Invariance in Kernel Methods - Distance and Integration Kernels

PASCAL Workshop on Machine Learning, SVM and Large Scale Optimization 2005

Dipl.-Math. Bernard Haasdonk

University of Freiburg, Germany
Computer Science Department
Chair of Pattern Recognition and Image Processing

Abstract

In pattern analysis with kernel methods, it is widely accepted, that the kernel function is the main ingredient to represent any kind of prior knowledge. A frequently observed kind of prior knowledge is invariance with respect to transformations of single patterns. For this case we present two generic methods of incorporating such knowledge into quite arbitrary kernels.

The first is based on invariant distances, the second on invariant integration. As the former method can not guarantee to result in positive definite kernels, we discuss the use of indefinite kernels in machine learning. Application on several classification problems with SVMs demonstrates the competitiveness in terms of recognition accuracy and computational complexity with existing methods.

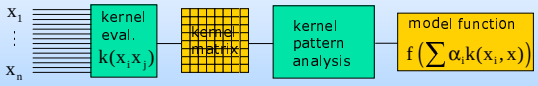
Overview

- Motivation and Notions
  - Kernel Methods and Invariance
- Invariant Distance Substitution Kernels
  - Distance Substitution Kernels
  - Tangent Distance Kernels
  - Application Raman-Microspectroscopy
- Transformation Integration Kernels
  - Application OCR
- Indefinite Kernels in SVM
  - Interpretation in pE spaces
- Summary and Perspectives

Motivation and Notions

Kernel Methods [SS02,SC04]

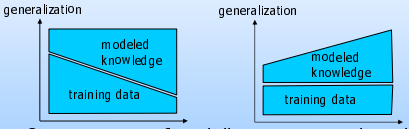
- Typical tasks:
  - Classification, Regression, Clustering, Novelty Detection,...
- Multitude of kernel methods: SVM, SVR, KPCA,...
- Analysis chain:



- Multitude of kernels for various datatypes
  - Vectorial, sequences, graphs, finite state machines...
- Kernel matrix is information „bottleneck"
  => importance of kernel choice!

Importance of Prior Knowledge

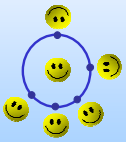
- Without prior assumptions no generalization possible
  There is no best classification method („No Free Lunch Thm.")
  There is no best feature representation („Ugly Duckling Thm.")
- Equivalence of training data and model complexity



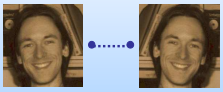
- Consequences of modelling more prior knowledge
  - less training samples required
  - better generalization in case of equal number of samples
  - feature representation or classification can be simplified

Invariance as Prior Knowledge

- transformation knowledge
  i.e. samples maintain inherent meaning under certain transformations
- continuous



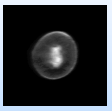
- discrete



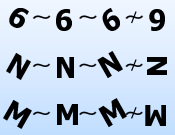
- continuous and discrete

Transformation Examples

- complete group

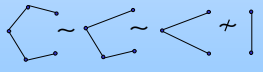


- subset of group



- reversible

- irreversible

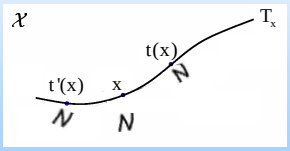


Transformation Formalization

- Patterns x stem from pattern space $\mathcal{X}$
- Transformations $t : \mathcal{X} \to \mathcal{X}$
- Set of transformations T
- Set of transformed patterns of x

$$T_x := \{t(x) \mid t \in T\}$$

= „similar" meaning as x



- Total invariance: $k(x, x') = k(t(x), t'(x'))$

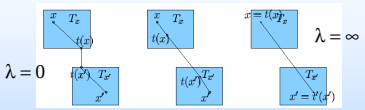
Specific Goals

- Wanted properties for general approach:
  - support for various kernel methods, not only SVM
  - support for many kernels, not only Gaussian
  - support for infinite set of transformations, exponentially many transformation combinations
  - support for discrete, continuous, group and non-group transformations
  - Applicability, good memory, time and model complexities
  - Adjustability of degree of invariance

