Semi-Supervised Novelty Detection using SVM entire solution path

Frank de Morsier, *Student Member, IEEE*, Devis Tuia *Member, IEEE*, Maurice Borgeaud, *Senior Member, IEEE*, Volker Gass, Jean-Philippe Thiran, *Senior Member, IEEE*

***Abstract*— Very often, the only reliable information available to perform change detection is the description of some unchanged regions. Since sometimes these regions do not contain all the relevant information to identify their counterpart (the changes), we consider the use of unlabeled data to perform Semi-Supervised Novelty detection (SSND). SSND can be seen as an unbalanced classification problem solved using the Cost-Sensitive Support Vector Machine (CS-SVM), but this requires a heavy parameter search. We propose here to use entire solution path algorithms for the CS-SVM in order to facilitate and accelerate the parameter selection for SSND. Two algorithms are considered and evaluated. The first one is an extension of the CS-SVM algorithm that returns the entire solution path in a single optimization. This way, the optimization of a separate model for each hyperparameter set is avoided. The second forces the solution to be coherent through the solution path, thus producing classification boundaries that are nested (included in each other). We also present a low density criterion for selecting the optimal classification boundaries, thus avoiding the recourse to cross-validation that usually requires information about the "change" class. Experiments are performed on two multitemporal change detection datasets (flood and fire detection). Both algorithms tracing the solution path provide similar performances than the standard CS-SVM while being significantly faster. The low density criterion proposed achieves results that are close to the ones obtained by cross-validation, but without using information about the changes.**

***Index Terms*— Change detection, Learning from Positive and Unlabeled examples (LPUE), Unsupervised parameter selection, Low density separation, Cost-Sensitive SVM, Nested SVM.**

I. Introduction

Change detection in remote sensing [1], [2] is an important field with a wide range of applications, from natural disaster damage assessments to urban expansion monitoring. In most situations, the lack of ground truth information remains the main challenge to overcome. The changes characteristics are often unknown beforehand or difficult to model from a varying spectral signature (e.g. burnt areas, areas on fire, collapsed buildings) [3]. For this reason, many unsupervised change detection algorithms have been proposed in recent literature, ranging from Change Vector Analysis (CVA) [4]–[8] to canonical correlation analysis [9], [10] or clustering [11], [12].

If information about the change is often difficult to obtain, information on the nature of the unchanged areas is easier to have beforehand. Considering these initial conditions (ignorance about the nature of the change and knowledge about some "unchanged" areas), the change detection problem can be reformulated as a Novelty Detection (ND) problem where only samples from "unchanged" areas are available to detect "changed" regions, also referred to as outliers or *novelties*. Since labeling is extremely costly, samples of "unchanged" areas are assumed to be available only in a limited amount.

Novelty detection is a field of machine learning that aims at modeling the distribution of the "unchanged" samples, usually called normal, typical or target samples [13] in order to detect what is abnormal, i.e., the novelties. This field was investigated extensively this last decade and several approaches have been proposed, which are often extensions of standard methods for classification. The Mixture of Gaussian Domain Description Classifier trains a mixture of Gaussians on the target class and sets a threshold level for detecting novelties [13], [14]. The One-Class Support Vector Machine (OC-SVM) proposed by Schölkopf *et al.* maximizes the margin between the data and the origin in a higher dimensional feature space [15], [16]. Similarly, the Support Vector Data Description (SVDD) defines a sphere around the target data in the induced feature space and detects outliers outside its boundary [17].

Novelty detection has been considered in remote sensing classification: after its introduction for tasks of anomaly detection in hyperspectral imagery [18], SVDD has been considered for multiclass classification and detection of outliers [14], [19]. Each class is described by a SVDD and the classification is based on the smallest distance to the different SVDD spheres. This approach also shows the advantage of detecting outliers when a test sample is outside all the different SVDD. In multitemporal analysis, novelty detection approaches have been considered for oil slick detection with SAR images using OC-SVM and wavelet decomposition [20], for landmines detection from Ground-Penetrating Radars using OC-SVM [21] or for fire detection using SVDD initialized with CVA [22].

In remote sensing, unlabeled data are abundant and can be acquired at no extra costs, contrarily to labeled data requiring expert's labeling time or expensive ground surveys. Semi-supervised techniques, exploiting unlabeled data, have shown great improvements for classification methods under

Manuscript received [];
This work has been partly supported by the Swiss National Science Foundation (grant PZ00P2-136827) and RUAG Schweiz AG.

FDM and JPT are with the LTS5 laboratory, École Polytechnique Fédérale de Lausanne, Switzerland. Email: {frank.demorsier, jean-philippe.thiran}@epfl.ch, web: http://lts5www.epfl.ch/ , Phone: +4121 693 26 01, Fax: +4121 693 76 00.

DT is with Laboratoire des Systèmes d'Information Géographique, École Polytechnique Fédérale de Lausanne, Switzerland. Email: devis.tuia@epfl.ch, web: http://lasig.epfl.ch/ , Phone: +4121 693 57 85, Fax: +4121 693 57 90.

MB is with European Space Agency, ESRIN, Frascati, Italy. Email: maurice.borgeaud@esa.int

VG is with the Space Center, École Polytechnique Fédérale de Lausanne, Switzerland. Email: space.center@epfl.ch, web: http://space.epfl.ch/ , Phone: +4121 693 69 48, Fax: +4121 693 69 40.

appropriate assumptions on the data distributions [23]. In remote sensing data classification, semi-supervised learning has driven a strong current of research, where methods exploiting graphs on manifolds [24], [25], low density areas [26], [27], clustering of data [28]–[30] have shown to be strongly beneficial for the classification performance. In [31], semi-supervised change detection is performed using the Semi-Supervised SVM (S3VM), which labels progressively the unlabeled samples from an initial classification requiring both "unchanged" and "changed" labels. Finally in [32], semi-supervised kernel orthogonal space projection is proposed to perform target detection without knowledge of the outlier class.

Semi-Supervised Novelty Detection (SSND) deals with situations having only labeled "unchanged" samples. These labeled "unchanged" pixels are exploited jointly with a large set of unlabeled samples. No information about changes, that lie among the unlabeled data, is available beforehand. In [33], a Cost-Sensitive SVM (CS-SVM) is proposed for text classification using the following principle: the positive samples are classified against all the unlabeled samples with different penalization on the respective errors. The asymmetry in the classes cost allows to penalize more the errors done on the labeled samples and less those on the unlabeled samples (since they contain both "unchanged" and "changed" samples). Fig. 1(b) illustrates this principle. This approach reduces SSND to a binary unbalanced classification problem (labeled vs. unlabeled) and has been proven to be very effective and general, since no assumptions have to be done on the distributions and on the proportion of novelties [34].

The CS-SVM optimization involves three main hyperparameters: a regularization parameter λ, the cost asymmetry between the two classes γ and the kernel parameters (e.g. σ for the Gaussian RBF kernel). The SVM solutions along different values of the two first parameters are piecewise linear [35], [36] contrarily to the third [37]. This allows to trace the entire solution paths along λ or γ at the same computational cost than a single SVM. Recently, a nested solution path of the CS-SVM was proposed to provide a more coherent classification along the solution path [38]. In this model, all the boundaries are included in each other, which means that the predicted class of a sample changes only once along the path of different cost asymmetries. Finally in [39], the entire regularization path (along λ) of the standard SVM was assessed in SSND situations but without exploiting a cost asymmetry.

In remote sensing literature, two SSND approaches were presented and compared in [40]. The first is the standard CS-SVM with labeled and unlabeled data, while the second is the SVDD with a kernel distorted by a graph Laplacian built on the unlabeled samples (S2OCSVM). CS-SVM was very efficient in difficult change detection scenarios (e.g. cloud vs. snow). In [41], the authors proposed a SSND approach aiming at learning the conditional probabilities of the "unchanged" class by training on labeled "unchanged" and unlabeled examples. The retrieved probabilities were finally normalized by a constant factor. This approach showed promising results in remote sensing one-class classification under the assumption that labeled samples are selected completely at random. Note that in such a setting (which is also the one proposed in this paper), the natural unbalance of the novelty detection problem is further increased, since we consider labeled pixels, that are "unchanged", and confront them against unlabeled pixels, that are a mixture of "changed" and (mostly) "unchanged" areas. As a consequence, the unbalance between "changed " and "unchanged" regions is stronger.

The SSND approaches considered in remote sensing so far show three main weaknesses: (i) they are all extremely time consuming, in particular when they reduce to an unbalanced two-classes classification problem which requires a heavy parameter selection to find the optimal cost asymmetry [40], [42]. (ii) Very often, the approaches are based on modeling the changes, thus requiring labeled "changed" samples that are difficult to obtain in sufficient quantity to be representative. (iii) Finally, even the approaches based only on "unchanged" regions (Novelty Detection approaches) select the optimal boundary and the free parameters through cross-validation, thus again using labels from both "unchanged" and "changed" classes.

