

# Unsupervised Change Detection via Hierarchical Support Vector Clustering

Frank de Morsier, Devis Tuia,  
Volker Gass, Jean-Philippe Thiran  
*EPFL, Lausanne, Switzerland*

Maurice Borgeaud  
*ESA (ESRIN), Frascati, Italy*

## Abstract

*When dealing with change detection problems, information about the nature of the changes is often unavailable. In this paper we propose a solution to perform unsupervised change detection based on nonlinear support vector clustering. We build a series of nested hierarchical support vector clustering descriptions, select the appropriate one using a cluster validity measure and finally merge the clusters into two classes, corresponding to changed and unchanged areas. Experiments on two multispectral datasets confirm the power and appropriateness of the proposed system.*

## 1. Introduction

Change detection in remote sensing has a wide range of applications from natural hazard damage assessments to urban expansion monitoring. The interest for unsupervised methods has always been present since ground truth information is often unavailable or difficult to retrieve in a timely manner after the event.

A problem that often arises is how to perform change detection with complex data, while no information about the structure of the changes is available. One approach is to transform the data in a way that highlights changed areas and then use a thresholding, like Change Vector Analysis (CVA) methods [2]. In [10], the difference image is represented in the feature space through the use of kernel functions. Another way to proceed is to model the support of the data distribution by acting on *regularization* (i.e. rejection of outliers from the support of the distribution). In this work, we model the support of the data using a one class SVM [8] and consider the disconnected regions in the input space to define clusters [1]. By computing the entire regularization path with a nested constraint on the boundaries [5], we retrieve a hierarchical suite of nested clusters.

To find the best clustering in the space of parameters (regularization and kernel bandwidth), we propose

a new cluster validity measure inspired by the one proposed in [3]. Finally, the clusters composing the selected solution are aggregated into two classes corresponding to changed and unchanged regions. To do so, we propose a cluster merging system accounting for cluster sizes, separation among clusters and closeness of outliers among clusters.

## 2 Nested OC-SVM with entire regularization path

The One-Class Support Vector Machine (OC-SVM) was proposed to estimate a level set (usually the support) of a distribution in a high dimensional space induced by a mapping function  $\Phi(\mathbf{x})$  [8]. The OC-SVM separates the data with a hyperplane at maximum distance from the origin. In order to derive the entire regularization path, the usual  $C$  parameter is replaced by a regularization parameter  $\lambda_m$ , where  $m$  is one point on the regularization path [4]. Given a dataset  $\{\mathbf{x}_i\}_{i=1}^N$ , its support is found with the following optimization :

$$\begin{aligned} \min_{w, \xi} \quad & \frac{\lambda_m}{2} \|w\|^2 + \frac{1}{N} \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & \langle w, \Phi(\mathbf{x}_i) \rangle \geq 1 - \xi_i, \xi_i \geq 0, \forall i \end{aligned} \quad (1)$$

Eq. (1) is solved through its Lagrangian dual formulation [4], leading to the function  $f_{\lambda_m}(\mathbf{x}) = \frac{1}{\lambda_m} \sum_i \alpha_{i,m} K(\mathbf{x}_i, \mathbf{x})$ , where  $K(\mathbf{x}_i, \mathbf{x}_j)$  is the kernel representing the dot product in the high-dimensional space induced by  $\Phi(\mathbf{x})$ . This last expression leads to the decision function

$$g_{\lambda_m}(\mathbf{x}) = \text{sign}(f_{\lambda_m}(\mathbf{x}) - 1) \quad (2)$$

which is positive for samples inside the support (inliers) and negative for those outside (outliers). Varying the  $\lambda_m$  parameter allows to retrieve the entire OC-SVM *regularization path*. In [4], the entire path is derived in a single optimization step. In the following, we extend the derivation of the OC-SVM solution along the path with a nestedness constraint [5].

**Nested regularization path.** The boundaries generated along the entire regularization path of the OC-SVM are not included in each other (or *nested*). Since level sets of data are naturally nested, it would make sense to get nested boundaries, thus leading to a true hierarchy among them. Nesting of the boundaries is achieved by adding a constraint to the optimization problem. Assuming  $M$  different level of interest on the path  $\lambda_1 > \lambda_2 > \dots > \lambda_M > 0$ , the optimization problem leading to the family of nested OC-SVM on the path is

$$\begin{aligned} \min_{\alpha_1, \dots, \alpha_M} \sum_{m=1}^M \left[ \frac{1}{2\lambda_m} \sum_{i,j} \alpha_{i,m} \alpha_{j,m} K(\mathbf{x}_i, \mathbf{x}_j) - \sum_i \alpha_{i,m} \right] \\ \text{s.t. } 0 \leq \alpha_{i,m} \leq \frac{1}{N}, \forall i, m \\ \frac{\alpha_{i,1}}{\lambda_1} \leq \frac{\alpha_{i,2}}{\lambda_2} \leq \dots \leq \frac{\alpha_{i,M}}{\lambda_M} \end{aligned} \quad (3)$$

