A Pseudometric for Gaussian Mixture Models

Linfei Zhou*, Wei Ye*, Bianca Wackersreuther*, Claudia Plant† and Christian Böhm*
* Institute for Computer Science, Ludwig-Maximilians-Universität München
† Department of Computer Science, University of Vienna
Email: *{zhou, ye, wackersb, boehm}@dbs.ifi.lmu.de, †claudia.plant@univie.ac.at

***Abstract*—Efficient similarity search for uncertain data is a challenging task in many modern data mining applications such as image retrieval, speaker recognition and stock market analysis. A common way to model the uncertainty of the objects is using probability density functions in the form of Gaussian Mixture Models (GMMs), which have the ability to approximate any arbitrary distribution. However, there is a lack of suitable similarity measures for GMMs. Hence, in this paper we propose a similarity measure, Pseudometric for GMMs (PmG). The advantage of PmG is that it is efficient in computation because of its closed-form expression for GMMs, and it fulfills the triangle inequality which is necessary for many techniques like clustering and embedding. Extensive experimental evaluations of the proposed similarity measure on various real-world and synthetic data sets demonstrate a considerably better performance than that of the existing similarity measures, in terms of run-time and result quality in classification and clustering.**

Keywords–Gaussian Mixture Models, Similarity Measures, Metric

I. INTRODUCTION

Information extraction systems capable of handling uncertain data objects is an actively investigated research field. Many modern applications like speaker recognition systems [1, 2], content-based image and video retrieval [3, 4], biometric identification and stock market analysis can be supported by uncertain data representation. As a general class of Probability Density Functions (PDF), Gaussian Mixture Model (GMM) consists of a weighted sum of univariate or multivariate Gaussian distributions, allowing a concise but exact representation of uncertain data objects [5]. A good example of using objects represented by GMMs is managing multimedia data [6]. A 90 minutes movie contains about 130,000 single images. It requires large storage capacities as well as enormous computational effort for content-based retrieval. Storing the movie as GMMs will dramatically reduce the resource consumption while guaranteeing a high accuracy of the search result.

Besides the modeling of uncertainty, the efficiency of similarity search on uncertain data is another important aspect. For data objects represented by GMMs, Rougui et al. [7] have built a bottom-up hierarchical tree based on the calculation of the complete similarity matrix for all GMMs. However, it is only usable for static data sets, since it has no convenient insertion and deletion strategy, which depends on a proper similarity measure and requires a corresponding custom-built structure. The suitable similarity measures of GMMs for the indexing trees are yet to be developed and tested. A competitive candidate for such a similarity measure has the competencies to guarantee high efficiency in its computation and to facilitate indexing and further analysis. As we will demonstrate, our technique proposed in this paper is highly efficient because it enables closed-form computation. Moreover, our technique has the property of being a pseudometric, thus indexing techniques like VP-tree [8] can be applied for efficient search and embedding techniques like multidimensional scaling facilitate the subsequent analysis of the data set. To our knowledge, several studies [9]–[14] have dealt with defining similarity measures for GMMs, but only a few of them have closed-form expressions and none of them is a metric or pseudometric.

The main contributions of this paper are:

- We propose a pseudometric for GMMs (PmG), which is a similarity measure for GMMs. We derive the closed-form expression and prove that it is a pseudometric. The closed-form expression has a great advantage in calculation, and the properties of pseudometric are required by many analysis techniques.
- We define Normalized Matching Probability (NMP), which can be constituted to form novel similarity measures that have closed-form expressions for GMMs.
- Experimental evaluation demonstrates both the effectiveness and efficiency of PmG.

The rest of this paper is organized as follows: Section II gives the basic definition of GMMs, metric and pseudometric. Section III defines NMP and PmG, and gives the proof of the pseudometric properties of PmG. Section IV shows the experimental studies for verifying the efficiency and effectiveness of the proposed similarity measure. In Section V, we survey the previous work. Finally, Section VI summarizes the paper and presents some ideas for further research.

II. FORMAL DEFINITIONS

In this section, we summarize the formal notations for GMM. GMM is a probabilistic model that represents the probability distribution of observations. The definition of GMM is shown as follow.

