Data analysis of (non-)metric proximities at linear costs

Frank-Michael Schleif and Andrej Gisbrecht

CITEC centre of excellence, Bielefeld University, 33615 Bielefeld, Germany,
{fschleif|agisbrec}@techfak.uni-bielefeld.de

Abstract. Domain specific (dis-)similarity or proximity measures, employed e.g. in alignment algorithms in bio-informatics, are often used to compare complex data objects and to cover domain specific data properties. Lacking an underlying vector space, data are given as pairwise (dis-)similarities. The few available methods for such data do not scale well to very large data sets. Kernel methods easily deal with *metric similarity* matrices, also at large scale, but costly transformations are necessary starting with non-metric (dis-) similarities. We propose an integrative combination of Nyström approximation, potential double centering and eigenvalue correction to obtain valid kernel matrices at linear costs. Accordingly effective kernel approaches, become accessible for these data. Evaluation at several larger (dis-)similarity data sets shows that the proposed method achieves much better runtime performance than the standard strategy while keeping competitive model accuracy. Our main contribution is an efficient linear technique, to convert (potentially non-metric) large scale *dissimilarity matrices* into approximated positive semi-definite kernel matrices.

1 Introduction

In many application areas such as bioinformatics, different technical systems, or the web, electronic data is getting larger and more complex in size and representation, using *domain specific* (dis-)similarity measures as a replacement or complement to Euclidean measures. Many classical machine learning techniques, have been proposed for Euclidean vectorial data. However, modern data are often associated to dedicated structures which make a representation in terms of Euclidean vectors difficult: biological sequence data, text files, XML data, trees, graphs, or time series [14, 10, 1] are of this type. These data are inherently compositional and a feature representation leads to information loss. As an alternative, a dedicated dissimilarity measure such as pairwise alignment, or kernels for structures can be used as the interface to the data. In such cases, machine learning techniques which can deal with pairwise similarities or dissimilarities have to be used [15]. Native methods for the analysis of dissimilarity data have been proposed in [15, 8, 7], but are widely based on non-convex optimization schemes and with quadratic to linear memory and runtime complexity, the later employing some of the approximation techniques discussed subsequently and additional heuristics.

Especially kernel methods, based on *metric similarity matrices*, revolutionized the possibility to deal with large electronic data, offering powerful tools to automatically extract regularities [19] in a convex optimization framework. But complex preprocessing steps are necessary, as discussed in the following, to apply them on non-metric (dis-) similarities. Large (dis-)similarity data are common in biology like the famous *UniProt/SwissProt*-DB with $\approx$ 500.000 entries or *GenBank* with $\approx$ 135.000 entries,

but there are many more (dis-)similarity data as discussed in the work based on [15, 16]. These growing data sets request effective modeling approaches. For protein and gene data recent work, proposed widely heuristically, strategies to improve the situation for large applications in unsupervised peptide retrieval [21].

Here we will show how potentially non-metric (dis-)similarities can be effectively processed by standard kernel methods with linear costs, also *in* the transformation step, which, to the authors best knowledge has not been reported before[1]. The proposed strategies permit the effective application of many kernel methods for these type of data under very mild conditions. Especially for metric dissimilarities the approach keeps the known guarantees like generalization bounds (see e.g. [3]) while for non-psd data corresponding proofs are still open, but our experiments are promising. The paper is organized as follows. First we give a short review about transformation techniques for dissimilarity data and discuss the influence of non-euclidean measures, by eigenvalue corrections. Subsequently, we discuss alternative methods for processing small dissimilarity data. We extend this discussion to approximation strategies, recalling the derivation of the low rank Nyström approximation for similarities and transfer this principle to dissimilarities. Then we link both strategies effectively to use kernel methods for the analysis of (non-)metric dissimilarity data and show the effectiveness by different exemplary supervised experiments. We also discuss differences and commons to some known approaches supported by experiments on simulated data.

2 Transformation techniques for dissimilarity data

Let $\mathbf{v}_j \in \mathbb{V}$ be a set of objects defined in some data space, with $|\mathbb{V}| = N$. We assume, there exists a dissimilarity measure such that $\mathbf{D} \in \mathbb{R}^{N \times N}$ is a dissimilarity matrix measuring the pairwise dissimilarities $D_{ij} = d(\mathbf{v}_i, \mathbf{v}_j)$ between all pairs $(\mathbf{v}_i, \mathbf{v}_j) \in \mathbb{V}$. Any reasonable (possibly non-metric) distance measure is sufficient. We assume zero diagonal $d(\mathbf{v}_i, \mathbf{v}_i) = 0$ for all i and symmetry $d(\mathbf{v}_i, \mathbf{v}_j) = d(\mathbf{v}_j, \mathbf{v}_i)$ for all i, j.

2.1 Analyzing dissimilarities by means of similarities for small N

For every dissimilarity matrix $\mathbf{D}$, an associated similarity matrix $\mathbf{S}$ is induced by a process referred to as double centering with costs of $\mathcal{O}(N^2)$[15]:

$$\mathbf{S} = -\mathbf{JDJ}/2$$
$$\mathbf{J} = (\mathbf{I} - \mathbf{1}\mathbf{1}^\top/N)$$

with identity matrix $\mathbf{I}$ and vector of ones $\mathbf{1}$. $\mathbf{D}$ is Euclidean if and only if $\mathbf{S}$ is positive semi-definite (psd). This means, we do not observe negative eigenvalues in the eigenspectrum of the matrix $\mathbf{S}$ associated to $\mathbf{D}$.

Many classification techniques have been proposed to deal with such psd kernel matrices $\mathbf{S}$ implicitly such as the support vector machine (SVM). In this case, preprocessing is *required to guarantee* psd. In [1] different strategies were analyzed to obtain

[1] Matlab code of the described transformations and test routines are available at: *kept blank for review*.

valid kernel matrices for a given similarity matrix $\mathbf{S}$, most popular are: *clipping, flipping, shift correction, vector-representation*. The underlying idea is to remove negative eigenvalues in the eigenspectrum of the matrix $\mathbf{S}$.

Assuming we have a symmetric similarity matrix $\mathbf{S}$, it has an eigenvalue decomposition $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, with orthonormal matrix $\mathbf{U}$ and diagonal matrix $\mathbf{\Lambda}$ collecting the eigenvalues. In general, p eigenvectors of $\mathbf{S}$ have positive eigenvalues and q have negative eigenvalues, $(p, q, N - p - q)$ is referred to as the *signature*.

The *clip*-operation sets all negative eigenvalues to zero, the *flip*-operation takes the absolute values, the *shift*-operation increases all eigenvalues by the absolute value of the minimal eigenvalue.

