Learning with Distance Substitution Kernels

Bernard Haasdonk[1] and Claus Bahlmann[2]

[1] Computer Science Department
Albert-Ludwigs-University Freiburg
79110 Freiburg, Germany
haasdonk@informatik.uni-freiburg.de
[2] Siemens Corporate Research, Inc.
755 College Road East
Princeton, NJ 08540, USA
claus.bahlmann@scr.siemens.com

Abstract. During recent years much effort has been spent in incorporating problem specific a-priori knowledge into kernel methods for machine learning. A common example is a-priori knowledge given by a distance measure between objects. A simple but effective approach for kernel construction consists of substituting the Euclidean distance in ordinary kernel functions by the problem specific distance measure. We formalize this *distance substitution* procedure and investigate theoretical and empirical effects. In particular we state criteria for definiteness of the resulting kernels. We demonstrate the wide applicability by solving several classification tasks with SVMs. Regularization of the kernel matrices can additionally increase the recognition accuracy.

1 Introduction

In machine learning so called kernel methods have developed to state-of-the-art for a variety of different problem types like regression, classification, clustering, etc. [14]. Main ingredient in these methods is the problem specific choice of a kernel function. This choice should ideally incorporate as much a-priori knowledge as possible. One example is the incorporation of knowledge about pairwise proximities. In this setting, the objects are not given explicitly but only implicitly by a distance measure.

This paper focusses on the incorporation of such distance measures in kernel functions and investigates the application in support vector machines (SVMs) as the most widespread kernel method. Up to now mainly three approaches have been proposed for using distance data in SVMs. One approach consists of representing each training object as vector of its distances to all training objects and using standard SVMs on this data [5, 12]. The second method is embedding the distance data in a vector space, regularizing the possibly indefinite space and performing ordinary linear SVM classification [5, 12]. These approaches have the disadvantage of losing the sparsity in the sense that all training objects have to be retained for classification. This makes them inconvenient for large scale data.

The third method circumvents this problem by using the Gaussian rbf-kernel and plugging in problem specific distance measures [1, 4, 8, 11]. The aim of this paper is to formalize and extend this approach to more kernel types including polynomial kernels.

The paper is structured as follows: We formalize distance substitution in the next section. Statements on theoretical properties of the kernels follow in Section 3 and comments on consequences for use in SVMs are given in Section 4. In Section 5 we continue with SVM experiments by distance substitution and investigate regularization methods for the resulting kernel matrices. We conclude with Section 6.

2 Distance Substitution Kernels

The term *kernel* refers to a real valued symmetric function $k(x, x)$ of objects x in a set $\mathcal{X}$. A kernel function is *positive definite* (pd), if for any n, any objects $x_1, \ldots, x_n \in \mathcal{X}$ and any vector $\mathbf{c} \in \mathbb{R}^n$ the induced *kernel matrix* $\mathbf{K} := (k(x_i, x_j))_{i,j=1}^n$ satisfies $\mathbf{c}^T \mathbf{K} \mathbf{c} \geq 0$. The larger set of *conditionally positive definite* (cpd) kernels consists of those which satisfy this inequality for all $\mathbf{c}$ with $\mathbf{c}^T \mathbf{1} = 0$. These pd/cpd kernel functions got much attention as they have nice properties, in particular they can be interpreted/related to inner products in Hilbert spaces.

In distance based learning the data samples x are not given explicitly but only by a *distance* function $d(x, x')$. We do not impose strict assumptions on this distance measure, but require it to be symmetric, have zero diagonal, i.e. $d(x, x) = 0$, and be nonnegative. If a given distance measure does not satisfy these requirements, it can easily be symmetrized by $\bar{d}(x, x') := \frac{1}{2}(d(x, x') + d(x', x))$, given zero diagonal by $\bar{d}(x, x') := d(x, x') - \frac{1}{2}(d(x, x) + d(x', x'))$ or made positive by $\bar{d}(x, x') := |d(x, x')|$. We call such a distance *isometric to an L^2-norm* if the data can be embedded in a Hilbert space $\mathcal{H}$ by $\Phi : \mathcal{X} \to \mathcal{H}$ such that $d(x, x') = \|\Phi(x) - \Phi(x')\|$. After choice of an origin $O \in \mathcal{X}$ every distance d induces a function

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

This notation reflects the idea that in case of d being the L^2-norm in a Hilbert space $\mathcal{X}$, $\langle x, x' \rangle_d^O$ corresponds to the inner product in this space with respect to the origin O.

