SEMI-SUPERVISED REMOTE SENSING IMAGE CLASSIFICATION VIA MAXIMUM ENTROPY

Ayşe Naz Erkan[1] , *Gustavo Camps-Valls*[2] *and Yasemin Altun*[3]

[1] Courant Institute, New York University, New York, USA, naz@cs.nyu.edu
[2] Image Processing Laboratory, Universitat de València, València, Spain, gustavo.camps@uv.es
[3] Max Planck Institute for Biological Cybernetics, Tübingen, Germany, altun@tuebingen.mpg.de

ABSTRACT

Remote sensing image segmentation requires multi-category classification typically with limited number of labeled training samples. While semi-supervised learning (SSL) has emerged as a sub-field of machine learning to tackle the scarcity of labeled samples, most SSL algorithms to date have had trade-offs in terms of scalability and/or applicability to multi-categorical data. In this paper, we evaluate semi-supervised logistic regression (SLR), a recent information theoretic semi-supervised algorithm, for remote sensing image classification problems. SLR is a probabilistic discriminative classifier and a specific instance of the generalized maximum entropy framework with a convex loss function. Moreover, the method is inherently multi-class and easy to implement. These characteristics make SLR a strong alternative to the widely used semi-supervised variants of SVM for the segmentation of remote sensing images. We demonstrate the competitiveness of SLR in multispectral, hyperspectral and radar image classification.

1. INTRODUCTION

Remote sensing is a discipline that studies and models the processes occurring on the Earth's surface and their interaction with the atmosphere [1]. Images acquired by airborne or satellite optical sensors measure the emergent radiation at different wavelengths, while active sensors measure the back-scattered energy emitted by the on-board antenna. In both cases, a pixel in the image can be defined as a potentially very high-dimensional vector characterizing the observed material. This information allows the characterization, identification, and classification of the land-cover classes. The main focus of remote sensing data analysis is image segmentation, however, its success is limited by the scarcity (and also the quality) of the labeled pixels. Collecting a sufficient amount of reliable labels requires a very costly terrestrial campaign, in terms of both time and human resources. As in most application domains, unlabeled remote sensing data are relatively easier to obtain as it does not require human annotator: one just has to select a set of unlabeled pixels in the image.

In the last decade, semi-supervised learning (SSL) has appeared as a promising tool for combining unlabeled data along with labeled data so as to increase the accuracy and robustness of predictions, e.g., [2] and references therein. Semi-supervised learning aims to exploit prior knowledge on the intrinsic geometry of the marginal data distribution. For instance, the intuition behind many of the SSL algorithms is that the model outputs should be smooth with respect to the structure of the data, i.e., the labels of two inputs that are similar with respect to the intrinsic geometry of data are likely to be the same. This idea is often further specified via the *cluster assumption* or the *manifold assumption*.

Two main families of SSL methods exist: generative and discriminative approaches. In this paper, we focus on discriminative models, which have been successfully used in remote sensing [3]. A general taxonomy of discriminative SSL methods can be given as follows: i. the Transductive Support Vector Machine (TSVM) [4], which maximizes the margin for labeled and unlabeled samples simultaneously; ii. Graph-based methods, in which each sample diffuses its label information to its neighbors until a global steady state is achieved on the whole data set [5, 6]; iii. the Laplacian SVM (LapSVM) [7, 8], which deforms the kernel matrix of a standard SVM (or least squares SVM) using the relations extracted from the graph Laplacian; iv. cluster and bagged kernels [9], which modify the eigenspectrum of the training kernel matrix with the spectrum of unsupervised kernels; and v. semi-supervised neural networks [10] proposed as an alternative method to overcome the high computational cost of shallow architectures such as kernel methods.

Most semi-supervised variants of SVM suffer from a high computational burden and consequently a limited number of unlabeled samples can be used for their training. This gives rise to a poor estimation of the marginal data distribution. Many heuristics have been proposed to reduce the computational cost of the TSVM. In [11], a mixed integer programming was proposed to find the labeling with

This research project was conducted while the first author was visiting Department Schölkopf at the Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

This work is supported by the Spanish Ministry of Innovation and Science under projects TEC2009-13696, AYA2008-05965-C04-03 and CSD2007-00018.

the lowest objective function. The optimization, however, is intractable for large data sets. In [12], a heuristic that iteratively solves a convex SVM objective function with alternate labeling of unlabeled samples was proposed. Yet, the algorithm is capable of dealing with a few thousand samples only. The ∇TSVM still has a cubic cost, and requires storing huge kernel matrices [13]. Several alternative proposals exist for the LapSVM, either by using a sparse manifold regularizer [14] or by using an ℓ_1 penalization term and a regularization path algorithm [15]. A second and important problem with LapSVM is related to the use of a functional form of the Laplacian eigenmaps, which yields a constrained optimization problem that is hard to solve.

