Semi-supervised Hyperspectral Image Segmentation

Jun Li, José M. Bioucas-Dias
Instituto de Telecomunicações, Instituto Superior Técnico, TULisbon, 1049-001, Lisboa, Portugal

Antonio Plaza
Department of Technology of Computers and Communications, University of Extremadura, E-10071 Caceres, Spain

ABSTRACT
This paper presents a new semi-supervised segmentation algorithm, suited to high dimensional data, of which hyperspectral images are an example. The algorithm implements two main steps: (a) semi-supervised learning, used to infer the class distributions, followed by (b) segmentation, by inferring the labels from a posterior density built on the learned class distributions and on a Markov random field. The class distributions are modeled with a multinomial logistic regression, where the regressors are learned using both labeled and, through a graph-based technique, unlabeled samples. The prior on the labels is a multi-level logistic model. The maximum a posterior segmentation is computed via a min-cut based integer optimization algorithm. We give experimental evidence that the spatial prior greatly improves the segmentation performance, with respect to that of a semi-supervised classifier. The effectiveness of the proposed method is demonstrated with simulated and real data.

Index Terms— Hyperspectral image segmentation, semi-supervised classification, Markov random field.

1. INTRODUCTION
In the recent years, important research efforts have been devoted to image classification and segmentation. Some of these efforts have been directed to remote sensing imagery, leading to good and effective results [1]. However, the classification and segmentation of high dimensional datasets, such as hyperspectral images, is still a difficult endeavor. Obstacles, such as the Hughes phenomenon, come out as the data dimensionality increases; in order to get an acceptable classification accuracy, large amount of training samples are required, which may be difficult, expensive, or sometimes plain impossible to get. These difficulties have fostered the development of new classification methods, which are able to deal with high dimensional data sets and limited training samples.

The discriminative approach to classification circumvents the difficulties in learning the class distributions in high dimensional spaces by inferring the boundaries between classes in the feature space [2, 3]. Support vector machines (SVMs) [4] and multinomial logistic regression [5], are among the state-of-the-art discriminative techniques to classification. Due to their ability to deal with large input spaces efficiently and to produce sparse solutions, SVMs have been successfully used for hyperspectral supervised classification [6–8]. The multinomial logistic regression has the advantage of learning the class distributions themselves. Effective sparse multinomial logistic regression methods are available [9]. These ideas have been applied to hyperspectral image classification [10]. In order to improve the classification accuracy, some methods have integrated spatial (contextual) information [7, 8, 10].

The acquisition of labeled training data is costly and time consuming, whereas unlabeled samples can be obtained easily. This has fostered active research on semi-supervised learning algorithms, which are trained with both labeled and unlabeled samples [11, 12]. This trend has also been followed in remote sensing applications [8,13–15]. Most semi-supervised learning algorithms use some type of regularization which encourages “close” features to belong to the same class. The effect of this regularization is to push the boundaries between classes toward regions of low data density [12]. An usual way of building such a regularizer is to associate the vertices of a graph to the complete set of samples and build a regularizer depending on variables defined the vertices.

In this paper, we introduce a new semi-supervised learning algorithm which exploits the spatial contextual information. The algorithm implements two main steps: (a) the semi-supervised learning algorithm [11] to infer the class distributions; and (b) segmentation, by inferring the labels from a posterior distribution built on the learned class distributions and on a multi-level logistic (MLL) prior. The class distributions are modeled with a multinomial logistic regression, where the regressors are learned using both labeled and, through a graph-based technique, unlabeled samples. The spatial contextual information is used both in building the graph accounting for the feature “closeness” and in the MLL prior. The maximum a posterior (MAP) segmentation is computed via a min-cut based integer optimization algorithm.

The remainder of the paper is organized as follows. Section 2 formulates the problem. Section 3 describes proposed approach. Section 4 reports segmentation results based on simulated and real hyperspectral datasets, in comparison with state-of-the-art competitors. Finally, section 5 concludes with some remarks.

