COMPARISON OF KULLBACK-LEIBLER DIVERGENCE APPROXIMATION METHODS BETWEEN GAUSSIAN MIXTURE MODELS FOR SATELLITE IMAGE RETRIEVAL

Shiyong Cui, Mihai Datcu

Remote Sensing Technology Institute (IMF)
German Aerospace Center (DLR)
Münchener Straße 20, 82234 Wessling
shiyong.cui, mihai.datcu@dlr.de

ABSTRACT

In many applications, such as image retrieval and change detection, we need to assess the similarity of two statistical models. As a distance measure between two probability density functions, Kullback-Leibler divergence is widely used for comparing two statistical models. Unfortunately, for some models such as Gaussian Mixture Model (GMM), Kullback-Leibler divergence has no analytically tractable formula. We have to resort to approximation methods. In this paper, we compare seven methods, namely Monte Carlo method, matched bond approximation, product of Gaussian, variational method, unscented transformation, Gaussian approximation, and min-Gaussian approximation, for approximating the Kullback-Leibler divergence between two Gaussian mixture models for satellite image retrieval. Two image retrieval experiments based on two publicly available datasets have been performed. The comparison is carried out in terms of both retrieval performance and computational time.

Index Terms— Gaussian Mixture Model (GMM), Kullback-Leibler Divergence, Image Retrieval.

1. INTRODUCTION

In Earth Observation (EO), a large amount of high-resolution satellite images are available at ground segments. Fast browsing and automatic interpretation of large data volumes is very challenging. Thus, content based image retrieval has been developed since years to solve this problem, such as the Knowledge-driven Information Mining (KIM) system [1] and the Geospatial Information Retrieval and Indexing (GeoIRIS) system [2]. To retrieval images, we have to solve two fundamental problems. The first problem is to find a method to describe the image content. The second problem is to compute the similarity values between a query and the remaining images in a database based the selected image content representation. Statistical feature space modeling [3] is an important method for image content representation. In this kind of methods, first we extract some local features from an image and then assume a parametric or semi-parametric model for the feature space. Next, we learn the parameters governing this model. This learned model can be considered as a statistical representation of the image content. Gaussian mixture model (GMM) is a popular choice due to its flexibility and the availability of Expectation-Maximization algorithm for parameter estimation. Gaussian mixture models were used in [4] for modeling the time-localized feature space and the dynamic feature space, where model selection was approached by the minimum description length principle.

To address the second problem, we have to find a similarity measure between two statistical models. The Kullback–Leibler divergence [5], defined as

$$D(X||Y) = \sum_x p_X(x) \log \frac{p_X(x)}{p_Y(x)},$$

is widely used as a similarity measure between two discrete probability distributions $p_X(x)$ and $p_Y(x)$. However, for some parametric models, it is hard to compute the integral involved in computing Kullback-Leibler divergence since it is not analytically tractable, which is the case for GMMs. Therefore, we have to resort to some kind of approximations to the Kullback-Leibler divergence between two GMMs. In the literature, there are a number of methods addressing this issue. Thus, in this paper, we compare seven methods for approximating the Kullback-Leibler divergence between two GMMs from a point view of content based satellite image retrieval.

2. KULLBACK-LEIBLER DIVERGENCE APPROXIMATION METHODS

2.1. Gaussian Mixture Model

A random variable X follows a Gaussian mixture distribution if its probability density function can be written as $p_X(\boldsymbol{x}) = \sum_{i=1}^{M} \pi_i \mathcal{N}_i(X; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where π_i is the prior probability of each component and $\mathcal{N}_i(X; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is a multivariate Gaussian distribution with a mean vector $\boldsymbol{\mu}_i$ and a covariance matrix $\boldsymbol{\Sigma}_i$. To apply this model to a feature space, we have to estimated the involved parameters

$\boldsymbol{\Theta} = (\alpha_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, ..., \alpha_M, \boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M)$ using a set of training data $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N)$. The standard method to estimate the parameters of a GMM is maximum likelihood estimation which is usually solved efficiently by the Expectation-Maximization (EM) algorithm [6]. However, we need to choose an appropriate number of Gaussian components before applying the EM algorithm to parameter estimation. To solve this problem, a number of different model selection methods, such as Bayes Information Criterion (BIC) [7], have been proposed in the literature. In this letter, BIC defined by (1) is used for choosing the number of Gaussian components.

