Bayesian Hyperspectral Image Segmentation with Discriminative Class Learning

Janete S. Borges[1], José M. Bioucas-Dias[2], and André R. S. Marçal[1]

[1] Faculdade de Ciências, Universidade do Porto
[2] Instituto de Telecomunicações, Instituto Superior Técnico, TULisbon
jsborges@fc.up.pt, bioucas@lx.it.pt, andre.marcal@fc.up.pt

Abstract. This paper presents a new Bayesian approach to hyperspectral image segmentation that boosts the performance of the discriminative classifiers. This is achieved by combining class densities based on discriminative classifiers with a Multi-Level Logistic Markov-Gibs prior. This density favors neighbouring labels of the same class. The adopted discriminative classifier is the Fast Sparse Multinomial Regression. The discrete optimization problem one is led to is solved efficiently via graph cut tools. The effectiveness of the proposed method is evaluated, with simulated and real AVIRIS images, in two directions: 1) to improve the classification performance and 2) to decrease the size of the training sets.

1 Introduction

In recent years much research has been done in the field of image segmentation. Several methods have been used in a wide range of applications in computer vision. However, its application to high dimensional data, such as hyperspectral images, is still a delicate task, namely owing to well known difficulties in learning high dimensional densities from a limited number of training samples.

The discriminative approach to classification circumvent these difficulties by modelling directly the densities of the labels, given the features. This framework have shown success in dealing with small class distances, high dimensionality, and limited training sets. As a consequence, discriminative classifiers hold the state-of-the art in supervised hyperspectral image classification (see, *e.g.*, [1]).

Real world images tend to exhibit piecewise spatial continuity of categorical properties (*i.e.*, classes). Thus, an intuitive way of improving the performance of discriminative classifiers (and others) consists in adding contextual information in the form of spatial dependencies. This direction has been pursued in [2], introducing the concept of discriminative random fields in the computer vision applications, and in [3] and [4] using composite kernels, in hyperspectral applications.

This paper introduces a new Bayesian segmentation approach for hyperspectral images. Spatial dependencies are enforced by a Multi-Level Logistic (MLL) Markov-Gibs prior, which favours neighbouring labels of the same class. The class densities are build on the discriminative Fast Sparse Multinomial Regression (FSMLR) [5], which a fast version of the Sparse Multinomial Regression

J. Martí et al. (Eds.): IbPRIA 2007, Part I, LNCS 4477, pp. 22–29, 2007.

(SMLR) [6]. The SMLR includes a Laplacian prior to control the complexity of the learned classifier and, therefore, to achieve good generalization capabilities.

To compute an approximation to the Maximum A Posteriori probability (MAP) segmentation, we adopt the α-Expansion graph cut based algorithm proposed in [7]. This tool is computationally efficient and yields nearly optimum solutions.

The paper is organized as follows. Section 2 formulates the problem, describe briefly the FSMLR classifier, the MLL Markov Gibs prior, and the α-Expansion optimization algorithm. Section 3 presents results based on simulated and real hyperspectral datasets.

2 Formulation

A segmentation is an image of labels $\mathbf{y} = \{y_i\}_{i\in\mathcal{S}}$, where $y_i \in \mathcal{L} = \{1, 2, \ldots, K\}$. Let $\mathbf{x} = \left\{x_i \in \mathbb{R}^d, i \in \mathcal{S}\right\}$ be the observed multi-dimensional images, also known as feature image. The goal of the segmentation is to estimate $\mathbf{y}$, having observed $\mathbf{x}$. In a Bayesian framework, this estimation is done by maximizing the posterior distribution $p(\mathbf{y}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{y})p(\mathbf{y})$, where $p(\mathbf{x}|\mathbf{y})$ is the likelihood function (or the probability of feature image) and $p(\mathbf{y})$ is the prior over the classes.

In the present approach, we use the discriminative FSMLR classifier [5] to learn the class densities $p(y_i|x_i)$. The likelihood is then given by $p(x_i|y_i) = p(y_i|x_i)p(x_i)/p(y_i)$. Noting that $p(x_i)$ does not depend on the labeling $\mathbf{y}$ and assuming $p(y_i) = 1/K$, we have

$$p(\mathbf{x}|\mathbf{y}) \propto \prod_{i\in\mathcal{S}} p(y_i|x_i), \tag{1}$$

where conditional independence is understood.

In the following sections, we briefly describe the FSMLR method yielding the density $p(\mathbf{y}|\mathbf{x})$, the MLL prior $p(\mathbf{y})$, and α-Expansion optimization algorithm.