2. PROBLEM FORMULATION
Let \( S \equiv \{1, \ldots, n\} \) denote a set of integers indexing the \( n \) pixels of an image; let \( L \equiv \{1, \ldots, K\} \) be a set \( K \) labels; let \( x = (x_1, \ldots, x_n) \in \mathbb{R}^{d \times n} \) denote an image in which the pixels are \( d \)-dimensional feature vectors; and finally let \( y = (y_1, \ldots, y_n) \in L^n \) denote an image of labels. With the above definitions in mind, the goal of image classification and segmentation is to estimate \( y \), having observed \( x \). In a Bayesian framework, this estimation is usually carried out by maximizing the posterior distribution \( P(y|x) \propto p(x|y)P(y) \), where \( p(x|y) \) is the likelihood function (i.e., the probability of feature image given the labels) and \( P(y) \) is the prior over the label image. Assuming conditional independency of the features given the labels, i.e., \( p(x|y) = \prod_{i=1}^{n} p(x_i|y_i) \), then the posterior

This work was supported by Marie Curie training Grant MEST-CT-2005-021175 from the European Commission.
\(P(y|x),\) as a function of \(y,\) may be written as

\[
P(y|x) = \frac{1}{p(x)} p(x|y) P(y) = \alpha(x) \prod_{i=1}^{n} \frac{p(y_i|x_i)}{p(y_i)} P(y),
\]

where \(\alpha(x) \equiv \prod_{i=1}^{n} p(x_i)/p(x)\) is a factor not depending on \(y.\) In this paper we assume, without loss of generality, that \(p(y_i) = 1/K.\) The maximum a posteriori (MAP) segmentation is then given by

\[
\hat{y} = \arg \max_{y \in \mathcal{Y}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log p(y_i|x_i) \right\} + \log P(y).
\]

In the present approach, the densities \(p(y_i|x_i)\) are learned and modeled as in [11], following a parametric semi-supervised approach. The next section summarizes the main aspects of the proposed method.

### 3. PROPOSED APPROACH

Let \(y_i = [y_{i1}, \ldots, y_{ik}]\) denote a “1-of-K” encoding of the \(K\) classes and \(\omega(k) \in \mathbb{R}^d\) be a vector of parameters associated with class \(k.\) Note the the variables \(y_i,\) just defined, and \(y_i\) have different structure but are equivalent [e.g., \(y_i = [0, 0, 1, 0]^T \Leftrightarrow (y_i = 3)].\) With this in mind, we can model the posterior densities \(p(y_i|x_i)\) with a multinomial logistic regression, i.e.,

\[
p(y_i = 1|x_i, \omega) = \frac{\exp(\omega(k)^T h(x_i))}{\sum_{k=1}^{K} \exp(\omega(k)^T h(x_i))}
\]

where \(h(x_i) = [h(1, x_i), \ldots, h(l, x_i)]^T\) is a vector of \(l\) fixed functions of the input, often termed features, \(\omega = [\omega(1)^T, \ldots, \omega(K-1)^T]^T,\) and, because the density (3) does not depend on translations on the regressors \(\omega(k),\) we take \(\omega(1) = 0.\)

It should be noted that the function \(h\) may be linear (i.e., \(h(x_i) = [1, x_i, x_i^2, \ldots, x_i^d]^T\)), where \(x_{i,j}\) is the \(j\)-th component of \(x_i\) or nonlinear. Kernels, i.e., \(h(x_i) = [1, K_{x_i,x_1}, \ldots, K_{x_i,x_d}]^T\), where \(K_{x_i,x_j} = K(x_i, x_j)\) and \(K(\cdot)\) is some symmetric kernel function, are a relevant example of the nonlinear case. Kernels have been largely used because they tend to improve the data separability in the transformed space. In this paper, we use a Gaussian Radial Basis Function (RBF) \(K(x, z) = -\exp(||x - z||^2/2\sigma^2\) kernel, which is widely used in hyperspectral image classification [6]). From now on, \(d\) denotes the dimension of \(h(x_i).\)