On a different note, in most SSL methods, unlabeled data is integrated directly in the dual problem, often in an ad-hoc manner, e.g., via a regularizer, which we believe lacks an intuitive interpretation. Convexity is also a concern for TSVM and related methods. Finally, for most SVM variants the issue of tackling classification problems for a vast number of categories has not been solved entirely. These methods use one-versus-all schemes and majority voting, but this approach is neither natural nor well-motivated.

Here we evaluate a recent discriminative probabilistic and multi-class SSL method originally presented in [16], which solves most of the aforementioned problems. This algorithm allows a natural interpretation of model weights, and has a convex loss function which is a significant advantage. The *semi-supervised logistic regression* (SLR) algorithm is founded on information-theoretic principles. In particular, it is based on modifications to the penalty functions of the generalized maximum entropy (MaxEnt) objective in the primal, such that the expectations of similarity features over local regions are consistent. These modifications along with the minimization of the Kullback-Leibler divergence yield the SLR loss. Encoding *prior* knowledge, e.g., label proportions, is straightforward and scalability is also ensured via sparse similarity features.

The remainder of the paper is outlined as follows. Section 2 revises the generalized maximum entropy framework for conditional distributions. This gives the basis for the derivations of the semi-supervised logistic regression algorithm presented in Section 3, as a particular instance of a family of semi-supervised learning methods motivated by MaxEnt. Section 4 describes the data collection, experimental setup and a discussion on the results on multispectral, hyperspectral, and very high resolution images. Finally, Section 5 concludes with some remarks and further research directions.

2. GENERALIZED MAXIMUM ENTROPY FRAMEWORK

Generalized Maximum Entropy (MaxEnt) aims to find a distribution that minimizes a divergence, $\mathbb{D}(p|q)$ between a target distribution p and a reference distribution q (or equivalently maximizes an entropy function when q is chosen as the uniform distribution) while respecting prior information represented as potential functions in miscellaneous forms of constraints and/or penalties. When the model distribution p, which is the primal variable of the MaxEnt objective, defines a conditional distribution over classes for each data point, MaxEnt leads to discriminative learning algorithms, e.g., Logistic regression (LR) [17], kernel logistic regression (KLR) [18], or conditional random fields (CRF) [19]. Using various forms of model spaces for p, $\mathbb{D}$, and approximation criteria yields a family of inference algorithms which is referred to as the Generalized MaxEnt framework [20]. In this section, we briefly summarize the Maximum Entropy (MaxEnt) framework for conditional distributions of the form,

$$\mathcal{P} = \left\{ p \,|\, p(y|x) \geq 0, \sum_{y \in \mathcal{Y}} p(y|x) = 1, \forall x \in \mathcal{X}, \forall y \in \mathcal{Y} \right\},$$

where $\mathcal{X}$ and $\mathcal{Y}$ are respectively the input and output spaces.

In the traditional supervised setting, the divergence minimization objective is penalized with the discrepancy between observed values $\tilde{\psi}$ of some pre-defined *model* feature functions $\psi : \mathcal{X} \times \mathcal{Y} \to \mathcal{B}$ and their expected values with respect to the target distribution. Often, $\tilde{\psi}$ is the empirical average of the features and it can be derived from a set of n input-output pairs $\{(x_i, y_i)|i = 1, \ldots, n\}$, e.g., $\tilde{\psi} = \frac{1}{n}\sum_{i=1}^{n} \psi(x_i, y_i)$. When the discrepancy functions are differentiable and defined over finite dimensional feature spaces, the maximum entropy problem can be solved using Lagrangian techniques. However, in the generalized MaxEnt framework with non-differentiable penalty functions or with infinite dimensional spaces, Fenchel's duality is required for a proper analysis of the primal-dual space relations. See [21] for details.

The following lemma shows the duality of generalized MaxEnt for conditional distributions and various supervised learning methods.

