ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Learning in indefinite proximity spaces - recent trends

Frank-Michael Schleif[1], Peter Tino[1] and Yingyu Liang[2]

1- University of Birmingham
School of Computer Science,
Edgbaston B15 2TT Birmingham,
United Kingdom.

2- Princeton University
School of Computer Science
Princeton, USA

Abstract. Efficient learning of a data analysis task strongly depends on the data representation. Many methods rely on symmetric similarity or dissimilarity representations by means of *metric* inner products or distances, providing easy access to powerful mathematical formalisms like kernel approaches. Similarities and dissimilarities are however often naturally obtained by non-metric proximity measures which can not easily be handled by classical learning algorithms. Major efforts have been undertaken to provide approaches which can either directly be used for such data or to make standard methods available for these type of data. We provide an overview about recent achievements in the field of learning with indefinite proximities.

1 Introduction

The notion of pairwise proximities plays a key role in many machine learning algorithms. The comparison of objects by a *metric*, often Euclidean, distance measure is a standard element in basically every data analysis algorithm. Based on work of [1] and others, the usage of similarities by means of metric inner products or kernel matrices has lead to a great success of similarity based learning. Thereby the data are represented by metric pairwise similarities only. We can distinguish similarities, indicating how close or similar two items are to each other and dissimilarities as measures of the unrelatedness of two items. Given a set of N data items, their pairwise proximity (similarity or dissimilarity) measures can be conveniently summarized in a $N \times N$ proximity matrix. In the following we will refer to similarity and dissimilarity type proximity matrices as $\mathbf{S}$ and $\mathbf{D}$, respectively. In general at least symmetry is expected for $\mathbf{S}$ and $\mathbf{D}$, sometimes accompanied with non-negativity. These notions enter into models by means of similarity or dissimilarity functions $f(\mathbf{x}, \mathbf{y}) \in \mathbb{R}$ where $\mathbf{x}$ and $\mathbf{y}$ are the compared objects. The objects $\mathbf{x}, \mathbf{y}$ may exist in a d-dimensional vector space, so that $\mathbf{x} \in \mathbb{R}^d$, but can also be given without an explicit vectorial representation, e.g. biological sequences. However, as pointed out in [2], proximities often occur to be non-metric and their usage in standard algorithms leads to invalid model formulations.

The function $f(\mathbf{x}, \mathbf{y})$ may violate the metric properties to different degrees. While in general proximities are symmetric, the triangle inequality is often violated, proximities are negative, or self-dissimilarities are not zero. Such violations can be attributed

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

to different sources. While some authors attribute it to noise [3], for some proximities and proximity functions f this may be purposely caused by the measure itself. If noise is the source, often a simple eigenvalue correction [4] can be used, although this can become costly for large datasets. A recent analysis of the possible sources of negative eigenvalues is provided in [5]. Such an analysis can be potentially helpful in, for example, selecting the appropriate eigenvalue correction method applied to the proximity matrix. Prominent examples for genuine non-metric proximity measures can be found in the field of bioinformatics where classical sequence alignment algorithms produce non-metric proximity values. Here the non-metric part of the data can actually contain valuable information and should not be removed [6].

For non-metric inputs the support vector machine formulation [7] no longer leads to a convex optimization problem, but provides a local optimum [8], only. Accordingly, dedicated strategies for non-metric data are very desirable and the focus of this tutorial paper. A recent survey paper on the topic has been published in [9].

The paper is organized as follows. First we outline some basic notation and some mathematical formalism, related to machine learning with non-metric proximities also used in the referenced literature. Subsequently we discuss different views and sources of indefinite proximities and addresses the respective challenges in more detail. We also link to appropriate algorithms for indefinite proximity learning. Section 5 concludes this paper.

2 Notation and basic concepts

We now briefly review some concepts typically used in proximity based learning.