Definition 1: (Gaussian Mixture Model) Let $\mathbf{x} \in \mathbb{R}^D$ be a variable in a D-dimensional space, $\mathbf{x} = (x_1, x_2, ..., x_D)$. A Gaussian Mixture Model $\mathcal{G}$ is the weighted sum of m Gaussian functions, defined as:

$$\mathcal{G}(\mathbf{x}) = \sum_{1 \leq i \leq m} w_i \cdot \mathcal{N}_i(\mathbf{x}) \quad (1)$$

where $\sum_{1 \leq i \leq m} w_i = 1$, $\forall i \in [1, m], w_i \geq 0$, and Gaussian component $\mathcal{N}_i(\mathbf{x})$ is the density of a Gaussian distribution with

 ISBN: 978-1-61208-558-6

a covariance matrix Σ_i:

$$\mathcal{N}_i(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^D|\Sigma_i|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i)\right)$$

Specially, when Σ_i is a diagonal matrix, $\mathcal{N}_i(\mathbf{x})$ can be reformulated as:

$$\mathcal{N}_i(\mathbf{x}) = \prod_{1\le l\le D}\frac{1}{\sqrt{2\pi\sigma_{i,l}^2}}\exp\left(\frac{-(x_l-\mu_{i,l})^2}{2\sigma_{i,l}^2}\right)$$

where $\sigma_{i,l}$ is the l-th element on the diagonal of Σ_i.

Most of dissimilarities are distances, and they are also metrics if the following definition is matched.

Definition 2: (Metric)

Given an nonempty set of objects $\mathscr{P}(\mathbb{R}^D)$, a mapping $d : \mathscr{P}(\mathbb{R}^D) \times \mathscr{P}(\mathbb{R}^D) \longrightarrow \mathbb{R}^+$ is a metric when the following properties always hold for any object $\mathcal{X}, \mathcal{Y}, \mathcal{Z} \in \mathscr{P}(\mathbb{R}^D)$.

- **Non-negativity**: $d(\mathcal{X}, \mathcal{Y}) \ge 0$
- **Identity of indiscernibles**: $d(\mathcal{X}, \mathcal{Y}) = 0 \Leftrightarrow \mathcal{X} = \mathcal{Y}$
- **Symmetry**: $d(\mathcal{X}, \mathcal{Y}) = d(\mathcal{Y}, \mathcal{X})$
- **Triangle inequality**: $d(\mathcal{X}, \mathcal{Y}) + d(\mathcal{Y}, \mathcal{Z}) \ge d(\mathcal{X}, \mathcal{Z})$

A pseudometric is a mapping that satisfies the axioms for a metric, except that instead of the identity of indiscernbles, $d(\mathcal{X}, \mathcal{X}) = 0$ but for some distinct objects $\mathcal{X} \ne \mathcal{Y}$, possibly $d(\mathcal{X}, \mathcal{Y}) = 0$.

The properties of the metric, especially the triangle inequality, are essential to index structures such as M-tree and VP-tree for efficient queries, and they are fully required by some techniques like DBSCAN. For similarity measures without the metric properties, specialized index and analysis methods are needed to guarantee the efficiency and the applicability of certain techniques.

III. Pseudometric for Gaussian Mixture Models

In this section, we extend Matching Probability (MP) into NMP, and derive its closed-form expression for GMMs. NMP provides a fundamental closed-form calculation for GMMs, and it can be used to define other similarity measures for GMMs. Specifically, we define PmG, a pseudometric for GMMs.

A. Normalized Matching Probability

MP considers all the possible positions of true feature vectors, and sums up the joint probabilities of two PDFs. Here we define NMP for GMMs.

Definition 3: (Normalized Matching Probability) Given two GMMs $\mathcal{G}_1(\mathrm{x})$ and $\mathcal{G}_2(\mathrm{x})$ with diagonal covariance matrices in space $\mathbb{R}^D$, we define the NMP $\langle\mathcal{G}_1, \mathcal{G}_2\rangle$ as follows:

$$\langle\mathcal{G}_1, \mathcal{G}_2\rangle = \int_{\mathbb{R}^D} \mathcal{G}_1'(\mathbf{x}) \cdot \mathcal{G}_2'(\mathbf{x})\mathrm{d}\mathbf{x} \qquad (2)$$

