Machine learning and Change detection

Gilles GASSO

LITIS EA 4108, INSA Rouen

CNET Colloqium, Toulouse

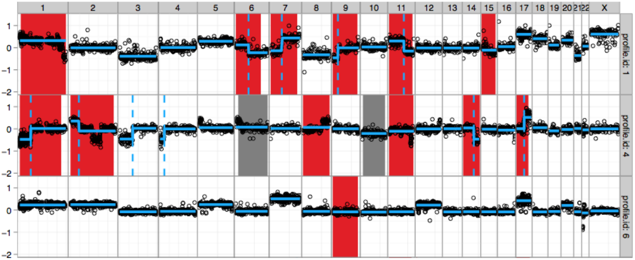
May 10, 2016

Outline

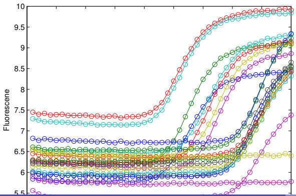
1. Introduction
   - Introducing examples
   - Tentative categorization of the problems
2. Kernel based approaches
   - One-Class SVM
   - Application to novelty and change detection
3. Classification approach
4. Conclusion

Introduction : different problems

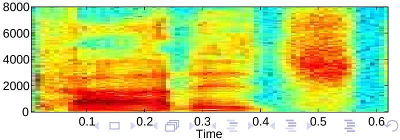
Identify breakpoints in DNA profiles of patients



Early detection of biological threats
based on fluorescence analysis



Speaker segmentation



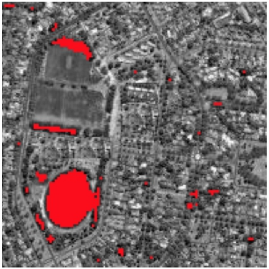
Introduction : different problems

Detect novel pixels in a image

Reference image

New image

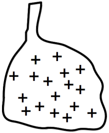
Detected new pixels



Introduction : different problems

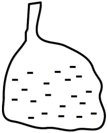
Guided robust discovery

$$\max \; TPR \quad \text{s.t.} \quad FPR \leq q\,TPR \qquad (q \ll 1 : \text{confidence level})$$



Possible positives (label $y = ?$)

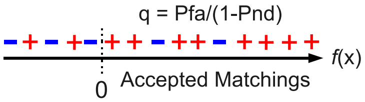
vs



Reliable Negatives (label $y = -1$)

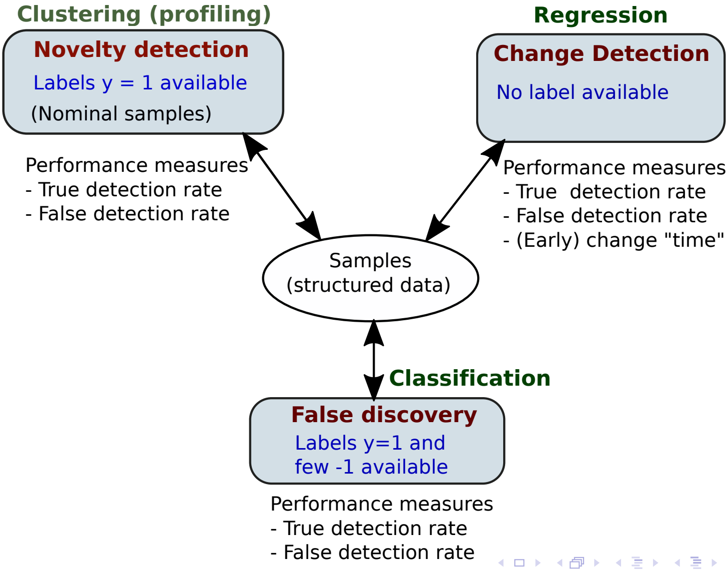
Application

- Matching spectrum with peptides (pieces of proteins)
- Fake spectra are well known (randomly generated)
- True spectra are conjectured