Invariant Distance Substitution Kernels

Invariant Distances

- Distances of $T_x, T_{x'}$ better than $d(x, x')$



- Invariant Distances
  - two-sided distance

$$d_{2S}(x, x') := \min_{t,t'} \left(d(t(x), t'(x')) + \lambda \Omega(t, t') \right)$$

  - cost function

$$\Omega(t, t') \geq 0, \quad \Omega(t, t') = 0 \Leftrightarrow t = t' = \mathrm{id}$$

  - λ is regularization parameter
  - similar: one-sided distance + symmetrization

Distance Substitution Kernels [HB04]

- Distance d: symmetric, nonnegative, zero-diagonal
- Distance-based kernels:

$$k\left(\|x-x'\|\right) \Rightarrow k_d(x,x') := k\left(d(x,x')\right)$$

- Inner-product case:

$$k\left(\langle x,x'\rangle\right) \Rightarrow k_d(x,x') := k\left(\langle x,x'\rangle_d^O\right)$$

where O is an arbitrary origin and

$$\langle x,x'\rangle_d^O := -\frac{1}{2}\left(d(x,x')^2 - d(x,O)^2 - d(x',O)^2\right)$$

- Examples:

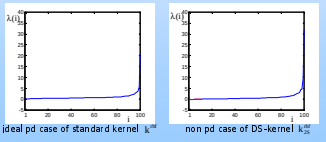
$$k_d^{lin}(x,x') := \langle x,x'\rangle_d^O \qquad k_d^{nd}(x,x') := -d(x,x')^\beta, \beta \in [0,2]$$

$$k_d^{pol}(x,x') := \left(1+\gamma\langle x,x'\rangle_d^O\right)^p \qquad k_d^{rbf}(x,x') := e^{-\gamma d(x,x')^2}, p \in \mathbb{N}, \gamma \geq 0$$

- Expectation: Similar behaviour as standard kernels

Positive Definiteness

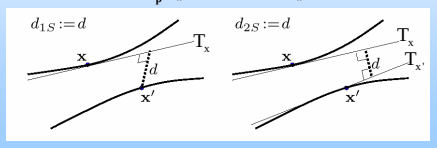
- Equivalent conditions:
  - k_d^{nd} is cpd for all $\beta \in [0,2]$
  - k_d^{rbf} is pd for all $\gamma \geq 0$
  - d is Hilbertian metric
  - k_d^{lin} is pd
  - k_d^{pol} is pd for all $\gamma \geq 0, p \in \mathbb{N}$
- DS-kernels mostly non-(c)pd:
  counterexample by invalid Δ-inequality
- Empirical observations:
  - only weak pd-ness violation on real world data,



Tangent Distance Kernels

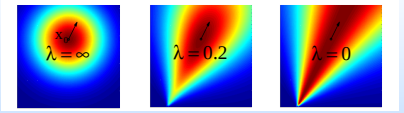
- Assumption: differentiable transformations
- Tangent Distance: local linear approximation [SLD93]
  - distance of tangent spaces instead of manifolds
  - symmetric two-sided distance $d_{2S}(x,x')$
  - nonsymmetric one-sided distance $d_{1S}(x,x')$

$$d_{1S}(x,x') = \min_p \left\| x + \sum_i p_i t_i - x' \right\|^2 + \lambda \|p\|^2$$

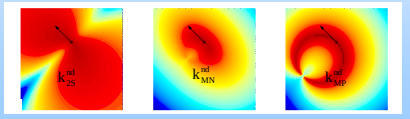


Invariance in 2D

- Scaling: $k_{MP}^{rbf}(x_0,x)$ varying x



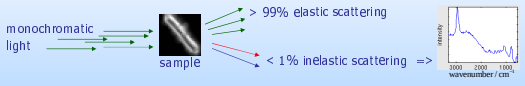
  - invariance can be adjusted by regularization λ
- Rotation:



