SEMI-SUPERVISED AND UNSUPERVISED NOVELTY DETECTION USING NESTED SUPPORT VECTOR MACHINES

Frank de Morsier, Maurice Borgeaud[‡], Christoph Küchler[†], Volker Gass[⋆] and Jean-Philippe Thiran

LTS 5, École Polytechnique Fédérale de Lausanne, Switzerland
[‡] European Space Agency, ESRIN, Frascati, Italy
[†] RUAG Schweiz AG, Emmen, Switzerland
[⋆] Space Center, École Polytechnique Fédérale de Lausanne, Switzerland

ABSTRACT

Very often in change detection only few labels or even none are available. In order to perform change detection in these extreme scenarios, they can be considered as novelty detection problems, semi-supervised (SSND) if some labels are available otherwise unsupervised (UND). SSND can be seen as an unbalanced classification between labeled and unlabeled samples using the Cost-Sensitive Support Vector Machine (CS-SVM). UND assumes novelties in low density regions and can be approached using the One-Class SVM (OC-SVM). We propose here to use nested entire solution path algorithms for the OC-SVM and CS-SVM in order to accelerate the parameter selection and alleviate the dependency to labeled "changed" samples. Experiments are performed on two multitemporal change detection datasets (flood and fire detection) and the performance of the two methods proposed compared.

Index Terms— Novelty detection, Semi-Supervised, Solution Path, Nested SVM, Low Density Criterion

1. INTRODUCTION

One of the major challenges in remote sensing classification and change detection is the lack of groundtruth information. In most of the change detection situations the characteristics of changes are difficult to model beforehand or even unknown. These situations are often reformulated as novelty detection or one-class classification problems: a few labels on "unchanged" regions are available and none on "changed" regions, which are detected as outliers (novelties) [1].

Kernel methods in remote sensing analysis have shown great performances, handling non-linear relationships, being robust to noise from intrinsic regularization and having good generalization properties [1]. More specifically, novelty detection problems such as anomaly detection [2] and one-class classification [3] have been approached in remote sensing using the One-Class Support Vector Machines (OC-SVM).

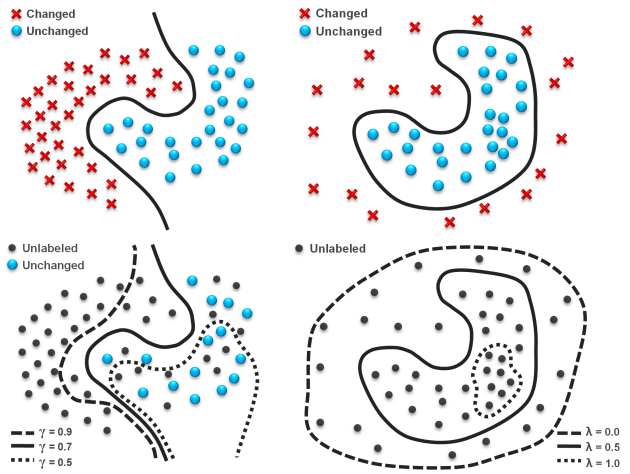
The availability of unlabeled samples can be used under certain assumptions to improve classification accuracy [4]. In [3], Semi-Supervised Novelty Detection (SSND) is performed by deforming the OC-SVM kernel using a graph Laplacian built on the unlabeled samples (*S2OC-SVM*). SSND can also be treated as an unbalanced two-class classification problem where a Cost-Sensitive SVM (CS-SVM) is trained with labeled samples classified against unlabeled samples [5]. The training errors costs are biased to penalize less the errors made on unlabeled samples than on labeled samples. The CS-SVM has given very good results but has a major drawback: the search for the optimal cost factors requires label information from both classes ("changed" and "unchanged") and is computationally intensive by solving each time a large optimization problem [3].

In Unsupervised Novelty Detection (UND) only unlabeled samples are available and novelties are assumed not clustered [6] but spread in low density regions in opposition to simple clustering. This happen in situations where changes are from multiple types (urban, vegetation, etc..) with different spectral signatures and can be emphasized with appropriate features spreading out the changes (e.g. image difference). In this case, the detection of changes is a density estimation problem where the novelties are present in the tails of the distribution [7].

