Hyperspectral Image Segmentation with Discriminative Class Learning

Janete S. Borges, José M. Bioucas-Dias and André R. S. Marçal
Faculdade de Ciências - Universidade do Porto - DMA, Rua do Campo Alegre 687, 4169-007 Porto, Portugal
Phone:+351220100840, Fax: +351220100809, e-mail: jsborges@fc.up.pt
Instituto de Telecomunicações, Instituto Superior Técnico, Technical University of Lisbon

***Abstract*— This paper presents a 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 neighboring 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 hyperspectral images, in two directions: 1) to improve the classification/segmentation performance and 2) to decrease the size of the training sets.**

I. Introduction

In recent years much research has been done in the field of image classification/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 in classification problems circumvents the difficulties in inferring the class densities by learning directly the boundaries between classes in the feature space. Discriminative approaches hold the state-of-the art in supervised hyperspectral image classification (see, e.g. [1]). These approaches have been successful in dealing with small class distances, high dimensionality, and limited training sets characteristic of hyperspectral imagery.

An intuitive way of improving the classification/segmentation performance of discriminative classifiers consists in adding contextual information in the form of spatial dependencies. This is, in a sense, the idea behind the *discriminative random fields* (DRFs), introduced by Kumar and Hebert [2]. This paper introduces a Bayesian approach for the segmentation of hyperspectral images. Spatial dependencies are enforced by a *multi-level logistic* (MLL) Markov-Gibs prior. This density favors neighboring labels of the same class. The class densities are build on the *fast sparse multinomial logistic regression* (FSMLR) [3], which is a fast implementation of the *sparse multinomial regression* algorithm [4]. This approach to segmentation departs substantially from that of based on the DRFs framework, since the latter adopts a supervised methodology to learn all the model parameters, leading to complex learning algorithms, still an object of research [2].

To compute an approximation of the *maximum a posteriori* probability (MAP) segmentation, we adopt the α-Expansion graph cut based algorithm proposed in [5]. This algorithm is very efficient from the numeric point of view and yields nearly optimum solutions.

The paper is organized as follows: Section 2 formulates the problem, presents a brief description of the FSMLR learning algorithm, of the MLL Markov-Gibs prior, and of the α-Expansion optimization tool. Section 3 presents results using simulated and real hyperspectral (AVIRIS) images.

II. 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 given the labels) and $p(\mathbf{y})$ is the prior over the classes.

In the present approach, we use the discriminative FSMLR classifier [3] 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 $\mathbf{y}$, we have

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

where conditional independence is understood.

The prior probabilities $p(y_i)$ associated with the training set may differ from those of the the data to classify. This deviation may be corrected by reweighting the posteriori class probabilities, as proposed in [6]. In this paper, we have assumed, however, that the classes are likely probable. We are well aware that this choice may not be the best. Nevertheless, it still leads to very good results, as shown in the Section Experimental Results.

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 the Graph-Cuts optimization algorithm.

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).

This work was supported by the Fundação para a Ciência e Tecnologia, under the project PDCTE/CPS/49967/2003, and by the Instituto de Telecomunicações under the project IT/LA/325/2005.

The authors acknowledge Vladimir Kolmogorov for the max-ow/min-cut C++ code made available on the web (see [12] for more details) and David Landgrebe for providing the AVIRIS data.

A. Class Density Estimation Using Fast-SMLR Method

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 [4].

The goal is to assign to each x_i the probability of belonging to each of the K classes, yielding K sets of feature weights, one for each class. 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

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

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 function are the linear (*i.e.*, $h(x_i) = [1, x_{i,1}, \ldots, x_{i,d}]^T$, where $x_{i,j}$ is the jth component of x_i) and the kernel (*i.e.*, $h(x) = [1, K(x, x_1), \ldots, K(x, x_n)]^T$, where $K(\cdot, \cdot)$ is some symmetric kernel function). Kernels are nonlinear mappings, thus ensuring that the transformed samples are more likely to be linearly separable.

The MAP estimate of w is

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

where $l(w)$ is the log-likelihood function and $p(w) \propto \exp(-\lambda \|w\|_1)$; λ is a regularization parameter controlling the degree of sparseness of $\hat{w}_{MAP}$. The inclusion of a Laplacian prior does not allow the use of the classical *iterative reweighted least squares* (IRLS) method [4] to learn the weights w. However, by using bound optimization tools [7], it is possible to perform exact MAP multinomial logistic regression under a Laplacian prior, with the same cost as the original IRLS algorithm for ML estimation (see [8]).

In practice, the application of SMLR to large data sets is often prohibitive. The FSMLR algorithm tackles this limitation by replacing the solution of a sequence large linear system of equations with a sequence of block Gauss-Seidel iterations [8]. More specifically, in each iteration, instead of solving the complete set of weights, only blocks corresponding to the weights belonging to the same class are solved [3]. The gain in number of floating point operations is of the order of $O(K^2)$, where K is the number of classes.