Lemma 1 (MaxEnt Duality for conditionals) *Let $p, q \in \mathcal{P}$ be conditional distributions and $\mathbb{D}$ be a divergence function that measures the discrepancy between two distributions,*

$$\mathbb{D}(p|q) = \sum_{x} \tilde{\pi}(x)\mathbb{D}_x\left(p_x|q_x\right). \tag{1}$$

Moreover, let $\psi : \mathcal{X} \times \mathcal{Y} \to \mathcal{B}$ be a feature map to a Banach space $\mathcal{B}$, g be a lower semi-continuous (lsc) convex function and $\mathbb{E}_p$ is the conditional expectation operator. Also define

$$t := \min_{p \in \mathcal{P}} \left\{ \mathbb{D}(p|q) + g\left(\mathbb{E}_p[\psi]; \tilde{\psi}, \epsilon\right) \right\}, \tag{2}$$

$$d := \max_{\lambda \in \mathcal{B}^*} \left\{ - \sum_{x} \tilde{\pi}(x)\mathbb{D}_x^*(\langle\psi(x, .), \lambda\rangle\, ; q_x) \right. \tag{3}$$

$$\left. - \, g^*(\lambda; \tilde{\psi}, \epsilon) \right\},$$

where q is a reference distribution (reflecting the prior *knowledge for target distribution), $\mathcal{B}^*$ is the dual space of $\mathcal{B}$ and g^* and D^* are the convex conjugates of g and D respectively. Then, $d = t$.*

The conditional expectation is defined as

$$\mathbb{E}_p[\psi] = \sum_x \tilde{\pi}(x) \mathbb{E}_{y\sim p(.|x)}[\psi(x,y)],$$

where $\tilde{\pi}$ refers to the empirical marginal distribution. See [16] for the proof of Lemma 1.

3. METHODOLOGY

This section presents the formulation of the semisupervised logistic regression and provides details on the implementation.

3.1. Semi-supervised Logistic Regression

The semi-supervised logistic regression can be derived as a specific instance of Lemma 1 imposing additional potential functions to the primal MaxEnt problem [16]. Here, we are particularly interested in the case of expectation penalties which gives the following objective function,

$$\min_{p\in\mathcal{P}} \left\{ \mathrm{KL}(p||q) + \frac{\|\tilde{\psi} - \mathbb{E}_p[\psi]\|_2^2}{2\epsilon} + \frac{\|\Phi p\|_2^2}{2\epsilon} \right\}, \quad (4)$$

where $\mathbb{D}_x$ is set to the Kullback-Leibler (KL) divergence, $\psi(x,y)$ are the model features as in the traditional LR setting, and g is the squared-norm penalty function. We take the reference distribution q as the uniform distribution.

If the linear operator $\Phi p = \sum_x \Phi_x p_x$ over similarity feature functions ϕ is defined as

$$\phi_{k,y}(x_i, y') = \begin{cases} s(x_k, x_i) & \text{if } y = y' \text{ and } i \neq k, \\ -\sum_j s(x_j, x_i) & \text{if } y = y' \text{ and } i = k, \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

for $i \in \{1, \ldots, n\}$, then $(\Phi p)_{i,y}$ yields the following additional potential function for x_i:

$$(\Phi p)_{i,y} = \sum_{\bar{x}\in S_x} \left(s(x_i, \bar{x}) p(y|x_i) - s(x_i, \bar{x}) p(y|\bar{x}) \right), \quad (6)$$

where $s(\cdot,\cdot)$ is a similarity function between samples, and a particular form is given in Section 4, Eq. (9). This additional potential enforces the weighted averages (with respect to a predefined similarity measure encoded via the similarity features ϕ) of the model outputs in local regions centered around each instance x_i, to match the model output for that instance. In other words, this penalty function manipulates the MaxEnt objective so that the model favors smooth output probabilities over local regions.

