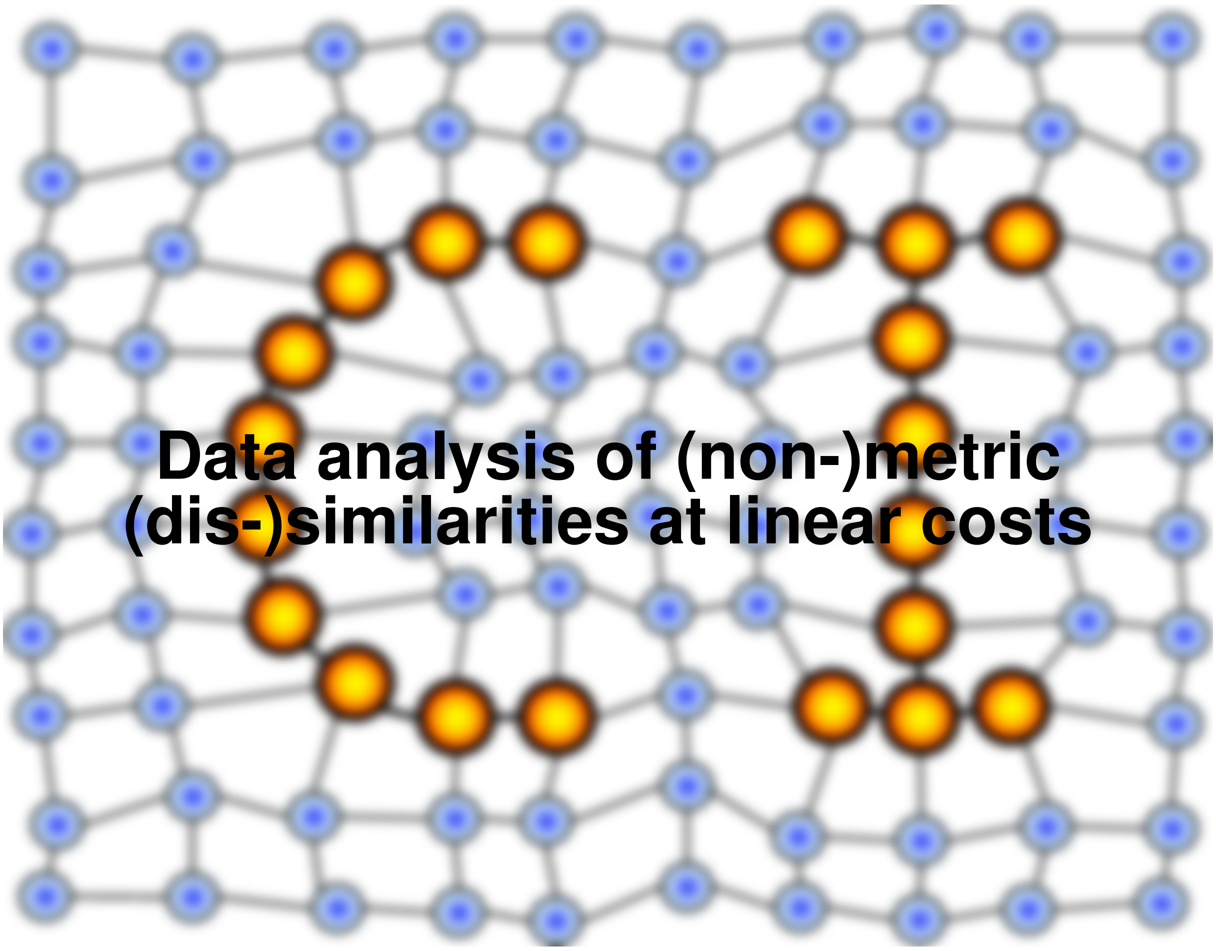
MACHINE LEARNING REPORTS



Data analysis of (non-)metric (dis-)similarities at linear costs

Report 04/2012
Submitted: 15.09.2012
Published: 20.09.2012

F.-M. Schleif*, A. Gisbrecht
(1) Theoretical Computer Science, University of Bielefeld, Universitätsstrasse 21-23, 33615
Bielefeld, Germany
* both authors contributed equally,
corresponding author: *email: fschleif@techfak.uni-bielefeld.de*