The corrected matrix $\mathbf{S}^*$ is obtained as $\mathbf{S}^* = \mathbf{U}\mathbf{\Lambda}^*\mathbf{U}^\top$, with $\mathbf{\Lambda}^*$ as the modified eigenvalue matrix using one of the above operations. The obtained matrix $\mathbf{S}^*$ can now be considered as a valid kernel matrix $\mathbf{K}$.

As an alternative, data points can be treated as vectors which coefficients or variables are given by the pairwise (dis-)similarity. These vectors can be processed using standard kernels. However, this view is changing the original data representation and leads to a finite data space, limited by the number of samples.

Interestingly, some operations such as shift do not affect the location of global optima of important cost functions such as the quantization error [12], albeit the transformation can severely affect the performance of optimization algorithms [9]. The analysis in [17] indicates that for non-Euclidean dissimilarities some corrections like above may change the data representation such that information loss occurs.

A schematic view of the relations between $\mathbf{S}$ and $\mathbf{D}$ and its transformations[2] is shown in Figure 1. Here we also report the complexity of the transformations using current typical approaches. Some of the steps can be done more efficiently by known methods, but with additional constraints or in under atypical settings as discussed in the following.

2.2 Analyzing dissimilarities by dedicated methods for small N

Alternatively, techniques have been introduced which directly deal with possibly non-metric dissimilarities. Given a symmetric dissimilarity with zero diagonal, an embedding of the data in a pseudo-Euclidean vector space determined by the eigenvector decomposition of the associated matrix $\mathbf{S}$ is always possible. A symmetric bilinear form in this space is given by $\langle \mathbf{x}, \mathbf{y} \rangle_{p,q} = \mathbf{x}^\top \mathbf{I}_{p,q} \mathbf{y}$ where $\mathbf{I}_{p,q}$ is a diagonal matrix with p entries 1 and q entries -1. Taking the eigenvectors of $\mathbf{S}$ together with the square root of the absolute value of the eigenvalues, we obtain vectors $\mathbf{v}_i$ in a pseudo-Euclidean space such that $D_{ij} = \langle \mathbf{v}_i - \mathbf{v}_j, \mathbf{v}_i - \mathbf{v}_j \rangle_{p,q}$ holds for every pair of data points. If the number of data is not limited, a generalization of this concept to Krein spaces with according decomposition is possible [15].

Vector operations can be directly transferred to the pseudo-Euclidean space, i.e. we can deal with center points (similar to k-means) as linear combinations of data in this space. Hence we can use multiple machine learning algorithms explicitly in pseudo-Euclidean space, relying on vector operations only. One problem of this explicit trans-

[2] Transformation equations are given also in the following sections.

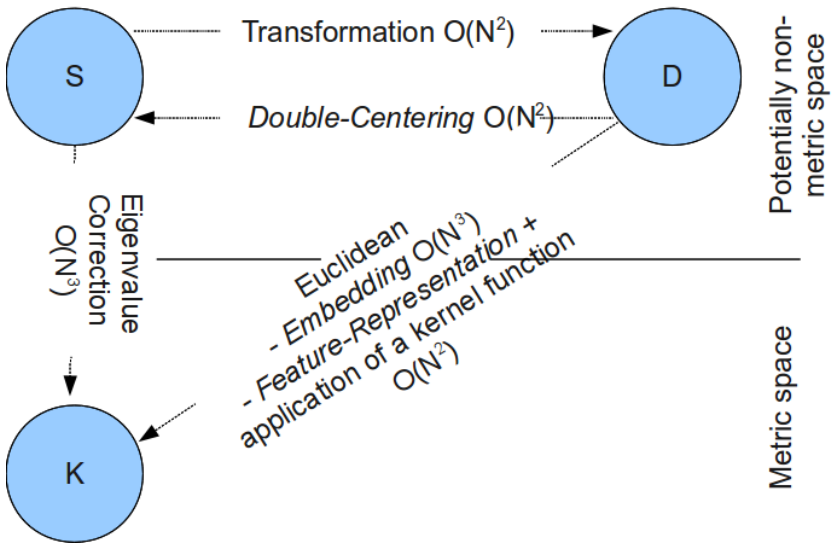

Fig. 1: Schema to illustrate the relation between similarities and dissimilarities.

fer is given by the computational complexity of the embedding which is $\mathcal{O}(N^3)$, and, further, the fact that out-of-sample extensions to new data points characterized by pairwise dissimilarities are not immediate. An improved strategy for learning a valid relational kernel from a matrix S was recently proposed in [13], employing latent wishart processes, but this approach does not scale for larger datasets. A further strategy is to employ so called relational or proximity learning methods as discussed e.g. in [7] The underlying models consist of prototypes, which are implicitly defined as a weighted linear combination of training points: $\mathbf{w}_j = \sum_i \alpha_{ji}\mathbf{v}_i$ with $\sum_i \alpha_{ji} = 1$. But this explicit representation is not necessary because the algorithms are solely based on a specific form of distance calculations using only the matrix $\mathbf{D}$, the potentially unknown vector space V is not needed. The basic idea is an implicit computation of distances $d(\cdot,\cdot)$ during the model calculation based on the dissimilarity matrix $\mathbf{D}$ using weights α:

$$d(\mathbf{v}_i, \mathbf{w}_j) = [\mathbf{D} \cdot \alpha_j]_i - \frac{1}{2} \cdot \alpha_j^\top \mathbf{D} \alpha_j \tag{1}$$

details can be found in the aforementioned paper. As shown e.g. in [9] the mentioned methods do not rely on a metric dissimilarity matrix $\mathbf{D}$, but it is sufficient to have a symmetric $\mathbf{D}$ in a pseudo-euclidean space, with constant self-dissimilarities.

The methods discussed before are suitable for data analysis based on similarity or dissimilarity data where the number of samples N is rather small, e.g. scales by some thousand samples. For larger N only for *metric, similarity data* (valid kernels) efficient approaches have been proposed before, e.g. low-rank linearized SVM [25] or the Core-Vector Machine (CVM) [22].

In the following we discuss techniques to deal with larger sample sets for, potentially non-metric similarity and especially dissimilarity data. Especially we show how standard kernel methods can be used, assuming that for non-metric data, the necessary transformations have no severe negative influence on the data accuracy. Basically also core-set techniques become accessible for large potentially non-metric (dis-)similarity data in this way, but at the cost of multiple additional intermediate steps.

