On Euclidean Corrections for Non-Euclidean Dissimilarities

Robert P.W. Duin[1], Elżbieta Pękalska[2], Artsiom Harol[1], Wan-Jui Lee[1], and Horst Bunke[3]

[1] Faculty of Electrical Engineering, Mathematics and Computer Sciences, Delft University of Technology, The Netherlands
[2] School of Computer Science, University of Manchester, United Kingdom
[3] Department of Computer Science, University of Bern, Switzerland
r.duin@ieee.org,pekalska@cs.man.ac.uk,
{a.harol,w.r.lee}@tudelft.nl,bunke@iam.unibe.ch

Abstract. Non-Euclidean dissimilarity measures can be well suited for building representation spaces that are more beneficial for pattern classification systems than the related Euclidean ones [1,2]. A non-Euclidean representation space is however cumbersome for training classifiers, as many statistical techniques rely on the Euclidean inner product that is missing there. In this paper we report our findings on the applicability of corrections that transform a non-Euclidean representation space into a Euclidean one in which similar or better classifiers can be trained. In a case-study based on four principally different classifiers we find out that standard correction procedures fail to construct an appropriate Euclidean space, equivalent to the original non-Euclidean one.

1 Introduction

In various pattern recognition applications the knowledge on a set of objects can be encoded with dissimilarity functions, which relate new objects to be classified to the training set. The main reason for such a preference is the difficulty of defining good features. In particular, structural descriptions of objects are an example of this. Instead of a feature-based approach, string matching or graph matching procedures can be applied [3,4,5,6], leading to (dis)similarity data.

In spite of the lack of a good feature representation, various tools have become available that build statistical pattern classifiers on dissimilarity data. This is possible thanks to the techniques that embed arbitrary dissimilarity functions into a fixed-dimensional vector space. Many successful examples are reported; see e.g. [7,8]. Some of these mappings are hampered by the use of non-Euclidean dissimilarity measures. These usually result from various matching procedures or when robustness or invariance are incorporated into the measure [9]. Consequently, a Euclidean embedding becomes imperfect and may loose information. In [1,2] we showed that classifiers based on non-Euclidean dissimilarity representations may lead to better results than those based on transformed dissimilarity measures that are either Euclidean or have reduced non-Euclidean component. In

N. da Vitora Lobo et al. (Eds.): SSPR&SPR 2008, LNCS 5342, pp. 551–561, 2008.

a series of 44 related experiments we found out that a significantly non-Euclidean measure performed best [2].

Non-Euclidean vector spaces, e.g. pseudo-Euclidean ones, are however not well equipped with the tools for training classifiers. Distances have to be computed in a specific way and are usually not invariant to orthogonal rotations. Densities may not be properly defined. Some density-based classifiers may still be used, albeit with some restrictions. So, Euclidean corrections become of interest. They have recently gained more attention, especially in relation to Support Vector Machines (SVMs) for indefinite kernels. Such kernels arise from similarity measures related to non-Euclidean dissimilarity ones. In general, they cannot guarantee a unique solution of SVM by quadratic programming [10]. Euclidean corrections may therefore be useful for SVMs if one deals with indefinite kernels, resulting e.g. from a kernel combining procedure or incorporation of invariance.

Many of the Euclidean corrections result in a continuous one-to-one correspondence between a non-Euclidean representation and the resulting Euclidean space. Consequently, vectors separating the classes in one space, separate the classes in the same way in the other space. This thereby gives hope that such corrections may determine a Euclidean space that can equally well (or even better) be used to train good classifiers.

In this paper we set out to study a set of Euclidean correction procedures on the basis of the performances before and after the correction. This topic is also discussed by [11]. We will present an experimental study based on four classifiers. These are local and global distance-based classifiers, such as 1-Nearest Neighbor rule (1-NN) and Nearest Mean Classifier (NMC), and local and global density-based classifiers, such as Parzen classifier with a small kernel width and Quadratic Discriminant Analysis (QDA). We focus on the same problem as studied in our previous papers on this topic [1,2]: the weighted edit distance between a set of contours obtained from the Chicken Pieces data [12].