$$BIC(\Theta) = -2 \log L(\Theta|\boldsymbol{X}) + K \log D \tag{1}$$

$L(\Theta|\boldsymbol{X})$ is the likelihood function, K is the number of parameters, and D is the dimensionality of the feature vector. If there is only one Gaussian component $p_X(\boldsymbol{x}) = \mathcal{N}(X; \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ and $p_Y(\boldsymbol{x}) = \mathcal{N}(Y; \boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$ in each GMM, the Kullback-Leibler divergence turns down to that between two Gaussian distributions. Unfortunately, if there are more than one Gaussian component, Kullback-Leibler divergence is not analytically tractable. Thus, we have to resort to approximation methods.

2.2. Approximation Methods

Given two GMMs $p_X(\boldsymbol{x}) = \sum_{i=1}^{M} \pi_i \mathcal{N}_i(X; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and $p_Y(\boldsymbol{x}) = \sum_{j=1}^{N} \alpha_j \mathcal{N}_j(Y; \boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j)$, our goal is to compute the Kullback-Leibler divergence between them. The methods we compared for approximation are presented as follows. In the following, for the brevity of presentation, we denote the Gaussian components of $p_X(\boldsymbol{x})$ and $p_Y(\boldsymbol{x})$ by $p_X^i(\boldsymbol{x}) = \mathcal{N}_i(X; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and $p_Y^j(\boldsymbol{x}) = \mathcal{N}_j(Y; \boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j)$.

2.2.1. Monte Carlo Sampling

The fundamental idea is to draw a large number of samples $\{\boldsymbol{x}_k\}_{k=1}^n$ from $p_X(\boldsymbol{x})$ and use these samples to replace the numerical integral by a summation over all samples. Thus, the Kullback-Leibler divergence can be approximated as

$$D_{MC}(X||Y) = \frac{1}{n} \sum_{i=1}^{n} \Big(\log p_X(x_i) - \log p_Y(x_i) \Big) \tag{2}$$

If the number of samples used for approximation goes to infinite, the approximation will be very close to the true value of Kullback-Leibler divergence. Practically, we need to draw a large number of samples $\{\boldsymbol{x}_k\}_{k=1}^n$ from a GMM. First we select a Gaussian component according to their prior distribution π_i. Then we draw a sample $\boldsymbol{x}_k$ from the selected Gaussian component $\mathcal{N}_i(X; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. We repeat this procedure for a large number of times to obtain enough samples. The Monte Carlo method is the only method that can really estimate the Kullback-Leibler divergence provided we have a large number of independent and identically distributed samples.

2.2.2. Gaussian Approximation

This method first approximates $p_X(x)$ and $p_Y(x)$ by two Gaussian distributions $\hat{p}_X(x) = \mathcal{N}(X; \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ and $\hat{p}_Y(x) = \mathcal{N}(Y; \boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y)$. The mean and covariance matrix can be estimated by $\boldsymbol{\mu}_X = \sum_M^{i=1} \pi_i \boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_X = \sum_M^{i=1} \pi_i (\boldsymbol{\Sigma}_i + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_X)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_X)^T)$. Then the Kullback-Leibler divergence between $p_X(x)$ and $p_Y(x)$ can be approximated by that of these two Gaussian distributions. Another popular choice of Gaussian approximation is to use the minimum Kullback-Leibler divergence between components of the two GMMs.

2.2.3. The Product of Gaussians Approximation

This method [8] is derived based on an upper bound on the likelihood resulted from Jensen's inequality. Since likelihood and Kullback-leibler divergence have the following relation

$$D(X||Y) = \mathbb{E}_{p_X(\boldsymbol{x})}[\log p_X(\boldsymbol{x})] - \mathbb{E}_{p_X(\boldsymbol{x})}[\log p_Y(\boldsymbol{x})], \tag{3}$$

where $\mathbb{E}[\cdot]$ denotes the expectation, Kullback-Leibler divergence can be approximated by an estimate of the likelihood. Based on the Jensen's inequality $f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$, an upper bound of likelihood can be derived as (4),

$$\mathbb{E}_{p_X(\boldsymbol{x})}[\log p_Y(\boldsymbol{x})] \leq \sum_{i=1}^{M} \pi_i \log \sum_{j=1}^{N} \alpha_j C_{ij} \tag{4}$$