For any kernel $k(\|\mathbf{x} - \mathbf{x}'\|)$ and distance measure d we call $k_d(x, x') := k(d(x, x'))$ its *distance substitution kernel* (DS-kernel). Similarly, for a kernel $k(\langle \mathbf{x}, \mathbf{x}' \rangle)$ we call $k_d(x, x') := k(\langle x, x' \rangle_d^O)$ its DS-kernel. This notion is reasonable as in terms of (1) indeed distances are substituted. In particular for the simple linear, negative-distance, polynomial, and Gaussian kernels, we denote their DS-kernels by

$$\begin{aligned} k_d^{\mathrm{lin}}(x, x') &:= \langle x, x' \rangle_d^O & k_d^{\mathrm{nd}}(x, x') &:= -d(x, x')^\beta, \beta \in [0, 2] \\ k_d^{\mathrm{pol}}(x, x') &:= \left(1 + \gamma \langle x, x' \rangle_d^O\right)^p & k_d^{\mathrm{rbf}}(x, x') &:= e^{-\gamma d(x, x')^2}, p \in \mathbb{N}, \gamma \in \mathbb{R}^+. \end{aligned} \tag{2}$$

Of course, more general distance- or dot-product based kernels exist and corresponding DS-kernels can be defined, e.g. sigmoid, multiquadric, B_n-spline [14], etc.

3 Definiteness of DS-Kernels

The most interesting question posed on new kernels is whether they are (c)pd. In fact, for DS-kernels given by (2) the definiteness can be summed up quite easily. The necessary tools and references can be found in [14].

Proposition 1 (Definiteness of Simple DS-Kernels). *The following statements are equivalent for a (nonnegative, symmetric, zero-diagonal) distance d:*

$i)$ *d is isometric to an L^2-norm*

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

$iv)$ *k_d^{rbf} is pd for all $\gamma \in I\!\!R^+$* $\quad$ $v)$ *k_d^{pol} is pd for all $p \in I\!\!N, \gamma \in I\!\!R^+$.*

Proof. $i)$ implies $ii)$: [14, Prop. 2.22] covers the case $\beta = 2$ and [14, Prop. 2.23] settles the statement for arbitrary $\beta \in [0,2]$. The reverse implication $ii) \Rightarrow i)$ follows by [14, Prop. 2.24]. Equivalence of $ii)$ and $iii)$ also is a consequence of [14, Prop. 2.22]. [14, Prop. 2.28] implies the equivalence of $ii)$ and $iv)$. Statement $v)$ follows from $iii)$ as the set of pd functions is closed under products and linear combinations with positive coefficients. The reverse can be obtained from the pd functions $\frac{1}{\gamma}k_d^{\mathrm{pol}}$. With $p = 1$ and $\gamma \to \infty$ these functions converge to $\langle x, x' \rangle_d^O$. Hence the latter also is pd.

Further statements for definiteness of more general dot-product or distance-based kernels are possible, e.g. by Taylor series argumentation.

For some distance measures, the relation to an L^2-norm is apparent. An example is the *Hellinger distance* $H(p,p')$ between probability distributions which is defined by $(H(p,p'))^2 := \int \left(\sqrt{p} - \sqrt{p'}\right)^2 dx$. However, the class of distances which are isometric to L^2-norms is much wider than the obvious forms $d = \|x - x'\|$. For instance, [2] proves very nicely that $k^{\mathrm{rbf}}_{\sqrt{\chi^2}}$ is pd, where

$$\chi^2(\mathbf{x}, \mathbf{y}) := \frac{1}{2} \sum_i \frac{(x_i - y_i)^2}{x_i + y_i}$$

denotes the χ^2-distance between histograms. Thus, according to Proposition 1, $\sqrt{\chi^2}$ is isometric to an L^2-norm. Only looking at the χ^2-distance, the corresponding Hilbert space is not apparent. In summary we can conclude that not only k_H^{lin} (Bhattacharyya's affinity) and $k^{\mathrm{rbf}}_{\sqrt{\chi^2}}$, but all DS-kernels given by (2) are pd/cpd when using $\sqrt{\chi^2}$ or H.