In this work, we present a methodology avoiding the heavy supervised parameter selection for the CS-SVM in SSND. We first show that algorithms providing the entire solution path for CS-SVM result in faster and coherent classifiers for SSND, being complex unbalanced situations because of the important overlap between labeled and unlabeled samples. To the authors' knowledge experiments exploiting the entire cost asymmetry path for SSND have not been presented previously. Finally, we provide bounds on the regularization parameter to restrict the search space and propose a way to estimate the optimal free parameters without resorting to cross-validation: we remind that this type of optimization should be avoided, since no reliable information on changes is usually available. We assume that the "changed" and "unchanged" distributions are clustered (*cluster assumption*) and select the cost asymmetry, as well as the kernel and regularization parameters, by searching the boundary passing through the low density regions. Exploiting the same intuition behind the TSVM model [26], our proposed low density criterion exploits the distance between pairs of samples across the boundary and is well adapted to unbalanced situations.

The remainder of the paper is as follows: Section II presents the semi-supervised novelty detection systems proposed. Section III presents our low density criterion for unsupervised parameter selection. Section IV details the datasets and the experimental setup of the experiments presented and discussed in Section V. Section VI concludes the paper.

II. Semi-Supervised Novelty Detection

In the following section, we introduce the Cost-Sensitive Support Vector Machine for Semi-Supervised Novelty Detection, with two different algorithms deriving the entire solution path.

A. Cost-Sensitive SVM

Support Vector Machines (SVM) are efficient kernel machines seeking for the hyperplane separating two classes with

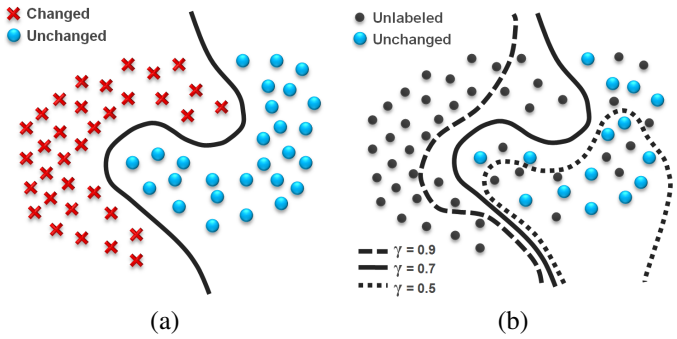


Fig. 1. (a) Supervised two-classes classification ("unchanged" vs. "changed") and (b) Semi-Supervised novelty detection by classifying labeled vs. unlabeled data with different cost asymmetries γ.

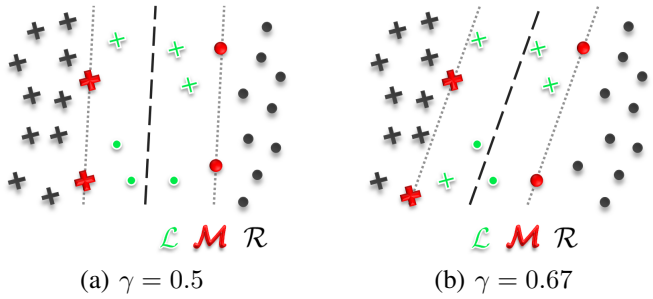


Fig. 2. The classification boundary and the active sets of samples ($\mathcal{L}$, $\mathcal{M}$ and $\mathcal{R}$) for two different cost asymmetries γ.

maximum margin in a high dimensional feature space induced by a mapping function Φ. This mapping function Φ allows having a non-linear separation in the input space by finding the linear separation in the induced feature space. If the classes are not separable, classification errors are allowed but penalized by a cost parameter C, which controls the trade-off between maximum margin and misclassifications. SVMs have been extremely efficient in classification problems in remote sensing [43]. SVMs are trained using a set of samples (multi-temporal pixels) with associated labels $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{+1, -1\}$.

In the standard SVM (Cost-Insensitive SVM), the cost parameter C penalizes equally the errors done on the two classes. This is usually desirable, but in some specific situations, like unbalanced problems, the two classes should be penalized with different strength. The Cost-Sensitive SVM (CS-SVM) has two costs C_+ and C_-, one for each class, combined into the cost asymmetry $\gamma = \frac{C_+}{C_+ + C_-}$ and the total amount of regularization $\lambda = \frac{1}{C_+ + C_-}$ [36].

For SSND, Let $I_+ = \{i : y_i = +1\}$ be the set of labeled "unchanged" samples and $I_- = \{i : y_i = -1\}$ the set of unlabeled samples containing a mixture of the two classes ("unchanged" and "changed"). The CS-SVM optimization problem is

$$\min_{w,\xi} \left\{ \frac{\lambda}{2} \|w\|^2 + \gamma \sum_{i \in I_+} \xi_i + (1-\gamma) \sum_{i \in I_-} \xi_i \right\} \quad (1)$$
$$\text{s.t. } y_i \langle w, \Phi(\mathbf{x}_i) \rangle \geq 1 - \xi_i,\ \xi_i \geq 0,\ \forall i$$

The cost asymmetry γ controls the trade-off between false positive and false negative rates. The entire set of classifiers along the Receiver Operator Curve (ROC) are obtained for γ ranging from 0 to 1. For $\gamma = 0.5$, the algorithm reduces to the standard SVM (Cost-Insensitive). It is important to notice that not all the cost asymmetries are interesting for SSND. Since the errors committed on the labeled "unchanged" samples should be penalized more than the errors done on the unlabeled samples (pixels in I_+ are certainly unchanged, while those in I_- are a mixture of changed and unchanged), γ typically ranges from 0.5 to 1. Therefore, the optimal γ separating the "unchanged" and "changed" samples is related to the balance between the two classes and to the number of labeled and unlabeled samples. An example of boundaries obtained in a SSND setting for different cost asymmetries γ are shown in Fig. 1(b).

The optimization problem in Eq. (1) is usually solved through its Lagrangian dual formulation, given by

$$\min_{\alpha} \left\{ \frac{1}{2\lambda} \sum_{i,j} \alpha_{i,\gamma} \alpha_{j,\gamma} y_i y_j K_{i,j} - \sum_i \alpha_{i,\gamma} \right\} \quad (2)$$
$$\text{s.t. } 0 \leq \alpha_{i,\gamma} \leq \mathbf{1}_{\{y_i < 0\}} + y_i \gamma,\ \forall i$$

where $K_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$ is the kernel representing the dot product in the high-dimensional space induced by Φ. The indicator function $\mathbf{1}_{\{y_i<0\}}$ returns 1 for $y_i = -1$ and 0 otherwise. The dual optimization problem is a quadratic programming (QP) problem. Many algorithms can solve QP problems efficiently by exploiting decomposition methods (e.g. Sequential Minimal Optimization [44]). It is important to notice that the solution of Eq. (2) is obtained for a single and fixed cost asymmetry γ, and thus that it becomes necessary to solve an additional optimization problem for each new cost asymmetry considered.

Similarly to the standard SVM, the class label of a test sample $\mathbf{x}_t$ is obtained from the sign of the decision function:

$$f_{\gamma,\lambda}(\mathbf{x}_t) = \frac{1}{\lambda} \sum_i \alpha^*_{i,\gamma} y_i K_{i,t} \quad (3)$$

where $\alpha^*_{i,\gamma}$ are the support vector coefficients solutions of Eq. (2). This decision function $f_{\gamma,\lambda}(x)$ can be interpreted as a distance to the boundary in the kernel induced space, becoming null for samples lying on it.

B. Entire solution path for Cost-Sensitive SVM

The CS-SVM has different parameters to be tuned: the global regularization parameter λ, the cost asymmetry γ and the kernel parameters. The Lagrangian multipliers $\boldsymbol{\alpha}$ are continuous piecewise linear along the different values of λ and γ [35], [36]. These are called the *solution path* of the CS-SVM ("regularization path" along λ and "cost asymmetry path" along γ). Let us split the samples $\mathbf{x}_i$ into three *active sets*, respecting the convention used in [36]

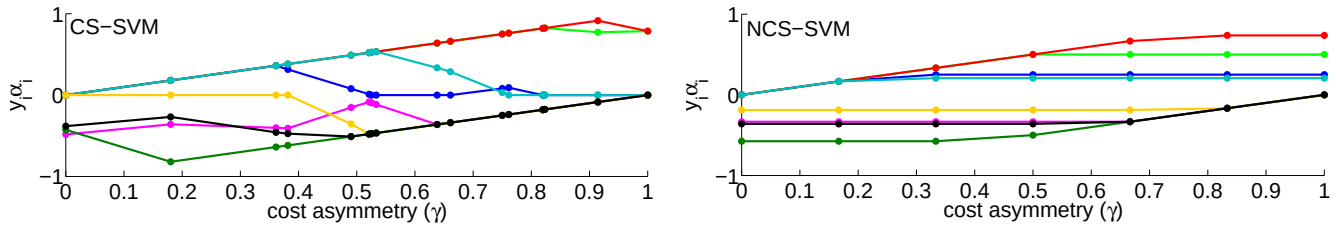


Fig. 3. The piecewise linear solutions α_i along the path of cost asymmetries γ for the CS-SVM (left) and NCS-SVM (right) for 8 samples (toy example of two overlapping classes with 4 samples each). Each color represents a different sample and each dot represents a breakpoint where a solution has been obtained. The number of breakpoints of the CS-SVM is usually around $2 \approx 3$ times the number of samples whereas the M breakpoints of the NCS-SVM are manually preset (here $M = 7$).