3 Nyström approximation

The aforementioned methods depend on the similarity matrix $\mathbf{S}$ or dissimilarity matrix $\mathbf{D}$, respectively. For kernel methods and more recently for prototype based learning the usage of the Nystöm approximation is a well known technique to obtain effective learning algorithms [23, 7].

3.1 Nyström approximation for similarities

The Nyström approximation technique has been proposed in the context of kernel methods in [23] with related proofs and bounds given in [3]. Here, we give a short review of this technique. One well known way to approximate a $N \times N$ Gram matrix, is to use a low-rank approximation. This can be done by computing the eigendecomposition of the kernel $\mathbf{K} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, where $\mathbf{U}$ is a matrix, whose columns are orthonormal eigenvectors, and $\mathbf{\Lambda}$ is a diagonal matrix consisting of eigenvalues $\mathbf{\Lambda}_{11} \geq \mathbf{\Lambda}_{22} \geq ... \geq 0$, and keeping only the m eigenspaces which correspond to the m largest eigenvalues of the matrix. The approximation is $\mathbf{K} \approx \mathbf{U}_{N,m}\mathbf{\Lambda}_{m,m}\mathbf{U}_{m,N}$, where the indices refer to the size of the corresponding submatrix. The Nyström method approximates a kernel in a similar way, without computing the eigendecomposition of the whole matrix, which otherwise is an $O(N^3)$ operation.

By the Mercer theorem kernels $k(\mathbf{x}, \mathbf{y})$ can be expanded by orthonormal eigenfunctions ψ_i and non negative eigenvalues λ_i in the form

$$k(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} \lambda_i \psi_i(\mathbf{x}) \psi_i(\mathbf{y}).$$

The eigenfunctions and eigenvalues of a kernel are defined as the solution of the integral equation

$$\int k(\mathbf{y}, \mathbf{x}) \psi_i(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \lambda_i \psi_i(\mathbf{y}),$$

where $p(\mathbf{x})$ is the probability density of $\mathbf{x}$. This integral can be approximated based on the Nyström technique by sampling $\mathbf{x}^k$ i.i.d. according to $p(\mathbf{x})$:

$$\frac{1}{m} \sum_{k=1}^{m} k(\mathbf{y}, \mathbf{x}^k) \psi_i(\mathbf{x}^k) \approx \lambda_i \psi_i(\mathbf{y}).$$

Using this approximation and the matrix eigenproblem equation

$$\mathbf{K}^{(m)}\mathbf{U}^{(m)} = \mathbf{U}^{(m)}\mathbf{\Lambda}^{(m)}$$

of the corresponding $m \times m$ Gram sub-matrix $\mathbf{K}^{(m)}$ we can derive the approximations for the eigenfunctions and eigenvalues of the kernel k

$$\lambda_i \approx \frac{\lambda_i^{(m)}}{m}, \quad \psi_i(\mathbf{y}) \approx \frac{\sqrt{m}}{\lambda_i^{(m)}} \mathbf{k}_y \mathbf{u}_i^{(m)}, \tag{2}$$

where $\mathbf{u}_i^{(m)}$ is the ith column of $\mathbf{U}^{(m)}$. Thus, we can approximate ψ_i at an arbitrary point $\mathbf{y}$ as long as we know the vector $\mathbf{k}_y = (k(\mathbf{x}^1, \mathbf{y}), ..., k(\mathbf{x}^m, \mathbf{y}))^\top$.

For a given $N \times N$ Gram matrix $\mathbf{K}$ we randomly choose m rows and respective columns. The corresponding indices's are also called landmarks, and should be chosen such that the data distribution is sufficiently covered. A specific analysis about selection strategies was recently discussed in [24]. We denote these rows by $\mathbf{K}_{m,N}$. Using the formulas (2) we obtain $\tilde{\mathbf{K}} = \sum_{i=1}^{m} 1/\lambda_i^{(m)} \cdot \mathbf{K}_{m,N}^\top \mathbf{u}_i^{(m)} (\mathbf{u}_i^{(m)})^\top \mathbf{K}_{m,N}$, where $\lambda_i^{(m)}$ and $\mathbf{u}_i^{(m)}$ correspond to the $m \times m$ eigenproblem. Thus we get, $\mathbf{K}_{m,m}^{-1}$ denoting the Moore-Penrose pseudoinverse, an approximation of $\mathbf{K}$ as

$$\tilde{\mathbf{K}} = \mathbf{K}_{m,N}^\top \mathbf{K}_{m,m}^{-1} \mathbf{K}_{m,N}.$$

This approximation is exact, if $\mathbf{K}_{m,m}$ has the same rank as $\mathbf{K}$.

3.2 Nyström approximation for dissimilarity data

For dissimilarity data, a direct transfer is possible, see [7] for earlier work on this topic. Earlier work in this line, but not equivalent, also appeared in the work around Landmark Multi-Dimensional-Scaling (LMDS) [20] which we address in the next section. According to the spectral theorem, a symmetric dissimilarity matrix $\mathbf{D}$ can be diagonalized $\mathbf{D} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ with $\mathbf{U}$ being a unitary matrix whose column vectors are the orthonormal eigenvectors of $\mathbf{D}$ and $\mathbf{\Lambda}$ a diagonal matrix with the corresponding eigenvalues of $\mathbf{D}$, Therefore the dissimilarity matrix can be seen as an operator

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N} \lambda_i \psi_i(\mathbf{x}) \psi_i(\mathbf{y})$$

where $\lambda_i \in \mathbb{R}$ correspond to the diagonal elements of $\mathbf{\Lambda}$ and ψ_i denote the eigenfunctions. The only difference to an expansion of a kernel is that the eigenvalues can be negative. All further mathematical manipulations can be applied in the same way and we can write in an analogy to the equation 3.1

$$\hat{\mathbf{D}} = \mathbf{D}_{N,m} \mathbf{D}_{m,m}^{-1} \mathbf{D}_{N,m}^\top.$$

It allows to approximate dissimilarities between a point $\mathbf{w}^k$ represented by a coefficient vector α_k and a data point $\mathbf{x}^i$, as discussed within Eq (1), in the way

$$\begin{aligned} d(\mathbf{x}^i, \mathbf{w}^k) \approx & \left[\mathbf{D}_{m,N}^\top \left(\mathbf{D}_{m,m}^{-1} \left(\mathbf{D}_{m,N} \boldsymbol{\alpha}_k\right)\right)\right]_i \\ & -\frac{1}{2} \cdot \left(\boldsymbol{\alpha}_k^\top \mathbf{D}_{m,N}^\top\right) \cdot \\ & \left(\mathbf{D}_{m,m}^{-1} \left(\mathbf{D}_{m,N} \boldsymbol{\alpha}_k\right)\right) \end{aligned}$$

with a linear submatrix of m rows and a low rank matrix $\mathbf{D}_{m,m}$. Performing these matrix multiplications from right to left, this computation is $\mathcal{O}(m^2N)$ instead of $\mathcal{O}(N^2)$, i.e. it is linear in the number of data points N, assuming fixed approximation m.