$C_{ij} = \int p_X^i(\boldsymbol{x}) p_Y^j(\boldsymbol{x}) d\boldsymbol{x}$ is the normalization constant of a product of two Gaussians that can be found from the appendix. Therefore, the Kullback-Leibler divergence can be approximated using the above upper bound

$$D_{PoG}(X||Y) = \sum_{i=1}^{M} \pi_i \log \frac{\sum_{j=1}^{N} \pi_j C_{ij}}{\sum_{j=1}^{N} \alpha_j C_{ij}} \tag{5}$$

2.2.4. The Unscented Transformation

The unscented transform [9] is a method to estimate an expectation $E_{f(x)}[h(x)]$ of a function $h(x)$ with a probability density function $f(x)$. Following the same idea as Monte Carlo, the expectation can be estimated by a sample x_i. However, we do not have to draw a large number of samples. Instead, we select only 2D "sigma" points $\{x_k\}_{k=1}^{2D}$ of the distribution $f(x)$. Thus, the expectation can be written as $E_{f(x)}[h(x)] = \frac{1}{2D} \sum_{k=1}^{2D} h(x_k)$. One popular choice of the sigma points for a $\mathcal{N}_i(X; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is $x_{i,k} = \boldsymbol{\mu}_i + \sqrt{D\lambda_{i,k}} \boldsymbol{e}_{i,k}$ and $x_{i,D+k} = \boldsymbol{\mu}_i - \sqrt{D\lambda_{i,k}} \boldsymbol{e}_{i,k}$ with $\lambda_{i,k}$ and $\boldsymbol{e}_{i,k}$ being the eigenvalues and eigenvectors of the covariance matrix $\boldsymbol{\Sigma}_i$. Therefore, we can draw a sample $x_{i,k}$ from each Gaussian component of $p_X(x)$ and use them to approximate the Kullback-Leibler divergence as follows.

$$D_{ustd}(X||Y) = \frac{1}{2D} \sum_{i=1}^{M} \pi_i \sum_{k=1}^{2D} \log \frac{p_X(x_{i,k})}{p_Y(x_{i,k})} \tag{6}$$

2.2.5. *The Matched Bound Approximation*

The matched bound approximation [9] computes the Kullback-Leibler divergence by minimizing a matching function that finds the closest Gaussian component of $p_Y(\boldsymbol{x})$ to that of $p_X(\boldsymbol{x})$. It has two steps. The first step is to find the closes Gaussian component of $p_Y(\boldsymbol{x})$ to each component of $p_X(\boldsymbol{x})$. Formally, we solve the following minimization problem (7) for each $p_X^i(\boldsymbol{x})$.

$$m(i) = \underset{j}{\operatorname{argmin}}\, D(p_X^i(\boldsymbol{x})||p_Y^j(\boldsymbol{x})) - \log \alpha_j \tag{7}$$

Then we use the matched pairs of Gaussian components to approximate the Kullback-Leibler divergence as follows

$$D_M(X||Y) = \sum_{i=1}^{M} \pi_i \Big(D(p_X^i(\boldsymbol{x})||p_Y^{m(i)}(\boldsymbol{x})) + \log \frac{\pi_i}{\alpha_{m(i)}} \Big) \tag{8}$$

2.2.6. *The Variational Approximation*

Variational approximation [8] is based on a variational lower bound on the likelihood $\mathbb{E}_{p_X(\boldsymbol{x})}[\log p_Y(\boldsymbol{x})]$ obtained by introducing a set of variational parameters $\phi_{j|i} > 0$ such that $\sum_j \phi_{j|i} = 1$. Based on the Jensen's inequality, we have the following lower bound:

$$\begin{aligned} \mathbb{E}_{p_X(\boldsymbol{x})}[\log p_Y(\boldsymbol{x})] &= \mathbb{E}_{p_X(\boldsymbol{x})} \Big[\log \sum_{j=1}^{N} \alpha_j p_Y^j(\boldsymbol{x}) \Big] \\ &\geq \sum_{i=1}^{M} \sum_{j=1}^{N} \pi_i \phi_{j|i} \Big(\log \frac{\alpha_j}{\phi_{j|i}} + \mathbb{E}_{p_X^i(\boldsymbol{x})}[\log p_Y^j(\boldsymbol{x})] \Big) \end{aligned} \tag{9}$$

