Robust Pseudo-hierarchical Support Vector Clustering

Michael Sass Hansen[1], Karl Sjöstrand[1], Hildur Ólafsdóttir[1],
Henrik B.W. Larsson[2], Mikkel B. Stegmann[3], and Rasmus Larsen[1]

[1] Informatics and Mathematical Modelling, Technical University of Denmark,
Lyngby, Denmark
[2] Hospital of Glostrup, Glostrup, Denmark
[3] 3shape A/S, Copenhagen, Denmark

Abstract. Support vector clustering (SVC) has proven an efficient algorithm for clustering of noisy and high-dimensional data sets, with applications within many fields of research. An inherent problem, however, has been setting the parameters of the SVC algorithm. Using the recent emergence of a method for calculating the entire regularization path of the support vector domain description, we propose a fast method for robust pseudo-hierarchical support vector clustering (HSVC). The method is demonstrated to work well on generated data, as well as for detecting ischemic segments from multidimensional myocardial perfusion magnetic resonance imaging data, giving robust results while drastically reducing the need for parameter estimation.

1 Introduction

Support Vector Clustering (SVC) was introduced by Ben-Hur et al. [1]. SVC uses the one-class Support Vector Domain Description (SVDD) as the basis of the clustering algorithm. SVDD was introduced by Tax and Duin [2] in 1999, and it is often calculated with a Gaussian kernel replacing the Euclidian inner product. The SVDD description maps the points into a high dimensional feature space dividing inliers from outliers, where the decision boundary consists of contours enclosing clusters of the data points.

The clustering is done with no assumption on the number of clusters or the shape of the clusters. Ben-Hur et. al. proposed to vary the parameters of the SVDD in a manner that increases the number of clusters while keeping the number of outliers and bounded support vectors (BSV) low. Strictly hierarchical support vector clustering was presented by Ben-Hur in [3]. This algorithm applies SVC subsequently on subsets of the data contained in clusters, and thus achieves a hierarchy of clusters. The clustering, however, depends on the initial steps of the division process.

Yang et al. have proposed improvements to the cluster labelling using proximity graph modelling [4], similar to that of the presented method.

Recently Sjöstrand and Larsen showed that the entire regularization path of the SVDD can be calculated efficiently [5]. This result is the backbone of the

B.K. Ersbøll and K.S. Pedersen (Eds.): SCIA 2007, LNCS 4522, pp. 808–817, 2007.

presented method, and allows for a robust pseudo-hierarchical support vector clustering (HSVC). Given a scale parameter of the Gaussian kernel, a clustering can be estimated efficiently for all values of the regularization parameter. From this ensemble of clusterings a more robust clustering estimate is calculated. To validate the method, the clustering was tested on both artificially generated data, and a real work example of a high dimensional clustering problem.

2 Methods

As other SVC algorithms the basis of the current algorithm is the one-class support vector classification. The recently emerged method for an efficient calculation of the entire regularization path of the SSVD is described briefly for completeness. It is shown that between events the discrimination function varies monotonically, and it is concluded that the description is complete.

2.1 Support Vector Domain Description

The support vector domain description was presented by Tax and Duin[2], posing it as a quadratic optimization problem for a fixed value of the regularization parameter. The criterion to be maximized, given a point set x_i, can be formulated as

$$\min_{R^2,\ \mathbf{a},\ \xi_i} \sum_i \xi_i + \lambda R^2 \quad , \qquad \text{Subject to}$$

$$(\mathbf{x}_i - \mathbf{a})(\mathbf{x}_i - \mathbf{a})^T \leq R^2 + \xi_i \quad \text{and} \quad \xi_i \geq 0 \ \ \forall\ i,$$

where the general idea is to find the minimal sphere that encapsulates the points, allowing some points to be outside the sphere. The regularization parameter λ penalizes the radius R^2 and for large values of λ the radius will tend to be smaller and vice versa. Some points, the outliers, are allowed to be outside the sphere, and the number of outliers is governed by the regularization parameter λ.