A benefit of the Nyström technique is that it can be decided priorly which linear parts of the dissimilarity matrix will be used in training. Therefore, it is sufficient to *compute only a linear part of the full dissimilarity matrix* $\mathbf{D}$ to use these methods. A drawback of the Nyström approximation is that a good approximation can only be achieved if the rank of $\mathbf{D}$ is kept as much as possible, i.e. the chosen subset should be representative. The specific selection of the m landmark points has been recently analyzed in [24]. It was found that best results can be obtained by choosing the potential cluster centers of the data distribution as landmarks, rather a random subset, to be able to keep m smallest at lowest representation error. However the determination of these centers can become complicated for large data sets, since it can be obviously not be based on a Nyström approximated set. However the effect is not such severe as long as m is not too small.

4 Transformations of (dis-)similarities with linear costs

For *metric* similarity data, kernel methods can be applied directly, or in case of large N, the Nyström approximation can be used. We will discuss *non*-metric data later and focus now on metric or almost metric *dissimilarity* data $\mathbf{D}$.

4.1 Transformation of dissimilarities to similarities

As pointed out before current methods for large dissimilarity matrix $\mathbf{D}$ are non-convex approaches. On the other hand multiple effective convex kernel methods are available for metric similarity data using a matrix $\mathbf{S} = \mathbf{K}$ which we will now make accessible for matrices $\mathbf{D}$ in an effective manner. This requests for a transformation of the matrix $\mathbf{D}$ to $\mathbf{S}$ using double-centering as discussed above. This transformation contains a summation over the whole matrix and thus has quadratic complexity, which would be prohibitive for larger data sets.

One way to achieve this transformation in linear time, is to use landmark multidimensional scaling (LMDS) [20] which was shown to be a Nyström technique as well [18]. The idea is to sample a small amount m of points, called landmarks, compute the corresponding dissimilarity matrix, apply double centering on this matrix and finally project the data to a low dimensional space using eigenvalue decomposition. The remaining points can then be projected into the same space, taking into account the distances to the landmarks, and applying triangulation. Having vectorial representation of the data, it is then easy to retrieve the similarity matrix as a scalar product between the points.

Another possibility arises if we take into account our key observation, that we can combine both transformations, double centering and Nyström approximation, and make use of their linearity. Instead of applying double centering, followed by the Nyström approximation we first approximate the matrix $\mathbf{D}$ and then transform it by double centering, which yields the approximated similarity matrix $\hat{\mathbf{S}}$.

Both approaches have the costs of $\mathcal{O}(m^2N)$ and produce the same results, up to shift and rotation. This is because LMDS, in contrast to our approach, makes double centering only on a small part of $\mathbf{D}$, and thus is unable to detect the mean and the primary components of the whole data set. This can result in an unreliable impact, since similarities which are not centered might lead to an inferior performance of the algorithms and, thus, our approach should be used instead[3]. Additionally LMDS implicitly assumes that the dissimilarities are metric, respectively the negative eigenvalues of the corresponding similarity matrix are automatically clipped. This can have a negative impact on the data analysis as we show in a synthetic example in the following. Further LMDS is proposed as a projection technique leading to a low-dimensional, typically $2-3$ dimensional embedding of the data. Higher dimensional embeddings by LMDS are possible (limited by the number of positive eigenvalues), but to our best knowledge neither used nor discussed so far. A Nyström approximated kernel, avoiding the calculations of all dissimilarities, as shown in the following is not directly obtained but only after embedding of the corresponding dissimilarities and subsequent calculation of the inner products. But for this kernel the negative eigenvalues are always clipped which can have a negative impact on the analysis. Accordingly, the connection of LMDS to our approach is rather weak[4], which will get more obvious in the following derivations.

As mentioned before double centering of a matrix $\mathbf{D}$ is defined as:

$$\mathbf{S} = -\mathbf{JDJ}/2$$

where $\mathbf{J} = (\mathbf{I} - \mathbf{1}\mathbf{1}^\top/N)$ with identity matrix $\mathbf{I}$ and vector of ones $\mathbf{1}$. $\mathbf{S}$ is positive semi-definite (psd) if and only if D is Euclidean.

Lets start with a dissimilarity matrix D where we apply double centering, subsequently we approximate the obtained $\mathbf{S}$ by integrating the Nyström approximation to the matrix $\mathbf{D}$.

$$\begin{aligned}
\mathbf{S} &= -\frac{1}{2}\mathbf{JDJ} \\
&= -\frac{1}{2}\left(\left(\mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\mathbf{D}\left(\mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right)\right) \\
&= -\frac{1}{2}\left(\mathbf{IDI} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\mathbf{DI} - \mathbf{ID}\frac{1}{N}\mathbf{1}\mathbf{1}^\top + \frac{1}{N}\mathbf{1}\mathbf{1}^\top\mathbf{D}\frac{1}{N}\mathbf{1}\mathbf{1}^\top\right) \\
&= -\frac{1}{2}\left(\mathbf{D} - \frac{1}{N}\mathbf{D}\mathbf{1}\mathbf{1}^\top - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\mathbf{D} + \frac{1}{N^2}\mathbf{1}\mathbf{1}^\top\mathbf{D}\mathbf{1}\mathbf{1}^\top\right)
\end{aligned}$$

[3] For domain specific dissimilarity measures and non-vectorial data as discussed here, it is, under practical conditions, hard to ensure that the underlying, implicit space is normalized to N(0,1), this is getting even more complicated if the measure is non-metric.