Margin:	$\mathcal{M}$	$= \{i, y_i f_{\gamma,\lambda}(\mathbf{x}_i) = 1\}$
Left of margin:	$\mathcal{L}$	$= \{i, y_i f_{\gamma,\lambda}(\mathbf{x}_i) < 1\}$
Right of margin:	$\mathcal{R}$	$= \{i, y_i f_{\gamma,\lambda}(\mathbf{x}_i) > 1\}$

Fig. 2(a)-(b) represents the three active sets around the margin, with the non-null α_i being inside ($i \in \mathcal{L}$) and on the margin ($i \in \mathcal{M}$) at two different cost asymmetries γ. The corresponding α_i for the three different sets are: $\forall i \in \mathcal{M}$: $0 \leq \alpha_{i,\gamma} \leq \mathbf{1}_{\{y_i<0\}} + y_i\gamma$, $\forall i \in \mathcal{L} : \alpha_{i,\gamma} = \mathbf{1}_{\{y_i<0\}} + y_i\gamma$, $\forall i \in \mathcal{R} : \alpha_{i,\gamma} = 0$.

For known active sets $\mathcal{M}$, $\mathcal{L}$ and $\mathcal{R}$, the optimal solutions $\alpha_{i,\gamma}$ can be derived from a linear system of equations. Actually, only the $\alpha_{i,\gamma} \in \mathcal{M}$ are really unknown, since $\alpha_{i,\gamma} \in \mathcal{L} \cup \mathcal{R}$ are either fixed at 0 or $\mathbf{1}_{\{y_i<0\}} + y_i\gamma$. The active sets remain unchanged on a certain range of the parameter resulting in a linear "segment" on the path where the α_i are linearly related to the parameter. The value at which the active sets change produces a breakpoint. The breakpoints can be incrementally computed by tracking the next parameter value for which the KKT conditions are no more satisfied (e.g. the events when samples enter or leave the margin). Fig. 3 represents the piecewise linear $\boldsymbol{\alpha}$ along the cost asymmetry path.

This algorithm (called CS-SVM PATH in the rest of the paper) starts considering the path in the situation where all the samples are inside the margin ($\forall i, i \in \mathcal{L}$, $\mathcal{M} \cup \mathcal{R} \in \emptyset$). In this situation, all the $\alpha_{i,\gamma} = \mathbf{1}_{\{y_i<0\}} + y_i\gamma$. This is achieved by the largest regularization enforcing the maximum margin, meaning a null penalization of the errors: $C_+ + C_- = 0 \Rightarrow \lambda = \infty$. However, this solution is reached already at a certain regularization parameter λ_{max} above which the solutions stop changing: $\forall \lambda \in [\lambda_{max}, \infty]$. This maximum regularization parameter can be obtained directly from the kernel matrix: since all the samples are inside and on the margin: $y_i f_{\gamma,\lambda}(\mathbf{x}_i) \leq 1$, $\forall i$, by replacing $f_{\gamma,\lambda}(\mathbf{x}_i)$ using Eq. (3) and isolating λ, we end up with

$$\lambda_{max} = \max \left[\gamma \sum_{j \in I_+} y_i y_j K_{i,j} + (1-\gamma) \sum_{j \in I_-} y_i y_j K_{i,j} \right] \quad (4)$$

See [38] for more details on the derivation of Eq. (4) and on the implementation of the CS-SVM PATH algorithm based on the *SVMPath* toolbox [35].

C. Nested Cost-Sensitive-SVM

The Nested Cost-Sensitive SVM (NCS-SVM) is another formulation for computing the entire cost asymmetry path of the CS-SVM proposed in [38]. The NCS-SVM forces the boundaries obtained for different cost asymmetries to be nested (see 1(b) and 8(a) for examples of nested boundaries). Let G_{γ_m} be the set of samples in the positive class for the cost asymmetry γ_m. For a $\gamma_k > \gamma_m$, $G_{\gamma_m} \subseteq G_{\gamma_k}$. As depicted in Fig. 1(b), a sample on the positive side of the boundary for a certain γ_m will remain on the positive side for all γ larger than γ_m. This ensures a coherence along the path of cost asymmetries and should provide less sensitivity to the free parameters (regularization and kernel parameters) and to the noise in training data.

Nested solution paths are also continuous piecewise linear function of the cost asymmetry parameter. The nested solution paths are monotonic due to the nestedness constraint, as it can be observed in Fig. 3. The breakpoints along the cost asymmetry path are pre-defined by the user and not derived from the data as in the CS-SVM PATH algorithm (see section II-B). The monotonicity of the path allows pre-defining only a small number of breakpoints. The ability of defining the breakpoints is advantageous for our SSND setting, where only the cost asymmetries ranging between 0.5 and 1 are of interest.

Let us define M different cost asymmetries $0.5 \leq \gamma_m \leq 1$ corresponding to the fixed breakpoints. In this setting, the NCS-SVM dual formulation is

$$\min_{\alpha_{i,1},\ldots,\alpha_{i,M}} \sum_{m=1}^{M} \left[\frac{1}{2\lambda} \sum_{i,j} \alpha_{i,m}\alpha_{j,m} y_i y_j K_{i,j} - \sum_i \alpha_{i,m} \right] \quad (5)$$

$$\text{s.t. } 0 \leq \alpha_{i,m} \leq \mathbf{1}_{\{y_i<0\}} + y_i\gamma_m, \ \forall i, m \quad (6)$$

$$y_i\alpha_{i,1} \leq \ldots \leq y_i\alpha_{i,M} \ \forall i \quad (7)$$

This dual formulation is extremely close to the dual formulation of the CS-SVM in Eq. (2). The differences lie in i) the simultaneous consideration of all the breakpoints and in ii) the additional constraints in Eq. (7) enforcing the nesting of the boundaries. The solutions $\alpha_{i,m+\eta}$ at an intermediate cost asymmetry $\gamma_{m+\eta}$ ($0<\eta<1$) can be obtained from a linear interpolation between the lower and upper solutions at γ_m and γ_{m+1}.

The optimization problem of Eq. (5) is complex since the constraints are imposed on the samples (6) and the cost asymmetries (7) resulting in $M \times N$ variables (N being the number

of labeled samples). A decomposition algorithm is proposed to split the QP problem into smaller subproblems. The sample $\mathbf{x}_k$ violating the KKT conditions the most is selected and its solutions $\alpha_{k,m}$ are optimized along the breakpoints while the other solutions remain fixed. The objective function of Eq. (5) is thus rewritten with $\alpha_{k,m}$ highlighted

$$\sum_{m=1}^{M}\left[\frac{1}{2\lambda}\sum_{i,j}\alpha_{i,m}\alpha_{j,m}y_iy_jK_{i,j}-\sum_i\alpha_{i,m}\right]=$$
$$\sum_{m=1}^{M}\left[\frac{1}{2\lambda}\alpha_{k,m}^2K_{k,k}+\alpha_{k,m}\left(\sum_{j\neq k}\alpha_{j,m}y_jy_kK_{k,j}-\lambda\right)\right]+Cst$$

where Cst contains the terms not related to $\alpha_{k,m}$. Samples are selected iteratively to update their solutions until the optimality condition error is under a predefined value. For more details on the sample selection and termination criterion, please refer to [38].

D. Computational complexity

Solving the QP problem for the standard SVM requires approximately $\mathcal{O}(s^2N+s^3)$ with s the number of samples inside the margin (e.g. support vectors). The complexity is dependent on the regularization parameter λ and end up between $\mathcal{O}(N^2)$ and $\mathcal{O}(N^3)$ for a large and a low regularization respectively [45]. Therefore, if P cost asymmetries γ are considered along the path, the maximal computational complexity would be $\mathcal{O}(PN^3)$.