Using Lagrange multipliers this optimization problem can be restated as

$$\max_{\alpha_i} \sum_i \alpha_i \mathbf{x}_i \mathbf{x}_i^T - \frac{1}{\lambda} \sum_i \sum_j \alpha_i \alpha_j \mathbf{x}_i \mathbf{x}_j^T,$$

$$0 \leq \alpha_i \leq 1, \ \sum_i \alpha_i = \lambda, \qquad (1)$$

where α_i are the Lagrange mulitpliers and as a consequence of the Karush-Kuhn-Tucker complimentary conditions is that for inliers $\alpha_i = 0$ and for outliers $\alpha_i = 1$. The dimensionality can be increased using a basis expansion and substituting the dot-product with an inner product, the inner products can be replaced by $K_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$, where K is some suitable kernel function. In the presented work the Gaussian kernel $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\gamma}}$ was used as a kernel function. The optimization problem is then given by

$$\max_{\alpha_i} \sum_i \alpha_i K_{i,i} - \frac{1}{\lambda} \sum_o \sum_j \alpha_i \alpha_j K_{i,j}$$
$$0 \leq \alpha_i \leq 1, \ \sum_i \alpha_i = \lambda. \tag{2}$$

For a given λ the squared distance from the center of the sphere to a point $\mathbf{x}$ is

$$f(\mathbf{x}; \lambda) = \|h(\mathbf{x}) - \mathbf{a}\|^2 = K(\mathbf{x}, \mathbf{x})$$
$$-\frac{2}{\lambda} \sum_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + \frac{1}{\lambda^2} \sum_i \sum_j \alpha_i \alpha_j K_{i,j} \tag{3}$$

The entire regularization path of the SVDD. Sjöstrand and Larsen have shown that the entire regularization path of the parameter λ can be calculated with approximately the same complexity as required for solving the initial optimization problem, posed by Tax and Duin [5]. This is because the regularization path of the parameters α_i is piecewise linear. This can be realized by examining the distance functions of two points on the boundary.

$$f(\mathbf{x}_h; \lambda) = f(\mathbf{x}_k; \lambda), \quad h, k \in B \tag{4}$$

where B is the set of points on the boundary. Formulating this equation for different points on the boundary and using the constraint of the sum of α_i gives a complete set of equations for estimating all the α_i. Let $\boldsymbol{\alpha}$ be a vector with the values α_i and let $\boldsymbol{p}$ and $\boldsymbol{q}$ be the slope and intersection respectively, then (refer to [5] for a detailed derivation)

$$\boldsymbol{\alpha} = \lambda \boldsymbol{p} + \boldsymbol{q}, \tag{5}$$

where $\boldsymbol{p}$ and $\boldsymbol{q}$ are constant on intervals $[\lambda_l; \lambda_{l+1}[$, which are defined as intervals between events where a point either leaves or joins the boundary. The division in inliers and outliers is illustrated in Figure 1.

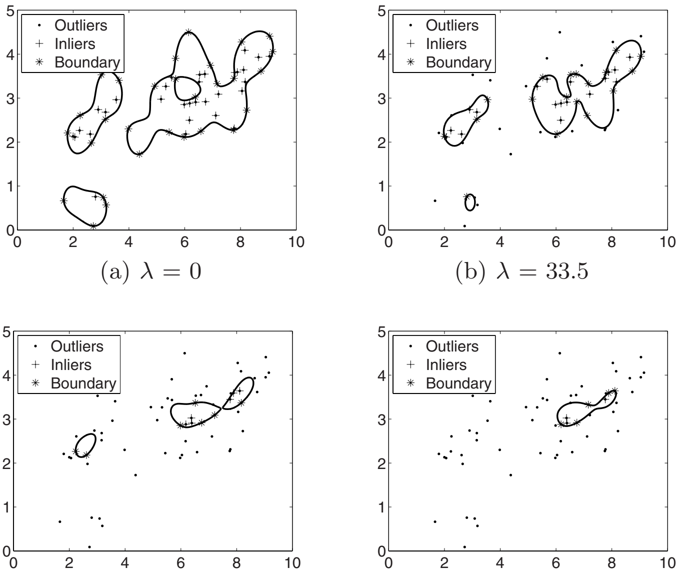
2.2 Support Vector Clustering

The SVDD yields an explicit expression for the distance given by Eq. (3). Now R can be calculated by

$$R = f(\mathbf{x}_k; \lambda) = K_{k,k} - \frac{2}{\lambda} \sum_i \alpha_i K_{k,i} + \frac{1}{\lambda^2} \sum_i \sum_j \alpha_i \alpha_j K_{i,j}.$$

Consider an arbitrary point $\boldsymbol{x}$, and define the distance function $g(\boldsymbol{x}, \lambda)$, as the distance to the boundary.

$$g(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}, \lambda) - R \quad . \tag{6}$$



(a) $\lambda = 0$

(b) $\lambda = 33.5$

(c) $\lambda = 48.3$

(d) $\lambda = 55.0$

Fig. 1. SVDD calculated for the entire regularization path. The line marks the boundary between inliers and outliers, the generalized circle.

