A New Approach to Unsupervised Classification of Hyperspectral Earth Observation Imagery Using a Gaussian Mixture Model

VICTOR-EMIL NEAGOE, VLAD BERBENTEA CHIRILA
Faculty of Electronics, Telecommunications, and Information Technology
"Politehnica" University of Bucharest
Splaiul Independenței 313, Bucharest
ROMANIA
email: victor.neagoe@upb.ro

Abstract: - This paper presents an approach for improving performances of the unsupervised classification by proposing a technique to combine the classical techniques of K-means clustering and the Gaussian mixture model (GMM) based on expectation-maximization (EM). The proposed model means to apply firstly K-means clustering, and to use the result of this technique to compute means μ_k, covariance matrices Σ_k and mixing coefficients π_k, considered as initialization parameters for the next stage of GMM-EM. The above mentioned algorithm has been successfully applied for clustering of Earth Observation (EO) hyperspectral imagery. The information content of hyperspectral images with hundreds of channels allows us to remotely identify ground materials, based on their spectral signature. The performances of the proposed method have been evaluated using the Pavia Centre hyperspectral database with 102 spectral bands and a resolution of 1.3 meters/pixel. The proposed combined clustering model K-means+GMM-EM, has led to a significant improvement in performance over any of the two single clustering techniques K-means and GMM-EMM.

Key-Words: - unsupervised classification, hyperspectral imagery, Earth Observation (EO), Gaussian mixture model (GMM), expectation-maximization (EM)

1 Introduction

Analysis of Earth Observation (EO) imagery has wide applications for generation of various kinds of civil or military maps: maps of vegetation, maps of mineral resources of the Earth, land-use maps (urban areas, agricultural fields, woods, rivers, lakes, buildings, airports, and highways), and so on [1], [2]. Clustering (also called unsupervised classification) has been used for a wide variety of tasks in remote sensing image analysis such as segmentation, feature extraction, dimensionality reduction, data visualization, and final classification [1], [2], [3], [4]. Advanced optical sensors on board of the satellites or spacecraft (e.g. HYDICE, Hyperion, AVIRIS) can be used to deal with the limit of multispectral sensors; they are able to capture scenes at hundreds of contiguous bands with bandwidth as narrow as several nanometers providing a sampling of the visible and near infrared spectrum (referred to as hyperspectral imagery or HSI) [2], [5], [6], [7], [8], [9]. As opposed to multispectral sensors that produce only a few radiance data points for a ground pixel, hyperspectral imaging sensors construct the near continuous quality radiance spectrum for each pixel in the scene. Each hyperspectral image pixel is a high dimensional vector corresponding to the sampling of the incoming radiance at this pixel. The information content of hyperspectral images with hundreds of channels allows us to remotely identify ground materials, based on their spectral signature [2], [5], [6], [7], [8], [9]. Hyperspectral remote sensing is used in a wide array of applications. It has been originally applied for mining and geology (mineral and oil identification). Now hyperspectral imaging has spread into new fields as ecology (global change research), surveillance (military aims, archaeological sites discovery), and agriculture (early detection of disease, soil characterization and inventory, harvest estimation). Unsupervised classification techniques do not require the user to specify any information about the features contained in the images. One of the most widely used clustering algorithm includes the iterative partitioning method called K-means and its family [10]. The K-means technique attempts to iteratively minimize an error criterion and terminate the iterations when a local minimum is reached. The K-Means clustering is one of the most common methods of unsupervised classification for data analysis, as in the fields of pattern recognition, data mining, image processing and so on [4], [5]. It is very useful in the area of remote sensing image analysis, where objects with similar spectrum values are clustered together without any former knowledge [4]. However, the k-means algorithm based on minimization of the sum of squared errors criterion is very limited in terms of its cluster modeling capability because it can only model spherical clusters with similar number of data points [3], [10]. The clustering based on the Gaussian mixture model (GMM) that are learned by maximizing the likelihood function using the expectation-maximization (EM) algorithm [10], [11], [12], [13], [14] has proved to be superior to k-means in the sense that it is capable of finding clusters of arbitrary ellipsoidal shapes with arbitrary number of data points. Whereas the *K*-means algorithm performs a *hard* assignment of data points to clusters, in which each data point is associated uniquely with one cluster, the GMM-EM algorithm makes a *soft* assignment based on the posterior probabilities [10]. Furthermore, both the k-means and the GMM-EM algorithm are very sensitive to initializations [10].