#### 3.1. Estimation of the Logistic Regressors

Under the present setup, learning the class densities amounts to estimating the logistic regressors \(\omega.\) Since we are assuming a semi-supervised scenario, this estimation is based on a small set of labeled samples \(D_L = \{(y_1, x_1), \ldots, (y_L, x_L)\}\) and a larger set of unlabeled samples \(D_U = \{x_{L+1}, \ldots, x_{L+U}\}\). Following [11], we compute the MAP estimate of \(\omega\) given by

\[
\hat{\omega} = \arg \max_{\omega} \left[ l(\omega) + \log p(\omega) \right],
\]

where \(l(\omega) \equiv \prod_{i=1}^{L} p(y_i|x_i, \omega)\) is the log-likelihood function of \(\omega\) given the labeled samples \(D_L\) and \(p(\omega)\) is a prior. In [11], this prior is Gaussian, with potential matrix \(\Gamma(\lambda)\) depending on a set of \(Kd\) scaling parameters \(\lambda\) and on a weighted graph \(\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{B}),\)

where \(\mathcal{V}\) is the set of vertices corresponding to labeled and unlabeled data, \(\mathcal{E}\) is a set of edges defined on \(\mathcal{V} \times \mathcal{V}\), and \(\mathcal{B}\) is a set of weights defined on \(\mathcal{E}.\) Therefore,

\[
p(\omega|\lambda) \propto \exp\left\{ -\frac{1}{2} \omega^T \Gamma(\lambda) \omega \right\},
\]

where \(\Gamma(\lambda) = \Lambda_0 \otimes \Lambda + \Lambda\) (symbol \(\otimes\) denotes the Kronecker product) and

\[
\Lambda_0 = \text{diag}(\lambda_1, \ldots, \lambda_0(K-1)),
\]

\[
\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_{(K-1)d}),
\]

\[
\mathbf{X} = [h(x_1), \ldots, h(x_{L+U})]
\]

\[
\Delta = \text{Laplacian of the graph } \mathcal{G}.
\]

In the definitions above, \(\lambda_1, \ldots, \lambda_0(K-1)\) and \(\lambda_1, \ldots, \lambda_{(K-1)d}\) are scale factors and the graph \(\mathcal{G}\) has vertices \(V = \{1, \ldots, U + L\},\) and weights \(\mathcal{B} = \{\beta_{ij}, (i, j) \in \mathcal{E}\}.\) In order to interpret the rule of the prior \(p(\omega),\) and having in mind the meaning of the Laplacian of a graph, we have

\[
\omega^{(k)^T} \Delta \omega^{(k)} = \sum_{(i,j) \in \mathcal{E}} \beta_{ij} [\omega^{(k)^T}(h(x_i) - h(x_j))]^2.
\]

Therefore, the lower values of \(\omega^{(k)^T} \Delta \omega^{(k)},\) corresponding to the most probable regressors \(\omega^{(k)},\) occur when both features \(x_i\) and \(x_j\) are in the same side of the separating hyperplane defined by \(\omega^{(k)}\). In this way, the prior acts as regularizers on \(\omega^{(k)},\) promoting those solutions for which the features connected with higher values of weights \(\beta_{ij}\) are given the same label. This implies that the boundaries among the classes tend to be pushed to the regions of low density, with respect to the underlying graph \(\mathcal{G} .\)