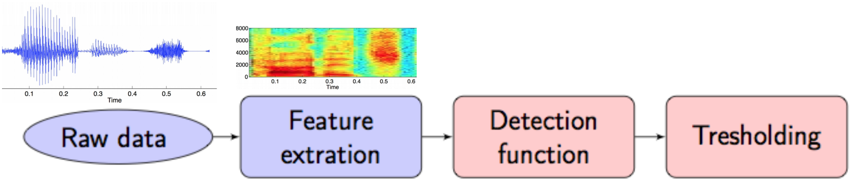


- Assume $q = 0.01$ and $n_+ = n_-$
- Expecting $TP = 1000 \rightarrow FP \leq 10$

Tentative taxonomy



Methodology



Taxonomies of detection approaches

- Homogeneity test based
- Non-parametric modeling
- Offline (batch) or online decisions

Focus of this talk

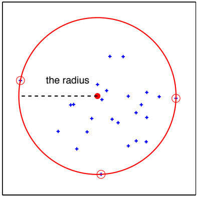
- One-class SVM for novelty and change detection
- Classification approach for false discovery

Kerne- based approaches : one-class SVM Smola and Schölkopf [1998]

Minimum enclosing ball problem : Support vector data description (SVDD)

Given N points, $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^N$, find

$$\min_{R\in\mathbb{R}, \boldsymbol{c}\in\mathbb{R}^d} \quad R^2$$
$$\text{s.t.} \quad \|\mathbf{x}_i - \boldsymbol{c}\|^2 \leq R^2, \quad \forall i$$



Rewritting SVDD

$$\min_{\rho\in\mathbb{R}, \boldsymbol{c}\in\mathbb{R}^d} \quad \frac{1}{2}\|\boldsymbol{c}\|^2 - \rho$$
$$\text{s.t.} \quad \boldsymbol{c}^\top \mathbf{x}_i \geq \rho + \|\mathbf{x}_i\|^2, \quad \forall i$$

with $\rho = \frac{1}{2}(\|\boldsymbol{c}\|^2 - R^2)$

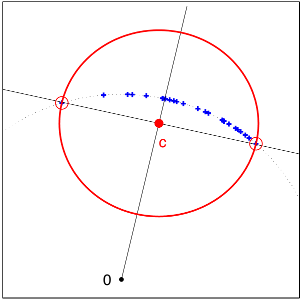
Linear OC-SVM

is recovered if $\|x_i\|^2 = \text{constant}$

$$\min_{\rho\in\mathbb{R}, \boldsymbol{c}\in\mathbb{R}^d} \quad \frac{1}{2}\|\boldsymbol{c}\|^2 - \rho'$$
$$\text{s.t.} \quad \boldsymbol{c}^\top \mathbf{x}_i \geq \rho', \quad \forall i$$

$\rightarrow$ OC-SVM is a particular case of SVDD

Illustration : the sphere and the hyperplane



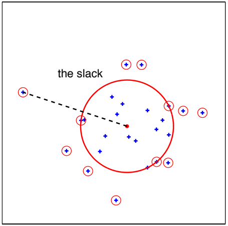
SVDD and OCSVM when $\forall i = 1, N,\ \|\mathbf{x}_i\|^2 = 1$

- $\|\mathbf{x}_i - \boldsymbol{c}\|^2 \leq R^2 \quad \Leftrightarrow \quad \boldsymbol{c}^\top \mathbf{x}_i \geq \rho$

  "Belonging to the ball" $\Leftrightarrow$ "being above" an hyperplane
- $\|\mathbf{x}_i\|^2 = 1 \quad \Leftrightarrow \quad$ samples $\mathbf{x}_i$ lie on a sphere

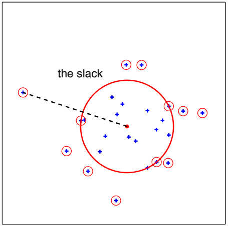
Dealing with outliers and non-linear case

Outliers : allow a proportion of reference samples to be out of the enclosing ball



$$\begin{array}{lll} \min\limits_{R,\mathbf{c},\xi} & R^2 + \mu \sum_{i=1}^{N} \xi_i & \\ \text{s.t.} & \|\mathbf{x}_i - \mathbf{c}\|^2 \leq R^2 + \xi_i, & i = 1, \ldots, N \\ \text{and} & \xi_i \geq 0, & i = 1, \ldots, N \end{array}$$