2.1 Kernels and kernel functions

Let $\mathcal{X}$ be a collection of N objects x_i, $i = 1, 2, ..., N$, in some input space. Further, let $\phi : \mathcal{X} \mapsto \mathcal{H}$ be a mapping of patterns from $\mathcal{X}$ to a high-dimensional or infinite dimensional Hilbert space $\mathcal{H}$ equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. The transformation ϕ is in general a non-linear mapping to a high-dimensional space $\mathcal{H}$ and may in general not be given in an explicit form. Instead a kernel function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is given which encodes the inner product in $\mathcal{H}$. The kernel k is a positive (semi) definite function such that $k(x, x') = \phi(x)^\top \phi(x')$ for any $x, x' \in \mathcal{X}$. The matrix $K := \Phi^\top \Phi$ is an $N \times N$ kernel matrix derived from the training data, where $\Phi : [\phi(x_1), \ldots, \phi(x_N)]$ is a matrix of images (column vectors) of the training data in $\mathcal{H}$. The motivation for such an embedding comes with the hope that the non-linear transformation of input data into higher dimensional $\mathcal{H}$ allows for using linear techniques in $\mathcal{H}$. Kernelized methods process the embedded data points in a feature space utilizing only the inner products $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ (kernel trick) [10], without the need to explicitly calculate ϕ. The specific kernel function can be very generic. Most prominent are the linear kernel with $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$ where $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$ is the Euclidean inner product or the rbf kernel $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right)$, with σ as a free parameter. Thereby it is assumed that the kernel function $k(\mathbf{x}, \mathbf{x}')$ is positive semi definite (psd).

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

2.2 Krein space

A Krein space is an *indefinite* inner product space endowed with a Hilbertian topology. Let $\mathcal{K}$ be a real vector space. An inner product space with an indefinite inner product $\langle \cdot, \cdot \rangle_{\mathcal{K}}$ on $\mathcal{K}$ is a bi-linear form where all $f, g, h \in \mathcal{K}$ and $\alpha \in \mathbb{R}$ obey the following conditions. Symmetry: $\langle f, g \rangle_{\mathcal{K}} = \langle g, f \rangle_{\mathcal{K}}$; linearity: $\langle \alpha f + g, h \rangle_{\mathcal{K}} = \alpha \langle f, h \rangle_{\mathcal{K}} + \langle g, h \rangle_{\mathcal{K}}$; and $\langle f, g \rangle_{\mathcal{K}} = 0$ implies $f = 0$. An inner product is positive definite if $\forall f \in \mathcal{K}$, $\langle f, f \rangle_{\mathcal{K}} \geq 0$, negative definite if $\forall f \in \mathcal{K}$, $\langle f, f \rangle_{\mathcal{K}} \leq 0$, otherwise it is indefinite. A vector space $\mathcal{K}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{K}}$ is called an inner product space.

An inner product space $(\mathcal{K}, \langle \cdot, \cdot \rangle_{\mathcal{K}})$ is a Krein space if we have two Hilbert spaces $\mathcal{H}_+$ and $\mathcal{H}_-$ spanning $\mathcal{K}$ such that $\forall f \in \mathcal{K}$ we have $f = f_+ + f_-$ with $f_+ \in \mathcal{H}_+$ and $f_- \in \mathcal{H}_-$ and $\forall f, g \in \mathcal{K}$, $\langle f, g \rangle_{\mathcal{K}} = \langle f_+, g_+ \rangle_{\mathcal{H}_+} - \langle f_-, g_- \rangle_{\mathcal{H}_-}$. A finite-dimensional Krein-space is a so called pseudo Euclidean space.

Indefinite kernels are typically observed by means of domain specific non-metric similarity functions (such as alignment functions used in biology [11]), by specific kernel functions - e.g. the Manhattan kernel $k(\mathbf{x}, \mathbf{y}) = -\|\mathbf{x} - \mathbf{y}\|_1$, tangent distance kernel [12] or divergence measures plugged into standard kernel functions [13]. Another source of non-psd kernels are noise artifacts on standard kernel functions [14].

For such spaces vectors can have negative squared "norm", negative squared "distances" and the concept of orthogonality is different from the usual Euclidean case. Given a symmetric *dissimilarity* matrix with zero diagonal, an embedding of the data in a pseudo-Euclidean vector space determined by the eigenvector decomposition of the associated similarity matrix $\mathbf{S}$ is always possible [15] [1]. Given the eigendecomposition of $\mathbf{S}$, $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, we can compute the corresponding vectorial representation $\mathbf{V}$ in the pseudo-Euclidean space by

$$\mathbf{V} = \mathbf{U}_{p+q+z} \left| \mathbf{\Lambda}_{p+q+z} \right|^{1/2} \tag{1}$$

where $\mathbf{\Lambda}_{p+q+z}$ consists of p positive, q negative non-zero eigenvalues and z zero eigenvalues. $\mathbf{U}_{p+q+z}$ consists of the corresponding eigenvectors. The triplet (p, q, z) is also referred to as the signature of the Pseudo-Euclidean space. A more detailed presentation of these mathematical aspects can be found in [2, 16, 17].