In practice however, problem specific distance measures often lead to DS-kernels which are not pd. A criterion for disproving pd-ness is the following corollary, which is a simple consequence of Proposition 1 as L^2-norms are in particular metrics. It allows to conclude missing pd-ness of DS-kernels that arise from distances which are non-metric, e.g. violate the triangle inequality. It can immediately be applied to kernels based on tangent-distance [8], dynamic-time-warping (DTW) distance [1] or Kullback-Leibler (KL) divergence [11].

Corollary 1 (Non-Metricity Prevents Definiteness). *If d is not metric then the resulting DS-kernel k_d^{nd} is not cpd and $k_d^{\mathrm{lin}}, k_d^{\mathrm{rbf}}, k_d^{\mathrm{pol}}$ are not pd.*

Note, that for certain values β, γ, p, the resulting DS-kernels are possibly (c)pd. Remind also that the reverse of the corollary is not true. In particular, the L^p-metrics for $p \neq 2$ can be shown to produce non-pd DS-kernels.

4 SVMs with Indefinite Kernels

In the following we apply DS-kernels on learning problems. For this we focus on the very successful SVM for classification. This method can traditionally be applied if the kernel functions are pd or cpd. If a given distance produces DS-kernels which are pd, these can be applied in SVMs as usual. But also in the non-pd case they can be useful, as non-cpd kernels have shown convincing results in SVMs [1,4,8,11]. This empirical success is additionally supported by several theoretical statements:

1. *Feature space:* Indefinite kernels can be interpreted as inner products in indefinite vector spaces, enabling geometric argumentation [10].
2. *Optimal hyperplane classifier:* SVMs with indefinite kernels can be interpreted as optimal hyperplane classifiers in these indefinite spaces [7].
3. *Numerics:* Convergence of SVM implementations to a (possibly local) stationary point can be guaranteed [9].
4. *Uniqueness:* Even with extreme non-cpd kernel matrices unique solutions are possible [7].

5 Experiments

We performed experiments using various distance measures. Most of them were used in literature before. We do not explicitly state the definitions but refer to the corresponding publications. For each distance measure we used several labeled datasets or several labelings of a single dataset.

The dataset *kimia* (2 sets, each 72 samples, 6 classes) is based on binary images of shapes. The dissimilarity is measured by the modified Hausdorff distance. Details and results from other classification methods can be found in [12]. We applied a multiclass-SVM. The dataset *proteins* (226 samples) consists of evolutionary distances between amino acid sequences of proteins [6]. We used 4 different binary labelings corresponding to one-versus-rest problems. The dataset *cat-cortex* (65 samples) is based on a matrix of connectivity strengths between cortical areas of a cat. Other experiments with this data have been presented in [5, 6]. Here we symmetrized the similarity matrix and produced a zero diagonal distance matrix. Again we used 4 binary labelings corresponding to one-versus-rest classification problems. The datasets *music-EMD* and *music-PTD* are based on sets of 50 and 57 music pieces represented as weighted point sets. The earth-mover's distance (EMD) and the proportional transportation distance (PTD) were chosen as distance measures, see [16]. As class labels we used the corresponding composers resulting in 2 binary classification problems per distance measure. The dataset *USPS-TD* (4 sets, 250 samples per set, 2 classes) uses a fraction of the well known USPS handwritten digits data. As distance measure we use the two-sided tangent distance [15], which incorporates certain problem specific transformation knowledge. The set *UNIPEN-DTW* (2 sets, 250 samples per set, 5 classes) is based on a fraction of the the huge UNIPEN online handwriting sequence dataset. Dissimilarities were defined by the DTW-distance [1], again we applied a multiclass-SVM.

These different datasets represent a wide spectrum from easily to difficultly separable data. None of these distances are isometric to an L^2-norm. The restricted number of samples is consequence of the size of the small original datasets or due to the fact, that regularization experiments presented in Section 5.2 are only feasible for reasonably sized datasets.

5.1 Pure Distance Substitution

In this section we present results with pure distance substitution and compare them with the 1-nearest-neighbour and best k-nearest-neighbour classifier. These are natural classifiers when dealing with distance data.