We can maximize the lower bound in (9) and solve for $\phi_{j|i}$, which is given as

$$\hat{\phi}_{j|i} = \frac{\alpha_j \exp\Big(- D(p_X^i(\boldsymbol{x})||p_Y^j(\boldsymbol{x})) \Big)}{\sum_{j=1}^{N} \alpha_j \exp\Big(- D(p_X^i(\boldsymbol{x})||p_Y^j(\boldsymbol{x})) \Big)} \tag{10}$$

Then the lower bound can be computed by substituting (10) into (9). Likewise, we can define a lower bound on $E_{p_X(x)}[\log p_X(x)]$ by introducing another set of variational parameters. Finally the Kullback-Leibler divergence between $p_X(\boldsymbol{x})$ and $p_Y(\boldsymbol{x})$ can be approximated as (11)

$$D_v(X||Y) = \sum_{i=1}^{M} \pi_i \log \frac{\sum_{j=1}^{M} \pi_j \exp\Big(- D(p_X^i(\boldsymbol{x})||p_Y^j(\boldsymbol{x})) \Big)}{\sum_{j=1}^{N} \alpha_j \exp\Big(- D(p_X^i(\boldsymbol{x})||p_Y^j(\boldsymbol{x})) \Big)} \tag{11}$$

3. EXPERIMENTS AND DISCUSSION

Two datasets are used for evaluation. The first one is the UC Merced land use dataset [10][1]. It contains contain 18 classes of scenes. Each class has 100 images with a size of 256×256 pixels. The second is the Wuhan high resolution satellite scene dataset [2]. This dataset contains contain 18 classes of scenes and for each class, there are 50 samples with a size of 600×600 pixels.

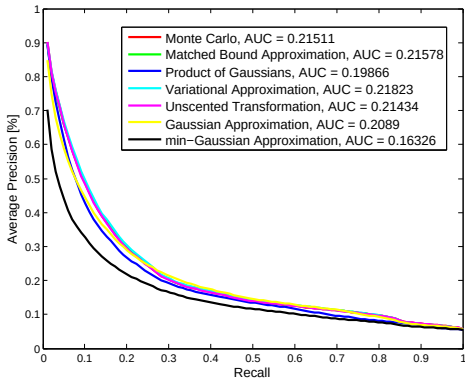

Fig. 1. Average AUC of the seven approximation methods.

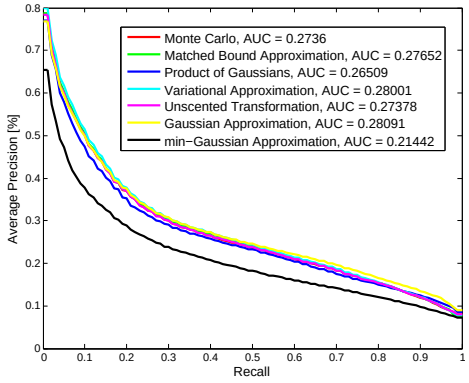

Fig. 2. Average AUC of the seven approximation methods.

We use each image as a query and search for similar images among the remaining images. For each query, we first learn a GMM using the RGB pixel values based on the algorithm presented in section 2.1 which can automatically estimate the number of components. Then we use the estimated parameters and the seven methods for approximating the Kullback-Leibler divergence between two GMMs as a similarity measure. To evaluate, we compute the precision and recall curve. The area under this curve (AUC) is also compüuted and compared. In addition, we also compare the com-

[1]The data is available at `http://vision.ucmerced.edu/datasets/landuse.html`

[2]The data is available at `http://dsp.whu.edu.cn/cn/staff/yw/HRSscene.html`

Table 1. Average AUC [%] and the average CPU time of the seven methods.

