UNIVERSITY
OF TRENTO

DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY

38050 Povo – Trento (Italy), Via Sommarive 14
http://www.dit.unitn.it

A NOVEL CONTEXT-SENSITIVE SVM FOR CLASSIFICATION OF REMOTE SENSING IMAGES

F. Bovolo, L. Buzzone, M. Marconcini, C. Persello.

June 2006

Technical Report DIT-06-040

A Novel Context-Sensitive SVM for Classification of Remote Sensing Images

Francesca BOVOLO, Lorenzo BRUZZONE, Mattia MARCONCINI, Claudio PERSELLO

Dept. of Information and Communication Technologies, University of Trento,
Via Sommarive, 14, I-38050 Trento, Italy
E-mail: lorenzo.bruzzone@ing.unitn.it

***Abstract* – In this paper, a novel context-sensitive classification technique based on Support Vector Machines (CS-SVM) is proposed. This technique aims at exploiting the promising SVM method for classification of 2-D (or n-D) scenes by considering the spatial-context information of the pixel to be analyzed. In greater detail, the proposed architecture properly exploits the spatial-context information for: i) increasing the robustness of the learning procedure of SVMs to the noise present in the training set (mislabeled training samples); ii) regularizing the classification maps. The first property is achieved by introducing a context-sensitive term in the objective function to be minimized for defining the decision hyperplane in the SVM kernel space. The second property is obtained including in the classification procedure of a generic pattern the information of neighboring pixels. Experiments carried out on very high geometrical resolution images confirm the validity of the proposed technique.**

***Keywords* – Support Vector Machine, Supervised Classification, Image Classification, Context-Sensitive Classification, Remote Sensing.**

I. INTRODUCTION

Image classification is one of the most common applications of the automatic analysis of remote sensing data. Although often in real applications and commercial software packages image classification problems are addressed according to pixel-based (context-insensitive) classifiers, from a theoretical and practical point of view it is very important to develop classification techniques capable to exploit the spatial-context information present in the images. In this framework, it seems particularly relevant to develop context-sensitive classification methods capable to properly exploit the most promising pixel-based classification methodologies recently proposed in the literature. Among them, one of the most effective approaches based on machine learning consists in Support Vector Machines (SVMs). SVMs, originated from the statistical learning theory formulated by Vapnik [1], are a distribution-free classification approach, which proved very effective in many context-insensitive classification problems. SVM-based classifiers have four main advantages with respect to standard machine learning techniques based on neural networks: i) simple architecture design; ii) relatively low computational complexity; iii) learning phase associated with the optimization of a convex cost function; iv) excellent generalization capability [1]. In particular, advantage iv) is very relevant in image classification problems. In greater detail, designing a classifier characterized by good generalization properties requires assuming the statistical independence of training samples. In image classification problems this assumption is frequently violated due to high dependency between neighboring pixels; thus, the excellent generalization ability of SVMs and their robustness to the Hughes phenomenon seem very suitable to the solution of such kind of problems. However, at the present, only very few preliminary investigations on the exploitation of SVM in the framework of a context-sensitive architecture have been proposed [2], [3]. These investigations are based on the estimation of the statistical terms of classes from the output of SVMs and on their integration in a Markov Random Field (MRF) framework. However, they are not related to the definition of an intrinsically context-

sensitive SVM technique.

In this paper, a novel Context-Sensitive SVM (CS-SVM) technique is proposed. This technique aims at exploiting the promising SVM method for classification of 2-D (or n-D) scenes by considering the spatial-context information of the pixel to analyze. The proposed novel approach has two main properties: i) it exploits the spatial-context information for the definition of the hyperplane in the SVM kernel space, by properly defining a context-sensitive cost function; ii) it considers the spatial-context information in the classification phase, relating the classification output for a generic pixel to the behavior of other pixels in a predefined neighborhood system of it. The first property, which is very important, results in an increased robustness of the learning procedure to the noise present in the training set (mislabeled training samples). The second property involves regularized classification maps, in which the noise is sharply reduced.