Dealing with outliers and non-linear case



$$\begin{array}{lll} \min\limits_{R,\mathbf{c},\xi} & R^2 + \mu \sum_{i=1}^N \xi_i & \\ \text{s.t.} & \|\mathbf{x}_i - \mathbf{c}\|^2 \leq R^2 + \xi_i, & i = 1, \ldots, N \\ \text{and} & \xi_i \geq 0, & i = 1, \ldots, N \end{array}$$

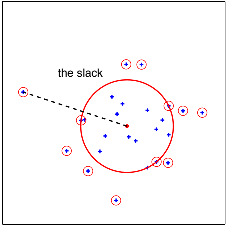
Handle non-linear case: use kernel

Definition (Kernel)

A function of two variable $k(\mathbf{x}, \mathbf{x}')$ with values in $\mathbb{R}$, symmetric positive

- Linear kernel: $k(\mathbf{x}, \mathbf{x}) = \mathbf{x}^\top \mathbf{z}$
- Gaussian kernel $k(\mathbf{x}, \mathbf{z}) = \exp(-\frac{\|\mathbf{x}-\mathbf{z}\|^2}{b})$

Dealing with outliers and non-linear case

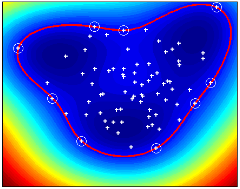


$$\begin{array}{lll} \min\limits_{R,\mathbf{c},\xi} & R^2 + \mu \sum_{i=1}^N \xi_i & \\ \text{s.t.} & \|\mathbf{x}_i - \mathbf{c}\|^2 \leq R^2 + \xi_i, & i = 1, \ldots, N \\ \text{and} & \xi_i \geq 0, & i = 1, \ldots, N \end{array}$$

Handle non-linear case: use kernel

Nonlinear mapping:

$$\begin{array}{rcl} \mathbb{R}^d & \longrightarrow & \mathcal{H} \\ c & \longrightarrow & f(\bullet) \\ \mathbf{x}_i & \longrightarrow & k(\mathbf{x}_i, \bullet) \\ \|\mathbf{x}_i - c\|^2_{\mathbb{R}^d} \leq R^2 & \longrightarrow & \|k(\mathbf{x}_i, \bullet) - f(\bullet)\|^2_{\mathcal{H}} \leq R^2 \end{array}$$



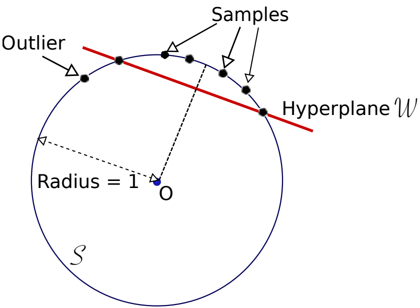
OC-SVM $\equiv$ SVDD with translation invariant kernel with $k(\mathbf{x}_i, \mathbf{x}_i) =$ constant

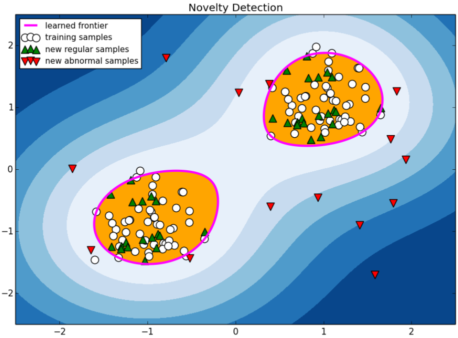
Applications

- Novelty detection
- Change detection

Novelty detection

Recall $k(\mathbf{x}, \mathbf{x}) = \|k(\mathbf{x}, \bullet)\|^2_{\mathcal{H}} = \text{constant} \longleftrightarrow$ data lie on a sphere





Novelty detection

- Learn the hyperplane using (reference) training data
- New samples: deemed novel if below the hyperplane

Novelty detection

Pros

- Avoid density estimation of nominal data
- Kernel OC-SVM estimates the distribution level set $\{\mathbf{x} \in \mathbb{R}^d \mid \mathbb{P}(\mathbf{x}) \geq \rho\}$
- Can handle vectorial or non-vectorial data (graphs, sequences. . .)
- Benefit from huge data