Machine Learning Reports
http://www.techfak.uni-bielefeld.de/∼fschleif/mlr/mlr.html

Abstract

The analysis of large datasets is often based on domain specific dissimilarity measures. Such measures can be found for example in the life-sciences but also in fields like web mining or social networks. These measures can generate similarities or dissimilarities and may be metric or non-metric, often without an explicite vector space. Metric similarity data are easily processed by kernel methods, whereas for dissimilarity data only few specific methods are available or costly transformations are needed to obtain a valid kernel matrix. If the data additionally are non-euclidean further corrections are needed and it is argued that transformations to a kernel space, are not only costly but further can lead to information loss. We consider the linear processing of metric similarity and dissimilarity data and show procedures to do data analysis with linear costs for both types of data.

Machine Learning Reports
http://www.techfak.uni-bielefeld.de/∼fschleif/mlr/mlr.html

1 Introduction

In many application areas such as bioinformatics, technical systems, or the web, electronic data sets are increasing rapidly with respect to size and complexity and domain specific (dis-)similarity measures, replacing or complementing Euclidean measures are more and more common. Machine learning has revolutionized the possibility to deal with large electronic data sets in these areas by offering powerful tools to automatically extract a regularity from given data. Popular approaches provide diverse techniques for data structuring and data inspection. Visualization, clustering, or classification still constitute one of the most common tasks in this context [1, 4, 20, 9].

Many classical machine learning techniques, have been proposed for Euclidean vectorial data. 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, for example [15, 7]. 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 [12].

Also kernel methods like the Support Vector Machine (SVM) (see e.g.[19]) can be used for dissimilarity data, but complex preprocessing steps are necessary as discussed in the following. In fact, as discussed in the work of Pekalska[13], dissimilarity data can encode information in the euclidean and non-euclidean space and transformations to obtain a valid kernel may be inappropriate[14]. Native methods for the analysis of dissimilarity data have been proposed in [13, 24, 6] with quadratic to linear memory and runtime complexity, the later employing approximation techniques discussed in the following. To obtain effective scaling these methods employ the Nystöm approximation, as discussed in the following and are based on prototypes. Prototypes are typical representants in the dataspace of the underlying problem, either in vector form or implicite by linear combination of known points. Prominent methods of this type are e.g. k-means, with cluster centers as prototypes or learning vector quantizers, where characteristic points of a class are determined as prototypes used to model the underlying classification problem see e.g. [18]. Prototype methods share some nice properties. They provide sparse models, show good interpretability and can be easily extended to new requirements [17]. However they are often determined by non-convex, iterative optimization methods.

Here we will show how metric dissimilarities 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. The paper is organized as follows. First we give a short review about transformation methods between similarity and dissimilarity data and discuss the influence of non-euclidean measures. Subsequently, we recall 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 metric dissimilarity data and show the effectiveness by different experiments. The paper is closed by a discussion about the results and a short outlook regarding non-metric cases.

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 $\mathcal{O}(N^2)$:

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

with identity matrix $\mathbf{I}$ and vector of ones $\mathbf{1}$. $\mathbf{D}$ is Euclidean if and only if $\mathbf{S}$ is positive semidefinite (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 [2] different strategies were analyzed to obtained valid kernel matrized for a given similarity matrix $\mathbf{S}$, most popular are:

- clipping
- flipping
- shift correction
- vector-representation

The underlying idea is to change the eigenvalue decomposition of the similarity matrix $\mathbf{S}$ such that negative eigenvalues are avoided.

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.

