Variational Inference of Infinite Generalized Gaussian Mixture Models using feature selection

Srikanth Amudala
Concordia Institute for
Information Systems Engineering
Concordia University
Montreal, Canada
srikanth.amudala@mail.concordia.ca

Nizar Bouguila
Concordia Institute for
Information Systems Engineering
Concordia University
Montreal, Canada
nizar.bouguila@concordia.ca

Abstract

This paper presents a variational Bayesian learning framework for the infinite generalized Gaussian mixture(IGGMM) model. Our work is motivated with the flexibility of shape parameter in the generalized Gaussian distribution(GGD) that has proven its capability to model complex multi-dimensional data. We also incorporate feature selection to consider the features that are most appropriate in constructing an approximate model in terms of classification and clustering accuracy. Experimental results on medical and image categorization dataset show the effectiveness of the proposed algorithm.

Keywords

Data clustering, Mixture models, Variational Bayesian inference, Generalized Gaussian Distribution, Classification, Image categorization.

1 Introduction

Statistical inference plays an important role in many research areas such as computer vision, signal processing, and pattern recognition. Mixture models are widely incorporated for dealing with Parameter learning. For example, the most famous technique [1] of all is the Expectation-Maximization (EM) algorithm. However, the EM algorithm also suffers from overfitting and has a high dependency on initialization [2]. This trade-off the efficiency of the learning algorithm and impacts the accuracy of the model.

Aside from this, an issue that normally emerges, particularly when dealing with high-dimensional data is the detection of the salient features. Naturally, salient features are those that encourage the modelling task and produce efficient outcomes. Uniform or unimodal features from these high dimensional data are irrelevant to clustering. Moreover, having higer number of features may confuse the model by increasing the model complexity [3] [4]. This suggests that choosing the features and selecting the number of mixutre components should be addressed simultaneously.

To address both feature and model selection, we present a novel mathematical model by extending Generalized Gaussian Mixture Model (GGMM) to the infinity to address the selection of right number of mixture components. This reduces the computation cost and also gives a good approximation of the underlying distribution for the data. We employ the model proposed in [5], a feature saliency determination process, where each feature is weighted up to a probability ranging between zero and one.

This paper is organized as follows. In Section 2 we introducte the mathematical model and variational learning of the IGGMM and the experimental results are explained in Section 3. The paper is concluded in Section 4.

2 Mathematical Model

The finite GGMM for a variable $\mathcal{X} \epsilon R$ with parameters mean(μ), precision(τ) and shape(λ) is defined as follows:

$$P(\mathcal{X}|\mu,\tau,\lambda) = \frac{\lambda\tau^{\frac{1}{\lambda}}}{2\Gamma(\frac{1}{\lambda})} e^{-\tau|(\mathcal{X}-\mu)|^{\lambda}} \quad (1)$$

where $\tau = \left(\frac{1}{\sigma}\sqrt{\frac{\Gamma(\frac{3}{\lambda})}{\Gamma(\frac{1}{\lambda})}}\right)^{\lambda}$, $\Gamma(.)$ denotes the gamma function given by $\Gamma(z) = \int_0^{\infty} p^{z-1}e^{-p}dp$, where z and p are real variables. The shape of the probability density function is determined by the shape parameter λ. The larger the value, the flatter the probability density function. This means that the decay rate of the density func-

tion is determined by λ. Note that for the two special cases, when $\lambda = 2$ and $\lambda = 1$, the GGD is reduced to the Gaussian and the Laplacian distributions, respectively. If $\mathcal{X}$ follows a mixture of K GGDs, then

$$P(\mathcal{X}|\Theta) = \sum_{k=1}^{K} P(\mathcal{X}|\mu_k, \tau_k, \lambda_k)\pi_k \tag{2}$$

where $\pi_k (0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1)$ are the mixing weights and $p(\mathcal{X}|\mu_k, \tau_k, \lambda_k)$ is the GGMM likelihood of component k. As for the symbol $\Theta = (\epsilon, \pi)$, it refers to the entire set of parameters to be estimated where $\epsilon = (\mu_1, \tau_1, \lambda_1, ..., \mu_K, \tau_K, \lambda_K)$ and $\pi = (\pi_1, ..., \pi_K)$.