This formulation requires nC additional optimization parameters to the standard logistic regression where n is the number of samples and C is the number of categories in the classification problem. Deriving the convex dual of (4) using Fenchel's duality yields the following semi-supervised logistic regression with ℓ_2^2 regularization,

$$\mathbf{Q}(\lambda, \gamma) = \sum_{x\in S_x} \log Z_x(\lambda;\gamma) - \left\langle \lambda, \tilde{\psi} \right\rangle + \epsilon \frac{\|\lambda\|_2^2}{2} + \epsilon \frac{\|\gamma\|_2^2}{2},$$

where λ and γ are the dual variables corresponding to the model and similarity features respectively. Z and F are given as follows,

$$Z_x(\lambda, \gamma) = \sum_y \exp\left(F(x, y; \lambda, \gamma)\right), \quad (7)$$

$$F(x, y; \lambda, \gamma) = \langle \lambda, \psi(x, y) \rangle + \sum_{\hat{x}} s(\hat{x}, x)\gamma_{xy} - \sum_{\bar{x}} s(x, \bar{x})\gamma_{\bar{x}y}. \quad (8)$$

The relation between the primal variable p and the dual variables λ and γ is given by $p(y|x) = \exp(F(x,y))/Z_x$, and the gradients with respect to λ and γ are

$$\frac{\partial \mathbf{Q}(\lambda,\gamma)}{\partial \lambda} = \mathbb{E}_{p_x}[\psi(x,y)] - \tilde{\psi} + \epsilon\lambda \,,$$

$$\frac{\partial \mathbf{Q}(\lambda,\gamma)}{\partial \gamma_{xy}} = \sum_{\breve{x}} p(y|\breve{x}) s(\breve{x}, x) - \sum_{\hat{x}} p(y|x) s(\hat{x}, x) + \epsilon\gamma \,.$$

Working in the dual space is advantageous as it yields an unconstrained optimization problem that can easily be solved via gradient descent methods. Note that the dual objective $\mathbf{Q}(\lambda, \gamma)$ is no longer the negative log-likelihood term. First, the similarity features and the corresponding optimization parameters don't exists in the inner product term involving the empirical expectation. Second, the log-partition function $\log Z_x$ is summed over both labeled and unlabeled data.

The similarity terms in F can be seen as a flow problem, where the weighted average of incoming flow from neighbors $s(\hat{x}, x)\gamma_{xy}$ is matched to the outgoing flow $s(x, \bar{x})\gamma_{\bar{x}y}$. When one discards the similarity terms, the rest of this loss function is identical to that of multinomial logistic regression.

The underlying motivation of SLR is similar to other SSL methods that use similarities to impose the smoothness criterion such as the graph based label propagation methods and Laplacian SVMs mentioned earlier. However, the resulting formulation is substantially different as it treats similarities as feature functions and associates parameters to these features individually rather than treating them uniformly as in LapSVMs.

3.2. Implementation

To train our classifier, we use the Toolkit for Advanced Optimization(TAO) software [22] which is designed for large-scale optimization problems. In particular, we have used the limited memory variable metric (LMVM) algorithm (also known as L-BFGS). Details on parameter tuning are given in Section 4.

4. EXPERIMENTS

This section presents experimental results of the SLR algorithm in various remote sensing image classification problems.

4.1. Data collection

We considered different kinds of remotely sensed images in the experiments:

- Salinas. The Salinas AVIRIS data set, collected over Salinas Valley, California. A total of 16 crop classes were labeled. However, we selected the 8 most representative classes (‘Broccoli’, ‘Celery’, ‘Corn’, ‘Fallow’, ‘Lettuce’, ‘Soil’, ‘Stubble’, and ‘Vinyard’) in the image to conduct the experiments. This is a high-resolution scene with pixels of 3.7 meters and the spectral similarity among the classes is also very high. This hyperspectral image is 217×512 and contains 224 spectral channels.

- FC1. The Flightline C1 data is a 12-bands multispectral image taken over Tippecanoe County, Indiana (USA) by the M7 scanner in June 1966. The image is 949 × 220 pixels and contains 10 classes, mainly crop types, from which we selected the 4 most represented classes in the scene.

- Naples99. Images from ERS2 synthetic aperture radar (SAR) and Landsat Thematic Mapper (TM) sensors were acquired in 1999 over Naples (Italy). The available features were the seven TM bands, two SAR backscattering intensities (0–35 days), and the SAR interferometric coherence. Since these features come from different sensors, the first step was to perform a specific processing and conditioning of optical and SAR data, which were then co-registered [23]. After pre-processing, all features were stacked at a pixel level.

- KSC. The image was acquired by the AVIRIS instrument over the Kennedy Space Center (KSC), Florida, on March 23rd, 1996. A total of 224 spectral bands of 10 nm width with center wavelengths from 400-2500 nm is acquired. The image was acquired from an altitude of 20 km and has a spatial resolution of 18 m. After removing low SNR bands and water absorption, a total of 176 bands remains for analysis. A total of 13 classes of interest were labeled representing the various land cover types of the environment. Classes were highly unbalanced, and different marsh subclasses were labeled which makes it a difficult classification problem.