This paper introduces to the remote sensing community the Nested SVMs, two algorithms based on solving the entire SVM solution path and having robust properties (enforcing nested boundaries). Finding the optimal classifier requires the tuning of "magic" parameters in both situations (SSND and UND), therefore we propose unspervised data-driven parameter selection methods avoiding the classical unrealistic cross-validation.

2. PROPOSED METHODS

In Fig. 1., the two different situations are illustrated in a 2D example. (a) In SSND situation, the "changed" samples are assumed clustered and separated from "unchanged" samples by a low density region helping the selection of the optimal boundary at a cost asymmetry γ. A set of classifiers is ob-



(a) (b)

Fig. 1. (a) SSND: boundaries for different cost asymmetry γ. (b) UND: boundaries of different density level sets λ.

tained for the entire range of cost asymmetries in a single optimization called the Nested Cost-Sensitive SVM. The optimal cost asymmetry is then selected via a low-density criterion based on samples close to the boundary.

(b) In UND situation, the "changed" samples are assumed spread in low density regions. A set of density levels (indexed by λ) is obtained in a single optimization called the Nested One-Class SVM. The averaged decision function over the different levels allows a density-based ranking of the unlabeled samples and the selection of a threshold value separating "changed" from "unchanged" samples.

2.1. Semi-Supervised Novelty Detection using Nested Cost-Sensitive SVM

The Nested Cost-Sensitive SVM (NCS-SVM) solves a single optimization problem to derive the full set of Cost-Sensitive SVMs [8]. The NCS-SVM constraints the boundaries at different cost asymmetries to be included in each other (nested). This ensures a certain coherence among the different boundaries and more robustness to parameters (less variations with different kernel bandwidth). As a recall, the primal optimization problem of the standard Cost-Sensitive SVM is

$$\min_{w,\xi} \left\{ \frac{\lambda}{2} \|w\|^2 + \gamma_m \sum_{I_+} \xi_i + (1-\gamma_m) \sum_{I_-} \xi_i \right\}$$
$$\text{such that } y_i \langle w, \Phi(\mathbf{x}_i) \rangle \geq 1 - \xi_i,\ \xi_i \geq 0,\ \forall i$$

As it can be seen in Fig. 1. (b), a cost asymmetry of $\gamma_m = 0$ penalizes only unlabeled samples, while $\gamma_m = 1$ penalizes only labeled samples. Finally $\gamma_m = 0.5$ penalizes equally both classes (standard SVM). Let us define M different cost asymmetries $0.5 \leq \gamma_m \leq 1$ and consider the pixel feature vector $\mathbf{x}_i \in \mathbb{R}^d$ with y_i its label corresponding to either the class of labeled "unchanged" samples ($I_+ = \{i : y_i = +1\}$) or the unlabeled samples ($I_- = \{i : y_i = -1\}$). The Lagrangian dual formulation of the NCS-SVM is

$$\min_{\alpha_{i,1},\ldots,\alpha_{i,M}} \sum_{m=1}^{M} \left[\frac{1}{2\lambda} \sum_{i,j} \alpha_{i,m}\alpha_{j,m} y_i y_j K_{i,j} - \sum_i \alpha_{i,m} \right]$$
$$\text{s. t. } 0 \leq \alpha_{i,m} \leq \mathbf{1}_{\{y_i<0\}} + y_i\gamma_m,$$
$$y_i\alpha_{i,1} \leq \ldots \leq y_i\alpha_{i,M}\ \forall i, m \quad (1)$$

with $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j) =< \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_i) >$ the kernel representing the dot product of samples mapped by Φ in a high-dimensional space and λ a global regularization parameter. The constraints of Eq. (1) enforce the boundaries to be nested. Nested solution paths are piecewise linear along the different cost asymmetries γ_m and require very few breakpoints along the path (usually $M \approx 10$ is enough). Intermediate solutions are obtained via linear interpolation of the Lagrangian multipliers $\alpha_{i,m}$.