We computed the leave-one-out (LOO) error of an SVM while logarithmically varying the parameter C along a line, respectively C, γ in a suitable grid. For the k_d^{pol} kernel a fixed polynomial degree p among $\{2, 4, 6, 8\}$ was chosen after simple initial experiments. The origin O was chosen to be the point with minimum squared distance sum to the other training objects. As $\frac{1}{2}k_d^{\mathrm{nd}}$ with $\beta = 2$ and k_d^{lin} are equivalent in SVMs (which follows by plugging (1) in the SVM optimization problem and making use of the equality constraint), we confine ourselves to using the former. We report the best LOO-error for all datasets in Table 1. Note that these errors might be biased compared to the true generalization error, as we did not use training/validation partitions for parameter optimization.

Table 1. Base LOO-errors [%] of classification experiments

dataset	k_d^{nd}	k_d^{pol}	k_d^{rbf}	1-nn	k-nn	dataset	k_d^{nd}	k_d^{pol}	k_d^{rbf}	1-nn	k-nn
kimia-1	15.28	11.11	4.17	5.56	5.56	music-EMD-1	40.00	22.00	20.00	42.00	42.00
kimia-2	12.50	9.72	9.72	12.50	12.50	music-EMD-2	42.11	43.86	10.53	21.05	21.05
proteins-H-α	0.89	0.89	0.89	1.33	1.33	music-PTD-1	34.00	30.00	32.00	46.00	34.00
proteins-H-β	3.54	2.21	2.65	3.54	3.54	music-PTD-2	31.58	33.33	28.07	38.60	38.60
proteins-M	0.00	0.00	0.00	0.00	0.00	USPS-TD-1	10.40	5.20	3.20	3.60	3.60
proteins-GH	0.00	0.44	0.00	1.77	1.77	USPS-TD-2	14.40	7.60	2.40	3.20	3.20
cat-cortex-V	3.08	1.54	0.00	3.08	3.08	USPS-TD-3	12.80	6.80	4.00	5.20	5.20
cat-cortex-A	6.15	3.08	4.62	6.15	6.15	USPS-TD-4	10.80	6.40	3.20	4.40	4.00
cat-cortex-S	6.15	3.08	3.08	6.15	3.08	UNIPEN-DTW-1	14.40	6.00	5.20	5.60	5.60
cat-cortex-F	7.69	6.15	4.62	4.62	3.08	UNIPEN-DTW-2	10.80	7.60	6.00	7.20	6.40

The identical low errors of the 1-nn and k-nn in the datasets *kimia, proteins, cat-cortex, USPS-TD,* and *UNIPEN-DTW* demonstrate that the data clusters well with the given labeling. For the music data sets the labels obviously not define proper clusters.

As SVMs with kernels k_d^{nd}, $\beta = 2$ can be interpreted as linear classifiers [7], the good performance of these on proteins and cat-cortex data is a hint on their linear separability. Simultaneously, the sets with higher error indicate that a nonlinear classifier in the dissimilarity space has to be applied. Indeed, the polynomial and Gaussian DS-kernel improve the results of the linear kernel for most datasets. The Gaussian DS-kernel even slightly outperforms the polynomial in most cases.

Compared to the nearest neighbour results, the nonlinear distance substitutions compare very favorable. The polynomial kernel can compete with or outperform the best k-nn for the majority of datasets, The Gaussian DS-kernel competes with or outperforms the best k-nn for all but one dataset.

For the last two distance measures, large scale experiments with certain distance substitution kernels have already been successfully presented in [1, 8]. In this respect, scalability of the results to large datasets is expected. To summarize, the experiments demonstrate the effectiveness of distance substitution kernels despite producing indefinite kernel matrices. The result is a sparse representation of the solution by training examples, that is, only a small subset of training objects has to be retained. Thus, it is particularly suited for large training sets.

5.2 Regularization of Kernel Matrices

In this section we investigate different regularization methods to eliminate the negative eigenvalues of the kernel matrices. Similar regularizations have been performed in literature, e.g. regularizing linear SVMs [5, 12] or embedding of non-metric data [13]. The method denoted off-diagonal addition (ODA) simply adds a suitable constant on the off-diagonal elements of the squared distance matrix, which results in a Euclidean distance matrix and therefore can be used for distance substitution resulting in pd kernels. Two other methods center the kernel matrix [12] and perform an eigenvalue decomposition. The approach (CNE) cuts off contributions corresponding to negative eigenvalues and (RNE) reflects the negative eigenvalues by taking their absolute values.

