Optimal Sparse Kernel Learning for Hyperspectral Anomaly Detection

Zhimin Peng [1], Prudhvi Gurram [2], Hessung Kwon [2], Wotao Yin [1]

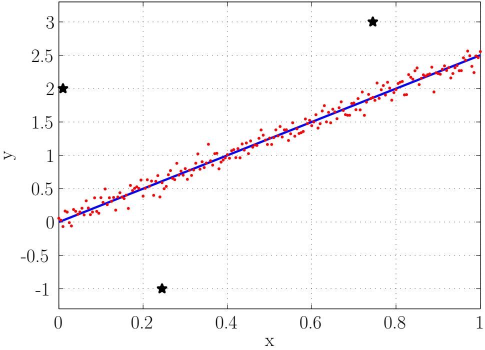
[1]Comp. and Applied Math, Rice University, Houston, TX, 77005

[2]U.S. Army Research Lab, Adelphi, Maryland, 20783

June 18, 2013

What are anomalies?

Anomalies are patterns in data that do not conform to a well defined notion of normal behavior.



Anomalies of hyperspectral images

normal behavior $\Leftrightarrow$ background

hyperspectral anomalies $\Leftrightarrow$ observations deviate from the background.

Support vector data description

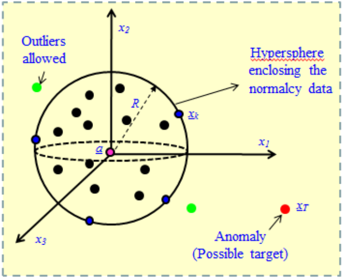
- one of most efficient anomaly detectors (D. Tax, R. Duin [2004])
- learns the support or boundary of the background normalcy data
- minimize the radius of enclosing hypersphere

Model:

$$\min_{\mathbf{a},R,\xi_i} \; L(\mathbf{a},R) = R^2 + C \cdot \sum_i \xi_i$$

$$\text{s.t. } \|\mathbf{x}_i - \mathbf{a}\|^2 \leqslant R^2 + \xi_i$$

$$\xi_i \geqslant 0, \quad \forall i = 1, 2, \cdots, N.$$



Dual problem:

$$\max_{\alpha} L(\alpha_i) = \sum_i \alpha_i \langle \mathbf{x}_i, \mathbf{x}_i \rangle - \sum_{i,j} \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

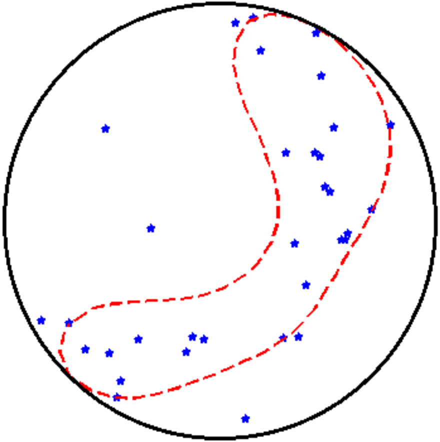
$$\text{s.t. } 0 \leqslant \alpha_i \leqslant C$$

$$\sum_i \alpha_i = 1$$

- if $\alpha_i^* = 0$, $\mathbf{x}_i$ is inside the hypersphere;
- if $\alpha_i^* = C$, $\mathbf{x}_i$ is outside the hypersphere;
- if $0 < \alpha_i^* < C$, $\mathbf{x}_i$ is a support vector.

center: $\mathbf{a} = \sum_i \alpha_i^* \mathbf{x}_i$

radius: $R^2 = \frac{1}{N_b} \sum_{k=1}^{N_b} \|\mathbf{x}_k - \mathbf{a}\|^2$



Kernel based SVDD

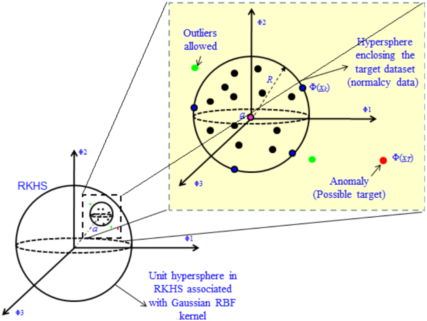
- linear SVDD fails in non-spherical boundary in the input space
- kernel functions map input space to high dimensional feature space
- learns the boundary of the background normalcy data in high dimensional feature space