[4] Although LMDS can be adapted to provide similar results, with the exception that the small inner matrix is calculated differently with the pre-discussed influence on unnormalized data.

$$\mathbf{S} \overset{Ny}{\approx} \hat{\mathbf{S}} = -\frac{1}{2}\left[\mathbf{D}_{N,m}\cdot\mathbf{D}_{m,m}^{-1}\cdot\mathbf{D}_{m,N} - \frac{1}{N}\mathbf{D}_{N,m}\right. \tag{3}$$
$$\cdot(\mathbf{D}_{m,m}^{-1}\cdot(\mathbf{D}_{m,N}\mathbf{1}))\mathbf{1}^\top - \frac{1}{N}\mathbf{1}((\mathbf{1}^\top\mathbf{D}_{N,m})\cdot\mathbf{D}_{m,m}^{-1})$$
$$\left.\cdot\mathbf{D}_{m,N} + \frac{1}{N^2}\mathbf{1}((\mathbf{1}^\top\mathbf{D}_{N,m})\cdot\mathbf{D}_{m,m}^{-1}\cdot(\mathbf{D}_{m,N}\mathbf{1}))\mathbf{1}^\top\right]$$

This equation can be rewritten for each entry of the matrix $\hat{\mathbf{S}}$

$$\hat{S}_{ij} = -\frac{1}{2}\left[\mathbf{D}_{i,m}\cdot\mathbf{D}_{m,m}^{-1}\cdot\mathbf{D}_{m,j} - \frac{1}{N}\sum_k \mathbf{D}_{k,m}\cdot\mathbf{D}_{m,m}^{-1}\cdot\mathbf{D}_{m,j}\right.$$
$$-\frac{1}{N}\sum_k \mathbf{D}_{i,m}\cdot\mathbf{D}_{m,m}^{-1}\cdot\mathbf{D}_{m,k}$$
$$\left.+\frac{1}{N^2}\sum_{kl}\mathbf{D}_{k,m}\cdot\mathbf{D}_{m,m}^{-1}\cdot\mathbf{D}_{m,l}\right],$$

as well as for the sub-matrices $\hat{\mathbf{S}}_{m,m}$ and $\hat{\mathbf{S}}_{N,m}$, in which we are interested for the Nyström approximation

$$\hat{\mathbf{S}}_{m,m} = -\frac{1}{2}\left[\mathbf{D}_{m,m} - \frac{1}{N}\mathbf{1}\cdot\sum_k \mathbf{D}_{k,m}\right.$$
$$-\frac{1}{N}\sum_k \mathbf{D}_{m,k}\cdot\mathbf{1}^\top$$
$$\left.+\frac{1}{N^2}\mathbf{1}\cdot\sum_{kl}\mathbf{D}_{k,m}\cdot\mathbf{D}_{m,m}^{-1}\cdot\mathbf{D}_{m,l}\cdot\mathbf{1}^\top\right]$$

$$\hat{\mathbf{S}}_{N,m} = -\frac{1}{2}\left[\mathbf{D}_{N,m} - \frac{1}{N}\mathbf{1}\cdot\sum_k \mathbf{D}_{k,m}\right.$$
$$-\frac{1}{N}\sum_k \mathbf{D}_{N,m}\cdot\mathbf{D}_{m,m}^{-1}\cdot\mathbf{D}_{m,k}\cdot\mathbf{1}^\top$$
$$\left.+\frac{1}{N^2}\mathbf{1}\cdot\sum_{kl}\mathbf{D}_{k,m}\cdot\mathbf{D}_{m,m}^{-1}\cdot\mathbf{D}_{m,l}\cdot\mathbf{1}^\top\right].$$

It should be noted that $\hat{\mathbf{S}}$ is only a valid kernel if $\hat{\mathbf{D}}$ is metric. The information loss obtained by the approximation is 0 if m corresponds to the rank of $\mathbf{S}$ and increases for smaller m.

4.2 Non-metric (dis-)similarities

In case of a non-metric $\mathbf{D}$ the transformation shown in equation 3 can still be used, but the obtained matrix $\hat{\mathbf{S}}$ is not a valid kernel. A strategy to obtain a valid kernel matrix $\hat{\mathbf{S}}$ is to apply an eigenvalue correction as discussed above. This however can be prohibitive for large matrices, since to correct the whole eigenvalue spectrum, the whole eigenvalue decomposition is needed, which has $\mathcal{O}(N^3)$ complexity. The Nyström approximation can again decrease computational costs dramatically. Since we now can apply the approximation on an arbitrary symmetric matrix, we can make the correction afterward. To correct an already approximated similarity matrix $\hat{\mathbf{S}}$ it is sufficient to correct the eigenvalues of $\mathbf{S}_{m,m}$. Altogether we get $\mathcal{O}(m^2N)$ complexity.

We can write for the approximated matrix $\hat{\mathbf{S}}$ its eigenvalue decomposition as

$$\hat{\mathbf{S}} = \mathbf{S}_{N,m}\mathbf{S}_{m,m}^{-1}\mathbf{S}_{N,m}^{\top} = \mathbf{S}_{N,m}\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^{\top}\mathbf{S}_{N,m}^{\top},$$

where we can correct the eigenvalues $\mathbf{\Lambda}$ by some technique as discussed in section 2.1 to $\mathbf{\Lambda}^*$. The corrected approximated matrix $\hat{\mathbf{S}}^*$ is then simply

$$\hat{\mathbf{S}}^* = \mathbf{S}_{N,m}\mathbf{U}\left(\mathbf{\Lambda}^*\right)^{-1}\mathbf{U}^{\top}\mathbf{S}_{N,m}^{\top}. \tag{4}$$

This approach can also be used to correct dissimilarity matrices $\mathbf{D}$ by first approximating them, converting to similarities $\hat{\mathbf{S}}$ using equation 3 and then correcting the similarities. If it is desirable to work with the corrected dissimilarities, then we should note, that it is possible to transform the similarity matrix $\mathbf{S}$ to a dissimilarity matrix $\mathbf{D}$: $D_{ij}^2 = S_{ii} + S_{jj} - 2S_{ij}$. This obviously applies as well to the approximated and corrected matrices $\hat{\mathbf{S}}^*$ and $\hat{\mathbf{D}}^*$ and we get by substitution:

$$\hat{\mathbf{D}}^* = \mathbf{D}_{N,m}^*\left(\mathbf{D}_{m,m}^*\right)^{-1}\mathbf{D}_{N,m}^{*\top}. \tag{5}$$

Usually the algorithms are learned on a so called training set and we expect them to perform well on the new unseen data, or the test set. In such cases we need to provide an out of sample extension, i.e. a way to compute the algorithm on the new data. This might be a problem for the techniques dealing with (dis)similarities. If the matrices are corrected, we need to correct the new (dis)similarities as well to get consistent results. Fortunately, it is quite easy in the Nyström framework. By examining the equations 4 and 5 we see, that we simply need to extend the matrices $\mathbf{D}_{N,m}$ or $\mathbf{S}_{N,m}$, respectively, by uncorrected (dis)similarities between the new points and the landmarks to obtain the full approximated and *corrected* (dis)similarity matrices, which then can be used by the algorithms to compute the out of sample extension.