  - More than linear invariances can be captured

Application Raman Spectra

- Project OMIB in BMBF framework „Biophotonics"
  - detection of clean room contaminations by Raman micro-spectroscopy
- Raman spectroscopy

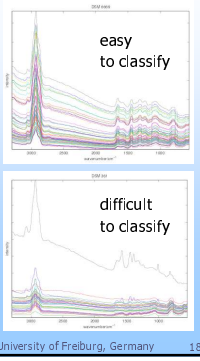


  - Use of the „Raman effect", inelastic scattering of light
  - Distribution of energy differences => Raman spectrum
  - non-destructive method, fingerprint of chemical bindings
- Pattern variations due to
  - measurement duration, background radiation
  - thickness of sample, heterogeneity, growing time, nutrition condition, temperature, photo bleaching

Application Raman Spectra

- Details on dataset [R&al05]:
  - 2545 spectra, 1833 features
  - 20 classes,
  - unequally distributed

Name	Code	Class Number	Number of Samples
B. pumilus	DSM 27	1	57
B. pumilus	DSM 361	5	43
B. sphaericus	DSM 28	2	53
B. sphaericus	DSM 396	6	42
B. subtilis subsp. subtilis	DSM 10	0	306
B. subtilis subsp. spizizenii	DSM 347	3	42
E. coli	DSM 423	7	51
E. coli	DSM 498	8	21
E. coli	DSM 499	9	20
M. luteus	DSM 348	4	619
M. luteus	DSM 20030	13	48
M. lylae	DSM 20315	16	20
M. lylae	DSM 20318	18	20
S. cohnii subsp. cohnii	DSM 6669	10	67
S. cohnii subsp. cohnii	DSM 20260	15	65
S. cohnii subsp. urealyticum	DSM 6718	11	65
S. cohnii subsp. urealyticum	DSM 6719	12	63
S. Epidermidis	DSM RP 62A	19	805
S. Warneri	DSM 20036	14	67
S. Warneri	DSM 20316	17	71

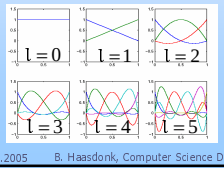


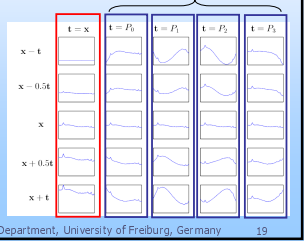
Application Raman Spectra

- Details on tangents:
  - intensity scale: scale tangent $t = x$
  - baseline shift:
    Lagrange polynomials of degree l : $t = P_m$

$$P_m(x) = \prod_{i \neq m} \frac{\left(x - \frac{i}{l}\right)}{\left(\frac{m}{l} - \frac{i}{l}\right)}$$

$$m = 0, \ldots, l$$





Application Raman Spectra

- Recognition results
  - LIBSVMTL, one-vs-one SVM, k_{2S}^{rbf}
  - Simultaneous scaling and Lagrange tangents
  - 8x8 grid-search for (C, γ)
  - LOO and class-wise average LOO minimization

	LOO-error [%], varying degree l					av-LOO-error [%], varying degree l				
reg.par λ	0	1	2	3	4	0	1	2	3	4
1.000000	4.72	4.56	4.56	4.44	4.28	10.71	10.33	10.20	9.83	9.54
0.100000	4.48	4.44	4.56	4.72	4.48	10.16	9.63	9.85	10.49	9.76
0.010000	4.05	4.09	3.85	3.93	3.97	9.06	9.18	8.53	8.71	9.49
0.001000	**3.65**	3.69	3.77	3.77	3.73	**8.61**	8.46	8.55	8.65	8.50
0.000100	3.85	**3.03**	**2.99**	**3.18**	**3.18**	8.87	**6.87**	**7.26**	**7.48**	**7.32**
0.000010	5.78	3.65	3.34	3.38	3.65	12.70	8.09	7.46	7.68	8.26
0.000001	11.43	7.27	6.05	5.38	5.38	26.96	16.68	14.17	12.86	12.51
base SVM	4.20					9.58				

