Classification with Invariant Distance Substitution Kernels

Bernard Haasdonk
Institute of Mathematics
University of Freiburg
Hermann-Herder-Str. 10
79104 Freiburg, Germany
haasdonk@mathematik.uni-freiburg.de

Hans Burkhardt
Institute of Computer Science
University of Freiburg
Georges-Köhler-Allee 52
79110 Freiburg, Germany
burkhardt@informatik.uni-freiburg.de

Abstract

Kernel methods offer a flexible toolbox for pattern analysis and machine learning. A general class of kernel functions which incorporates known pattern invariances are *invariant distance substitution (IDS)* kernels. Instances such as tangent distance or dynamic time-warping kernels have demonstrated the real world applicability. This motivates the demand for investigating the elementary properties of the general IDS-kernels. In this paper we formally state and demonstrate their invariance properties, in particular the adjustability of the invariance in two conceptionally different ways. We characterize the definiteness of the kernels. We apply the kernels in different classification methods, which demonstrates various benefits of invariance.

1 Introduction

Kernel methods have gained large popularity in the pattern recognition and machine learning communities due to the modularity of the algorithms and the data representations by kernel functions, cf. (Schölkopf and Smola (2002)) and (Shawe-Taylor and Cristianini (2004)). It is well known that prior knowledge of a problem at hand must be incorporated in the solution to improve the generalization results. We address a general class of kernel functions called IDS-kernels (Haasdonk and Burkhardt (2007)) which incorporates prior knowledge given by pattern invariances.

The contribution of the current study is a detailed formalization of their basic properties. We both formally characterize and illustratively demonstrate their adjustable invariance properties

[0] Copyright notice: Permission for this electronic version granted by Springer. The article is to appear in: Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme, Reinhold Decker (eds.): Data Analysis, Machine Learning, and Applications. Proc. of the 31th Annual Conference of the German Classification Society (GfKl), University of Freiburg, March 7-9, 2007. Springer-Verlag, Heidelberg-Berlin, 2008.

in Sec. 3. We formalize the definiteness properties in detail in Sec. 4. The wide applicability of the kernels is demonstrated in different classification methods in Sec. 5.

2 Background

Kernel methods are general nonlinear analysis methods such as the *kernel principal component analysis*, *support vector machine*, *kernel perceptron*, *kernel Fisher discriminant*, etc. (Schölkopf and Smola (2002)) and (Shawe-Taylor and Cristianini (2004)). The main ingredient in these methods is the kernel as a similarity measure between pairs of patterns from the set $\mathcal{X}$.

Definition 1 (Kernel, Definiteness). *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which is symmetric is called a* kernel*. A kernel k is called* positive definite (pd)*, if for all n and all sets of observations $(x_i)_{i=1}^n \in \mathcal{X}^n$ the kernel matrix $\mathbf{K} := (k(x_i, x_j))_{i,j=1}^n$ satisfies $\mathbf{v}^T\mathbf{K}\mathbf{v} \geq 0$ for all $\mathbf{v} \in \mathbb{R}^n$. If this only holds for all $\mathbf{v}$ satisfying $\mathbf{v}^T\mathbf{1} = 0$, the kernel is called* conditionally positive definite (cpd)*.*

We denote some particular l^2-inner-product $\langle \cdot, \cdot \rangle$ and l^2-distance $\|\cdot - \cdot\|$ based kernels by $k^{\text{lin}}(\mathbf{x}, \mathbf{x}') := \langle \mathbf{x}, \mathbf{x}' \rangle$, $k^{\text{nd}}(\mathbf{x}, \mathbf{x}') := -\|\mathbf{x} - \mathbf{x}'\|^\beta$ for $\beta \in [0, 2]$, $k^{\text{pol}}(\mathbf{x}, \mathbf{x}') := (1 + \gamma \langle \mathbf{x}, \mathbf{x}' \rangle)^p$, $k^{\text{rbf}}(\mathbf{x}, \mathbf{x}') := e^{-\gamma\|\mathbf{x}-\mathbf{x}'\|^2}$ for $p \in \mathbb{N}, \gamma \in \mathbb{R}_+$. Here, the linear k^{lin}, polynomial k^{pol} and Gaussian radial basis function (rbf) k^{rbf} are pd for the given parameter ranges. The negative distance kernel k^{nd} is cpd (Schölkopf and Smola (2002)). We continue with formalizing the prior knowledge about pattern variations and corresponding notation:

Definition 2 (Transformation Knowledge). *We assume to have transformation knowledge for a given task, i.e. the knowledge of a set $T = \{t : \mathcal{X} \to \mathcal{X}\}$ of transformations of the object space including the identity, i.e.* id $\in T$*. We denote the set of transformed patterns of $x \in \mathcal{X}$ as $T_x := \{t(x) | t \in T\}$ which are assumed to have identical or similar inherent meaning as x.*

The set of concatenations of transformations from two sets T, T' is denoted as $T \circ T'$. The n-fold concatenation of transformations t are denoted as $t^{n+1} := t \circ t^n$, the corresponding sets denoted as $T^{n+1} := T \circ T^n$. If all $t \in T$ are invertible, we denote the set of inverted functions as T^{-1}. We denote the semigroup of transformations generated by T as $\bar{T} := \bigcup_{n \in \mathbb{N}} T^n$. The set $\bar{T}$ induces an equivalence relation on $\mathcal{X}$ by $x \sim x' :\Leftrightarrow$ there exist $\bar{t}, \bar{t}' \in \bar{T}$ such that $\bar{t}(x) = \bar{t}'(x')$. The equivalence class of x is denoted with E_x and the set of all equivalence sets is $\mathcal{X}/_\sim$.

Learning targets can often be modelled as functions of several input objects, for instance depending on the training data and the data for which predictions are required. We define the desired notion of invariance:

Definition 3 (Total Invariance). *We call a function $f : \mathcal{X}^n \to \mathcal{H}$* totally invariant *with respect to T, if for all patterns $x_1, \ldots, x_n \in \mathcal{X}$ and transformations $t_1, \ldots, t_n \in T$ holds $f(x_1, \ldots, x_n) = f(t_1(x_1), \ldots, t_n(x_n))$.*

As the IDS-kernels are based on distances, we define:

Definition 4 (Distance, Hilbertian Metric). *A function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a* distance*, if it is symmetric and nonnegative and has zero diagonal, i.e. $d(x, x) = 0$. A distance is a* Hilbertian metric *if there exists an embedding into a Hilbert space $\Phi : \mathcal{X} \to \mathcal{H}$ such that $d(x, x') = \|\Phi(x) - \Phi(x')\|$.*

So in particular the triangle inequality does not need to be valid for a distance function in this sense. Note also that a Hilbertian metric can still allow $d(x, x') = 0$ for $x \neq x'$.

Assuming some distance function d on the space of patterns $\mathcal{X}$ enables to incorporate the invariance knowledge given by the transformations T into a new dissimilarity measure.

Definition 5 (Two-Sided Invariant Distance)**.** *For a given distance d on the set $\mathcal{X}$ and some cost function $\Omega : T \times T \to \mathbb{R}_+$ with $\Omega(t,t') = 0 \Leftrightarrow t = t' = \mathrm{id}$, we define the* two-sided invariant distance *as*

$$d_{2S}(x,x') := \inf_{t,t' \in T} d(t(x), t'(x')) + \lambda\Omega(t,t'). \tag{1}$$

For $\lambda = 0$ the distance is called *unregularized.* In the following we exclude artificial degenerate cases and reasonably assume that $\lim_{\lambda\to\infty} d_{2S}(x,x') = d(x,x')$ for all x, x'. The requirement of precise invariance is often too strict for practical problems. The points within T_x are sometimes not to be regarded as identical to x, but only as similar, where the similarity can even vary over T_x. An intuitive example is optical character recognition, where the similarity of a letter and its rotated version is decreasing with growing rotation angle. This approximate invariance can be realized with IDS-kernels by choosing $\lambda > 0$.

With the notion of invariant distance we define the *invariant distance substitution kernels* as follows:

Definition 6 (IDS-Kernels)**.** *For a* distance-based kernel*, i.e.* $k(\mathbf{x},\mathbf{x}') = f(\|\mathbf{x} - \mathbf{x}'\|)$*, and the invariant distance measure d_{2S} we call $k_{IDS}(x,x') := f(d_{2S}(x,x'))$ its* invariant distance substitution kernel *(IDS-kernel). Similarly, for an* inner-product-based *kernel k, i.e.* $k(\mathbf{x},\mathbf{x}') = f(\langle \mathbf{x},\mathbf{x}' \rangle)$*, we call $k_{IDS}(x,x') := f(\langle x,x' \rangle^O)$ its IDS-kernel, where $O \in \mathcal{X}$ is an arbitrary* origin *and a generalization of the inner product is given by* $\langle x,x' \rangle^O := -\frac{1}{2}(d_{2S}(x,x')^2 - d_{2S}(x,O)^2 - d_{2S}(x',O)^2)$.