3 Indefinite proximity functions

Proximity functions can be very generic but are often restricted to fulfill metric properties to simplify the mathematical modeling and especially the parameter optimization. In [16] a large variety of such measures was reviewed and basically most nowadays public methods make use of metric properties. While this appears to be a reliable strategy researchers from different disciplines have criticized this restriction as inappropriate in multiple cases [18, 19, 5, 20].

In fact in [20] multiple examples from real problems show that many real life problems are better addressed by proximity measures which are not restricted to be metric.

[1] The associated similarity matrix can be obtained by double centering [2] of the dissimilarity matrix. $\mathbf{S} = -\mathbf{JDJ}/2$ with $\mathbf{J} = (\mathbf{I} - \mathbf{1}\mathbf{1}^\top/N)$, identity matrix $\mathbf{I}$ and vector of ones $\mathbf{1}$.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

Measure	Application field
Dynamic Time Warping (DTW) [30]	Time series or spectral alignment
Inner distance [31]	Shape retrieval e.g. in robotics
Compression distance [32]	Generic used also for text analysis
Smith Waterman Alignment [33]	Bioinformatics
Divergence measures [13]	Spectroscopy and audio processing
Non-metric modified Hausdorff [34]	Template matching

Table 1: List of commonly used non-metric proximity measures in various domains

The triangle inequality is most often violated if we consider object comparisons in daily life problems like the comparison of text documents, biological sequence data, spectral data or graphs [21, 22, 23]. These data are inherently compositional and a feature representation leads to information loss. As an alternative, tailored dissimilarity measures such as pairwise alignment functions, kernels for structures or other domain specific similarity and dissimilarity functions can be used as the interface to the data [24, 25]. But also for vectorial data, non-metric proximity measures are common in some disciplines. An example of this type is the use of divergence measures [13] which are very popular for spectral data analysis in chemistry, geo- and medical sciences [26, 27, 28, 29], and are not metric in general. Also the popular Dynamic Time Warping (DTW) [30] algorithm provides a non-metric alignment score which is often used as a proximity measure between two one-dimensional functions of different length. In image processing and shape retrieval indefinite proximities are often obtained by means of the inner distance [31].

A list of non-metric proximity measures is given in Table 1. Most of these measures are very popular but often violate the symmetry or triangle inequality condition or both. Hence many standard proximity based machine learning methods like kernel methods are not easy accessible for these data.

4 Learning models for indefinite proximities

A large number of algorithmic approaches assume that the data are given in a metric vector space, typically an Euclidean vector space, motivated by the strong mathematical framework which is available for metric Euclidean data. But with the advent of new measurement technologies and many non-standard data this strong constraint is often violated in practical applications and non-metric proximity matrices are more and more common.

This is often a severe problem for standard optimization frameworks as used e.g. for the Support Vector Machines (SVM), where psd matrices or more specific mercer kernels, are expected [7]. The naive usage of non-psd matrices in such a context invalidates the guarantees of the original approach (like ensured convergence to a convex or stationary point or the expected generalization accuracy to new points).

In [14] it was shown that the SVM not any longer optimizes a global convex function but is minimizing the distance between reduced convex hulls in a pseudo-Euclidean

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

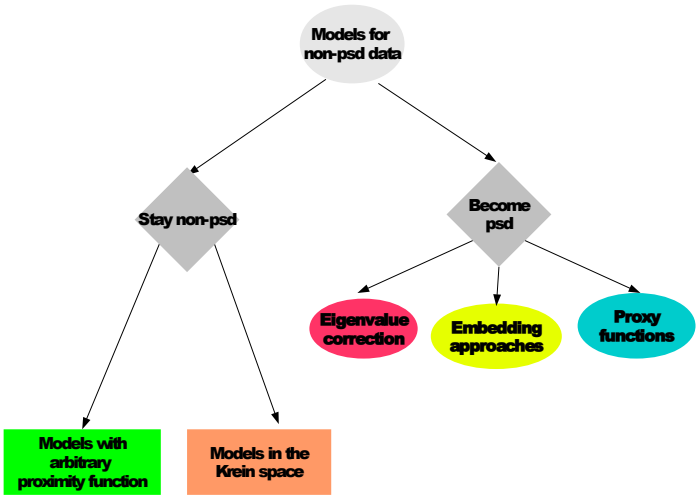

Fig. 1: Schematic view of different approaches to analyze non-psd data

space leading to a local optimum.