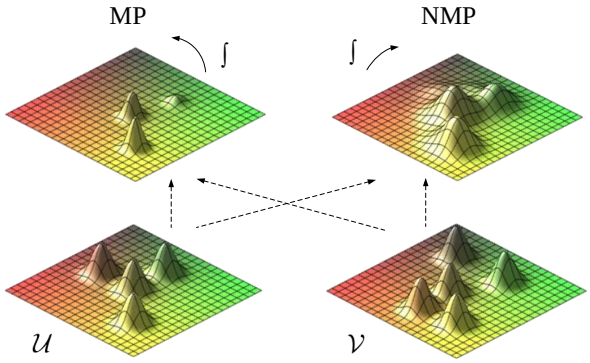


Figure 1: Demonstration of MP and NMP between GMM objects $\mathcal{U}, \mathcal{V}$ in a two-dimensional space.

where $\mathcal{G}'(\mathrm{x}) = \sum_{1\le i\le m}\prod_{1\le l\le D}\mathcal{N}(\mu_{i,l}, \sigma_{i,l}^2/w_i)$. m, μ and σ are the parameters of GMM $\mathcal{G}$ (see Definition 1).

Figure 1 demonstrates MP and NMP between two GMM objects $\mathcal{U}, \mathcal{V}$. As the measures of similarity, both MP and NMP integrate the similar parts of Gaussian components, as shown in the top of the figure. Because of the normalization operation, NMP gains a greater values than MP and emphasizes the shared parts.

For the closed-form expression of $\langle\mathcal{G}_1, \mathcal{G}_2\rangle$, we can derive it from the following equation.

$$\begin{aligned}
&\langle\mathcal{G}_1, \mathcal{G}_2\rangle \\
&= \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\prod_{l=1}^{D}\frac{\sqrt{w_{1,i}w_{2,j}}}{2\pi\sqrt{\sigma_{1,i,l}^2\sigma_{2,j,l}^2}}\int \mathrm{e}^{-\frac{(x-\mu_{1,i,l})^2}{2\sigma_{1,i,l}^2/w_{1,i}}-\frac{(x-\mu_{2,j,l})^2}{2\sigma_{2,j,l}^2}}\mathrm{d}x \\
&= \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\prod_{l=1}^{D}\frac{\mathrm{e}^{-\frac{(\mu_{1,i,l}-\mu_{2,j,l})^2}{2(\sigma_{1,i,l}^2/w_{1,i}+\sigma_{2,j,l}^2/w_{2,j})}}}{\sqrt{2\pi(\sigma_{1,i,l}^2/w_{1,i}+\sigma_{2,j,l}^2/w_{2,j})}} \\
&= \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\prod_{l=1}^{D}\mathcal{N}(\mu_{1,i,l}, \mu_{2,j,l}, \frac{\sigma_{1,i,l}^2}{w_{1,i}}+\frac{\sigma_{2,j,l}^2}{w_{2,j}})
\end{aligned}$$

If two GMMs are very disjoint, NMP between them is close to zero. To obtain a higher NMP, it is required that two GMMs have similar shapes, i.e. similar parameters.

A closed-form expression is intrinsically valuable for computation. It saves extra efforts to get a good approximation by avoiding simulation methods like Monte-Carlo, which may cause a significant increase in computation time and the loss of precision. Therefore, closed-form expressions are well received in many applications, especially in real-time applications.

Based on NMP, we can get a set of similarity measures with closed-form expressions for GMMs. For example, we can define a distance as follows.

$$d(\mathcal{G}_1, \mathcal{G}_2) = 1 - \frac{\langle\mathcal{G}_1, \mathcal{G}_2\rangle}{\sqrt{\langle\mathcal{G}_1, \mathcal{G}_1\rangle\langle\mathcal{G}_2, \mathcal{G}_2\rangle}}$$

 ISBN: 978-1-61208-558-6

There are several similarity measures based on MP have been proposed [10, 12, 13], and we can easily extend NMP on them.

B. Pseudometric for GMMs

On the basis of NMP, we define a pseudometric for GMMs, PmG. Likewise, PmG determines the square differences between normalized GMMs, sums them up by integration and returns the root of the integration. The definition of PmG is shown as follows.

Definition 4: (Pseudometric for GMMs) Given two GMMs $\mathcal{G}_1(\mathrm{x})$ and $\mathcal{G}_2(\mathrm{x})$ with diagonal covariance matrices in space $\mathbb{R}^D$, we define the PmG of them as follows:

$$d_{\mathrm{PmG}}(\mathcal{G}_1, \mathcal{G}_2) = \sqrt{\langle \mathcal{G}_1, \mathcal{G}_1 \rangle + \langle \mathcal{G}_2, \mathcal{G}_2 \rangle - 2 \cdot \langle \mathcal{G}_1, \mathcal{G}_2 \rangle} \quad (3)$$

Obviously, PmG has a closed-form expression for GMMs with diagonal covariances. Then we give the proof that PmG fulfills three properties of a metric.