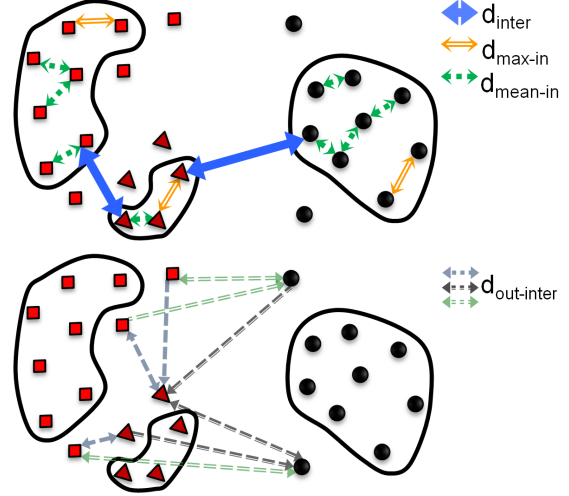
with  $\alpha_{i,m}$  the Lagrangian multiplier corresponding to the sample  $\mathbf{x}_i$  at level  $\lambda_m$ . All solutions between two points  $m$  on the path are obtained by linear interpolation. The optimization problem is iteratively solved by splitting Eq. (3) into subproblems solving for a single sample  $\mathbf{x}$  at each iteration (see [5] for more details).

### 3 Hierarchical Support Vector Clustering

Once the nested OC-SVM solution is retrieved, we have a series of descriptions ( $M$  or more) separating the changed and unchanged areas from the noise with different degrees of regularization. In the input space the boundaries obtained from an OC-SVM enclose several disconnected components. Each of these components is a cluster core. The decision function  $g_{\lambda_m}(\mathbf{x})$  defines if a sample  $\mathbf{x}$  is an inlier or an outlier but does not indicate if two inliers are in the same cluster core or not. The path linking a pair of inlier samples  $\mathbf{x}_i$  and  $\mathbf{x}_j$  from two different clusters in the input space crosses an outlier region ( $g_{\lambda_m}(\mathbf{x}_{\text{onpath}(i,j)}) < 0$ ). Following this intuition, an adjacency matrix  $A_m$  at a specific  $\lambda_m$  can be built by sampling the path between each pairs of inliers and checking  $g_{\lambda_m}$ .

$$A_m(i, j) = \left\{ \begin{array}{ll} 1 & \text{if } g_{\lambda_m}(\mathbf{x}_i + \epsilon(\mathbf{x}_j - \mathbf{x}_i)) \geq 0 \\ 0 & \text{otherwise} \end{array} \right\} \quad (4)$$

with  $\epsilon \in [0; 1]$ . Several methods have been proposed to accelerate the computation of the matrix  $A$ . Graphs such as Delaunay diagram or Minimum Spanning Tree can be built on the data and allow to reduce the number of sample pairs checked [6].



**Figure 1. Distances involved in CVM/CMS.**

The cluster memberships are retrieved from the adjacency matrix  $A_m$  by following the existing links between inliers samples. For outliers, memberships are assigned to the closest cluster using Euclidean distance.

**Clustering validity measure (CVM).** A single solution in the hierarchy should be selected to perform change detection. The quality of a given clustering can be assessed using a measure relating the scatter matrix of samples for each cluster and the scatter matrix between clusters [9, 10]. However this qualifies the global clustering with the outliers and not only the quality of the clustering of the cores. In [3], a cluster validity measure valid for arbitrary cluster shapes tries to maximize the compactness and separability of the cluster cores for a fixed regularization. Consider a distance  $d_{\text{in}}(k)$  describing the core size of cluster  $k$  as the average of the maximal distances among each member of the cluster.

$$d_{\text{in}}(k) = \frac{1}{|C_k|} \sum_{i \in C_k} \max_{j \in C_k} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \quad (5)$$

Moreover, we also define a measure of distance between clusters,  $d_{\text{inter}}(k)$ , as the minimal distance to a member of any other cluster  $l$ .

