21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

BAYESIAN FEATURE SELECTION AND MODEL DETECTION FOR STUDENT'S T-MIXTURE DISTRIBUTIONS

Hui Zhang[1,2], Q. M. Jonathan Wu[1], Thanh Minh Nguyen[1]
[1]*Department of Electrical and Computer Engineering, University of Windsor, Canada*
[2]*School of Computer & Software, Nanjing University of Information Science & Technology, China*

ABSTRACT

In this paper, we propose a novel method for feature selection and model detection using Student's t-distributions based on the variational Bayesian (VB) approach. First, our method is based on the Student's t-mixture model which has heavier tails than the Gaussian distribution and is therefore less sensitive to small numbers of data points and consequent precision-estimates of the components number. Second, the number of components, the local feature saliency and the parameters of the mixture model are simultaneously estimated by Bayesian variational learning.

1. INTRODUCTION

Variational methods have their origins in the 18th century with the work of Euler, Lagrange, and other scientists focus on the calculus of variations [1]. According to the variational methodology an approximation to the posterior of the hidden variables given the observations is used. Corduneanu and Bishop use Gaussian mixture model (GMM) and lower bound marginal likelihood for fitting mixture numbers [2]. Student's t-mixture model (SMM) has been proposed by Peel and Mclachlan recently as an alternative to GMM providing high robustness against noises [3]. The significant tolerance of SMM to training data has been experimentally depicted in many recent articles where it has been shown that SMM can model the hidden patterns of data significantly well [4]. Svensen and Bishop propose a Bayesian approach to Student's t-mixture model which is more robust than VB-GMM [5]. VB-GMM is a special case of VB-SMM which is less sensitive to outliers and can obtain robust estimates for the mean of a set of data points.

Methods mentioned above consider all available features with equal weight of the dataset. In practice, this can not be always true as some features can be irrelevant. An effective solution is feature selection which chooses the best feature for clustering. Recently, Law et al. [6] define feature saliency as a probability and assume these features are independent and follow a common distribution with Gaussian mixture models for clustering. The complement of this probability is the measure of feature saliency and the estimation is carried out by the Maximum Likelihood (ML) or Maximum A Priori (MAP) with the EM algorithm. Constantinopoulos et al. [7] and Li et al. [8, 9] adept the same Gaussian mixture model as in [6] to describe the feature saliency, but the former employ Variational Bayesian algorithm and the later propose localized feature saliency measure instead of global feature approaches.

2. FEATURE SELECTION FOR VBSMM

The Student's t-distribution provides a heavy-tailed alternative to the Gaussian distribution for potential outliers. The probability density function (pdf) of a Student's t-distribution with mean μ, precision τ and degrees of freedom v is [1]

$$t(x;\mu,\tau,v)=\frac{\Gamma(v/2+1/2)}{\Gamma(v/2)}\left(\frac{\tau}{\pi v}\right)^{1/2}\left[1+\frac{\tau(x-\mu)^2}{v}\right]^{-v/2-1/2}, \quad (1)$$

where $\Gamma(s)$ is the Gamma function,

$$\Gamma(s)=\int_0^\infty t^{s-1}e^{-t}dt\,. \quad (2)$$

It can be shown that the Student's t distribution $t(\mu, \tau, v)$ is equivalent to a Gaussian distribution $N(\mu, u\tau)$ with the precision scaling factor u which is given as

$$x\,|\,u\sim N(\mu,u\tau), \quad (3)$$

where the scaling factor u is a Gamma-distributed latent variable, depending on the degree of freedom v, defined

$$u\sim G\left(\frac{v}{2},\frac{v}{2}\right), \quad (4)$$

the pdf of the Gamma distribution $G(u;\alpha,\beta)$ is given by

$$G(u;\alpha,\beta)=\frac{1}{\Gamma(\alpha)}\beta^{\alpha}u^{\alpha-1}e^{-\beta u}\,. \quad (5)$$

We now consider variational Bayesian method to Student's t-mixture model with localized feature saliency through the binary latent variables z_{jn}, where

$z_{jn} \in \{0,1\}$ and $\sum_{j=1}^{J} z_{jn} = 1$. If a given data point x_{in} is generated from component j then the value of z_{jn} is one; otherwise, it is zero. The saliency of localized feature is expressed through the hidden variables s_{ijn}, where s_{ijn}=1 indicates that for data x_n, feature i is associated with component j, and s_{ijn}=0 otherwise. Then w_{ij}=Pr(s_{ijn}=1) is the probability that feature i is relevant to component j for all observed data X. The priori distribution is given as follows

$$p(X \mid \Theta) = \prod_{n=1}^{N} \prod_{j=1}^{J} \left[\prod_{i=1}^{d} t\left(x_{in}; \mu_{ij}, \tau_{ij}, v_{ij}\right)^{s_{ijn}} t\left(x_{in}; \varepsilon_{ij}, \gamma_{ij}, \eta_{ij}\right)^{1-s_{ijn}} \right]^{z_{jn}} . \quad (6)$$

