Gaussian Mixture Models for Supervised Classification of Remote Sensing Multispectral Images

Ana Claudia Oliveira de Melo[1], Ronei Marcos de Moraes[1], and
Liliane dos Santos Machado[2]

[1] UFPB – Federal University of Paraíba - Department of Statistics,
Cidade Universitária s/n, 58051-900, João Pessoa, PB, Brazil,
{anaclaudia, ronei}@de.ufpb.br
http://www.de.ufpb.br/
[2] UFPB – Federal University of Paraíba - Department of Computer Sciences,
Cidade Universitária s/n, 58051-900, João Pessoa, PB, Brazil
liliane@di.ufpb.br
http://www.di.ufpb.br/

Abstract. This paper proposes the use of Gaussian Mixture Models as a supervised classifier for remote sensing multispectral images. The main advantage of this approach is provide more adequated adjust to several statistical distributions, including non-symmetrical statistical distributions. We present some results of this method application over a real image of an area of Tapajós River in Brazil and the results are analysed according to a reference image. We perform also a comparison with Maximum Likelihood classifier. The Gaussian Mixture classifier obtained best adjust about image data and best classification performance too.

1 Introduction

Several times, researchers in image processing justify the use of Gaussian approach due to the data volume. In the practice of image classification, this approach should not be applied without verify the Gaussian distribution hypothesis. It implicates in a low percentage of correct classification, as example, when the Maximum Likelihood classifier is utilized. To minimize this effect, several approachs was experimented, using: fuzzy membership functions [1], multiple classifiers [2], Neural Networks [3], expert systems [4], etc.

The Gaussian Mixture Models (GMM), in theory, can be applied to modeling a large number of statistical distributions, including non-symmetrical distributions. These models have been used to data classification [5] and speech and speaker recognition [6]. In image processing applications, researchers commonly use the unsupervised version of mixture models [7], [8].

In this paper, we proposed the utilization of ellipsoidal GMM [9] as a supervised classifier for remote multispectral sensing images. We applied this method in a real image of an area of Tapajós River in Brazil and compare the results with a Maximum Likelihood classifier.

2   Theoretical Aspects

In this section, we present the Gaussian Mixture Models method for image classification. The clusters associated to the mixture components are ellipsoidal, centered at the means μ_k and variances matrix Σ_k determine their geometric characteristics of the ellipses. Parameter estimation equations for classes models are presented first. After, the GMM method for classification is then described as a maximum likelihood classifier [6].

2.1   Gaussian Mixture Models (GMM)

Let $X = \{x_1, x_2, ..., x_T\}$ be a set of T vectors, extracted from attribute space, obtained from the samples areas of the image. These information can be extracted at T different spectral bands of the image. Since the distribution of these vectors is unknown, it is approximately modeled by a mixture of Gaussian densities as the weighted sum of c component densities, given by the equation

$$p(x_t|\lambda) = \sum_{i=1}^{c} w_i N(x_t, \mu_i, \Sigma_i), \qquad t = 1, ..., T \tag{1}$$

where λ denotes a prototype consisting of a set of model parameters $\lambda = \{w_i, \mu_i, \Sigma_i\}$, w_i, $i = 1, ..., c$ are the mixture weights and $N(x_t, \mu_i, \Sigma_i)$, are the T–variate Gaussian component densities with mean vectors μ_i and covariance matrices Σ_i:

$$N(x_t, \mu_i, \Sigma_i) = \frac{\exp\left\{-\frac{1}{2}\left(x_t - \mu_i\right)' \Sigma_i^{-1} \left(x_t - \mu_i\right)\right\}}{(2\pi)^{d/2} |\Sigma_i|^{1/2}}. \tag{2}$$

To train the GMM, these parameters are estimated such that they best match the distribution of the samples from the image. The maximum likelihood estimation is widely used as a training method. For a sequence of sample vectors X for a λ, the likelihood of the GMM is done by:

$$p(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda). \tag{3}$$

The aim of maximum likelihood estimation is to find a new parameter model $\overline{\lambda}$ such that $p(X|\overline{\lambda}) \geq p(X|\lambda)$. Since the expression in (3) is a nonlinear function of parameters in λ, its direct maximisation is not possible. However, these parameters can be obtained iteratively using the Expectation-Maximisation algorithm[10]. In this algorithm, we use an auxiliary function Q done by:

$$Q(\lambda, \overline{\lambda}) = \sum_{t=1}^{T} \sum_{i=1}^{c} p(i|x_t, \lambda) \log\left[\overline{w_i} N(x_t, \overline{\mu_i}, \overline{\Sigma_i}\right] \tag{4}$$

where $p(i|x_t, \lambda)$ is the a *posteriori* probability for each mixture component of image class i, $i = 1, ..., c$ and satisfies

$$p(i|x_t, \lambda) = \frac{w_i N(x_t, \mu_i, \Sigma_i)}{\sum_{k=1}^{c} w_k N(x_t, \mu_k, \Sigma_k)}. \tag{5}$$

