On Not Making Dissimilarities Euclidean

Elżbieta Pękalska[1], Robert P.W. Duin[1], Simon Günter[2], and Horst Bunke[2]

[1] ICT Group, Faculty of Electrical Engineering,
Mathematics and Computer Sciences,
Delft University of Technology, The Netherlands
{e.pekalska,r.p.w.duin}@ewi.tudelft.nl
[2] Department of Computer Science, University of Bern, Switzerland
{gunter,bunke}@iam.unibe.ch

Abstract. Non-metric dissimilarity measures may arise in practice e.g. when objects represented by sensory measurements or by structural descriptions are compared. It is an open issue whether such non-metric measures should be corrected in some way to be metric or even Euclidean. The reason for such corrections is the fact that pairwise metric distances are interpreted in metric spaces, while Euclidean distances can be embedded into Euclidean spaces. Hence, traditional learning methods can be used.
The k-nearest neighbor rule is usually applied to dissimilarities. In our earlier study [12, 13], we proposed some alternative approaches to general dissimilarity representations (DRs). They rely either on an embedding to a pseudo-Euclidean space and building classifiers there or on constructing classifiers on the representation directly. In this paper, we investigate ways of correcting DRs to make them more Euclidean (metric) either by adding a proper constant or by some concave transformations. Classification experiments conducted on five dissimilarity data sets indicate that non-metric dissimilarity measures can be more beneficial than their corrected Euclidean or metric counterparts. The discriminating power of the measure itself is more important than its Euclidean (or metric) properties.

1 Introduction

For learning purposes, objects can be described by dissimilarities to some chosen examples. Such representations can be derived from raw (sensor) measurements, e.g. images or spectra [10, 7], feature-based representations, e.g. for objects represented by mixed variables, or they can result from structural descriptions, e.g. when objects are defined by strings or trees [2].

Assume a collection of objects, a representation set $R := \{p_1, p_2, \ldots, p_r\}$ and a dissimilarity measure d, capturing the notion of closeness between two objects. d is required to be nonnegative and to obey the reflexivity condition, $d(x, x) = 0$, yet, it might be non-metric. A dissimilarity representation (DR) of an object x is defined as a vector of dissimilarities between x and the objects of R, i.e. $D(x, R) = [d(x, p_1), d(x, p_2), \ldots, d(x, p_r)]$. Hence, for a set of objects from T, it

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 1145–1154, 2004.

extends to a dissimilarity matrix $D(T, R)$. The set R ($R \subseteq T$ or $R \cap T = \emptyset$), consisting of representative objects for the domain, should be relatively small.

A direct approach to dissimilarities leads to the k-nearest neighbor (k-NN) method. This rule is applied here to $D(T_t, R)$, so test objects of T_t become members of the class the most frequently occurring among the k nearest neighbors from R. The k-NN rule can learn complex boundaries and generalize well for large representation sets, yet, at high computational costs. In practice, it might also be difficult to get a sufficiently large R to reach a satisfactory accuracy. Moreover, the performance of the k-NN rule may be affected by presence of noisy examples.

Alternative approaches to DRs can be more computationally advantageous than the k-NN method, especially for a small R. The *embedding* approach builds an embedded pseudo-Euclidean configuration such that the dissimilarities are preserved. In the *dissimilarity space* approach, $D(x, R)$ is considered as a data-depending mapping to the so-called dissimilarity space, where each dimension corresponds to a dissimilarity to a particular object from R [12]. Various classifiers can be constructed in both embedded and dissimilarity spaces [11–13].

The k-NN method is often applied to metric distances, where based on metric properties also fast approximating NN rules can be constructed; see e.g. [9]. Our approaches to DRs can handle quite arbitrary measures. Still, an open question refers to possible benefits of correcting a measure to make it metric or even Euclidean [4, 14]. Metric or Euclidean distances can be interpreted in appropriate spaces, which posses many algebraical properties and where an arsenal of discrimination functions exists. Here, we investigate some ways of making a dissimilarity measure 'more' Euclidean (or 'more' metric) and the influence of such corrections on the performance of some classifiers. We will show that the corrected measures do not necessarily guarantee better performances.