2.1 Class Density Estimation Using Fast-SMLR Method

Given the training set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, the SMLR algorithm learns a multi-class classifier based on the multinomial logistic regression. By incorporating a Laplacian prior, this method performs simultaneously feature selection, to identify a small subset of the most relevant features, and learns the classifier itself [6]. The goal is to assign to each site $i \in \mathcal{S}$ the probability of $y_i = k$, for $k = 1, \ldots, K$. In particular, if $y_i = [y^{(1)}, \ldots, y^{(K)}]^T$ is a 1-of-K encoding of the K classes, and if $w^{(k)}$ is the feature weight vector associated with class k, then the probability of $y_i^{(k)} = 1$ given x_i is

$$P\left(y_i^{(k)} = 1|x_i, w\right) = \frac{\exp\left(w^{(k)^T} h(x_i)\right)}{\sum_{k=1}^{K} \exp\left(w^{(k)^T} h(x_i)\right)}, \tag{2}$$

where $w = [w^{(1)^T}, \ldots, w^{(K)^T}]^T$ and $h(x) = [h_1(x), \ldots, h_l(x)]^T$ is a vector of l fixed functions of the input, often termed features. Possible choices for this

vector are $h(x_i) = [1, x_{i,1}, \ldots, x_{i,d}]^T$, where $x_{i,j}$ is the jth component of x_i, and $h(x) = [1, K(x, x_1), \ldots, K(x, x_n)]^T$, where $K(\cdot, \cdot)$ is some symmetric kernel function. The latter nonlinear mapping guarantees that the transformed samples are more likely to be linearly separable. Nevertheless, in this paper, we consider only the linear mapping, because it is much lighter from the computational point of view (note that the linear and the nonlinear mapping have $l = d+1$ and $l = n + 1$, respectively, and, usually, $n \gg d$) yet it leads to competitive results.

The MAP estimate of w is

$$\hat{w}_{MAP} = \arg\max_w L(w) = \arg\max_w \left[l(w) + \log p(w) \right], \tag{3}$$

where $l(w)$ is the log-likelihood function and $p(w) \propto \exp(-\lambda \|w\|_1)$, where λ is a regularization parameter controlling the degree of sparseness of $\hat{w}_{MAP}$. The inclusion of the Laplacian prior does not allow the use of the classical IRLS method. However, the bound optimization framework [8] supplies a tool that makes it possible to perform exact MAP multinomial logistic regression, with the same cost as the original IRLS algorithm for ML estimation (see [6] for details).

In practice, the application of SMLR to large datasets is often prohibitive. A solution for this problem consists in using the Block Gauss-Seidel method [9] to solve the system used in the IRLS method. In each iteration, instead of solving the complete set of weights, only blocks corresponding to the weights belonging to the same class are solved [5], resulting in an improvement of the order of $O(K^2)$, where K is the number of classes.

2.2 The MLL Markov-Gibs Prior

The MLL prior is a MRF which models the piecewise continuous nature of the image elements, considering that adjacent pixels are likely to belong to the same class. According to the Hammersly-Clifford theorem, the prior probability of an MRF is a Gibb's distribution [10]. Thus

$$p(\mathbf{y}) = \frac{1}{Z} \exp\left(- \sum_{c \in C} V_c(\mathbf{y}) \right), \tag{4}$$

where Z is a normalizing constant and the sum is over the prior potentials $V_c(\mathbf{y})$ for the set of cliques[1] C over the image, and

$$- V_c(\mathbf{y}) = \begin{cases} \alpha_{y_i} \quad if \ |c| = 1 \text{ (single clique)} \\ \beta_c \quad if \ |c| > 1 \text{ and all sites in } c \text{ have the same label} \\ -\beta_c \ if \ |c| > 1 \text{ at least one site has a different label,} \end{cases} \tag{5}$$

where β_c is a nonnegative constant.

[1] A clique is a set of pixels that are neighbours of one another.

Let $\alpha_k = \alpha$ and $\beta_c = \frac{1}{2}\beta > 0$. This choice gives no preference to any label nor to any direction. Under this circumstances, (4) can be written as

$$p(\mathbf{y}) = \frac{1}{Z} e^{\beta n(\mathbf{y})}, \tag{6}$$

where $n(\mathbf{y})$ denotes the number of cliques having the same label. The conditional probability is given by

$$p(y_i = k | y_{\mathcal{N}_i}) = \frac{e^{\beta n_i(k)}}{\sum_{k=1}^{K} e^{\beta n_i(k)}}, \tag{7}$$

where $n_i(k)$ is the number of sites in the neighbourhood of site i, $\mathcal{N}_i$, having the label k.