The Expectation-Maximisation algorithm is such that if $Q(\lambda, \overline{\lambda}) \geq Q(\lambda, \lambda)$ then $p(X|\overline{\lambda}) \geq p(X|\lambda)$[13]. Setting derivatives of the Q function with respect to $\overline{\lambda}$ to zero, we found the following reestimation formulas:

$$\overline{w_i} = \frac{1}{T} \sum_{t=1}^{T} p(i|x_t, \lambda), \tag{6}$$

$$\overline{\mu_i} = \frac{\sum_{t=1}^{T} p(i|x_t, \lambda) x_t}{\sum_{t=1}^{T} p(i|x_t, \lambda)}, \tag{7}$$

$$\overline{\Sigma_i} = \frac{\sum_{t=1}^{T} p(i|x_t, \lambda) (x_t - \mu_i) (x_t - \mu_i)'}{\sum_{t=1}^{T} p(i|x_t, \lambda)}. \tag{8}$$

The algorithm for training the GMM is described as follows:

1. Generate the a *posteriori* probability $p(i|x_t, \lambda)$ at random satisfying (5);

2. Compute the mixture weight, the mean vector, and the covariance matrix following (6), (7) and (8);

3. Update the a posteriori probability $p(i|x_t, \lambda)$ according to (5) and compute the Q function using (4);

4. Stop if the increase in the value of the Q function at the current iteration relative to the value of the Q function at the previous iteration is below a chosen threshold, otherwise go to step 2.

The GMM classification. To provide GMM classification, we need several classes of image λ. So, let λ_k, $k = 1, ..., N$, denote models of N possible classes of image. Given a feature vector sequence X, a classifier is designed to classify X into N classes of image by using N discriminant functions $g_k(X)$, computing the similarities between the unknown X and each class of image λ_k and selecting the class of image λ_{k^*} if

$$k^* = \arg \max_{1 \leq k \leq N} g_k(X) \tag{9}$$

In the minimum–error–rate classifier, the discriminant function is the a *posteriori* probability:

$$g_k(X) = p(\lambda_k|X). \tag{10}$$

We can use the Bayes' rule

$$p(\lambda_k|X) = \frac{p(\lambda_k)p(X|\lambda_k)}{p(X)} \tag{11}$$

and we can assume equal likelihood of all performances, i.e., $p(\lambda_k) = 1/N$. Since $p(X)$ is the same for all performance models, the discriminant function in (10) is equivalent to the following[14]:

$$g_k(X) = p(X|\lambda_k) \qquad (12)$$

Finally, using the log-likelihood, the decision rule used for class of performance identification is:

Select performance model k^* *if*

$$k^* = \arg \max_{1 \leq k \leq N} \sum_{t=1}^{T} \log p(x_t|\lambda_k) \qquad (13)$$

where $p(x_t|\lambda_k)$ is given by (1) for each k, $k = 1, ..., N$.

2.2   The Bayesian Information Criterion

Several information criteria are presented in literature as Bayesian Information Criterion (BIC) [9] and Akaike Information Criterion (AIC) [11]. However, Biernacki and Govaert [12] concluded that BIC has a better performance that others to choose a good mixture model after the comparison of some information criteria.

The Bayesian Information Criterion (BIC) was proposed by Schwartz [9] to evaluate the quality of estimations. This criterion compares models with different parameters or different number of components. Formaly, let a data set X and a model m, $p(X|\lambda_m)$, $m = 1, 2, ...$ with c components, $c = 1, 2, ...$, the BIC for this model is done by:

$$BIC(m, c) = -2log[p(X|\lambda_m)] + \nu_{m,c} ln(T) \qquad (14)$$

where $p(X|\lambda_m)$ is done by (3), $\nu_{m,c}$ is the number of independent parameters to be estimed in the model m with c components and T is the number of components in vector X.

The best model to represent the data set X is the model m^* that

$$m^* = \arg \max_m BIC(m, c) \qquad (15)$$

3   A Case Study

To test the classifier proposed in Sect. 2.2, we performed a case study with a real image of Brazilian state of Pará at Tapajós River. The image was obtained from the number 4 spectral sensor of Landsat satellite. The image was composed by three classes: Tapajós River (class 1), Contact Areas (class 2) and Human Occupation Areas (class 3).