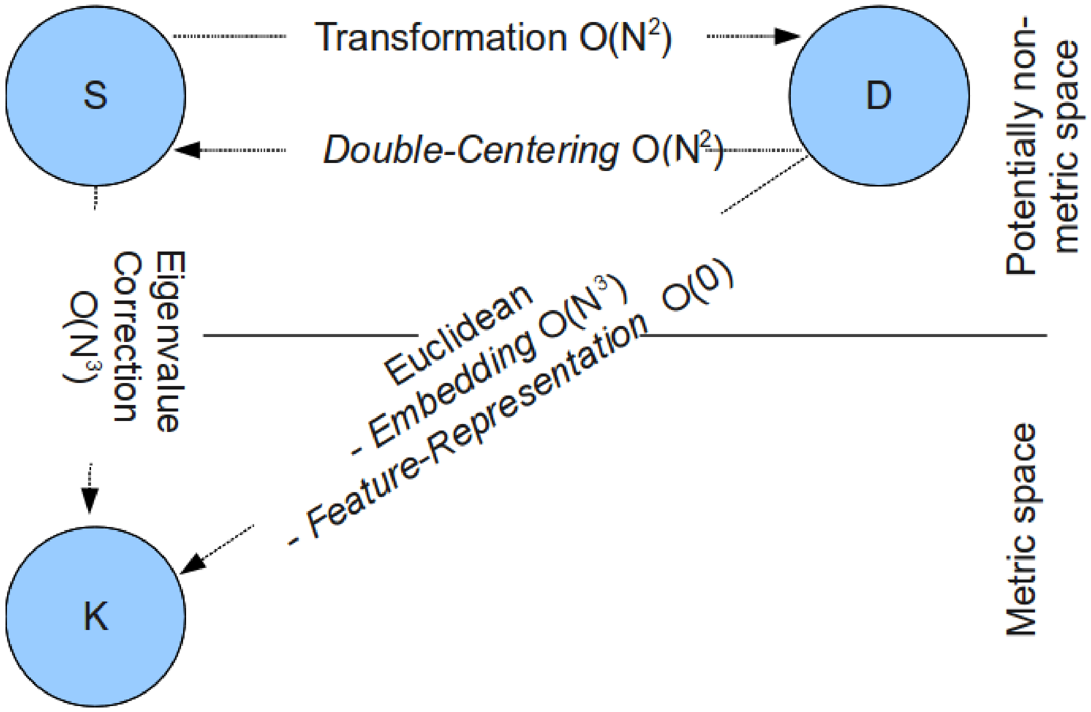

S
D
K
Transformation O(N²)
Double-Centering O(N²)
Eigenvalue Correction O(N³)
Euclidean
- Embedding O(N³)
- Feature-Representation O(0)
Potentially non-metric space
Metric space

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

Interestingly, some operations such as shift do not affect the location of global optima of important cost functions such as the quantization error [11], albeit the transformation can severely affect the performance of optimization algorithms [8]. The analysis in [14] indicates that for non-Euclidean dissimilarities corrections like above should be avoided.

A schematic view of the relations between $\mathbf{S}$ and $\mathbf{D}$ and its transformations is shown in Figure 1.

2.2 Analyzing dissimilarities by dedicated methods for small N

Alternatively, techniques have been introduced which directly deal with possibly non-psd dissimilarities. Given a symmetric dissimilarity with zero diagonal, an embedding of data in a pseudo-Euclidean vector space determined by the eigenvector decomposition of 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 S together with the square root of the absolute value of the eigenvalues, we obtain vectors $\mathbf{v}_i$ in 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 [13].

Vector operations can be directly transferred to pseudo-Euclidean space, i.e. we can deal with prototypes as linear combinations of data in this space. Hence we can use prototype-based learning explicitly in pseudo-Euclidean space since it relies on vector operations only. One problem of this explicit transfer 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. Because of this fact, we are interested in efficient techniques which implicitly refer to this embedding only or can be effectively employ the above mentioned techniques for metric spaces.