Cons

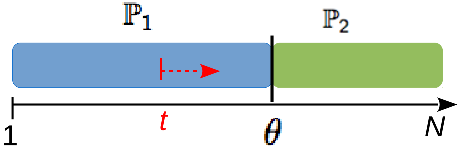
- Complexity of the underlying optimization problem
- Choice of the kernel parameter(s) and hyper-parameter μ

Application domains (see for instance Pimentel et al. [2014])

- Electronics IT security, industrial system surveillance
- Medical diagnosis

Change detection: principle

- $\mathbb{H}_0 : \{\mathbf{x}_1, \cdots, \mathbf{x}_N\} \sim \mathbb{P}_1$
- $\mathbb{H}_1$: there exists θ such that $\{\mathbf{x}_1, \cdots, \mathbf{x}_\theta\} \sim \mathbb{P}_1$ and $\{\mathbf{x}_{\theta+1}, \cdots, \mathbf{x}_N\} \sim \mathbb{P}_2$



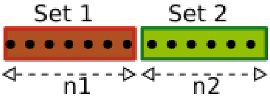
Issues

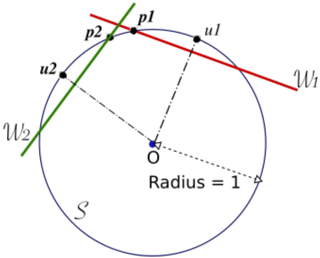
- Find test statistic S_t
- Find threshold γ in order to maximize probability of detection $\mathbb{P}(S_t \geq \gamma | \mathbb{H}_1)$ for a fixed false alarm rate $\mathbb{P}(S_t < \gamma | \mathbb{H}_0) = \alpha$
- Test : Decide a change occurs if there is $1 < t < N$ such that $S_t > \gamma$

Change detection with OC-SVM Desobry et al. [2006]

Learn two OC-SVM on sets $\mathcal{X}_1 = \{\mathbf{x}_1, \cdots, \mathbf{x}_t\}$ and $\mathcal{X}_2 = \{\mathbf{x}_{t+1}, \cdots, \mathbf{x}_N\}$

$\rightarrow$ test for homogeneity of their level sets

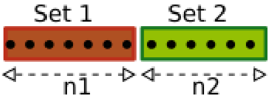


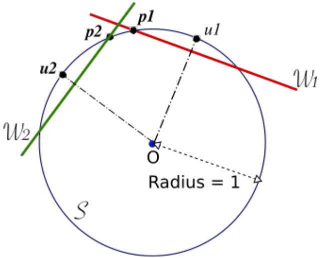


Change detection with OC-SVM Desobry et al. [2006]

Learn two OC-SVM on sets $\mathcal{X}_1 = \{\mathbf{x}_1, \cdots, \mathbf{x}_t\}$ and $\mathcal{X}_2 = \{\mathbf{x}_{t+1}, \cdots, \mathbf{x}_N\}$

$\rightarrow$ test for homogeneity of their level sets





Change detection

- Based on inter-region/intra-region ratio of the level sets
- Decide a change (sets $\mathcal{X}_1$ and $\mathcal{X}_2$ are statistically different) if

$$S_t = \frac{\widehat{u_1 u_2}}{\widehat{u_1 p_1} + \widehat{u_2 p_2}} > \gamma$$

Change detection with OC-SVM Desobry et al. [2006]

Learn two OC-SVM on sets $\mathcal{X}_1 = \{\mathbf{x}_1, \cdots, \mathbf{x}_t\}$ and $\mathcal{X}_2 = \{\mathbf{x}_{t+1}, \cdots, \mathbf{x}_N\}$

$\rightarrow$ test for homogeneity of their level sets

Change detection

- Based on inter-region/intra-region ratio of the level sets
- Decide a change (sets $\mathcal{X}_1$ and $\mathcal{X}_2$ are statistically different) if

$$S_t = \frac{\widehat{u_1 u_2}}{\widehat{u_1 p_1} + \widehat{u_2 p_2}} > \gamma$$