The paper is organized as follows. Section 2 explains the Euclidean correction procedures. Section 3 summarizes the classifiers we use. Experiments are presented in Section 4, while conclusions are discussed in Section 5.

2 Euclidean Correction Procedures

Our starting point is a set of objects $\mathcal{X} = \{x_1, \ldots, x_m\}$ and a symmetric dissimilarity function that compares pairs of objects. This leads to a symmetric $m \times m$ dissimilarity matrix $D := (d_{ij})$, where $d_{ij} = d(x_i, x_j)$ and $d_{ii} = 0$. Such a matrix can be perfectly embedded in a pseudo-Euclidean space (PES) $\mathcal{E}$ by an isometric, distance-preserving, mapping [13,7]. $\mathcal{E} = \mathbb{R}^{(p,q)} = \mathbb{R}^p \oplus \mathbb{R}^q$ is a real vector space equipped with a non-degenerate indefinite inner product $\langle \cdot, \cdot \rangle_{\mathcal{E}}$ which is positive definite on $\mathbb{R}^p$ and negative definite on $\mathbb{R}^q$. $\mathcal{E}$ is characterized by the signature (p, q). This PES is determined by eigendecomposition of an (indefinite) Gram matrix $G = -\frac{1}{2} J D^{*2} J$ derived from D. J is the centering matrix, while $D^{*2} = (d_{ij}^2)$; see [13,7] for details. In this decomposition p positive and q negative eigenvalues arise, indicating the signature of $\mathcal{E}$. The axes of $\mathcal{E}$ are

constituted by $\sqrt{|\lambda_i|}\mathbf{q}_i$, where $(\lambda_i, \mathbf{q}_i)$ is the corresponding eigenvalue-eigenvector pair of G. Note that the resulting configuration is uncorrelated and λ_i capture the variances of the embedded data.

Distances in $\mathcal{E}$ are defined as the square pseudo-Euclidean distance $d^2_{\mathcal{E}}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{y} \rangle_{\mathcal{E}} = d^2(\mathbf{x}_p, \mathbf{y}_p) - d^2(\mathbf{x}_q, \mathbf{y}_q)$ for $\mathbf{x}, \mathbf{y} \in \mathcal{E}$. The "-" sign distinguishes the pseudo-Euclidean distance from the Euclidean one. $p+q = m-1$ holds in a full embedding. It is perfect because the distances between the embedded PES vectors are equal to the original ones. In practice, the dimensions of PES related to small eigenvalues are discarded, which leads to more stable and better defined classifiers. Here we restrict ourselves to a full embedding because we want to make the comparisons with Euclidean corrections independent from the estimation of intrinsic dimension (i.e. number of significant eigenvalues). This has however consequences for some classifiers as appropriate regularization may be necessary. Euclidean correction procedures are described below.

The Positive part of the Pseudo Euclidean Space (PPES)
The most obvious correction for a pseudo-Euclidean space $\mathbb{R}^{(p,q)}$ is to neglect the negative definite subspace. This results in a p-dimensional Euclidean space $\mathbb{R}^p$ with many-to-one mappings to $\mathcal{E}$. Consequently, it is possible that the class overlap increases. It may, however, be worthwhile if the negative eigenvalues in the embedding procedure are mainly the result of noise and not informative for the class separation. In that case this correction may improve the classification.

The Associated Euclidean Space (AES)
Since $\mathbb{R}^{(p,q)}$ is a vector space, we can equip it with the traditional inner product, which leads to the so-called associated Euclidean space $\mathbb{R}^{p+q}$. It means that the vector coordinates are identical to those of PES, but now we use the norm and distance measure that are Euclidean. This is consistent with the natural topology of a vector space. This solution is identical to the one obtained by classical scaling based on the magnitudes of eigenvalues [14,7].