A further strategy is to employ so called relational or proximity learning methods as discussed in [24] for unsupervised data and in [6] for supervised prototype learning methods. The underlying models consist of prototypes, which are implicitely defined as a weighted linear combination of training points:

$$\mathbf{w}_j = \sum_i \alpha_{ji}\mathbf{v}_i \text{ with } \sum_i \alpha_{ji} = 1\,. \tag{1}$$

But this explicite 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 implicite computation of distances $d(\cdot,\cdot)$ during the model calculation based on the dissimilarity matrix $\mathbf{D}$ using weights α:

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

The prototypes identified therein, build the main parameters of the corresponding models, details can be found in the aforementioned papers. As shown e.g. in [8] 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 as 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 novel methods have been proposed quite recently, e.g. based on core-set techniques which can easily deal with multiple million points at very low costs, but this is only valid for metric similarity data.

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

Standard kernel and prototype methods for dissimilarity data 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 [21, 6, 6].

3.1 Nyström approximation for similarity data

Nyström approximation technique has been proposed in the context of kernel methods in [21]. 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

$$\lambda_i \approx \frac{\lambda_i^{(m)}}{m}, \quad \psi_i(\mathbf{y}) \approx \frac{\sqrt{m}}{\lambda_i^{(m)}}\mathbf{k}_y\mathbf{u}_i^{(m)}, \tag{3}$$

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 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 [22]. We denote these rows by $\mathbf{K}_{m,N}$. Using the formulas (3) 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

$$\tilde{\mathbf{K}} = \mathbf{K}_{m,N}^\top\mathbf{K}_{m,m}^{-1}\mathbf{K}_{m,N}. \tag{4}$$

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 [5] for preliminary work on this topic. 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 eigenvalues of $\mathbf{D}$, which can be negative for non-Euclidean distances. 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 4

$$\hat{\mathbf{D}} = \mathbf{D}_{N,m}\mathbf{D}_{m,m}^{-1}\mathbf{D}_{N,m}^\top. \tag{5}$$

It allows to approximate dissimilarities between a point $\mathbf{w}^k$ represented by a coefficient vector α_k and a data point $\mathbf{x}^i$ in the way

$$\begin{aligned} d(\mathbf{x}^i, \mathbf{w}^k) \approx & \left[\mathbf{D}_{m,N}^\top \left(\mathbf{D}_{m,m}^{-1}\left(\mathbf{D}_{m,N}\boldsymbol{\alpha}_k\right)\right)\right]_i \\ & -\frac{1}{2} \cdot \left(\boldsymbol{\alpha}_k^\top \mathbf{D}_{m,N}^\top\right) \cdot \\ & \left(\mathbf{D}_{m,m}^{-1}\left(\mathbf{D}_{m,N}\boldsymbol{\alpha}_k\right)\right) \end{aligned} \tag{6}$$

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. The amount m of the landmark points can be differed during the training, whether one is satisfied with the results or not.

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 [22]. It was found that best results can be obtained by chosing 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 appoximation can be used, as known already very well. 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 a dissimilarity matrix $\mathbf{D}$ can be effectively processed by prototype learning techniques, again using Nyström approximation in case of large N. Having however a huge set of effective, in parts convex kernel methods available for similarity data, we are also interested to use kernel methods for dissimilarity data. 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 whole matrix and thus has quadratical complexity, which would be prohibitive, if we were to use a linear time technique.

One way to achieve this transformation in linear time, is to use landmark multidimensional scaling (LMDS) [3] which was shown to be a Nyström technique as well [16]. The idea is to sample 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[1], that we can combine both transformations, double centering and Nyström approximation, and make use of the linearity of the both operations. 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 unwanted impact, since non mean centered similarities might lead to an inferior performance of the algorithms and, thus, our approach should be used instead.

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

$$\mathbf{S} = -\mathbf{JDJ}/2 \tag{7}$$

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 Euclidean if and only if $\mathbf{D}$ is positive semidefinite (psd).

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 $\mathbf{D}$.

