Feature Selection for Density Level-Sets

Marius Kloft† Shinichi Nakajima‡ Ulf Brefeld†

† Machine Learning Group, TU Berlin
‡ Optical Research Laboratory, Nikon Corporation, Tokyo, Japan

Motivating Example

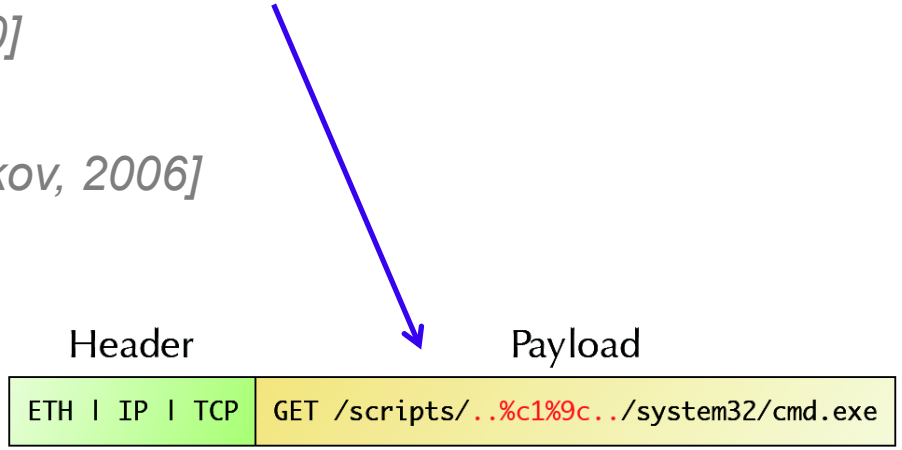
Network intrusion detection.

➔ find new attacks in incoming network traffic

➔ no knowledge about future attacks

Various feature representations of *HTTP requests* possible, e.g.

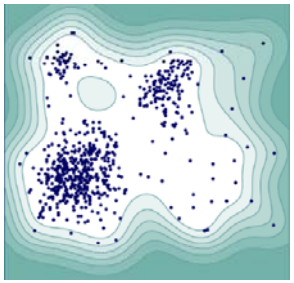
- *expert features* *[Lee & Stolfo, 2000]*
- *headers* *[Mahoney & Chan, 2003]*
- *n-grams of payloads* *[Rieck & Laskov, 2006]*



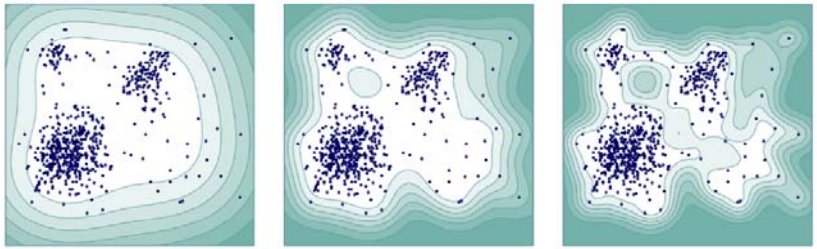
Problem Setting

Density level-set estimation

for the detection of novel or *anomalous* instances, e.g. *network attacks.*



Given: data x_i and several feature embeddings, each encoded by a kernel $K_j,\ j = 1, ..., m,$ e.g. *n*-grams payloads for various *n.*



Goal:

- find a density level-set $D = \{\mathbf{x} : f(\mathbf{x}) \geq \rho\}$
- and an optimal kernel mixing $K = \sum_{j=1}^{m} \beta_j K_j$

What is the optimal kernel mixture?

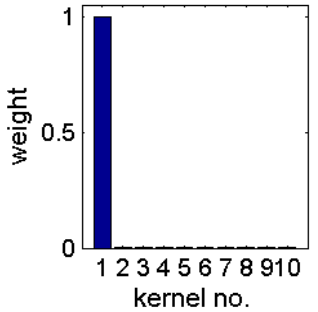
We focus on linear kernel mixtures $K = \sum_{j=1}^{m} \beta_j K_j$ and 1-class SVMs.