In this work, we exploit the fact that the data to be segmented are arranged in the form of (multi-dimensional) images and, therefore, it is very likely that two neighboring pixels have the same label. Accordingly, we have set

\[
\beta_{ij} = e^{-2 ||h(x_i) - h(x_j)||^2}, \quad \text{if } j \in \mathcal{N}_i
\]

where \(\mathcal{N}_i\) is the set of first-order neighbors of pixel \(i.\) In order to compute the MAP estimate of \(\omega^{(k)},\) we proceed as in [16], i.e. by applying the EM algorithm, where the scale factors \(\lambda\) are the missing variables with Gamma density, i.e. the conjugate priors of the inverse of a variance. The E and M steps are respectively accomplished as follows (see [16] for details):

\[
\mathbf{Y}(\mathbf{\omega}) = \mathbb{E}[\mathbf{Y}(\mathbf{\lambda})|\mathbf{\omega}],
\]

\[
\hat{\omega} = \arg \max_{\omega} \left\{ l(\omega) - (1/2) \omega^T \mathbf{Y}(\mathbf{\omega}) \right\}.
\]

Given the matrix \(\mathbf{Y}(\mathbf{\omega}),\) each M-step reduces to a logistic regression problem with a quadratic regularizer. To this end, we adopt the bound optimization approach (see details in [5]), which amounts to solve a linear system with \(d(K - 1)\) unknowns, thus with \(O(d(K - 1)^3)\) complexity. Instead of solving exactly each quadratic problem, we implement the Block Gauss-Seidel method. In practice, in each iteration we solve \(K\) systems of dimension \(d,\) thus, resulting in an improvement of the order of \(O(K^2).\) Additional details about the proposed methodology are available in [10].
In order to model spatial context, we adopt the Markov random field (MRF) isotropic Multi-Level Logistic (MLL) prior given by:

$$p(y) = \frac{1}{Z} e^{-\sum_{i,j} \delta(y_i - y_j)},$$

(9)

where $Z$ is a normalizing constant, $\delta(y)$ is the unit impulse function, and $\mu > 0$ is a parameter controlling the likelihood that two neighboring pixels belong to the same class. Note that the pairwise interaction terms $\delta(y_i - y_j)$ attach higher probability to equal neighboring labels than the other way around. In this way, the MLL prior promotes piecewise smooth segmentations.

3.2. The Multi-Level Logistic spatial prior

In order to model spatial context, we adopt the Markov random field (MRF) isotropic Multi-Level Logistic (MLL) prior given by:

$$p(y) = \frac{1}{Z} e^{-\sum_{i,j} \delta(y_i - y_j)},$$

(9)

where $Z$ is a normalizing constant, $\delta(y)$ is the unit impulse function, and $\mu > 0$ is a parameter controlling the likelihood that two neighboring pixels belong to the same class. Note that the pairwise interaction terms $\delta(y_i - y_j)$ attach higher probability to equal neighboring labels than the other way around. In this way, the MLL prior promotes piecewise smooth segmentations.

3.3. Computing the MAP Estimate via Graph-Cuts

Using the parametric semi-supervised classifier to learn $p(y|x_i)$ and the MLL prior $p(y)$, and according to (2), the MAP segmentation is finally given by:

$$\hat{y} = \arg \min_{y \in \mathcal{Y}} \left\{ \sum_{i \in \mathcal{S}} -\log p(y_i | \hat{\omega}) - \mu \sum_{i,j \in \mathcal{C}} \delta(y_i - y_j) \right\},$$

(10)

where $p(y_i | \hat{\omega}) \equiv p(y_i | x_i, \hat{\omega})$, computed at $\hat{\omega}$. Minimization of expression (10) is a combinatorial optimization problem, involving unary and pairwise interaction terms. The exact solution for $K = 2$ was introduced by mapping the problem into the computation of a min-cut on a suitable graph [17]. This approach has been intensely investigated recently. A key element in graph-cut based approaches to integer optimization is the so-called sub-modularity of

$$\sum_{i,j \in \mathcal{C}} \delta(y_i - y_j) = \sum_{i,j \in \mathcal{C}} \delta(y_i - y_j) + \sum_{i \in \mathcal{S}} \delta(y_i - 0) = \sum_{i,j \in \mathcal{C}} \delta(y_i - y_j) + \sum_{i \in \mathcal{S}} \delta(y_i - y_j) = \sum_{i,j \in \mathcal{C}} \delta(y_i - y_j) + \sum_{i \in \mathcal{S}} \delta(y_i - 0) = \sum_{i,j \in \mathcal{C}} \delta(y_i - y_j).$$

(10)

The proposed approach was also applied to the well-known AVIRIS Indian Pines scene, collected over Northwestern Indiana in June of 2001.