This paper proposes a new unsupervised classification model for satellite hyperspectral imagery consisting of the clustering cascade K-means+GMM-EM. This means to firstly apply the K-means clustering and to use the final result of this technique to compute the initialization parameters for the next stage of GMM-EM. The performances of the proposed algorithm for hyperspectral EO image unsupervised classification using a Pavia Centre dataset are evaluated.

2 K-Means+GMM-EM Clustering Algorithm

Denote by {(x_n), n=1,...,N} the set of N multispectral n-band pixels to be clustered in K classes.

The stages of the proposed EM clustering algorithm are the following:

A. K-means clustering for a reduced number of iterations using a random parameter initialization.

B. GMM-EM clustering using the parameter initialization given by the results of the previous K-means clustering phase. Assuming expectation maximization for a Gaussian mixture model (GMM-EM), the goal is to maximize the likelihood function with respect to the parameters (comprising the means and covariances of the components and the mixing coefficients). The steps of the EM clustering are further presented.

α) Compute the means μ_k, covariance matrices Σ_k and mixing coefficients π_k (where k=1,....K) as a result of previous phase of K-means clustering, by considering them as initialization parameters for the present GMM- EM phase and evaluate the initial value of the log likelihood.

β) **E step**. Evaluate the responsibilities using the current parameter values 10]

$$\gamma(z_{nk}) = \frac{\pi_k N(x_n|\mu_k,\Sigma_k)}{\sum_{l=1}^{K} \pi_l N(x_n|\mu_l,\Sigma_l)} \quad (1)$$

ϒ) **M step**. Re-estimate the parameters using the current responsibilities [10]

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n \quad (2)$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T \quad (3)$$

$$\pi_k = \frac{N_k}{N} \quad (4)$$

where

$$\sum_{n=1}^{N} \gamma(z_{nk}) = N_k \quad (5)$$

δ) Evaluate the log likelihood

$$\ln p(X|\pi,\mu,\Sigma) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k N(x|\mu_k,\Sigma_k) \quad (6)$$

ε) Check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step b.

C. Calibration

After clustering, we assign to each of the K resulted clusters the label of the corresponding majority class.

D. Evaluation of the correct classification score

3 Experimental Results

3.1. Pavia Centre database

The Pavia Centre scene is acquired by the ROSIS sensor during a flight campaign over Pavia, northern Italy.

The number of spectral bands is 102 and the scene corresponds to a 1096*1096 pixels image, but some of the samples contain no information and have to be discarded before the analysis. Since then, the image to be analyzed corresponds to 1096x715 pixels. The geometric resolution is 1.3 meters and the groundtruth differentiates nine classes. There are a lot of discarded samples in the figures as abroad black strips (Fig. 1). The class structure is given in Table 1 and the groundtruth scene is shown in Fig. 2.

Table 1. Groundtruth Class Structure of the Pavia Centre Scene Dataset (102 Spectral BANDS, resolution of 1.3 meters/pixel).