Lemma 3.1: PmG is a pseudometric.

Proof: **Non-negativity:**

According to the definition of NMP, for any $\mathcal{G}_1, \mathcal{G}_2$, $\langle \mathcal{G}_1, \mathcal{G}_1 \rangle + \langle \mathcal{G}_2, \mathcal{G}_2 \rangle - 2 \cdot \langle \mathcal{G}_1, \mathcal{G}_2 \rangle \geq 0$. Thus $d_{\mathrm{PmG}}(\mathcal{G}_1, \mathcal{G}_2) \geq 0$. If $\mathcal{G}_1 = \mathcal{G}_2$, $d_{\mathrm{PmG}}(\mathcal{G}_1, \mathcal{G}_2) = 0$.

Symmetry:

Obviously, $d_{\mathrm{PmG}}(\mathcal{G}_1, \mathcal{G}_2) = d_{\mathrm{PmG}}(\mathcal{G}_2, \mathcal{G}_1)$.

Triangle Inequality:

Since $d_{\mathrm{PmG}}(\mathcal{G}_1, \mathcal{G}_2)$ can be reformed to the ℓ^2 norm of $\mathcal{G}_1^{'}$ and $\mathcal{G}_2^{'}$, its triangle inequality can be proved easily. ■

Having the property of triangle inequality, metric or pseudometric can employ various of metric trees to make accesses to the data objects more efficient. Otherwise, specialized index and analysis methods are needed to guarantee the efficiency and applicability of index, which means extra efforts.

IV. Experimental Evaluation

In this section, we provide experiments on both real-world and synthetic data sets to show the effectiveness and efficiency of the proposed pseudometric for GMMs. Classification and clustering, which are the major subdivisions of pattern recognition techniques, as well as the run-time of similarity calculation are used in the evaluation.

For Kullback-Leibler (KL) divergence [15] based similarity measures, only matching based approximation (KLm) [9] is included in this paper since it is one of the best-performing approximations [16].

All the experiments are implemented with Java 1.7, and executed on a regular workstation PC with 3.4 GHz dual core CPU equipped with 32 GB RAM. For all the experiments, we use the 10-fold cross validation and report the average results over 100 runs.

A. Data Sets

Synthetic data and four kinds of real-world data, including activity data, image data, audio data, and weather data, are used in the experiments. For the data objects, GMMs are estimated using Expectation-Maximization (EM) algorithm (implementation provided by WEKA).

Activity data [17] is collected from 15 participants performing seven activities. Assuming that participants complete a single activity in three seconds, we regard the 150 continuous measurements of acceleration on three axes as one data object. Thus 1083 objects are generated for participant 1.

Image data [18] is a collection of images taking under various viewpoints. In this paper, we use the gray images recording 100 objects from 72 viewpoints. Every image (192×144) is smoothed by a Gaussian filter with a standard deviation of five pixels.

Audio data [19] consists of speech from ten speakers, the names of who are shown as follows: Aaron, Abdul Moiz, Afshad, Afzal, Akahansson, Alexander Drachmann, Afred Strauss, Andy, Anna Karpelevich and Anniepoo. Every wav file is split into ten fragments, transformed into frequency domain by Fast Fourier Transform.

Weather data [20] is the historical weather data of 908 airports in Europe from 2005 to 2014. The features of Weather data are temperature and humidity, and only the average values of each day are used.