Model:

$$\min_{\mathbf{a},R,\xi_i} \quad L(\mathbf{a},R) = R^2 + C \cdot \sum_i \xi_i$$

$$s.t. \quad \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2 \leqslant R^2 + \xi_i$$

$$\xi_i \geqslant 0, \quad \forall i = 1, 2, ..., N.$$



Dual problem

$$\max\ L(\alpha_i) = \sum_i \alpha_i k(x_i, x_i) - \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j)$$

$$s.t. \quad 0 \leqslant \alpha_i \leqslant C, \quad \forall i = 1, 2, .., N.$$

$$\sum_i \alpha_i = 1$$

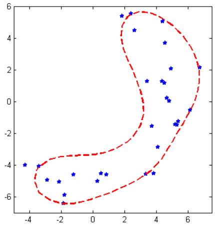
where $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$
Data description:

$$\text{Center:} \quad \mathbf{a} = \sum_i \alpha_i^* \Phi(x_i)$$

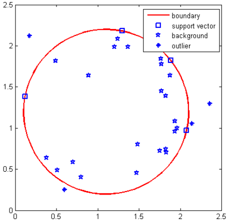
$$\text{Radius:} \quad R^2 = \frac{1}{N_b} \sum_{k=1}^{N_b} \|\Phi(x_k) - a\|^2$$

- overfitting!

input space:



feature space:



Optimal sparse kernel learning (OSKLAD)

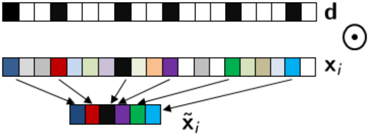
$$\min_{d} \min_{R,\xi_i,a} R^2 + C \cdot \sum_{i=1}^{N} \xi_i$$

$$\text{subject to } \|\Phi(\tilde{\mathbf{x}}_i) - a\|^2 \leqslant R^2 + \xi_i$$

$$\xi_i \geqslant 0$$

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i \odot \mathbf{d}, \quad i = 1, 2, ..., N$$

where $\mathbf{d} \in \mathbb{D} = \{\mathbf{d} | d_j \in \{0, 1\}, \sum_{j=1}^{M} d_j = B\}$.



Dual problem:

$$\min_{\mathbf{d}} \max_{\alpha} \; S(\alpha, \mathbf{d})$$

$$\text{subject to } \sum_{i=1}^{N} \alpha_i = 1$$

$$0 \leqslant \alpha_i \leqslant C$$

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i \odot \mathbf{d}, \;\; i = 1, 2, \ldots, N$$

$$\text{where } S(\alpha, \mathbf{d}) = \sum_{i=1}^{N} \alpha_i k(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_i) - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j k(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$$

- Mixed Integer Programming (MIP)
- \# of possible $\mathbf{d} = \binom{M}{B}$
- if $M = 150$, $B = 75$, # of possible $\mathbf{d} \simeq 9.28 \times 10^{43}$
- NP-complete $\rightarrow$ Hard to solve!

Algorithm

min $\rightleftharpoons$ max

$$
\begin{aligned}
\max_{\alpha_i} \min_{\mathbf{d}} \; & S(\alpha, \mathbf{d}) \\
\text{s.t.} \; & \sum_{i=1}^{N} \alpha_i = 1 \\
& 0 \leqslant \alpha_i \leqslant C \\
& \tilde{\mathbf{x}}_i = \mathbf{x}_i \odot \mathbf{d}
\end{aligned}
$$

QCLP

Introduce slack variable t:

$$
\begin{aligned}
\max_{\alpha, t} \; & t \\
\text{s.t.} \; & \sum_{i=1}^{N} \alpha_i = 1 \\
& 0 \leqslant \alpha_i \leqslant C \\
& t \leqslant S(\alpha, \mathbf{d}), \;\; \mathbf{d} \in \mathbb{D}
\end{aligned}
$$

where $\mathbb{D} = \{\mathbf{d} | d_j \in \{0, 1\}, \sum_{j=1}^{M} d_j = B\}$

Lagrange with respect to t is: $L(t,\mu) = t + \sum_{l=1}^{p} \mu_l (S(\alpha, \mathbf{d}^l) - t)$.

Setting $\frac{\partial L}{\partial t} = 0$

$$\begin{aligned} \max_{\alpha} \min_{\mu} \; & \sum_{l=1}^{p} \mu_l S(\alpha, \mathbf{d}^l) \\ \text{subject to} \; & \sum_{i=1}^{N} \alpha_i = 1 \\ & 0 \leqslant \alpha_i \leqslant C \text{ for } i = 1, 2..., N \\ & \sum_{l=1}^{p} \mu_l = 1 \\ & \mu_l \geqslant 0 \text{ for } l = 1, 2..., p \end{aligned}$$

- solved by the existing algorithm SKAD;[1]
- a large number of kernels $\rightarrow$ Inefficient to solve!