4. EXPERIMENTAL RESULTS

In all experiments, the spectral vectors are normalized and the RBF scale parameter is set to $\sigma = 0.6$. The prior regularization parameter is set to $\mu = 2$. The number of unlabeled samples, $U$, is set to $U = 4L$, i.e., four times the number of the labeled samples. Although these are not optimal choices, they lead to very good estimate.

4.1. Results with simulated data

A simulated data set has been created with spectral vectors generated according to a Gaussian density, $x_i \sim \mathcal{N}(m_i, \sigma^2 I)$, for $i \in \mathcal{S}$, and the means selected from a set of $K = 10$ spectral signatures according to the MLL prior $p(y)$ given by (9), with $\mu = 2$. The value of variance $\sigma^2$ is set as a function of the desired signal-to-noise ratio $\text{SNR} = E[\|m_i\|^2]/(\sigma^2 L)$ ($L$ is the number of bands). The image size is 120 (lines) $\times$ 120 (columns) $\times$ 224 (bands).

Fig. 1(top) shows the overall accuracy (OA) as a function of SNR with five training samples per class. The remaining samples are used for testing. On the other hand, Fig. 1 (bottom) shows the OA results as a function of the size of the training set, with SNR = 5 dB. For illustrative purposes, Fig. 2(left) shows the spectral band at 0.5 nm wavelength of the simulated data set with SNR= 5 dB. The difficulty of the segmentation problem is evident. Finally, Fig. 2(right) shows the segmented image obtained with 20 training samples per class. It should be noted that, in experiments, each value reported for the OA is obtained after 10 Monte Carlo runs. As expected, the proposed semi-supervised algorithm with a MLL prior outperforms the parametric semi-supervised classification [11] approach in all situations, with better results as the complexity of the problem increases.

4.2. Results with real data

The proposed approach was also applied to the well-known AVIRIS Indian Pines scene, collected over Northwestern Indiana in June of 2001.

\footnote{These signatures were randomly selected from the U.S. Geological Survey (USGS) digital spectral library, available online: \url{http://speci4lab.cr.usgs.gov}.}
1992 [1]. The scene is available online³, and contains 145 × 145 pixels and 224 spectral bands. A total of 20 spectral were removed prior to experiments due to noise and water absorption in those channels. The ground truth data contains 16 classes, out of which 7 were discarded due to their very limited size. During the experiments, we randomly choose 3, 5, 10, 15, 20 samples per class from the complete training set as the initial training samples to evaluate our algorithm. The results are compared with a semi-supervised classification approach [11], which is equivalent to our method without spatial prior, and with the semi-supervised method presented in [15]. As shown in Fig. 3, the proposed method yields competitive results, outperforming the parametric semi-supervised classification approach [11] in all cases with an OA larger than 10%. With only 3 pixel per class, it yields an OA of 63.97%. With the same experimental setup and a the number of training samples of only 5 per class, the overall accuracy of the proposed method is 74.76%, while the best result presented in [15] is 66.04%.

5. CONCLUSIONS

In this paper, a semi-supervised classification/segmentation approach for hyperspectral images has been introduced. Unlabeled samples are used to improve the estimation of the class distributions. By adopting a spatial multi-level logistic prior and computing the maximum a posteriori segmentation with the $\alpha$-expansion graph-cut based algorithm, the overall accuracy greatly improves with respect to the classification based on the learned class distributions. A comparison with a state-of-art method in [15] indicates that the proposed method yields an overall accuracy of 8.7% higher using only 5 training samples. A more detailed evaluation of the competitiveness of the proposed method will be carried out in future research.

6. REFERENCES


³http://cobweb.ecn.purdue.edu/~biehl/MultiSpec/