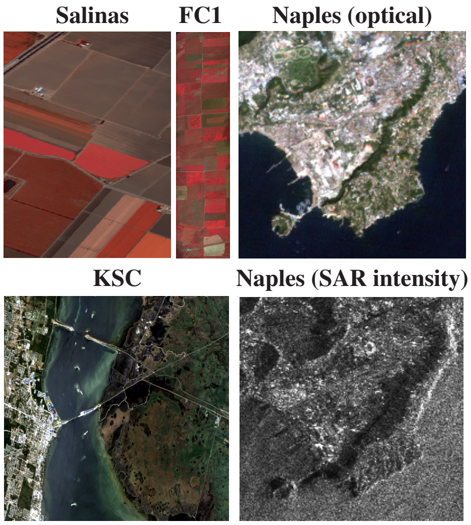

Fig. 1. RGB composition of the considered five scenes, ranging from multispectral to hyperspectral, radar and very high spatial resolution imagery.

Note that the selected images cover the most significant remote sensing situations and sensors: hyperspectral (Salinas, KSC), multispectral (FC1, Naples), radar (Naples99), and very high geometrical resolution imagery (FC1). RGB compositions along with the spectral dimensionality and spatial resolution for the considered scenes are given in Fig. 1.

4.2. Experimental setup

For all considered classification problems, we have generated three data sets: training (t), validation (v) and unlabeled (u) sets. Both training and validation sets contained the same number of labeled samples (variable in the range $[100, 500]$) and the unlabeled data set contained a total of 2000 samples (500 for the KSC data) to be used by the semi-supervised methods. The data partitioning was done for 10 different realizations and we report the averaged overall accuracy, OA[%] in inductive (validation) and transductive (unlabeled) sets in all cases. Data was scaled in the range $[0, 1]$ before training.

4.3. Model Selection for SLR

With regard the proposed SLR method, the following similarity definition was adopted:

$$s(x_i, x_j) = \begin{cases} K(x_i, x_j) & \text{if } x_j \in N_{\kappa_{x_i}}, \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where K is a Gaussian radial basis function (RBF) kernel, $K(x_i, x_j) = \exp\left(-\|x_i - x_j\|^2/2\sigma^2\right)$, where $\sigma \in \mathbb{R}^+$ is the kernel width, and $N_{\kappa_{x_i}}$ is the κ-nearest neighborhood of x_i with respect to K. Note that this similarity metric is sparse and non-symmetric. Several values for κ were tested.

The hyper-parameters of the SLR algorithm are the neighborhood size κ in (9), the regularization constant ϵ and the kernel bandwidth σ for the RBF kernel. We performed cross validation on a subset of labeled samples for model selection. From each data split we transferred 25% of the labeled samples to the corresponding unlabeled data split and found the model parameters that give the best average transduction performance on these samples only. In other words, *model selection is completely blind to the true labels of the unlabeled samples* in order to reflect the real-life scenario as closely as possible. We considered a range of hyper-parameters for model selection, $\kappa \in \{20, 30, 40\}$ and $\epsilon \in \{e^{-1}, e^{-2}, e^{-3}, e^{-4}\}$. We set σ to the median of pairwise distances.

4.4. Model Selection for Other Methods

We compare SLR with standard methods in the literature: classical SVM, regularized least squares SVM (RLSC), Laplacian SVM (LapSVM), and the Laplacian RLSC. Note that LapSVM contains other SSL methods as particular cases, and hence we are implicitly testing, for instance, spectral clustering, graph-based regularization or label propagation algorithms [7].

For all the semi-supervised SVM variants, we used the RBF kernel. The graph Laplacian consisted of labeled plus unlabeled nodes connected using κ nearest neighbors, and computed the edge weights using the Euclidean distance among samples. Two more free parameters are tuned in Laplacian methods: γ_L is the standard regularization parameter for the decision function and γ_M controls its complexity in the intrinsic geometry of the marginal data distribution. Both parameters were varied in the range $[10^{-4}, 10^4]$, the number of neighbors κ used to compute the graph Laplacian was varied from 3 to 9, and the kernel width was tuned in the range $\sigma = \{10^{-2}, \ldots, 10\}$. The selection of the best subset of free parameters was carried out via cross-validation on the training set.