The sets $\{\mu_{ij}\}$, $\{\tau_{ij}\}$, and $\{v_{ij}\}$ denote the mean, precision and degrees of freedom of the "useful" subcomponents. Correspondingly, $\{\varepsilon_{ij}\}$, $\{\gamma_{ij}\}$, and $\{\eta_{ij}\}$ are the sets of parameters for the "noise" subcomponents. With the latent variables u_{ijn} and λ_{ijn}, Student's t distribution can be equivalent to a Gaussian distribution as follows

$$p(X \mid \Theta) = \prod_{n=1}^{N} \prod_{j=1}^{J} \left[\prod_{i=1}^{d} N\left(x_{in}; \mu_{ij}, u_{ijn}\tau_{ij}\right)^{s_{ijn}} N\left(x_{in}; \varepsilon_{ij}, \lambda_{ijn}\gamma_{ij}\right)^{1-s_{ijn}} \right]^{z_{jn}} \quad (7)$$

The hidden variable Z is given the distribution governed by the mixing probabilities π=$\{\pi_j\}$ and the hidden variable S is given the distribution governed by the probabilities $w = \{w_{ij}\}$ as $p(Z \mid \pi) = \prod_{n,j=1}^{N,J} \pi_j^{z_{jn}}$, $p(S \mid w) = \prod_{n,j,i=1}^{N,J,d} w_{ij}^{s_{ijn}} \left(1 - w_{ij}\right)^{1-s_{ijn}}$, The model selection is accomplished by introducing conjugate priors over the means, precisions and degrees of freedom for "useful" subcomponents

$$p(\mu) = \prod_{j=1}^{J} \prod_{i=1}^{d} N\left(\mu_{ij}; m_i, c\right), \quad (8)$$

$$p(T) = \prod_{j=1}^{J} \prod_{i=1}^{d} G\left(\tau_{ij}; \alpha, \beta\right), \quad (9)$$

$$p(U) = \prod_{n=1}^{N} \prod_{j=1}^{J} \prod_{i=1}^{d} G\left(u_{ijn}; \frac{v_{ij}}{2}, \frac{v_{ij}}{2}\right). \quad (10)$$

Here, the hyperparameters m_i, c, α, β control the prior distributions over μ_{ij} and are chosen to give broad prior for τ_{ij} with "useful" subcomponents. Moreove, the actual model parameters are not sensitive to these hyperparameters on the near zero scale. It is noticed that there is no conjugate prior for v_{ij} since no prior is imposed on it. However, since there is only one optimal parameter for each mixture component, the learning approach used for the proposed model is discussed in the next part.

3. VARIATIONAL APPROXIMATION

In order to obtain the estimation of parameters, we maximize the marginal likelihood $p(X)$ by integrating out the variables as follows

$$p(X) = \int p(X, \Theta) d\Theta, \quad (11)$$

The integral denotes the joint integration over observed variables and the summation over hidden variables Z and S. Since the integration in the equation above is intractable, an alternative way to solve this problem is variational Bayes method which aims to maximize a lower bound L of the logarithmic marginal likelihood:

$$L(Q) = \int Q(\Theta) \log \frac{p(X, \Theta)}{Q(\Theta)} d\Theta \le \log p(X) . \quad (12)$$

where Q is an arbitrary distribution which provides an approximation to the true posterior distribution. Although the computation of original log likelihood function $logp(X)$ is not tractable, the lower bound $L(Q)$ may be tractable to compute through choosing a suitable form for the Q distribution. The difference between the lower bound $L(Q)$ and the true log likelihood $logp(X)$ is Kullback-Leibler (KL) divergence. For this purpose, we approximate P with optimizing the Q by minimizing the KL divergence.