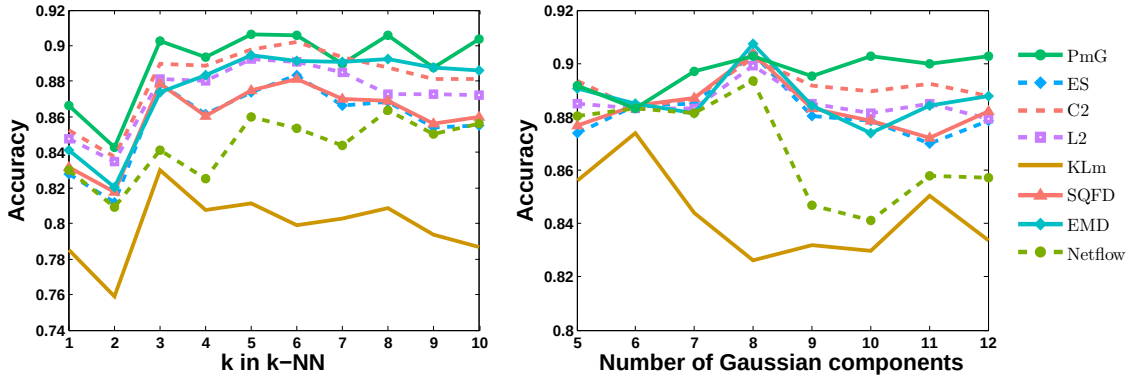
The synthetic data sets [21] are generated by randomly choosing mean values between 0 and 100 and standard deviations between 0 and 5 for each Gaussian component. The weights are randomly assigned, and they sum up to one within each GMM. Since there is no intuitive way to assign class labels for GMMs in advance, here we use the synthetic data sets only for the run-time evaluation.

B. Effectiveness Evaluation

In the evaluation of classification, we employ the simplest and widely used algorithm k-Nearest Neighbors (k-NN), rather than the other more complex techniques, to compare the effectiveness of the similarity measures, since we are not interested in tuning classification accuracy to its optimum.

Varying k in k-NN and the number of Gaussian components in each GMM, we start with experiments on Activity data. As shown in Figure 2, PmG has a better performance than the other similarity measures for different k. With the increase number of Gaussian components in each GMM, the classification accuracies of PmG sightly increase, and generally outperform the other measures. For all the four real-world data sets, we fix the numbers of Gaussian components and report the results of 1-NN classification accuracies in Table I, where the highest accuracies of each column are marked in bold and with •, while the second highest ones are just marked in bold. We can see that PmG achieves the highest or second highest accuracies among all the similarity measures.

To compare the usability of the proposed similarity measure for unsupervised data mining, we perform clustering experiments on all four real-world data sets. Instead of k-means algorithm, the k-medoids is used here since it works



(a) (b)

Figure 2: Classification results of Activity data. The numbers of Gaussian components in GMMs generated from data objects in (a) are fixed as ten. In (b), k in k-NN is fixed as three.

TABLE II. CLUSTERING RESULTS OF k-MEDOIDS.

	Activity (k=7)			**Image** (k=100)			**Audio** (k=10)			**Weather** (k=14)		
	Purity	NMI	FM	Purity	NMI	FM	Purity	NMI	FM	Purity	NMI	FM
PmG	**.82±.03**	**.49±.04**	**.56±.07**	**.43±.00**	**.69±.00**	**.30±.01**	**.56±.04**	**.45±.03**	**.40±.03**	**.58±.01**	**.23±.01**	.22±.03
ES	.73±.05	.37±.03	.39±.03	.40±.01	.67±.00	.28±.01	.49±.04	.37±.02	.33±.03	.55±.02	.18±.02	.21±.03
C2	.77±.06	.39±.07	.45±.07	.40±.00	.67±.00	.27±.00	.46±.05	.34±.03	.32±.03	.53±.02	.16±.02	.22±.04
L2	.77±.03	.37±.04	.44±.06	.40±.01	.66±.00	.27±.01	.48±.04	.36±.03	.33±.02	.54±.02	.17±.01	.20±.02
KLm	.74±.03	.41±.03	.50±.05	.40±.00	.66±.00	.27±.00	.46±.05	.38±.03	.33±.04	.47±.01	.15±.01	**.26±.03**
SQFD	.72±.04	.35±.03	.39±.04	.40±.01	.66±.00	.27±.00	.51±.04	.38±.03	.35±.03	.54±.01	.17±.01	.21±.02
EMD	.79±.04	.45±.05	.48±.07	.33±.01	.58±.00	.21±.01	.47±.03	.40±.03	.34±.03	.57±.01	.21±.01	.23±.03
Netflow	.74±.07	.36±.08	.44±.05	.39±.00	.66±.00	.27±.00	.49±.03	.45±.04	.38±.04	.56±.01	.20±.01	.19±.02

TABLE I. 1-NN CLASSIFICATION RESULTS OF REAL-WORLD DATA.

