SEMI-SUPERVISED DISCRIMINATIVE RANDOM FIELD FOR HYPERSPECTRAL IMAGE CLASSIFICATION

Jun Li[1,2], *José M. Bioucas-Dias*[2], and *Antonio Plaza*[1]

[1]Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, University of Extremadura, E-10071 Caceres, Spain.
[2]Instituto de Telecomunicações, Instituto Superior Técnico, TULisbon, 1900-118, Lisbon, Portugal.

ABSTRACT

Remotely sensed hyperspectral imaging allows for the detailed analysis of the surface of the Earth using advanced imaging instruments which can produce high-dimensional images with hundreds of spectral bands. Supervised hyperspectral image classification is a difficult task due to the unbalance between the high dimensionality of the data and the limited availability of labeled training samples in real analysis scenarios. While the collection of labeled samples is generally difficult, expensive and time-consuming, unlabeled samples can be generated in a much easier way. This observation has fostered the idea of adopting semi-supervised learning (SSL) techniques in hyperspectral image classification. The main assumption of such techniques is that the new (unlabeled) training samples can be obtained from a (limited) set of available labeled samples without significant effort/cost. In this work, we propose a new semi-supervised discriminative random field (SSDRF) technique for spectral-spatial hyperspectral image classification. The proposed approach is validated using a hyperspectral dataset collected using NASA's Airborne Visible Infra-Red Imaging Spectrometer (AVIRIS) over the Indian Pines region. The obtained results indicate that, by automatically generating unlabeled information, the proposed SSDRF algorithm exhibits very good performance in terms of accuracies in comparison with supervised algorithms. In terms of computational cost, the proposed SSDRF algorithm self learns the classifier with the same complexity as the supervised algorithm, and converges very efficiently.

Index Terms— Hyperspectral image classification, spectral-spatial analysis, semi-supervised learning, discriminative random field, loopy belief propagation (LBP).

1. INTRODUCTION

Remotely sensed hyperspectral image classification [1] takes advantage of the detailed information contained in each pixel (vector) to generate thematic maps from detailed spectral signatures. A relevant challenge for supervised classification techniques is the limited availability of labeled training sets, since their collection generally involves expensive ground campaigns [2]. While the collection of labeled samples is generally difficult, expensive and time-consuming, unlabeled samples can be generated in a much easier way. This observation has fostered the idea of adopting semi-supervised learning (SSL) techniques in which new (unlabeled) training samples can be obtained from a (limited) set of available labeled samples without significant effort/cost [3]. The area of SSL has experienced a significant evolution in terms of the adopted models, which comprise complex generative models [4–7], self-learning models [8, 9], multi-view learning models [10, 11], transductive support vector machines (SVMs) [12, 13], and graph-based methods [14]. A survey of SSL algorithms is available in [15]. Most of these algorithms use some type of regularization which encourages the fact that "similar" features are associated to the same class. The effect of such regularization is to push the boundaries between classes towards regions with low data density [16], where the usual strategy adopted first associates the vertices of a graph to the complete set of samples and then builds the regularizer depending on variables defined on the vertices. This trend has been successfully adopted in several remote sensing image classification studies [2, 17–21].

In general, the computational complexity of SSL algorithms depends on the number of labeled and unlabeled samples, such that it is very difficult to use the whole observed image or even a large number of unlabeled samples. Therefore, one of the main difficulties for SSL is finding a trade-off between the computational complexity and the number of unlabeled samples. With a reduced number of unlabeled samples, convergence problem may be observed. Hence, it is important to use all available unlabeled samples without adding significant computational cost in comparison with the supervised algorithm which only uses the labeled information. In order to address this issue, we introduce in this work a new semi-supervised Bayesian algorithm for hyperspectral image classification which can learn from all available (labeled and unlabeled) samples without significantly increasing the computational cost with regards to supervised learning. Our newly proposed method, called semi-supervised discriminative random field (SSDRF), models the posterior class probability distributions using multinomial logistic regression (MLR) [22, 23], where the regressors are efficiently inferred by a variable splitting and augmented algorithm [24]. The spatial-contextual information in the hyperspectral scene is modeled by a Markov random field (MRF) multi-level logistic prior, which enforces neighboring pixels to belong to the same class.

