Adaptive spectrum transformation by topology preserving on indefinite proximity data

Khadijeh Sadatnejad[a], Saeed Shiry Ghidary [a]*

[a] *Computer Engineering and Information Technology Department, Amirkabir University of Technology,Hafez ave., Tehran, Iran*

ABSTRACT

Similarity-based representation extensively generates indefinite matrices, which are not consistent with the framework of classical kernel-based learning methods. In this paper, we present an adaptive spectrum transformation method to provide a positive semi-definite (*psd*) kernel, which is consistent with the intrinsic geometry of proximity data. In the proposed method, an indefinite similarity matrix is rectified by maximizing EF criterion, which represents the similarity of resulting feature space to the Euclidean space. This is achieved by modifying volume elements by applying a conformal transform over the similarity matrix. Several experiments have been performed to evaluate the performance of the proposed method in comparison with *flip*, *clip*, *shift*, and *square* spectrum transformation techniques on some (dis)similarity matrices. Applying the resulting *psd* matrices as kernels in dimensionality reduction and clustering setups confirms the success of the proposed approach in adapting to the data and preserving the topological information. The experiments show that in classification setup, the superiority of the proposed method is considerable where the negative eigen-fraction of the similarity matrix is significant.

1. Introduction

In this paper, we propose an adaptive mapping to transform the indefinite similarity matrices to positive semi-definite (*psd*) kernels. Similarity matrices are used for representing the similarity between instances in different applications, such as natural language processing, information retrieval, Bioinformatics, and computer vision (Schleif and Tino, 2015). Similarity-based representation extensively generates non-*psd* matrices. In a general categorization, three sources can be identified for negative eigenvalues of similarity matrices; 1) using non-Euclidean metrics, where data points lie on a nonlinear manifold, 2) using non-metric distances, for example, in the case of extended objects, and 3) the noise, which is the result of numerical or measurement inaccuracies (Xu, 2013). For example, measuring pairwise similarity between protein sequences and DNA in Bioinformatics applications using Dynamic Time Warping (DTW) (Noma and Shimodaira, 2002), the Smith-Waterman algorithm, or BLAST (Altschul et al., 1990) generates indefinite (non-*psd*) kernel matrices. Dynamic partial function (Qamra et al., 2005) and earthmover's distance function (Rubner et al., 2000), which are efficient measures for representing perceptual dissimilarity between instances in image/video retrieval application, are non-metric. An effective measure for calculating the distance between two characters in handwritten-digit recognition is the tangent distance, which despite its strength in overcoming the translation and rotation phenomenon (Simard et al., 1993), its induced kernel might be indefinite (Haasdonk, 2005).

The two major approaches for analyzing proximity data are treating the similarity matrix as inner products between samples and considering the similarities with each sample as its feature vector (Duin et al., 1997; Pękalska and Duin, 2002; Schleif and Tino, 2015). One popular technique in the former approach, which has led to great success, is using the kernel-based learning techniques (Pekalska et al.,2002; Pelillo, 2013; Schleif and Tino, 2015; Wu et al., 2005). In this approach, for learning over the similarity matrix, it is assumed that the proximities represent the inner product in a Hilbert space. Projection to Reproducing Kernel Hilbert space (RKHS) using an implicit feature mapping $\varphi(.)$, imposes the *psd* constraint to kernel functions. This constraint is required for finding the optimal solution for optimization problem in induced feature space. Using indefinite kernels in methods, which relied on empirical risk minimization (ex. Support Vector Machine (SVM) (Vapnik, 2013)) leads to a non-convex optimization problem. Solving this problem, which

* Corresponding author. Tel.: +0-98-21-64542737; fax: +0-98-21-66495521; e-mail: shiry@aut.ac.ir

produces a saddle point solution, does not guarantee the minimization of risk function (Wu et al., 2005). In addition, embedding into real-valued Euclidean space is not possible due to negative eigenvalues of the non-*psd* similarity matrices. In such cases, projection to pseudo-Euclidean space, as a solution that does not distort the original distances, leads to a non-metric embedding space. Therefore, the popular geometric learning techniques, which are adapted to a vector space, cannot be used for learning in resulting embedding space. The problem of learning with non-*psd* similarity matrix has been addressed by two major approaches: transforming a non-*psd* similarity matrix to become a *psd* matrix, and providing methods that can adapt to non-metric data without being sensitive to violations of metric conditions.

Spectrum *clip*, *flip*, *shift*, and *square* transformations, which are some spectrum transformation techniques, provide *psd* similarity matrices from indefinite kernels. All of these methods somehow neglect the negative eigenvalues; *clip* replaces them with zero, *flip* replaces them with their absolute value, *shift* adds a constant value to eigenvalues to make them positive, and *square* uses their squared values. Neglecting negative eigenvalues or transforming them without topological consideration will lead to missing geometrical information (Pekalska et al., 2002; Pękalska et al., 2006).

On the other hand, topology of data points plays an essential role in different machine learning techniques. For instance, representing and preserving the structure of the dataset are the main challenges in dimensionality reduction techniques (Lee and Verleysen, 2007). In addition, solving a classification problem can be based on the compactness hypothesis (Arkadev and Braverman, 1967; Duin, 1999), which states similar objects have close representations, in other words, distribution of classes are affected by the topology of data points.

These considerations motivate us to handle negative eigenvalues using spectrum transformation from a geometrical point of view. For this purpose, we propose an adaptive approach for rectifying an indefinite similarity matrix while preserving geometrical information provided by all eigenvalues. We manipulate the indefinite similarity matrix using a conformal transform such that the non-Euclidean characteristics of data points decrease. Emphasis on preserving the topology in spectrum transformation results in considerable performance improvement of classical machine learning techniques for analyzing the proximity data in comparison with common competitors.