Dissimilarity Enlargement by a Constant (DEC)
Instead of modifying the embedding procedure, the dissimilarity matrix may be adapted such that it is embeddable into a Euclidean space. A simple way to avoid the negative eigenvalues is to increase all off-diagonal elements of the dissimilarity matrix such that $d^2_c(x_i, x_j) = d^2(x_i, x_j) + 2c, \forall_{i \neq j}$. The value of c is chosen such that $c \geq -\lambda_{min}$, where λ_{min} is the smallest negative eigenvalue in the pseudo-Euclidean embedding of D. As a result, all eigenvalues are increased by c [7]. In our experiments we set $c = -\lambda_{min}$. Since the eigenvalues reflect the variances of the embedded data, the dimensions of the resulting Euclidean space are unevenly scaled by $\sqrt{\lambda_i + c}$. Note that the dimension with the largest negative contribution in PES has now a zero variance. In this way, dimensions related to noisy negative eigenvalues are more pronounced. [7]

Relaxation by a Power Transformation (Relax)
Another way to adapt the dissimilarity matrix such that it is embeddable into a Euclidean space is to suppress the influence of large distances by a suitable concave transformation. Here, we relax the dissimilarity values by taking a small

power: $d_r(x_i, x_j) = d^\alpha(x_i, x_j), \forall_{i,j}$. For small values of α, $0 < \alpha < 1$, the dissimilarity values become increasingly more alike, by which the objects can eventually be embedded in a Euclidean space. We set $\alpha = 0.1$ in all experiments as it is just a sufficient value for the worst cases.

Laplace Transformation (Laplace)
We focus now on Euclidean corrections based on differential geometric properties of the given problem. We look for a smooth transformation T that changes large distances, but preserves the local structure as well as possible. Minimizing with respect to an objective functional, one may find a relation between the original distances, approximated by geodesics on a smooth manifold, and the Euclidean ones. For this purpose, we consider the following Hilbert-Schimdt operator $T\colon L_2 \to L_2$ acting in a Hilbert space of functions from L_2:

$$Tf(x) = \int_X \mathcal{D}(x, y) f(y) d\mu(y). \tag{1}$$

It is a compact linear operator associated to the kernel $\mathcal{D}(x, y)$, represented by the dissimilarity matrix D. T is defined on all configurations of our data points as specified by the function f. We assume that the data are distributed according to some unknown but finite probability measure $\mu(y)$. Hence, by defining a probability density function (pdf) $p(y)$, we have $d\mu(y) = p(y)dy$. Now a normalized kernel of the operator (1) can be constructed as $\mathcal{D}_p(x, y) = \frac{\mathcal{D}(x,y)}{p(x)p(y)}$. Consider now a normalization of T as

$$T_\nu f(x) = \frac{1}{\nu^2(x)} \int_X \mathcal{D}_p(x, y) f(y) dy, \tag{2}$$

with respect to the global scaling specified by the function $\nu^2(x) = \int_X \mathcal{D}_p(x, y)dy$, which is well defined iff $\mathcal{D}_p(x, y) \geq 0$. The resulting stochastic operator T_ν is widely used in spectral graph theory and usually referred to as a normalized graph Laplacian $\mathcal{L} = I - T_\nu$ [15]. $\mathcal{L}$ has a positive spectrum, by construction. Indeed, thanks to normalization, the spectrum of T_ν belongs to a sphere of radius 1. Moreover, the eigenfunctions of T_ν are identical to those of $\mathcal{L}$, leading to the same data configuration. T_ν has an asymmetric kernel due to the applied normalization. We can overcome this by working with its symmetric conjugate. For that, we conjugate $\mathcal{D}_p(x, y)$ with ν, and get a new normalized kernel $\mathcal{D}_\nu(x, y) = \frac{\mathcal{D}_p(x,y)}{\nu^2(x)} \cdot \frac{\nu(x)}{\nu(y)} = \frac{\mathcal{D}_p(x,y)}{\nu(x)\nu(y)}$. The associated operator is defined as