The remainder of the paper is organized as follows. Section 2 formulates the considered problem and provides a supervised strategy to address it. Section 3 first introduces the proposed SSDRF algorithm. Section 4 describes an experimental evaluation of the proposed integrated approach, conducted using a real hyperspectral dataset respectively collected by AVIRIS over the Indian Pines region in NW Indiana. Comparisons with supervised techniques for hyperspectral image classification are also given. Finally, section 5 concludes with some remarks and hints at plausible future research.

This work has been supported by the European Community's Marie Curie Research Training Networks Programme under contract MRTN-CT-2006-035927, Hyperspectral Imaging Network (HYPER-I-NET). Funding from the Portuguese Science and Technology Foundation, project PEst-OE/EEI/LA0008/2011, and from the Spanish Ministry of Science and Innovation (CEOS-SPAIN project, reference AYA2011-29334-C02-02) is also gratefully acknowledged.

2. PROBLEM FORMULATION: SUPERVISED LEARNING

First of all, we define the notation that will be adopted throughout the paper. Let $\mathcal{K} \equiv \{1, \ldots, K\}$ denote a set of K class labels; let $\mathcal{S} \equiv \{1, \ldots, n\}$ denote a set of integers indexing the n pixels of a hyperspectral image; let $\mathbf{x} \equiv (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{d \times n}$ denote such hyperspectral image made up of d-dimensional feature vectors; let $\mathbf{y} \equiv (y_1, \ldots, y_n)$ denote an image of labels; let $\mathcal{D}_L \equiv \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_L, y_L)\}$ be the labeled set.

With these definitions in place, we now build the posterior density $p(\mathbf{y}|\mathbf{x})$ of the class labels $\mathbf{y}$ given the features $\mathbf{x}$, which is the engine for the class labels inference. We follow a discriminative approach, i.e., we model the distribution $p(\mathbf{y}|\mathbf{x})$ directly instead of the joint distribution $p(\mathbf{y}, \mathbf{x})$, which quite often implies simplistic assumptions about the data generation mechanism. Furthermore, because the discriminative models are less complex than the corresponding generative models, learning the former models yields often better results, in particularly when the training data are limited. Specifically, we adopt the following model for our posterior density[1]:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\boldsymbol{\omega}, \mathbf{x})} \exp\left(\sum_{i \in \mathcal{S}} \log p(y_i|\mathbf{x}_i, \boldsymbol{\omega}) + \mu \sum_{(i,j) \in \mathcal{C}} \delta(y_i - y_j) \right), \tag{1}$$

where $Z(\boldsymbol{\omega}, \mathbf{x})$ is the normalizing factor, also known as a partition function:

$$Z(\boldsymbol{\omega}, \mathbf{x}) = \sum_{\mathbf{y}} \exp\left(\sum_{i \in \mathcal{S}} \log p(y_i|\mathbf{x}_i, \boldsymbol{\omega}) + \mu \sum_{(i,j) \in \mathcal{C}} \delta(y_i - y_j) \right), \tag{2}$$

where $p(y_i|\mathbf{x}_i, \boldsymbol{\omega})$ denotes the MLR [22] with regression vector $\boldsymbol{\omega}$, μ is a parameter controlling the degree of smoothness on the image of labels, $\delta(y)$ is the unit impulse function[2], and $\mathcal{C}$ is set of cliques[3]. In the discriminative model (1), the term $p(y_i|\mathbf{x}_i, \boldsymbol{\omega})$ is itself a discriminative classifier, which gives the probability of label y_i given the feature vector $\mathbf{x}_i$, and the pairwise interaction term $\mu\, \delta(y_i - y_j)$ encodes spatial-contextual information by attaching higher probability to equal neighboring labels than the other way around. Therefore, the term $\mu\, \delta(y_i - y_j)$ promotes piecewise smooth labelings, where μ controls the degree of smoothness.