The CS-SVM PATH algorithm has a complexity of $\mathcal{O}(m^2N+N^2m)$ for an entire solution path, with m the maximum number of samples on the margin $\mathcal{M}$ along the path [35]. The maximum value for m is N leading to a computational cost equivalent to two standard SVM. Therefore the CS-SVM PATH algorithm has already a lower complexity than the standard CS-SVM when more than two cost asymmetries are solved ($P>2$).

The NCS-SVM has a complexity linked to the number of iterations required to converge which is proportional to the number of samples $\mathcal{O}(N)$. The size of the QP subproblem solved at each iteration is proportional to the number of breakpoints $\mathcal{O}(M^2)$. Moreover checking the KKT conditions requires $\mathcal{O}(NM)$, leading to a total complexity of approximately: $\mathcal{O}(M^2N+N^2M)$. Experimentally, $M\approx 10$ and the iterations required are $\approx 5\times N$ for the NCS-SVM [38]. Thus the NCS-SVM will have a lower computational cost than the standard CS-SVM and CS-SVM PATH algorithms for datasets larger than ≈ 50 samples.

III. Unsupervised parameter selection based on cluster assumption

In the SSND settings presented, no assumptions were made on the distribution of the novelties. There is a trade-off between restricting to certain assumptions and performing the parameter selection: either no assumptions are made but then a parameter selection based on cross-validation becomes necessary (thus requiring labeled pixels corresponding to "changed" areas), or an assumption is made on the two distributions and the parameter selection can be done in an unsupervised way.

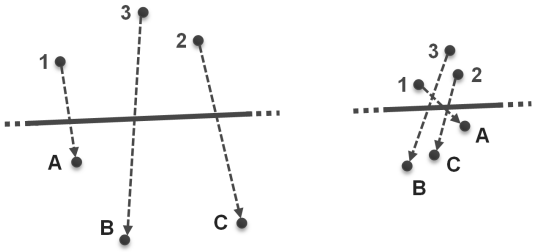


Fig. 4. Low-density criterion principle for $k=3$ for low density (left) and high density (right). 3 unique pairs of samples with minimum distance across the boundary are formed. First the closest sample (1) is paired with its closest sample on the other side (A). Then sample (2) is linked with (C) since the closest sample (A) is already linked. Finally sample (3) is paired with (B) the closest remaining sample.

We propose an approach for selecting the optimal cost asymmetry and the other free parameters based on the *cluster assumption*, an extensively used assumption in semi-supervised learning [23], [46]. This assumption states that the two classes ("changed" and "unchanged") are clustered in the input space. Therefore the boundary of the optimal classifier should not pass through the clusters but in the region of low density between them. The Transductive SVM (TSVM) exploits the cluster assumption by iteratively labeling the unlabeled samples and retraining with the augmented labeled set of samples [47], while the Semi-Supervised SVM (S3VM) penalizes the unlabeled samples lying inside the margin of the SVM directly in its objective function [31], [48].

In our SSND context, unlabeled data are already involved in the training process to help obtaining a better discrimination between the "unchanged" and "changed" classes. The only remaining step is the selection of the optimal solution along the cost asymmetry path. This selection is usually done by cross-validation, training on a subset of the data and testing on a separate subset. It is however impossible to get a reliable accuracy since no labels are available for the "changed" areas. To overcome this issue, the boundary selection can be performed based on the local density around the boundary. The distances between the closest samples across the boundary are inversely related to the density in these regions: the larger the distances, the lower the density. The other parameters influencing the boundary (kernel and regularization parameters) can be selected in the same way by maximizing the distances across the boundary.

Let us define H_{O+} and H_{O-}, the sets of samples ordered by their distance to the boundary ($|f_{\gamma,\lambda}(\mathbf{x}_i)|$) on the positive and negative side respectively. A set of k unique pairs of samples are built across the boundary (each sample is linked only once). The first sample of H_{O+} is linked with the sample at minimum distance from H_{O-}. The second closest sample from H_{O+} is linked with the closest of the $k-1$ remaining samples on the other side of the boundary. This linkage goes on until the k^{th} sample has been linked. This principle is illustrated in Fig. 4. The procedure is then repeated considering the other side of the boundary, and finally provides $2\cdot k$ pairs of samples across the boundary. The set of Euclidean distances between

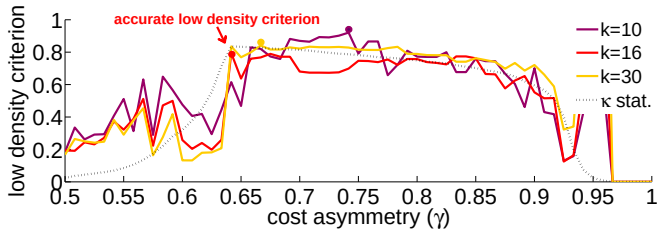


Fig. 5. Low density criterions along the cost asymmetries γ for different number of pairs of samples across the boundary ($k = 10, 16, 30$). The minimal maximum peak indicates the appropriate k for the low density criterion (here $k = 16$), since $k = 30$ adds pairs of samples far from the boundary and $k = 10$ misses samples along the boundary. The low density criterions have been rescaled for visualization purposes with the overlayed κ statistic. The maximum κ is obtained for $\gamma \approx 0.64$, matching with the low density criterion peak at $k = 16$.

the pairs of samples is denoted by $D_{pair}(k, \gamma)$.

The estimated average distance across the boundary is obtained using the median of the paired distances to avoid possible biases related to isolated samples linked with remaining distant samples. The density criterion across the boundary is defined as $DC(k, \gamma) = median\{D_{pair}(k, \gamma)\}$. A small DC indicates that boundary passes through high density regions, while a large DC indicates that boundary passes through low density region. The maximum DC value indicates the optimal boundary passing through lowest density regions . For this reason, the final criterion LDC is defined as $LDC(k) = \max_\gamma(DC(k, \gamma))$.

Even if a robust average is obtained using the median, the choice of the parameter k influences the extent of the density measure (from local to global for increasing k). A too large k could add pairs of samples actually far from the boundary (inside the clusters), whereas a too small k could miss pairs close to the boundary. Both situations could result into an overestimated median distance across the boundary. To avoid such situations, the most accurate distance across the boundary is searched through a range of k values (k_{range}). The value minimizing LDC over k_{range} is retained as the most robust value, k^*. Using this optimal k^*, the boundary passing through the lowest density region is localized at $\gamma^* = \arg\max_\gamma(DC(k^*, \gamma))$. The full procedure is summarized in Algorithm 1.

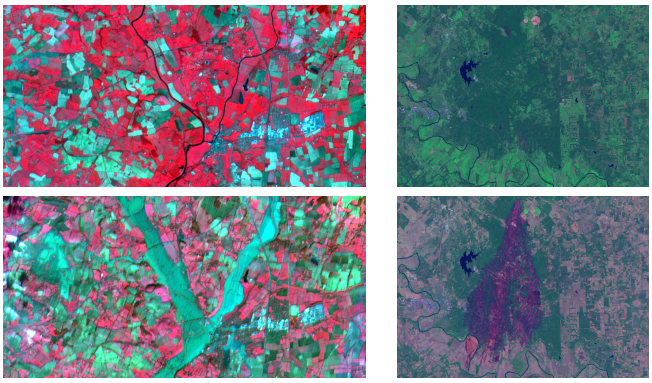
Algorithm 1 Low density boundary selection

Require: (N)CS-SVM solutions at M cost asymmetries γ_m

for each $k \in k_{range}$ **do**
  for each $\gamma \in \gamma_m$, $m = 1, ..., M$ **do**
    Get $H_{O+} = sort(f_{\gamma,\lambda}(\mathbf{x}_i))$ with $\{i, f_{\gamma,\lambda}(\mathbf{x}_i) > 0\}$
    Get $H_{O-} = sort(|f_{\gamma,\lambda}(\mathbf{x}_i)|)$ with $\{i, f_{\gamma,\lambda}(\mathbf{x}_i) < 0\}$
    Build unique pairs of samples (from H_{O+} and H_{O-})
    $D_{pair}(k, \gamma) \leftarrow$ set of distances between the pairs
    $DC(k, \gamma) \leftarrow median\{D_{pair}(k, \gamma)\}$
  end for
  $LDC(k) \leftarrow \max_\gamma(DC(k, \gamma))$ ▷ low density criterion
end for

$k^* = \arg\min_k(LDC(k))$ ▷ optimal k not overestimating

return $\gamma^* \leftarrow \arg\max_\gamma(DC(k^*, \gamma))$ ▷ optimal boundary



(a) Gloucester floods (b) Bastrop fires

Fig. 6. The two multitemporal datasets considered in the experiments.

IV. Experiments

Two multitemporal datasets are used to compare and validate the proposed approaches.