The IDS-kernels capture existing approaches such as tangent distance or dynamic time-warping kernels which indicates the real world applicability, cf. (Haasdonk (2005)) and (Haasdonk and Burkhardt (2007)) and the references therein.

Crucial for efficient computation of the kernels is to avoid explicit pattern transformations by using or assuming some additional structure on T. An important computational benefit of the IDS-kernels must be mentioned, which is the possibility to precompute the distance matrices. By this, the final kernel evaluation is very cheap and ordinary fast model selection by varying kernel or training parameters can be performed.

3 Adjustable Invariance

As first elementary property, we address the invariance. The IDS-kernels offer two possibilities for controlling the transformation extent and thereby interpolating between the invariant and non-invariant case. Firstly, the size of T can be adjusted. Secondly, the regularization parameter λ can be increased to reduce the invariance. This is summarized in the following:

Proposition 7 (Invariance of IDS-Kernels)**.**

i) If $T = \{\mathrm{id}\}$ and d is an arbitrary distance, then $k_{IDS} = k$.

ii) If all $t \in T$ are invertible, then distance-based unregularized IDS-kernels $k_{IDS}(\cdot, x)$ are constant on $(T^{-1} \circ T)_x$.

iii) If $T = \bar{T}$ and $\bar{T}^{-1} = \bar{T}$, then unregularized IDS-kernels are totally invariant with respect to $\bar{T}$.

iv) If d is the ordinary Euclidean distance, then $\lim_{\lambda\to\infty} k_{IDS} = k$.

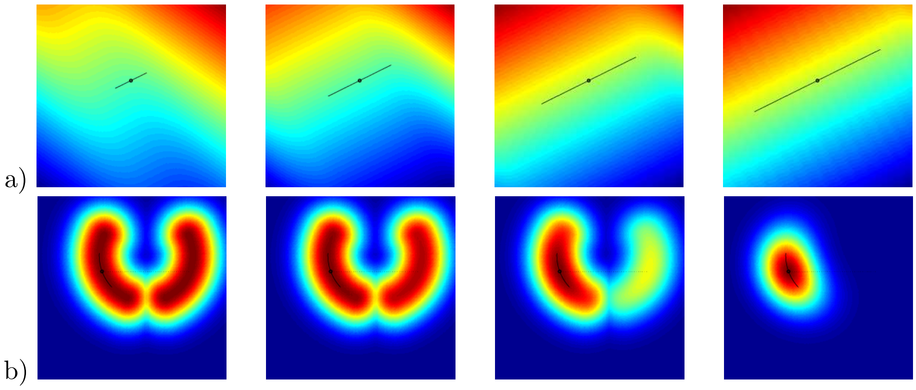

Figure 1: Adjustable invariance of IDS-kernels. a) Linear kernel k_{IDS}^{lin} with invariance wrt. linear shifts, adjustability by increasing transformation extent by the set T, $\lambda = 0$, b) kernel k_{IDS}^{rbf} with combined nonlinear and discrete transformations, adjustability by increasing regularization parameter λ.

Proof. Statement i) is obvious from the definition, as $d_{2S} = d$ in this case. Similarly, iv) follows as $\lim_{\lambda\to\infty} d_{2S} = d$. For statement ii), we note that if $x' \in (T^{-1} \circ T)_x$, then there exist transformations $t, t' \in T$ such that $t(x) = t'(x')$ and consequently $d_{2S}(x, x') = 0$. So any distance-based kernel k_{IDS} is constant on this set $(T^{-1} \circ T)_x$. For proving iii) we observe that for $\bar{t}, \bar{t}' \in \bar{T}$ holds $d_{2S}(\bar{t}(x), \bar{t}'(x')) = \inf_{t,t'} d(t(\bar{t}(x)), t'(\bar{t}'(x'))) \geq \inf_{t,t'} d(t(x), t'(x')) = d_{2S}(x, x')$. Using the same argumentation with $\bar{t}(x)$ for x, $\bar{t}^{-1}$ for $\bar{t}$ and similar replacements for $x', \bar{t}'$ yields $d_{2S}(x, x') \geq d_{2S}(\bar{t}(x), \bar{t}'(x'))$, which gives the total invariance of d_{2S} and thus for all unregularized IDS-kernels. □

Points i) to iii) imply that the invariance can be adjusted by the size of T. Point ii) implies that the invariance occasionally exceeds the set T_x. If for instance T is closed with respect to inversions, i.e. $T = T^{-1}$, then the set of constant values is $(T^2)_x$. Point iii) and iv) indicate that λ can be used to interpolate between the full invariant and non-invariant case.