The function g is the decision criteria determining if a point is an inlier or an outlier. In Figure 1 the discriminating function g is calculated to create the contour dividing inliers from outliers. Though the optimization problem is to find a circle in the space of the expanded basis, the result appears very little like a circle in the input-space, which in this case has two dimensions. The different enclosed areas could be considered as clusters, denoted support vector clusters.

Assigning clusters. While evaluating $g(\boldsymbol{x}, \lambda)$ reveals if $\boldsymbol{x}$ is an inlier or outlier, it does not contain any specific information on the assignment of clusters. Inspired from Figure 1 it is observed that all paths connecting two points in two different clusters have some points outside the clusters. The current implementation uses an adjacency matrix to determine which points are connected, and which are not. The connection graph is sparsely built, similar to the approach chosen by Yang et. al. [4].

$$A_{ij} = \left\{ \begin{array}{ll} 1\,, & \text{if } g(\mathbf{x}_i + \mu(\mathbf{x}_j - \mathbf{x}_i)) < 0 \;\; \forall \;\; \mu \in [0;1] \\ 0\,, & \text{else} \end{array} \right\}, \tag{7}$$

Connected clusters are detected from the adjacency matrix by using standard graph theory concepts. Outliers are by definition not adjacent to any points, but are assigned to the closest detected cluster.

2.3 Regularized SVC Based on the Entire Regularization Path

Given a λ the clustering can be determined from the adjacency matrix (7), but λ on the interval [0;n] gives rise to changes in the distance function, and thus potentially the clustering. In Section 2.3 it is shown that the distance function (3) is monotonic in the interval $[\lambda_l; \lambda_{l+1}[$ between two events, which means that an almost complete description is obtained by detecting the clusters in the points of the events.

Completeness of the hierarchical description. To ensure that the complete description of the clustering path has been obtained, the distance function is analyzed as a function of the regularization parameter λ.

$$\begin{aligned} g(\boldsymbol{x}, \lambda) &= f(\boldsymbol{x}, \lambda) - R = f(\boldsymbol{x}, \lambda) - f(\boldsymbol{x}_k, \lambda)\ ,\ k \in B, \\ &= K(\boldsymbol{x}, \boldsymbol{x}) - K_{k,k} - \frac{2}{\lambda} \sum_i \alpha_i (K(\boldsymbol{x}, \boldsymbol{x}_i) - K_{k,i}). \end{aligned} \tag{8}$$

Equation (5) states the linear relation between α and λ is given by $\boldsymbol{\alpha} = \lambda \boldsymbol{p} + \boldsymbol{q}$. Let each Lagrange multiplier be given by $\alpha_i = \lambda p_i + q_i$, and the derivative $\frac{\delta g}{\delta \lambda}$ can be calculated as

$$\begin{aligned} \frac{\delta g}{\delta \lambda} &= \frac{\delta}{\delta \lambda} \left[-2 \sum_i (p_i + \frac{q_i}{\lambda})(K(\boldsymbol{x}, \boldsymbol{x}_i) - K_{k,i}) \right] \\ &= \frac{2}{\lambda^2} \sum_i q_i (K(\boldsymbol{x}, \boldsymbol{x}_i) - K_{k,i}), \qquad \lambda \in]\lambda_l; \lambda_{l+1}[\ . \end{aligned} \tag{9}$$

The only dependence on λ in Eq. (9) is on a (inverse squared) multiplicative term. From this, it is concluded that $g(\boldsymbol{x}, \lambda)$ can only change sign once on the interval $[\lambda_l; \lambda_{l+1}]$, so all changes in the clustering are observed in the clustering calculated at every event.