These operations particularly imply that the same operations have to be performed for the testing data. If the testing data is known beforehand, this can be used during training for computing and regularizing the kernel matrix. Note that this is *not* training on the testing data, as only the data points but not the labels are used for the kernel computations. Such training is commonly called *transductive* learning. If a test sample is not known at the training stage, the vector of kernel evaluations has to undergo the same regularization transformation as the kernel matrix before. Hence the diagonalizing vectors and eigenvalues have to be maintained and involved in this remapping of each testing vector. Both methods have the consequence that the computational complexity is increased during training and testing and the sparsity is lost, i.e. the solution depends on all training instances. So these regularization methods only apply, where computational demands are not so strict and sparsity is not necessary. For the experiments we used the transductive approach for determining the LOO errors.

If one can do without sparsity, another simple method is used for comparisons: Representing each training instance by a vector of squared distances to all training points makes a simple linear or Gaussian SVM applicable. We denoted these approaches as lin-SVM resp. rbf-SVM in Table 2, which lists the classification results.

The experiments demonstrate that regularization of kernel matrices can remarkably improve recognition accuracies and compete with or outperform SVMs on distance-vectors. The ODA regularization can increase accuracies, but it is clearly outperformed by the CNE and RNE methods which maintain or increase accuracy in 52 resp. 50 of the 60 experiments. Regularization seems to be advantageous for linear and polynomial kernels. For the Gaussian DS-kernels only few improvements can be observed. A comparison to the last columns indicates that the (non-ODA) regularized k_d^{nd} classifiers can compete with the linear SVM. The latter however is clearly inferior to the regularized

Table 2. LOO-errors [%] of classification experiments with regularized kernel matrices

dataset	k_d^{nd} ODA	k_d^{nd} CNE	k_d^{nd} RNE	k_d^{pol} CNE	k_d^{pol} RNE	k_d^{rbf} CNE	k_d^{rbf} RNE	lin-SVM	rbf-SVM
kimia-1	13.89	8.33	4.17	8.33	4.17	4.17	4.17	8.33	6.94
kimia-2	16.67	9.72	8.33	9.72	8.33	9.72	8.33	8.33	8.33
proteins-H-α	0.44	0.89	0.89	0.89	0.89	0.89	0.89	1.33	0.44
proteins-H-β	3.10	3.54	3.98	2.21	2.21	2.65	2.65	5.75	2.65
proteins-M	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
proteins-GH	0.00	0.00	0.00	0.44	0.44	0.00	0.00	0.00	0.00
cat-cortex-V	6.15	3.08	3.08	3.08	3.08	1.54	3.08	4.62	3.08
cat-cortex-A	6.15	4.62	6.15	4.62	6.15	4.62	6.15	1.54	1.54
cat-cortex-S	6.15	3.08	4.62	3.08	3.08	3.08	3.08	3.08	3.08
cat-cortex-F	6.15	4.62	4.62	4.62	4.62	4.62	4.62	1.54	1.54
music-EMD-1	44.00	38.00	40.00	30.00	40.00	30.00	30.00	44.00	20.00
music-EMD-2	42.11	15.79	21.05	12.28	12.28	14.04	10.53	21.05	15.79
music-PTD-1	38.00	44.00	40.00	40.00	38.00	32.00	32.00	40.00	28.00
music-PTD-2	47.37	29.82	38.60	26.32	22.81	28.07	17.54	29.82	21.05
USPS-TD-1	9.60	4.00	6.00	4.80	4.00	3.20	3.20	6.80	4.80
USPS-TD-2	14.40	9.60	7.20	5.60	4.00	2.40	2.40	6.00	4.40
USPS-TD-3	12.00	6.80	8.00	4.00	4.40	4.00	4.40	6.80	4.40
USPS-TD-4	11.20	7.60	6.40	5.20	5.60	3.20	3.20	7.20	1.60
UNIPEN-DTW-1	13.20	8.40	8.40	5.20	5.60	4.40	4.80	8.00	6.80
UNIPEN-DTW-2	11.20	7.60	9.60	6.80	6.40	6.00	5.60	9.60	8.40

nonlinear DS-kernels k_d^{pol} and k_d^{rbf}. In comparison to the rbf-SVM the k_d^{nd} experiments can not compete. The k_d^{pol}-CNE experiments also perform worse than the rbf-SVM in 12 cases. But the k_d^{pol}-RNE, k_d^{rbf}-CNE resp. k_d^{nd}-RNE settings obtain identical or better results than the rbf-SVM in the majority of classification problems.