Currently, the vast majority of approaches encodes such comparisons by enforcing metric properties into these measures or by using alternative, in general less expressive measures, which do obey metric properties. With the continuous increase of non-standard and non-vectorial data sets non-metric measures and algorithms in Krein or pseudo-euclidean spaces are getting more popular and have recently raised wide interest in the research community [35, 36, 37, 38, 39, 40].

A schematic view summarizing the major research directions of the field is show in Figure 1 .

Basically, there exist two main directions:

(A) Transforming the non-metric proximities to become metric

(B) Stay in the non-metric space by providing a method which is insensitive to metric violations or can naturally deal with non-metric data

The first direction (A) can again be divided to the following sub-strategies:

- *Applying direct eigenvalue corrections*. The original data are decomposed by an Eigenvalue decomposition and the eigenspectrum is corrected in different ways to obtain a corrected psd matrix. A very simple form, avoiding the eigen-decomposition, is to consider $\mathbf{S} \cdot \mathbf{S}'$ which effectively squares the eigenvalues of the matrix S while the eigenvectors remain the same.

- *Embedding of the data in a metric space*. Here, the input data are embedded into a (in general Euclidean) vector space. A very simple strategy is to use Multi-Dimensional Scaling (MDS) to get a two- dimensional representation of the distance relations encoded in the original input matrix. If an eigen-decomposition of

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

the proximity matrix is available a pseudo-Euclidean embedding as shown in (1) can be used. Given the matrix $\mathbf{S}$ has true low rank this can be done with moderate costs using the approach proposed in [41].

- *Learning of a proxy function to the proximities*. These approaches learn an alternative (proxy) psd representation with maximum alignment to the non-psd input data. A prominent example is a SVM for indefinite input kernels provided by [3].

The second branch (B) is less diverse but one can identify at least two sub-strategies:

- *Model definition based on the non-metric proximity function*. Recent theoretical work on generic dissimilarity and similarity functions is used to define models which can directly employ the given proximity function with only very moderate assumptions. Key work in this line can be found in [42, 43, 44].

- *Krein space model definition*. The Krein space is the natural representation for non-psd data and some approaches have been formulated within this much less restrictive, but hence more complicated, mathematical space. A recent proposal with very generic implications can be found in [45].

As a general comment the approaches covered in (B) stay closer to the original input data whereas for the strategy (A) the input data are in parts substantially modified which can lead to a reduced interpretability and also limits a valid out-of sample extension in many cases. A detailed comparison, including algorithmic derivations is provided in [9]. While many approaches are focused on similarities, also dissimilarities can be employed in a similar form as shown in [41]. Specific realization of the aforementioned concepts can also be found in these proceedings. The work of [46] makes use of eigenvalue correction techniques to obtained positive definite kernels for a dimension reduction approach. The approach by [47] is based on a median prototype approach, following the branch (B) of our schema, where mixtures of dissimilarities are used in a supervised learning task. Finally the paper [48] is addressing the potentially negative impact of the positivation of graph kernels.

5 Conclusions

In this tutorial we briefly reviewed challenges and approaches common in the field of learning with indefinite proximities. The more recent proposals in this domain focus on the immediate processing of the data in the Krein space as well as on large scale problems. Those methods which are directly working on the indefinite proximity matrix are often non-sparse and use costly matrix operations [14, 49]. Large scale problems for proximity data are typically addressed by matrix approximation approaches [50, 51], now also available for indefinite proximities [41]. More recently also large scale sparse probabilistic models where proposed which do not rely on *metric* proximity data but use the empirical feature space [52, 53, 54].

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

Acknowledgment

This work was funded by a Marie Curie Intra-European Fellowship within the 7th European Community Framework Program (PIEF-GA-2012-327791). Peter Tino was supported by the EPSRC grant EP/L000296/1. Yingyu Liang was supported in part by NSF grants CCF-0832797, CCF-1117309, CCF-1302518, DMS-1317308, Simons Investigator Award, and Simons Collaboration Grant.

References

[1] B. Schoelkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

[2] E. Pekalska and R. Duin. *The dissimilarity representation for pattern recognition*. World Scientific, 2005.