Considering N observations, $\mathcal{X} = (X_1, X_2, ..., X_N)$, and supposing that the number of components K is known, the data likelihood is denoted as follows:

$$P(\mathcal{X}|\Theta) = \prod_{n=1}^{N} \sum_{k=1}^{K} P(X_n|\epsilon_k)\pi_k \tag{3}$$

where $\epsilon_k = (\mu_k, \tau_k, \lambda_k)$. For each variable X_i, let Z_i be K-dimensional vector known by the unobserved vector that assigns the appropriate mixture component X_i belongs to. Then, Z_{ik} is equal to 1 if X_i belongs to class k and 0, otherwise. Hence, the complete-data likelihood is given as follows:

$$P(\mathcal{X}|\Theta) = \prod_{n=1}^{N} \sum_{k=1}^{K} (P(X_n|\epsilon_k)\pi_k)^{Z_{nk}} \tag{4}$$

2.0.1 Feature Selection

Feature selection is an essential process in a mixture model as some features in the data do not necessarily have importance in clustering. Assume $\mathcal{X} = (X_1, X_2, ..., X_N)$ is a dataset of points, each X_N is a real factor vector in a d-dimensional space. We wish to train a mixture model by modelling this data and we expect that each mixture component density is factorized over the features. Hence, the features are considered to be independent for each mixture component. As we know, Not all the features might be relevant for modelling while some of the features may be more useful. Rather than expecting that there is a deterministic separation among the useful and non-useful features, we assume that a feature is useful up to a weight ranging from between 0 and 1.

Thus, given a few mixture components, we assume that a feature of $\mathcal{X}$ is drawn from a mixture of two univariate subcomponents, as proposed in [4]. The first subcomponent that is distinctive for every mixture component produces necessary information, while the second subcomponent that is basic to all mixture components generates the "noisy" information. Hence, the features follow the following distribution:

$$\begin{aligned} p(X|Z, \Theta, \zeta, S) = \prod_{n=1}^{N} \prod_{k=1}^{K} \Big[\prod_{i=1}^{d} p(X_i|\Theta_{ik})^{s_n} \\ p(X|\zeta_{ik})^{1-s_n} p(S|\epsilon) \Big]^{z_{nk}} \\ \text{where } \Theta = \{\mu, \tau, \lambda\}, \zeta = \{\epsilon, \delta, \omega\} \end{aligned} \tag{5}$$

$$\begin{aligned} p(X, Z, &\pi, \mu, \tau, \lambda, \epsilon, \delta, \omega, S) = \\ &\prod_{n=1}^{N} \prod_{k=1}^{K} \Big[\prod_{i=1}^{d} p(X_i|Z_{nk}, \mu_{ki}, \tau_{ki}, \lambda_{ki})^{s_i^n} \\ &p(X_n|Z_{nk}, \epsilon_{ki}, \delta_{ki}, \omega_{ki})^{1-s_i^n} \Big] \end{aligned} \tag{6}$$

The model parameter mean(μ), precision(τ) and shape(λ) of the relevant subcomponents. Respectively, ϵ, δ, and ω are the sets of parameters for the irrelevant subcomponent. The saliency of features is expressed through the hidden variables s_i^n , where $s_i^n \in \{0, 1\}$. If the value of s_i^n is one, then the i^{th} feature of X_N has been generated from the relevant subcomponent; otherwise, it has been generated from the irrelevant subcomponent. The distribution of the hidden variable S given the probabilities $w = \{w_i\}$ (feature saliencies) is given as follows:

$$p(S|w) = \prod_{n=1}^{N} \prod_{i=1}^{d} w_i^{s_i^n} (1 - w_i)^{1-s_i^n} \tag{7}$$

2.1 Bayesian Learning

With in the framework of Variational Expectation-Maximization (VEM) we have proposed a variational inference approach for the GGMM [6] [7]. The closed-form updates and the automatic determination of the mixture components (k) is obtained by optimizing the Kullback–Leibler (KL) divergence between the true posterior p and the appropriate distribution q [7]. The stronger the relationship between the distribution the smaller is the KL divergence. The KL divergence is denoted by:

$$\begin{aligned} KL(p \parallel q) &= -\int q(Z) \ln\{\frac{p(Z, \mathcal{X})}{q(Z)} - \ln p(\mathcal{X})\} dZ \\ &= -\int q(Z) \ln\{\frac{p(Z, \mathcal{X})}{q(Z)}\} dZ + \ln p(\mathcal{X}) \end{aligned} \tag{8}$$