2.3 Energy Minimization Via Graph Cuts

Using the FSMLR to learn $p(\mathbf{x}|\mathbf{y})$ and the MLL prior $p(\mathbf{y})$, the MAP segmentation is given by

$$\begin{aligned} \hat{\mathbf{y}} &= \arg\max_{\mathbf{y}} p(\mathbf{x}|\mathbf{y}) p(\mathbf{y}) \\ &= \arg\max_{\mathbf{y}} \sum_{i \in \mathcal{S}} \log p(x_i|y_i) + \beta n(\mathbf{y}) \\ &= \arg\min_{\mathbf{y}} \sum_{i \in \mathcal{S}} - \log p(x_i|y_i) - \beta \sum_{i,j \in c} \delta(y_i - y_j). \end{aligned} \tag{8}$$

Minimization (8) is a combinatorial optimization problem, where the right hand side pairwise interaction term is equivalent to a metric[2] and thus α-Expansion algorithm can be applied, yielding very good approximations to the MAP segmentation [7].

3 Results

3.1 Experimental Setup

Simulated datasets were used to test the proposed method. Images of labels were generated using a MLL distribution, with a 2nd order neighbourhood. The shape of these label images depends on a parameter (β_0) that controls the spatial continuity (β_0 represents the β in (6)). This parameter takes values between 1 and 2, with increments of 0.2. Figure 1 shows four examples of these label images with 4 classes, for $\beta_0 = 1$, $\beta_0 = 1.4$, $\beta_0 = 1.6$ and $\beta_0 = 2.0$. Images with 4 and 10 classes were generated, for each value of β_0, resulting in a total of 12 different label images. The feature images were obtained by adding zero-mean Gaussian independent noise with standard deviation σ to a source matrix of

[2] A metric is obtained by adding β to terms $-\beta\delta(y_i - y_j)$.

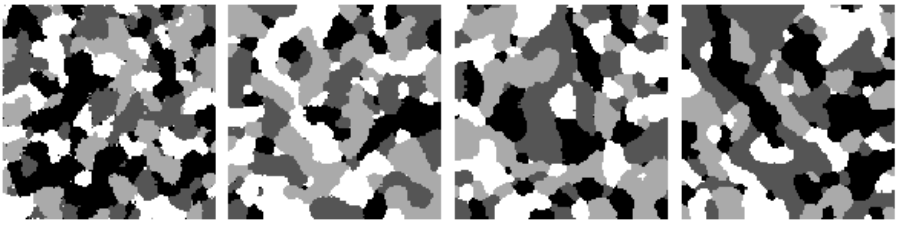


Fig. 1. Image labels with four classes generated by a MLL distribution with $\beta_0 = 1$, $\beta_0 = 1.4$, $\beta_0 = 1.6$ and $\beta_0 = 2$ (from left to right, respectively)

mineral signatures. This source matrix is provided by a Matlab data file [11], and was extracted from the USGS spectral library. Each signature is evaluated in 221 spectral bands, resulting in a dataset of dimension 120×120×221 (120×120 is the spatial size of the simulated images). Datasets with added noise standard deviation of 0.01, 0.1 and 1 were generated. The variation of the parameters β and σ was made to evaluate the response of the proposed method to the spatial continuity of label images and to the amount of noise present in the feature data. To evaluate the method performance depending on the size of the training sets, tests were made using 10%, 30%, 50%, 70% and 90% of the training set.

Datasets with the characteristics described above were simulated 10 times for each set of parameters, in order to better evaluate the segmentation results.

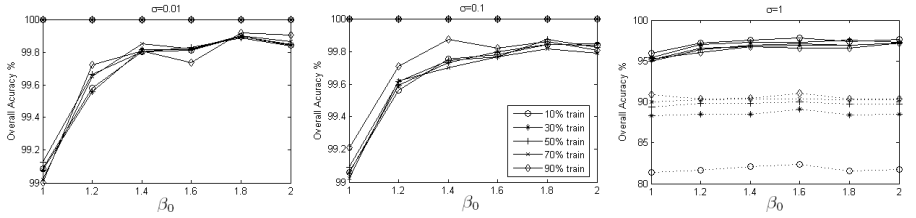
3.2 Results on Simulated Data

The ratio of the correct classified pixels over the total number of pixels, termed overall accuracy (OA), was computed for each dataset. The regularization parameter is $\beta = 1.5$.