	Activity ($m = 10$)	**Image** ($m = 5$)	**Audio** ($m = 5$)	**Weather** ($m = 10$)
PmG	**.865±.030** •	**.852±.010** •	**.851±.044**	**.761±.032** •
ES	.834±.032	**.827±.010**	.804±.035	.740±.030
C2	**.859±.032**	.827±.010	.803±.039	.730±.033
L2	.853±.032	.826±.010	.802±.039	.737±.028
KLm	.793±.039	.823±.010	.825±.037	.717±.030
SQFD	.843±.030	.825±.014	.808±.028	.724±.035
EMD	.848±.033	.519±.014	.809±.032	**.758±.047**
Netflow	.838±.030	.813±.010	**.857±.032** •	.757±.051

with arbitrary similarity measure, making it more suitable in our situation. We evaluate the clustering results using three widely used criteria, Purity, Normalized Mutual Information (NMI) and F1 Measure (FM).

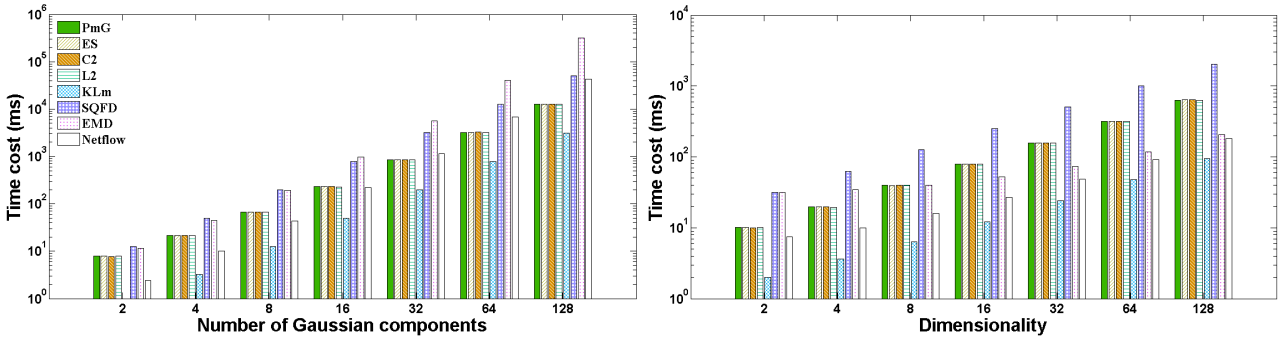
Table II illustrates the evaluation of clustering results when using different similarity measures on four real-world data sets. PmG achieves the best performance among all the measures on all three criteria, except for FM on Weather data.

C. *Efficiency Evaluation*

Every similarity measure evaluated in this paper has a time complexity of $O(m_1 \cdot m_2 \cdot D)$, where m_1 and m_2 are the numbers of Gaussian components in GMMs that are used for similarity calculating, and D is the dimensionality of data space. To support the theoretical time complexity, we provide comparisons by scaling m and D on synthetic data sets.

We calculate distance matrices for all the synthetic data sets, and as mentioned before, the average time cost of 100 runs are reported. Figure 3 shows the comparison of run-time between all the similarity measures on synthetic data sets with different numbers of Gaussian components in each GMM and different data dimensionality. The run-time of all similarity measures has a quadratic relation with the component number and a linear dependence with data dimensionality. PmG has a similar performance with ES, C2 and L2. With the increase of components number, EMD and Netflow gain more than PmG in time-cost. For varying dimensionality, the tendencies of all the similarity measures are very similar.

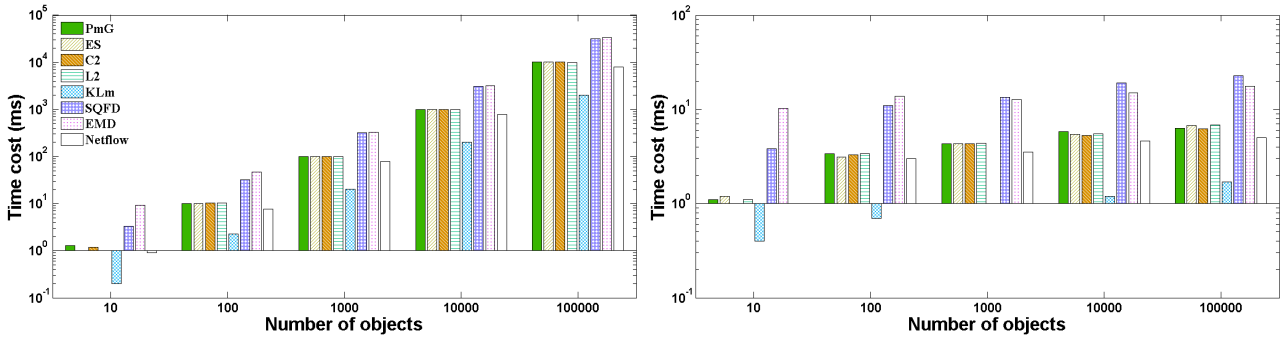
Comparing the query efficiency of linear scan and metric tree, we illustrate the time-cost of queries in Figure 4. Figure 4(a) demonstrates the linear relation between the query time and the number of data objects in linear scan queries. When



(a) Varying the number of Gaussian components.