4.5. Results

Results are shown in Fig. 2 for the *inductive* (prediction on the validation set) and *transductive* (prediction on the unlabeled set) settings for all considered images.

For the particular cases of FC1 and Salinas, we observe that a clear gain is obtained by all semi-supervised methods in the inductive setting, and SLR outperforms the rest with an average gain of +1.1% (FlightLine C1) and +2% (Salinas). The gain over supervised approaches is more noticeable when working with a low number of training samples. In the transductive settings, a marginal gain is obtained over LapSVM/LapRLSC for the FlightLine C1 image. For the Salinas image, a dramatic improvement is obtained when working with low-sized data sets ($n < 300$), but performance saturates for $n > 300$ and unlabeled data can worsen the results.

In the case of Naples, the SLR transduction error is lower than 1% and largely outperforms the rest of the methods while the induction accuracy is too low. Both results match with the data characteristics: we are merging features of different nature (optical and radar) so we observe that, first, the data set is very sensitive to the non-linear similarity features, and second, that a linear logistic regression may not be sufficient to solve the problem.

Finally, in the case of the hyperspectral KSC image, we observe poor performance in the inductive setting (using unlabeled samples here may even harm the solution) but high accuracy is noticed in transduction, with an average gain over the (nonlinear) Laplacian methods of around +1.5%.

5. CONCLUSIONS

We have empirically evaluated the semi-supervised logistic regression (SLR), a recently introduced information-theoretic semi-supervised algorithm, in the domain of remote sensing image classification. SLR has shown to be well suited as it is inherently a multi-class discriminative algorithm and hence it does not require 1-vs-rest inference scheme as the majority of discriminative SSL algorithms, particularly the semi-supervised variants of SVMs. The method performs well for a wide range of sensor types: ranging from low-dimensional multispectral images to very high-dimensional hyperspectral imagery, and also in the case of high spatial redundancy. Therefore, we conclude that it constitutes a powerful method for image classification.

The proposed SLR can be easily extended to semi-supervised kernel logistic regression when the features ψ are defined in a reproducing kernel Hilbert space $\mathcal{H}$. Further work will consider development and experimental comparison of this method. Finally, through the experiments, we have identified that a proper selection of training samples would play a role in the classification accuracy. Active learning or simple spatial image sampling could be considered to improve the results.

6. REFERENCES

[1] T. M. Lillesand, R. W. Kiefer, and J. W. Chipman, *Remote Sensing and Image Interpretation*, John Wiley, New York, 5th edition, 2004.

[2] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*, MIT Press, Cambridge, MA, USA, 2006.

[3] G. Camps-Valls and L. Bruzzone, Eds., *Kernel methods for Remote Sensing Data Analysis*, Wiley, UK, Dec 2009.

[4] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.

[5] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, Dec 2004, vol. 16, MIT Press.