2.4 Pseudo-hierarchical Support Vector Clustering

The calculated clusters are often only changing slowly with changes in the regularization parameter λ. When an event consists of a point leaving the boundary to become an outlier, this does not necessarily alter the boundary much elsewhere. Since the point is still close to the same cluster, and may be associated with this, still, many clusters are close to identical. The similarity can be observed in Figure 1. Moreover, the same clusters may appear again for a different value of the regularization parameter.

The idea presented in this paper, is to collect all the similar clusterings, and build a hierarchy of clusters, which can be thought of as being composed of other, and obviously bigger, clusters. The toy example illustrated in Figure 1 only has a few clusterings that are actually different, and there is a strong relation between the different clusterings of the data, which is illustrated in Figure 2.

There is, however, no guarantee that different clusters, calculated for different values of the regularization parameter, are nested in a strict hierarchical way. In

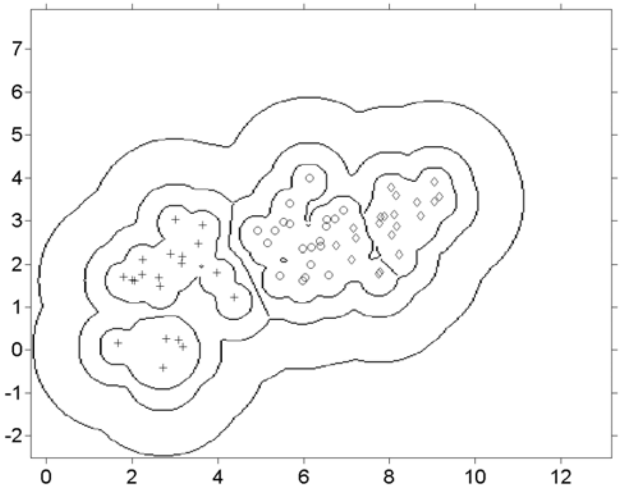


Fig. 2. Hierarchical clustering: From coarse to detailed description

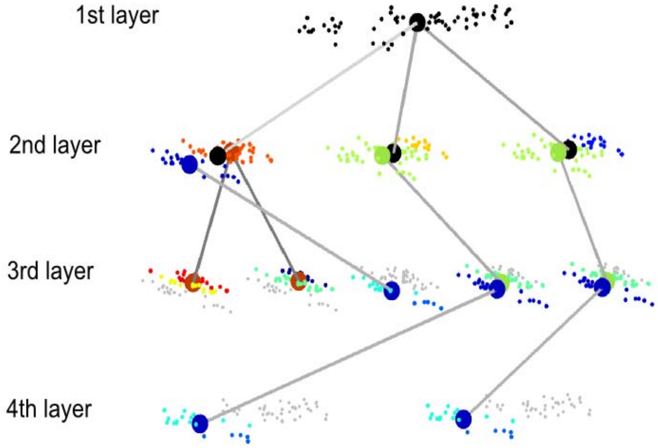


Fig. 3. Competing hierarchical clusterings, obtained from the entire regularization path of the SVDD. The lines show how a cluster is split into smaller clusters. The light gray pixels are the ones not included to describe the subclustering of the cluster. The gray and colored points form together the whole reference data set illustrated in Figure 1.

fact multiple different hierarchical clustering may be proposed. This is illustrated in Figure 3. Each branch of these different cluster representations demonstrate two or more ways, the cluster could be split in smaller clusters. For each cluster, it is known for which intervals of the regularization parameter, the cluster is present. Also it is possible to record if the points forming the cluster are inliers or outliers.

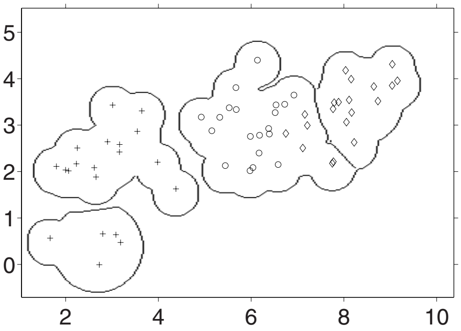


Fig. 4. Clustering of the reference data set by HSVC using the cluster discrimination feature, based on a generalized within covariance matrix