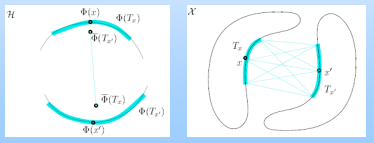
  - LOO and av-LOO clearly improved by tangents
  - Regularization required due to scaling tangent

Transformation Integration Kernels

Transformation Integration Kernels

- Motivation:
  - Maintain pd-ness by avoiding min/max-operations
  - Extend Haar-integration framework [S94] to kernels
  - Adjustability of degree of invariance
- Definition of TI-Kernels [HVB05]:

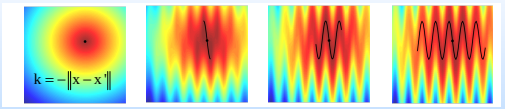
$$k_{TI}(x, x') := \left\langle \int_T \Phi(t(x))dt, \int_T \Phi(t'(x'))dt' \right\rangle = \iint_{T \times T} k(t(x), t'(x'))dtdt'$$



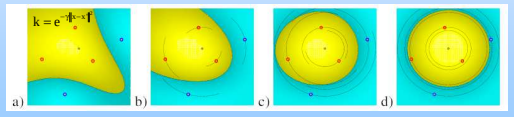
- (c)pd-ness transfered from base-kernel

Invariance in 2D

- Invariance adjustability:
  highly nonlinear sinus shifts



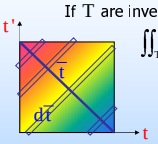
- Effect in SVM: rotation



- Similar: Total invariance of linear, polynomial kernels

Acceleration

- Single Kernel Evaluation: Integral Reduction (IR)
  If T are invertible and compatible with k



$$\iint_{T \times T} k(t(x), t'(x'))dtdt' = \iint_{T \times T} k(x, t^{-1} \circ t'(x'))dtdt'$$

$$= \int_{T^{-1} \circ T} k(x, \bar{t}(x'))d\bar{t}$$

=> squareroot complexity reduction

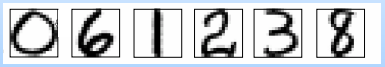
- SVM-Training: SV-extraction (SV)
  - Perform initial ordinary SVM-training
  - Extract SVs
  - Train invariant SVM on the SVs

=> linear complexity reduction

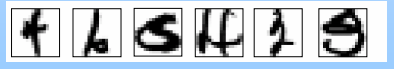
Application USPS-Digits

- Details on dataset:
  - standard benchmark dataset
  - 7291 training-, 2007 testpatterns of handwritten digits given by 16x16 grayvalue bitmaps
  - samples:

easy to classify



difficult to classify



Application USPS-Digits

- SVM Recognition Results:

increasing rotation integration range 

ϕ-range [rad]	k^{rbf} test error [%]
0	4.5
$\pm 0.04\pi$	4.1
$\pm 0.08\pi$	4.2
$\pm 0.12\pi$	3.9
$\pm 0.16\pi$	4.2

increasing x-y translation integration range 

x-y-range [pixels]	k^{rbf} test error [%]
0	4.5
± 1	3.7
± 2	3.2
± 3	3.3
± 4	3.2

- Both TI-kernels superior over base kernel
- x-y-integration better than rotations
- TI-kernels yield state-of-the-art result as VSV [SBV96]

Application USPS-Digits

- Complexity comparison to VSV:
  - shared non-optimized parameters,

x-y-translation, +-2 pixels, 3x3 shifts per sample

Method	test-error [%]	train-time [sec]	test-time [sec]	average #SV
TI-SVM	3.6	1771	479	412
TI-SVM, IR	3.6	810	176	412
TI-SVM, SV	3.6	113 + 130 + 297	466	410
TI-SVM, SV + IR	3.6	113 + 130 + 91	172	410
VSV-SVM	3.5	113 + 864 + 1925	177	4240