The predicted label of a pixel $\mathbf{x}$ at a cost asymmetry γ_m is obtained from the sign of the decision function: $f_{\gamma_m}(\mathbf{x}) = \frac{1}{\lambda} \sum_i \alpha_{i,m} y_i k(\mathbf{x}_i, \mathbf{x})$.

The selection of the optimal cost asymmetry γ^* is based on the low-density principle, an extensively used assumption in semi-supervised learning [4]. This assumption means that the boundary of the optimal classifier should not cut a cluster but pass through low density regions only. We propose a low-density criterion based on the samples that are close to the boundary. The average distance among k unique pairs of samples across the boundary will reflect the inverse of the density: a large average distance meaning a low density around the boundary.

The optimal cost asymmetry γ^* passing through low density region is selected as follows

$$\gamma_m^* = \arg\max_{\gamma} \left(\frac{1}{k} \sum_{i=1}^{k} D_{pair}(i, \gamma) \right) \quad (2)$$

where $D_{pair}(i, \gamma)$ is the distance between the i^{th} closest sample to the boundary (on the positive side) and its unique pair at minimum distance across the boundary for the cost asymmetry γ. The closest samples to the boundary are found using the $|f_{\gamma,\lambda}(\mathbf{x}_i)|$ values and paired progressively with closest samples on the other side. A large range of k (e.g. from 10 to 100) is used and the most frequent γ^* obtained is selected. In practice the γ obtained for different k are quite stable.

2.2. Unsupervised Novelty Detection using Nested One-Class SVM

The One-class SVM (OC-SVM) with a Gaussian kernel can lead to an efficient estimation of the support of a distribution.

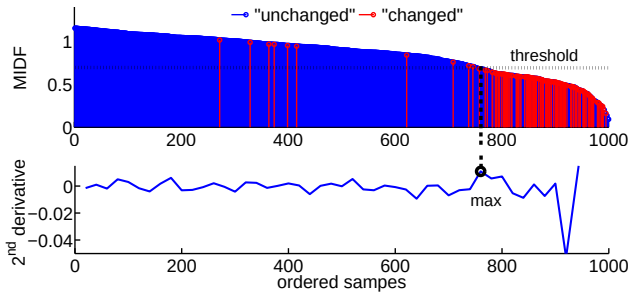


Fig. 2. MIDF and its 2^{nd} derivative for *Gloucester floods* (NDVI) for 1000 random samples. High MIDF values correspond to high density and low MIDF values to low density.

Varying the regularization parameter of the OC-SVM rejects a certain number of training samples and define a particular density level set. The Nested One-Class SVM (NOC-SVM) solves a single optimization problem to derive the entire set of OC-SVMs at different regularization levels [8]. Moreover the NOC-SVM ensures that the boundaries are nested like the true density levels (hierarchically included in each other).

The dual formulation of the NOC-SVM is

$$\min_{\alpha_{i,1},\ldots,\alpha_{i,M}} \sum_{m=1}^{M} \left[\frac{1}{2\lambda_m} \sum_{i,j} \alpha_{i,m}\alpha_{j,m}K_{i,j} - \sum_i \alpha_{i,m} \right]$$
$$\text{s.t. } 0 \leq \alpha_{i,m} \leq \frac{1}{N}, \quad \frac{\alpha_{i,1}}{\lambda_1} \leq \ldots \leq \frac{\alpha_{i,M}}{\lambda_M} \ \forall i,m$$

The decision function at the regularization level λ_m is $f_{\lambda_m}(\mathbf{x}) = \frac{1}{\lambda_m}\sum_i \alpha_{i,m}k(\mathbf{x}_i,\mathbf{x})$. A sample $\mathbf{x}$ is inside the boundary if $f_{\lambda_m}(\mathbf{x}) > 1$, on the boundary if $f_{\lambda_m}(\mathbf{x}) = 1$ and outside otherwise.