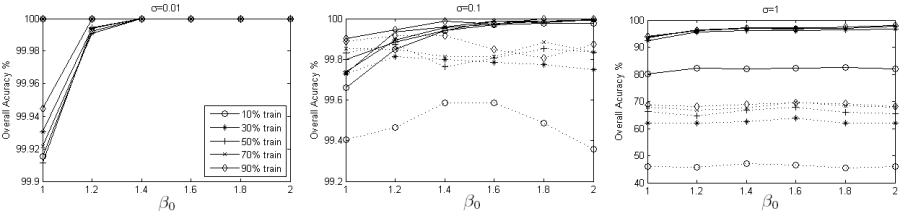
Graphics with the overall accuracies as function of β_0 are presented in Figures 2(a) and 2(b), for 4 and 10 classes, respectively. Lines corresponds to the overall accuracies for the MRF segmentation and dotted lines for the FSMLR classification. For each method, results for different training set size are also presented. The graphics are displayed for different values of feature noise: $\sigma = 0.01, 0.1$ and 1. In some plots, the overall accuracy of the FSMLR classifier (dotted lines) are nearly 100%, and the lines are thus not visible in those plots.

In the case of $K = 4$ (Fig.2(a)) and for larger values of β_0 and $\sigma = 0.01$ or 0.1, the results from FSMLR classifier and MRF segmentation are similar. When the noise increases, the MRF outperforms the FSMLR by over 5%. As expected for values of low values of β_0, the performance of the MRF segmentation is slightly worse. However, our method clearly outperforms the FSMLR classification when the noise is high ($\sigma = 1$). The different sizes used for the training set do not seems to affect the results, except for the case of $\sigma = 1$, where the use of a smaller training set degrades the performance of the FSMLR.

In the case of $K = 10$ (Fig.2(b)), the results for $\sigma = 0.01$ and $\sigma = 0.1$ are very similar for both methods. For $\sigma = 0.1$, it is nevertheless possible to see a small improvement of the accuracy achieved with MRF segmentation. Once again,



(a) $K = 4$



(b) $K = 10$

Fig. 2. Overall accuracies as function of spatial continuity (β_0) of the label images. Lines represent the MRF segmentation results and dotted lines the FSMLR classification results.

when higher noise in the feature image is considered ($\sigma = 1$), MRF segmentation clearly outperforms FSMLR classifier, by over 30%.

3.3 Results on Real Data

We applied the proposed MAP segmentation to an AVIRIS spectrometer image, the Indian Pines 92 from Northern Indiana, taken on June 12, 1992 [12]. The ground truth data image consists of 145 x 145 pixels of the AVIRIS image in 220 contiguous spectral bands. Experiments were carried out without 20 noisy bands. Due to the insufficient number of training samples, seven classes were discarded, leaving a dataset with 9 classes distributed by 9345 elements. This dataset was randomly partitioned into a set of 4757 training samples and 4588 validation samples. The spatial distribution of the class labels is presented in Figure 3. Each of the nine land cover classes is represented in one of nine grey levels. The black areas are the areas with unknown classes.

The results presented in this section are the overall accuracy measured in the independent (validation) dataset with 4588 samples. Experiments were evaluated using 10%, 20% and the complete training set. As in the previous section, we use a linear mapping h in the FSMLR. Parameter β was learned in a supervised fashion leading to $\beta = 1.5$ and $\beta = 4$, when the complete and subsets of the training set were used, respectively.

The results of overall accuracy from FSMLR classification and segmentation with MRF are presented in Table 3.3. From these results we observe that, regardless the size of the training set used to learn the density function, the MRF segmentation with a linear mapping h outperforms all other methods compared. The

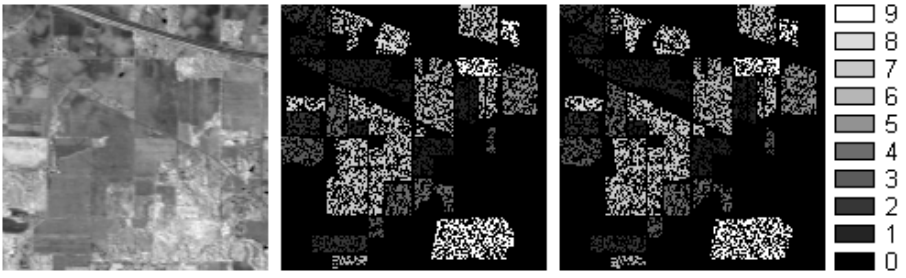


Fig. 3. AVIRIS image used for testing. Left: original image band 50 (near infrared); Centre: training areas; Right: validation areas

Table 1. Overall accuracy of the proposed MRF segmentation with linear mapping h, the SVM [1], the LDA, and the FSMLR, using 10%, 20% and 100% of the complete training set. The number of bands selected by the FSMLR is shown in brackets.