The remainder of this paper is organized as follows: We first review the major methods in spectrum transformation in section 2. Then, in section 3 we review some mathematical preliminaries, which are required for having a good understanding of the proposed approach for adaptive spectrum transformation that is described in section 4. The experimental results for comparing the proposed approach with other spectrum transformation techniques on standard benchmarks are reported and discussed in section 5. We conclude the results in section 6.

2. Related Works

In this section we review four major spectrum-transformation techniques, which generate *psd* matrices from non-*psd* ones.

2.1. Spectrum clip

Spectrum *clip* generates a *psd* matrix from an indefinite similarity matrix by setting all negative eigenvalues to zero (Chen and Gupta et al., 2009; Wu et al., 2005). The main idea behind *clip* is that the negative eigenvalues of similarity matrix are generated due to the noise and therefore *clip* acts as a denoising process (Wu et al., 2005).

Let $S = U\ \Lambda\ U^T$ represents the similarity matrix, where Λ is a diagonal matrix that the eigenvalues of S (denoted as λ_i) are its diagonal entries, and U is the matrix of eigenvectors of S. Applying a *clip* transformation on S will produce:

$$S_{clip} = U\ diag(max(\lambda_1, 0)\ , \dots , max(\lambda_n, 0))U^T. \qquad (1)$$

Obviously, in the cases that negative eigenvalues are not negligible, the *clip* spectrum transformation leads to losing a significant part of the information that were provided by negative eigenvalues (Chen and Garcia, et al., 2009). Cliping a matrix is equal to approximating a non-*psd* matrix by a *psd* matrix in terms of Frobenious norm (Wu et al., 2005).

2.2. Spectrum flip

In contrast with researches that consider negative eigenvalues of similarity matrix as the result of noise; multiple researches manifest that negative eigenvalues may convey significant topological and discrimination information (Laub and Müller, 2004; Laub, 2006). In order to preserve the information of negative eigenvalues, spectrum *flip* replaces each eigenvalue by its absolute value (Wu et al., 2005):

$$S_{flip} = U\ diag(|\lambda_1|, \dots , |\lambda_n|)U^T. \qquad (2)$$

Flip transformation is equal to projecting data into Krein space $\kappa = \mathcal{H}_+ \oplus \mathcal{H}_-$ where the similarity is defined as $< x, y > =< x_+, y_+ >_{\mathcal{H}_+} - < x_-, y_- >_{\mathcal{H}_-}$ in this space. The space κ is the direct sum of two disjoint Hilbert space denoted by $\mathcal{H}_+$ and $\mathcal{H}_-$, where for any x and y that are the members of κ we have $x = x_+ + x_-$ and $y = y_+ + y_-$such that $y_+, x_+ \in \mathcal{H}_+$ and $y_-, x_- \in \mathcal{H}_-$.

2.3. Spectrum shift

Spectrum *shift* is a popular approach for providing a *psd* matrix from non-*psd* kernel by adding a constant value to all eigenvalues. In this approach, any eigenvalue is shifted by the magnitude of the minimum of the eigenvalues:

$$S_{shift} = U\ diag(\lambda_1 + |min(\lambda_{min}(S), 0)|\ , \dots , \lambda_n + |min(\lambda_{min}(S), 0)\ |)\ U^T. \qquad (3)$$

In comparison with *clip* and *flip* spectrum transformation techniques, this approach merely changes self-similarity and does not modify similarity between different samples (Roth et al., 2003).

2.4. Spectrum square

The *square* strategy for spectrum transformation is recently developed by Muñoz and Diego, 2006.This approach changes the eigenvalues of similarity matrix by squaring them:

$$S_{square} = U\ diag\left({\lambda_1}^2, \dots , {\lambda_n}^2\right)U^T. \qquad (4)$$

It is claimed that this transformation produces a kernel, which if it is used as a kernel for SVM classifier it would lead to promising results.

3. Mathematical preliminaries

In this section, we describe some basic concepts of kernel's geometry that are required for better understanding of this paper. First, we review the formulation of generating a similarity matrix from distances, and then describe the relationship between a kernel and volume element corresponding to the induced metric in input space.

Let D be the pairwise dissimilarity matrix between N samples, which are denoted by $\{x_i\}_{i=1}^{N}$. The similarity matrix (including the similarities between the pairs of data points) is computed by applying double centering method (Cox and Cox, 1994) to the dissimilarity matrix:

$$S = -(1/2)JD^2J \tag{5}$$
$$J = I_{N\times N} - (1/N)1_N \times {1_N}^T$$

where S is the similarity matrix, $I_{N\times N}$ is an $N\times N$ identity matrix, and 1_N is a column vector that all its elements are 1.

A positive semi definite similarity matrix can be considered as a kernel ($K = [k(x_i, x_j)]$) in an empirical feature space, which is corresponding to an inner product in Hilbert space (i.e., $k(x_i, x_j) =< \varphi(x_i), \varphi(x_j) >$, where $\varphi(.)$ is an implicit feature mapping from input space to an implicit feature space).

The kernel function $K(.,.)$ induces a Riemannian metric to input space using feature mapping $\varphi(.)$, which is computed as (Amari and Wu, 1999; Wu and Amari, 2002):

$$g_{ij}(x, x') = \partial/\partial v_i\, \partial/\partial v'_j K(x, x')|_{x=x'} \tag{6}$$

where v_i denotes i^{th} basis of x vector. Eq. (6) is written in Einstein summation notation. The volume element corresponding to the induced metric in input space is computed as (Wu, Amari, 2002):

$$dV = \sqrt{g(x)}dv_1 \dots dv_n \tag{7}$$

where $g(x)$ represents the determinant of the matrix whose elements are g_{ij} and dV denotes the volume element. The expression $\sqrt{g(x)}$ is a factor that controls the expansion and contraction of volume elements (Williams et al., 2007). This equation confirms the influence of modifying the feature mapping and consequently kernel function on volume element.