$$\widehat{T}_\nu f(x) = \int_X \mathcal{D}_\nu(x, y) f(y) dy, \tag{3}$$

and the Laplacian is $\widehat{\mathcal{L}} = I - \widehat{T}_\nu$. The optimal solution for f yields the largest correlation between f and $\widehat{T}_\nu$. As (3) is an averaging operator, this corresponds to the flattest configuration of the data in a Hilbert space. To obtain Euclidean distances from the Laplacian, we solve the following optimization problem

$$\begin{aligned} &\min\ \gamma \\ &\text{s.t. } \gamma I - \widehat{\mathcal{L}} \preceq 0,\ \gamma \geq 0, \end{aligned}$$

where the first constraint is equivalent to $(\gamma - 1)I + \widehat{T}_\nu \preceq 0$, while $\preceq$ denotes non-positive definiteness. This is equivalent to adding a constant to the spectrum of $\widehat{T}_\nu$ to make it non-positive definite, i.e. to represent Euclidean distances [14].

3 Classification Procedures

Generalization capabilities of classifiers in a vector space are based on two principles: densities and distances, related to generative and discriminative approaches, respectively. Objects are either classified into the most probable class derived from densities and prior probabilities, or into the class that is in some sense the nearest. Many classification algorithms rely on both principles, e.g. density estimators like Parzen and k-NN use distances between objects while distance-based classifiers like Fisher's Linear Discriminant and SVMs make use of parameter optimization procedures that somehow depend on the probability density of training vectors and not just their distances.

Both general densities and distances are not yet well formulated in PES, differently than in AES. E.g. we are not aware of the definition of a Gaussian distribution directly in PES, although the related quadratic classifier can easily be computed as it appears to be independent of the signature of the space. In fact, it can be computed directly in AES. Formally, the definition of a pdf does not make use of a metric, as the computation of integrals over volumes is sufficient. For the estimation of a pdf from a finite set of points, however, the metric may be very useful.

Pairwise distances between vectors are well defined in PES. Consequently, the 1-NN rule and also the Parzen classifier can be used. The interpretation of the latter as a consistent Bayes classifier for an appropriate asymptotic choice of its kernel function has still to be investigated. Distances between vectors and class boundaries, even if they are linear, are not well formulated in PES. Vectors on a linear boundary may have arbitrary large negative distances to the vectors on both sides of that boundary. An exception is the Nearest Mean Classifier, given that the contribution of the $\mathbb{R}^q$-subspace to the distances is smaller than of the $\mathbb{R}^p$-subspace. It produces a linear boundary, but its interpretation as the nearest class mean remains valid, because class means are properly computed.

We will use a consistent set of classifiers in order to compare different representation vector spaces. Four essentially different classifiers are selected in a Euclidean space such that they can also be computed in PES, but whose properties have to be still analyzed in the future. These are:

- 1-NN, the Nearest-Neighbor rule. It can directly be applied to a given dissimilarity matrix. Hence, embedding is not necessary.
- NMC, the Nearest Mean classifier. It is a global distance-based classifier (as it depends on all objects). Class means and distances of objects to these means have to be computed in the embedded PES.
- Parzen, a local-density based classifier. We use the same radial basis function as a kernel for all classes. The resulting classifier is thereby independent of

the dimensionality and is a function of the distances between training and test objects. Consequently, it can be computed directly on the dissimilarity matrix. In order to have a consistent choice for the kernel width over various representation spaces we set it to 0.1 of the 1% percentile of the distance distribution in the training set. This is a small value, in agreement with our intention to create a local density-based classifier.

- QDA, a global density-based classifier in a Euclidean space, equivalent to the Bayes classier assuming Gaussian class distributions. It is a global classifier as it everywhere depends on all training objects. As the dimension of PES is $m-1$ for a set of m objects it is necessary include some regularization. In order to compare over different representation spaces with different scalings we use QDA in a reduced PCA-space with a fixed dimension.

It is not our intention here to find the best classifier for a given problem. Instead, we want to investigate how different classifiers behave over several transformations, in order to characterize these transformations. We believe that the wide spectrum of the properties of the chosen classifiers serves thereby our target.