6 Conclusion and Perspectives

We have characterized a class of kernels by formalizing distance substitution. This has so far been performed for the Gaussian kernel. By the equivalence of inner product and L^2-norm after fixing an origin, distances can also be used in inner-product kernels like the linear or polynomial kernel. We have given conditions for proving/disproving (c)pd-ness of the resulting kernels. We have concluded that DS-kernels involving e.g. the χ^2-distance are (c)pd, and others, e.g. resulting from KL-divergence, are not.

We have investigated the applicability of the DS-kernels by solving various SVM-classification problems with different data sets and different distance measures, which are not isometric to L^2-norms. The conclusion of the experiments was, that good classification is possible despite indefinite kernel matrices. Disadvantages of other methods are circumvented, e.g. test-data involved in training, approximate embeddings, non-sparse solutions or explicit working in feature space. This indicates that distance substitution kernels in particular are promising for large datasets. In particular the Gaussian and polynomial DS-kernels are good choices for general datasets due to their nonlinearity. If sparsity of the solution is not necessary and computational demands during classification are not so strict, then regularizations of the kernel matrices and the test-kernel evaluations can be recommended. It has been shown that this procedure can substantially improve recognition accuracy for e.g. the linear and polynomial DS-kernels.

Perspectives are to apply distance substitution on further types of kernels, further distance measures and in other kernel methods. This would in particular support recent promising efforts to establish non-cpd kernels for machine learning [3].

Acknowledgements

The authors want to thank Rainer Typke and Elzbieta Pekalska for making their distance data available and Harald Stepputtis for initial experiments.

References

1. C. Bahlmann, B. Haasdonk, and H. Burkhardt. On-line Handwriting Recognition with Support Vector Machines—A Kernel Approach. In *Proc. of the 8th IWFHR*, pages 49–54, 2002.
2. S. Belongie, C. Fowlkes, F. Chung, and J. Malik. Spectral partitioning with indefinite kernels using the Nyström extension. In *ECCV*, volume 2, pages 21–31, 2002.
3. S. Canu. Learning with non-positive kernels. Submitted to ICML, 2004.
4. O. Chapelle, P. Haffner, and V. Vapnik. Support vector machines for histogram-based image classification. *IEEE-NN*, 10(5):1055–1064, September 1999.
5. T. Graepel, R. Herbrich, P. Bollmann-Sdorra, and K. Obermayer. Classification on pairwise proximity data. In *NIPS 12*, pages 438–444. MIT Press, 1999.
6. T. Graepel, R. Herbrich, B. Schölkopf, A. Smola, P. Bartlett, K.-R. Müller, K. Obermayer, and B. Williamson. Classification on proximity data with LP–machines. In *Proc. of the 9th ICANN*, pages 304–309, 1999.
7. B. Haasdonk. Feature space interpretation of SVMs with non positive definite kernels. Internal report 1/03, IIF-LMB, University Freiburg, October 2003. Submitted to IEEE TPAMI.
8. B. Haasdonk and D. Keysers. Tangent distance kernels for support vector machines. In *Proc. of the 16th ICPR*, volume 2, pages 864–868, 2002.
9. H.-T. Lin and 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.
10. X. Mary. *Hilbertian subspaces, subdualities and applications*. PhD thesis, INSA Rouen, 2003.
11. P.J. Moreno, P. Ho, and N. Vasconcelos. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In *NIPS 17*, 2003.
12. E. Pekalska, P. Paclik, and R. Duin. A generalized kernel approach to dissimilarity based classification. *J. of Mach. Learn. Research*, 2:175–211, 2001.
13. V. Roth, J. Laub, M. Kawanabe, and J.M. Buhmann. Optimal cluster preserving embedding of nonmetric proximity data. *IEEE TPAMI*, 25(12):1540–1551, 2003.
14. B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, 2002.
15. P.Y. Simard, Y.A. LeCun, and J.S. Denker. Efficient pattern recognition using a new transformation distance. In *NIPS 5*, pages 50–58. Morgan Kaufmann, 1993.
16. R. Typke, P. Giannopoulos, R.C. Veltkamp, F. Wiering, and R. van Oostrum. Using transportation distances for measuring melodic similarity. In *ISMIR*, pages 107–114, 2003.