B. The MLL Markov-Gibs Prior

The MLL prior is a MRF that models the piecewise smooth nature of the real world images; *i.e.*, it favors neighboring labels of the same class.

According to the Hammersly-Clifford theorem, the density associated with a MRF is a Gibb's distribution [9]. Therefore, the prior model for segmentation has the structure

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

where Z is the 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

$$-V_c(\mathbf{y}) = \begin{cases} \alpha_{y_i} & if \quad |c| = 1 \quad \text{(single clique)} \\ \beta_c & if \quad |c| > 1 \quad \text{and } \forall_{i,j \in c}\, y_i = y_j \\ -\beta_c & if \quad |c| > 1 \quad \text{and } \exists_{i,j \in c}\, y_i \neq y_j \end{cases} \quad (5)$$

where β_c is a nonnegative constant.

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

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

where $n(\mathbf{y})$ denotes the number of cliques having the same label. The conditional probability $p(y_i = k | y_j, j \in \mathcal{S} - i)$ is then given by

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

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

C. Energy Minimization Via Graph Cuts

The MAP segmentation is given by

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

where $p(\mathbf{x}|\mathbf{y}) \propto \prod_i p(y_i | x_i)$ was learned using the FSMLR algorithm. Minimizing (8) is a hard combinatorial optimization problem. To compute a very good approximation for it, we run the graph cut α-Expansion based algorithm [5], which can be applied because the pairwise interaction term on the right hand side of (8) is equivalent to a metric[2].

III. Experimental Results

In this section, we present experimental results based on simulated and real hyperspectral data sets. The simulated feature images, $\mathbf{x}$, were generated according to the a Gaussian density $p(\mathbf{x}|\mathbf{y})$, where the prior $p(\mathbf{y})$ follows an MLL densities. The real data was acquired with the AVIRIS spectrometer.

A. Simulated Hyperspectral Images

The simulated spectral vector x_i for $i \in \mathcal{S}$, given the label y_i, is Gaussian distributed with mean $\mu(y_i)$ and covariance matrix $\sigma^2\mathbf{I}$, i.e., $x_i \sim \mathcal{N}[\mu(y_i), \sigma^2\mathbf{I}]$. The means $\mu(y_i)$, playing the role of spectral signatures, were extracted from the USGS spectral library [10].

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

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

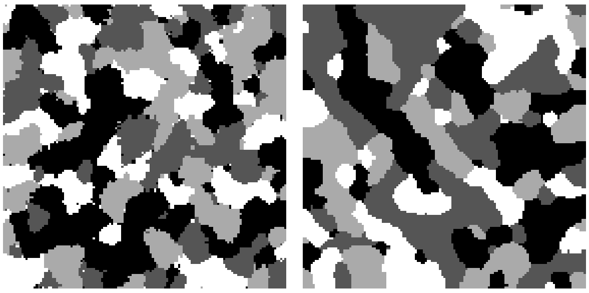

Fig. 1. Image labels with four classes generated by a MLL distribution with $\beta = 1$ and $\beta = 2$ (from left to right, respectively

The images of labels $\mathbf{y}$ were generated according to the MLL density (6) using a 2nd order neighbourhood[3]. The shape of these label images depends on the parameter β that controls the spatial continuity. In this work we considered $\beta = 1$ and $\beta = 2$. Figure 1 shows two samples of these image of labels with 4 classes and size 120×120.

The noise variance is set to $\sigma^2 = 1$, corresponding to a signal-to-noise ratio $\|\mu(y_i)\|^2/(L\sigma^2)$ less than one, and thus to a hard classification problem.

Tables I and II show, for h linear, the overall accuracy (i.e., the ratio of the correct classified pixels over the total number of pixels) as a function of the number of classes, the smoothness parameter $\widehat{\beta}$ used in the α-Expansion algorithm, and the size of the training set, whose size is 120×120 spectral vectors.

Observe that, in every case, the proposed method largely outperforms the FSMLR classification. When $K = 4$, the difference in the accuracies are larger when small training data sets are considered. The improvements in this situation ranges from 15% (for 10% of the training set) to around 7% (for 90% of the training set). When we consider $K = 10$, the improvements of the proposed method over the FSMLR classification are much higher, ranging from 42% (for 10% of the training set) to around 27% (for 90% of the training set).

We also note that when using 50% or 90% of the training data, the performance of the segmentation method proposed does not vary too much with the number of classes. On the opposite, the FSMLR decreases its performance over 30% when images with higher number of classes are considered.

Table III shows the overall accuracy obtained with a RBF kernel. In this experience, we considered only 4 classes and 10% of the training set. As with the linear kernels, the segmentation method proposed improved the results over 13%.