At this point, we note that the posterior (1) is a particular case of a discriminative random field (DRF) [25] with association potentials given by $\log p(y_i|\mathbf{x}_i, \boldsymbol{\omega})$ and *interaction potentials* given by $\mu\, \delta(y_i - y_j)$. The DRF is based on the concept of conditional random field [26]. In a sense, it is a generalization of the Markov random field (MRF) offering several advantages, namely: i) the relaxation of conditional independence of the observed data, ii) the exploitation of probabilistic discriminative models instead of generative MRFs, and iii) the simultaneous estimation of all DRF parameters from the training data unlike the MRF framework, where the class parameters are usually estimated independently from the field parameters. In our case, we are mainly exploiting properties ii) and iii). Concerning property i), the association potential $\log p(y_i|\mathbf{x}_i, \boldsymbol{\omega})$ –which is strongly linked with the conditional independence– yields an excellent balance between model complexity and the quality of the results.

[1] To keep the notation simple, we use $p(\cdot)$ to denote both continuous densities and discrete distributions of random variables. The meaning should be clear from the context.

[2] *i.e.,* $\delta(0) = 1$ and $\delta(y) = 0$, for $y \neq 0$.

[3] A clique is a set of labels which are neighbors of each other.

2.1. Multinomial Logistic Regression (MLR)

The posterior densities $p(y_i|\mathbf{x}_i, \boldsymbol{\omega})$ that appear in the association potentials of (1) are modeled by an MLR [22], formally given by:

$$p(y_i = k|\mathbf{x}_i, \boldsymbol{\omega}) = \frac{\exp({\boldsymbol{\omega}^{(k)}}^T \mathbf{h}(\mathbf{x}_i))}{\sum_{k=1}^{K} \exp({\boldsymbol{\omega}^{(k)}}^T \mathbf{h}(\mathbf{x}_i))}, \tag{3}$$

where $\boldsymbol{\omega} = [{\boldsymbol{\omega}^{(1)}}^T, ..., {\boldsymbol{\omega}^{(K-1)}}^T]^T$. Since the density in (3) does not depend on translations of the regressors $\boldsymbol{\omega}^{(k)}$, we take $\boldsymbol{\omega}^{(K)} = \mathbf{0}$. The term $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), ..., h_l(\mathbf{x})]^T$ is a vector of l fixed functions of the input, often termed *features*. In this paper, we use a Gaussian Radial Basis Function (RBF) kernel given by $K(\mathbf{x}, \mathbf{z}) = \exp(-\|\mathbf{x} - \mathbf{z}\|^2 / 2\sigma^2)$, which is widely used in hyperspectral image classification problems [27]. In order to control the machine complexity and its generalization capacity, we model $\boldsymbol{\omega}$ as a random vector with Laplacian density [23]:

$$p(\boldsymbol{\omega}) \propto \exp(-\lambda \|\boldsymbol{\omega}\|_1), \tag{4}$$

where λ is the regularization parameter controlling the degree of sparsity of $\boldsymbol{\omega}$. In the present problem, under a supervised scenario, learning the class density amounts to estimating the logistic regressors $\boldsymbol{\omega}$, which can be done by computing the maximum a posterior (MAP) estimate of $\boldsymbol{\omega}$:

$$\widehat{\boldsymbol{\omega}} \quad = \quad \arg\max_{\boldsymbol{\omega}} \quad \ell(\boldsymbol{\omega}) + \log p(\boldsymbol{\omega}), \tag{5}$$

where $\ell(\boldsymbol{\omega})$ is the log-likelihood function over the labeled training samples $\mathcal{D}_L$, for supervised learning, given by:

$$\ell(\boldsymbol{\omega}) \quad \equiv \quad \sum_{i=1}^{L} \log p(y_i = k|\mathbf{x}_i, \boldsymbol{\omega}). \tag{6}$$

Problem (5), although convex, it is difficult to compute because the term of $l(\boldsymbol{\omega})$ is non-quadratic and the term $\log p(\boldsymbol{\omega})$ is non-smooth. In this work, we take advantage from the logistic regression via variable splitting and augmented Lagrangian (LORSAL) algorithm [24], which allows replacing a difficult non-smooth convex problem with a sequence of quadratic plus diagonal l_2-l_1 problems very efficiently.