2 Interpretations of the Dissimilarity Data

Embedding. Given any symmetric $D(R, R)$, a configuration X can be found such that the distances between the vectors of X reflect the original ones. In general, a Euclidean space is not 'large enough' for such a distance-preserving mapping, but a pseudo-Euclidean space is [5]. It is a $(p+q)$-dimensional non-degenerate indefinite inner product space $\mathcal{E} := \mathcal{R}^{(p,q)}$ such that the inner product $\langle \cdot, \cdot \rangle_{\mathcal{E}}$ is positive definite (pd) on $\mathcal{R}^p$ and negative definite on $\mathcal{R}^q$. Therefore, $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{E}} = \sum_{i=1}^{q} x_i y_i - \sum_{i=p+1}^{p+q} x_i y_i = \boldsymbol{x}^T \mathcal{J}_{pq} \boldsymbol{y}$,, where $\mathcal{J}_{pq} = \mathrm{diag}\,(I_{p\times p}; -I_{q\times q})$ and I is the identity matrix. Consequently, $d^2_{\mathcal{E}}(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}-\boldsymbol{y}, \boldsymbol{x}-\boldsymbol{y} \rangle_{\mathcal{E}} = d^2_{\mathcal{R}^p}(\boldsymbol{x}, \boldsymbol{y}) - d^2_{\mathcal{R}^q}(\boldsymbol{x}, \boldsymbol{y})$. Since $\mathcal{E}$ is a linear space, many inner product based properties can be appropriately extended from the Euclidean case. Yet, the interpretations are different [5, 11].

The inner product (Gram) matrix S of the underlying configuration X can be expressed by using the square dissimilarities $D^{*2} = (d^2_{ij})$ as $S = -\frac{1}{2} J D^{*2} J$, where $J = I - \frac{1}{r}\mathbf{1}\mathbf{1}^T$ [5, 13, 11]. So, X is determined by the eigendecomposion of $S = Q \Lambda Q^T = Q|\Lambda|^{1/2}\mathrm{diag}(\mathcal{J}_{p'q'}; 0)\,|\Lambda|^{1/2} Q^T$, where $|\Lambda|$ is a diagonal matrix of

first decreasing p' positive eigenvalues, then decreasing magnitudes of q' negative eigenvalues, followed by zeros. Q is a matrix of the corresponding eigenvectors. X is then uncorrelated [5, 13] and represented in $\mathcal{R}^k$, $k = p' + q'$, as $X = Q_k |\Lambda_k|^{1/2}$. Since only some eigenvalues are large (in magnitude), the remaining ones, if close to zero, can be disregarded as non-informative. By their removal, the data are not only de-noised, but the curse of dimensionality is also avoided. So, the reduced representation $X_{red} = Q_m |\Lambda_m|^{1/2}$, $m = p + q < k$, is determined by the largest p positive and the smallest q negative eigenvalues. New objects $D(T_t, R)$ are orthogonally projected onto $\mathcal{R}^m$; see [5, 13, 11] for details.

Inner product based classifiers can appropriately be redefined in a pseudo-Euclidean space. A linear classifier $f(\boldsymbol{x}) = \boldsymbol{v}^T \mathcal{J}_{pq} \boldsymbol{x} + v_0$ is e.g. constructed by addressing it as $f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + v_0$, where $\boldsymbol{w} = \mathcal{J}_{pq} \boldsymbol{v}$; see also [5, 13, 11].

Dissimilarity Spaces. In a dissimilarity space, each dimension corresponds to a dissimilarity $D(\cdot, p_i)$. The property that dissimilarities should be small for similar objects (belonging to the same class) and large for distinct objects, gives a possibility for a discrimination. Thereby, $D(\cdot, p_i)$ can be interpreted as an attribute. This reasoning justifies the usage of traditional classifiers, e.g. linear ones, built in dissimilarity spaces. They can outperform the k-NN rule since they become more global in their decisions by making use of a larger training set T, while maintaining a small R. By using weighted combinations of dissimilarities, such classifiers suppress the influence of noisy examples [12, 13].