[3] R. Luss and A. d'Aspremont. Support vector machine classification with indefinite kernels. *Mathematical Programming Computation*, 1(2-3):97–118, 2009.

[4] Y. Chen, E.K. Garcia, M.R. Gupta, A. Rahimi, and L. Cazzanti. Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*, 10:747–776, 2009.

[5] W. Xu, R.C. Wilson, and E.R. Hancock. Determining the cause of negative dissimilarity eigenvalues. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6854 LNCS(PART 1):589–597, 2011.

[6] Elzbieta Pekalska, Robert P. W. Duin, Simon Günter, and Horst Bunke. On not making dissimilarities euclidean. In *Structural, Syntactic, and Statistical Pattern Recognition, Joint IAPR International Workshops, SSPR 2004 and SPR 2004, Lisbon, Portugal, August 18-20, 2004 Proceedings*, pages 1145–1154, 2004.

[7] V.N. Vapnik. *The nature of statistical learning theory*. Statistics for engineering and information science. Springer, 2000.

[8] Hsuan tien Lin and Chih-Jen Lin. A study on sigmoid kernels for svm and the training of non-psd kernels by smo-type methods. Technical report, 2003.

[9] F.-M. Schleif and P. Tino. Indefinite proximity learning - a review. *Neural Computation*, 27(10):2039–2096, 2015.

[10] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis and Discovery*. Cambridge University Press, 2004.

[11] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, March 1981.

[12] Bernard Haasdonk and Daniel Keysers. Tangent distance kernels for support vector machines. In *ICPR (2)*, pages 864–868, 2002.

[13] A. Cichocki and S.-I. Amari. Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

[14] B. Haasdonk. Feature space interpretation of svms with indefinite kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):482–492, 2005.

[15] L. Goldfarb. A unified approach to pattern recognition. *Pattern Recognition*, 17(5):575 – 582, 1984.

[16] M.M. Deza and E. Deza. *Encyclopedia of Distances*. Encyclopedia of Distances. Springer, 2009.

[17] C.S. Ong, X. Mary, S. Canu, and A.J. Smola. Learning with non-positive kernels. pages 639–646, 2004.

[18] C.J. Hodgetts and U. Hahn. Similarity-based asymmetries in perceptual matching. *Acta Psychologica*, 139(2):291–299, 2012.

[19] T. Kinsman, M. Fairchild, and J. Pelz. Color is not a metric space implications for pattern recognition, machine learning, and computer vision. pages 37–40, 2012.

[20] Robert P. W. Duin and Elzbieta Pekalska. Non-euclidean dissimilarities: Causes and informativeness. In *Structural, Syntactic, and Statistical Pattern Recognition, Joint IAPR International Workshop, SSPR&SPR 2010, Cesme, Izmir, Turkey, August 18-20, 2010. Proceedings*, pages 324–333, 2010.

[21] Yihua Chen, Eric K. Garcia, Maya R. Gupta, Ali Rahimi, and Luca Cazzanti. Similarity-based classification: Concepts and algorithms. *JMLR*, 10:747–776, 2009.

[22] T. Kohonen and P. Somervuo. How to make large self-organizing maps for non-vectorial data. *Neural Networks*, 15(8-9):945–952, 2002.

[23] M. Neuhaus and H. Bunke. Edit distance based kernel functions for structural pattern classification. *Pattern Recognition*, 39(10):1852–1863, 2006.

[24] Thomas Gärtner, John W. Lloyd, and Peter A. Flach. Kernels and distances for structured data. *Machine Learning*, 57(3):205–232, 2004.

[25] Aleksandar Poleksic. Optimal pairwise alignment of fixed protein structures in subquadratic time. pages 367–382, 2011.

[26] E. Mwebaze, P. Schneider, F.-M. Schleif, J.R. Aduwo, J.A. Quinn, S. Haase, T. Villmann, and M. Biehl. Divergence based classification in learning vector quantization. *NeuroComputing*, 74:1429–1435, 2010.

[27] N.Q. Nguyen, C.K. Abbey, and M.F. Insana. Objective assessment of sonographic: Quality ii acquisition information spectrum. *IEEE Transactions on Medical Imaging*, 32(4):691–698, 2013.

[28] J. Tian, S. Cui, and P. Reinartz. Building change detection based on satellite stereo imagery and digital surface models. *IEEE Transactions on Geoscience and Remote Sensing*, 2013.