II. Proposed Context-Sensitive SVM Classifier

Let $\mathbf{X}$ denote the d-dimensional remote-sensing image (of size $I \times J$ pixels) to be classified. Let us assume that the available training set T is made up of n patterns, i.e. $T = \{\mathbf{x}_i\}_{i=1}^{n}, \mathbf{x}_i \in \mathbf{X}$. For the sake of simplicity, since SVMs are binary classifiers, we focus the attention on the two-class case. Let $Y = \{y_i\}_{i=1}^{n}$ denote the set of labels associated with training samples, where $y_i \in \{-1, +1\}$. For the generalization to the multiclass case, many different architectures can be adopted (e.g., One-Against-One, One-Against-All [4]). We define with $\Delta_m(\mathbf{x})$ a local neighborhood system (whose shape and size depend on the specific investigated image and application) of the generic pattern $\mathbf{x}$, where m represents the number of pixels considered in the neighborhood.

In order to characterize the spatial-context information of a generic pattern $\mathbf{x}$, we define the *mean contextual value* $\tilde{\mathbf{x}}$ as follows:

$$\tilde{\mathbf{x}} = \frac{1}{m} \sum_{j \in \Delta_m(\mathbf{x})} \mathbf{x}_j \tag{1}$$

The main idea at the basis of the proposed methodology is to include the spatial-context

information in two different phases of the SVM algorithm: i) the learning phase (i.e., definition of the separation hyperplane in the kernel space); ii) the testing phase (i.e., production of the classification map). In the following, these two phases are analyzed in detail.

A. Context-Sensitive Learning Phase

The rationale of defining a context-sensitive learning phase of the classifier consists in reducing the effects of possible mislabeled pixels (i.e., pixels with a wrong label) present in the training set in the estimation of the classifier parameters. This is a critical problem of supervised classification of remote sensing images, because often training sets, which are defined according to ground truth surveys (or photointerpretation), are affected by a non-negligible number of mislabeled samples. The proposed methodology properly weights the importance of each pattern in the training procedure according to the behavior of unlabeled pixels in its neighborhood system. This is achieved by defining a proper cost function, which is composed of three main terms: i) a standard term that expresses the concept of margin maximization; ii) a standard penalization term, which regularizes the cost function with respect to classification errors on training samples (considered without their contextual information); iii) a novel context term (composed of a number of contributions depending on the order of the considered neighborhood system) that regularizes the learning process with respect to the behavior of pixels in the neighborhood of the pixel under investigation. The resulting bound minimization problem is the following:

$$\begin{cases} \min\limits_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\varphi}} \left\{ \dfrac{1}{2}\|\mathbf{w}\|^2 + C\sum\limits_{i=1}^{n}\xi_i + K\sum\limits_{i=1}^{n}\varphi_i \right\} \\ y_i\left(\left\langle \phi(\mathbf{x}_i), \mathbf{w} \right\rangle + b\right) \geq 1 - \xi_i \\ y_i\left(\left\langle \phi(\tilde{\mathbf{x}}_i), \mathbf{w} \right\rangle + b\right) \geq 1 - \varphi_i \quad \forall i = 1,\ldots,n \\ \xi_i, \varphi_i \geq 0 \end{cases} \tag{2}$$

where $\mathbf{w}$ is the vector normal to the separating hyperplane, b is a constant such that $\frac{b}{\|\mathbf{w}\|}$ represents the distance of the hyperplane from the origin, $\phi(\cdot)$ is a non-linear mapping function, $\{\xi_i\}_{i=1}^{n}$ are slack variables that control the empirical risk (i.e., the number of training errors) and

$C \in \mathbb{R}_0^+$ is a regularization parameter which tunes the trade-off between the empirical error and the complexity term (i.e., the generalization capability). With respect to standard context-insensitive SVMs, we also introduce the slack variables $\{\varphi_i\}_{i=1}^n$, defined as:

$$\varphi_i = \varphi_i(\tilde{\mathbf{x}}_i, y_i, \mathbf{w}, b) = \max\left\{0, 1 - y_i\left(\left\langle \phi(\tilde{\mathbf{x}}_i), \mathbf{w}\right\rangle + b\right)\right\} \tag{3}$$

which depend on $\tilde{\mathbf{x}}_i$ and thus permit to consider the contextual information. The rationale of the use of this term in the learning phase is based on the assumption that pixels in the neighborhood of x_i have a high probability to have the same label y_i. This is modeled according to the use of the mean contextual value, which imposes a constraint in the learning of the classifier. The term $K \in \mathbb{R}_0^+$ tunes the penalty for the contextual information. In particular, for a small value of the ratio C/K the spatial context has the highest relevance in defining the separation hyperplane; otherwise, the greatest importance is given to the error on the training patterns.