2.2. Maximum a posteriori marginal (MPM) labeling

The MPM estimate minimizes the Bayesian risk associated to the sitewise zero-one loss function. As in the previous section, suppose we are given the estimates $\widehat{\boldsymbol{\omega}}$ and $\widehat{\mu}$ of $\boldsymbol{\omega}$ and μ, respectively. The marginal estimate for $\mathbf{x}_i$ with respect to class k is given by:

$$\widehat{y}_i^{(k)} \quad = \quad q(y_i = k|\mathbf{x}, \widehat{\boldsymbol{\omega}}, \widehat{\mu});\ i \in \mathcal{S} \tag{7}$$

where $p(y_i|\mathbf{x})$ is the marginal density of $p(\mathbf{y}|\mathbf{x})$ with respect to $\mathbf{x}_i$. Therefore, the MPM estimate is:

$$\widehat{y}_i \quad = \quad \arg\max_{y_i} \quad q(y_i|\mathbf{x}, \widehat{\boldsymbol{\omega}}, \widehat{\mu});\ i \in \mathcal{S} \tag{8}$$

Finding a MPM solution for (8) is a combinatorial task, and thus very hard to solve exactly. In this work, we use the belief propagation (BP) algorithm to estimate the MPM solution, where BP is an efficient inference approach to estimate Bayesian beliefs [28] in graphical models.

3. PROPOSED SEMI-SUPERVISED APPROACH

In this section we revisit the SSDRF algorithm for hyperspectral classification. Since our approach is semi-supervised, we learn the classifier from the labeled data (usually a set of small size) and from the unlabeled data, which is usually a larger set. In contrast with this general scenario, in this work we use the whole image for SSL. Also in contrast with conventional SSL algorithms, in which the computational complexity generally depends on the size of the labeled and unlabeled training sets [18, 20], the proposed algorithm learns the classifier without too much additional cost with regards to the supervised case. In the following, we introduce the detailed algorithm exhibiting these advantages. A usual way to do inference with unobserved data is the expectation maximization (EM) algorithm, an iterative scheme that computes in each iteration the so-called E-step (for mean value) and the M-step (for maximization). In the present context, the E-step and M-step of this algorithm are given by:

E-step:

$$Q(\boldsymbol{\omega}|\widehat{\boldsymbol{\omega}}_t) \equiv E[\sum_{i\in\mathcal{S}} \log q(y_i|\mathbf{x},\boldsymbol{\omega})|\widehat{\boldsymbol{\omega}}_t], \tag{9}$$

M-step:

$$\widehat{\boldsymbol{\omega}}_{t+1} \equiv \arg\max Q(\boldsymbol{\omega}|\widehat{\boldsymbol{\omega}}_t), \tag{10}$$

where the E-step computes the mean value of the posterior density given by expression (1) and the M-step estimates the logistic regressors by maximizing the objective function (9). By adopting the DRF model, we explicitly include the spatial information and unlabeled information. Next, we describe these steps in more detail.

3.1. E-step

From (1) and (4), we can obtain the posterior marginal as

$$q(y_i|\mathbf{x},\boldsymbol{\omega}) \quad = \quad \frac{1}{Z(\boldsymbol{\omega},\mathbf{x}_i)} p(y_i|\mathbf{x}_i,\boldsymbol{\omega}) p(\boldsymbol{\omega}) c(\mathbf{y}) \tag{11}$$

where term $c(\mathbf{y})$ is independent from $\mathbf{x}$ and $\boldsymbol{\omega}$, and $p(\boldsymbol{\omega})$ is give by (4) which only depends on $\boldsymbol{\omega}$. We then have the E-step:

$$\begin{aligned} Q(\boldsymbol{\omega}|\widehat{\boldsymbol{\omega}}_t) \quad &\equiv \quad -E[\log Z(\boldsymbol{\omega},\mathbf{x})|\widehat{\boldsymbol{\omega}}_t] && (12)\\ &\quad + E[\sum_{i\in\mathcal{S}} \log p(y_i|\mathbf{x}_i,\boldsymbol{\omega})|\widehat{\boldsymbol{\omega}}_t] && (13)\\ &\quad + \log p(\boldsymbol{\omega}) + \log c(\mathbf{y}). && (14) \end{aligned}$$