In [1] a similar approach is taken. First, the whole similarity matrix is corrected by means of a projection matrix. Then this projection matrix is applied to the new data, so that the corrected similarity between old and new data can be computed. This technique is in fact the Nyström approximation, where the whole similarity matrix $\mathbf{S}$ is treated as the approximation matrix $\mathbf{S}_{m,m}$ and the old data, together with the new data build the matrix $\mathbf{S}_{N,m}$. Rewriting this in the Nyström framework makes it clear and more obvious, without the need to compute the projection matrix and with an additional possibility to compute the similarities between the new points. In Figure 2 we depict

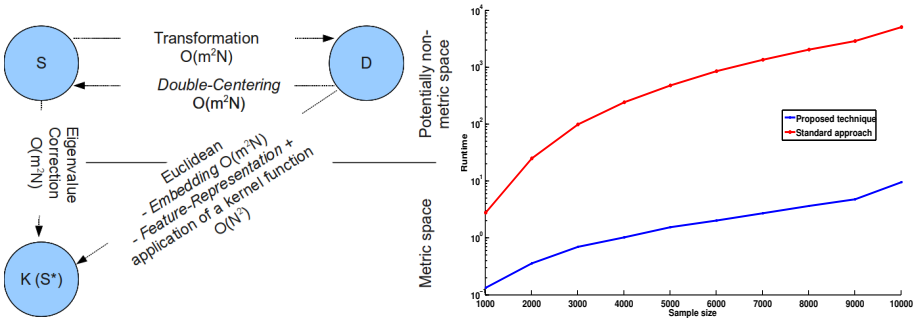

Fig. 2: Left: Updated schema from Figure 1 using the discussed approximation. The costs are now substantially smaller $m \ll N$. Right: Runtime in seconds at log-scale for the SwissProt-Runtime experiment. The standard approach is two magnitudes slower than the proposed technique.

schematically the new situation for similarity and dissimilarity data incorporating the proposed approach.

As a last point it should be mentioned that corrections like flipping, clipping or others are still under discussion and not always optimal [15]. Additionally the selection of landmark points can be complicated as discussed in [24]. Further for very large data sets (e.g. some 100 million points) the Nyström approximation may still be too costly and some other strategies have to be found.

We close this section by a small experiment on the ball dataset as proposed in [5]. It is an artificial dataset based on the surface distances of randomly positioned balls of two classes having a slightly different radius. The dataset is non-euclidean with substantial information encoded in the negative part of the eigenspectrum. We generated the data with 100 samples per class leading to a dissimilarity matrix $D = N \times N$, with $N = 200$. Now the data have been processed in four different ways to obtain a valid kernel matrix S. First we converted D into a valid kernel matrix by a full eigenvalue decomposition, followed by flipping of the negative eigenvalues and a reconstruction of the similarity matrix $K = S$, denoted as $SIM1$. This approach has a complexity of $\mathcal{O}(N^3)$. Further we generated an approximated similarity matrix $\hat{S}$ by using the proposed approach, flipping in the eigenvalue correction and 10 landmarks for the Nyström approximation. This dataset is denoted as $SIM2$ and was obtained with a complexity of $\mathcal{O}(m^2N)$. The third dataset $SIM3$ was obtained in the same way but the eigenvalues were clipped. The dataset $SIM4$ was obtained using landmark MDS with the same landmarks as for $SIM2$ and $SIM3$. The data are processed by a Support Vector Machine in a 10-fold crossvalidation results on the test sets are shown in Table 1. As mentioned the data con-

Table 1: Test set results of a 10-fold SVM run on the ball dataset using the different encodings.

	$SIM1$	$SIM2$	$SIM3$	$SIM4$
Test-Accuracy	100 ± 0	87.00 ± 7.53	68.00 ± 6.32	52.00 ± 11.83

tain substantial information in the negative fraction of the eigenspectrum, accordingly one may expect that this eigenvalues should not be removed. This is also reflected in the results. LMDS removed the negative eigenvalues and the classification model based on this data shows random prediction accuracy. The SIM3 encoding is slightly better. Also in this case the negative eigenvalues are removed but the limited amount of class separation information, encoded in the positive fraction was better preserved, probably due to the different calculation of the matrix $\hat{S}_{mm}$. The SIM2 data used the flipping strategy and shows already quite good prediction accuracy, taking into account that the kernel matrix is only approximated by 10 landmarks and the relevant (original negative) eigenvalues are of small magnitude.

5 Experiments

We now apply the priorly derived approach to three non-metric dissimilarity and similarity data and show the effectiveness for a classification task. The considered data are (1) the SwissProt similarity data as described in [10] (DS1, 10988 samples, 30 classes, imbalanced, signature: $[8488, 2500, 0]$) (2) the chromosome dissimilarity data taken from [14] (DS2, 4200 samples, 21 classes, balanced, signature: $[2258, 1899, 43]$) and the proteom dissimilarity data set [4] (DS3, 2604 samples, 53 classes, imbalanced, signature: $[1502, 682, 420]$). All datasets are non-metric, multiclass and contain multiple thousand objects, such that a regular eigenvalue correction with a prior double-centering for dissimilarity data, as discussed before, is already very costly. The data are analyzed in two ways, employing either the flipping strategy as an eigenvalue correction, or by not-correcting the eigenvalues[5]. To be effective for the large number of object we also apply the Nyström approximation as discussed before using a sample rate of $1\%, 10\%, 30\%$[6], by selecting random landmarks from the data. Other sampling strategies have been discussed in [24, 6], also the impact of the Nyström approximation with respect to kernel methods has been discussed recently in [2], but this is out of the focus of this paper.

To get comparable experiments, the same randomly drawn landmarks are used in each of the corresponding sub-experiments (along a column in the table). New landmarks are only drawn for different Nyström approximations and sample sizes like in Figure 3. Classification rates are calculated in a 10-fold crossvalidation using the Core-Vector-Machine (CVM) and the Support-Vector-Machine (SVM) (see [22, 19]). The crossvalidation does not include a new draw of the landmarks, to cancel out the selection bias of the Nyström approximation, accordingly SVM and CVM use the same kernel matrices. However, our objective is not maximum classification performance (which is only one possible application) but to demonstrate the effectiveness of our approach for dissimilarity data of larger scale. The classification results are summarized

[5] Clipping and flipping were found similar effective, with a little advance for flipping. With flipping the information of the negative-eigenvalues is at least somewhat kept in the data representation so we focus on this representation. Shift correction was found to have a negative impact on the model as already discussed in [1].

[6] A larger sample size did not lead to further substantial improvements in the results.

Table 2: Average test set accuracy for SwissProt (DS1), Chromosome (DS2), Proteom (DS3) using a Nyström approximation of 1% and 10% and no or flip eigenvalue correction. Kernel matrices have been Nyström approximated either, as proposed during the eigenvalue correction, or later on, like in the standard approach. The signatures are based on the approximated kernel matrices.