By exploiting the Lagrange theory, we define the Lagrange function for the primal problem as follows:

$$\begin{aligned} &L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\varphi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{r}, \mathbf{s}) = \\ &= L = \begin{pmatrix} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n \xi_i + K\sum_{i=1}^n \varphi_i + \\ -\sum_{i=1}^n \alpha_i\left[y_i\left(\left\langle \phi(\mathbf{x}_i), \mathbf{w}\right\rangle + b\right) - 1 + \xi_i\right] - \sum_{i=1}^n r_i\xi_i + \\ -\sum_{i=1}^n \beta_i\left[y_i\left(\left\langle \phi(\tilde{\mathbf{x}}_i), \mathbf{w}\right\rangle + b\right) - 1 + \varphi_i\right] - \sum_{i=1}^n s_i\varphi_i \end{pmatrix} \end{aligned} \tag{4}$$

where $\alpha_{i=1}^n$, $\beta_{i=1}^n$, $r_{i=1}^n$, $s_{i=1}^n$ are multipliers associated with training patterns, mean contextual values, non-contextual slack variables $\{\xi_i\}_{i=1}^n$ and contextual slack variables $\{\varphi_i\}_{i=1}^n$, respectively.

Accordingly, it is possible to reformulate (2) as:

$$\begin{cases} \max\limits_{\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\varphi}, \boldsymbol{\alpha}, \boldsymbol{\beta}} \{L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\varphi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{r}, \mathbf{s})\} \\ \xi_i, \varphi_i, \alpha_i, \beta_i, r_i, s_i \geq 0 \quad \forall i = 1, \ldots, n \end{cases} \tag{5}$$

On the basis of the proposed cost function, we adjust the Karush-Kuhn-Tucker conditions,

which are necessary and sufficient conditions for solving (5):

$$\begin{cases} \dfrac{\partial L}{\partial \mathbf{w}} = \dfrac{\partial L}{\partial b} = \dfrac{\partial L}{\partial \xi_i} = \dfrac{\partial L}{\partial \varphi_i} = 0 \\ y_i \left(\left\langle \phi(\mathbf{x}_i), \mathbf{w} \right\rangle + b \right) - 1 + \xi_i \geq 0 \\ y_i \left(\left\langle \phi(\mathbf{x}_i), \mathbf{w} \right\rangle + b \right) - 1 + \varphi_i \geq 0 \\ \xi_i, \varphi_i, \alpha_i, \beta_i, r_i, s_i \geq 0 \\ \alpha_i \left[y_i \left(\left\langle \phi(\mathbf{x}_i), \mathbf{w} \right\rangle + b \right) - 1 + \xi_i \right] = 0 \\ \beta_i \left[y_i \left(\left\langle \phi(\tilde{\mathbf{x}}_i), \mathbf{w} \right\rangle + b \right) - 1 + \varphi_i \right] = 0 \\ \xi_i (\alpha_i - C) = 0 \\ \varphi_i (\beta_i - K) = 0 \end{cases} \quad \forall i = 1, \ldots, n \tag{6}$$

Finally, we can formulate the dual problem as follows:

$$\begin{cases} \max\limits_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \left\{ \sum\limits_{i=1}^{n} (\alpha_i + \beta_i) - \dfrac{1}{2} \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} y_i y_j \cdot \begin{bmatrix} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) + \\ \beta_i \beta_j k(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) + \\ 2\alpha_i \beta_j k(\mathbf{x}_i, \tilde{\mathbf{x}}_j) \end{bmatrix} \right\} \\ \sum\limits_{i=1}^{n} y_i (\alpha_i + \beta_i) = 0 \\ 0 \leq \alpha_i \leq C \qquad \forall i = 1, \ldots, n \\ 0 \leq \beta_i \leq K \end{cases} \tag{7}$$