Let the following expression:

$$\ell(\boldsymbol{\omega}) = \sum_{i\in\mathcal{S}} \log p(y_i|\mathbf{x}_i,\boldsymbol{\omega}) \tag{15}$$

be the loglikelihood function (see Eq. (13)). Notice that, for supervised learning, the learning process only uses the labeled set $\mathcal{D}_L$ [see (6]. In turn, for SSL the learning process uses the whole image $\mathbf{x}$ [see (15)]. Let $\varphi(\boldsymbol{\omega}) = \log Z(\boldsymbol{\omega},\mathbf{x})$ be the logarithm of the partition function [see (12)]. With these assumptions in mind, the cost function of the E-step can be defined as:

$$Q(\boldsymbol{\omega}|\widehat{\boldsymbol{\omega}}_t) \quad \equiv \quad -\varphi(\boldsymbol{\omega}|\widehat{\boldsymbol{\omega}}_t) + \ell(\boldsymbol{\omega}|\widehat{\boldsymbol{\omega}}_t) + \log p(\boldsymbol{\omega}). \tag{16}$$

3.2. M-step

For the present problem, the M-step amounts to maximize objective function (16) which is, although convex, difficult to compute because the terms $\ell(\boldsymbol{\omega})$ and $\varphi(\boldsymbol{\omega})$ are non-quadratic and the term $\log p(\omega)$ is non-smooth. In this work, following the previous work [22,23], we first approximate the loglikelihood function and partition function by quadratic functions. However, the problem is still difficult as the term $\log p(\omega)$ is non-smooth. Recently, this type of l_2-l_1 optimization problems have been efficiently solved by the variable splitting and augmented Lagrangian algorithms, which allow replacing a difficult complexity non-smooth convex problem with a sequence of quadratic plus diagonal $l_2 - l_1$ problems [24, 29, 30]. In this work, we use an algorithm of this kind to learn the regressors.

3.3. Semi-supervised Discriminative Random Field (SSDRF)

This subsection presents the proposed SSDRF algorithm, which is described in the form of a pseudo-code in Algorithm 1. In line 3 of Algorithm 1 the posterior marginals are estimated by the BP algorithm, which simultaneously includes both the spectral information given by the MLR model and the spatial-contextual information given by the MRF model which enforces neighboring pixels to the same class with regularization parameter μ controlling the degree of smoothness. In line 4, we learn the regressors by the proposed EM algorithm which exploits the available unlabeled information in the observed image. Notice that, as can be observed from the loglikelihood function (15), the logistic regressors $\boldsymbol{\omega}$ remain of the same size as obtained from the supervised algorithm, which only uses the labeled information. This is important as we can learn the regressors from the whole image without paying too much additional computational cost. Finally, we have empirically demonstrated that the proposed SSDRF algorithm converges very fast (less than 20 iterations).

Algorithm 1 Semi-supervised Discriminative Random Field (SSDRF)

Require: $\mathbf{x}$, $\mathcal{D}_L$, $\boldsymbol{\omega}_0$, μ, γ, λ

1: $\widehat{\boldsymbol{\omega}} := \boldsymbol{\omega}_0$
2: **repeat**
3: $\quad \widehat{\mathbf{y}} := \text{BP}(\mathcal{D}_L, \mathbf{x}, \widehat{\boldsymbol{\omega}}, \mu)$
4: $\quad \widehat{\boldsymbol{\omega}} := \text{EM}(\mathbf{x}, \widehat{\mathbf{y}}, \gamma, \lambda)$
5: **until** convergence