Calculating the evidence $\ln p(\mathcal{X})$ will give us the KL divergence. However, this is complex to calculate which motivates the proposed algorithm.

$$\ln p(\mathcal{X}) = KL(p \parallel q) + \underbrace{\int q(Z) \ln\{\frac{p(Z,\mathcal{X})}{q(Z)}\} dZ}_{\text{Evidence Lower Bound}} \quad (9)$$

$$\mathcal{L} = \int q(Z) \ln\{\frac{p(Z,\mathcal{X})}{q(Z)}\} dZ \quad (10)$$

Maximizing the Evidence Lower Bound (ELBO) is equivalent to minimizing the KL divergence. By applying Jensen's inequality, the ELBO serves as a lower-bound for the log-evidence, $\ln p(\mathcal{X}) \geq \text{ELBO}(q)$ for any $q(Z)$, which is approximate of the posterior. To maximize the ELBO, we need to choose a variational family q. The complexity of the family determines the flexibility in providing an appropriate approximation to the true posterior distribution.

As a result of the nonlinearity of the shape parameter, the conjugate prior distribution can not be found directly. Therefore, we considered utilizing the Taylor estimation to find an approximate lower bound of the complete data log-likelihood to determine whether an appropriate prior exists in the exponential family.

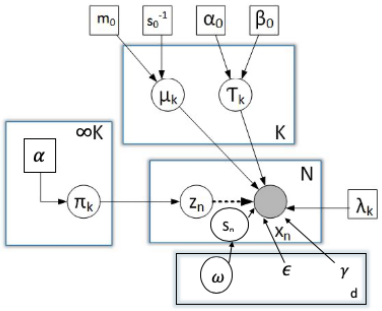

Figure 1: Graphical model for the VIGGMM with feature selection. The filled circle, unfilled circles and squares represent observations, random variables, and parameters, respectively. The dependency among the variables is represented by directional arrows.

However, the second-order derivative of the function $q(\lambda)$ is negative making the function concave, resulting in an upper bound rather than a lower bound; which is required. Hence, λ is considered as a parameter and is not assigned any prior distribution [8]. Normal and Gamma distributions are assigned as conjugate exponential priors for μ and τ. This can be observed in Fig. 1 and the conjugate priors for all the model parameters are given as follows:

$$q^\star(\mu) = \prod_{k=1}^{K} \prod_{i=1}^{d} N(\mu_{ik} | m_{ik}, s_{ik}^{-1}) \quad (11)$$

$$q^\star(\tau) = \prod_{k=1}^{K} \prod_{i=1}^{d} G(\tau_{ik} | \alpha_{ik}, \beta_{ik}) \quad (12)$$

$$q^\star(S) = \prod_{n=1}^{N} \prod_{i=1}^{d} \eta_{in}^{s_i^n} (1-\eta_{in})^{1-s_i^n} \quad (13)$$

$$q^\star(\pi) = Dir(\pi|\gamma) \quad (14)$$

$$\gamma_k = \gamma_0 + N_k \quad (15)$$

$$\mathbb{E}[\ln \pi_k] = \psi(\gamma_k) - \psi(\hat{\gamma}) \quad (16)$$

$$\hat{\gamma} = \sum_{k=1}^{K} \gamma_k \quad (17)$$

We consider a variational distribution which factorizes into the latent variables and the parameters:

$$\ln q^\star(Z) = \mathbb{E}_{\mu,\tau,\pi,s}[\ln p(\mathcal{X}, \pi, \mu, \tau, \lambda, \epsilon, \delta, \omega, S)] + const. \quad (18)$$

where $\mathbb{E}$ represents the expectation with respect to the subscripted parameter and $const$ denotes an additive constant. Substituting the two conditional distributions, and absorbing any terms that are independent of Z into the additive constant, we have:

$$\ln q^\star(Z) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + const \quad (19)$$

where we define:

$$\begin{aligned} \ln \rho_{nk} = & \mathbb{E}_\pi[\ln \pi_k] + \mathbb{E}_{\mu,\tau,s} \Bigg[\\ & s_{nk} \Bigg(\ln \frac{\lambda_k \tau_k^{\frac{1}{\lambda_k}}}{2\Gamma(\frac{1}{\lambda_k})} - \tau_k |X - \mu_k|^{\lambda_k} \Bigg) + \\ & (1 - s_{nk}) \Bigg(\ln \frac{\Omega_k \Lambda_k^{\frac{1}{\delta_k}}}{2\Gamma(\frac{1}{\Omega_k})} - \Lambda_k |X_n - \delta_k|^{\Omega_k} \Bigg) \Bigg] \end{aligned} \quad (20)$$

Normalizing the distribution, noting for each value of n the values of Z_{nk} are binary and add up to 1 overall values of k, we obtain:

$$q^\star(Z) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{z_{nk}} \quad (21)$$

The variational parameters $r_{nk}, m_{ik}, s_{ik}^{-1}, \alpha_{ik}, \beta_{ik}, \eta_{in}$ are obtained by maximizing and determining the densities involved in q. The variational parameters are

defined using the expected values of $z_{nk}, \mu_{ik}, s_{ik}^{-1}, s_i^n$ and functions of them. The following equations are obtained after deriving the expectation from $q^\star(Z), q^\star(\mu), q^\star(\tau), q^\star(S)$

$$r_{nk} = \frac{\rho_{nk}}{\sum_{k=1}^{K} \rho_{nk}} \tag{22}$$

$$\eta_{in} = \frac{w_i \hat{\eta_{in}}}{w_i \hat{\eta_{in}} + (1 - w_i)\varepsilon_{in}} \tag{23}$$

$$\begin{aligned}\hat{\eta}_{in} = \exp\Bigg\{ & \frac{1}{2}\sum_{k=1}^{K} r_{nk}[\psi(\alpha_{ik}) - \log\beta_{ik}] \\ & - \frac{1}{2}\sum_{k=1}^{K} r_{nk}\frac{\alpha_{ki}}{\beta_{ki}}[(x_i^n - m_{ik})^2 + \tau_{ki}]\Bigg\}\end{aligned} \tag{24}$$

$$\varepsilon_{in} = \exp\left\{ -\frac{1}{2}\gamma_i(x_i^n - \epsilon_i)^2 + \frac{1}{2}\log\gamma_i \right\} \tag{25}$$

$$N_k = \sum_{n=1}^{N} r_{nk} \tag{26}$$

$$m_{ik} = \frac{\frac{s_0 m_0}{2} + t_1}{s_{ik}} \tag{27}$$

$$s_{ik} = \frac{s_0}{2} + t_2 \tag{28}$$

where t_1, t_2 have two different cases as follows:

$$t_1 = \begin{cases} \sum_{n=1}^{N}(r_{nk}\bar{s}_n\bar{\tau}_{ik}\frac{\lambda_{ik}}{4}(\lambda_{ik}-1)\mu_{ik}^{\lambda_{ik}-3}x_n^2 + \\ \sum_{n=1}^{N}(r_{nk}\bar{s}_n\bar{\tau}_{ik}\frac{\lambda_k}{2}\mu_{ik}^{\lambda_k-2}x_n)), \text{if } X_n < m_k \\ \\ \sum_{n=1}^{N} r_{nk}\bar{s}_n\bar{\tau}_k\lambda_k\frac{|x_n|^{\lambda_k}}{x_n}, \text{otherwise} \end{cases}$$

$$t_2 = \begin{cases} \sum_{n=1}^{N}(r_{nk}\bar{s}_n\bar{\tau}_{ik}\mu_{ik}^{\lambda_{ik}-2}), \text{if } X_n < m_{ik} \\ \\ \sum_{n=1}^{N}(r_{nk}\bar{s}_n\bar{\tau}_{ik}\frac{\lambda_{ik}}{2}(\lambda_{ik}-1)\frac{|x_n^{\lambda_{ik}}|}{x_n^2}), \text{otherwise} \end{cases}$$

Where $\bar{\tau}$ represents $\mathbb{E}_\tau[\tau]$.