No	Class	# pixels
1	Water	824
2	Trees	820
3	Asphalt	816
4	Self-Blocking Bricks	808
5	Bitumen	808
6	Tiles	1260
7	Shadows	476
8	Meadows	824
9	Bare Soil	820

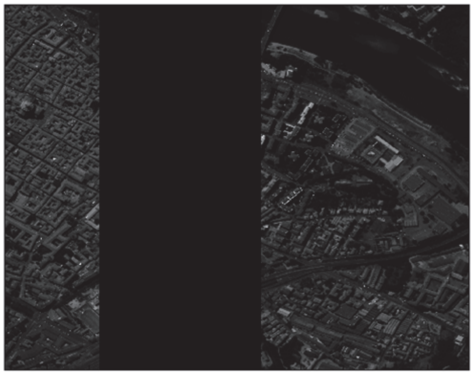

Fig. 1. Sample band of Pavia Centre dataset.

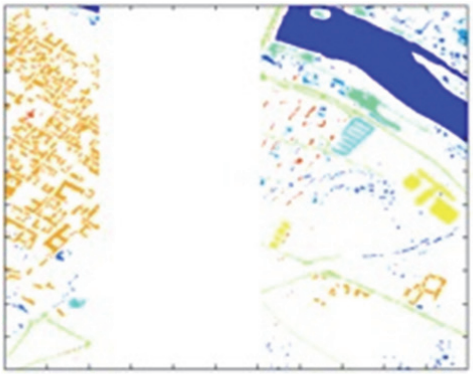

Fig. 2. Groundtruth of Pavia Centre dataset.

3.2 Experimental performances

From all the labeled pixels shown in Table 1, we have selected a balanced set to be used for evaluation of the proposed clustering model. Using the above described Pavia Centre hyperspectral balanced data set of 102-band with 9-class pixels, we have comparatively evaluated the performances of the following clustering algorithms:

- K-means
- EM
- K-means+EM

The global experimental results are shown in Table 2. The confusion matrix for the new variant K-Means+GMM-EM is shown in Table 3.

Table 2. Global clustering performances of the experimented algorithms (Hyperspectral Pavia Centre dataset; 102 spectral bands; 9 classes).

Clustering algorithm	Correct classification score (%)	Number of iterations
K-Means	67.64	21
EM	74.36	36
K-Means+GMM-EM	81.73	50+24

Table 3. Confusion matrix of the proposed clustering cascade {K-means+GMM-EM} (Hyperspectral Pavia Centre dataset; 102 spectral bands; 9 classes; (50+24) iterations; classification score: 81.73%).

Assigned class Real class	1	2	3	4	5	6	7	8	9
1	**99.85%**	0.15%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
2	0.00%	**96.54%**	3.46%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
3	0.00%	11.32%	**88.60%**	0.00%	0.04%	0.00%	0.00%	0.00%	0.04%
4	0.00%	0.00%	0.00%	**62.79%**	20.26%	0.19%	6.11%	10.61%	0.04%
5	0.00%	0.04%	0.00%	59.07%	**28.01%**	0.00%	0.00%	0.04%	12.85%
6	0.00%	0.00%	0.00%	0.74%	0.15%	**96.24%**	1.42%	1.42%	0.04%
7	0.00%	0.00%	0.00%	0.30%	0.04%	8.12%	**91.32%**	0.19%	0.04%
8	0.00%	0.00%	0.00%	0.04%	1.68%	0.11%	0.04%	**72.77%**	25.36%
9	0.00%	0.60%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**99.40%**

4 Concluding Remarks

By comparing K-means clustering versus clustering with GMM-EM, when both techniques use random parameter initialization, one can deduce a better score for GMM-EM over K-means with about 7% advantage for GMM-EM (74.36% versus 67.64%) as shown in Table 2.

The proposed clustering algorithm means to firstly apply K-means clustering and to use the result of this stage to compute means μ_k, covariance matrices Σ_k and mixing coefficients π_k, considered as initialization parameters for the next stage of GMM-EM. The model of K-means+GMM-EM has led to a significant improvement in performance over any of the two single clustering techniques K-means and GMM-EMM, when both use random parameter initialization. The proposed combined clustering K-means+GMM-EM leads to 81.73% recognition score, meaning an increasing with about 14% over K-means and respectively a score with more than 7% better than the single GMM-EM.