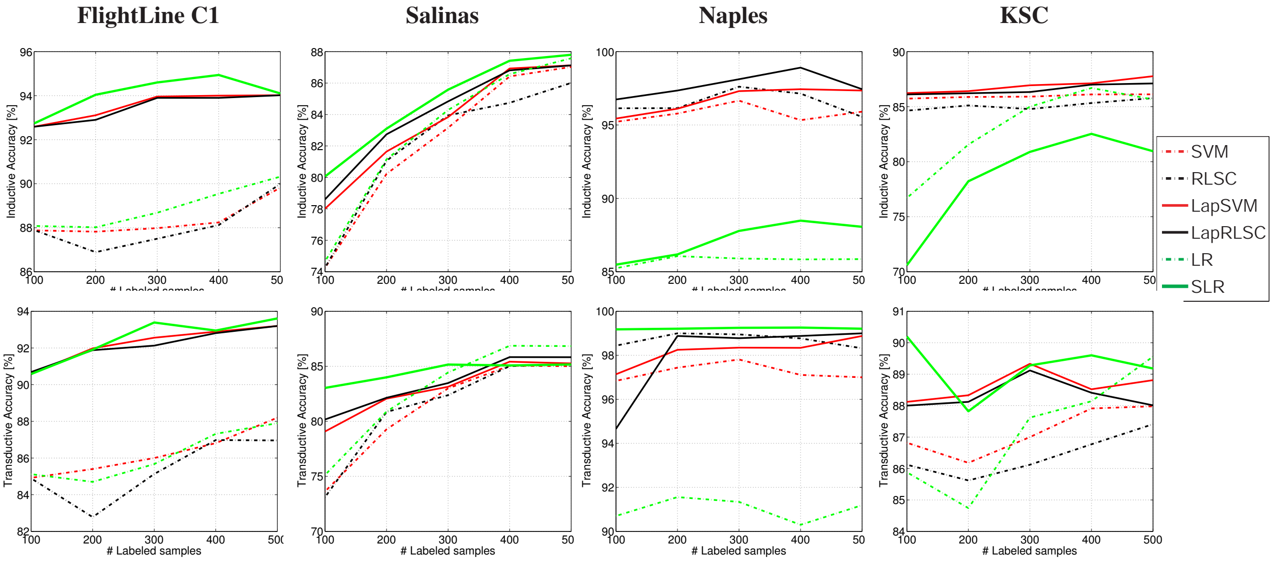

Fig. 2. Overall accuracy, [%]OA, for the considered images in both inductive (top row) and transductive (bottom row) settings, and for both supervised (dashed lines) and semisupervised (solid lines) methods.

[6] G. Camps-Valls, T. Bandos, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Trans. Geosc. Rem. Sens.*, vol. 45, no. 10, pp. 2044–3054, 2007.

[7] V. Sindhwani, P. Niyogi, and M. Belkin, "Beyond the point cloud: from transductive to semi-supervised learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, New York, USA, 2005, pp. 824–831, ACM Press.

[8] L. Gómez-Chova, G. Camps-Valls, J. Muñoz-Marí, and J. Calpe-Maravilla, "Semi-supervised image classification with Laplacian support vector machines," *IEEE Geosc. Rem. Sens. Lett.*, vol. 5, no. 3, pp. 336–340, July 2008.

[9] O. Chapelle, J. Weston, and B. Schölkopf, "Cluster kernels for semi-supervised learning," in *Advances in Neural Information Processing Systems (NIPS)*, S. Becker, S. Thrun, and K. Obermayer, Eds., Cambridge, MA, USA, 2003, vol. 15, pp. 585–592.

[10] J. Weston, F. Ratle, and R. Collobert, "Deep learning via semi-supervised embedding," in *Proceedings of the International Conference on Machine Learning (ICML)*, Helsinki, Finland, 2008, pp. 1168–1175.

[11] K. Bennet and A. Demiriz, "Semi-supervised support vector machines," in *Advances in Neural Information Processing Systems (NIPS)*, Cambridge, MA, USA, Dec 1998.

[12] T. Joachims, *Making large-scale support vector machine learning practical*, pp. 169–184, MIT Press, 1999.

[13] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2005, pp. 57–64.

[14] I. W. Tsang and J. T. Kwok, "Large-scale sparsified manifold regularization," in *Advances in Neural Information Processing Systems (NIPS)*, 2006, pp. 1401–1408.

[15] G. Gasso, K. Zapien, and S. Canu, "L1-norm regularization path for sparse semi-supervised Laplacian SVM," in *International Conference on Machine Learning and Applications (ICMLA)*, 2007.

[16] A. Erkan and Y. Altun, "Semi-supervised learning via generalized maximum entropy," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010, vol. 9, pp. 209–216.

[17] Joseph M. Hilbe, *Logistic Regression Models*, Chapman & Hall/CRC Press, New York, 2009.

[18] J. Zhu and T. Hastie, "Kernel logistic regression and the import vector machine," in *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, Dec 2002, vol. 12, MIT Press.

[19] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2001, pp. 282–289.

[20] Y. Altun and A. J. Smola, "Unifying divergence minimization and statistical inference via convex duality," in *Proceedings of the Annual Conference on Computational Learning Theory (COLT)*, 2006, pp. 139–153.

[21] J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization Theory and Examples*, Springer, 2006.

[22] S. Benson, L. C. McInnes, J. Moré, T. Munson, and J. Sarich, "TAO user manual (revision 1.9)," Tech. Rep. ANL/MCS-TM-242, Mathematics and Computer Science Division, Argonne National Laboratory, 2007, http://www.mcs.anl.gov/tao.

[23] L. Gómez-Chova, D. Fernández-Prieto, J. Calpe, E. Soria, J. Vila-Francés, and G. Camps-Valls, "Urban monitoring using multitemporal SAR and multispectral data," *Patt. Rec. Lett.*, vol. 27, no. 4, pp. 234–243, Mar 2006.