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

[1] Taking the authors results about Nyström approximation for dissimilarity data into account.

$$\begin{aligned}\mathbf{S} \overset{Ny}{\approx} \hat{\mathbf{S}} \;&=\; -\frac{1}{2}\Big[\mathbf{D}_{N,m} \cdot \mathbf{D}^{-1}_{m,m} \cdot \mathbf{D}_{m,N} - \frac{1}{N}\mathbf{D}_{N,m} \cdot (\mathbf{D}^{-1}_{m,m} \cdot (\mathbf{D}_{m,N}\mathbf{1}))\mathbf{1}^\top \\ &\quad -\frac{1}{N}\mathbf{1}((\mathbf{1}^\top \mathbf{D}_{N,m}) \cdot \mathbf{D}^{-1}_{m,m}) \cdot \mathbf{D}_{m,N} + \frac{1}{N^2}\mathbf{1}((\mathbf{1}^\top \mathbf{D}_{N,m}) \cdot \mathbf{D}^{-1}_{m,m} \cdot (\mathbf{D}_{m,N}\mathbf{1}))\mathbf{1}^\top\Big]\end{aligned} \tag{8}$$

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

$$\begin{aligned}\hat{S}_{ij} \;&=\; -\frac{1}{2}\Big[\mathbf{D}_{i,m} \cdot \mathbf{D}^{-1}_{m,m} \cdot \mathbf{D}_{m,j} - \frac{1}{N}\sum_k \mathbf{D}_{k,m} \cdot \mathbf{D}^{-1}_{m,m} \cdot \mathbf{D}_{m,j} \\ &\quad -\frac{1}{N}\sum_k \mathbf{D}_{i,m} \cdot \mathbf{D}^{-1}_{m,m} \cdot \mathbf{D}_{m,k} + \frac{1}{N^2}\sum_{kl} \mathbf{D}_{k,m} \cdot \mathbf{D}^{-1}_{m,m} \cdot \mathbf{D}_{m,l}\Big],\end{aligned}$$

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

$$\begin{aligned}\hat{\mathbf{S}}_{m,m} \;&=\; -\frac{1}{2}\Big[\mathbf{D}_{m,m} - \frac{1}{N}\mathbf{1} \cdot \sum_k \mathbf{D}_{k,m} \\ &\quad -\frac{1}{N}\sum_k \mathbf{D}_{m,k} \cdot \mathbf{1}^\top + \frac{1}{N^2}\mathbf{1} \cdot \sum_{kl} \mathbf{D}_{k,m} \cdot \mathbf{D}^{-1}_{m,m} \cdot \mathbf{D}_{m,l} \cdot \mathbf{1}^\top\Big]\end{aligned}$$

$$\begin{aligned}\hat{\mathbf{S}}_{N,m} \;&=\; -\frac{1}{2}\Big[\mathbf{D}_{N,m} - \frac{1}{N}\mathbf{1} \cdot \sum_k \mathbf{D}_{k,m} \\ &\quad -\frac{1}{N}\sum_k \mathbf{D}_{N,m} \cdot \mathbf{D}^{-1}_{m,m} \cdot \mathbf{D}_{m,k} \cdot \mathbf{1}^\top + \frac{1}{N^2}\mathbf{1} \cdot \sum_{kl} \mathbf{D}_{k,m} \cdot \mathbf{D}^{-1}_{m,m} \cdot \mathbf{D}_{m,l} \cdot \mathbf{1}^\top\Big].\end{aligned}$$

It should be noted that $\hat{\mathrm{S}}$ 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 S and increases for smaller m.

4.2 Non-metric (dis-)similarities

In case of a non-metric D the transformation shown in equation 8 can still be used, but the obtained matrix $\hat{\mathrm{S}}$ is not a valid kernel. A strategy to obtain a valid kernel matrix $\hat{\mathrm{S}}$ 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 the needed computation dramatically. Since we now can apply the approximation on an arbitrary symmetric matrix, we can make the correction afterwards. To correct an already approximated similarity matrix $\hat{\mathrm{S}}$ it is sufficient to correct the eigenvalues of $\mathrm{S}_{m,m}$. Additionally to the Nyström approximation complexity $\mathcal{O}(m^2N)$, this results in an extra cost of $\mathcal{O}(m^3)$, which can be seen as well as a part of the matrix inversion, resulting altogether in $\mathcal{O}(m^2N)$ 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 discussed in section 2.1 to $\mathbf{\Lambda}^*$. The corrected approximated matrix $\hat{\mathbf{S}}^*$ is then simply

$$\hat{\mathbf{S}}^* = \mathbf{S}_{N,m}\mathbf{U}\left(\mathbf{\Lambda}^*\right)^{-1}\mathbf{U}^{\top}\mathbf{S}_{N,m}^{\top}. \tag{9}$$