Minimizing the KL divergence with respect to all possible function form of Q, the standard variational approach provides the following general form of the solutions

$$Q(\Theta_i) = \frac{\exp\left\langle p(X, \Theta) \right\rangle_{k \ne i}}{\int \exp\left\langle p(X, \Theta) \right\rangle_{k \ne i} d\Theta_i}, \quad (13)$$

where $\langle \cdot \rangle_{k \ne i}$ denotes an expectation with respect to the distributions $Q_k(\Theta_k)$ for all $k \ne i$. The factors of the variational posterior are given by calculation of (13) as follows

$$Q(Z) = \prod_{n=1}^{N} \prod_{j=1}^{J} r_{jn}^{z_{jn}}, \quad (14)$$

$$Q(\mu) = \prod_{j=1}^{J} \prod_{i=1}^{d} N\left(\mu_{ij}; m_{ij}^{\mu}, c_{ij}^{\mu}\right), \quad (15)$$

$$Q(T) = \prod_{j=1}^{J} \prod_{i=1}^{d} G\left(\tau_{ij}; \alpha_{ij}^{\tau}, \beta_{ij}^{\tau}\right), \quad (16)$$

$$Q(U) = \prod_{n=1}^{N} \prod_{j=1}^{J} \prod_{i=1}^{d} G\left(u_{ijn}; \alpha_{ijn}^{u}, \beta_{ijn}^{u}\right), \quad (17)$$

$$Q(S) = \prod_{n=1}^{N} \prod_{j=1}^{J} \prod_{i=1}^{d} \rho_{ijn}^{s_{ijn}} \left(1 - \rho_{ijn}\right)^{1-s_{ijn}} . \quad (18)$$

The variational parameters are given by maximizing and determining the density involved in Q

$$r_{jn} = \frac{\tilde{r}_{jn}}{\sum_{j=1}^{J} \tilde{r}_{jn}}, \quad (19)$$

$$\tilde{r}_{jn} = \pi_j \exp \times \left(\frac{1}{2} \sum_{i=1}^{d} \left\langle s_{ijn} \right\rangle \left\{ \left\langle \log \tau_{ij} \right\rangle + \left\langle \log u_{ijn} \right\rangle - \left\langle \left(x_{in} - \mu_{ij}\right)^2 \right\rangle \left\langle u_{ijn} \right\rangle \left\langle \tau_{ij} \right\rangle \right\} \right), \quad (20)$$

$$m_{ij}^{\mu}=\frac{cm_i+\langle\tau_{ij}\rangle\sum_{n=1}^{N}\langle z_{jn}\rangle\langle s_{ijn}\rangle\langle u_{ijn}\rangle x_{in}}{c+\langle\tau_{ij}\rangle\sum_{n=1}^{N}\langle z_{jn}\rangle\langle s_{ijn}\rangle\langle u_{ijn}\rangle}, \tag{21}$$

$$c_{ij}^{\mu}=c+\langle\tau_{ij}\rangle\sum_{n=1}^{N}\langle z_{jn}\rangle\langle s_{ijn}\rangle\langle u_{ijn}\rangle, \tag{22}$$

$$\alpha_{ij}^{\tau}=\alpha+\frac{1}{2}\sum_{n=1}^{N}\langle z_{jn}\rangle\langle s_{ijn}\rangle, \tag{23}$$

$$\beta_{ij}^{\tau}=\beta+\frac{1}{2}\sum_{n=1}^{N}\langle z_{jn}\rangle\langle s_{ijn}\rangle\langle u_{ijn}\rangle\langle(x_{in}-\mu_{ij})^2\rangle, \tag{24}$$

$$\alpha_{ijn}^{u}=\frac{v_{ij}+\langle z_{jn}\rangle\langle s_{ijn}\rangle}{2}, \tag{25}$$

$$\beta_{ijn}^{u}=\frac{v_{ij}+\langle z_{jn}\rangle\langle s_{ijn}\rangle\langle(x_{in}-\mu_{ij})^2\rangle\langle\tau_{ij}\rangle}{2}, \tag{26}$$

$$\rho_{ijn}=\frac{w_{ij}\tilde{\rho}_{ijn}}{w_{ij}\tilde{\rho}_{ijn}+(1-w_{ij})\xi_{ijn}}, \tag{27}$$

$$\tilde{\rho}_{ijn}=\exp\left\{\frac{1}{2}\langle z_{jn}\rangle\left[\langle\log\tau_{ij}\rangle+\langle\log u_{ijn}\rangle-\langle(x_{in}-\mu_{ij})^2\rangle\langle u_{ijn}\rangle\langle\tau_{ij}\rangle\right]\right\}, \tag{28}$$

$$\xi_{ijn}=\exp\left\{\frac{1}{2}\langle z_{jn}\rangle\left[\log\gamma_{ij}+\log\lambda_{ijn}-(x_{in}-\varepsilon_{ij})^2\lambda_{ijn}\gamma_{ij}\right]\right\}. \tag{29}$$