(b) Varying data dimensionality.

Figure 3: Time cost of linear scan queries on synthetic data sets. In (a), the data dimensionality is fixed as two. The numbers of Gaussian components are set as ten in (b).



(a) Time cost of linear scan queries.

(b) Time cost of VP-tree queries.

Figure 4: Time cost of queries when varying the number of data objects. Each GMM used here has ten Gaussian components in a two-dimensional space. The capacity of nodes in the VP-tree is set to 32.

using VP-tree, as shown in Figure 4(b), there are great improvements in query efficiency for all the similarity measures. However, only PmG, EMD and Netflow can guarantee query accuracies among them.

V. Related Work

This section gives a survey and discussion of similarity measures for GMMs in previous work. Firstly we summarize approximation approaches for GMMs, then we discuss similarity measures that have closed-form expressions.

The Kullback-Leibler divergence [15] is a common way to measure the distance between two PDFs. It is given by $d_{\mathrm{KL}}(f_1\|f_2) = \int f_1(x)\log\frac{f_1(x)}{f_2(x)}\mathrm{d}x$. For the properties of metric, the KL divergence only satisfies the non-negativity property, although its symmetric version ($|\frac{1}{2}d_{\mathrm{KL}}(f_1\|f_2)+\frac{1}{2}d_{\mathrm{KL}}(f_2\|f_1)|$) also satisfies the symmetry property. Moreover, it has a closed-form expression for Gaussian distributions, but no such expression for GMMs exists.

To compute the distance between GMMs by the KL divergence, several approximation methods have been proposed. A commonly used approximation to $d_{\mathrm{KL}}(f_1\|f_2)$, the Gaussian approximation, replaces f_1 and f_2 with two Gaussian distributions, whose means and covariance matrices depend on those of GMMs. Another popular way is to use the minimum KL divergence of Gaussian components that are included in two GMMs. Moreover, Hershey et al. [11] have proposed the product of Gaussian approximation and the variation approximation, but the former tends to greatly underestimate $d_{\mathrm{KL}}(f_1\|f_2)$ while the latter does not satisfy the positivity property. Besides, Goldberger et al. [9] have proposed KLm and the unscented transformation based KL divergence(KLt). KLm works well when the Gaussian elements are far apart, but it cannot handle the overlapping situations, which are very common in real-world data sets. KLt solves the overlapping problem based on a non-linear transformation. Cui et al. [16] have compared the six approximation methods for KL divergence with Monte Carlo sampling, where the variation approximation achieves the best result quality, while KLm give a comparable result with a much faster speed.

Besides the approximation similarity methods for GMMs, several methods with closed-form expression have been proposed. Helén et al. [10] have described a squared Euclidean distance, which integrates the squared differences over the whole feature space. It has a closed-form expression for GMMs. Sfikas et al. [13] have presented a KL divergence based distance C2 and a Bhattacharyya-based distance for GMMs. Jensen et al. [12] used a normalized L2 distance to

 ISBN: 978-1-61208-558-6

measure the similarity of GMMs in mel-frequency cepstral coefficients from songs. Beecks et al. [22] have proposed SQFD for GMMs to model the similarities between images. However, none of these similarity measures with closed-form expression for GMMs obeys the triangle inequality.

VI. Discussion and Conclusions

In this paper, we have proposed PmG for GMMs. As a metric, PmG enables storing GMMs in any metric tree and applying analysis techniques that require the properties of triangle inequality. In our experimental evaluations on real-world data sets, we have demonstrated the effectiveness of the proposed similarity measure. PmG outperform the other measures on different types of data sets in both classification and clustering.