4 Experiments

We focus on a single problem, also studied by us before [1,2], based on the Chicken Pieces data [12]. It consists of 446 binary images representing five classes of chicken parts: wing (117 examples), back (76), drumstick (96), thigh and back (61), and breast (96). We estimate class prior probabilities by class frequencies here. Object contours are first approximated by straight line segments of a fixed length L; $5 \leq L \leq 40$ pixels. The sequence of angles between the neighboring segments becomes the initial string representation for which a family of edit distances [3] is derived. Insertion and deletion costs are set to 45 degrees, while the substitution cost is the magnitude of angle differences. The dissimilarity matrices are available from `http://www.iam.unibe.ch/fki/databases/string-edit-distance-matrices/`. The asymmetric dissimilarities are made symmetric by averaging, $d_{ij} = (d_{ij} + d_{ji})/2$.

Every dissimilarity matrix D_L from the family of edit-distances, $5 \leq L \leq 40$, is perfectly embedded in PES. The fraction of negative eigenvalues grows with increasing L indicating that the dissimilarity measure becomes increasingly non-Euclidean. After embedding (defined by all data), the performance of the four classifiers is computed for all values of L by averaging over ten runs of the 10-fold cross-validation. Experiments are handled with care; the same objects are used for training and testing in identical runs over different classifiers and different values of L. This results in smooth curves with reliable differences in classification performance for the given dissimilarity data.

The performance of four classifiers in six representation spaces is presented in Fig. 1 and 2. They show the same curves, organized either by classifier (Fig. 1) or by space (Fig. 2). The 1-NN results are essentially identical for the PES, DEC and Relax representations as monotonic transformations preserve the ranking of dissimilarity values. The 1-NN rule performs identically for the PPES and AES

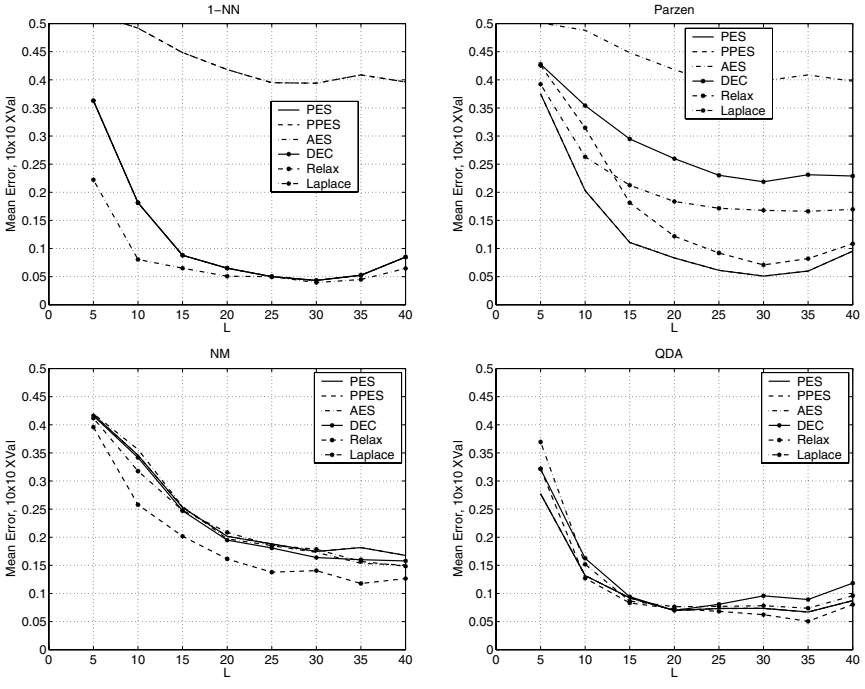

Fig. 1. Averaged 10-fold cross-validation performance of four classifiers in six representation vector spaces, shown by classifier. L (indicating different edit-distance data) is varied along the horizontal axis. In the top-left plot (1-NN), the curves for the PES, DEC and Relax cases are identical as well as the curves for AES and PPES. The curves for PES and PPES coincide in the top-right plot (Parzen). Standard deviations over the 10 repetitions are smaller than 0.01 everywhere.