$$d_{\text{inter}}(k) = \min_{i \in C_k, j \in C_l, k \neq l} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \quad (6)$$

In both Equations,  $C_k$  is the set of inliers in the  $k$ -th cluster. Minimizing the sum of  $d_{\text{in}}$  over the sum of  $d_{\text{inter}}$  works well for a fixed regularization  $\lambda_m$  [3], but does not work when searching for the best kernel parameter  $\sigma$  and regularization  $\lambda_m$  simultaneously. Looking for the most compact and separated clusters would let them

shrink and result in a clustering solution where almost all the samples are considered outliers apart from the cluster centers. In order to take into account the regularization parameter, both the size of the clusters  $d_{in}$  and the distance  $d_{inter}$  between clusters should be maximized. Therefore the cluster validity index will be high for solutions having all samples as inliers apart from the ones being in the overlap region between two clusters. In order to penalize solutions including noisy samples in the cluster core description, we add a ratio  $R_{in}$  between the maximum and the average of the smallest inner distances in each cluster, inspired by [3]. This ratio reflects the difference in density inside a cluster.

$$R_{in}(k) = \frac{d_{\max\text{-in}}(k)}{d_{\text{mean-in}}(k)} \quad (7)$$

$$= \frac{\max\left(\min_{i,j \in C_k, i \neq j} \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)}{\frac{1}{|C_k|} \sum_{j \in C_k} \left(\min_{i \in C_k, i \neq j} \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)}$$

The cluster validity measure (CVM) proposed rescales the cluster sizes by  $R_{in}$  and becomes small for solutions including many outliers.

$$\text{CVM} = \arg \max_{\lambda, \sigma} \sum_{k=1}^M \frac{d_{in}(k)}{R_{in}(k)} \cdot \sum_{k=1}^M d_{inter}(k) \quad (8)$$

**Cluster merging system (CMS).** In change detection analysis, the final number of classes of interest is two: unchanged and changed. The solution chosen by the CVM may show more disconnected clusters belonging to either class. Therefore the clusters found in the data should be merged to end up with two ensemble of clusters representing the two classes. The merging is mainly based on the distance among outliers of different cores.

$$d_{\text{out-inter}}(k, l) = \frac{1}{|O_k|} \sum_{i \in O_k} \min_{j \in O_l, l \neq k} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \quad (9)$$

where  $O_k$  is the set of outliers assigned to the  $k$ -th cluster. The cluster merging score between two clusters is defined as the ratio between the mean minimum distance among the outliers of the two clusters and the mean size of the cluster cores (represented by  $d_{in}(\cdot)$ ). This quantity is then multiplied by the distance between the cores  $d_{inter}(i, j)$ , to favor close clusters to merge. Fundamentally, this system merges neighboring cores having most of their outliers in the space between them.

$$\text{CMS} = \arg \min_{i, j, i \neq j} \left( \frac{d_{\text{out-inter}}(i) + d_{\text{out-inter}}(j)}{\frac{1}{2}(d_{in}(i) + d_{in}(j))} \right) d_{inter}(i, j) \quad (10)$$

The distance between clusters  $d_{inter}$  is evaluated between two arbitrary clusters  $i$  and  $j$ , instead of between  $i$  and its closest neighbor, as in Eq. (6). Clusters  $i$  and  $j$  minimizing this score are merged and the merging is repeated until only two classes remain.

## 4 Experiments

**Setup.** Two sets of multitemporal images are used to validate the unsupervised change detection approach proposed. *Gloucester floods* consists in two SPOT images acquired before and after a flood in 2000 in UK [7]. Images are  $800 \times 1600$  pixels with a spatial resolution of 20m and 3 spectral bands (NIR-R-G). *Bastrop fires* consists in two Landsat 5 TM images acquired before and after large fires in Texas (USA) in 2011. Images are  $785 \times 929$  pixels with 6 spectral bands (from 450 nm to 2350 nm) at a spatial resolution of 30m. Unsupervised clustering is performed on two different feature sets: the difference image (DIFF), and the stack of Normalized Difference Vegetation Indices (NDVI) since the vegetation response can dramatically changes after floods or fires. Comparisons with standard CVA and Kernel  $K$ -Means ( $K = 2$ ) initialized using 500 samples from CVA are performed [10]. The nested support vector clustering uses the CVA initialization only with the DIFF features and randomly selected samples with the NDVI features. The features have been centered (zero-mean) and normalized by their standard deviation independently for the Nested Support Vector Clustering (N-SVC). A RBF kernel is used and its bandwidth  $\sigma$  searched among ten values:  $\sigma = [\sigma_0/20, \dots, \sigma_0/2]$ , with  $\sigma_0$  the mean standard deviation of the samples. The optimal  $\sigma$  and  $\lambda$  are selected using CVM and used in ten experiments having different sets of 500 samples.