4. Adaptive conformal spectrum transformation

In this section, we describe our proposed approach to transform an indefinite similarity matrix to a kernel that satisfies Mercer's condition (Burges, 1998).

We first describe the proposed transformation to rectify non-Euclidean characteristics of an indefinite similarity matrix, then introduce a criterion for defining an objective function, and finally find appropriate optimization technique for solving the objective function.

4.1. Conformal mapping

Given an indefinite similarity matrix S_0, we wish to find a *psd* matrix using an adaptive spectrum transformation that preserves the topology of the data.

We begin by applying a conformal transform, which is a local topology preserving transformation, over the centralized similarity matrix (S_0). The conformal transformation preserves the structure by keeping angles unchanged. Eq. (6) and Eq. (7) imply that modifying the similarity matrix leads to changes in the induced Riemannian metric and, consequently, the volume element.

A conformal map C, applied to similarity matrix S_0, will produce matrix S:

$$S = C \times S_0 \times C, \tag{8}$$
$$C = diag([c(x_1), \dots, c(x_N)])$$

where C is an $N \times N$ diagonal matrix with $c(x_i)$ as its diagonal entries (Eq. (9)). N represents the number of training data, which are denoted by x_i , and $c(x_i)$ is the conformal transformation of x_i. It is calculated based on dissimilarity between the corresponding sample and some or all of the other samples using following formula (Xiong at al., 2005):

$$c(x) = \alpha_0 + \sum_{j=1}^{M} \alpha_j e^{-\delta\|x-m_j\|^2} = S_1(x)^T\alpha, \tag{9}$$
$$S_1(x) = \left[1 \ e^{-\delta\|x-m_1\|^2} \dots \ e^{-\delta\|x-m_M\|^2}\right]^T$$
$$\alpha = [\alpha_0 \dots \alpha_M]^T$$

where m_js called empirical cores can be selected randomly or based on the geometry of the training dataset and M denotes the number of cores. The $\alpha_0, \dots, \alpha_M$, which denote the weight or contribution of dissimilarity to each core (i.e. $\|x - m_j\|$) in $c(x)$, are the unknown parameters of our transformation.

To achieve our goal, the expansion and contraction of the volume element using conformal transformation should result a Euclidean similarity matrix. Therefore, we encounter with a kernel parameter selection problem. For an appropriate model selection, we express this problem as an optimization problem and introduce a proper criterion to modify the metric such that it results in a *psd* similarity matrix.

4.2. Euclidean factor criterion

To deal with negative eigenvalues of the similarity matrix, which are the result of non-Euclidean characteristics of the feature space, we introduce a criterion based on this fact that magnitudes of negative eigenvalues represent the departure from Euclidean behavior. For this purpose, we introduce Euclidean Factor (EF) criterion, which shows similarity of the feature space to the Euclidean space. This criterion is based on this fact that dataset shows Euclidean behavior if and only if its corresponding grammian matrix is *psd*. Therefore, we define EF criterion to include the overall contribution of negative eigenvalues of the similarity matrix:

$$EF\ (S) = \sum_{\lambda_i<0} \lambda_i(\text{S}) \Big/ \sum_i |\lambda_i(\text{S})|. \tag{10}$$

*4.3. Optimizing the **EF** criterion*

To maximize $EF(S(\alpha))$, first we show that $EF(S(\alpha))$ is compatible with the following fractional programming problem:

$$maximize\ EF(S(\alpha)) = maximize\ \ f(\alpha)/g(\alpha) \tag{11}$$
$$= \alpha^T P\alpha / \alpha^T Q\alpha$$

where $f(\alpha)$ and $g(\alpha)$ are continuous and real values in $\mathbb{R}^n \backslash \{0\}$.

n denotes the length of vector α or the number of unknown parameters, $\alpha_i s$. In addition, $g(\alpha) > 0$ for all $\alpha \in A$, where A is a convex set.

Lemma 1: let $\lambda_1(\alpha),\ \lambda_2(\alpha), \ldots, \lambda_N(\alpha)$ be the eigenvalues of $S(\alpha)$ matrix; then $\lambda_i(\alpha)$ maximally has degree 2 in α.

Proof: Given that, $\lambda_i(\alpha)$s are the roots of the characteristic polynomial of matrix S, which depends on unknown parameter α, the characteristic polynomial, $p(\lambda, \alpha)$, can be written as (Horn and Johnson, 2012):

$$\begin{gathered} p(\lambda, \alpha) = det(S(\alpha) - \lambda I) \\ = \sum_{i=1}^{N} (-1)^{i+l}\ t_{il}\ T_{il}\ \ 1 \leq l \leq N \\ T = [t_{ij}] = S(\alpha) - \lambda I \end{gathered} \tag{12}$$

where T_{il}, which denotes the sub-matrix of T, is derived by removing the i^{th} column and l^{th} row of matrix S. Computation of this inductive presentation can begin by computing the determinant of a single entry matrix. For a 1×1 matrix we have:

$$\begin{gathered} \mathrm{T}_{1\times 1} = [\mathrm{c(x)\ S_0(x,x)c(x)} - \lambda] \\ \mathrm{p}(\lambda, \alpha) = \alpha^{\mathrm{T}}\ \mathrm{S'(x,x)}\alpha - \lambda = 0, \\ \mathrm{S'(x_i, x_j)} = \ \mathrm{S_1^T(x_i)\ \ S_0(x_i, x_j)\ S_1(x_j, :)} \ \rightarrow \\ \lambda = \ \alpha^{\mathrm{T}}\mathrm{S'(x,x)}\alpha. \end{gathered} \tag{13}$$