cases and the Parzen classifier gives the same results for the PES and PPES cases. The following observations can be made w.r.t. various transformations:

- PPES (neglecting all 'negative' dimensions in PES) deteriorates the 1-NN results, but does not influence other classifiers. So it destroys very local distances without influencing the large ones. It also does not affect the densities.
- AES interprets PES as Euclidean, worsens the performance of locally-sensitive classifiers (1-NN, Parzen), but leaves the globally sensitive classification unaffected.
- DEC. Although this monotonic transformation preserves the ranking of dissimilarities, and so the 1-NN performance, it negatively influences the local densities, and thereby the Parzen results.
- Relax. The power transformation of the dissimilarities leads to better results than the DEC space. Parzen is less affected and NMC is even improved.
- Laplace. This shows an interesting result. The 1-NN performance improves but other classifiers deteriorate (Parzen and NM) or stay equal (QDA).

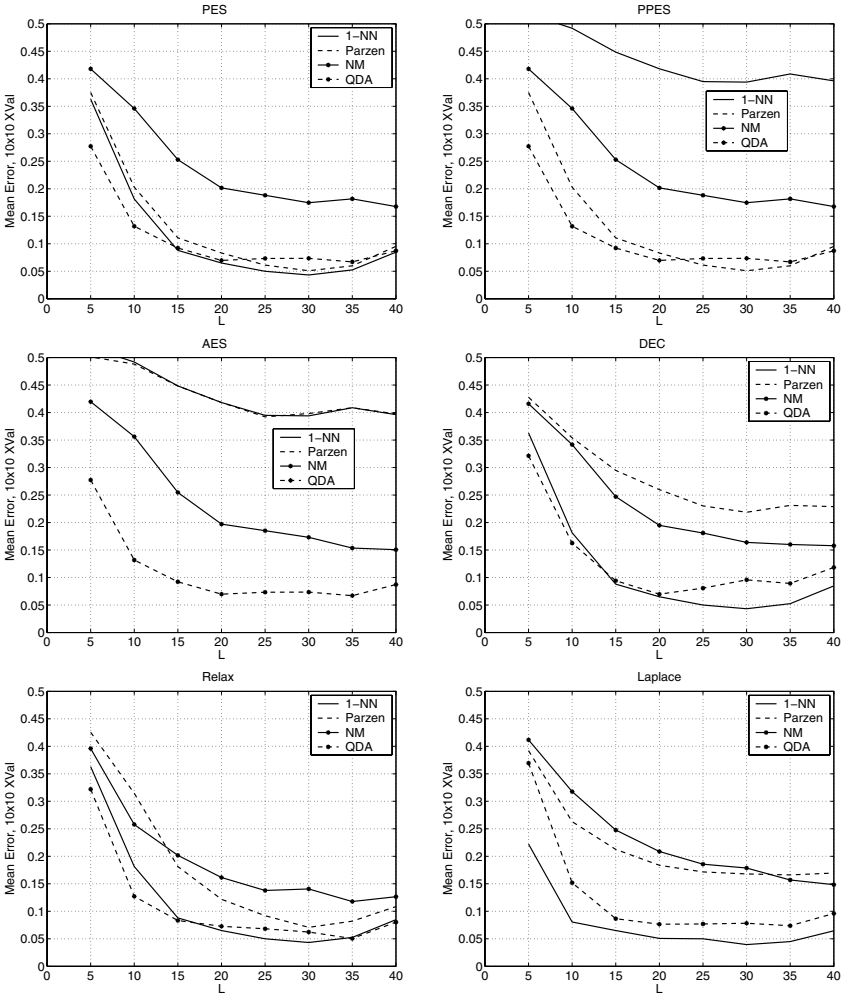

Fig. 2. Averaged 10-fold cross-validation performance of four classifiers in six representation vector spaces, shown by space. The curves for the 1-NN rule and Parzen (almost) coincide in AES. Standard deviations are smaller than 0.01 everywhere.