Since no conjugate prior is imposed on the degree of freedom, we update v_{ij} with the log-marginal maximum likelihood estimates, by setting the corresponding gradient to zero and solving the non-linear equations as follows

$$\log\left(\frac{v_{ij}}{2}\right)-\psi\left(\frac{v_{ij}}{2}\right)+1+\frac{\sum_n\langle z_{jn}\rangle\langle s_{ijn}\rangle\left[\langle\log u_{ijn}\rangle-\langle u_{ijn}\rangle\right]}{\sum_n\langle z_{jn}\rangle\langle s_{ijn}\rangle}=0. \tag{30}$$

where $\psi(x)$ is the digamma function $\psi(x)=d\log\Gamma(x)/dx$. Degree of freedom v_{ij} decides the shape of Student's distribution. It tends to a Gaussian distribution for large values of its degree of freedom and takes in to account the outliers with heavier tails for small value. Compared with previous methods based on GMM, our method needs additional iterative computation for the degrees of freedom.

With the variational factors in (14) to (18) and estimated variational parameters in (19) to (30), we can maximize the tight bound of the log marginal likelihood which equals to the minimization of the KL divergence. It is noticed that the variational factors are mutually coupled through their dependence on moments of the other factors, these parameters can be achieved by optimizing each factor in turn, holding the others fixed, with iteratively updating. We first initialize the approximating distributions, cycle round each factor in turn and then replace its current estimate by its optimal solution given the current estimates for the other factors.

It is noticed that the solutions for the factors of the variational posterior, and hence the value of the lower bound, depended on the values π_j, w_{ij}, ε_{ij}, and γ_{ij}. These parameters are obtained by setting the derivative of L with respect to π_j, w_{ij}, ε_{ij}, and γ_{ij} equal to zero, which leads to the following updated rules

$$\pi_j=\frac{1}{N}\sum_{n=1}^{N}r_{jn}, \tag{31}$$

$$w_{ij}=\frac{1}{N}\sum_{n=1}^{N}\rho_{ijn}, \tag{32}$$

$$\varepsilon_{ij}=\frac{\sum_{n=1}^{N}(1-\rho_{ijn})r_{jn}\lambda_{ijn}x_{in}}{\sum_{n=1}^{N}(1-\rho_{ijn})r_{jn}\lambda_{ijn}}. \tag{33}$$

$$\frac{1}{\gamma_{ij}}=\frac{\sum_{n=1}^{N}(1-\rho_{ijn})r_{jn}\lambda_{ijn}(x_{in}-\varepsilon_{ij})^2}{\sum_{n=1}^{N}(1-\rho_{ijn})r_{jn}}. \tag{34}$$

The previous equations are repeated alternatively until convergence, which can be monitored through inspection of the variational lower bound. The lower bound value does not decrease with the increasing of the iterative times. Here, parameters π_j and w_{ij} can be considered as "contribution" parameters for cluster numbers and feature saliencies. The property of variational Bayesian algorithm guaranties that the components with similar parameters fitting the same Student's t-distributions generating a dominant cluster. Thus, we can start the model with a large number of components; the competition among components finally yields a model which is composed by dominant components and the redundant components are removed. Meanwhile, feature saliency determines the subcomponent to be a "useful" subcomponent with value one or "noise" subcomponent with value zero. The update of the parameters π_j and w_{ij} identify the cluster numbers and feature saliencies simultaneously.

4. EXPERIMENTAL RESULTS

In this section, we experimentally evaluate our method (FSVB-SMM) in a set of synthetic data and real data. For comparison, we also evaluate the Bayesian Student's t-mixture model (VB-SMM) [5] and feature selection for the Gaussian mixture model (FSVB-GMM) [8]. The synthetic data set consists of 600 data points from a mixture of three Gaussian components, where m_1=(6, -1.5), m_2=(6, 1.5), m_3=(0,0). Here, each component contains 200 data points. These Gaussians are then embedded into subsets of 10-dimensional background (zero values) with standard Gaussian noise. Components can have a different number of features for the feature saliency test. Thus, we select features 1 and 3 from the background data and then replace the first

200 positions with component 1. Similarly, we embed component 2 with features 4 and 5, and component 3 with features 2 and 5 into the background. The test data is further augmented by 10% outliers, which are drawn from a uniform distribution on the interval [-10, 10]. We ran three different methods 10 times independently with the stopping threshold sets to 10^{-8}. All 10-dimentional data cannot be shown on a 2-D display, thus we selected data with "useful" features for effective display. We give the average mean values estimation in 10 run-times for the three methods in Table 1, where we can see that proposed method outperforms other methods. Our method yields a nearly perfect result, with the estimated mean values almost identical to the actual mean values of the modeled data distributions.