3 Going More Euclidean or More Metric

The Gram matrix $S = -\frac{1}{2} J D^{*2} J$ is pd iff D is Euclidean [12, 11, 5]. If S has negative eigenvalues, then D is non-Euclidean and a Euclidean configuration X preserving the distances perfectly cannot be constructed. However, D can be corrected to be Euclidean, which makes the corresponding S pd. Some possible approaches to address this issue are [4, 14, 11]:

- *Clipping* - only p positive eigenvalues are considered yielding a p-dimensional configuration $X = Q_p \Lambda_p^{1/2}$. Now, after neglecting the negative contributions, the resulting Euclidean representation overestimates the actual dissimilarities.
- *Adding* 2τ - there exists a positive $\tau \geq -\lambda_{\min}$, where $\lambda_{\min}$ is the smallest (negative) eigenvalue of S, such that $D_{corr} = [D^{*2} + 2\,\tau\,(\mathbf{1}\mathbf{1}^T - I)]^{*1/2}$ is Euclidean [6, 13, 11]. This means that the corresponding S_{corr} is pd. In practice, the eigenvectors of S and S_{corr} are identical, but the value τ is added to the eigenvalues, giving rise to the new diagonal eigenvalue matrix $\Lambda_{cor} := \Lambda_k + \tau I$. The distortion is significant if τ is large. If reduced representations of a fixed dimensionality are considered, different eigenvectors will be selected (based on significant eigenvalues) for the original and corrected dissimilarities.
- *Adding* κ - there exists a positive $\kappa \geq \lambda_{\max}$, where $\lambda_{\max}$ is the largest eigenvalue of $\begin{bmatrix} O_{n\times n} & 2\,S(D^{*2}) \\ -I_{n\times n} & -4\,S(D) \end{bmatrix}$, $S(A) := -\frac{1}{2} JAJ$, such that $D_{corr} = D + \kappa\,(\mathbf{1}\mathbf{1}^T - I)$ is Euclidean [6, 13, 11]. The corresponding Gram matrix S_{corr} yields eigenvectors which are different than these of S.

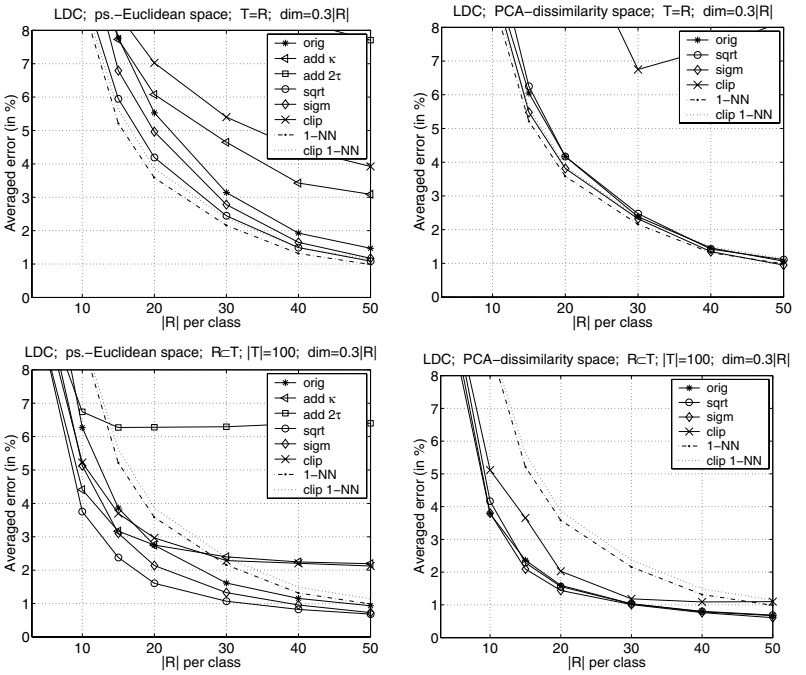

Fig. 1. The performance of the LDC for the Pen-angle dissimilarity data.