1) *Gloucester floods*: this dataset consists in two SPOT images acquired before and after a flood in 2000 near Gloucester in UK [49]. The considered subset (around the city of Tewkesbury) is of size 800×1600 pixels with a spatial resolution of 20m and 3 spectral bands (NIR-R-G). The two images are presented in Fig. 6(a).
2) *Bastrop fires*: the second dataset consists in two Landsat 5 TM images acquired a year before (03/10/2010) and a week after (11/09/2011) large fires near Bastrop in Texas, USA. The considered subset is of size 780×1085 pixels with 6 spectral bands (from 450 nm to 2350 nm) at a spatial resolution of 30m. A (MIR-G-B) composition representing well the fire is presented in Fig. 6(b). This second image is particularly challenging, since the unchanged areas show radiometric differences related to seasonal vegetation changes. Therefore, methods based on image differencing should return a large amount of false alarms.

A. Experimental Setup

In this work the results for the CS-SVM PATH and NCS-SVM methods are compared with the standard CS-SVM, the SVDD and the unsupervised CVA [8]. Moreover, the proposed methods are evaluated both using cross-validation (CV) and low density (LD) parameter selection criteria.

In a change detection scenario, either the difference or ratio image are often used [50]. Here we consider the difference of the spectral bands of the two images. Another set of features more physically inspired is a stack of the Normalized Difference Vegetation Index ($NDVI = (NIR - R)/(NIR + R)$) obtained separately at each acquisition. It is important in such scenarios to choose the set of features, which is best adapted to the situation [1]. The floods over non-urban areas impact the

vegetation which is reflected by a change of the response of flooded vegetation, which is better visible in the NDVI than in the difference or stacked images. For this reason, the stacked NDVI will be used with the *Gloucester floods* dataset. For the Bastrop dataset, the NDVI features could be ambiguous for regions having dried grass or cut crops. This type of land, being well spread among the unburnt areas, would avoid detecting burnt areas. Therefore the difference image is used in the *Bastrop fires* experiments. The datasets have been centered (zero-mean) and normalized by their standard deviation on each feature independently.

Ground truth have been established for the whole images both by visual interpretation and using documents assessing the range of damages. Some small ambiguous zones ($\approx 3\%$) have been left aside from the *Bastrop fires* ground truth. The training set is composed of $N_{labeled}$ samples randomly selected from the "unchanged" class and of $N_{unlabeled}$ samples randomly selected among all the remaining samples. The number of labeled and unlabeled samples are varied in the range: $[50, 100, 200, 300, 400, 500]$.

The validation set, composed of 10000 samples, allows to select the parameters σ (bandwidth of RBF kernel) and λ (regularization) returning the highest averaged Cohen's kappa statistic (κ) [51] ('CV' hereafter). Our proposed alternative to cross-validation selects the optimal parameters by finding the maximum low density criterion ('LD' hereafter). The average over three different random training and validation sets is used for both CV and LD parameter selection. An initial guess on the bandwidth of the RBF kernel σ_0 is obtained from the median standard deviation among 1000 random "unchanged" samples. The search is then realized over 15 values: $\sigma = [0.1 \times \sigma_0, ..., 1.5 \times \sigma_0]$. The regularization parameter λ, which balances the importance between maximum margin and error penalization, is searched among different values lower than the maximum regularization parameter λ_{max}: $\lambda = [0.01 \times \lambda_{max}, 0.1 \times \lambda_{max}, \lambda_{max}]$.

Finally, Cohen's κ, overall accuracy, F1-score [52], False alarm rate (False changes detected/total "unchanged") and Missed alarm rate (Missed changes/total "changed") are assessed using the best parameters (σ,λ,γ^*,k^*). Ten independent training and validation sets are used to validate the approach on the test set composed of the remaining samples in the image.

For the NCS-SVM, $M = 7$ breakpoints are preselected equally spaced in the range $\gamma_m = [0.5; 1]$, then the solutions are interpolated by a factor of 10 resulting in $7 + 6 \times 9 = 61$ solutions on half of the cost asymmetry path.

The different number k of sample pairs across the boundary for the low density criterion are: $k_{range} = [10, 11, ..., 40]$.

To ensure fair comparison, the standard CS-SVM is trained and evaluated on the same cost asymmetries than the NCS-SVM interpolated breakpoints ($P = 61$). Finally, the SVDD is trained with the same training set than the other methods, but using the labels of the samples considered unlabeled by the other approaches ("changed" samples considered as outliers during training [41]). This allows comparison with the best fully supervised classification achievable using standard novelty detection classifier.

Experiments were designed in the MATLAB environment based on the SVMPath and Nested SVM toolboxes (downloadable at http://www.eecs.umich.edu/∼cscott/code.html)

V. Results

This section reports and discusses the experimental results obtained on the two multitemporal datasets (flood and fire detection) for the different methods presented.

A. Numerical results

Table I summarizes the different statistics for the best parameters averaged over 10 independent experiments (common to all the methods, except the CVA which uses the whole image directly). Fig. 7. presents the averaged change detection maps for the different methods. For CS-SVM PATH and NCS-SVM we report the results obtained by the two strategies for parameters optimization: crossvalidation (CV) and the proposed unsupervised low density criterion (LD).

In general, the SSND methods considered provide accurate detection maps in both datasets.

The SVDD, which is the only fully supervised classifier, provides results with more false detections compared with the standard CS-SVM (see Table I, False Alarm Rate (FAR)). This can be due to the trade-off for the sphere between fitting precisely the distribution (requiring a small σ) and covering an important volume of the feature space (requiring a large σ). Here the sphere is small enough to recover most of the "changed" regions at the price of many other false detection outside the sphere. The SSND approach using the standard CS-SVM is more discriminative and thus performs better than the SVDD in these two change detection scenarios. The best performances of the CS-SVM PATH and NCS-SVM algorithms using CV parameter selection are equivalent to those obtained by the standard CS-SVM, confirming the efficiency of solution path algorithms in SSND situations.

Note that the NCS-SVM, with its interpolated solutions can reach the same accuracy than the CS-SVM PATH.

Concerning the proposed low density criterion (LD), we observe that it provides results not far from those obtained by cross-validation. The NCS-SVM LD solutions have slightly less false alarms but more missed alarms, which results in κ values $\approx 0.02 - 0.05$ lower than those obtained with cross-validation (CV). This demonstrates that the *cluster assumption* holds in these scenarios and that a low density separation exists between the classes. The LD criterion is less stable than the CV resulting in higher κ standard deviation. As it can be observed in Fig. 5, the criterion along the cost asymmetries γ is not smooth, which could result in selecting a suboptimal boundary.

The NCS-SVM LD gives better results than the CS-SVM PATH LD and is more stable. We observed that for cost asymmetries close to 0.5 the nesting of the boundaries could induce boundaries with holes in the middle of the distribution of the "unchanged" class. This phenomenon is illustrated in Fig. 8(a), representing the nested boundaries at different cost asymmetries. Fig. 8(b) illustrates the corresponding nested detection maps. These boundaries with holes in the middle of the distribution have the advantage of being difficultly selected

TABLE I
MEAN (STANDARD DEVIATION) STATISTICS OVER 10 DIFFERENT REALIZATIONS USING 500 PIXELS FOR EACH CLASS.

	Methods		κ	κ diff.	OA	F1-score	FAR	MAR
Gloucester	SVDD		0.738 (0.03)	-0.083	94.40 (1.15)	0.88 (0.03)	4.09 (1.97)	17.58 (6.50)
	CS-SVM		0.821 (0.02)	/	96.58 (0.42)	0.84 (0.02)	1.28 (0.51)	20.27 (3.89)
	CS-SVM PATH	CV	0.808 (0.03)	-0.013	95.98 (0.61)	0.83 (0.02)	1.37 (0.64)	22.30 (1.54)
		LD	0.752 (0.08)	-0.069	94.44 (2.99)	0.78 (0.07)	2.83 (4.35)	24.40 (10.0)
	NCS-SVM	CV	0.819 (0.02)	**-0.002**	96.21 (0.36)	0.84 (0.02)	1.30 (0.31)	20.99 (1.81)
		LD	0.791 (0.04)	-0.030	95.86 (0.65)	0.81 (0.04)	0.76 (0.40)	27.55 (7.02)
	CVA		0.407 (—)	-0.414	80.18 (—)	0.51 (—)	20.0 (—)	18.35 (—)
Bastrop	SVDD		0.819 (0.04)	-0.064	95.785 (1.13)	0.84 (0.04)	2.77 (1.26)	13.85 (4.35)
	CS-SVM		0.883 (0.02)	/	97.412 (0.50)	0.90 (0.02)	1.13 (0.257)	12.30 (3.35)
	CS-SVM PATH	CV	0.897 (0.02)	**+0.017**	97.342 (0.39)	0.94 (0.01)	1.44 (0.336)	9.41 (1.96)
		LD	0.803 (0.10)	-0.080	95.26 (1.79)	0.87 (0.11)	1.91 (0.85)	20.36 (14.9)
	NCS-SVM	CV	0.870 (0.02)	-0.013	96.69 (0.50)	0.89 (0.02)	1.70 (0.48)	12.31 (2.82)
		LD	0.812 (0.07)	-0.071	95.59 (1.38)	0.84 (0.07)	1.06 (0.65)	23.06 (11.5)
	CVA		-0.023 (—)	-0.897	38.44 (—)	0.23 (—)	65.50 (—)	39.80 (—)

by the low density criterion (since part of the boundaries are in the high density region inside the "unchanged" cluster). Therefore nested boundaries resulting in false detections are less likely to be selected than un-nested ones.