From Table 3, we can evaluate the confusion matrix for the new K-means+EM clustering cascade. The best recognized classes (with a recognition score better than 95%) are water, bare soils, trees and tiles. The minimum recognition score (28%) is registered for the class of bitumen.

References:

[1] J.A. Richards, X. Jia, *Remote Sensing Digital Image Analysis, An Introduction*, Springer, New York, 2005.

[2] C.H. Chen, *Signal and Image Processing for Remote Sensing*, CRC Press, Boca Raton-Fl-USA, 2012.

[3] C. Ari, S. Aksoy, Unsupervised Classification of Remotely Sensed Images Using Gaussian Mixture Models and Particle Swarm Optimization, *Proc. IEEE Intern. Symposium Geoscience and Remote Sensing (IGARSS 2010)*, Honolulu, Hawaii, 25-30 July 2010, pp. 1859-1862.

[4] Z. Lv, Y. Hu, H. Zhong, J. Wu, B. Li, and H. Zhao, Parallel K-Means Clustering of Remote Sensing Images Based on MapReduce, *Proc. of the 2010 Int. Conf. Web Information Systems and Mining (WISM'10)*, Sanya, China, October 23-24, 2010, Springer, Berlin, 2010, pp. 162-170.

[5] F. Melgani, L. Bruzzone, Classification of hyperspectral remote sensing images with support vector machines, *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, Aug. 2004, pp. 1778–1790.

[6] A. Plaza, J.A. Benediktsson, J.W. Boardman, J. Brazile, L. Bruzzone, G. Camps-Valls, J. Chanussot, M. Fauvel, P. Gamba, A. Gualtieri, M. Marconcini, J.C. Tilton, and G. Trianni, Recent advances in techniques for hyperspectral image processing, *Remote Sensing of Environment*, vol. 113, Sept. 2009, pp. 110-122.

[7] A. Backhaus, F. Bollenbeck, and U. Seiffert, Robust classification of the nutrition state in crop plants by hyperspectral imaging and artificial neural networks, *Proc. 3rd Workshop Hyperspectral Image, Signal Process. (WHISPERS 2011)*, Lisbon, Portugal, June 2011, pp. 1-4.

[8] V.E. Neagoe, A.D. Ciotec, A New Approach for Accurate Classification of Hyperspectral Images Using Virtual Sample Generation by Concurrent Self-Organizing Maps, *2013 Proc. IEEE International Geoscience & Remote Sensing Symposium, (IGARSS'13)*, Melbourne, Australia, 21-26 July 21-26, 2013, pp. 1031-1034.

[9] M. Khodadadzadeh, J. Li, A. Plazza, P. Gamba, J.A. Benediktsson, and J.M. Bioucas-Dias, A new framework for hyperspectral image classification using multiple spectral and spatial features, *2014 Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS'14)*, Quebec-City, Canada, July 2014, pp. 4628-4631.

[10] [10] M. Bishop, *Pattern Recognition and Machine Learning*, New York, NY, USA: Springer, 2006.

[11] A.P. Dempster, N.M. Laird, and D.B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, Series B (Methodological), Vol. 39, No. 1, 1977, pp. 1-38.

[12] R.A. Redner, H.F. Walker, Mixture Densities, Maximum Likelihood and the EM Algorithm, *SIAM Review*, Vol. 26, Issue 2 (April), 1984, pp. 195-239.

[13] G. McLachlan, T. Krishnan, *The EM Algorithm and Extensions*, 2nd Edition, Wiley, New York, 2008.

[14] V.E. Neagoe, V. Berbentea-Chirila, Improved Gaussian Mixture Model with Expectation-Maximization for Clustering of Remote Sensing Imagery, *Proc. 2016 IEEE Int. Geosci. Remote Sens. Symp. (IGARSS'16)*, Beijing, China, July 10-15, 2016, pp. 3063-3065.