where, according to the Mercer's theorem, $k(\cdot,\cdot)$ is a kernel function such that $k(\cdot,\cdot) = \langle \phi(\cdot), \phi(\cdot) \rangle$. One can note that, whether $\alpha_{i=1}^{n}$ are superiorly bounded by C, the upper bound for $\beta_{i=1}^{n}$ becomes K. It is possible to prove that the cost function maintains the important convexity property (which is typical of SVMs). This is a fundamental aspect, because it results in the possibility of using quadratic programming methods for solving the dual optimization problem. In particular, in the proposed technique, a properly modified version of the Sequential Minimal Optimization (SMO) algorithm [5] has been used. After the Lagrange multipliers $\alpha_{i=1}^{n}$ and $\beta_{i=1}^{n}$ are fixed, for the generic pixel $\mathbf{x}$ the output of the discriminant function is given by:

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b = \sum_{i=1}^{n} \left[y_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + y_i \beta_i k(\tilde{\mathbf{x}}_i, \mathbf{x}) \right] + b \tag{8}$$

B. Context-Sensitive Classification Phase

The rationale of the context-sensitive classification phase consists in producing regularized

classification maps, in which the prior information on the spatial autocorrelation of images in the scene is properly considered. To this end, two different strategies can be used: i) to extract the probabilistic terms of classes from the output of SVMs according to a logistic regression procedure and integrate these probabilities in a Markov Random Field (MRF) framework [2], [3]; ii) to include directly in the definition of the output of SVM a regularization term. In the approach presented in this paper, we considered the latter strategy.

For a given pixel $\mathbf{x}$, the final predicted label of the proposed CS-SVM, $\hat{y}$, is obtained according to:

$$\hat{y} = \text{sgn}\left[\hat{f}(\mathbf{x})\right] = \text{sgn}\left[f(\mathbf{x}) + \frac{Q}{m}\sum_{j\in\Delta_m(\mathbf{x})} f(\mathbf{x}_j)\right] \tag{9}$$

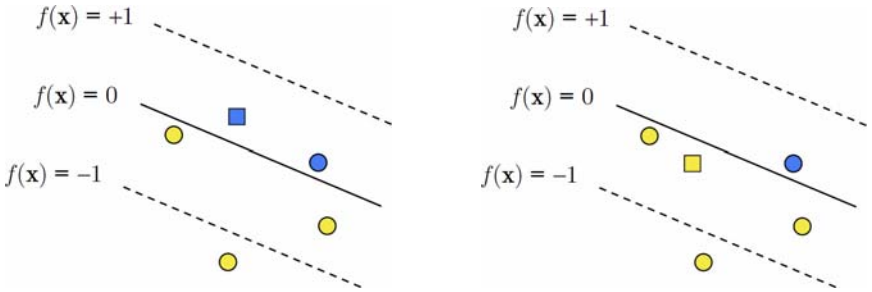
where the decision function depends on two terms:

a term related to the position of the considered pixel with respect to the discriminant function (i.e., the hyperplane in the kernel space derived according to the context-sensitive learning procedure described in the previous paragraph);

a term that properly considers the positions (with respect to the discriminant function) of the pixels in the neighborhood system of the analyzed pattern.

The second term plays the role of a regularization term driven from the spatial-context information included in the neighborhood system of the analyzed pixel $\mathbf{x}$. In particular, it depends on the average of the outputs of the discriminant function for the pixels in the neighborhood (see Fig. 1).

The above-mentioned terms are related by the spatial-regularization parameter $Q \in \mathbb{R}_0^+$, which tunes the effects of the contextual information on the classification map (i.e., it tunes the trade-off between the regularization of the map and the preservation of geometrical details). For high values of Q the final classification map may exhibit a loss of precision in representing the details of the image; accordingly, a proper tuning phase of Q is necessary.



(a) **(b)**

Figure 1 Example of classification outputs obtained by: (a) a standard context-insensitive SVM; (b) the proposed CS-SVM. A first order neighborhood system (i.e., $m = 4$) is considered. The investigated pixel is represented as a square, whereas its 4 neighboring pixels are represented as circles. Class "+1" is associated with the blue color; class "-1" is associated with yellow color.