[1] P. Gurram, H. Kwon and T. Han, *Sparse Kernel-based Hyperspectral Anomaly Detection*

- quadratic constraints: $t \leqslant S(\alpha, \mathbf{d})$ where $\mathbf{d} \in \mathbb{D}$;
- only a pool of sparse feature subsets is needed;
- find the most violated d by solving:

$$\min_{\mathbf{d}} S(\alpha, \mathbf{d})$$

$$\text{subject to } \sum_{i=1}^{p} d_i = B$$

$$\mathbf{d} \in \mathbb{D}$$

- If linear kernel $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$ is used, then we have

$$S(\alpha, \mathbf{d}) = \sum_{j=1}^{M} d_j c_j$$

where $c_j = \sum_{i=1}^{N} \alpha_i x_{ij}^2 + (\sum_{i=1}^{N} \alpha_i x_{ij})^2$

- If $k(\mathbf{x}, \mathbf{y}) = \exp(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2})$, use empirical kernel map.

Empirical kernel feature space

empirical kernel map:

$$\Phi_N : \mathbb{R}^M \to \mathbb{R}^N, \text{ where } x \mapsto (k(\mathbf{x}_1, \mathbf{x}), ..., k(\mathbf{x}_N, \mathbf{x}))^T$$

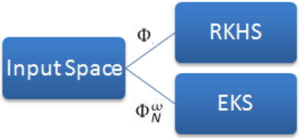
"whitening" empirical kernel map

$$\Phi_N^\omega : x \mapsto K^{-\frac{1}{2}}(k(\mathbf{x}_1, \mathbf{x}), ..., k(\mathbf{x}_N, \mathbf{x}))^T$$

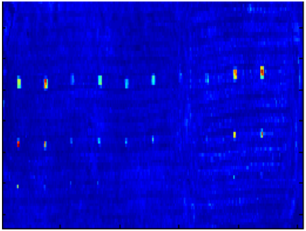
which satisfy

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi_N^\omega(\mathbf{x}_i), \Phi_N^\omega(\mathbf{x}_j) \rangle$$

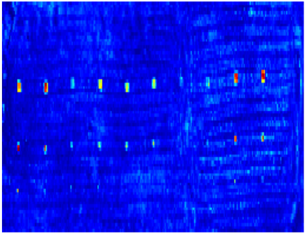
benefit: find a N-dimensional feature space associate with a given kernel $k(\cdot, \cdot)$.



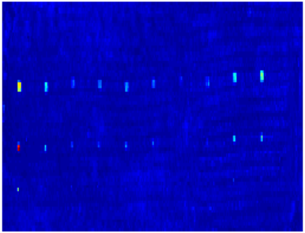
Results



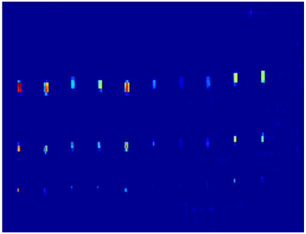
(a) SVDD – linear kernel



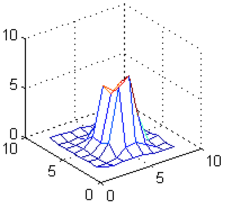
(b) SVDD – RBF kernel



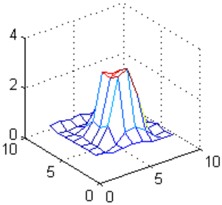
(c) OSKLAD – linear kernel



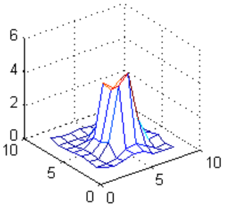
(d) OSKLAD – EKFS



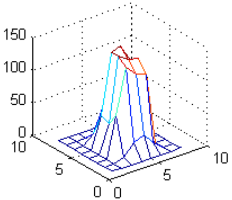
(a) SVDD – linear kernel



(b) SVDD – RBF kernel



(c) OSKLAD – linear kernel



(d) OSKLAD – EKFS

Conclusions:

- a novel framework for anomaly detection;
- features are optimally selected in nonlinear feature space.

Future work:

- local spectral anomaly detection;
- parallelize OSKLAD.