Methods	Monte Carlo	Matched Bound	Prod. of Gauss.	Variational	Unscented	Gaussian	min-Gauss.
Average AUC [%]	27.36	27.65	26.51	28.00	27.38	28.09	21.44
CPU time [s]	0.4439	0.0303	0.0468	0.0447	0.0330	0.0020	0.0217

Table 2. Average AUC [%] and the average CPU time of the seven methods.

Methods	Monte Carlo	Matched Bound	Prod. of Gauss.	Variational	Unscented	Gaussian	min-Gauss.
Average AUC [%]	21.51	21.58	19.87	21.82	21.43	20.89	16.33
CPU time [s]	0.2157	0.0100	0.0150	0.0358	0.0297	0.0014	0.0265

putational time. For the method of Monte Carlo sampling, we use a sample of $80,000$ points for approximation.

The results of the comparison using the first dataset (UC Merced land use dataset) is shown in Fig. 1. The average AUC and CPU time are shown in Table. 2. From the results, we can observe that the variational approximation is the best one with an average AUC of 0.2182. Matched bound approximation ranks second but with a much faster speed than the variational method. In addition, although we used a large sample of $80,000$ points in Monte Carlo sampling, they are still not enough if we compare it with the variational approximation. Furthermore, Monte Carlo method is computationally very slow, which limits its use in some applications such as change detection because it has to be computed for each pixel. min-Gaussian is the most inferior method and Gaussian approximation is the fastest one. Additionally, we can also observe that the matched bound approximation and the unscented transformation have similar

The results of the experiments using the second dataset (Wuhan High-resolution Satellite Scene Dataset) is presented in Fig. 2. Similar as the previous experiment, min-Gaussian performs least among the seven methods. However, Gaussian approximation performs best with an average AUC of 0.2687 and has the lowest computational complexity of 0.0020. The main reason is that, for some homogeneous classes such as forest, meadow, desert, etc., the assumed GMM distribution boils down to a Gaussian distribution. The variational method performs only slightly worse than the Gaussian approximation. As in the first experiment, we can also observe that the matched bound approximation and the unscented transformation have similar performances that are only slightly lower than the variational approximation method. but they can be computed faster than the variational method. The method of product of Gaussian performs similarly as in the first experiment.

4. CONCLUSION

In this paper, we compare seven methods for approximating the Kullback-Leibler divergence between two Gaussian mixture models for satellite image retrieval. Two image retrieval experiments based on two publicly available datasets have been performed. In principle, Monte Carlo method can achieve high accuracy provided a large number of samples are available. However, practically, it is not applicable in many cases due to its high computational complexity. Variational approximation seems a good compromise between computation and accuracy. If the images are homogeneous, Gaussian approximation will be a good choice. The matched bound approximation and the unscented transformation performs slightly worse than the variational method. min-Gaussian is generally not a good choice.

5. REFERENCES

[1] M. Datcu and K. Seidel, "Human-centered concepts for exploration and understanding of earth observation images," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 601–609, March 2005.

[2] Chi-Ren Shyu, M. Klaric, G. J. Scott, A. S. Barb, C. H. Davis, and K. Palaniappan, "GeoIRIS: Geospatial information retrieval and indexing system — content mining, semantics modeling, and complex queries," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 4, pp. 839–852, 2007.

[3] D. A. Lisin, M. A. Mattar, M. B. Blaschko, M. C. Benfield, and E. G. Learned-Miller, "Combining local and global image features for object class recognition," in *Proc. of IEEE Workshop on Learning in Computer Vision and Pattern Recognition (CVPR)*, June 2005, pp. 47–47.

[4] P. Heas and M. Datcu, "Modeling Trajectory of Dynamic Clusters in Image Time-Series for Spatio-Temporal Reasoning," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 7, pp. 1635–1647, Jul. 2005.

[5] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 49–86, 1951.

[6] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[7] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[8] J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, Apr. 2007, vol. 4, pp. IV–317–IV–320.

[9] J. Goldberger, S. Gordon, and H. Greenspan, "An efficient image similarity measure based on approximations of KL-divergence between two gaussian mixtures," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 2003.

[10] Y. Yang and S. Newsam, "Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification," in *Proc. 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '10*, New York, NY, 2010, pp. 270–279.