Due to the potentially different number of Gaussian components in GMMs, there is still not much specialized indexing structure for GMMs exist. Using matching probability as the similarity measure, Böhm et al. [23] and Zhou et al. [24] have decomposed each GMM into its components to support the indexing of GMMs. A specialized dynamic index for GMMs using a metric or pseudometric is a promising perspective.

Acknowledgment

We thank Thomas Abeel for sharing the code implementations of k-medoids algorithm, Damien Di Fede for sharing his implementation of Fast Fourier Transform, Mzechner for his/her implementation of audio file processing.

References

[1] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," IEEE Signal Process. Lett., vol. 13, no. 5, 2006, pp. 308–311.

[2] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," Digital Signal Processing, vol. 10, no. 1-3, 2000, pp. 19–41.

[3] P. KaewTraKulPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in Video-based surveillance systems. Springer, 2002, pp. 135–144.

[4] Z. Zivkovic, "Improved Adaptive Gaussian Mixture Model for Background Subtraction," in ICPR, 2004, pp. 28–31.

[5] D. Reynolds, "Gaussian mixture models," Encyclopedia of Biometrics, 2015, pp. 827–832.

[6] S. Ou, C. Lee, V. S. Somayazulu, Y. Chen, and S. Chien, "Low complexity on-line video summarization with gaussian mixture model based clustering," in ICASSP, 2014, pp. 1260–1264.

[7] J. E. Rougui, M. Gelgon, D. Aboutajdine, N. Mouaddib, and M. Rziza, "Organizing Gaussian mixture models into a tree for scaling up speaker retrieval," Pattern Recognition Letters, vol. 28, no. 11, 2007, pp. 1314–1319.

[8] P. N. Yianilos, "Data structures and algorithms for nearest neighbor search in general metric spaces," in ACM/SIGACT-SIAM SODA, 1993, pp. 311–321.

[9] J. Goldberger, S. Gordon, and H. Greenspan, "An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures," in ICCV, 2003, pp. 487–493.

[10] M. L. Helén and T. Virtanen, "Query by example of audio signals using euclidean distance between gaussian mixture models," in ICASSP (1), 2007, pp. 225–228.

[11] J. R. Hershey and P. A. Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in ICASSP, 2007, pp. 317–320.

[12] J. H. Jensen, D. P. W. Ellis, M. G. Christensen, and S. H. Jensen, "Evaluation of distance measures between gaussian mixture models of mfccs," in ISMIR, 2007, pp. 107–108.

[13] G. Sfikas, C. Constantinopoulos, A. Likas, and N. P. Galatsanos, "An analytic distance metric for gaussian mixture models with application in image retrieval," in ICANN (2), 2005, pp. 835–840.

[14] S. Zeng, R. Huang, H. Wang, and Z. Kang, "Image retrieval using spatiograms of colors quantized by gaussian mixture models," Neurocomputing, vol. 171, 2016, pp. 673–684.

[15] S. Kullback, Information theory and statistics. Courier Dover Publications, 2012.

[16] S. Cui and M. Datcu, "Comparison of kullback-leibler divergence approximation methods between gaussian mixture models for satellite image retrieval," in IGARSS, 2015, pp. 3719–3722.

[17] "Uci archive: Activity recognition data," http://archive.ics.uci.edu/ml/machine-learning-databases/00287/, accessed: 2017-03-27.

[18] "Aloi image data," http://aloi.science.uva.nl/, accessed: 2017-03-27.

[19] "Speaker recognition data," http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Audio/Main/16kHz_16bit/, accessed: 2017-03-27.

[20] "Weather of airports data," https://drive.google.com/open?id=0B3LRCuPdnX1BZ0d4RXIxMDVzakE, accessed: 2017-03-27.

[21] "Synthetic data sets," https://drive.google.com/open?id=0B3LRCuPdnX1BMzlHaUJYSFlpU1U, accessed: 2017-03-27.

[22] C. Beecks, A. M. Ivanescu, S. Kirchhoff, and T. Seidl, "Modeling image similarity by gaussian mixture models and the signature quadratic form distance," in ICCV, 2011, pp. 1754–1761.

[23] C. Böhm, P. Kunath, A. Pryakhin, and M. Schubert, "Querying objects modeled by arbitrary probability distributions," in SSTD, 2007, pp. 294–311.

[24] L. Zhou, B. Wackersreuther, F. Fiedler, C. Plant, and C. Böhm, "Gaussian component based index for GMMs," in ICDM, 2016, pp. 1365–1370.

 ISBN: 978-1-61208-558-6