Related method: kernel Fisher ratio

Sets: $\mathcal{X}_1 = \{\mathbf{x}_1, \cdots, \mathbf{x}_t\}$ and $\mathcal{X}_2 = \{\mathbf{x}_{t+1}, \cdots, \mathbf{x}_N\}$

Mapping: $\mathbf{x}_i \longrightarrow k(\mathbf{x}_i, \bullet)$

Change detection statistics Harchaoui et al. [2009a,b]

- Intuition: maximize the separation of sets $\mathcal{X}_1$ and $\mathcal{X}_2$
- Statistics : $S_t \propto \|(\mathbf{\Sigma} + \lambda \mathbf{I})^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\|^2_{\mathcal{H}}$

  $\boldsymbol{\mu}_j$: mean vector of $\mathcal{X}_j$ in $\mathcal{H}$

  $\mathbf{\Sigma}$ covariance operator defined as $\mathbf{\Sigma} \propto \beta \mathbf{\Sigma}_1 + (1-\beta)\mathbf{\Sigma}_2$

	Semantic seg.		Speaker seg.	
	Precision	Recall	Precision	Recall
KFDR	0.72	0.63	0.89	0.90
MMD	0.71	0.58	0.76	0.73
KCD	0.65	0.63	0.78	0.74
HMM	0.73	0.65	0.93	0.96

Change detection using kernel approach

Pros

- Same as for novelty detection

Cons

- Computation time of the statistics
- Choice of the kernel parameter(s)
- Setting the threshold

Applications

- Signals, videos segmentation
- Bscan images, remote sensing images . . .

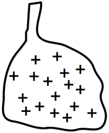
Classification approach

- Controlling false discovery

Classification approach

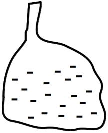
Controlling false discovery Gasso et al. [2011]

$$\min_f \Omega(f) + \lambda\ FNR(f) \quad \text{s.t.} \quad FPR(f) \leq q(1 - FNR(f)) \quad (q \ll 1 : \text{confidence level})$$



Possible positives (label $y =?$)

vs



Reliable Negatives (label $y = -1$)

Classification approach

Controlling false discovery Gasso et al. [2011]

$$\min_f \Omega(f) + \lambda\, FNR(f) \quad \text{s.t.} \quad FPR(f) \leq q(1 - FNR(f)) \quad (q \ll 1 : \text{confidence level})$$

Estimation of probabilities of error

- Data set $\mathcal{X}_+ = \{(\mathbf{x}_i, y_i = 1)\}_{i=1}^{n_+}, \quad \mathcal{X}_- = \{(\mathbf{x}_i, y_i = -1)\}_{i=1}^{n_-}$
- f: decision function to be learned
- Empirical probability errors $(0 - 1$ errors)

$$FNR(f) = \frac{1}{n_+} \sum_{i \in \mathcal{X}_+} \mathbb{I}_{f(\mathbf{x}_i) \leq 0}, \quad FPR(f) = \frac{1}{n_-} \sum_{i \in \mathcal{X}_-} \mathbb{I}_{f(\mathbf{x}_i) \geq 0}$$

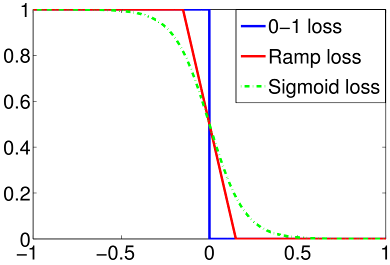
Using $0 - 1$ errors leads to NP hard problem

Classification approach

Dealing with the probabilities of errors

- Non-convex approximation of the 0-1 errors

$$\hat{FPR}(f) = \frac{1}{n_+} \sum_{i \in \mathcal{X}_+} \ell(y_i f(\mathbf{x}_i)) \,, \quad \hat{FNR}(f) = \frac{1}{n_-} \sum_{i \in \mathcal{X}_-} \ell(y_i f(\mathbf{x}_i)) \,.$$