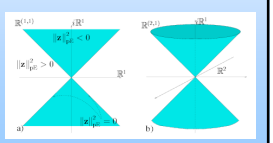
  - TI-kernels produce small models
  - No recognition degradation by SV-acceleration
  - Acceleration techniques clearly successful
  - Accelerated TI-SVM consistently faster than VSV

Indefinite Kernels in SVM

Pseudo-Euclidean Spaces [G85]

- Real finite dimensional vector spaces
  $\mathbb{R}^{(p,q)} := \mathbb{R}^p \times i\mathbb{R}^q$ of signature (p,q)
- symmetric (indefinite) inner-product
  $\langle \mathbf{z}, \mathbf{z}' \rangle_{pE} := \mathbf{z}_p^T \mathbf{z}'_p - \mathbf{z}_q^T \mathbf{z}'_q = \mathbf{z}^T \mathbf{M} \mathbf{z}' \quad \mathbf{M} = \mathrm{diag}(\mathbf{1}_p, -\mathbf{1}_q)$
- squared norm
  $\|\mathbf{z}\|_{pE}^2 := \langle \mathbf{z}, \mathbf{z} \rangle_{pE} = \mathbf{z}^T \mathbf{M} \mathbf{z}$
- squared distance
  $\|\mathbf{z} - \mathbf{z}'\|_{pE}^2 := \langle \mathbf{z} - \mathbf{z}', \mathbf{z} - \mathbf{z}' \rangle_{pE}$
- orthogonality
  $\langle \mathbf{z}, \mathbf{z}' \rangle_{pE} = \mathbf{z}^T \mathbf{M} \mathbf{z}' = 0$
- hyperplanes
  $\mathcal{H} : \langle \mathbf{z}, \mathbf{w} \rangle_{pE} + b = 0$

can be negative:



pE Feature Space Embedding

- Data dependent pE-embedding:

Given data $\{x_i\}_{i=1}^n$ + sym. kernel k $\Longrightarrow$ Existence of pE space $\mathbb{R}^{(p,q)}$ + embedding $\Phi : \mathcal{X} \to \mathbb{R}^{(p,q)}$

  - Representation of kernel
    $k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_{pE}$
- Construction by Eigendecomposition [LM04, GHBO99, PPD01]:

$\mathbf{K} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T \qquad p := \dim(\lambda^+), q := \dim(\lambda^-)$

$\mathbf{\Lambda} = \mathrm{diag}(\lambda^+, \lambda^-) \qquad \Phi(x_i) := \left(\sqrt{|\Lambda|}\mathbf{U}^T\right)_i$

Geometrical Interpretation of Kernel Operations

- Norm of Feature vector
  $$k(x_i, x_i) = \|\Phi(x_i)\|^2_{pE}$$
- Kernel induced distance
  $$d^2(x_i, x_j) := k(x_i, x_i) - 2k(x_j, x_j) + k(x_j, x_j)$$
  $$d^2(x_i, x_j) = \|\Phi(x_i) - \Phi(x_j)\|^2_{pE}$$
- Linear combinations
  $$\sum \alpha_i k(x_i, x) = \left\langle \sum \alpha_i \Phi(x_i), x \right\rangle_{pE}$$
- Centering of Kernel Matrix
  $$\mathbf{K}_C = \mathbf{J} \cdot \mathbf{K} \cdot \mathbf{J} \quad \mathbf{J} := \mathbf{I} - \mathbf{1} \cdot \mathbf{1}^T / n \quad \Rightarrow \quad \sum \Phi_C(x_i) = \mathbf{0}$$
- Projections, variance analysis, Eigendecompositions, Optimization problems ... general kernel methods?