– *Power/Sigmoid* - there exists a parameter p such that $D_p = (g(d_{ij}; p))$ is Euclidean for a concave function g such as $g(x) = x^p$ with $p < 1$ or a sigmoid $g(x) = 2/(1 + e^{-x/s}) - 1$ [4, 11]. In practice, p is determined by trial and error.

These approaches transform D such that a Euclidean configuration X can be found. It is, however, still possible that the corrections applied are less than required for Euclideaness. In such cases, the measure is simply made 'more' Euclidean (hence, also 'more' metric), since the influence of negative eigenvalues will become smaller after applying the above transformations.

4 Experiments

Five dissimilarity data sets are used in our study. The first two refer to DRs built on the contours of pen-based handwritten digits [1]. All digits are represented by strings of vectors between the contour points for which an edit distance with a fixed insertion and deletion costs and with some substitution cost is computed. The substitution costs such as an angle and a Euclidean distance between vectors lead to two different DRs, denoted as Pen-dist and Pen-angle, respectively; see also [2]. Here, only a part of the data of 3488 examples, is considered. The values are also scaled by some constant to bound the dissimilarities. The digits are unevenly represented; the class cardinalities vary between 334 and 363.

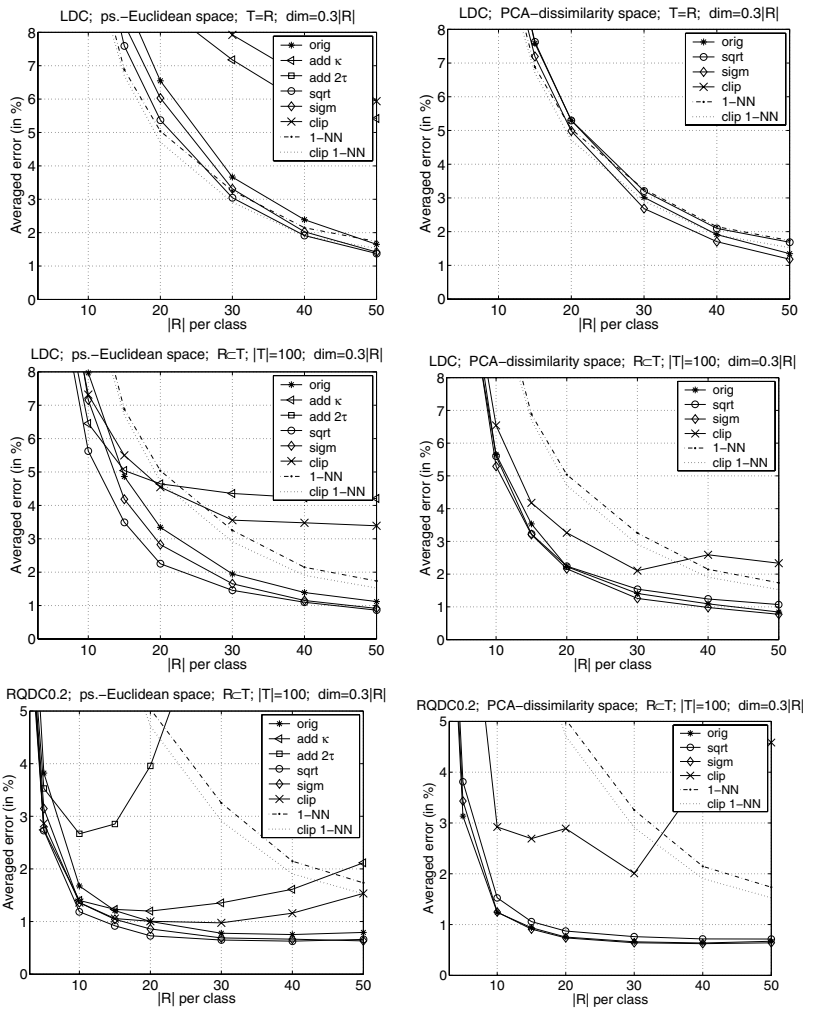

Fig. 2. The performances of the LDC and RQDC for the Pen-dist dissimilarity data.