Heuristic 1: use a single kernel

which is optimal in model selection (e.g. cross-validation)

Disadvantage:

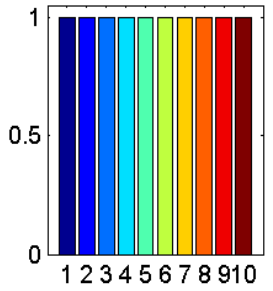
useful information discarded.



Heuristic 2: the uniform kernel mixture $\beta_1 = \cdots = \beta_m = \frac{1}{m}$

Disadvantage:

arbitrary choice → irrelevant kernels considered.



What is the optimal kernel mixture?

Heuristic 3: Brute-Force: try out all possible mixtures (e.g. grid search)

Infeasible: *computationally too demanding.*

Can we do better?

➔ integrate feature/kernel selection into density level-set estimator, i.e. 1-class SVM.

Learning an optimal kernel mixture

Multiple kernel learning (MKL):

Simultaneously determine

- an optimal kernel combination $K = \sum_j \beta_j K_j$,
- and a density level set $D = \{\mathbf{x} : f(\mathbf{x}) \geq \rho\}$

➔ such that a scoring function *(here 1-class SVM objective)*, $s(f(K))$, is minimal in *K.*

[Lanckriet et al., 2004]

Optimization Problem (MKL)

$$\min_{\boldsymbol{\beta}} \quad \text{svm}(\sum_j \beta_j K_j), \quad \text{s.t.} \quad \boldsymbol{\beta} \geq \mathbf{0}, \; ||\boldsymbol{\beta}||_1 = 1$$

➔ $\beta_i = 0$ for most *i*: classical MKL finds a sparse combination of kernels

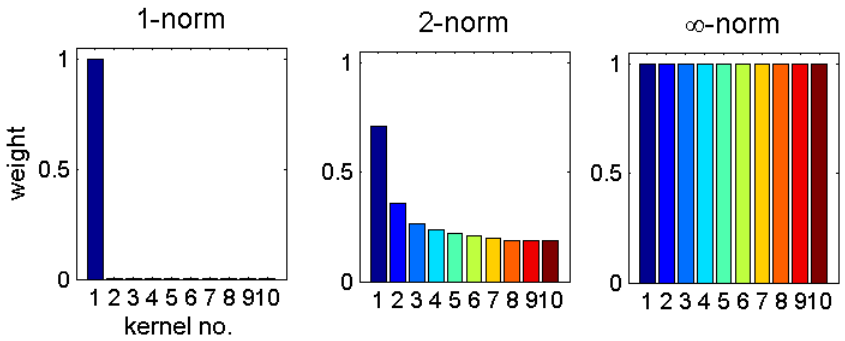
Our Contribution

Observation:

- MKL focuses on single kernel solutions
- all other mixing coefficients are set to zero [Shawe-Taylor & Hussain, 2008]

Remedy: generalize the $||\beta||_1 = 1$ constraint to $||\beta||_p = 1$ with p≥1.

➔ Enables smooth kernel mixtures:



Generalized 1-class MKL

Optimization Problem (Min-Max)

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\alpha}} \quad -\frac{1}{2}\boldsymbol{\alpha}^\top \sum_j \beta_j K_j \boldsymbol{\alpha}$$

$$\text{s.t.} \quad \mathbf{0} \leq \boldsymbol{\alpha} \leq \frac{1}{\nu n}; \quad \mathbf{1}^\top \boldsymbol{\alpha} = 1; \quad \boldsymbol{\beta} \geq 0; \quad \boxed{||\boldsymbol{\beta}||_p^p = 1.}$$

- The ℓ_p-norm equality constraint ruins convexity.
  Remedy: we relax the equality constraint to an inequality: $||\boldsymbol{\beta}||_p^p \leq 1$

- How can we optimize above OP?