As any Gaussian kernel method, the standard deviation σ has to be properly tuned. An unsupervised way of selecting it is using the Minimum Integrated Volume (MIV) criterion [7]. The optimal σ is resulting in the minimum area under the curve, represented by the volume of the different level sets as a function of the percentage of enclosed samples. The volume is obtained through sampling in a box around the data and the percentage of enclosed samples through Cross-Validation. The separation between "unchanged" and "changed" (novelties) samples is based on the ranking of the samples using the Mean Integrated Decision Function (MIDF) [9]: $MIDF(x) = \frac{1}{M}\sum_{m=1}^{M} f_{\lambda_m}(x)$, a novelty will have most of its $f_{\lambda_m} < 1$ and a sample lying in high density region will have most of $f_{\lambda_m} > 1$. The observation of $MIDF(x)$ values in descending order allows to localize the breakpoint separating the "changed" from "unchanged" samples at the maximum of the second derivative of the ordered $MIDF$ (see Fig. 2 for an example of unsupervised breakpoint selection).

3. EXPERIMENTS

Here are presented the results of experiments on two different image datasets: *Gloucester floods* consists in two SPOT images issued from the IEEE GRSS Data Fusion Contest (DFC) in 2009 [10]. The considered subset is $800\times1600px$, has a spatial resolution of $20m$ and 3 spectral bands (NIR-R-G). The images have been acquired before and after the floods. *Bastrop fires* consists in two Landsat 5 TM images acquired before and after large fires in Texas (USA) in 2011. Images are 785×929 pixels with 6 spectral bands (from 450 nm to 2350 nm) at a spatial resolution of 30m. Normalized Difference Vegetation Index (NDVI) features and difference image features (DIFF) have been considered alternatively in the experiments.

The following methods are compared: the NCS-SVM with the unsupervised cost asymmetry γ selection based on low density (NCS-SVM LD) and with the supervised selection through cross-validation (NCS-SVM CV), the NOC-SVM with unsupervised breakpoint selection (NOC-SVM BKP) and with supervised threshold selection (NOC-SVM CV).

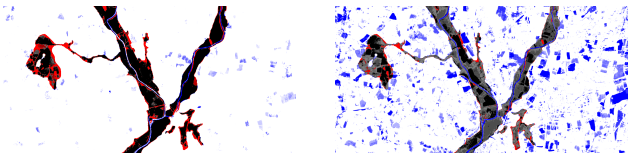
All the results are reported in Fig. 4. The best results for *Gloucester* are obtained with the NDVI features and for *Bastrop* with the DIFF features. The NDVI is in these cases less ambiguous for flooded areas than for burnt areas. For *Gloucester* the DIFF features are loosing too much information resulting in many false detections. The NCS-SVM give accurate results for appropriate features and the unsupervised cost asymmetry selection works very well (NSC-SVM LD has κ lower than CV of max. 0.03). Meaning that the cluster assumption is reasonable for the two datasets. The NOC-SVM performed worse since it is less discriminant (unsupervised and not semi-supervised) and because the assumption of changes spread in low density regions is not really respected. In opposition to our claim the difference features are not providing better results with the NOC-SVM. The unsupervised breakpoint selection (NOC-SVM BKP) results in less accurate and less stable κ accuracies than the supervised upper bound (NOC-SVM CV). The breakpoint separating the two classes is often difficult to localize. The number of samples is a critical issue for the NCS-SVM in order to localize the low density region between clusters but impacts less the NOC-SVM, where it only refines the density level sets. Detection map are presented in Fig. 3. and 5.

4. CONCLUSIONS

We presented two methods for Semi-Supervised Novelty Detection (SSND) and Unsupervised Novelty Detection (UND) based on Nested SVM which solve the entire path of regularization in a single optimization and gains in robustness. We proposed unsupervised data-driven parameter selection for each method. The experiments show the effectiveness of