We give simple illustrations of the proposed kernels and these adjustability mechanisms in Fig. 1. For the illustrations, our objects are simply points in two dimensions and several transformations define sets of points to be regarded as similar. We fix one argument $\mathbf{x}'$ (denoted with a black dot) of the kernel, and the other argument $\mathbf{x}$ is varying over the square $[-1, 2]^2$ in the Euclidean plane. We plot the different resulting kernel values $k(\mathbf{x}, \mathbf{x}')$ in gray-shades. All plots generated in the sequel can be reproduced by the MATLAB library *KerMet-Tools* (Haasdonk (2005)).

In Fig. 1 a) we focus on a linear shift along a certain slant direction while increasing the transformation extent, i.e. the size of T. The figure demonstrates the behaviour of the linear unregularized IDS-kernel, which perfectly aligns to the transformation direction as claimed by Prop. 7 i) to iii). It is striking that the captured transformation range is indeed much larger than T and very accurate for the IDS-kernels as promised by Prop. 7 ii).

The second means for controlling the transformation extent, namely increasing the regularization parameter λ, is also applicable for discrete transformations such as reflections and even in combination with continuous transformations such as rotations, cf. Fig. 1 b). We see that the interpolation between the invariant and non-invariant case as claimed in Prop. 7 ii) and iv) is nicely realized. So the approach is indeed very general concerning types of transformations, comprising discrete, continuous, linear, nonlinear transformations and combinations thereof.

4 Positive Definiteness

The second elementary property of interest, the positive definiteness of the kernels, can be characterized as follows by applying a finding from (Haasdonk and Bahlmann (2004)):

Proposition 8 (Definiteness of Simple IDS-Kernels)**.** *The following statements are equivalent:* *i)* d_{2S} *is a Hilbertian metric*

ii) k^{nd}_{IDS} *is cpd for all* $\beta \in [0,2]$ *iii)* k^{lin}_{IDS} *is pd*

iv) k^{rbf}_{IDS} *is pd for all* $\gamma \in \mathbb{R}_+$ *v)* k^{pol}_{IDS} *is pd for all* $p \in \mathbb{N}, \gamma \in \mathbb{R}_+$.

So the crucial property, which determines the (c)pd-ness of IDS-kernels is, whether the d_{2S} is a Hilbertian metric. A practical criterion for disproving this is a violation of the triangle inequality. A precise characterization for d_{2S} being a Hilbertian metric is obtained from the following.

Proposition 9 (Characterization of d_{2S} as Hilbertian Metric)**.** *The unregularized* d_{2S} *is a Hilbertian metric if and only if* d_{2S} *is totally invariant with respect to* $\bar{T}$ *and* d_{2S} *induces a Hilbertian metric on* $\mathcal{X}/_\sim$.