Translation into Semi-Infinite Program

Optimization problem (SIP). Given kernel matrices $K_1, ..., K_m$,

$$\min_{\lambda,\boldsymbol{\beta}} \quad \lambda$$

$$\text{s.t.} \quad \lambda \geq -\frac{1}{2}\boldsymbol{\alpha}^\top \sum_j \beta_j K_j \boldsymbol{\alpha}, \quad \forall \boldsymbol{\alpha} : \mathbf{0} \leq \boldsymbol{\alpha} \leq \frac{1}{\nu n}; \quad \mathbf{1}^\top \boldsymbol{\alpha} = 1;$$

$$\boldsymbol{\beta} \geq 0; \quad ||\boldsymbol{\beta}||_p^p \leq 1.$$

Above OP is convex and can be optimized by alternating α- and β-steps:

α-step: *solve SVM(α) with minor SVM iterations for actual mixture β to generate new constraint for SIP*

β-step: *solve SIP to refine actual β*

Problem: the *p*-norm constraint arising in step (2) cannot be handled by standard optimization toolboxes

Handling of p-norm Constraint

- We approximate the *p*-norm constraint by a quadratic Taylor Expansion, i.e.

$$||\boldsymbol{\beta}||_p^p \approx 1 - \frac{p(3-p)}{2} - \sum_j p(p-2)(\beta_j^{\text{old}})^{p-1}\ \beta_j + \frac{p(p-1)}{2}\sum_j (\beta_j^{old})^{p-2}\ \beta_j^2$$

➔ Solve a sequence of quadratically constrained subproblems (QCQPs).

Experiment 1: Intrusion Detection

Normal data

- 2500 unsanitized HTTP requests recorded at Fraunhofer Institute
- randomly drawn from two months of incoming HTTP traffic.

Attack data

- 30 instances of 10 different recent attacks (see Metasploit):

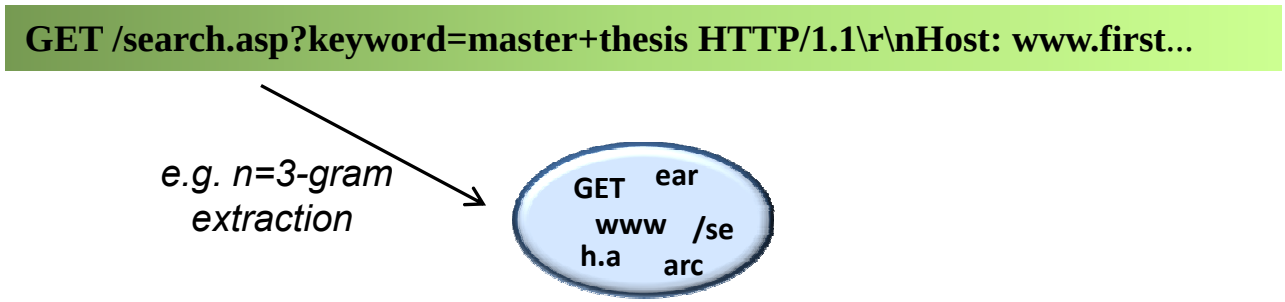
  6 buffer overflow exploits, 4 PHP vulnerabilities, 1 Nessus Scan.
- normalized to match the characteristics of normal HTTP requests.

Experimental setup:

- randomly drawn distinct training, tuning, and test sets
- attacks of the same class occur only either in validation or test set
- model selection, 100 repetitions, AUC over *interval [0,0.01].*

Intrusion Detection Experiment: Feature Representation

- Employed several *n*-gram kernels, for *n=1,…,10.*
  - Features are frequencies of occurrences of *n*-grams in HTTP requests:



- All kernels are normalized: $K(\mathbf{x}, \tilde{\mathbf{x}}) \longmapsto \frac{K(\mathbf{x}, \tilde{\mathbf{x}})}{\sqrt{K(\mathbf{x}, \mathbf{x}) K(\tilde{\mathbf{x}}, \tilde{\mathbf{x}})}}$
- At all, 10 different kernel functions

➔ Compare ℓ_p-norm MKL, $p = 1, \frac{4}{3}, 2, 4$,
with the uniform kernel mixture (∞-norm MKL).