$$\alpha_{ik} = \sum_{n=1}^{N} \bar{s}_n r_{nk} + \alpha_0 - 1 \tag{29}$$

$$\beta_{ik} = \beta_0 + \sum_{n=1}^{N} \bar{s}_n r_{nk}\, \mathbb{E}_\mu[|X_n - \mu_{ik}|^{\lambda_{ik}}] \tag{30}$$

$$\mathbb{E}_\mu[|X_n - \mu_{ik}|^{\lambda_{ik}}] = \begin{cases} |X_n|^{\lambda_{ik}} - \lambda_{ik}\frac{|X_n|^{\lambda_{ik}}}{X_n}m_{ik} + \\ \frac{\lambda_{ik}(\lambda_{ik}-1)}{2}\frac{|X_n|^{\lambda_{ik}}}{X_n^2}(\frac{1}{s_{ik}} + m_{ik}^2), \\ \text{if } X_n > \mu_{ik} \\ \\ \mathbb{E}[|\mu_{ik}|^{\lambda_{ik}} - \lambda_{ik}\mu_v^{\lambda_{ik}-1}X_n + \\ \frac{\lambda_{ik}}{2}(\lambda_{ik}-1)\mu_{ik}^{\lambda_{ik}-2}X_n^2], \text{otherwise} \end{cases}$$

Then using confluent hypergeometric function:

$$\begin{aligned}&\mathbb{E}\left[|\mu_{ik}|^{\lambda_{ik}}\right] = \\ &(\frac{1}{\sqrt{s_{ik}}})^{\lambda_{ik}} \cdot 2^{\lambda_{ik}/2}\frac{\Gamma\left(\frac{1+\lambda_{ik}}{2}\right)}{\sqrt{\pi}}{}_1F_1\left(-\frac{\lambda_{ik}}{2}, \frac{1}{2}, -\frac{1}{2}(m_{ik})^2 s_{ik}\right).\end{aligned} \tag{31}$$

After the maximization of Lowerbound $\mathcal{L}$ with respect to Q, the second step of the method requires maximization of $\mathcal{L}$ with respect to π_k, w_i, ϵ_i, and γ_i. Setting the derivative of $\mathcal{L}$ with respect to the parameters equal to zero [5], we get the following update rules:

$$\pi_k = \frac{1}{N}\sum_{n=1}^{N} r_{nk} \tag{32}$$

$$w_i = \frac{1}{N}\sum_{n=1}^{N} \eta_{in} \tag{33}$$

$$\epsilon_i = \frac{\sum_{n=1}^{N}\eta_{in}x_i^n}{\sum_{n=1}^{N}\eta_{in}} \tag{34}$$

$$\frac{1}{\gamma_i} = \frac{\sum_{n=1}^{N}\eta_{in}(x_i^n - \epsilon_i)^2}{\sum_{n=1}^{N}\eta_{in}} \tag{35}$$

Given the posterior distributions from the VE-step, the VM-step updates the parameters by maximizing the approximate lower bound $\mathcal{L}$. To estimate the parameters of the GGMM (i.e. λ),

$$\begin{aligned}\lambda_k^\star &= \lambda_k + s\Delta\lambda_k \\ \text{where } \Delta\lambda_k &= -\frac{\mathcal{L}_k^{'}(q,\Theta)}{\mathcal{L}_k^{''}(q,\Theta)}\end{aligned} \tag{36}$$

where s is determined by the backtracking line search [9].

2.2 Infinite Mixture Model

The mixing weights, π_k, are considerded to be a symmetric Dirichlet prior with concentration parameter α/k

$$p(Z|\pi) = \prod_{k=1}^{K}\pi_k^{n_k} \tag{37}$$

$$\begin{aligned}p(\pi|\alpha) &\sim Dirichlet\left(\frac{\alpha}{K}, .., \frac{\alpha}{K}\right) \\ &= \frac{\Gamma(\alpha)}{\Gamma(\frac{\alpha}{K})^K}\prod_{k=1}^{K}\pi_k^{\frac{\alpha}{K}-1}\end{aligned} \tag{38}$$

Using the standard Dirichlet integral, we can integrate the mixing weights and the prior can be directly written

in terms of indicators.