**Results and discussion.** The results on the two datasets and feature sets are presented in Figure 2. The  $\sigma$  selected by the CVM is usually  $\approx \sigma_0/10$ , resulting in 2 to 4 clusters. In all experiments, the N-SVC provided better results than the Kernel  $K$ -Means and CVA, both in terms of accuracy and stability across experiments. In the case of *Gloucester* (DIFF), the best results for N-SVC were obtained without the CVA initialization. The CVA initialization biased the initial OC-SVM estimation of the support, as it already rejected the noisy samples that should be detected as outliers. Let us remind that N-SVC finds the cluster cores that contain both the changed and unchanged areas and discriminates them from the outliers corresponding to noisy samples. It is only during the merging phase that the difference between changed and unchanged clusters is done. On the contrary, the  $K$ - $K$ -Means requires CVA to enforce a

		NDVI	DIFF	
Gloucester floods	N-SVC	OA	<b>94.76 (0.39)</b>	<b>92.80 (0.98)</b>
		$\kappa$	<b>0.738 (0.02)</b>	0.597 (0.09)
	K-K-Means	OA	83.65 (8.83)	90.91 (1.86)
		$\kappa$	0.527 (0.14)	<b>0.603 (0.05)</b>
	CVA	OA	80.36 (-)	80.18 (-)
		$\kappa$	0.412 (-)	0.407 (-)
Bastrop fires	N-SVC	OA	<b>91.79 (1.62)</b>	<b>90.64 (0.63)</b>
		$\kappa$	<b>0.656 (0.09)</b>	<b>0.599 (0.04)</b>
	K-K-Means	OA	86.16 (9.18)	82.90 (3.77)
		$\kappa$	0.608 (0.21)	0.269 (0.16)
	CVA	OA	87.66 (-)	75.18 (-)
		$\kappa$	0.646 (-)	0.436 (-)

Averaged prediction: 100% 50% 0% 50% 100%

ground truth: "unchanged" "changed"

**Figure 2. Averaged (standard deviation) results over 10 different runs.**

two-cluster structure, since the algorithm separates the dataset in two clusters from the beginning and is sensible to the presence of outliers. The N-SVC results in less false detections (red tones in Figure 2), thus giving a higher confidence in the detected changes. The nest-

edness of the N-SVC solutions actually helps finding solutions centered around the true cluster cores.

## 5. Conclusions

We presented an unsupervised change detection method based on a nested hierarchical support vector clustering with an appropriate cluster validity measure for parameter selection and a cluster merging system to end up with two main classes (changed and unchanged samples). Two multitemporal change detection experiments demonstrate the accuracy and stability of the proposed method compared with the kernel  $K$ -Means and standard CVA methods. Our method giving less false detections benefits from its intrinsic regularization and nested boundaries.

## References

- [1] A. Ben-Hur, D. Horn, H. Siegelmann, and V. Vapnik. Support vector clustering. *J. Mach. Learn. Res.*, 2:125–137, 2002.
- [2] F. Bovolo and L. Bruzzone. A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain. *IEEE Trans. Geosci. Remote Sens.*, 45(1):218–236, 2007.
- [3] W. Jeen-Shing and C. Jen-Chieh. A cluster validity measure with outlier detection for support vector clustering. *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, 38(1):78–89, 2008.
- [4] G. Lee and C. Scott. The one class support vector machine solution path. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, volume 2, pages II–521. IEEE, 2007.
- [5] G. Lee and C. Scott. Nested support vector machines. *IEEE Trans. Signal Process.*, 58(3):1648–1660, 2010.
- [6] J. Lee and D. Lee. An improved cluster labeling method for support vector clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(3):461–464, 2005.
- [7] 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.*, 5(1):331–342, 2012.
- [8] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7):1443–1471, 2001.
- [9] K. Sjöstrand, M. Hansen, H. Larsson, and R. Larsen. A path algorithm for the support vector domain description and its application to medical imaging. *Med. Image Anal.*, 11(5):417–428, 2007.
- [10] M. Volpi, D. Tuia, G. Camps-Valls, and M. Kanevski. Unsupervised change detection with kernels. *IEEE Geosci. Remote Sens. Lett.*, 9(6):1026–1030, 2012.