The sensitivity of the results with respect to the training set is reflected by the standard deviation of the κ statistics reported in Table I. The random selection of the training samples can induce a different interpretation of the changes to be detected: if a particular land cover is not present among the labeled "unchanged" samples but it is present in the unlabeled data, it will be interpreted as changed (e.g. the circular field at the top right of the *Bastrop fires*). It is important to remind that the maximum training set size considered ($N_{labeled} = N_{unlabeled} = 500$) corresponds to $\approx 0.1\%$ of the available image pixels. Experiments with a higher number of training samples ($N_{labeled} = N_{unlabeled}$ =1000, 2000, 5000 samples) for the NCS-SVM resulted in κ improvement of 0.03 and a false alarm rate slightly lower (under 1%) both for CV and LD. Despite the increase in stability, the missed alarm rate did not change. Visual inspection of the detection maps for 5000 samples in each class (not reported) shows that most of the false detection regions have disappeared where the few remaining could easily be filtered out and that the few regions still miss detected are actually "unchanged" areas mislabeled.

Finally, we compared with state of the art unsupervised CVA method. For the two case studies, the CVA solution has an important false detection rate, particularly for the Bastrop case study. The method, being completely unsupervised, is unable to properly target the changes of interest in presence of radiometric differences between the images or seasonal changes. In *Gloucester*, this corresponds to the presence of wet areas and clouds in the first acquisition, which become challenging "unchanged" regions. For *Bastrop*, the important seasonal changes in the vegetation make the "unchanged" pixels nonzero in the magnitude vector, thus preventing CVA to function correctly. This demonstrates the benefit of SSND, which allows to define typical "unchanged" areas, even if they show radiometric differences.

B. Free parameters sensitivity

Fig. 9 illustrates the optimization surface for the range of λ and σ parameters, as well as for the number of labeled "unchanged" pixels ($N_{labeled}$). For the two algorithms tracing the entire path, larger regularization gives the best results and reduces the sensitivity to the parameters (to the level of the standard CS-SVM). The low density criterion changes rapidly with σ and shows the best performance between 400 and 500 labeled samples. The matching between the maximum of the low density criterion (LD) and the maximum κ from cross-validation (CV) can be well observed in Fig. 9 (bottom row). Therefore, the parameters σ and λ selected by the unsupervised LD criterion are equivalent than using the supervised CV.

The dependence on the number of labeled and unlabeled samples is further studied in Fig. 10. NCS-SVM is slightly less sensitive than CS-SVM PATH when few labeled samples are used

The low density criterion is in comparison more sensitive to the number of unlabeled and labeled samples. A certain amount of unlabeled data is required to observe a cluster of pixels belonging to the "changed" regions. Moreover, the number of labeled samples should be large enough to properly cover the unlabeled samples from the "unchanged" distribution. NCS-SVM being more robust than CS-SVM PATH, it provides more accurate solutions in very unbalanced situations (i.e. $N_{labeled} = 50$ and $N_{unlabeled} = 500$). Experiments with a fixed number of labeled samples ($N_{labeled} = 500$) and a larger number of unlabeled samples ($N_{unlabeled} = 1000, 2000, 5000$) for the NCS-SVM did improve the stability but not significantly the accuracy. The approach is discriminative enough to ensure an accurate detection with a small amount of unlabeled samples. Adding more unlabeled samples allow a better discrimination of the boundary passing through low density regions and enforces the unbalanced situation between

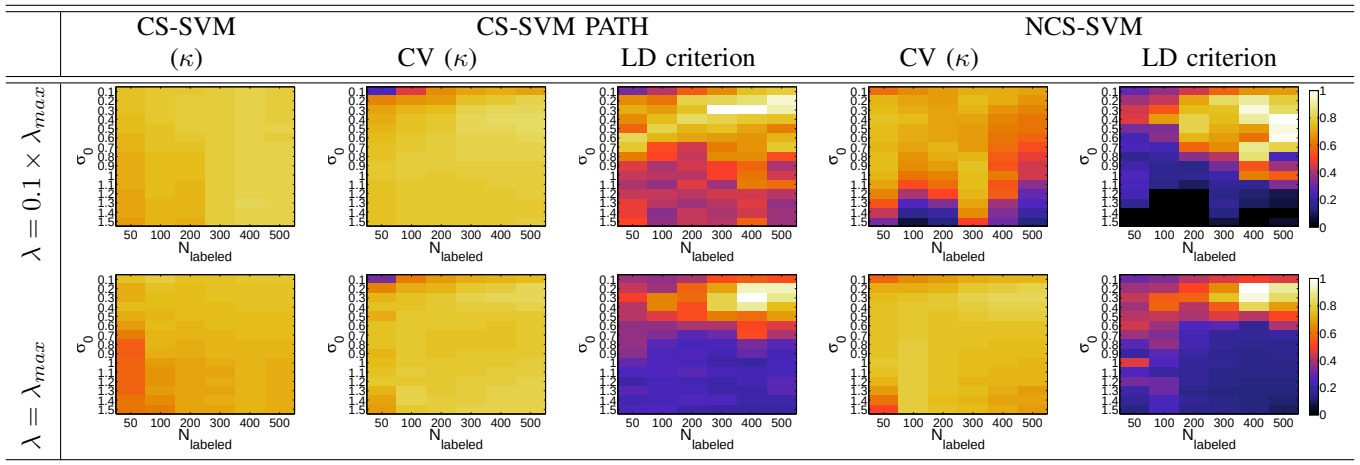


Fig. 9. Averaged κ statistic (CV) and low density criterion normalized (LD) over 3 random runs for the *Gloucester* dataset with 500 unlabeled samples and different number of labeled samples $N_{labeled}$, regularization parameters λ and kernel parameters σ (in terms of σ_0).

				CS-SVM Path		NCS-SVM	
		SVDD	CS-SVM	CV	LD	CV	LD
$N = 100$	Training (s)	11.2 (2.0)	11.2 (2.0)	0.7 (0.3)		9.9 (6.9)	
	Testing (s)	19.3 (0.3)	19.2 (0.3)	14.1 (1.5)	1.2 (0.5)	22.5 (1.5)	0.66 (0.03)
	Total (s)	30.5 (2.3)	30.4 (2.3)	14.8 (1.8)	**1.9 (0.8)**	32.4 (8.4)	10.6 (6.9)
$N = 500$	Training (s)	40000 (17125)	1505 (700)	36.4 (4.6)		95 (51)	
	Testing (s)	89.9 (0.6)	89.5 (0.9)	82.3 (1.6)	31.4 (5.5)	75 (2.9)	1.3 (0.06)
	Total (s)	40090 (17126)	1595 (702)	118 (6)	**67 (10)**	170 (54)	96 (51)

Fig. 11. Runtimes for the different methods with different training set size ($N = N_{labeled} = N_{unlabeled}$) for *Gloucester*.

labeled and unlabeled samples. This is not to be seen as a drawback, since by doing so the proportions of the two classes "unchanged" and "changed" are converging towards their true proportions in the dataset.

C. Algorithms runtime and convergence analysis

In the experiments, the two algorithms tracing the entire solution path converged to accurate solutions with low runtimes. In this section we provide further observations on the training and testing runtimes and convergence of the algorithms.

Fig. 11 reports the training and testing runtimes for different size of the training set. The solution path algorithms are effectively faster for training and grow linearly compared to the standard CS-SVM and SVDD (exploiting both the standard *quadprog* Matlab routine for fair comparisons). The iterative procedure of the NCS-SVM is arbitrarily limited to a maximum of 5000 iterations (corresponding to $5 \times \max(N_{labeled} + N_{unlabeled})$, which is a trade-off between computational cost and accuracy). When setting a too large σ or a too small λ the breakpoints close to $\gamma = 0.5$ can be difficult to obtain: in this case, the algorithm may not converge in the maximum number of iterations allowed.