The characteristic polynomial of a 2×2 transformed matrix is defined as:

$$\begin{gathered} T_{2\times 2} = \begin{bmatrix} \alpha^T S'(x,x)\alpha - \lambda & \alpha^T S'(x,y)\ \alpha \\ \alpha^T\ S'(y,x)\alpha & \alpha^T\ S'(y,y)\ \alpha - \lambda \end{bmatrix} \\ p(\lambda, \alpha) = \alpha^T\ S'(x,x)\alpha\ \alpha^T\ S'(y,y)\ \alpha + \lambda^2 \\ -\lambda(\alpha^T\ S'(x,x)\alpha + \alpha^T\ S'(y,y)\alpha) \\ -\ \alpha^T\ S'(x,y)\ \alpha\ \alpha^T S'(y,x)\alpha \end{gathered} \tag{14}$$

and so on. Due to equations 12, 13, and 14, the characteristic polynomial of an $N \times N$ matrix has degree N in λ and degree $2N$ in α. Therefore, the roots of the characteristic polynomial cannot be larger than 2 (i.e. $\lambda_i(\alpha) \in O(\alpha^2)$).

Theorem 1: All the eigenvalues of $S(\alpha)$ are quadratic polynomials of α.

Proof: The determinant of $S(\alpha)$ is computed using the following inductive representation:

$$det(S(\alpha)) = \sum_{i=1}^{N} (-1)^{i+l}\ s_{il}\ S_{il}\,, \qquad 1 \leq l \leq N \tag{15}$$

where s_{il} representes $(i,l)^{\text{th}}$ entry of S matrix and S_{il} is a sub-matrix of S resulted from removing i^{th} column and l^{th} row of S matrix. Computing $det(S(\alpha))$ using its inductive representation shows that it has degree of $2N$ in α. On the other hand, $det(S(\alpha))$ is equal to the product of eigenvalues of $S(\alpha)$:

$$det\ (S(\alpha)) = \prod_{i=1}^{N} \lambda_i(\alpha) \tag{16}$$

where $\lambda_i(\alpha)$ is an eigenvalue of $S(\alpha)$.

Now assume that:

$$\exists\ \lambda_i(\alpha): \lambda_i(\alpha) \in O(\alpha^l), l < 2. \tag{17}$$

Since $det\big(S(\alpha)\big) = \prod_{i=1}^{N} \lambda_i(\alpha) \in O(\alpha^{2N})$, there should be at least one eigenvalue, such that:

$$\exists\ \lambda_i(\alpha): \lambda_i(\alpha) \in \theta(\alpha^l), l > 2 \tag{18}$$

which is in contradiction with lemma 1. This contradiction proves that the proposition (18) is false. Therefore, we infer that

$$\nexists\ \lambda_i(\alpha): \lambda_i(\alpha) \in O(\alpha^l), l < 2. \tag{19}$$

From Eq. (19) and lemma 1 we infer that all eigenvalues of $S(\alpha)$ are quadratic polynomials (i.e., $\lambda_i(\alpha) =\ \alpha^T\ S_i^{''}\ \alpha$). Therefore, for positive eigenvalues ($\lambda_i(\alpha) > 0$), we have $\alpha^T\ S_i^{''} \alpha > 0$. According to the definition of positive definiteness (Horn and Johnson, 2012), we can infer $S_i^{''}$ is positive definite. In the same way, $S_i^{''}$ is negative definite for a negative eigenvalue.

Lemma 2. $EF(S(\alpha))$ is a concave-convex quadratic polynomial.

Let $\lambda_i(\alpha)$ be the eigenvalues of $S(\alpha)$ that are sorted in ascending order; $\lambda_1(\alpha) \leq \lambda_2(\alpha) \leq \cdots \leq\ \lambda_N(\alpha)$. Assume $nNeg$ denotes the number of negative eigenvalues of $S(\alpha)$ matrix. Since C is a non-singular matrix, $nNeg$ is equal to the number of negative eigenvalues of S_0 (Horn and Johnson, 2012). Assume

$$\begin{gathered} f(\alpha) =\ \textstyle\sum_{i=1}^{nNeg} \lambda_i(\alpha)\ \ , \\ g(\alpha) =\ \textstyle\sum_{i=1}^{N} |\lambda_i(\alpha)| = -\sum_{i=1}^{nNeg} \lambda_i(\alpha) + \\ \textstyle\sum_{i=nNeg+1}^{N} \lambda_i(\alpha). \end{gathered} \tag{20}$$

The summation of PD/ND matrices produces a PD/ND matrix. Therefore, we have:

$$\sum_{i=1}^{nNeg} \lambda_i(\alpha) =\ \alpha^T \sum_{i=1}^{k} S_i^{''}\ \alpha$$

where $\sum_{i=1}^{nNeg} S_i^{''}$ is a ND matrix. In addition, considering that the negation of an ND matrix is a PD matrix, we can show that the denominator of $EF(S(\alpha))$ is a quadratic polynomial with PD coefficient matrix:

$$\begin{aligned} \sum_{i=1}^{N} |\lambda_i(\alpha)| &= \sum_{i=1}^{nNeg} -\alpha^T S_i^{''} \alpha \\ &\quad + \sum_{i=nNeg+1}^{N} \alpha^T S_i^{''}\ \alpha \\ &= \ \alpha^T \left(-\sum_{i=1}^{nNeg} S_i^{''}\right) \alpha +\ \alpha^T \sum_{i=nNeg+1}^{N} S_i^{''}\ \alpha \\ &= \ \alpha^T (-\sum_{i=1}^{nNeg} S_i^{''} + \sum_{i=nNeg+1}^{N} S_i^{''})\alpha \end{aligned} \tag{22}$$

where $-\sum_{i=1}^{nNeg} S_i^{"}$ and $\sum_{i=nNeg+1}^{N} S_i^{"}$ are PD matrices and therefore $(-\sum_{i=1}^{nNeg} S_i^{"} + \sum_{i=nNeg+1}^{N} S_i^{"})$ is PD.