4. EXPERIMENTAL RESULTS

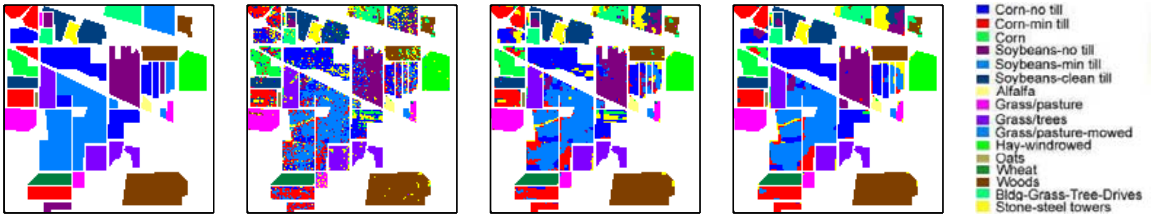
In this section, we evaluate the proposed SSDRF algorithm using a real hyperspectral dataset collected by the AVIRIS sensor over the Indian Pines region in NW Indiana[4]. This scene, with a size of 145 lines by 145 samples, was acquired over a mixed agricultural/forest area, early in the growing season. The scene comprises 202 spectral channels in the wavelength range from 0.4 to 2.5 μm, nominal spectral resolution of 10 nm, moderate spatial resolution of 20 meters by pixel, and 16-bit radiometric resolution. The ground-truth map available for the scene has 16 mutually exclusive ground-truth classes (in total, $l = 10366$ labeled samples).

Table 1 illustrates the obtained classification results as a function of the number of labeled samples. It is noticeable that the proposed SSDRF algorithm obtained very good results in terms of accuracies, in particular when a limited number of labeled samples were available. For illustrative purpose, Fig. 1 shows the obtained classification maps obtained using $l = 310$ labeled samples. Effective results can be observed from these maps despite the limited training information used for classification purposes.

[4]Available online: http://dynamo.ecn.purdue.edu/biehl/MultiSpec

Table 1. Overall accuracy (OA), average accuracy (AA) and kappa statistic (κ) (OA/AA/κ, [%]) as a function of the number of labeled samples per class (the total number of labeled samples is given in the parentheses) for the AVIRIS Indian Pines data set.

Methods	Number of labeled samples					
	5 per class ($l = 80$)	10 per class ($l = 160$)	15 per class ($l = 240$)	20 per class ($l = 310$)	25 per class ($l = 385$)	30 per class ($l = 443$)
MLR	49.92/62.10/44.97	61.16/72.03/56.98	66.58/77.85/62.79	70.29/80.11/66.70	72.84/82.30/69.50	74.58/83.24/71.39
MPM	53.06/65.82/48.55	66.60/77.20/63.00	73.08/83.52/69.98	77.71/86.32/74.95	80.42/88.80/77.96	82.56/89.79/80.30
SSDRF	59.03/64.39/54.33	75.22/79.18/71.91	77.97/86.02/75.17	82.25/87.44/79.81	83.44/89.67/81.23	86.09/90.20/84.15



(a)Ground-truth (b)MLR (70.16%) (c)MPM (78.87%) (d)SSDRF (83.61%)

Fig. 1. Classification maps for the AVIRIS Indian Pines image using $l = 310$ labeled samples along with the overall accuracies (OA).

5. CONCLUSIONS AND FUTURE RESEARCH LINES

In this paper, we introduced a semi-supervised discriminative random field (SSDRF) algorithm for hyperspectral classification, which exhibits state-of-the-art classification performance. Although the results obtained are very encouraging, further tests with additional scenes and comparison methods should be conducted. In the future, we will also develop computationally efficient implementations of the proposed approaches by resorting to parallel computer architectures such as commodity clusters or graphical processing units.

6. REFERENCES

[1] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. Hoboken, NJ: John Wiley, 2003.

[2] F. Bovolo, L. Bruzzone, and L. Carline, "A novel technique for subpixel image classification based on support vection machine," *IEEE Transactions on Image Processing*, vol. 19, pp. 2983–2999, 2010.

[3] L. Bruzzone, M. Chi, and M. Marconcini, "A novel transductive svm for the semisupervised classification of remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 11, pp. 3363–3373, 2006.