Proof. Let d_{2S} be a Hilbertian metric, i.e. $d_{2S}(x,x') = \|\Phi(x) - \Phi(x')\|$. For proving the total invariance wrt. $\bar{T}$ it is sufficient to prove the total invariance wrt. T due to transitivity. Assuming that for some choice of patterns/transformations holds $d_{2S}(x,x') \neq d_{2S}(t(x), t'(x'))$ a contradiction can be derived: Note that $d_{2S}(t(x), x')$ differs from one of both sides of the inequality, without loss of generality the left one, and assume $d_{2S}(x,x') < d_{2S}(t(x),x')$. The definition of the two-sided distance implies $d_{2S}(x, t(x)) = \inf_{t',t''} d(t'(x), t''(t(x))) = 0$ via $t' := t$ and $t'' = \text{id}$. By the triangle inequality, this gives the desired contradiction $d_{2S}(x,x') < d_{2S}(t(x),x') \leq d_{2S}(t(x),x) + d_{2S}(x,x') = 0 + d_{2S}(x,x')$. Based on the total invariance, $d_{2S}(\cdot, x'')$ is constant on each $E \in \mathcal{X}/_\sim$: For all $x \sim x'$ transformations $\bar{t}, \bar{t}'$ exist such that $\bar{t}(x) = \bar{t}'(x')$. So we have $d_{2S}(x,x'') = d_{2S}(\bar{t}(x), x'') = d_{2S}(\bar{t}'(x'), x'') = d_{2S}(x', x'')$, i.e. this induces a well defined function on $\mathcal{X}/_\sim$ by $\bar{d}_{2S}(E,E') := d_{2S}(x(E), x(E'))$. Here $x(E)$ denotes one representative from the equivalence class $E \in \mathcal{X}/_\sim$. Obviously, $\bar{d}_{2S}$ is a Hilbertian metric. via $\bar{\Phi}(E) := \Phi(x(E))$. The reverse direction of the proposition is clear by choosing $\Phi(x) := \bar{\Phi}(E_x)$. □

Precise statements for or against pd-ness can be derived, which are solely based on properties of the underlying T and base distance d:

Proposition 10 (Characterization by d and T)**.** *i) If* T *is too small compared to* $\bar{T}$ *in the sense that there exists* $x' \in \bar{T}_x$, *but* $d(T_x, T_{x'}) > 0$, *then the unregularized* d_{2S} *is not a Hilbertian metric.*

ii) If d *is the Euclidean distance in a Euclidean space* $\mathcal{X}$ *and* T_x *are parallel affine subspaces of* $\mathcal{X}$ *then the unregularized* d_{2S} *is a Hilbertian metric.*

Proof. For i) we note that $d(T_x, T_{x'}) = \inf_{t,t' \in T} d(t(x), t'(x')) > 0$. So d_{2S} is not totally invariant with respect to $\bar{T}$ and not a Hilbertian metric due to Prop. 9. For statement ii) we can define the orthogonal projection $\Phi : \mathcal{X} \to \mathcal{H} := (T_O)^\perp$ on the orthogonal complement of the linear subspace through the origin O, which implies that $d_{2S}(x,x') = d(\Phi(x), \Phi(x'))$ and all sets T_x are projected to a single point $\Phi(x)$ in $(T_O)^\perp$. So d_{2S} is a Hilbertian metric. □

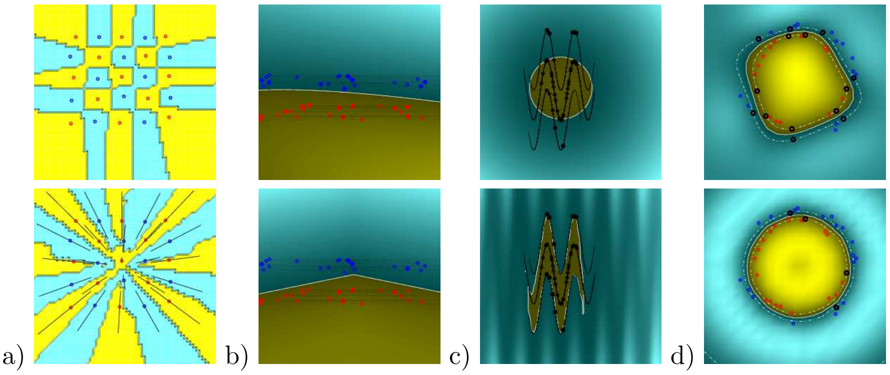

Figure 2: Illustration of non-invariant (upper row) versus invariant (lower row) kernel methods. a) Kernel k-nn classification with k^{rbf} and scale-invariance, b) kernel perceptron with k^{pol} of degree 2 and y-axis reflection-invariance, c) one-class-classification with k^{lin} and sine-invariance, d) SVM with k^{rbf} and rotation invariance.

In particular, these findings allow to state that the kernels on the left of Fig. 1 are not pd as they are not totally invariant wrt. $\bar{T}$. On the contrary, the extension of the upper right plot yields a pd kernel, as soon as T_x are complete affine subspaces. So these criteria can practically decide about the pd-ness of IDS-kernels.