Therefore, we can formulate the objective function as:

$$\begin{aligned} EF\big(S(\alpha)\big) &= f(\alpha)/g(\alpha) \\ &= \alpha^T \left(\sum_{i=1}^{nNeg} S_i^{"}\right)\alpha \\ &/ \alpha^T \left(-\sum_{i=1}^{nNeg} S_i^{"} + \sum_{i=nNeg+1}^{N} S_i^{"}\right)\alpha \end{aligned} \tag{23}$$

where coefficient matrices in the numerator and the denominator are ND and PD, respectively and the $EF\big(S(\alpha)\big)$ is a concave-convex fractional problem.

As proved in lemma 2, maximizing the EF criterion is compatible with a concave-convex fractional problem with quadratic numerator and denominator that its optimal solution can be found using Dinkelbach's algorithm (Dinkelbach, 1967). Although lemma 2 and Dinkelbach's algorithm guarantee finding the optimum of Eq. (11), but increasing the size of similarity matrix and consequently computing the eigenvalues of S matrix, which depends on parameter α, would be an intractable problem. Therefore, we use numerical methods for approximating the optimal value of the objective function.

5. Experimental evaluations

To assess the proposed spectrum transformation method, we examine it in three different experimental setups. First, we evaluate it in a dimensionality reduction problem with an artificial dataset to illustrate the effectiveness of the proposed method in preserving geometrical information. We run two other evaluations of the proposed approach over real datasets represented by the dissimilarities between instances in clustering and classification setups. In this section, we first describe the datasets involved in our experiments, and then express more details about the experiments.

5.1. Datasets

To evaluate our proposed method we use one dataset composed of dissimilarity between points lying over a spherical manifold, and six datasets that are given in terms of proximity data.

First, we run our experiments on a fishbowl dataset composed of about 1000 equi-distance instances that are sampled from a spherical manifold with radius 1.

For real world problems, we have chosen six different datasets representing dissimilarities in a wide range of applications:

- Catcortex dataset describes connection strengths between 65 cortical areas of a cat from auditory, somatosensory, visual, and frontolimbic regions (Scannell et al., 1995). This dataset is represented as a 65×65 dissimilarity matrix and is used in classification (Graepel et al., 1999) and clustering (Denœux and Masson, 2004) applications.
- Proteins dataset consists of dissimilarity between 226 protein sequences that belong to four classes of globins, including heterogeneous globin (G), hemoglobin-A (HA), hemoglobin-B (HB), and myoglobin (M). The dissimilarity between the protein sequences are compared based on the concept of evolutionary distance ((Graepel et al., 1999). This dataset is used in both clustering and classification problems.
- Music-EMD and Music-PTD contain distances between music pieces that are measured by the Earth Mover's Distance (EMD) and the Proportional Transportation Distance (PTD) respectively. The dataset contains dissimilarities between 22 music pieces from Georg Friedrich Händel and 28 pieces from Joseph class (Typke et al., 2003).
- Kimia dataset contains dissimilarity between 72 binary images belonging to 6 different classes. A modified Hausdorff distance is used for measuring the pairwise dissimilarity between distances (Pekalska et al., 2002; Sebastian et al., 2001).
- UNIPEN-DTW contains a fraction composed of 250 handwritten sequences from 5 different classes of the original UNIPEN dataset. Dynamic-time-warping measure is used for measuring dissimilarities (Bahlmann et al., 2002; I Guyon et al., 1994).
- USPS-TD composed of 250 samples, including 1-250, 251-500, 501-750 and 751-1000 subsets of the original USPS dataset. Tangent distance is used for measuring dissimilarities (Haasdonk and Keysers, 2002; Keysers et al., 2004). This problem is considered as a binary classification problem by assigning the digits 0,1,2,3, and 4 to class 1 and considering the digits 5, 6, 7, 8, and 9 as class 2.

The specifications of these datasets are briefly reviewed in table 1.

Table 1. Specifications of benchmark datasets

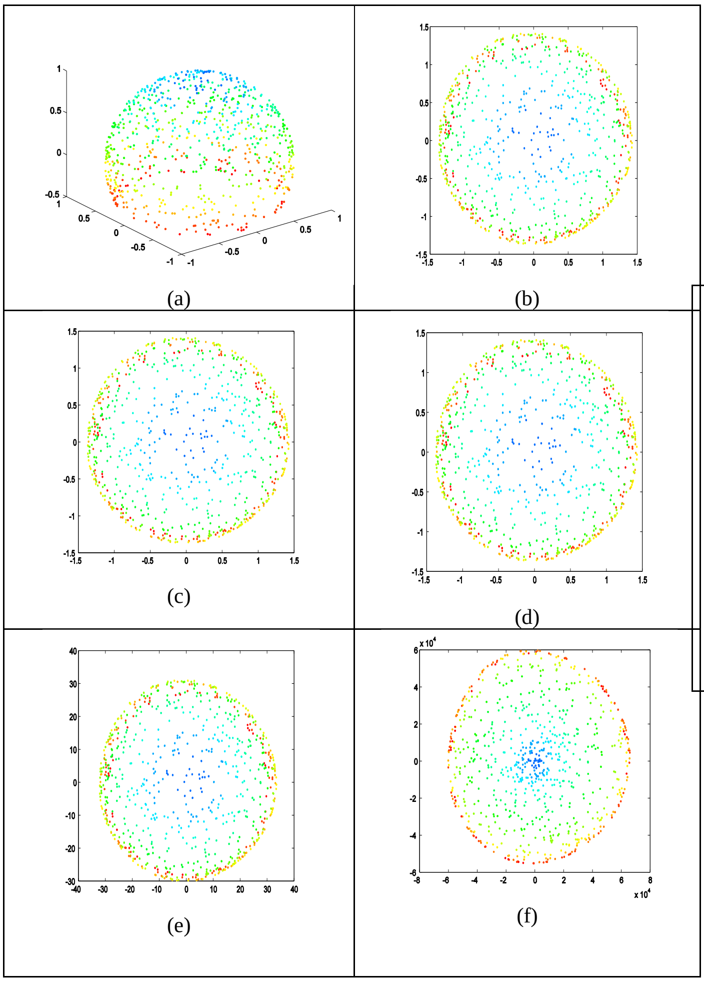
Datasets	**No. of classes**	**No. of samples**	$\frac{\sum_{\lambda_i<0}\lvert\lambda_i(S)\rvert}{\sum_i\lvert\lambda_i(S)\rvert}$
Catcortex	**4**	**65**	**0.2082**
Proteins	**4**	**226**	**7.4148e-04**
Music-EMD	**2**	**50**	**0.2819**
Music-PTD	**2**	**50**	**0.2047**
Kimia	**6**	**72**	**0.0745**
Unipen-DTW	**5**	**250**	**0.3129**
USPS-TD	**2**	**250**	**0.1486**