III. Experimental Results

To assess the effectiveness of the proposed approach, several experiments were carried out on a data set made up of a very high-resolution multispectral image acquired by the Ikonos satellite over the city of Ypenburg (Netherlands) (Fig. 1 (a)). The 4 [m] spatial resolution spectral bands have been reported to a 1 [m] spatial resolution according to a Pansharpening procedure. The available ground truth was used to derive a training set and a test set for the considered image (see Table. I).

TABLE I. Number of Patterns in the Training and Test Sets

Class		Training Set	Test set
Grass		920	1073
Roads		925	752
Buildings	Small-aligned	489	399
	White-roof	919	819
	Gray-roofs	800	671
	Red-roof	253	184
Shadow		500	461

The experiments were conducted in two different conditions: i) considering the available original

training set; ii) simulating the inclusion of samples with wrong labels in the training set (different training sets with increasing percentage of wrong samples were defined). The results provided by the proposed CS-SVM were compared with those achieved by a standard context-insensitive SVM classifier. A One-Against-All strategy has been adopted for defining a multiclass classifier from binary SVMs.

In the first experiment, both the proposed CS-SVM and the standard SVM were trained on the original training set. We used in both cases *Gaussian* kernels and optimized the regularization parameters and the spread of the kernels according to a grid search model-selection strategy. The same parameters were considered for all the binary SVMs included in each multiclass architecture. Different trials varying the weight of the contextual term Q (see (9)) in the CS-SVM were performed. For space constraint here only the results for $Q=2$ are reported. As one can see from Table II, the use of the contextual information allowed to increase of 2% the overall accuracy of the classification process. Furthermore, by a qualitative analysis of the classification maps (compare Fig. 1 (b) and (c)) it is possible to conclude that, as expected, the map obtained with the proposed CS-SVM is more regularized than the one obtained with the standard context-insensitive SVM.

In the second experiment, we compared the effectiveness of the proposed CS-SVM with that of the standard SVM, introducing in the training set different percentages of wrong samples. In order to better understand the properties of the proposed CS-SVM, in this experiment we also analyzed separately the effect of the context-sensitive learning of SVM and that of the use of the contextual information both in the learning and in the classification phases. By analyzing Table III one can see that the higher is the percentage of wrong patterns in the training set, the higher is the accuracy improvement obtained using the spatial context information in the training and classification phases. For example, with 20% wrong patterns in the training set, the use if the context-sensitive learning procedure increased the Kappa coefficient of 0.03, while the use of the context information in both the learning and the classification phases sharply increased the Kappa coefficient of more than 0.06.

IV. Discussion and Conclusion

In this paper, a novel context-sensitive image-classification approach based on SVMs has been proposed. The proposed architecture properly exploits the spatial-context information in both the learning and the classification phases of the SVM. This involves: i) an increase of the robustness of the learning procedure of SVMs to the noise present in the training set (mislabeled training samples); and ii) proper regularization of the classification map, in which isolated errors are eliminated. In order to obtain these properties we defined in the proposed CS-SVM: i) a modified cost function for identifying the decision hyperplane in the SVM kernel space

TABLE II. Kappa Coefficient and Overall Accuracy Provided by the Proposed CS-SVM and the Standard SVM

Classifier	Kappa coefficient	Overall accuracy (%)
Standard SVM	0.9147	92.92
CS-SVM ($Q=2$)	0.9376	94.83

TABLE III. Kappa Coefficient of Accuracy Provided by the Standard SVM and the Proposed CS-SVM

Wrong Pattern (%)	Standard SVM	Proposed CS-SVM (only training)	Proposed CS-SVM
5%	0.8954	0.9077	0.9350
13%	0.9010	0.9078	0.9307
20%	0.8766	0.9070	0.9418

during the learning phase, which contains a proper spatial-context cost term; ii) a decision strategy that considers the information of patterns in the neighborhood of the analyzed pixels. Experimental results carried out on high spatial resolution remote sensing images confirmed the effectiveness of the proposed context-sensitive approach, which significantly outperformed the standard context-insensitive SVM classifier on different data sets (for space constraints in this paper we reported results obtained only on one of them).