If IDS-kernels are involved in learning algorithms, one should be aware of the possible indefiniteness, though it is frequently no relevant disadvantage in practice. Kernel principal component analysis can work with indefinite kernels, the SVM is known to tolerate indefinite kernels and further kernel methods are developed that accept such kernels. Even if an IDS-kernel can be proven by the preceding to be non-(c)pd in general, for various kernel parameter choices or a given dataset, the resulting kernel matrix can occasionally still be (c)pd.

5 Classification Experiments

For demonstration of the practical applicability in kernel methods, we condense the results on classification with IDS-kernels from (Haasdonk and Burkhardt (2007)) in Fig. 2. That study also gives summaries of real-world applications in the fields of optical character recognition and bacteria-recognition.

A simple kernel method is the kernel nearest-neighbour algorithm for classification. Fig. 2 a) is the result of the kernel 1-nearest-neighbour algorithm with the k^{rbf} and its scale-invariant k^{rbf}_{IDS} kernel, where the scaling sets T_x are indicated with black lines. The invariance properties of the kernel function obviously transfer to the analysis method by IDS-kernels.

Another aspect of interest is the convergence speed of online-learning algorithms exemplified by the kernel perceptron. We choose two random point sets of 20 points each lying uniformly distributed within two horizontal rectangular stripes indicated in Fig. 2 b). We incorporate the y-axis reflection invariance. By a random data drawing repeated 20 times, the non-invariant kernel k^{pol} of degree 2 results in 21.00 ± 6.59 update steps, while the invariant kernel k^{pol}_{IDS} converges much faster after 11.55 ± 4.54 updates. So the explicit invariance knowledge leads to improved convergence properties.

An unsupervised method for novelty detection is the optimal enclosing hypersphere algorithm (Shawe-Taylor and Cristianini (2004)). As illustrated in Fig. 2 c) we choose 30 points randomly lying on a sine-curve, which are interpreted as normal observations. We randomly add 10 points on slightly downward/upward shifted curves and want these points to be detected as novelties. The linear non-invariant k^{lin} results in an ordinary sphere, which however gives an average of 4.75 ± 1.12 false alarms, i.e. normal patterns detected as novelties, and 4.35 ± 0.93 missed outliers, i.e. outliers detected as normal patterns. As soon as we involve the sine-invariance by the IDS-kernel we consistently obtain 0.00 ± 0.00 false alarms and 0.40 ± 0.50 misses. So explicit invariance gives a remarkable performance gain in terms of recognition or detection accuracy.

We conclude the 2D experiments with the SVM on two random sets of 20 points distributed uniformly on two concentric rings, cf. Fig. 2 d). We involve rotation invariance explicitly by taking T as rotations by angles $\phi \in [-\pi/2, \pi/2]$. In the example we obtain an average of 16.40 ± 1.67 SVs (indicated as black points) for the non-invariant k^{rbf} case, whereas the IDS-kernel only returns 3.40 ± 0.75 SVs. So there is a clear improvement by involving invariance expressed in the model size. This is a determining factor for the required storage, number of test-kernel evaluations and error estimates.

6 Conclusion

We investigated and formalized elementary properties of IDS-kernels. We have proven that IDS-kernels offer two intuitive ways of adjusting the total invariance to approximate invariance until recovering the non-invariant case for various discrete, continuous, infinite and even non-group transformations. By this they build a framework interpolating between invariant and non-invariant machine learning. The definiteness of the kernels can be characterized precisely, which gives practical criteria for checking positive definiteness in applications.

The experiments demonstrate various benefits. In addition to the model-inherent invariance, when applying such kernels, further advantages can be the convergence speed in online-learning methods, model size reduction in SV approaches, or improvement of prediction accuracy. We conclude that these kernels indeed can be valuable tools for general pattern recognition problems with known invariances.

References

HAASDONK, B. (2005): *Transformation Knowledge in Pattern Analysis with Kernel Methods - Distance and Integration Kernels.* PhD thesis, University of Freiburg.

HAASDONK, B. and BAHLMANN, B. (2004): Learning with distance substitution kernels. In *Proc. of 26th DAGM-Symposium.* Springer, 220–227.

HAASDONK, B. and BURKHARDT, H. (2007): Invariant kernels for pattern analysis and machine learning. *Machine Learning*, 68, 35–61.

SCHÖLKOPF, B. and SMOLA, A. J. (2002): *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond.* MIT Press.

SHAWE-TAYLOR, J. and CRISTIANINI, N. (2004): *Kernel Methods for Pattern Analysis.* Cambridge University Press.