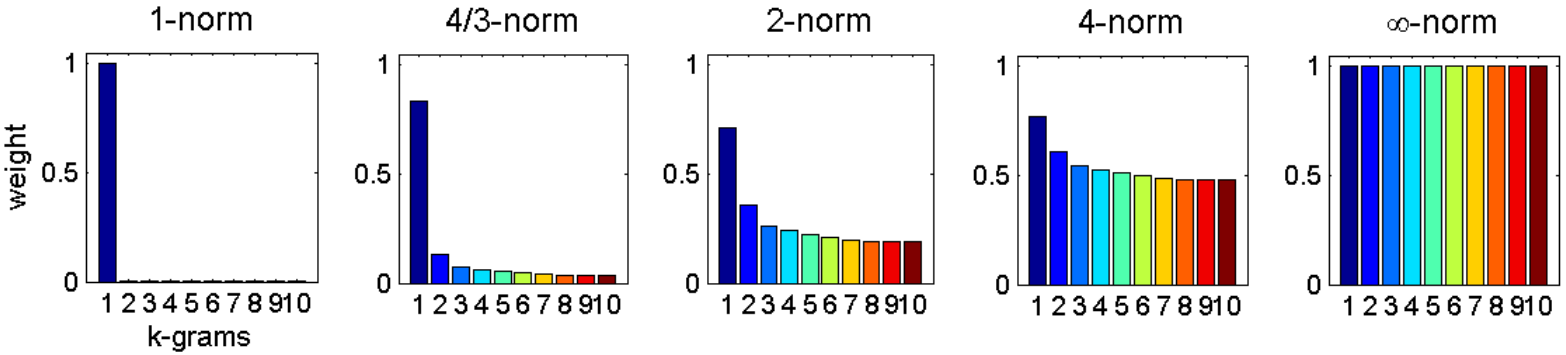
Intrusion Detection Experiment: Empirical Results

AUC Performance:

MKL	$\text{AUC}_{0.01}$
uniform	89.4 ± 0.7
1-norm	79.4 ± 0.9
$\frac{4}{3}$-norm	85.7 ± 0.8
2-norm	90.7 ± 0.8
4-norm	88.9 ± 0.9

- ℓ_2 *-MKL outperforms the uniform kernel mixture*
- ℓ_1 *-MKL performs worst.*

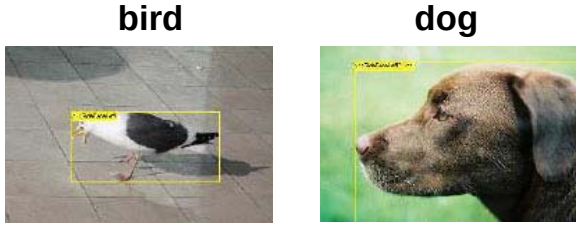
Optimal kernel mixtures:



Experiment 2:
Multi-Label Image Categorization

Data set

- VOC 2008 challenge data set:
  - 8780 images
  - 20 categories (*aeroplane, bird*, *dog*, …).



Feature extraction

- employed 12 domain-specific kernels (variation: 30 kernels)
  - based on several combination of basic features:
    e.g. histogram of visual words, color sets, pyramid level tilings
- all kernels are normalized.

Experimental setup

- train a single model for each category
- randomly drawn distinct training, tuning, and test sets
- 10 repetitions, model selection.

Image Categorization Experiment: Empirical Results

Average precision (AP) performance:

	1-norm	p^*-norm	2-norm	∞-norm
mean AP (K12)	**17.6±0.8**	**17.8±1.0**	17.1±0.8	17.0±0.6
mean AP (K30)	16.3±0.5	**17.1±0.9**	**17.1±0.6**	**17.0±0.7**

*p**-norm = *p* optimized class-wise

K12:
- ℓ_1 *-MKL outperforms the uniform mixture (*ℓ_∞*-norm)*
- ℓ_p *-MKL performs equally well but lacks interpretability*