5.2. Experiments

5.2.1. Dimensionality reduction setup

Fig.1. (a) shows the original 3 dimensional fishbowl dataset. The spectrum of this dataset has strong negative components. The dissimilarities between samples are computed along the manifold (i.e. geodesic distance). We compute the similarity matrix for this dataset by applying Eq. (5) over the dissimilarity matrix. We apply our method that we call it "Adaptive Topology Preserving Spectrum Transformation method (ATPST)" to generate a *psd* similarity matrix from proximity data and compare it with *flip*, *clip*, *shift*, and *square* spectrum transformation approaches in a dimensionality reduction setup.

For this purpose, we use resulting *psd* similarity matrices as the kernels for Kernel Principal Component Analysis (KPCA) method. As illustrated in Fig.1. (f), projection in 2-dimensional space using ATPST preserves the local geometry completely, while the other approaches, although providing *psd* matrices, lead to overlapping of the samples (note overlapping of red points over green and orange points in Fig. 1. (b), (c), (d), and (e)). Overlapping of the samples indicates the shortcoming of *clip*, *flip*, *shift*, and *square* techniques in preserving geometry.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 1. (a) 3-D fishbowl dataset. 2-D representations of it using kernel PCA, the kernel is rectified by (b) *clip* (c) *flip* (d) *shift* (e) *square* (f) ATPST spectrum transformation.

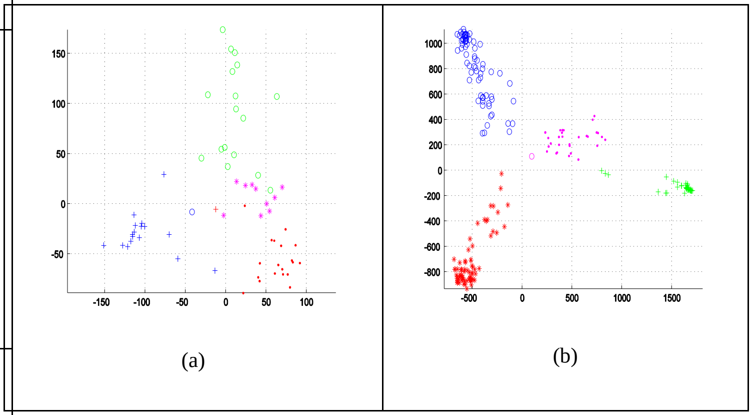
5.2.2. Clustering setup

To show the performance of the proposed algorithm in clustering problems, we compare the performance of ATPST with EVCLUS algorithm (Denœux and Masson, 2004), which is a reference method for clustering the proximity data, in the same experimental setting (Denœux and Masson, 2004). EVCLUS has shown to have good results as compared with several state of the art clustering techniques (Denœux and Masson, 2004). Proximity matrices rectified by ATPST method are used as kernels for kernel k-means algorithm to make them applicable for clustering the Catcortex and Proteins datasets. Two-dimensional representations of different groups found by this algorithm for Catcortex and Proteins datasets are shown in Fig.2. The different clusters of these two datasets are specified by different symbols for each group in corresponding figure. As illustrated in Fig. 2. kernel k-means using ATPST kernel leads to misclassification of 2 instances among 65 instances of Catcortex dataset and misclassification of 1instance among 226 instances of Proteins dataset. The EVCLUS algorithm has 3 misclassifications out of 65 points of Catcortex dataset and 1 misclassification in Proteins dataset.

5.2.3. Classification setup

As the last experiment, we evaluate ATPST in classification setup using the resulting similarity matrix as a kernel for the SVM classifier (Table. 2) and using it for running 1- Nearest Neighbor classifier (1-NN) (Table. 3).

For each dataset, we apply 10-fold cross validation and report the mean accuracy and standard deviation. The lib-svm (Chang and Lin, 2011) package is used for implementing SVM. In multi-class cases, one-against-one scheme is used for classification. Tuning the C parameter of SVM classifier is done by changing it in a wide range of values $[10^{-6}, 10^{-4}, \ldots, 10^{6}]$. The value of C, which leads to the best result in 10-fold cross-validation on the training set, is used in the evaluation of the test set. In our experiments, the ATPST's δ parameter is selected by applying cross-validation and is proportional to the variance between instances (Amari and Wu, 1993).