	10%	20%	100%
MRF (no. of bands)	**88.40**% (24)	**89.56**% (39)	**95.51**% (37)
SVM [1]	82.70%	86.70%	94.44%
FSMLR	75.57%	79.69%	85.77%
LDA	69.40%	78.40%	82.08%

gains with respect to the FSMLR and the linear discriminant analysis (LDA) are larger than 10%. Based on this results, we foresee that the proposed approach using kernel functions h will achieve much better performance than that of SVD [1].

4 Conclusions

A new segmentation technique for hyperspectral images was introduced. The procedure uses a sparse method for the estimation of feature densities, and includes statistical spatial information using a MLL Markov-Gibs based prior. The α-Expansion optimization tool is used to estimate the optimal segmentation.

Experiments were done using simulated datasets and an AVIRIS image. When compared with the support vector machines (SVM), the sparse multinomial logistic regression (SMLR), and the linear discriminant analysis (LDA), the proposed segmentation approach outperformed them all.

We believe that there is clearly room for improvement, namely by adopting kernel functions in the multinomial regression and by implementing accurate supervised learning of the model parameters.

Acknowledgments. The first author would like to thank the Fundação para a Ciência e a Tecnologia (FCT) for the financial support (PhD grant SFRH/BD/17191/2004). The authors acknowledge Vladimir Kolmogorov for the max-ow/min-cut C++ code made available on the web. See [13] for more details; and David Landgrebe for providing the AVIRIS data.

This work was supported by the Fundacão para a Ciência e Tecnologia, under the project PDCTE/CPS/49967/2003,and by the Instituto de Telecomunicaões under the project IT/LA/325/2005, and by CICGE under POCI 2010 programme.

References

1. Camps-Valls, G., Bruzzone, L.: Kernel-based methods for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing 43(6), 1351–1362 (2005)
2. Kumar, S., Hebert, M.: Discriminative Random Fields. International Journal of Computer Vision 68(2), 179–202 (2006)
3. Camps-Valls, G., Gomez-Chova, L., Muñoz-Marí, J., Vila-Francés, J., Calpe-Maravilla, J.: Composite kernels for hyperspectral image classification. IEEE Geoscience and Remote Sensing Letters 3(1), 93–97 (2006)
4. Plaza, A., Benediktsson, J., Boardman, J., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, A., Tilton, J., Trianni, G.: Advanced Processing of Hyperspectral Images. IEEE IGARSS Proceedings, vol. IV, pp. 1974–1979 (2006)
5. Borges, J.S., Bioucas-Dias, J., Marçal, A.R.S.: Fast Sparse Multinomial Regression Applied to Hyperspectral Data. In: Campilho, A., Kamel, M. (eds.) ICIAR 2006. LNCS, vol. 4142, pp. 700–709. Springer, Heidelberg (2006)
6. Krishnapuram, B., Carin, L., Figueiredo, M.A.T., Hartemink, A.J.: Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(6), 957–968 (2005)
7. Boykov, Y., Veksler, O., Zabih, R.: Fast Approximate Energy Minimization via Graph Cuts. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23(11), pp. 1222–1239. IEEE Computer Society Press, Los Alamitos (2001)
8. Hunter, D., Lange, K.: A Tutorial on MM algorithms. The American Statistician 58, 30–37 (2004)
9. Quarteroni, A., Sacco, R., Saleri, F.: Numerical Mathematics. TAM Series, vol. 37. Springer, Heidelberg (2000)
10. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721–741 (1984)
11. The MathWorks : MATLAB The Language of Technical Computing - Using MATLAB : version 6. The Math Works, Inc. (2000)
12. Landgrebe, D.A.: NW Indiana's Indian Pine. Available at `http://dynamo.ecn.purdue.edu/~biehl/MultiSpec/` (1992)
13. Boykov, Y., Kolmogorov, V.: An experimental comparison of mincut/max-flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(9), 1124–1137 (2004)