$$p(Z|\alpha) = \int p(Z|\pi)p(\pi|\alpha)d\pi \\ = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{k=1}^{K} \frac{\Gamma(\frac{\alpha}{K} + n_k)}{\Gamma(\frac{\alpha}{K})} \tag{39}$$

The conditional prior for the single indicator from Eq. 39 by keeping all but a single indicator fixed:

$$p(Z_{nk} = 1|\alpha, Z_{-n}) = \frac{n_{-nk} + \alpha/K}{N - 1 + \alpha} \tag{40}$$

where the subscript $-n$ indicates all indexes except n and n_{-nk} is the number of observations, excluding X_i, that are associated with component k.

So far, we considered the number of mixtures K as a fixed finite quantity, we will extend the model by updating the posteriors in Eq. 40 with $K \to \infty$

$$p(Z_{nk} = 1|\alpha, Z_{-n}) = \begin{cases} \frac{n_{-nk}}{N-1+\alpha} \text{ if } n_{-nk} > 0 \\ \frac{\alpha}{N-1+\alpha} \text{ if } n_{-nk} = 0 \end{cases} \tag{41}$$

where $n_{nk} > 0$ occurs when mixture k is represented. A sample X_i is associated with an existing component by a certain probability proportional to the number of samples already allocated to this component; while an unrepresented mixture component is proportional to α and N. We combine the likelihood from Eq. 1 on the indicators with the prior from Eq. 42 to obtain the conditional posteriors for the indicators

$$p(Z_{nk} = 1|...) = \\ \begin{cases} \frac{n_{-nk}}{N-1+\alpha} p(\mathcal{X}|\Theta) & \text{if } n_{-nk} > 0 \\ \frac{\alpha}{N-1+\alpha} \int p(\mathcal{X}|\Theta)p(\Theta|\mu,\tau)d\Theta & \text{if } n_{-nk} = 0 \end{cases} \tag{42}$$

Our complete algorithm can then be summarized as follows:

Algorithm

1. Initialize the parameters and the assignments.
2. **loop**
3. Update mixture parameters $\mu_k, \tau_k, S, \lambda_k$ from the posteriors in Eq. 11, Eq. 12, Eq. 13 and Eq. 36
4. Update hyperparameters $m_k, s_k, \alpha_{ik}, \beta_{ik}, \eta_{in}$ and Dirichlet process concentration parameter α from Eq. 23 to Eq. 36
5. Update the indicators conditioned on the other indicators and the hyperparameters from Eq. 42
6. The convergence criteria is reached when the difference of the current value of joint posteriors and the previous value is less than $1e-9$. Otherwise, repeat above loop until convergence
7. **end**

3 Experimental results and discussion

We evaluate the built variational IGGMM (VIGGMM) using two different datasets focused on image categorization and binary classification. We compare the effectiveness of the model based on Gaussian mixutre model (GMM) and Variational Gaussian mixture model (VGMM).

3.1 Image categorization

Image categorization in terms of automation plays an important role in multimedia applications [10]. Determining the patterns plays a vital role in multimedia applications. For our application we choose the Caltech 101 [11] objects dataset.

Among the 101 categories, we chose four categories: Bikes, Yin Yang, Sunflowers, Aeroplanes. All the categories have 60 images each to have a balanced dataset. Sample images of these categories are shown in Fig. 2. Also, to evaluate the robustness of our model, all the categories that are conidered have a similar landscape.

To implement our model on the images we need to initially create a bag of visual words model [12] [13]. To create a bag of visual words model, we need to initially extract some kind of descriptors from the images. The most commonly utilized descriptors are SIFT [14], SURF [15], HOG [16], and so forth. For our situation we found the SIFT descriptors to be an effective choice. Consequently, we first extract the SIFT features from the images and perform K-means clustering over the extracted SIFT descriptors to form the bag of the words feature vector for each image. This is utilized as input to our model. Table 1 shows the performance of our model by considering 200 features from the bag of words feature vector and compared to the other models. We can see that our model VIGGMM with feature selection has performed better than all the other models. Table 2 shows similar results when we considered using 300 features from the bag of words feature vector.

