 863 China National High Technology Development Project (No.20060101Z4073 and No.2006AA01Z194), and the National Grand Fundamental Research 973 Program of China (No. 2004CB318105).

References

1. Wold, E., Blum, T., Keislar, D., Wheaton, J.: Content-based classification, search, and retrieval of audio. IEEE Multimedia 3(2) (1996)
2. Aucouturier, J.J., Defreville, B., Pachet, F.: The bag-of-frames approach to audio pattern recognition: a sufficient model for urban soundscapes but not for Polyphonic Music. The Journal of Acoustic Society of America 122(2), 881–891 (2007)
3. Helen, M., Virtanen, T.: Query by example of audio signals using Euclidean distance between Gaussian mixture models. In: IEEE International Conference on Acoustic Speech and Signal Processing (ICASSP), Hawaii, USA (2007)
4. Foote, J.: Content-based retrieval of music and audio. Multimedia Storage Archiving Systems II 3229, 138–147 (1997)
5. Sundaram, S., Narayanan, S.: Audio Retrieval by Latent Perceptual Indexing. In: IEEE International Conference on Acoustic Speech and Signal Processing (CASSP), Las Vegas, USA (2008)
6. Kim, S., Sundaram, S., Narayanan, S.: Acoustic topic models for audio information retrieval. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (2009)
7. Kim, S., Sundaram, S., Georgiou, P., Narayanan, S.: Audio scene understanding using topic models. In: Neural Information Processing System (NIPS) Workshop (2009)
8. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3 (2003)
9. Winn, J.M.: Variational Message Passing and its Applications. PHD thesis, University of Cambridge (2003)
10. Attias, H.: A variational Bayesian framework for graphical models. In: Advances in Neural Information Processing Systems (2000)
11. `http://gibbslda.sourceforge.net/`