This approach can also be used to correct dissimilarity matrices $\mathbf{D}$ by first approximating them, converting to similarities $\hat{\mathbf{S}}$ using equation 8 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 $\mathbf{D}$ using Equations from [13]

$$D_{ij}^2 = S_{ii} + S_{jj} - 2S_{ij}. \tag{10}$$

This obviously applies as well to the approximated and corrected matrices $\hat{\mathbf{S}}^*$ and $\hat{\mathbf{D}}^*$.

$$\hat{\mathbf{D}}^* = \mathbf{D}_{N,m}^*\left(\mathbf{D}_{m,m}^*\right)^{-1}\mathbf{D}_{N,m}^{*\top}. \tag{11}$$

Usually the algorithms are learned on 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 extensions, 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 9 and 11 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 [2] 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 together with the new data build the matrix $\mathbf{S}_{N,m}$. Rewriting this in the Nyström framework makes it more cleaner, without the need to compute the projection matrix and with an additional possibility to compute the similarities between the new points.

As a last point it should be mentioned that corrections like flipping, clipping or others are still under discussion and cases are reported were the transformations had negative impact on the model performance [14]. Accordingly, it is still best to avoid such transformations and focus either on the appropriate methods, less sensitive to such effects, e.g. dissimilarity learners for dissimilarity data or to use (dis-)similarity measures which imply a metric space. Additionally the selection of landmark points can be complicated for more complex data sets and only random samples maybe not sufficient to cover the whole data properties as discussed in [22]. Further for very large data sets (e.g. some 100 million points or more) the Nyström approximation may still be too costly and some other strategies have to be found.