[29] F. van der Meer. The effectiveness of spectral similarity measures for the analysis of hyperspectral imagery. *International Journal of Applied Earth Observation and Geoinformation*, 8(1):3–17, 2006.

[30] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43–49, Feb 1978.

[31] Haibin Ling and David W. Jacobs. Using the inner-distance for classification of articulated shapes. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pages 719–726. IEEE Computer Society, 2005.

[32] Rudi Cilibrasi and Paul M. B. Vitányi. Clustering by compression. *IEEE Transactions on Information Theory*, 51(4):1523–1545, 2005.

[33] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997.

[34] M.-P. Dubuisson and A.K. Jain. A modified hausdorff distance for object matching. In *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision amp; Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, pages 566–568 vol.1, Oct 1994.

[35] G. Gnecco. Approximation and estimation bounds for subsets of reproducing kernel krein spaces. *Neural Processing Letters*, pages 1–17, 2013.

[36] J. Yang and L. Fan. A novel indefinite kernel dimensionality reduction algorithm: Weighted generalized indefinite kernel discriminant analysis. *Neural Processing Letters*, pages 1–13, 2013.

[37] S. Liwicki, S. Zafeiriou, and M. Pantic. Incremental slow feature analysis with indefinite kernel for online temporal video segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7725 LNCS(PART 2):162–176, 2013.

[38] S. Gu and Y. Guo. Learning svm classifiers with indefinite kernels. volume 2, pages 942–948, 2012.

[39] Stefanos Zafeiriou. Subspace learning in krein spaces: Complete kernel fisher discriminant analysis with indefinite kernels. In Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *ECCV (4)*, volume 7575 of *Lecture Notes in Computer Science*, pages 488–501. Springer, 2012.

[40] P. Kar and P. Jain. Supervised learning with similarity functions. volume 1, pages 215–223, 2012.

[41] A. Gisbrecht and F.-M. Schleif. Metric and non-metric proximity transformations at linear costs. *Neurocomputing*, 167:643–657, 2015.

[42] Huanhuan Chen, Peter Tino, and Xin Yao. Efficient probabilistic classification vector machine with incremental basis function selection. *IEEE Trans. Neural Netw. Learning Syst.*, 25(2):356–369, 2014.

[43] Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. A theory of learning with similarity functions. *Machine Learning*, 72(1-2):89–112, 2008.

[44] Liwei Wang, Masashi Sugiyama, Cheng Yang, Kohei Hatano, and Jufu Feng. Theory and algorithm for learning with dissimilarity functions. *Neural Computation*, 21(5):1459–1484, 2009.

[45] G. Loosli, S. Canu, and C. S. Ong. Learning svm in krein spaces. *TPAMI*, page to appear, 2016.

[46] A. Schulz and B. Hammer. Discriminative dimensionality reduction in kernel space. In *Proceedings of ESANN 2016*, page in these proc., 2016.

[47] M. Kaden, D. Nebel, and T. Villmann. Adaptive dissimilarity weighting for prototype-based classification optimizing mixtures of dissimilarities. In *Proceedings of ESANN 2016*, page in these proc., 2016.

[48] G. Loosli. Study on the loss of information caused by the "positivation" of graph kernels for 3d shapes. In *Proceedings of ESANN 2016*, page in these proc., 2016.

[49] G. Loosli and S. Canu. Non positive svm. In *Proc. of 4th International Workshop on Optimization for Machine Learning*, 2011.

[50] Mu Li, James T. Kwok, and Bao-Liang Lu. Making large-scale nyström approximation possible. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 631–638, 2010.

[51] F.-M. Schleif and A. Gisbrecht. Data analysis of (non-)metric proximities at linear costs. In *Proceedings of SIMBAD 2013*, pages 59–74, 2013.

[52] Huanhuan Chen, Peter Tino, and Xin Yao. Probabilistic classification vector machines. *IEEE Transactions on Neural Networks*, 20(6):901–914, 2009.

[53] F.-M. Schleif, H. Chen, and P. Tino. Incremental probabilistic classification vector machine with linear costs. In *Proceedings of IJCNN 2015*, pages 1–8, 2015.

[54] F.-M. Schleif, A. Gisbrecht, and P. Tino. Large scale indefinite kernel fisher discriminant. In *Proceedings of Simbad 2015*, pages 160–170, 2015.