Another dissimilarity data set, consisting of 2000 examples evenly distributed in ten classes, represents the NIST digits [15]. Here, the asymmetric similarity measure, based on deformable template matching, as defined in [8], is used. Let $S = (s_{ij})$ denote the similarities. The symmetric dissimilarities $D = (d_{ij})$ are derived as $d_{ij} = (s_{ii} + s_{jj} - s_{ij} - s_{ji})^{1/2}$ for $i \neq j$ and $d_{ii} = 0$.

The last two DRs are derived for randomly generated polygons. They consist of convex quadrilaterals and general heptagons. The polygons are first scaled and then the Hausdorff and modified Hausdorff distances [10] between their corners are computed. The two classes are equally represented by 2000 objects.

If a dissimilarity d is Euclidean, then for a symmetric $D = (d_{ij})$, all eigenvalues λ_i of the corresponding Gram matrix S are non-negative. Hence, the magnitudes

Table 1. Non-Euclidean and non-metric aspects of some DRs. The ranges of r_{mm}, r_{neg} and c indicate the smallest and largest values found for $D(R,R)$, where $|R|$ varies between $30-500$ or $10-200$ for the digit and polygon data, respectively. As a reference, the last two columns show the average and maximum dissimilarity for the complete data.

DR	r_{mm} (in %)	r_{neg} (in %)	c	avr. dissim.	max dissim.
Pen-angle	[10.6, 12.2]	[9.4, 24.1]	[0.0, 0.3]	7.1	20.0
Pen-dist	[13.8, 14.3]	[14.2, 27.8]	[0.3, 1.0]	4.0	12.5
NIST-matching	[27.5, 35.5]	[10.6, 35.5]	[0.1, 0.5]	0.6	1.0
Polygon-hausd	[13.0, 25.5]	[5.4, 31.6]	0	1.2	3.1
Polygon-mhausd	[5.0, 13.0]	[1.8, 24.6]	[0.0, 0.1]	0.7	1.6

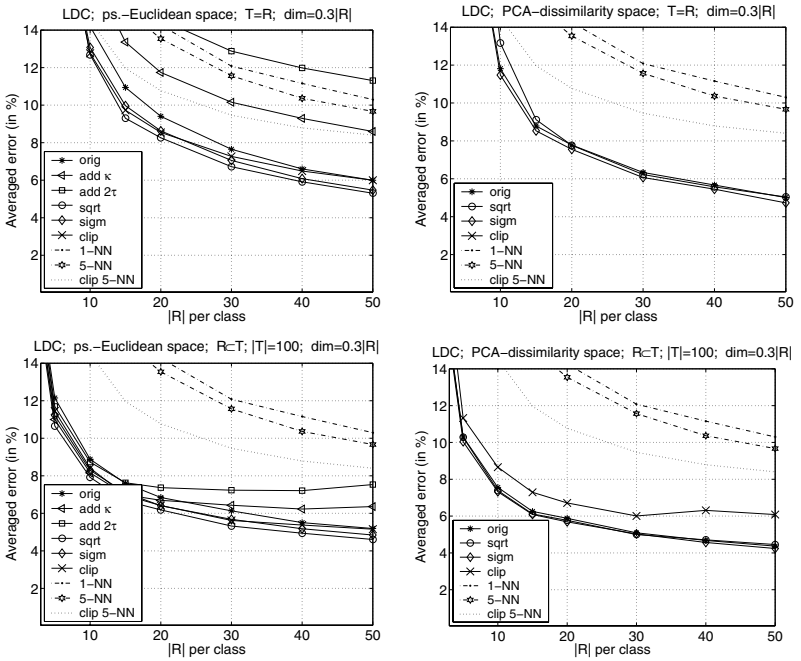

Fig. 3. The LDC performance for the NIST-matching dissimilarity data.