(a)

(b)

Fig. 2. Two-dimensional representation of (a) Catcortex and (b) Proteins datasets. A different color is used for each group found by the kernel k-means algorithm using ATPST for rectifying proximity data. True class membership is specified using different symbols.

Since computing eigenvalues of a matrix that depends on parameters is NP-hard, it would be intractable to find an optimal transformation by increasing the size of the similarity matrix. Therefore, we use a numerical method for approximating optimal values of the unknown parameters. For this purpose, we use Matlab's nonlinear multi-variable solver for approximating the optimal solution.

Table 2. Classification accuracy and standard deviation of similarity matrices rectified by ATPST, *clip*, *flip*, *shift*, and *square* approaches and used as the kernels for SVM

	ATPST	*Clip*	*Flip*	*Shift*	*Square*
Catcortex	**95.48± 7.31**	92.38±10.52	89.29± 7.47	89.29±10.05	91.19±10.17
Proteins	**96.46± 5.84**	95.59± 6.57	96.03± 6.72	95.59± 6.57	94.25± 6.24
Music-EMD	**60.00±13.33**	48.00±19.32	54.00±21.19	48.00±25.30	46.00±1897
Music-PTD	**60.00±18.86**	50.00±21.60	52.00±19.32	52.00±19.32	50.00±17.00
Kimia	90.36±11.64	86.07±13.49	**93.21± 9.85**	76.43±16.10	79.82±19.74
Unipen-DTW	**91.20± 6.75**	90.40± 6.85	90.80± 5.98	87.20± 4.92	90.80± 5.67
USPS-TD	**94.80± 5.67**	92.80± 6.75	92.00± 6.25	76.40±11.84	90.80± 5.67

Table 3. Classification accuracy and standard deviation of 1-NN based on similarity matrices that are rectified using ATPST, *clip*, *flip*, *shift*, and *square* approaches

	ATPST	*Clip*	*Flip*	*Shift*	*Square*
Catcortex	**97.14± 9.04**	96.90± 6.55	95.71± 6.90	95.24± 7.69	92.38± 8.08
Proteins	**97.37± 3.05**	**97.37± 3.05**	**97.37± 3.05**	**97.37± 3.05**	**88.54± 5.53**
Music-EMD	**64.00±18.38**	54.00±25.03	58.00±17.51	56.00±24.59	48.00±25.30
Music-PTD	**60.00±18.86**	52.00±13.98	56.00±12.65	48.00±13.98	56.00±18.38
Kimia	88.57±13.13	88.57±13.13	**92.86±12.14**	74.46±17.70	80.18±18.14
Unipen-DTW	**83.60± 6.92**	81.60±10.36	82.80± 7.07	77.20±10.34	70.80± 6.55
USPS-TD	**98.40± 2.07**	98.00± 2.11	96.00± 2.67	**98.40± 2.07**	84.00± 5.66

The statistical significance of the classification accuracy of the proposed method is reported in Table 4. Comparing it with respect to the other competitors is computed by one-sided Wilcoxon signed-rank test.

Table 4. P-values resulted from applying one-sided Wilcoxon signed-rank test over a classification rate resulted by ATPST for rectifying

similarity matrices and SVM/1NN classifier versus *clip, flip, shift,* and *square* spectrum transformation+ SVM/1NN

	ATPST/Clip	*ATPST/Flip*	*ATPST/Shift*	*ATPST/square*
SVM	0.008991	0.045514	0.008991	0.008991
1NN	0.021588	0.013868	0.021588	0.008991

As the results in Table 2 and 3 show, the superiority of the proposed method is considerable where the negative eigen-fraction of the similarity matrix is significant. In the case of Proteins dataset with very small negative eigen-fraction, all approaches lead to similar accuracy for the 1-NN classifier. The significant superiority of ATPST+SVM and ATPST-1NN over competitors are confirmed by p-values<0.05 reported in Table 4.

6. Conclusion

In this paper, we proposed an adaptive approach for rectifying indefinite proximity data using spectrum transformation. Modifying the volume element by applying a conformal transform is our contribution for generating a *psd* kernel from an indefinite similarity matrix. This goal is achieved by maximizing EF criterion, which shows similarity of the feature space to the Euclidean space.

As the conformal transform keeps angles unchanged, it can preserve the spatial relationship between the data points. Therefore, it can avoid overlapping problem, which is caused by missing topological information conveyed by negative eigenvalues and reach superior results in classification problems. The superiority of the proposed method is especially in the cases, where the spectrum of proximity data has strong negative components. The experimental evidences confirm that the semantic preservation, provided by preserving the topology of data points, leads to better results in dimensionality reduction and clustering setups.

Acknowledgments

The first author specially acknowledges Dr. Mansour rezghi, Dr. Elżbieta Pękalska, and Mr. Per Rutquist for their helps and useful discussions.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., 1990. Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410.

Amari, S. I., Wu, S., 1999. Improving support vector machine classifiers by modifying kernel functions. Neural Networks, 12(6), 783-789.

Arkadev, A. G., Braverman, È. M., 1967. Computers and pattern recognition. Thompson Book Co..

Bahlmann, C., Haasdonk, B., Burkhardt, H., 2002. Online handwriting recognition with support vector machines-a kernel approach. In Frontiers in handwriting recognition, 2002. proceedings. eighth international workshop on (pp. 49-54). IEEE.

Burges, C. J., 1998. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2), 121-167.

Chang, C. C., Lin, C. J., 2011. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3), 27.

Chen, Y., Gupta, M. R., Recht, B., 2009. Learning kernels from indefinite similarities. In Proceedings of the 26th Annual International Conference on Machine Learning (pp. 145-152). ACM.

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

Cox, T., Cox, M., 1994. Multidimensional Scaling. Chapman & Hall, London.

Denœux, T., Masson, M. H., 2004. EVCLUS: evidential clustering of proximity data. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 34(1), 95-109.