Classification approach

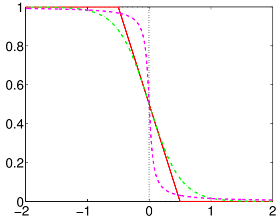
Proposed Algorithms

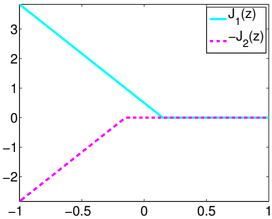
- Kernel machine (SVM)
  - Ramp loss approximation
    $\ell(z) = \max\left\{0, \frac{1}{2}(1-z)\right\} - \max\left\{0, -\frac{1}{2}(1+z)\right\}$
  - Remark: non-convex and non-differentiable
  - Batch learning for non-linear SVM: tool = DC programming (Tao and An [1998], Gasso et al. [2009])
  - Online learning for linear SVM (large scale datasets): tool = stochastic gradient
- Deep network
  - Sigmoid loss approximation $\ell(z) = \frac{1}{1+e^z}$
  - Online learning with stochastic gradient

Dealing with the non-convexity: elements of the solution

Decompose the loss as the difference of two convex functions

$$\ell(z) = \max\left\{0, \tfrac{1}{2}(1-z)\right\} - \max\left\{0, -\tfrac{1}{2}(1+z)\right\} = \ell_1(z) - \ell_2(z)$$





Principle: successive convex relaxations

- At each iteration t, define the convex majorization function

$$J_{cvx}(f) = J_1(f) - J_2(f_t) - \langle f - f_t, \boldsymbol{\alpha}_t \rangle \quad \text{with} \quad \boldsymbol{\alpha}_t \in \partial J_2(f_t)$$

- Next solution: $f_{t+1} = \text{argmin}_f \, J_{cvx}(f)$

Performance evaluation: q-value

Setup

- Peptides-spectrum matching (PSM) verification
- Goal: identify consistently true positive matchings
- Models investigated : non-linear SVM (qSVMOpt), deep network (qNNOpt)

q	qRanker	qSVMOpt	qNNOpt
0.0025	4,449	4,947	**5,005**
0.01	5,462	5666	**5,707**
0.1	7,473	**7,954**	7,491

Table: Number of true positives correctly identified (over 34,852).

Classification approach

Pros

- Benefit from labeled data
- Grounded in well known empirical risk minimization
- Extensions to Neyman-Pearson classification (learning under probability constraint on the false alarm)

Cons

- Non-convex optimization problems
- Dealing with probability constraints

Applications

- Bioinformatics
- Network surveillance (Distributed deni of service)
- Text mining

Conclusion: related work of the team

- Non-convex optimization
  - Learning with probability constraints
  - Robust (to outliers) SVDD
- Metric learning
  - Choice of the kernel in SVDD
  - Optimal transport to learn adapted metric to the data
  - Exploit manifold information for change detection
- Early change detection
  - Classification based detection using incomplete sequence

F. Desobry, M. Davy, A. Gretton, and C. Doncarli. An online support vector machine for abnormal events detection. *Signal Processing*, 86:2009–2025, 2006.

Gilles Gasso, Alain Rakotomamonjy, and Stéphane Canu. Recovering sparse signals with a certain family of nonconvex penalties and dc programming. *Signal Processing, IEEE Transactions on*, 57(12):4686–4698, 2009.

Gilles Gasso, Aristidis Pappaioannou, Marina Spivak, and Léon Bottou. Batch and online learning algorithms for nonconvex neyman-pearson classification. *ACM Trans. Intell. Syst. Technol.*, 2:28:1–28:19, May 2011. ISSN 2157-6904.

Zaid Harchaoui, Eric Moulines, and Francis R Bach. Kernel change-point analysis. In *Advances in Neural Information Processing Systems*, pages 609–616, 2009a.

Zaid Harchaoui, Félicien Vallet, Alexandre Lung-Yut-Fong, and Olivier Cappé. A regularized kernel-based approach to unsupervised audio segmentation. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 1665–1668. IEEE, 2009b.

Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

Alex J Smola and Bernhard Schölkopf. *Learning with kernels*. Citeseer, 1998.

P. D. Tao and L. T. Hoai An. Dc optimization algorithms for solving the trust region subproblem. *SIAM Journal of Optimization*, 8(2):476–505, 1998.