Quality measure of competing clusterings. The analysis described in the previous sections results in a number of competing cluster representations of the data. This analysis, however, does not directly indicate which clustering is the preferred one. We propose a scheme similar to using the 'within' and the 'between' covariance matrices, trace($S_W^{-1} S_B$). Instead of S_W we argue that a weighted within matrix S_W^* should be calculated, weighted by the length of the interval where a given point is an inlier, or an outlier associated with the cluster.

$$S_W^* = \sum_{j=1}^{n_{\text{clusters}}} \sum_{i \in \mathbf{C}_j} \frac{1}{\Pi_j} (\mathbf{x}_i - \mu_i)^T \Lambda_{j,i} (\mathbf{x}_i - \mu_i) \ , \tag{10}$$

where Λ_j defines the weighting of the point, which depends linearly on the length of the interval of λ where the point is an inlier and where it is an outlier. Π_j is a normalization constant. A potential clustering can now be assessed using the measure trace(${S_W^*}^{-1} S_B$), which evaluates the variance within clusters, compared to the introduced distance between clusters. In Figure 4 this is done for the same generated data that was used in Figures 1 and 2. The reference data is actually generated from three random independent distributions, generated as mixtures of Gaussian and uniform distributions. The three different sets are marked by the symbols '+', 'o' and '$\Diamond$' respectively. It can be observed that the clusters 'o' and '$\Diamond$' overlap to some extent, whereas '+' seems more separated from the other groups, and is split in two parts. In Figure 1 small values of λ, corresponding to a high confidence in the data, results in a separation of the two parts of the '+' cluster, whereas the other groups are merged into one cluster. This is opposite for high values of the regularization parameter, where the smaller clusters only appear to be outliers, but the two overlapping clusters are divided. The discrimination feature removes the need to select one value of λ, and appears to adapt to clusters of different variance. The criterion for accepting a subclustering is introduced as a threshold on the cluster separation, given by trace(${S_W^*}^{-1} S_B$). The lower the threshold, the more clusters are accepted.

2.5 Complexity

The complexity of the algorithm is vastly reduced by calculating the entire regularization path of the SVDD in an efficient sequential way, as described. The complexity for the referenced algorithm is $O(n^2)$ for each step between two events. For each event, the clusters are detected from the adjacency matrix, which can also be calculated with a complexity of the order of $O(n^2)$. Comparing with other clusters is done with complexity $O(n \cdot n_{clusters})$. Since the number of events is typically in the vicinity of 3-5 n the overall complexity is polynomial with a degree around 3. On the tested example, with about 500 points in 50 dimensions the algorithm took minutes.

3 Example Application: Detection of Ischemic Segments

To test the capability of the presented clustering algorithm, it has been applied to detect ischemic segments from perfusion MR images. In Figure 5 selected frames from a registered sequence of perfusion MR images of the myocardium are shown. The segmentation was performed previously, with satisfying results [6]. Intensity curves can be obtained pixel-wise from the intensity images, because of the pixel-wise correspondence. Previously ischemic segments have usually been detected using the measures *time-to-peak*, *maximum-upslope* and *peak value* [7]. In a previous study we showed that a generalized version of the distances obtained in the SVDD description corresponded well to the usual measures [8]. In Figure 6(a) the measures are illustrated. The developed HSVC method was applied on the data, which consisted of little less than 500 pixels, and 50 time steps were available for the intensity curve. HSVC divided data in a very few clusters, and in Figure6(b) the curves belonging to each cluster is colored in distinct colors.

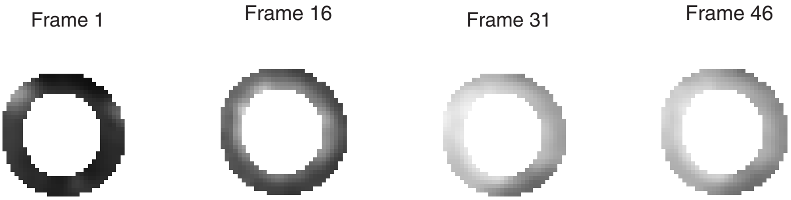
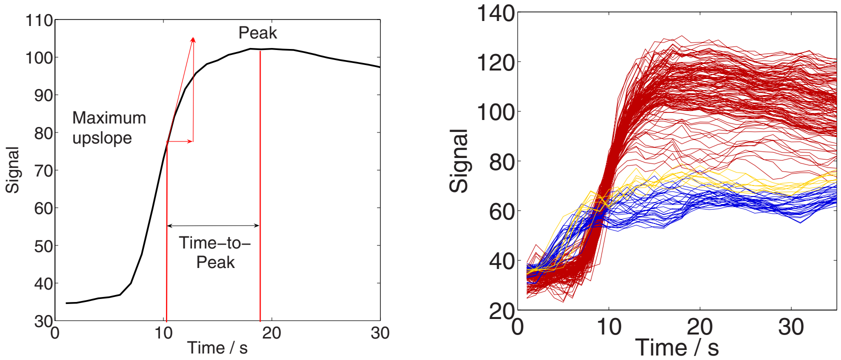