TABLE I

OVERALL ACCURACIES FOR THE PROPOSED SEGMENTATION METHOD AND FSMLR CLASSIFICATION WITH h, LINEAR $K = 4$, AND $\sigma = 1$.

		SIZE OF TRAINING SET		
		10%	50%	90%
	$\beta = 0.9$	96.26%	98.91%	98.96%
$\beta = 1$	$\beta = 1$	96.33%	98.92%	99.48%
	$\beta = 1.1$	96.55%	98.93%	98.82%
	FSMLR	82.13%	89.59%	92.27%
	$\beta = 1.9$	98.63%	99.36%	99.38%
$\beta = 2$	$\beta = 2$	98.49%	99.27%	98.96%
	$\beta = 2.1$	98.75%	99.27%	99.38%
	FSMLR	82.06%	89.82%	90.11%

TABLE II

OVERALL ACCURACIES FOR THE PROPOSED SEGMENTATION METHOD AND FSMLR CLASSIFICATION, WITH h LINEAR, $K = 10$, AND $\sigma = 1$.

		SIZE OF TRAINING SET		
		10%	50%	90%
	$\beta = 0.9$	66.18%	95.27%	95.85%
$\beta = 1$	$\beta = 1$	71.76%	94.99%	94.91%
	$\beta = 1.1$	68.54%	94.49%	94.77%
	FSMLR	46.18%	65.87%	69.47%
	$\beta = 1.9$	89.59%	96.61%	97.40%
$\beta = 2$	$\beta = 2$	89.38%	97.56%	97.54%
	$\beta = 2.1$	87.68%	95.71%	94.88%
	FSMLR	47.14%	67.20%	70.41%

B. Experiments on a real Hyperspectral Image

Experiments were also performed with a real hyperspectral AVIRIS spectrometer image, the Indian Pines 92 from Northern Indiana, taken on June 12, 1992 [11]. 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 9345 elements distributed by 9 classes. This dataset was randomly divided into a set of 4757 training samples and 4588 validation samples. The spatial distribution of the class labels is presented in Figure 2.

The results presented in this section are the overall accuracy measured in the independent (validation) dataset with 4588 samples. For the density estimation task, as well as for the FSMLR classification task, linear and RBF kernels were considered.

For the linear kernel case, experiments were carried out using 10%, 20% and the complete training set (100%). In the αExpansion method a $\beta = 1.5$ was defined when the complete training set is used, and $\beta = 4$ for subsets of training set. The results of overall accuracy from FSMLR classification and segmentation with MRF are presented in table IV.

From these results we can see that, regardless of the size

TABLE III

OVERALL ACCURACIES USING A RBF KERNEL IN THE ESTIMATION OF CLASS DENSITIES, FOR THE PROPOSED SEGMENTATION METHOD AND FSMLR CLASSIFICATION, USING 10% OF PIXELS AS TRAINING DATA.

	$\beta = 0.9$	$\beta = 1$	$\beta = 1.1$	FSMLR
$\beta = 1$	96.27%	96.91%	97.23%	83.53%
	$\beta = 1.9$	$\beta = 2$	$\beta = 2.1$	FSMLR
$\beta = 2$	97.88%	97.82%	97.77%	83.77%

[3] The 2nd order neighborhood of a site (i, j) is the set of sites $\mathcal{N}_{i,j} = \{(i, j+1), (i-1, j+1), (i-1, j), (i-1, j-1), (i, j-1), (i+1, j-1), (i+1, j), (i+1, j+1)\}$.

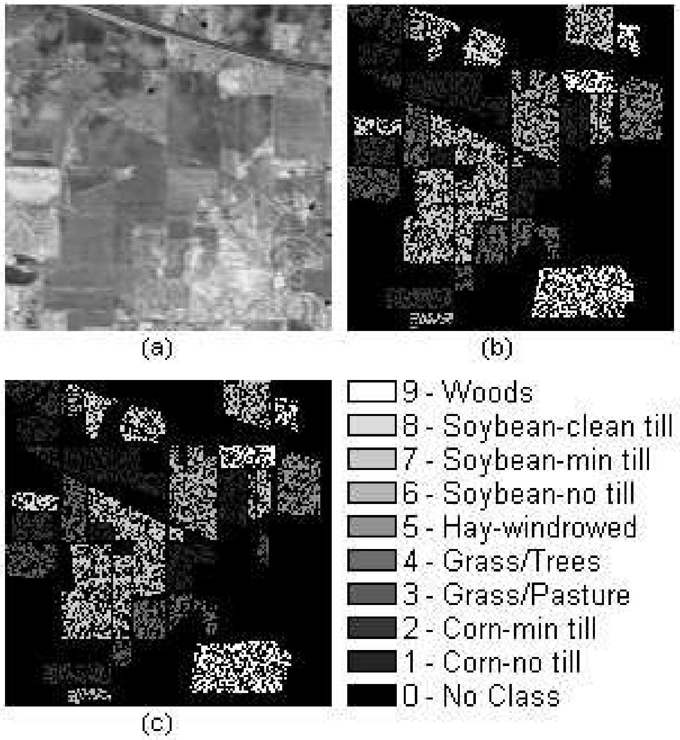

Fig. 2. AVIRIS image used for testing. (a): original image band 50 (near infrared); (b): training areas; (c): validation areas.