We observe that locally-sensitive classifiers are especially affected by the transformations. NMC and QDA remain stable or even improve (NMC by Relax). Parzen always deteriorates (except for PPES), and 1-NN deteriorates by PPES and AES, remains unaffected by DEC and Relax, and improves by Laplace.

Additional experiments are run for the best case, $L = 30$. The learning curves in Fig. 3 show no surprises: the relative performances of the classifiers are rather stable for the used sizes of the training set (90% from 445 objects). This indicates that the used data set is sufficiently large for the comparisons.

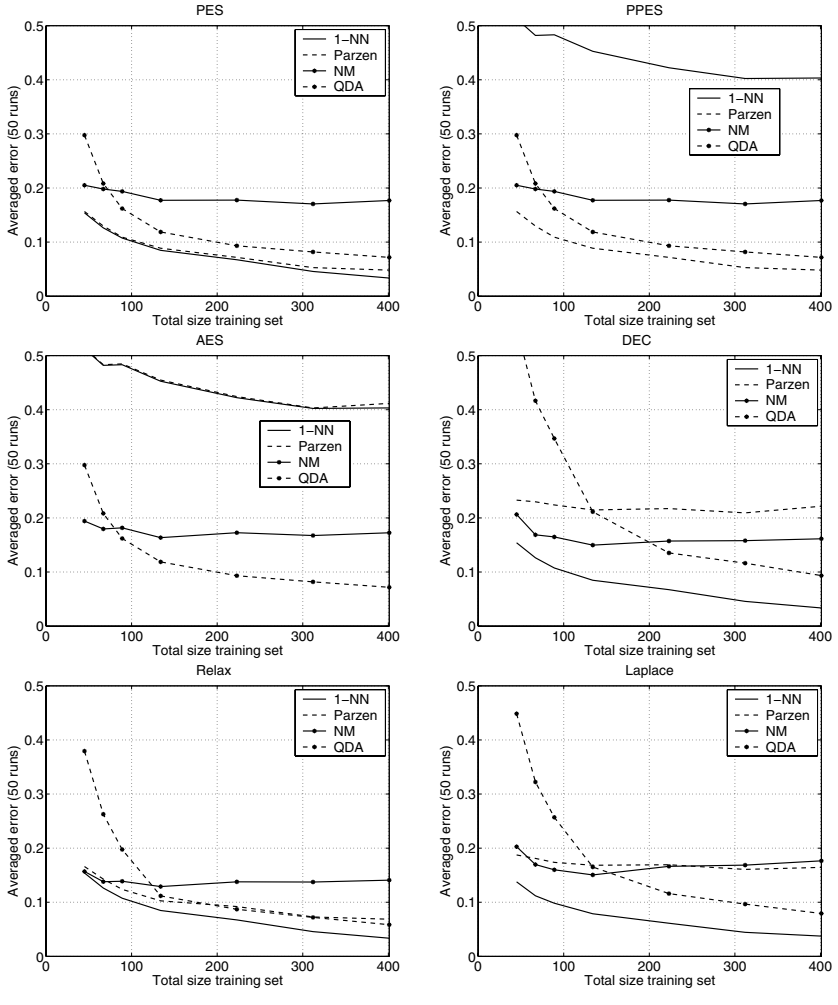

Fig. 3. Learning curves for the four classifiers in six representation spaces. Standard deviations are smaller than 0.002 everywhere.

5 Conclusions

In this paper we analyze the effect of Euclidean corrections for non-Euclidean dissimilarity data embedded in a pseudo-Euclidean space. We use a consistent and diverse set of classifiers that could be computed in the pseudo-Euclidean and Euclidean spaces. Our experiments rely on embedding both training and test data simultaneously (without using the labels of the test data). It is to be expected that all classifiers will perform worse if the test data are projected afterwards. We found that globally-sensitive classifiers are hardly affected by the correction procedures. The local distance-based 1-NN rule is insensitive to

correction procedures that rely on monotonic transformations of the original dissimilarities. Other transformations that neglect or correct the contribution of negative eigenvalues (scaling dimensions of the embedded space), severely deteriorate the performance. In addition, we also introduced the Laplace correction for dissimilarity data and found that it might increase the 1-NN accuracy.