Capacity Estimate

- VC-bound for pE hyperplanes in $\mathbb{R}^{(p,q)}$
  - data embedded s.th. $r^2 \|\Phi(x_i)\|^2_E \le \|\Phi(x_i)\|^2_{pE} \le R^2$
  - $\mathcal{F} := \left\{ \operatorname{sgn} \langle \mathbf{w}, \mathbf{z} \rangle_{pE} \right\}$ set of hyperplanes, which are canonical wrt the data and satisfy
    $$\lambda^2 \|\mathbf{w}\|^2_E \le \|\mathbf{w}\|^2_{pE} \le \Lambda^2$$
    then holds
    $$h(\mathcal{F}) \le \left(\frac{R\Lambda}{r\lambda} \right)^2$$
- Proof: Vapnik [V95] + modif. Cauchy Schwartz.
- $\Rightarrow$ SRM: minimize $\|\mathbf{w}\|^2_{pE} = \mathbf{w}^T \mathbf{M} \mathbf{w}$ while maintaining strict positivity

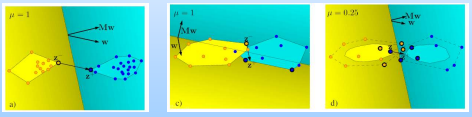
Optimal Separation of Convex Hulls

[H05, extension of BB00,CB00]

- Formalization

  train:
  $$\min_{z^+, z^-} \|z^- - z^+\|^2_{pE}$$
  $$\text{s.t. } z^\pm \in \text{conv}_\mu \{\Phi(x_i) \mid y_i = \pm 1\}$$

  classify: sign of
  $$f(x) = \|\Phi(x) - z^-\|^2_{pE} - \|\Phi(x) - z^+\|^2_{pE}$$
- Separable case
- Non-separable case



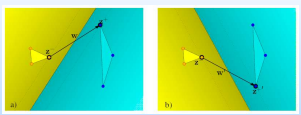
- Training and test avoid explicit pE-embedding
- Reasonable solution: $\mathbf{w}^T \mathbf{M} \mathbf{w} \ge 0$
- Optimum exists and obtained in span$\{\Phi(x_i)\}$

Indefinite SVM in pE Space

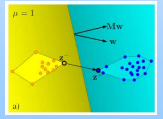
- SVM-"primal":
  $$\min_{w,b} \ \tfrac{1}{2} \mathbf{w}^T \mathbf{M} \mathbf{w}$$
  $$\text{s.t.} \ \ y_i \left(\mathbf{w}^T \mathbf{M} \Phi(x_i) + b \right) \ge 1$$
  - Usual interpretation of SV + coefficients
  - Non-SV are correctly classified,
  - Only bounded SV can be misclassified
- SVM is CH-classifier:
  - Feasible, stationary points + local optima transfer between SVM and CH, if $\mathbf{w}^T \mathbf{M} \mathbf{w} \ge 0$
  - E.g Nonzero SVM-solution $\boldsymbol{\alpha}$ => CH-solution $\bar{\boldsymbol{\alpha}}$
    $$\bar{\boldsymbol{\alpha}} := 2\boldsymbol{\alpha} / \textstyle\sum_i \alpha_i \qquad \mu := 2C / \textstyle\sum_i \alpha_i$$
  - CH and SVM hyperplanes are parallel, even identical if coefficients are not bounded

Numerics

- Convergence to stationary point, libsvm [LL03]
- Multiple solutions

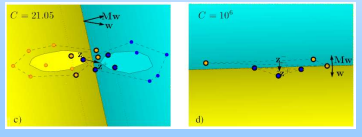


- Uniqueness in extreme indefinite cases



Practical Criteria for Indefinite SVM

- Criterion for suitability: #bSV
  - No (few) bounded $\alpha_i \Rightarrow$ no (few) training errors
- Criterion for unsuitability: $\mathbf{w}^T \mathbf{M} \mathbf{w} \le \mathbf{0}$
  - after training: $\sum_{i,j} \alpha_i \alpha_j y_i y_j k(x_i, x_j)$
  - before training:
    signature (p,q)
- Criterion for suitability: Distance of Class Means
  - If DCM is positive, sufficiently low C yields solution



$$DCM^2 = \sum_{i,j} c_i c_j k(x_i, x_j)$$
$$c_i = \begin{cases} 1/n^+ \text{ for } y_i = +1 \\ -1/n^- \text{ for } y_i = -1 \end{cases}$$