To evaluate the quality of estimations, we use Bayesian Information Criterion (BIC) [9], which compares models with different parameters or different number

of components. This criteria points to the best model. So, according to the BIC, the classes 1 and 3 were modeled using two components ($K = 2$), because those original data distributions were not Gaussian distributions, and the class 2 was modeled using one component. The Fig. 1 shows all the models.
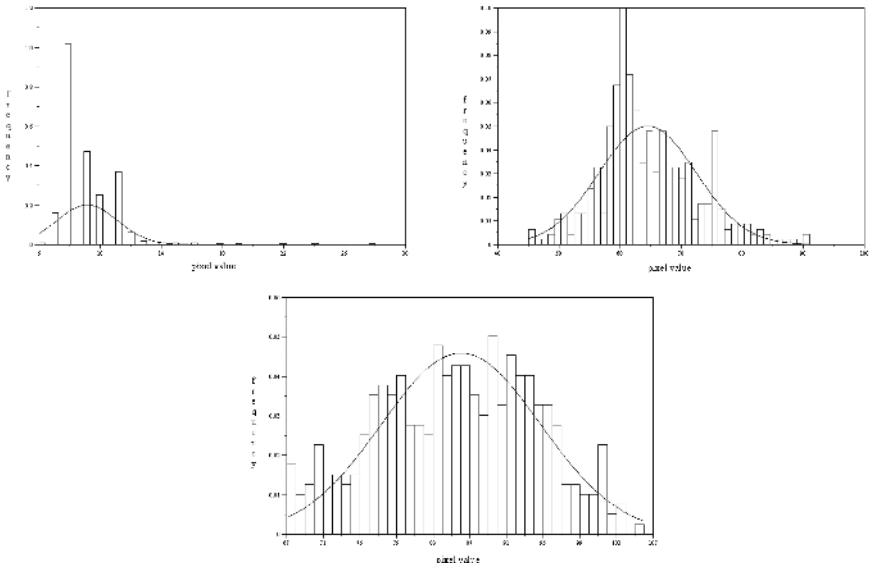

Fig. 1. Gaussian Mixture Models for classes 1, 2 and 3. Class 1 (a) was modeled using $K = 2$, class 2 (b) was modeled using $K = 1$ and Class 3 (c) was modeled using $K = 2$.

A statistical comparison between classification and a reference map of the same area was performed using Kappa coefficient. The Kappa result was 90.6793% with variance 9.5815×10^{-5}. The same coefficient for each class results were: 99.6929% for class 1, 84.5049% for class 2 and 86.7262% for class 3. The percentuals of correct classification for each class were respectively: $99, 8134\%$, 89.3032% and 90.6801%. The main misclassification occurred with classes 2 and 3. In fact, there is an overlapping of those classes distributions. The Table 1 shows the matrix classification used for calculations we presented above.

We can to note, in Table 1, that only one pixel of class 1 was classified as class 2, 46 pixles of class 2 were classified as class 3 and 37 pixles of class 3 were classified as class 2. The cause of these facts is the overlapping of classes 2 and 3 distributions.

Thus, the classification can be considered satisfactory in statistical terms, even if we consider the overlapping of class 2 and class 3 distributions. It is important to note that these results were obtained using only one original spectral band.

Table 1. Matrix classification for GMM classifier using a reference map.

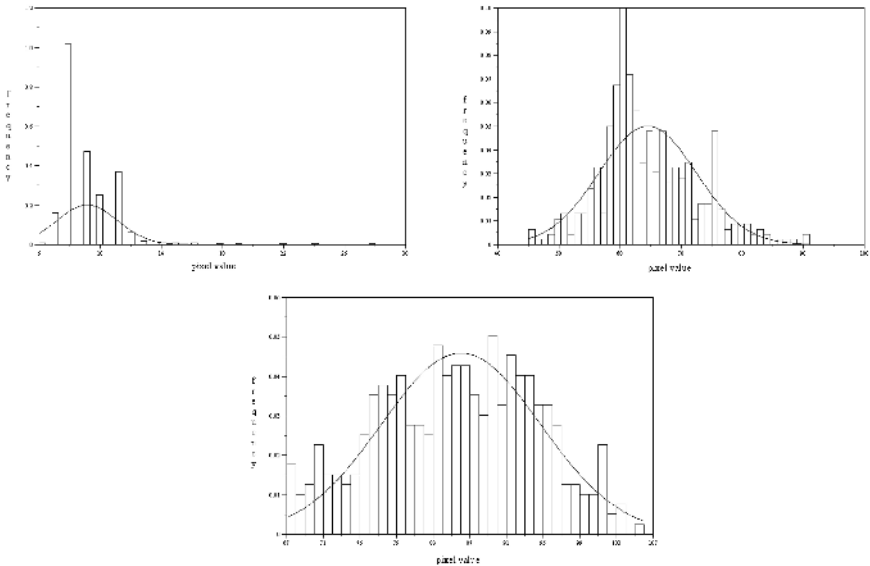| Classes	class 1	class 2	class 3
class 1	535	1	0
class 2	0	384	46
class 3	0	37	360

Fig. 2. Maximum Likelihood Models for classes 1 (a), 2 (b) and 3 (c). All classes were modeled as a Gaussian distribution.