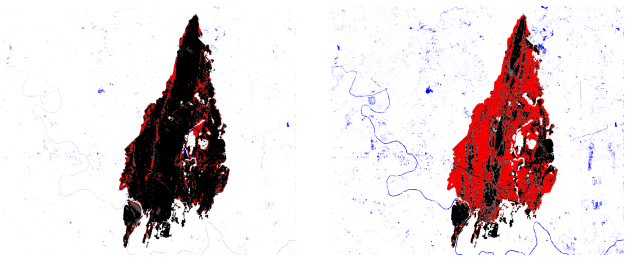
		Gloucester floods				*Bastrop fires*			
		NCS-SVM		NOC-SVM		NCS-SVM		NOC-SVM	
		CV	LD	CV	BKP	CV	LD	CV	BKP
NDVI	κ	0.82 (0.02)	0.82 (0.02)	0.57 (0.04)	0.45 (0.13)	0.88 (0.02)	0.87 (0.03)	0.69 (0.02)	0.66 (0.09)
	OA	96.2 (0.35)	96.2 (0.33)	88.3 (0.98)	86.7 (2.26)	96.73 (0.50)	90.7 (0.48)	90.7 (0.48)	90.2 (1.69)
DIFF	κ	0.63 (0.04)	0.50 (0.09)	0.39 (0.77)	0.26 (0.07)	0.94 (0.01)	0.91 (0.01)	0.63 (0.02)	0.48 (0.09)
	OA	93.1 (0.56)	91.9 (0.94)	86.6 (0.27)	85.6 (1.71)	98.1 (0.19)	97.6 (0.36)	89.5 (0.43)	87.8 (1.06)

Fig. 4. Averaged results over ten random runs. In parenthesis are reported the standard deviation.



(a) NCS-SVM LD (NDVI) (b) NOC-SVM BKP (NDVI)

Fig. 3. *Gloucester*: Averaged results over ten random runs. Black= 100% detected, white=0% detected, red=missed detection, blue=false detection



(a) NCS-SVM LD (DIFF) (b) NOC-SVM BKP (DIFF)

Fig. 5. *Bastrop*: Averaged results over ten random runs. Black= 100% detected, white=0% detected, red=missed detection, blue=false detection

our unsupervised selection of cost-asymmetry for the Nested Cost-Sensitive SVM based on the low density principle, confirming the cluster assumption in most of the cases. This induced that the novelties are more clustered than being spread in low-density regions, resulting in worse performances for the experiments with the Nested One-Class SVM. The NOC-SVM would require situations with changes of very different types. Further perspectives for the NOC-SVM are towards feature representations spreading more the changes and on a more robust unsupervised breakpoint localization.

5. REFERENCES

[1] G. Camps-Valls and L. Bruzzone, *Kernel methods for remote sensing data analysis*, Wiley Online Library, 2009.

[2] A. Banerjee, P. Burlina, and C. Diehl, "A support vector method for anomaly detection in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 8, pp. 2282–2291, 2006.

[3] J. Muñoz-Marí, F. Bovolo, Gómez-Chova, L. Bruzzone, and G. Camp-Valls, "Semisupervised one-class support vector machines for classification of remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 8, pp. 3188 –3197, aug. 2010.

[4] X. Zhu and A.B. Goldberg, *Introduction to semi-supervised learning*, Morgan & Claypool Publishers, 2009.

[5] G. Blanchard, G. Lee, and C. Scott, "Semi-supervised novelty detection," *The Journal of Machine Learning Research*, vol. 9999, pp. 2973–3009, 2010.

[6] F. Bovolo, G. Camps-Valls, and L. Bruzzone, "A support vector domain method for change detection in multitemporal images," *Pattern Recognition Letters*, vol. 31, no. 10, pp. 1148–1154, 2010.

[7] C.D. Scott and E.D. Kolaczyk, "Annotated minimum volume sets for nonparametric anomaly discovery," in *Statistical Signal Processing, 2007. SSP'07. IEEE/SP 14th Workshop on*. IEEE, 2007, pp. 234–238.

[8] G. Lee and C. Scott, "Nested support vector machines," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1648–1660, 2010.

[9] K. Sjöstrand, M.S. Hansen, H.B. Larsson, and R. Larsen, "A path algorithm for the support vector domain description and its application to medical imaging," *Medical image analysis*, vol. 11, no. 5, pp. 417–428, 2007.

[10] N. Longbotham, F. Pacifici, T. Glenn, A. Zare, M. Volpi, D. Tuia, E. Christophe, J. Michel, J. Inglada, J. Chanussot, and Q. Du, "Multi-modal change detection, application to the detection of flooded areas: outcome of the 2009-2010 data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ.*, vol. 5, no. 1, pp. 331–342, 2012.