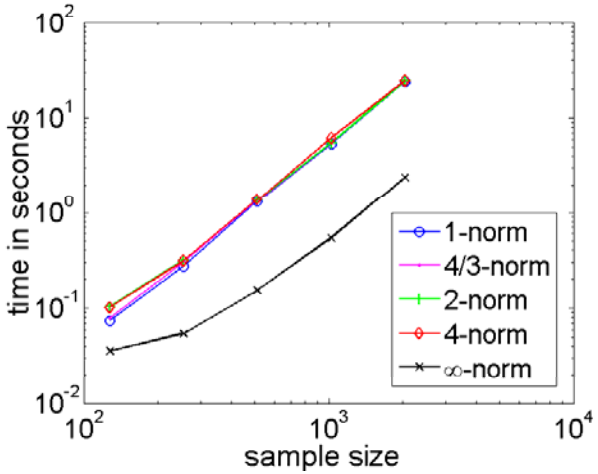
K30:
- *non-sparse* ℓ_p *-norms and uniform mixture perform best*

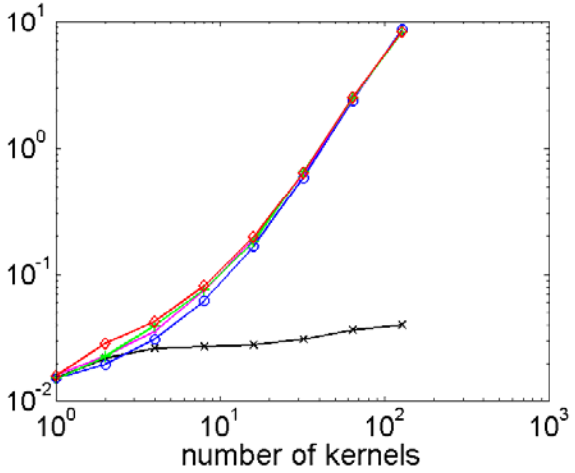
Runtime Experiment

Experimental setup

- generated a 10-dimensional gaussian sample; employed various RBF kernels.
- sample size and number of kernels varied; duality gap optimized up to 10^3.

Empirical results:





- uniform mixture baseline is the fastest method
- all *p*-norms perform similar; longer computation time for large values of *p*.

Conclusion

Observation:

- classical MKL for 1-class SVMs performs poorly.

Contributions:

- non-sparse multiple kernel learning for 1-class SVMs
- semi-infinite linear programming
- interleaved (alpha/beta) optimization
- empirically, ℓ_1-norm almost always outperformed by non-sparse mixtures.

Free implementation:

- http://www.shogun-toolbox.org/

The End

Thank you for your attention!

References

- Lee, W., Stolfo, S.: **A framework for constructing features and models for intrusion detection systems.** In: *ACM Transactions on Information and Systems Security,* 3(4):207-226, 2000.

- Lanckriet, G., Christianini, N., Bartlett, P., El Ghaoui, L., Jordan, M.: **Learning the kernel matrix with demidefinite programming.** In: *Journal of Machine Learning Research,* 5(Jan):27-72, 2004.

- Mahoney, M., Chan, P.: **Learning rules for anomaly detection of hostile network traffic.** In: *Proceedings of the 3rd IEEE International Conference on Data Mining,* 601-604, 2003.

- Rieck, K., Laskov, P.: **Language models for detection of unknown attacks in network traffic.** In: *Journal in Computer Virology,* 2(4): 243-256, 2007.

- Shawe-Taylor, J., Hussain, Z.: **Kernel learning for novelty detection.** In: *Proceedings of the NIPS Workshop on Kernel Learning,* 2008.

- Sonnenburg, S., Rätsch, G., Schäfer, C., Schölkopf, B.: **Large scale multiple kernel learning.** In: *Journal of Machine Learning Research,* 7(Jul):1531-1565, 2006.

Appendix A: MKL Optimization

1-step optimization techniques

[Lanckriet et al., 2004]: SDP, SOCP.

[Bach et al., 2004]: SMO.

2- step optimization techniques

[Sonnenburg et al.]: SIP

[Rakotomamonjy et al., 2007]: projected gradient.

[Xu et al., 2008]: level set method.