All correction procedures, including Laplace, deteriorate the results of the Parzen classifier for small kernels. This shows that the local structure in the data is damaged even though the nearest neighbor relations remain constant or are improved. The local separation capabilities based on densities may suffer from all attempts to correct the dissimilarities into a set of distances embeddable into a Euclidean space.

With a few minor exceptions we did not find any improvements of the classifiers after the Euclidean correction. So, such corrections do not seem worthwhile as general procedures for improving the generalization possibilities of the data in our example. Corrections need to be studied in relation with a classifier. Note that the above conclusions need to be handled with some care. There is a plethora of classifiers available in Euclidean spaces, much richer than the few ones currently studied in pseudo-Euclidean spaces, even though they still cover a wide range of generalization principles. It might be possible that very good classifiers can be found in corrected Euclidean spaces that have no counterpart (yet) in pseudo-Euclidean space.

Although our analysis is based on a careful study of just a single example, it still allows us to draw the conclusion that a general profitable correction procedure for non-Euclidean dissimilarities has not been found yet. Such procedures need thereby to be analyzed in relation with the classifier.

Acknowledgments. This work is supported by the Dutch Organization for Scientific Research (NWO) and the Engineering and Physical Sciences Research Council in the UK.

References

1. Pekalska, E., Duin, R., Günter, S., Bunke, H.: On not making dissimilarities Euclidean. In: S+SSPR, pp. 1145–1154 (2004)
2. Pekalska, E., Harol, A., Duin, R., Spillmann, B., Bunke, H.: Non-Euclidean or non-metric measures can be informative. In: S+SSPR, pp. 871–880 (2006)
3. Bunke, H., Sanfeliu, A. (eds.): Syntactic and Structural Pattern Recognition Theory and Applications. World Scientific, Singapore (1990)
4. Bunke, H., Shearer, K.: A graph distance metric based on the maximal common subgraph. Pattern Recognition Letters 19, 255–259 (1998)
5. Torsello, A., Hancock, E.: Computing approximate tree edit distance using relaxation labeling. Pattern Recognition Letters 24, 1089–1097 (2003)
6. Robles-Kelly, A., Hancock, E.: String edit distance, random walks and graph-matching. IJPRAI 18, 315–327 (2004)
7. Pekalska, E., Duin, R.: The Dissimilarity Representation for Pattern Recognition. Foundations and Applications. World Scientific, Singapore (2005)
8. Wilson, R., Luo, B., Hancock, E.: Pattern vectors from algebraic graph theory. IEEE Trans. on PAMI 27, 1112–1124 (2005)

9. Jacobs, D., Weinshall, D., Gdalyahu, Y.: Classification with Non-Metric Distances: Image Retrieval and Class Representation. IEEE TPAMI 22, 583–600 (2000)
10. Haasdonk, B.: Feature space interpretation of SVMs with indefinite kernels. IEEE TPAMI 25, 482–492 (2005)
11. Muñoz, A., de Diego, I.M.: From indefinite to positive semi-definite matrices. In: SSPR/SPR, pp. 764–772 (2006)
12. Andreu, G., Crespo, A., Valiente, J.M.: Selecting the toroidal self-organizing feature maps (TSOFM) best organized to object recogn. In: ICNN, pp. 1341–1346 (1997)
13. Goldfarb, L.: A new approach to pattern recognition. In: Kanal, L., Rosenfeld, A. (eds.) Progress in Pattern Recognition, vol. 2, pp. 241–402. Elsevier, Amsterdam (1985)
14. Gower, J.: Metric and Euclidean Properties of Dissimilarity Coefficients. J. of Classification 3, 5–48 (1986)
15. Chung, F.: Spectral Graph Theory. In: CBMS Regional Conference Series in Mathematics, vol. 92. American Mathematical Society (1997)