Summary and Perspectives

Summary

- Principle „Incorporation of transformation knowledge in kernels“ for improving kernel methods
- General framework I: IDS-Kernels
  - Distance substitution kernels with invariant distances
  - Application: Tangent Distance Kernels on Spectra
  - Efficient for multiple invariances, but indefiniteness
- General framework II: TI-Kernels
  - Integration over transformations
  - Application: Offline HWR (USPS digits)
  - Efficient for few invariances, but positive definiteness
- Indefinite SVM
  - Constructive geometric interpretation: pE CH-separation
  - Convergent implementation required
  - Practical criteria for suitability

Perspectives

- Invariance
  - combination with invariant representations
  - Learning of transformation directions
- Kernel design
  - Relaxing the pd-ness of kernels yields much wider flexibility
  - Applications with proximity data: nonlinear analysis!
- Kernel methods
  - Acceptance/robustness against indefinite kernels?
  - Interpretation in pE-spaces as basis for geometrical/numerical/statistical analysis and new methods
- Non-convex optimization
  - Large scale implementations with convergence statements
  - number and quality of solutions.

References

[BB00] K.P. Bennett, E.J.Bredensteiner. „Duality and geometry in SVM classifiers“, Proc. 17th ICML, pp. 57-64, 2000.
[CB00] D.J. Crisp, C.J.C. Burges. „A geometric interpretation on nu-svm classifiers“, NIPS 12, pp.223-229, MIT-Press, 2000.
[G85] L. Goldfarb. „A new approach to pattern recognition“, Progress in Pattern Recognition 2, pp. 241-402, Elsevier, 1985.
[GHBO99] T. Graepel, R. Herbrich, P. Bollmann-Sdorra,K. Obermayer. „Classification on pariwise proximity data“, NIPS'11, pp. 438-444, MIT Press, 1999.
[HB04] B. Haasdonk, C. Bahlmann. „Learning with Distance Substitution Kernels“, Proc. 26th DAGM, pp. 220-227, Springer, 2004.
[H05] B. Haasdonk. „Feature space interpretation of SVMs with indefinite kernels“, IEEE TPAMI, 27(4):482-492, april 2005.
[HVB05] B. Haasdonk, A. Vossen, H. Burkhardt. „Invariance in Kernel Methods by Haar-Integration Kernels“. Proc. 12th SCIA, 2005. submitted.
[LM04] J. Laub, K.-R. Müller. „Faeture Discoveryin non-metric pairwise data“ Journal of Machine Learning Research, 5:810-818,2004.
[LL03] H.-T. Lin, C.-J. Lin. „A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods.“ Technical Report, National Taiwan University, March 2003.
[PPD01] E. Pekalska, P. Paclik, R. Duin. „A generalized kernel approach to dissimilarity based classification“ Journal of Machine Learning Research, 2:175-211, 2001.
[R&al05] P. Rösch, M. Schmitt, K.-D. Peschke, O. Ronneberger, H. Burkhardt, H.-W. Motzkus, M- Lankers, S. Hofer, H. Thiele, J. Popp. „Chemotaxonomic identification of single bacteria relevant for clean room production by micro-Raman spectroscopy“, Applied and Environmental Microbiology, 2005. to appear.
[SS02] B. Schölkopf, A. Smola. „Learning with Kernels“, MIT-Press, 2002.
[S94] H. Schulz-Mirbach. „Constructing invariant features by averaging techniques“, Proc. 12th ICPR, IEEE, 1994
[SC04] J. Shawe-Taylor, N. Cristianini. „Kernel Methods for Pattern Analysis“, Cambridge University Press, 2004.
[SLD93] P.Y.Simard, Y.A. LeCun, J.S. Denker. „Efficient Pattern Recognition using a new transformation distance“, NIPS 5, pp. 50-58, Morgan Kaufman 1993.
[V95] V. Vapnik. „The nature of statistical learning theory“, Springer, New York, 1995.

Thank You!

Any questions?