	$DS1_{1\%}$	$DS2_{1\%}$	$DS3_{1\%}$	$DS1_{10\%}$	$DS2_{10\%}$	$DS3_{10\%}$
Signature	[109,1,10878]	[41,1,4158]	[25,1,2492]	[1078,19,9891]	[296,123,3781]	[235,10,2273]
CVM-No	92.81 ± 0.74	94.64 ± 0.88	64.42 ± 2.89	75.53 ± 0.90	40.43 ± 2.12	23.95 ± 2.4
SVM-No	92.82 ± 0.90	94.24 ± 1.00	45.59 ± 3.01	82.92 ± 2.00	47.21 ± 2.42	27.56 ± 2.93
Signature	[110,0,10878]	[42,0,4158]	[26,0,2492]	[1097,0,9891]	[419,0,3781]	[245,0,2273]
CVM-Flip	92.78 ± 0.74	94.62 ± 0.85	91.62 ± 1.57	97.01 ± 0.54	96.98 ± 0.77	96.98 ± 1.28
SVM-Flip	93.02 ± 0.70	94.31 ± 1.37	93.65 ± 1.52	97.56 ± 0.51	96.98 ± 0.88	97.34 ± 0.73

Table 3: Average test set accuracy for SwissProt (DS1), Chromosome (DS2), Proteom (DS3) using a Nyström approximation of 30% and no or flip eigenvalue correction. Kernel matrices have been Nyström approximated (with $L = 30\% \cdot N$) either, as proposed during the eigenvalue correction, or later on, like in the standard approach. The signatures are based on the approximated kernel matrices.

	DS1		DS2		DS3	
Signature	[2995,300,7693]		[759,493,2948]		[577,118,1823]	
CVM-No	72.14 ± 2.01		60.24 ± 3.12		56.75 ± 2.56	
SVM-No	77.01 ± 3.03		66.36 ± 2.94		49.21 ± 2.51	
Signature	[3295,0,7693]		[1252,0,2948]		[695,0,1823]	
CVM-Flip	96.85 ± 0.53		96.90 ± 0.66		99.17 ± 0.28	
SVM-Flip	97.49 ± 0.36		96.98 ± 0.45		98.85 ± 0.78	

in Table 2-3 for the different Nyström approximations $1\%, 10\%$ and 30%. First one observes that the eigenvalue correction has a strong, positive effect on the classification performance consistent with earlier findings [1]. However in case of a small number of landmarks the effect of the eigenvalue correction is less pronounced compared to the uncorrected experiment as shown in Table 2 for DS1 and DS2. In these cases the Nyström approximation has also reduced the number of non-negative eigenvalues, as shown by the corresponding signatures, such that an implicit eigenvalue correction is obtained. For DS3 the remaining eigenvector has a rather high magnitude and a strong impact accordingly, such that the classification performance is sub-optimal for the uncorrected experiment. Raising the number of landmarks Table 2-3 also the classification performance improves for the experiments with eigenvalue correction. The experiments without eigenvalue correction show however a degeneration in the performance, because more and more negative eigenvalues are still kept by the Nyström approximation as shown in the signatures[7].

[7] Comparing signatures at different Nyström approximations also shows that many eigenvalues are close to zero and are sometimes counted as positive,negative or zero.

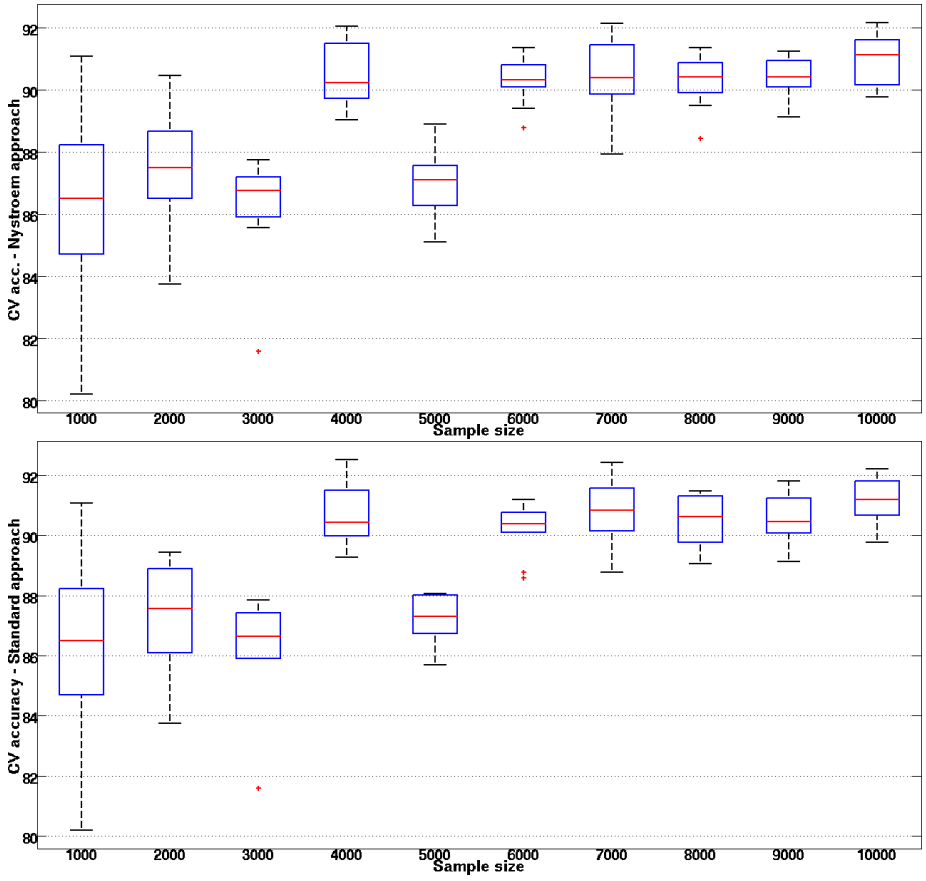

Fig. 3: Top: box-plots of the classification performance for different sample sizes of DS1 using the proposed approach with 100 landmarks. Bottom: The same experiment but with the standard approach. Obviously our approach does not sacrifice performance for computational speed.