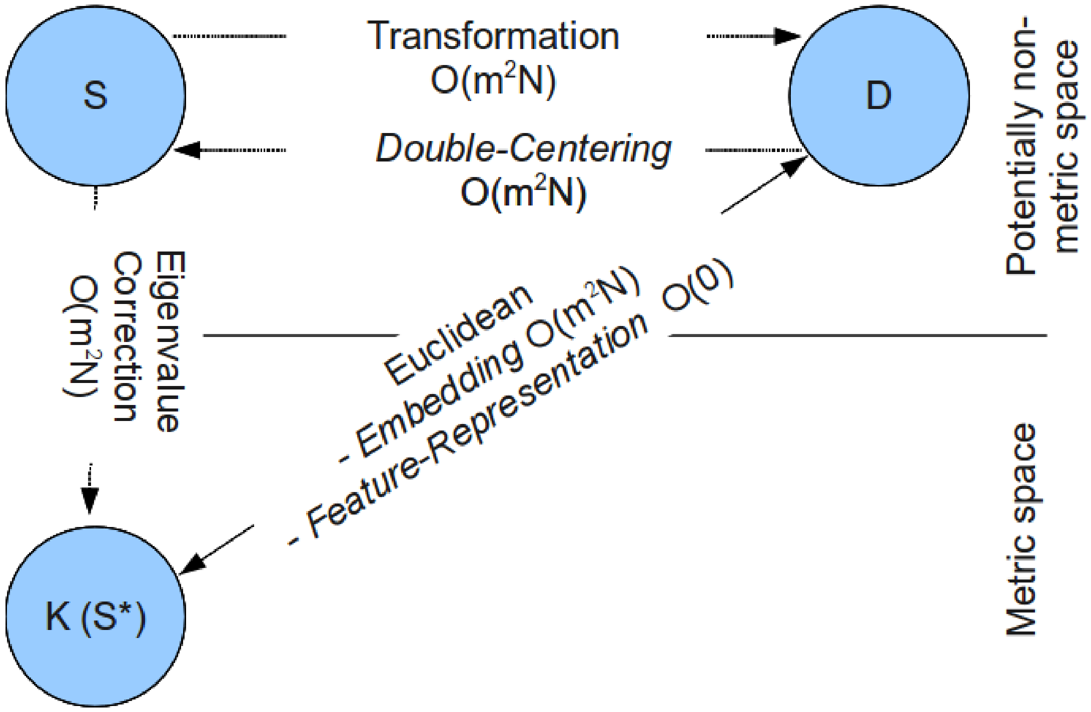


Figure 2: Schema to illustrate the relation between similarities and dissimilarities using the Nyström approximation. The costs are now substantially smaller $m \ll N$.

5 Outlook and Conclusions

In this report we discussed the relation between similarity and dissimilarity data and effective ways to move between the different representations in a systematic way. Using the Nyström approximation in the discussed way, effective and accurate transformations are possible, in contrast to former slightly linked work like L-MDS. Methods from both domains, namely kernel approaches but also dissimilarity learners become accessible for both types of data. The specific parametrization of the Nyström approximation is already studied elsewhere [23, 10, 22] but there are still multiple open issues which we will focus on in later work. Especially the, potentially negative, impact of the eigenvalue correction is still not sufficiently discuessed, although some initial steps in this line are published in [14]

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] Wesam Barbakh and Colin Fyfe. Online clustering algorithms. *Int. J. Neural Syst.*, 18(3):185–194, 2008.

[2] Yihua Chen, Eric K. Garcia, Maya R. Gupta, Ali Rahimi, and Luca Cazzanti. Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*, 10:747–776, 2009.

[3] Vin de Silva and Joshua B. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *NIPS*, pages 705–712. MIT Press, 2002.

[4] Roberto Gil-Pita and Xin Yao. Evolving edited k-nearest neighbor classifiers. *Int. J. Neural Syst.*, 18(6):459–467, 2008.

[5] A. Gisbrecht, B. Mokbel, and B. Hammer. The Nystrom approximation for relational generative topographic mappings. In *NIPS Workshop*, 2010.

[6] A. Gisbrecht, B. Mokbel, F.-M. Schleif, X. Zhu, and B. Hammer. Linear time relational prototype based learning. *Journal of Neural Systems*, page accepted, 2012.

[7] Alexander N. Gorban and Andrei Yu. Zinovyev. Principal manifolds and graphs in practice: from molecular biology to dynamical systems. *Int. J. Neural Syst.*, 20(3):219–232, 2010.

[8] Barbara Hammer and Alexander Hasenfuss. Topographic mapping of large dissimilarity data sets. *Neural Computation*, 22(9):2229–2284, 2010.

[9] Vilen Jumutc, Pawel Zayakin, and Arkady Borisov. Ranking-based kernels in applied biomedical diagnostics using a support vector machine. *Int. J. Neural Syst.*, 21(6):459–473, 2011.

[10] Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. On sampling-based approximate spectral decomposition. In Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman, editors, *ICML*, volume 382 of *ACM International Conference Proceeding Series*, page 70. ACM, 2009.

[11] Julian Laub, Volker Roth, Joachim M. Buhmann, and Klaus-Robert Müller. On the information and representation of non-euclidean pairwise data. *Pattern Recognition*, 39(10):1815–1826, 2006.

[12] Britta Mersch, Tobias Glasmachers, Peter Meinicke, and Christian Igel. Evolutionary optimization of sequence kernels for detection of bacterial gene starts. *Int. J. Neural Syst.*, 17(5):369–381, 2007.

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

[14] Elzbieta Pekalska, Robert P. W. Duin, Simon Günter, and Horst Bunke. On not making dissimilarities euclidean. In Ana L. N. Fred, Terry Caelli, Robert P. W. Duin, Aurélio C. Campilho, and Dick de Ridder, editors, *SSPR/SPR*, volume 3138 of *Lecture Notes in Computer Science*, pages 1145–1154. Springer, 2004.