Fig. 5. Different registered frames of one of the slices of the perfusion MR images

The perfusion measures were calculated previously, and they are illustrated in Figure 7(a-c).

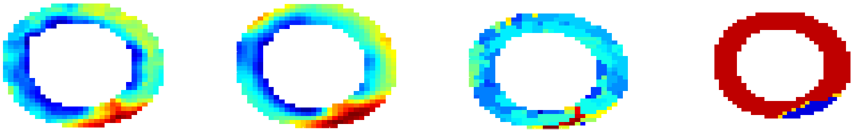
The correspondence between the areas is good, and the clustering is seen to provide a very good base for a simple cluster classification. All noise is suppressed by the HSVC, so the cluster covers a connected region in the image. It is worth noting that the only parameter which has been changed in this example instead of the previous example is the width of the Gaussian kernel. So using the statistical term trace$({S_W^*}^{-1} S_B)$ as cluster separation measure helps to reduce the dimensionality of the estimation problem.



(a) Parameter description.

(b) Intensity plots.

Fig. 6. Pixel-wise intensity plots. (a) Idealized plot, describing the perfusion parameters (b) Intensity curves for the 3 detected clusters, colors correspond to clusters.



(a) Max. upslope.

(b) Peak values.

(c) Time-to-peak.

(d) Clustering.

Fig. 7. Ischemic segment detection with standard measures compared to clustering

4 Conclusion

The proposed robust pseudo-hierarchical support vector clustering (HSVC) is demonstrated to give good results on both a random data set and in real application, and this with the same parameters though the two data sets are very different in range, n and dimensionality.

The proposed clustering algorithm has only one parameter, which is the threshold for splitting clusters, and this parameter correlates strongly with the number of clusters (and their quality in terms of separation). We therefore believe that HSVC can be a very useful tool in many applications where it is possible to define a kernel.

References

1. Ben-Hur, A., Horn, D., Siegelmann, H.T., Vapnik, V.: Support vector clustering. Journal of Machine Learning 2, 125–137 (2001)
2. Tax, D.M.J., Duin, R.P.W.: Support vector domain description. Pattern Recognition Letters 20, 1191–1199 (1999)

3. Ben-Hur, A., Horn, D., Siegelmann, H.T., Vapnik, V.: A support vector clustering method. In: Proceedings of conference on Advances in Neural Information Processing Systems (2001)
4. Yang, J., Estivill-Castro, V., Chalup, S.K.: Support vector clustering through proximity graph modelling (2002)
5. Sjöstrand, K., Larsen, R.: The entire regularization path for the support vector domain description. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, Springer, Heidelberg (2006)
6. Ólafsdóttir, H., Stegmann, M.B., Larsson, H.B.: Automatic assessment of cardiac perfusion MRI. In: Barillot, C., Haynor, D.R., Hellier, P. (eds.) MICCAI 2004. LNCS, vol. 3217, pp. 1060–1061. Springer, Heidelberg (2004)
7. Stegmann, M.B., Ólafsdóttir, H., Larsson, H.B.W.: Unsupervised motion-compensation of multi-slice cardiac perfusion mri. Medical Image Analysis 9(4), 394–410 (2005)
8. Hansen, M.S., Ólafsdóttir, H., Sjöstrand, K., Erbou, S.G., Larsson, H.B., Stegmann, M.B., Larsen, R.: Ischemic segment detection using the support vector domain description. In for Optical Engineering (SPIE), T.I.S., ed.: International Symposium on Medical Imaging. (February 2007)