As shown exemplary in Figure 3 the classification performance on eigenvalue-corrected data is approximately the same using our proposed strategy or the standard technique, but the runtime performance (right plot in Figure 2) is drastically better for an increase in the number of samples. To show this we selected subsets from the SwissProt data with different sizes from 1000 to 10000 points and calculated the runtime and classification performance using the CVM classifier in a 10-fold crossvalidation, with a fixed Nyström approximation of $L = 100$ and a flipping eigenvalue correction. The results of the proposed approach are shown in the left box-plots of Figure 3 and the results for the standard technique are shown in the right plot. The corresponding runtimes are shown in Figure 3, with the runtime of our approach as the curve on the bottom and the runtime of the standard method on the top, two magnitudes larger on log-scale.

6 Outlook and Conclusions

In this paper we discussed the relation between similarity and dissimilarity data and effective ways to move across the different representations in a systematic way. Using the presented approach, effective and *accurate* transformations are possible. Kernel approaches but also dissimilarity learners are now accessible for both types of data. While the parametrization of the Nyström approximation is already studied in [11, 24] there are still different open issues. In future work we will analyze more deeply the handling of extremely large (dis-)similarity sets and transfer our approach to unsupervised problems. While the proposed strategy was found to be very effective e.g. to improve supervised learning of non-metric dissimilarities by kernel methods, it is however also limited again by the Nyström approximation, which may fail to provide sufficient approximation. Accordingly it is still very interesting to provide dedicated methods for such data as argued in [17]. For non-psd data the error introduced by the Nyström approximation is not yet fully understood and bounds similar as proposed in [3] are still an open issue. In our experiments we observed that flipping was an effective approach to keep the relevant structure of the data but this are only heuristic findings and not yet completely understood, we will address this in future work. **Acknowledgments**: We would like to thank the *Max-Planck-Institute for Physics of Complex Systems in Dresden* and Michael Biehl, Thomas Villmann and Manfred Opper as the organizer of the *Statistical Inference: Models in Physics and Learning*-Workshop for providing a nice working atmosphere during the preparation of this manuscript. Frank-Michael Schleif was supported by the "German Science Foundation (DFG)" under grant number HA-2719/4-1. Financial support from the Cluster of Excellence 277 Cognitive Interaction Technology funded by the German Excellence Initiative is gratefully acknowledged. Further we would like to thank Fabrice Rossi for helpful discussions and suggestions.

References

[1] Chen, Y., Garcia, E.K., Gupta, M.R., Rahimi, A., Cazzanti, L.: Similarity-based classification: Concepts and algorithms. JMLR 10, 747–776 (2009)
[2] Cortes, C., Mohri, M., Talwalkar, A.: On the impact of kernel approximation on learning accuracy. JMLR - Proceedings Track 9, 113–120 (2010)
[3] Drineas, P., Mahoney, M.W.: On the nyström method for approximating a gram matrix for improved kernel-based learning. Journal of Machine Learning Research 6, 2153–2175 (2005)
[4] Duin, R.P.: PRTools (march 2012), http://www.prtools.org
[5] Duin, R.P.W., Pekalska, E.: Non-euclidean dissimilarities: Causes and informativeness. In: Hancock, E.R., Wilson, R.C., Windeatt, T., Ulusoy, I., Escolano, F. (eds.) SSPR/SPR. Lecture Notes in Computer Science, vol. 6218, pp. 324–333. Springer (2010)
[6] Farahat, A.K., Ghodsi, A., Kamel, M.S.: A novel greedy algorithm for nyström approximation. JMLR - Proceedings Track 15, 269–277 (2011)
[7] Gisbrecht, A., Mokbel, B., Schleif, F.M., Zhu, X., Hammer, B.: Linear time relational prototype based learning. Journal of Neural Systems 22(5) (2012)

[8] Graepel, T., Obermayer, K.: A stochastic self-organizing map for proximity data. Neural Computation 11(1), 139–155 (1999)
[9] Hammer, B., Hasenfuss, A.: Topographic mapping of large dissimilarity data sets. Neural Computation 22(9), 2229–2284 (2010)
[10] Kohonen, T., Somervuo, P.: How to make large self-organizing maps for nonvectorial data. Neural Networks 15(8-9), 945–952 (2002)
[11] Kumar, S., Mohri, M., Talwalkar, A.: On sampling-based approximate spectral decomposition. In: ICML. ACM International Conference Proceeding Series, vol. 382, p. 70. ACM (2009)
[12] Laub, J., Roth, V., Buhmann, J.M., Müller, K.R.: On the information and representation of non-euclidean pairwise data. Pattern Recognition 39(10), 1815–1826 (2006)
[13] Li, W.J., Zhang, Z., Yeung, D.Y.: Latent wishart processes for relational kernel learning. JMLR - Proceedings Track 5, 336–343 (2009)
[14] Neuhaus, M., Bunke, H.: Edit distance based kernel functions for structural pattern classification. Pattern Recognition 39(10), 1852–1863 (2006)
[15] Pekalska, E., Duin, R.: The dissimilarity representation for pattern recognition. World Scientific (2005)
[16] Pekalska, E., Duin, R.P.W.: Beyond traditional kernels: Classification in two dissimilarity-based representation spaces. IEEE Transactions on Systems, Man, and Cybernetics, Part C 38(6), 729–744 (2008)
[17] Pekalska, E., Duin, R.P.W., Günter, S., Bunke, H.: On not making dissimilarities euclidean. In: SSPR/SPR. Lecture Notes in Computer Science, vol. 3138, pp. 1145–1154. Springer (2004)
[18] Platt, J.: Fastmap, metricmap, and landmark mds are all nyström algorithms (2005)
[19] Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis and Discovery. Cambridge University Press (2004)
[20] de Silva, V., Tenenbaum, J.B.: Global versus local methods in nonlinear dimensionality reduction. In: NIPS. pp. 705–712. MIT Press (2002)
[21] Tan, J., Kuchibhatla, D., Sirota, F.L.: Tachyon search speeds up retrieval of similar sequences by several orders of magnitude. Bioinformatics p. online (23.04.2012) (2012)
[22] Tsang, I.W., Kocsor, A., Kwok, J.T.: Simpler core vector machines with enclosing balls. In: ICML. ACM International Conference Proceeding Series, vol. 227, pp. 911–918. ACM (2007)
[23] Williams, C.K.I., Seeger, M.: Using the nyström method to speed up kernel machines. In: NIPS. pp. 682–688. MIT Press (2000)
[24] Zhang, K., Kwok, J.T.: Clustered nyström method for large scale manifold learning and dimension reduction. IEEE Transactions on Neural Networks 21(10), 1576–1587 (2010)
[25] Zhang, K., Lan, L., Wang, Z., Moerchen, F.: Scaling up kernel svm on limited resources: A low-rank linearization approach. JMLR - Proceedings Track 22, 1425–1434 (2012)