of negative eigenvalues manifest the deviation from Euclideaness. An indication of such a deviation is given by $r_{mm} := |\lambda_{min}|/\lambda_{max}$, i.e. the ratio of the smallest negative eigenvalue to the largest positive one. The overall contribution of negative eigenvalues can be estimated by $r_{neg} := \sum_{\lambda_i<0} |\lambda_i| / \sum_{j=1}^{r} |\lambda_j|$. Any symmetric D can also be made metric by adding a suitable value c to all off-diagonal elements of D. Such a constant can be found as $c = \max_{p,q,t} |d_{pq} + d_{pt} - d_{qt}|$. A smaller value making D metric was determined by us in a binary search. Table 1 provides suitable information on the Euclidean and metric aspects of the measures considered:

- Pen-angle is moderately non-Euclidean and nearly metric.
- Pen-dist is both moderately non-Euclidean and non-metric.
- NIST-matching is highly non-Euclidean and highly non-metric.
- Polygon-hausd is highly non-Euclidean, yet metric.
- Polygon-mhausd is moderately non-Euclidean and slightly non-metric.

The experiments are repeated 50 times for representations sets of various sizes and the results are averaged. $|R|$ varies from 3 to 50 examples per class (ten classes) for the digit DRs and from 5 to 100 examples per class (two classes) for the polygon DRs. For each $|R|$, two cases for the training set T are considered: $T = R$ or T consists of 100/200 objects per class for the digit/polygon DRs, respectively. In the latter case, the ratio of $|T|/|R|$ becomes smaller with a growing $|R|$. The test sets consist of 2488/1000/3600 examples for the pendigit/NIST/polygon data, correspondingly. For each DR, the k-NN rule is considered, as well as the linear discriminant built in both embedded and dissimilarity spaces. The embedding is derived from $D(R, R)$, but additional objects $T \backslash R$, if available, are projected there and used for constructing classifiers. To denoise the data and avoid the curse of dimensionality, the dimensionality of the embedded space was fixed to $0.3|R|$, so the dimensions corresponding to small eigenvalues (in magnitude) are neglected. Also the principal component analysis was applied in the dissimilarity space $D(\cdot, R)$ to reduce the dimensionality to $0.3|R|$. In both cases, although the dimensionalities are reduced, the spaces are still defined by all the objects of R.

Adding a constant to the dissimilarities or applying a concave transformation preserves their order, hence it does not influence the k-NN rule. However, by clipping (neglecting all negative eigenvalues in the embedding), the re-computed Euclidean distances differ non-monotonically from the original ones, hence the k-NN rule behaves differently. Also both embedded and dissimilarity spaces change, so a linear classifier will change as well. (Adding a constant is not worth doing in dissimilarity spaces, since a constant shift is applied to all D_{ij}, but the self-dissimilarity $D_{ii} = 0$. This is expected to worsen a classifier performance). In our experiments, we study the influence of such corrections on the given measures for various R. For this purpose, proper κ and τ guaranteeing Euclideaness are chosen. Two concave transformations are considered: the square root (which makes the measures close to Euclidean, yet still not Euclidean) and the sigmoid with the slope $s := avr(D(R, R))$. Such measures are non-Euclidean, but less than the original ones as judged by magnitudes of negative eigenvalues in the embeddings.

The results of our experiments compare the averaged performance of the linear discriminant (LDC) and 1-NN rule and the best k-NN rule. They are presented in Fig. 1-5. The standard deviations (for all the data) reach on average 0.3% and maximally $0.8 - 1.4\%$ for very small R. Due to lack of space, the performance of the RQDC02 (regularized quadratic classifier with a relative regularization of 0.2) is shown in Fig. 2 for the Pen-dist data only to indicate that such a classifier can reach even better accuracy than the LDC. The notation in figures refers to:

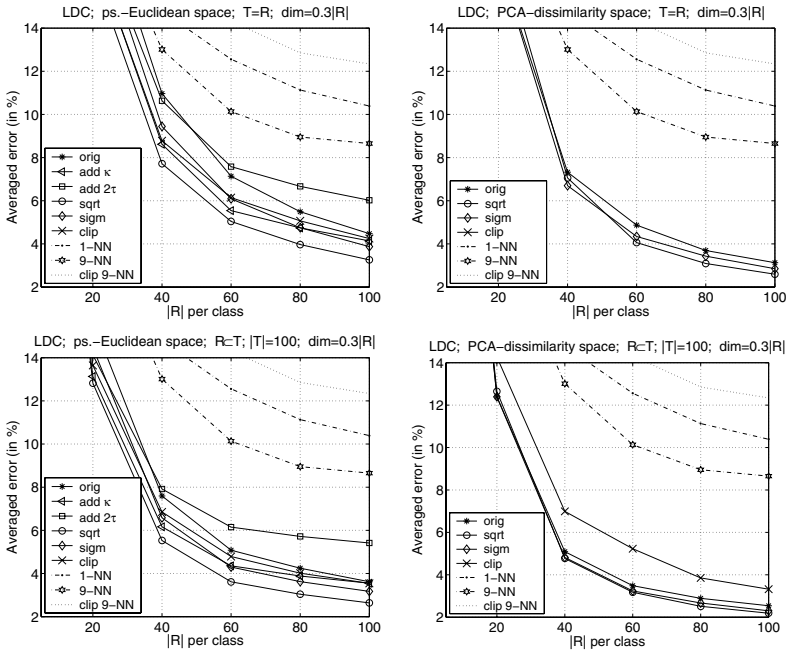

Fig. 4. The LDC performance for the Polygon-hausd dissimilarity data.

- *orig* - original dissimilarities; no transformation applied.
- *add* κ/2τ - a constant added to the dissimilarities; makes $D(R, R)$ Euclidean.
- *sqrt/sigm* - a square root/sigmoid transformation of the dissimilarities; makes $D(R, R)$ 'more' Euclidean.
- *clip* - only positive eigenvalues are used; a new Euclidean D_{eu} is derived from D.

The following general conclusions can be made by analyzing our results:

1. The correction by adding 2τ yields worse results than by adding κ (the former results are missing on some plots since they are out of the given scales).
2. The LDC and the RQDC in (corrected or not) dissimilarity spaces perform similarly or better than in pseudo-Euclidean spaces (compare right vs. left columns in all the figures).
3. For larger T and smaller R, the LDC/RQDC in both embedded and dissimilarity spaces (original or transformed by a square root or a sigmoid function) significantly outperform the k-NN and clip k-NN rules (bottom rows in all the figures). For $T = R$, this phenomenon is much less pronounced; the k-NN might even become somewhat better as observed for the Pen-angle data, Fig. 1.
4. Concave transformations of dissimilarities have a minor effect on the LDC/RQDC constructed in dissimilarity spaces. On the contrary, 'clipping' can deteriorate their performance.

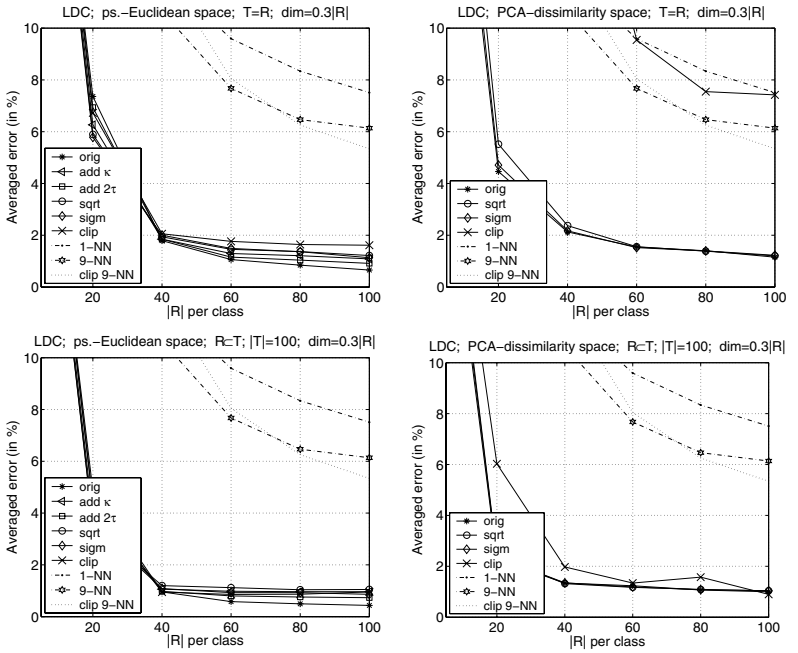

Fig. 5. The LDC performance for the Polygon-mhausd dissimilarity data.