4   Comparison with a Maximum Likelihood Classifier

To compare performances between GMM classifier with another classifier, we performed a classification using a Maximum Likelihood classifier. Using the same reference map, the statistical Kappa coefficient for Maximum Likelihood classification was 89.2355 with variance 1.09389×10^{-4}. The same coefficient for each class results were: 99.0809% for class 1, 84.2878% for class 2 and 82.9756% for class 3. The percentuals of correct classification for each class were respectively: $99,4403\%$, 89.3032% and 87.9093%. Again, the overlapping of 2 and 3 classes distributions was responsable for misclassification. The Table 2 shows the matrix classification used for calculations we presented above.

We can to note, in Table 2, that 3 pixels of class 1 was classified as class 2, 46 pixles of class 2 were classified as class 3 and 48 pixels of class 3 were classified as class 2. Besides the overlapping of 2 and 3 classes distributions, we can observe that Maximum Likelihood has obtained satisfactory results too. However, it is

Table 2. Matrix classification for Maximum Likelihood classifier using a reference map.

Classes	class 1	class 2	class 3
class 1	533	3	0
class 2	0	384	46
class 3	0	48	349

important to note that all results are not better than GMM classifier. The Fig. 2 shows all the models.

5   Conclusions and Future Works

In this paper, we proposed the use of Gaussian Mixture Models for supervised image classification. We obtained results of a real data classification using an area of Tapajós River in Brazil. We did a comparison of this classifier with the Maximum Likelihood classifier.

The Gaussian Mixture Models classifier is adequated to classify images which classes can not be adjusted by Gaussian distributions. These models provide more adequated adjust to several distributions with lower variance. This property in particular is welcome in image classification for reduce misclassification, but these properties are already used in others applications, as data analysis and speech and speaker recognition.

We presented an application using only one original spectral image. In this application we observed better statistical results of GMM classifier. It is important to note that classes 1 and 3 do not have Gaussian distribution as showed in the Fig. 1 and Fig. 2. This fact is relevant to explain the inferior performance of Maximum Likelihood classifier, mainly to classify the class 3.

As future works we intend to do a performance statistical comparison with others statistical classifiers.

References

1. Kent, J. T.; Mardia, K. V. Spatial classification using fuzzy membership models. IEEE Trans. on Patt. Anal. and Mach. Intel. **10** (1988) 659–671
2. Ho, T. K.; Hull, J. J.; Srihari, S. N.; Decision combination in multiple classifier systems. IEEE Trans. on Patt. Anal. and Mach. Intel. **16** (1994) 66–75
3. Simpson, P. K. Artificial Neural Systems. Pergamon Press, New York (1990)
4. Moraes, R. M.; Sandri, S. A.; Banon, G. J. F. Expert Systems Architecture for Image Classification Using Mathematical Morphology Operators. Information Science **142** (2002) 7–21
5. Yang, M. H., Ahuja, N. Gaussian Mixture Model for Human Skin Color and Its Application in Image and Video Databases, Proc. of the SPIE, vol. 3656: Conf. on Storage and Retrieval for Image and Video Databases, San Jose (1999) 458–466

6. Tran, D., Pham, T., Wagner, M., Speaker recognition using Gaussian mixture models and relaxation labeling. in Proceedings of the 3rd World Multiconference on Systemetics, Cybernetics and Informatics/ The 5th Int. Conf. Information Systems Analysis and Synthesis (SCI/ISAS99), USA **6** (1999) 383–389

7. H. Caillol, W. Pieczynski and A. Hillon, Estimation of fuzzy Gaussian mixture and unsupervised statistical image segmentation, IEEE Trans. on Image Proces. **6** (1977) 425–440

8. W. Pieczynski, J. Bouvrais, and C. Michel, Estimation of generalized mixture in the case of correlated sensors, IEEE Trans. on Image Proces., **9**,(2000) 308–311

9. G. Schwartz, Estimating the Dimension of a Model. The Ann. of Stat. **6** (1978) 461–464

10. A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm. J.R.Statis.Soc. B, **39** (1977) 1–38

11. Akaike, H. A new look at the statistical identification model. IEEE Trans. on Automatic Control, **19** (1974) 716–723.

12. Biernacki, C. and Govaert, G. Choosing models in model-based clustering and discriminant analysis. Technical Report INRIA no. 3509, (1998).

13. Rabiner, L. R., A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition. Proceedings of the IEEE, **77** (1989)

14. Johnson, R. A. and Wichern, D. W., Applied Multivariate Statistical Analysis. Prentice Hall (1998)