The testing runtimes are equivalent for the cross-validation (CV) through the different mehtods but much lower using the low density criterion (LD). The CS-SVM PATH LD is slower than the NCS-SVM LD for testing since it evaluates a larger number of breakpoints γ (several times the number of training samples) compared to the 61 interpolated breakpoints of the NCS-SVM. The convergence of the CS-SVM PATH algorithm is guaranteed by the initialization of the algorithm with an arbitrary cost asymmetry in the range $[0.5; 1]$.

VI. Conclusion

In this paper, we presented and evaluated two methods for Semi-Supervised Novelty Detection (SSND). In both cases, SSND is reduced to an unbalanced binary classification between labeled and unlabeled samples, without using any labeled "changed" samples. The methods, based on the Cost-Sensitive Support Vector Machine (CS-SVM), assign different error costs for the two classes (cost asymmetry). The errors done on the unlabeled samples containing both "changed" and "unchanged" samples are less penalized than those committed on labeled "unchanged" samples. The novelty of the proposed methods reside in the retrieval of the solutions for different cost asymmetries in a single optimization.

The CS-SVM PATH algorithm traces the solution path for all the cost asymmetries at approximately the same computational cost than training two standard CS-SVM. This is extremely advantageous, since the optimal cost asymmetry separating the two classes is unknown a-priori and currently proposed approaches require solving an additional CS-SVM for each cost asymmetry tested.

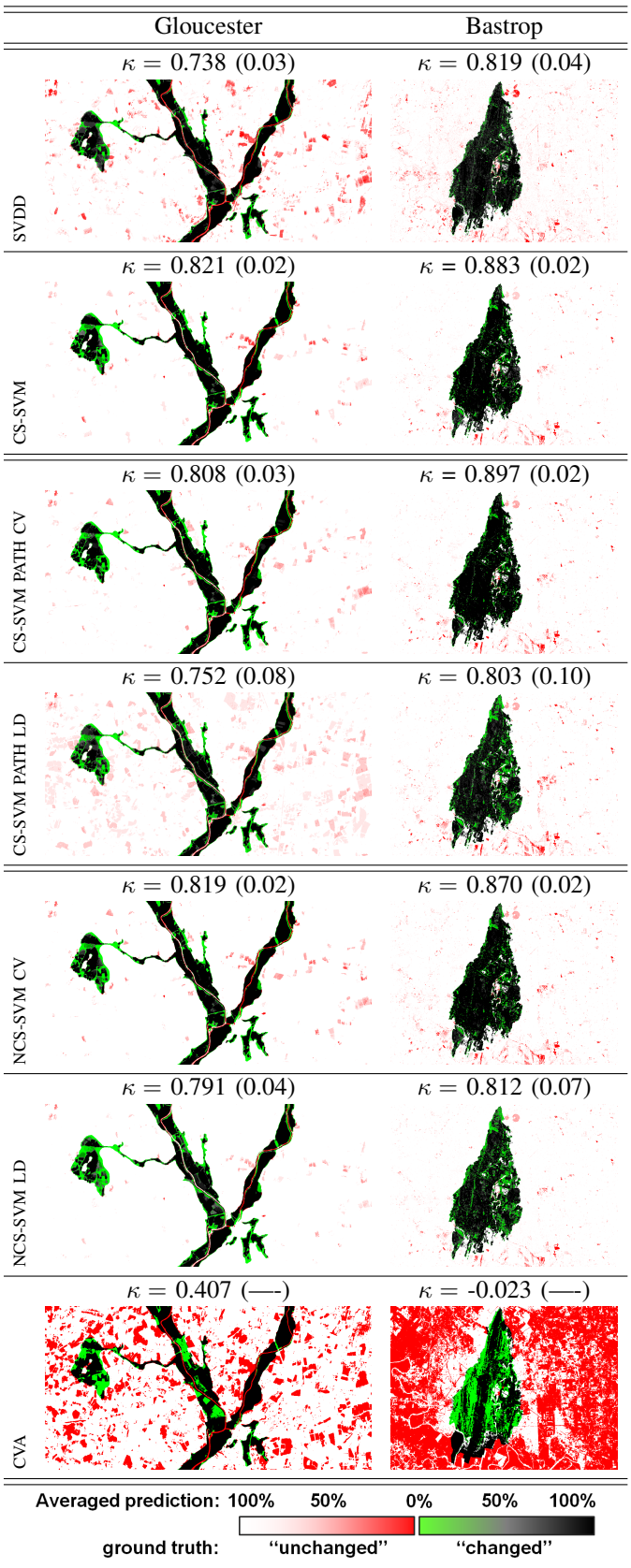
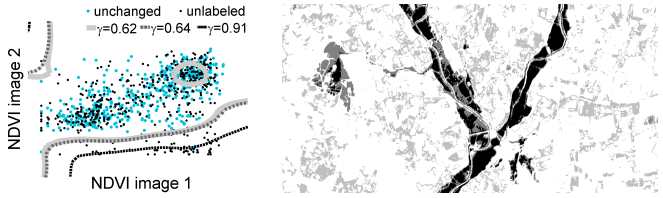


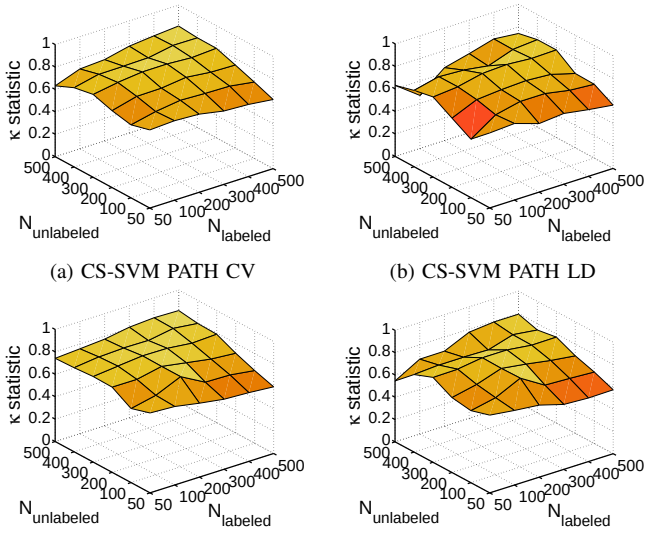
Fig. 7. Mean (standard deviation) statistics over 10 different realizations using 500 labeled and 500 unlabeled pixels.

(Refer to the electronic version for colors: white= 100% detected "unchanged", black=100% detected "changed", red=false detection, green=missed detection and gradient levels for percentages in between).



(a) (b)

Fig. 8. (a) 2D plot of 3 nested boundaries along the path and (b) the overlay of the 3 corresponding detection maps for the *Gloucester (NDVI)* feature set. In (a) a "hole" can be visualized in the boundary at $\gamma = 0.62$.



(a) CS-SVM PATH CV (b) CS-SVM PATH LD

(c) NCS-SVM CV (d) NCS-SVM LD

Fig. 10. Surface of the averaged κ over three random runs as a function of the number of labeled and unlabeled data for the *Gloucester* dataset using: (a) CS-SVM PATH cross-validation, (b) CS-SVM PATH low-density, (c) NCS-SVM cross-validation and (d) NCS-SVM low-density.

The Nested CS-SVM (NCS-SVM) algorithm also derives the entire solution path, but with the additional constraint of nested boundaries (the boundaries are included in each other). This ensures a certain coherence along the path where samples change only once of class.

We also proposed a low density criterion, which allows to select the optimal cost asymmetry and the other free parameters (kernel and regularization) in an unsupervised way. Such criterion estimates the local density along the boundary, based on pairwise distances across the boundary.

The results on two multitemporal change detection scenarios (flood and fire detection) showed the efficiency of these SSND approaches that only exploit "unchanged" information and unlabeled data. The two algorithms deriving the entire solution path performed better than the supervised SVDD and generally equivalently to the CS-SVM (but at a much lower computational cost). NCS-SVM also has the advantage of being less sensitive to the choice of parameters and the size of the training set. Using the low density criterion usually decreases the false alarm rate and slightly the global performance with respect to cross-validation ($\kappa \approx 0,02 - 0.05$ lower). However, we remind that, contrarily to cross-validation, no labeled information about the change is used, and that this is the price to pay to maintain the optimization unsupervised. Nonetheless,

the good results obtained confirmed that the two classes are separated by a low density region and showed the potential for this fully unsupervised method.

Future investigations will focus on the sampling of the training set by considering active learning methods [53]. In this sense, the work of Li et al. [54], where the authors selected unlabeled examples for semi-supervised learning using an active learning criterion, may be of great interest to find the pixels discriminating changed from unchanged areas.