5. The LDC/RQDC built in pseudo-Euclidean spaces derived from concave transformations of the dissimilarities may perform better than for the original dissimilarities or than the LDC/RQDC built in Euclidean spaces obtained from the corrections by clipping or by adding a constant. Still, the results reached by the LDC/RQDC in dissimilarity spaces are comparable or better.

5 Conclusions

If the k-NN is far from optimal for small representation sets, it can be significantly outperformed by linear (quadratic) classifiers built in both embedded or dissimilarity spaces. Concave transformations of dissimilarities are somewhat beneficial for classifiers in the embedded spaces, however, they may have no essential effect in dissimilarity spaces. None of the transformations considered here allows for reaching a considerably better performance than the results in original dissimilarity spaces. However, the transformations may influence the error and reject tradeoff [3]. We conclude that the potential advantages of imposed Euclideaness are doubtful. It is simply more important that the measure itself describes compact classes. This can be influenced by concave transformations which aim at diminishing the relative effect of large dissimilarities and not by making them really Euclidean or metric.

Acknowledgments

This work is supported by the Dutch Organization for Scientific Research (NWO). The authors thank prof. A. Jain and dr D. Zongker for providing the NIST dissimilarity data.

References

1. C.L. Blake and C.J. Merz. UCI repository of machine learning databases. University of California, Irvine, Dept. of Information and Computer Sciences, 1998.
2. H. Bunke, S. Günter, and X. Jiang. Towards bridging the gap between statistical and structural pattern recognition: Two new concepts in graph matching. In *Conf. on Advances in Pattern Recognition; LNCS 2013*, pages 1–11, 2001.
3. C.K. Chow. On optimum recognition error and reject tradeoff. *IEEE Trans. on Information Theory*, IT-16(1):41–46, 1970.
4. P. Courrieu. Straight monotonic embedding of data sets in Euclidean spaces. *Neural Networks*, 15:1185–1196, 2002.
5. L. Goldfarb. A new approach to pattern recognition. In *Progress in Pattern Recognition*, volume 2, pages 241–402. Elsevier Science Publishers B.V., 1985.
6. J.C. Gower. Metric and Euclidean Properties of Dissimilarity Coefficients. *Journal of Classification*, 3:5–48, 1986.
7. D.W. Jacobs, D. Weinshall, and Y. Gdalyahu. Classification with Non-Metric Distances: Image Retrieval and Class Representation. *IEEE Trans. on PAMI*, 22(6):583–600, 2000.
8. A.K. Jain and D. Zongker. Representation and recognition of handwritten digits using deformable templates. *IEEE Trans. on PAMI*, 19(12):1386–1391, 1997.
9. F. Moreno-Seco, L. Micó, and J. Oncina. A modification of the LAESA algorithm for approximated k-nn clasification. *Pattern Recogn. Letters*, 24(1-3):47–53, 2003.
10. Dubuisson M. P. and Jain A. K. Modified Hausdorff distance for object matching. In *12th Int. Conf. on Pattern Recognition*, volume 1, pages 566–568, 1994.
11. E. Pękalska. *working title: Dissimilarity-based pattern recognition.* PhD thesis, Delft University of Technology, The Netherlands, to appear, 2004.
12. E. Pękalska and R.P.W. Duin. Dissimilarity representations allow for building good classifiers. *Pattern Recogn. Letters*, 23(8):943–956, 2002.
13. E. Pękalska, P. Paclík, and R.P.W. Duin. A Generalized Kernel Approach to Dissimilarity Based Classification. *J. of Mach. Learn. Research*, 2:175–211, 2001.
14. V. Roth, J. Laub, J.M. Buhmann, and K.-R. Müller. Going metric: Denoising pairwise data. In *Advances in NIPS 15*, pages 841–856. MIT Press, 2003.
15. C.L. Wilson and M.D. Garris. Handprinted character database 3. Technical report, National Institute of Standards and Technology, February 1992.