TABLE IV
OVERALL ACCURACY OF FSMLR CLASSIFICATION (h LINEAR) AND MRF SEGMENTATION USING 10%, 20% AND THE COMPLETE TRAINING SET.

	10%	20%	100%
FSMLR	75.57%	79.69%	85.77%
MRF	88.40%	89.56%	95.51%

of the training set used to learn the density function, the MRF segmentation largely outperforms (by over 9%) the MAP classification with FSMLR.

In the case of RBF kernel, experiments were done using only 10% of training set and $\beta = 2$. In this conditions the proposed methof yielded an overall accuracy of 91.85%, while the FSMLR classifier yielded an overall accuracy of 84.98%.

Note that these experiments with the segmentation method were done in sub-optimal conditions, since no extensive search for the optimal parameter β was performed.

It also important to note that the presented segmentation procedure outperformed in over (10%) the results achieved in [1] with linear kernels, using the complete training set. Using only 10% of the training data and without tuning all the parameters we achieved an overall accuracy over 5% higher than the one achieved in [1] with the complete training set. When RBF kernels were considered, with only 10% of training data we achieved the same results in [1] with 50% of training data, and without tuning all the parameters.

IV. CONCLUSIONS

A segmentation technique for hyperspectral images is presented. This procedure uses a sparse method for the estimation of features densities, and includes statistical spatial information using a MLL Markov-Gibs based prior. The graph cuts αExpansion algorithm is used to estimate the optimal segmentation.

Experiments were done using simulated datasets and a real hyperspectral image from AVIRIS sensor. From the results with simulated datasets, and when compared with the FSMLR classification, the segmentation method reached higher accuracies independently of the number of classes considered. Also when a higher degree of noise is considered, the segmentation method largely outperformed the FSMLR classifier. When used over a benchmark dataset, the method here proposed, outperformed (over 9.5%) the results by Camps-Valls [1], using linear kernels. Higher accuracies were also achieved using only 10% and 50% of the training set and without tuning all the parameters. When using RBF kernels the results by Camps-Valls [1] with 50% of training data were achieved using only 10% of training data, also without tuning all the parameters.

It should be noticed that these are preliminary results since no extensive search of all parameters used in this work was done. The results achieved so far are very promising and the estimation of the optimum parameters is a problem that will be addressed in future work.

REFERENCES

[1] Camps-Valls, G. and Bruzzone, L. : Kernel-based methods for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing, Vol. 43, Issue 6. (2005) 1351–1362.

[2] Kumar, S. and Hebert, M. : Discriminative Random Fields. International Journal of Computer Vision, Vol. 68, Issue 2. (2006) 179–202.

[3] Borges, J.S., Bioucas-Dias, J., Marçal, A.R.S.: Fast Sparse Multinomial Regression Applied to Hyperspectral Data. Image Analysis and Recognition. Lecture Notes in Computer Science, Vol. 4142. Springer Berlin / Heidelberg (2006) 700–709

[4] Krishnapuram, B. Carin, L. Figueiredo, M.A.T. and Hartemink, A.J.: Sparse Multinomial Logistic Regression: Fast Algorithms and Generalization Bounds. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, Issue 6. (2005) 957–968

[5] Boykov, Y., Veksler, O., Zabih, R.: Fast Approximate Energy Minimization via Graph Cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23 Issue 11. IEEE Computer Society (2001) 1222–1239

[6] McLachlan, G. J. : Discriminant Analysis and Statistical Pattern Recognition John Wiley & Sons (1992)

[7] Lange, K.: Optimization. New York: Springer Verlag (2004)

[8] Quarteroni, A., Sacco, R. and Saleri, F.: Numerical Mathematics. Springer-Verlag, New-York. (2000) TAM Series n. 37.

[9] Geman, S. and Geman, D.: Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 6. (1984) 721–741

[10] USGS Spectroscopy Lab http://speclab.cr.usgs.gov/

[11] Landgrebe, D.A. : NW Indiana's Indian Pine (1992). Available at http://dynamo.ecn.purdue.edu/˜biehl/MultiSpec/.

[12] Boykov, Y. and Kolmogorov V.: An experimental comparison of mincut/max-flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 26, Issue 9 (2004) 1124–1137