In order to verify the performance of our model in binary classification situation, we choose just the Bikes and the Aeroplan images from the Caltech 101 dataset which has an almost similar landscape. Table 3 shows the performance of our model by considering 200 features from

the bag of words feature vector and compared to the other models.

(a) Bike

(b) Yin Yang

(c) Sunflower

(d) Aeroplan

Figure 2: Sample images from different categories of Caltech 101 dataset

Table 1: Accuracy of different models for Caltech 101 dataset

Method	**Features**	**Accuracy**(%)
GMM	200	33
VGMM	200	23
VIGGMM	200	74

Table 2: Accuracy of different models for Caltech 101 dataset

Method	**Features**	**Accuracy**(%)
GMM	300	24
VGMM	300	23
VIGGMM	300	72

3.2 Binary Classification

We apply our VIGGMM estimation algorithm for binary classification in medical applications involving detection of heart diseases[1]. The heart disease data set provides all the potential symptoms of a person with positive heart disease. This database contains 76 attributes, but all distributed tests refer to employing a subset of 14. The objective field alludes to the presence of heart infection within the patient.

[1]https://www.kaggle.com/ronitf/heart-disease-uci.

Table 3: Accuracy of different models for binary classification using Caltech 101 dataset

Method	**Features**	**Accuracy**(%)
GMM	200	60
VGMM	200	68
VIGGMM	200	92

We have implemented our VIGGMM classifier using cross-validation with the split size of 4. The label for each data point is determined with the largest component among the likelihood of the data point belonging to the classes. Table 4, presents the model accuracy in comparison with Gaussian Mixture Model(GMM), Variational Gaussian Mixture Model(VGMM).

Table 4: Accuracy of different models for binary classification using Heart Disease UCI dataset

Method	**Features**	**Accuracy**(%)
GMM	13	50
VGMM	13	57
VIGGMM	13	76

4 Conclusion

We have presented an variational inference approach for IGGMM with feature selection by considering the shape parameter as a variable and using Binomial Expansion, we estimate the expectation of the distributions. Hence, the posterior distributions of the inference can be updated by the corresponding hyperparameters. In the VM-step, the shape parameter is updated using the single-step update of Newton's method. Also, we address the challenges of parameter learning and choosing the correct number of mixture components by introducing Bayesian learning and extension of the GGMM model to infinity.

Experimental results show that the VIGGMM with feature selection is an accurate model for image categorization and medical applications by effectively estimating the parameters. Moreover, our model outperformed Gaussian and variational Gaussian mixture models which are known to be industry standards.

References

[1] N. Bouguila and D. Ziou, "Unsupervised selection of a finite dirichlet mixture model: an mml-based

approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 8, pp. 993–1009, 2006.

[2] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, vol. 2. IEEE, 1999, pp. 246–252.

[3] M. Harrison and A. Shirom, *Organizational diagnosis and assessment: Bridging theory and practice*. Sage Publications, 1998.

[4] M. H. Law, M. A. Figueiredo, and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 9, pp. 1154–1166, 2004.

[5] C. Constantinopoulos, M. K. Titsias, and A. Likas, "Bayesian feature and model selection for gaussian mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 6, pp. 1013–1018, 2006.

[6] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for bayesian inference," *IEEE Signal Processing Magazine*, vol. 25, no. 6, pp. 131–146, 2008.

[7] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[8] C. Liu, H.-C. Li, K. Fu, F. Zhang, M. Datcu, and W. J. Emery, "Bayesian estimation of generalized gamma mixture model based on variational em algorithm," *Pattern Recognition*, vol. 87, pp. 269–284, 2019.

[9] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[10] K. Maanicshah, N. Bouguila, and W. Fan, "Variational learning for finite generalized inverted dirichlet mixture models with a component splitting approach," in *2019 IEEE 28th International Symposium on Industrial Electronics (ISIE)*. IEEE, 2019, pp. 1453–1458.

[11] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *2004 conference on computer vision and pattern recognition workshop*. IEEE, 2004, pp. 178–178.

[12] T. Li, T. Mei, I.-S. Kweon, and X.-S. Hua, "Contextual bag-of-words for visual categorization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 4, pp. 381–392, 2010.

[13] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, no. 1-22. Prague, 2004, pp. 1–2.

[14] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[15] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.