Table 1. Estimated mean value for different methods.

	VB-SMM	FSVB-GMM	FSVB-SMM
Component1	[6.14 -1.18]	[5.97 -0.22]	[5.96 -1.62]
Component2	[6.21 1.14]	[0.74 -0.21]	[6.02 1.43]
Component3	[-0.01 0.01]	[0.05 0.06]	[0.03 0.05]

Table 2. Average error rate and estimated components.

	FSVB-GMM		FSVB-SMM	
	Error rate (%)	Estimated *J*	Error rate (%)	Estimated *J*
Wine	7.4(1.96)	5.2(0.45)	6.2(0.78)	3.2 (0.45)
Heart	43.85(9.12)	5(0.71)	37.38(4.35)	4.8(0.84)
WPBC	38.28(5.77)	2.8 (0.84)	34.6(1.97)	2.6(0.55)
Yeast	64.28(1.50)	5.8 (1.10)	61.32(4.73)	7.4(1.14)

The numbers in parenthesis are the standard deviations of the corresponding quantities.

The second experiment is implemented with real data with four data sets, Wine, Heart, WPBC and Yeast, available from the well-known UCI machine learning repository [10]. After the normalization of UCI data, we add standard Gaussian noise to the available data to make the classification task more challenging. Here, all methods use the same initialization with the common *k*-means algorithm for 500 runs and start with initial 20 components. The error rate is calculated according to the samples in true class labels, and the estimated samples in evaluated components. In our experiment, it is calculated based on the average of 5 random runs and the lower value indicates better performance. The mean and standard deviation of the error rate, and the class components over 5 runs for different methods are shown in Table 2. It is clear that FSVB-SMM outperforms FSVB-GMM, with lower error rates and variances. The high error rate for the Yeast data set is because Proteins are inherently more difficult to cluster. Another reason is that some classes contain only 20 samples, and some even fewer, at 5 samples of total 1484 samples, which makes the classification more difficult.

5. CONCLUSION

In this paper, we propose a new VB-SMM model with the consideration of feature selection methodology. The parameters of SMM, the number of components and the localized feature saliency can be automatically determined by the constructed model. The experiments conducted under synthetic and real dates demonstrate the discriminative power of the proposed descriptors, especially in noise conditions, which clearly outperform existing methods.

ACKOWLEDGEMENTS

This work was supported in part by the Canada Chair Research Program and the Natural Sciences and Engineering Research Council of Canada. This work was supported in part by the National Natural Science Foundation of China under Grant 61105007 and 61103141.

REFERENCES

[1] C.M. Bishop, Pattern Recognition and Machine Learning. *Springer*, 2006.

[2] A. Corduneanu and C. M. Bishop, "Variational Bayesian model selection for mixture distributions," in *Artificial Intelligence and Statistics* 2001. San Mateo, CA: Morgan Kaufmann, pp. 27–34, 2001.

[3] D. Peel and G. McLachlan, "Robust Mixture Modeling Using the t Distribution," *Statistics and Computing*, vol. 10, no. 4, pp. 335-344, Oct. 2000.

[4] S.P. Chatzis, D.I. Kosmopoulos and T.A. Varvarigou, "Robust Sequential Data Modeling Using an Outlier Tolerant Hidden Markov Model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1657-1669, Sep. 2009.

[5] M. Svensen and C.M. Bishop, "Robust Bayesian Mixture Modelling," *Neurocomputing*, vol. 64, pp. 235-252, Mar. 2005.

[6] M.H. Law, M.A.T. Figueiredo, and A.K. Jain, "Simultaneous Feature Selection and Clustering Using Mixture Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1154-1166, Sept. 2004.

[7] C. Constantinopoulos, M.K. Titsias, and A. Likas, "Bayesian Feature and Model Selection for Gaussian Mixture Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 6, pp. 1013-1018, Jun. 2006.

[8] Y.H. Li, M. Dong and J. Hua, "Simultaneous Localized Feature Selection and Model Detection for Gaussian Mixtures," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 953-960, May. 2009.

[9] Y.H. Li, M. Dong and Y.Q. Ma, "Localized feature selection for Gaussian mixtures using variational learning," *ICPR*, Dec, 2008.

[10] A. Asuncion and D. Newman, UCI Machine Learning Repository, http://www.ics.uci.edu/~mlearn/MLRepository.html, 2007.