Dinkelbach, W., 1967. On nonlinear fractional programming. Management Science, 13(7), 492-498.

Duin, R. P., de Ridder, D., Tax, D. M., 1997. Experiments with a featureless approach to pattern recognition. Pattern Recognition Letters, 18(11), 1159-1166.

Duin, R. P. W., 1999. Compactness and complexity of pattern recognition problems. In International Symposium on Pattern Recognition'In Memoriam Pierre Devijver (pp. 124-128).

Graepel, T., Herbrich, R., Bollmann-Sdorra, P., Obermayer, K., 1999. Classification on pairwise proximity data. Advances in neural information processing systems, 438-444.

Graepel, T., Herbrich, R., Schölkopf, B., Smola, A., Bartlett, P., Müller, K. R., ... Williamson, R., 1999. Classification on proximity data with LP-machines. In Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470) (Vol. 1, pp. 304-309). IET.

Haasdonk, B., 2005. Feature space interpretation of SVMs with indefinite kernels. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(4), 482-492.

Haasdonk, B., Keysers, D., 2002. Tangent distance kernels for support vector machines. In Pattern Recognition, 2002. Proceedings. 16th International Conference on (Vol. 2, pp. 864-868). IEEE.

Horn, R. A., Johnson, C. R., 2012. Matrix analysis. Cambridge university press.

I Guyon, I., Schomaker, L., Plamondon, R., Liberman, M., Janet, S., 1994. UNIPEN project of on-line data exchange and recognizer benchmarks. In Pattern Recognition, 1994. Vol. 2-Conference B: Computer Vision &amp; Image Processing., Proceedings of the 12th IAPR International. Conference on (Vol. 2, pp. 29-33). IEEE.

Keysers, D., Macherey, W., Ney, H., 2004. Adaptation in statistical pattern recognition using tangent vectors. IEEE Transactions on Pattern Analysis & Machine Intelligence, (2), 269-274.

Laub, J., Roth, V., Buhmann, J. M., Müller, K. R., 2006. On the information and representation of non-Euclidean pairwise data. Pattern Recognition, 39(10), 1815-1826.

Laub, J., Müller, K. R., 2004. Feature discovery in non-metric pairwise data. The Journal of Machine Learning Research, 5, 801-818.

[12] Lee, J. A., Verleysen, M., 2007. Nonlinear dimensionality reduction. Springer Science & Business Media.

Muñoz, A., n de Diego, I. M., 2006. From indefinite to positive semi-definite matrices. In Structural, Syntactic, and Statistical Pattern Recognition (pp. 764-772). Springer Berlin Heidelberg.

Noma, H. S. K. I., Shimodaira, K. 2002. Dynamic time-alignment kernel in support vector machine. Advances in neural information processing systems, 14, 921.

Pękalska, E., Harol, A., Duin, R. P., Spillmann, B., Bunke, H. 2006. Non-Euclidean or non-metric measures can be informative. In Structural, Syntactic, and Statistical Pattern Recognition (pp. 871-880). Springer Berlin Heidelberg.

Pekalska, E., Paclik, P., Duin, R. P., 2002. A generalized kernel approach to dissimilarity-based classification. The Journal of Machine Learning Research, 2, 175-211.

Pękalska, E., Duin, R. P., 2002. Dissimilarity representations allow for building good classifiers. Pattern Recognition Letters, 23(8), 943-956.

Pelillo, M., 2013. Similarity-based pattern analysis and recognition. Springer.

Qamra, A., Meng, Y., Chang, E. Y., 2005. Enhanced perceptual distance functions and indexing for image replica recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(3), 379-391.

Roth, V., Laub, J., Kawanabe, M., Buhmann, J. M., 2003. Optimal cluster preserving embedding of nonmetric proximity data. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 25(12), 1540-1551.

Rubner, Y., Tomasi, C., Guibas, L. J., 2000. The earth mover's distance as a metric for image retrieval. International journal of computer vision, 40(2), 99-121.

Scannell, J. W., Blakemore, C., Young, M. P., 1995. Analysis of connectivity in the cat cerebral cortex. The Journal of Neuroscience, 15(2), 1463-1483.

Schleif, F. M., Tino, P., 2015. Indefinite proximity learning: a review. Neural computation.

Sebastian, T., Klein, P., Kimia, B., 2001. Recognition of shapes by editing shock graphs. In null (p. 755). IEEE.

Simard, P., LeCun, Y., Denker, J. S., 1993. Efficient pattern recognition using a new transformation distance. In Advances in neural information processing systems (pp. 50-58).

Typke, R., Giannopoulos, P., Veltkamp, R. C., Wiering, F., Van Oostrum, R., 2003. Using transportation distances for measuring melodic similarity. In ISMIR.

Vapnik, V., 2013. The nature of statistical learning theory. Springer Science & Business Media.

Williams, P., Li, S., Feng, J., Wu, S., 2007. A geometrical method to improve performance of the support vector machine. Neural Networks, IEEE Transactions on, 18(3), 942-947.

Wu, G., Chang, E. Y., Zhang, Z., 2005, March. An analysis of transformation on non-positive semidefinite similarity matrix for kernel machines. In Proceedings of the 22nd International Conference on Machine Learning (Vol. 8).

Wu, S., Amari, S. I., 2002. Conformal transformation of kernel functions: A data-dependent way to improve support vector machine classifiers. Neural Processing Letters, 15(1), 59-67.

Xiong, H., Swamy, M. N. S., Ahmad, M. O., 2005. Optimizing the kernel in the empirical feature space. Neural Networks, IEEE Transactions on, 16(2), 460-474.

Xu, W., 2013. Non-Euclidean Dissimilarity Data in Pattern Recognition (Doctoral dissertation, University of York).