VII. Acknowledgements

The authors would like to acknowledge the EPFL Space Center and RUAG Schweiz AG (particularly C. Küchler) for supporting this research. They would also acknowledge J. Inglada (CNES) for preparing and providing the Gloucester dataset and G. Lee for providing the "Nested SVM" toolbox and his implementation of the non-nested CS-SVM based on the *SVMPath* toolbox. Finally, they acknowledge the anonymous reviewers for helping improving this manuscript.

References

[1] A. Singh, "Digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, 1989.

[2] R.J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: A systematic survey," *IEEE Trans. Im. Proc.*, vol. 14, no. 3, pp. 294–307, 2005.

[3] J.F. Mas, "Monitoring land-cover changes: a comparison of change detection techniques," *Int. J. Remote Sens.*, vol. 20, no. 1, pp. 139–152, 1999.

[4] F. Bovolo and L. Bruzzone, "A split-based approach to unsupervised change detection in large size multitemporal images: application to Tsunami-damage assessment," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1658–1671, 2007.

[5] M. Dalla Mura, J. A. Benediktsson, F. Bovolo, and L. Bruzzone, "An unsupervised technique based on morphological filters for change detection in very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 5(3), pp. 433 – 437, 2008.

[6] F. Bovolo, "A multilevel parcel-based approach to change detection in very high resolution multitemporal images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 1, pp. 33–37, 2009.

[7] J. Im, J. R. Jensen, and M. E. Hodgson, "Optimizing the binary discriminant function in change detection applications," *Remote Sens. Environ.*, vol. 112, no. 6, pp. 2761–2776, 2008.

[8] J. Chen, X. Chen, X. Cui, and J. Chen, "Change vector analysis in posterior probability space: A new method for land cover change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. PP, no. 99, pp. 317 –321, 2010.

[9] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (MAD) and MAF post-processing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sens. Environ.*, vol. 64, pp. 1–19, 1998.

[10] A. A. Nielsen, "The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data," *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 463–478, 2006.

[11] T. Celik, "Multiscale change detection in multitemporal satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 820 –824, oct. 2009.

[12] M. Volpi, D. Tuia, G. Camps-Valls, and M. Kanevski, "Unsupervised change detection with kernels," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 6, pp. 1026–1030, 2012.

[13] M. Markou and S. Singh, "Novelty detection: a reviewpart 1: statistical approaches," *Signal Process.*, vol. 83, no. 12, pp. 2481–2497, 2003.

[14] J. Muñoz-Marí, L. Bruzzone, and G. Camps-Valls, "A support vector domain description approach to supervised classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 8, pp. 2683–2692, 2007.

[15] B. Scholkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," *Advances in Neural Information Processing Systems*, vol. 12, pp. 582–588, 2000.

[16] B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computat.*, vol. 13, no. 7, pp. 1443–1471, 2001.

[17] D.M.J. Tax and R.P.W. Duin, "Support vector data description," *Machine Learn.*, vol. 54, no. 1, pp. 45–66, 2004.

[18] A. Banerjee, P. Burlina, and C. Diehl, "One-class svm for hyperspectral anomaly detection," in *Kernel methods for remote sensing data analysis*, G. Camps-Valls and L. Bruzzone, Eds., pp. 169–192. J. Wiley & Sons, NJ, USA, 2009.

[19] C. Sanchez-Hernandez, D.S. Boyd, and G.M. Foody, "One-class classification for mapping a specific land-cover class: Svdd classification of fenland," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 4, pp. 1061–1073, 2007.

[20] G. Mercier and F. Girard-Ardhuin, "Partially supervised oil-slick detection by sar imagery using kernel expansion," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 10, pp. 2839–2846, 2006.

[21] D. Potin, P. Vanheeghe, E. Duflos, and M. Davy, "An abrupt change detection algorithm for buried landmines localization," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 2, pp. 260–272, 2006.

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

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

[24] G. Camps-Valls, T.V. Bandos Marsheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3044–3054, 2007.

[25] L. Gómez-Chova, G. Camps-Valls, J. Muñoz-Marí, and J. Calpe, "Semi-supervised image classification with Laplacian support vector machines," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 4, pp. 336–340, 2008.

[26] L. Bruzzone, M. Chi, and M. Marconcini, "A novel transductive SVM for semisupervised classification of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3363–3373, 2006.

[27] L. Bruzzone and C. Persello, "A novel context-sensitive semisupervised svm classifier robust to mislabeled training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2142 –2154, july 2009.

[28] D. Tuia and G. Camps-Valls, "Semi-supervised remote sensing image classification with cluster kernels," *IEEE Geosci. Remote Sens. Lett*, vol. 6, no. 2, pp. 224–228, 2009.

[29] L. Gómez-Chova, G. Camps-Valls, L. Bruzzone, and J. Calpe-Maravilla, "Mean map kernel methods for semisupervised cloud classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 1, pp. 207–220, 2010.

[30] D. Tuia and G. Camps-Valls, "Urban image classification with semisupervised multiscale cluster kernels," *IEEE J. Sel. Topics Appl. Earth Observ.*, vol. 4, no. 1, pp. 65–74, 2011.

[31] F. Bovolo, L. Bruzzone, and M. Marconcini, "A novel approach to unsupervised change detection based on a semisupervised SVM and a similarity measure," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 7, pp. 2070–2082, 2008.

[32] L. Capobianco, A. Garzelli, and G. Camps-Valls, "Target detection with semisupervised kernel orthogonal subspace projection," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3822–3833, Jul 2009.

[33] B. Liu, Y. Dai, X. Li, W.S. Lee, and P.S. Yu, "Building text classifiers using positive and unlabeled examples," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*. IEEE, 2003, pp. 179–186.

[34] G. Blanchard, G. Lee, and C. Scott, "Semi-supervised novelty detection," *J. Mach. Learn. Res.*, vol. 9999, pp. 2973–3009, 2010.

[35] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, "The entire regularization path for the support vector machine," *J. Mach. Learn. Res.*, vol. 5, pp. 1391–1415, 2004.

[36] F.R. Bach, D. Heckerman, and E. Horvitz, "Considering cost asymmetry in learning classifiers," *J. Mach. Learn. Res.*, vol. 7, pp. 1713–1741, 2006.

[37] G. Wang, D.Y. Yeung, and F.H. Lochovsky, "A kernel path algorithm for support vector machines," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 951–958.

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

[39] L. Yao, J. Tang, and J. Li, "Entire solution path for support vector machine for positive and unlabeled classification*," *Tsinghua Science & Technology*, vol. 14, no. 2, pp. 242–251, 2009.

[40] 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.

[41] Wenkai Li, Qinghua Guo, and C. Elkan, "A positive and unlabeled learning algorithm for one-class classification of remote-sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 2, pp. 717 –725, feb. 2011.

[42] C. Elkan and K. Noto, "Learning classifiers from only positive and unlabeled data," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. ACM, 2008, pp. 213–220.

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

[44] J.C. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," *Advances in Kernel Methods: Support Vector Learning*, vol. 208, no. MSR-TR-98-14, pp. 1–21, 1998.

[45] A. Bordes, S. Ertekin, J. Weston, and L. Bottou, "Fast kernel classifiers with online and active learning," *J. Mach. Learn. Res.*, vol. 6, pp. 1579–1619, 2005.

[46] O. Chapelle, B. Schölkopf, A. Zien, et al., *Semi-supervised learning*, vol. 2, MIT press Cambridge, MA:, 2006.

[47] V.N. Vapnik, *The nature of statistical learning theory*, Springer Verlag, 2000.

[48] K. Bennett, A. Demiriz, et al., "Semi-supervised support vector machines," *Advances in Neural Information processing systems*, pp. 368–374, 1999.

[49] N. Longbotham, F. Pacifici, T. Glenn, A. Zare, M. Volpi, D. Tuia, E. Christophe, J. Michel, J. Inglada, J. Chanussot, and Q. Du, "Multimodal 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.

[50] G. Camps-Valls, L. Gomez-Chova, J. Muñoz-Mari, L. Alonso, J. Calpe-Maravilla, and J. Moreno, "Multitemporal image classification and change detection with kernels," in *SPIE International Symposium Remote Sensing XII*, 2006, vol. 6365.

[51] G.M. Foody and A. Mathur, "A relative evaluation of multiclass image classification by support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 6, pp. 1335 – 1343, june 2004.

[52] L. Meng and J.P. Kerekes, "Object tracking using high resolution satellite imagery," *IEEE J. Sel. Topics Appl. Earth Observ.*, , no. 99, pp. 1–1, 2012.

[53] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Muñoz-Marí, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 3, pp. 606–617, 2011.

[54] J. Li, J.M. Bioucas-Dias, and A. Plaza, "Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4085 –4098, 2010.