[4] S. Baluja, "Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data," in *Neural Information Procesing systems (NIPS '98)*, 1998.

[5] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training." Morgan Kaufmann Publishers, 1998, pp. 92–100.

[6] T. M. Mitchell, "The role of unlabeled data in supervised learning," in *In Proceedings of the Sixth International Colloquium on Cognitive Science*, 1999.

[7] A. Fujino, N. Ueda, and K. Saito, "A hybrid generative/discriminative approach to semi-supervised classifer degin," in *AAAI'05 Proceedings of the 20th national coference on Artificial intelligence*, vol. 2, 2005.

[8] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ser. ACL'95, 1995, pp. 189–196.

[9] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-supervised self-training of object detection models," in *Seventh IEEE Workshop on Applications of Computer Vision*, January 2005.

[10] I. Cowan, T. G, and V. R. D. Sa, "Learning classification with unlabeled data," 1994.

[11] U. Brefeld, T. Gärtner, T. Scheffer, and S. Wrobel, "Efficient co-regularised least squares regression," in *Proceedings of the 23rd international conference on Machine learning*, ser. ICML'06, 2006, pp. 137–144.

[12] V. Vapnik, *Statistical Learning Theory*. New York: John Wiley, 1998.

[13] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proceedings of the Sixteenth International Conference on Machine Learning*, ser. ICML'99, 1999, pp. 200–209.

[14] A. Blum and S. Chawla, "Learning from labeled and unlabeled data using graph mincuts," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML'01, 2001, pp. 19–26.

[15] X. Zhu, "Semi-supervised learning literature survey," Computer Sciences, University of Wisconsin-Madison, Tech. Rep. 1530, 2005.

[16] O. Chapelle, M. Chi, and A. Zien, "A continuation method for semi-supervised svms," in *Proceedings of the 23rd International Conference on Machine Learning*. ACM Press, 2006, pp. 185–192.

[17] L. Bruzzone, M. Chi, and M. Marconcini, "A novel transductive svm for the semisupervised classification of remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 11, pp. 3363–3373, 2006.

[18] G. Camps-Valls, T. Bandos, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, pp. 3044–3054, Oct 2007.

[19] L. Bruzzone and C. Persello, "A novel context-sensitive semisupervised svm classifier robust to mislabeled training samples," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 47, no. 7, pp. 2142 –2154, 2009.

[20] J. Li, J. Bioucas-Dias, and A. Plaza, "Semi-supervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 11, pp. 4085–4098, 2010.

[21] F. Ratle, G. Camps-Valls, and J. Weston, "Semisupervised neural networks for efficient hyperspectral image classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 48, no. 5, pp. 2271 –2282, may 2010.

[22] D. Böhning, "Multinomial logistic regression algorithm," *Annals of the Institute of Statistical Mathematics*, vol. 44, pp. 197–200, 1992.

[23] B. Krishnapuram, L. Carin, M. Figueiredo, and A. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 957–968, 2005.

[24] J. M. Bioucas-Dias and M. Figueiredo, "Logistic regression via variable splitting and augmented lagrangian tools," Instituto Superior Técnico, TULisbon, Tech. Rep., 2009.

[25] S. Kumar and M. Hebert, "Discriminative random fields," *International Journal of Computer Vision*, vol. 68, pp. 179–201, 2006.

[26] J. Lafferty, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *18th International Conference on Machine Learning ICML*. Morgan Kaufmann, 2001, pp. 282–289.

[27] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, pp. 1351–1362, 2005.

[28] J. Yedidia, W. Freeman, and Y.Weiss, "Understanding belief propagation and and its generalizations," in *Proceedings of International Joint Conference on Artificial Intelligence*, 2001.

[29] M. V. Afonso, J. Bioucas-Dias, and M. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *IEEE Transactions on Image Processing*, vol. 19, pp. 2345–2356, 2010.

[30] J. Li, J. Bioucas-Dias, and A. Plaza, "Hyperspectral image segmentation using a new bayesian approach with active learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 10, pp. 3947 –3960, 2011.