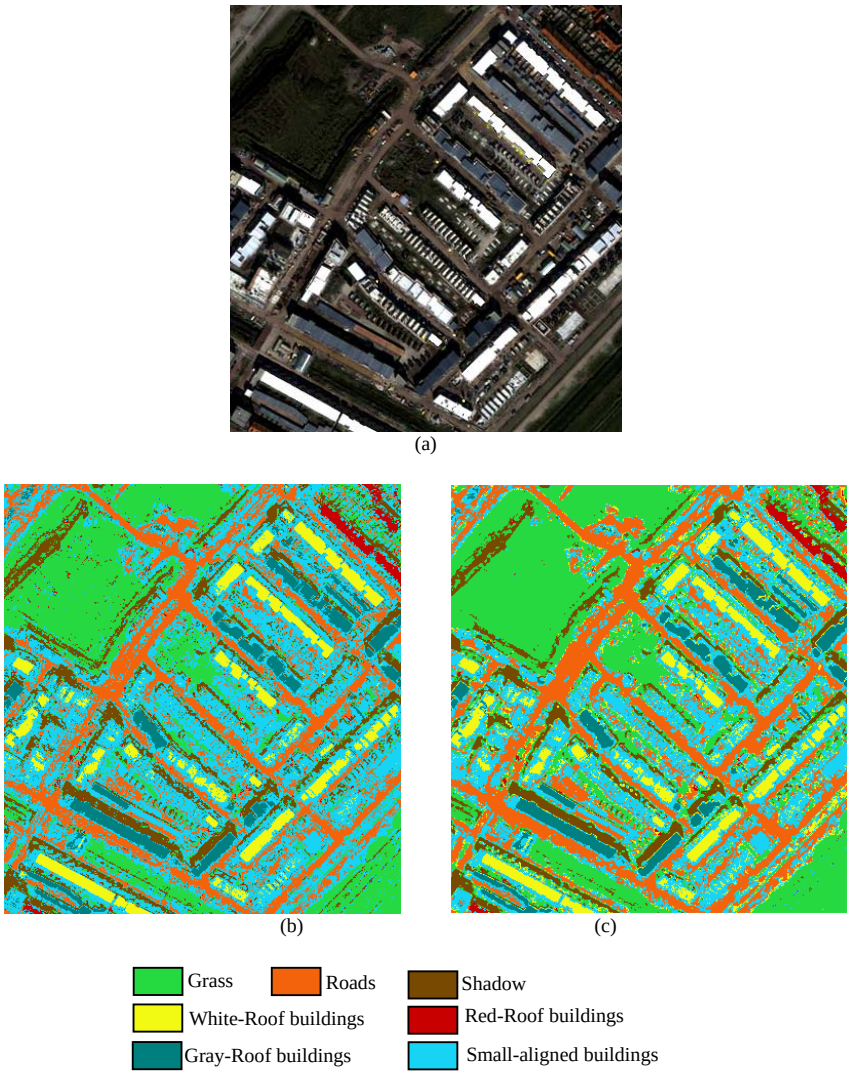


Figure 2 (a) Pansharpened Ikonos color composite image. (b) Classification map obtained with the standard SVM. (c) Classification map obtained with the proposed CS-SVM.

V. REFERENCES

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

[2] F. Bovolo, L. Bruzzone, "A Context-Sensitive Technique Based on Support Vector Machines for Image Classification," *Proc. IEEE Pattern Recognition and Machine Intelligence Conference* (PReMI 2005), Lecture Notes in Computer Science, Vol: 3776, Kolkata-India, 18-22 December, 2005.

[3] A.A. Farag, R.M. Mohamed, A. El-Baz, "A unified framework for MAP estimation in remote

sensing image segmentation," *IEEE Trans. Geosci. Rem. Sens.*, Vol. 43, No. 7, July 2005, pp. 1617-1634.

[4] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University press, U.K., 1995.

[5] J. Platt, "Fast training of support vector machines using sequential minimal optimization," in Advances in Kernel Methods: Support Vector Learning. MIT Press, pp. 185-208, 1998.