[15] François Petitjean, Florent Masseglia, Pierre Gançarski, and Germain Forestier. Discovering significant evolution patterns from satellite image time series. *Int. J. Neural Syst.*, 21(6):475–489, 2011.

[16] J. Platt. Fastmap, metricmap, and landmark mds are all nyström algorithms, 2005.

[17] F.-M. Schleif, T. Villmann, B. Hammer, and P. Schneider. Efficient kernelized prototype-based classification. *Journal of Neural Systems*, 21(6):443–457, 2011.

[18] F.-M. Schleif, X. Zhu, A. Gisbrecht, and B. Hammer. Fast approximated relational and kernel clustering. In *Proceedings of ICPR 2012*, page accepted, 2012.

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

[20] Liangdong Shi, Yinghuan Shi, Yang Gao, Lin Shang, and Yubin Yang. Xcsc: a novel approach to clustering with extended classifier system. *Int. J. Neural Syst.*, 21(1):79–93, 2011.

[21] Christopher K. I. Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *NIPS*, pages 682–688, 2000.

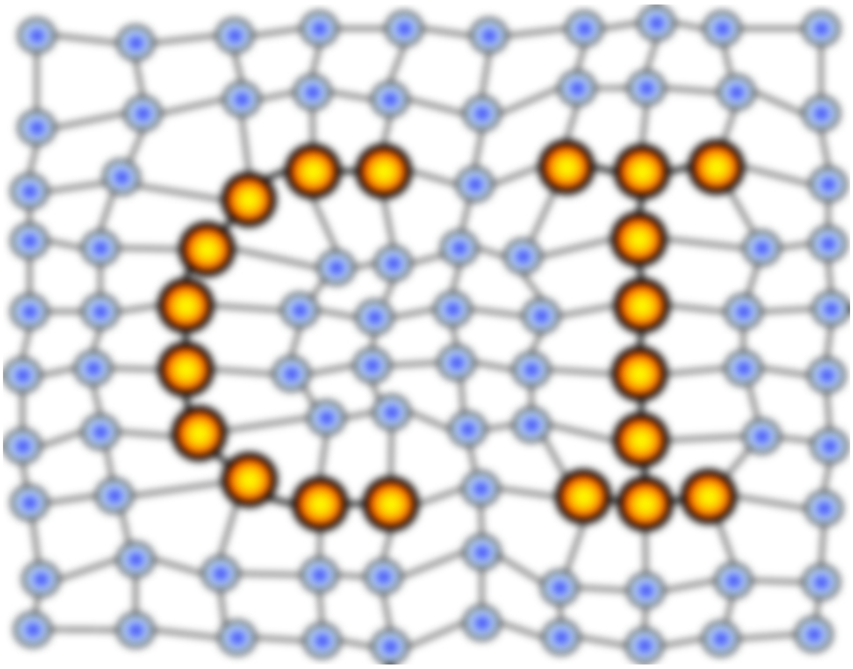
[22] Kai Zhang and James T. Kwok. Clustered nyström method for large scale manifold learning and dimension reduction. *IEEE Transactions on Neural Networks*, 21(10):1576–1587, 2010.

[23] Kai Zhang, Ivor W. Tsang, and James T. Kwok. Improved Nystrom low-rank approximation and error analysis. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 1232–1239, New York, NY, USA, 2008. ACM.

[24] X. Zhu, A. Gisbrecht, F.-M. Schleif, and B. Hammer. Approximation techniques for clustering dissimilarity data. *Neuro Computing*, 90:72–84, 2012.

MACHINE LEARNING REPORTS

Report 04/2012



Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

Prof. Dr. rer. nat. Thomas Villmann
University of Applied Sciences Mittweida
Technikumplatz 17, 09648 Mittweida, Germany
• http://www.mni.hs-mittweida.de/

Dr. rer. nat. Frank-Michael Schleif
University of Bielefeld
Universitätsstrasse 21-23, 33615 Bielefeld, Germany
• http://www.cit-ec.de/tcs/about

▽ Acknowledgments

We would like to thank the reviewers for their time and patience.