Spatial Representation of Dissimilarity Data via Lower-Complexity Linear and Nonlinear Mappings

Elżbieta Pekalska and Robert P. W. Duin

Pattern Recognition Group, Department of Applied Physics
Faculty of Applied Sciences, Delft University of Technology
Lorentzweg 1, 2628 CJ Delft, The Netherlands
{ela,duin}@ph.tn.tudelft.nl

Abstract. Dissimilarity representations are of interest when it is hard to define well-discriminating features for the raw measurements. For an exploration of such data, the techniques of multidimensional scaling (MDS) can be used. Given a symmetric dissimilarity matrix, they find a lower-dimensional configuration such that the distances are preserved. Here, Sammon nonlinear mapping is considered. In general, this iterative method must be recomputed when new examples are introduced, but its complexity is quadratic in the number of objects in each iteration step. A simple modification to the nonlinear MDS, allowing for a significant reduction in complexity, is therefore considered, as well as a linear projection of the dissimilarity data. Now, generalization to new data can be achieved, which makes it suitable for solving classification problems. The linear and nonlinear mappings are then used in the setting of data visualization and classification. Our experiments show that the nonlinear mapping can be preferable for data inspection, while for discrimination purposes, a linear mapping can be recommended. Moreover, for the spatial lower-dimensional representation, a more global, linear classifier can be built, which outperforms the local nearest neighbor rule, traditionally applied to dissimilarities.

1 Introduction

An alternative to the feature-based description is a representation based on dissimilarity relations between objects. Such representations are useful when features are difficult to obtain or when they have little discriminative power. Such situations are encountered in practice, especially when shapes, blobs, or some particular image characteristics have to be recognized [6,8]. The use of dissimilarities is, therefore, dictated by the application or data specification.

For an understanding of dissimilarity data, techniques of multidimensional scaling (MDS) [1,10] can be used. MDS refers to a group of methods mainly used for visualizing the structure in high-dimensional data by mapping it onto a 2- or 3-dimensional space. The output of MDS is a spatial representation of the data, i.e. a configuration of points, representing the objects, in a space. Such a

T. Caelli et al. (Eds.): SSPR&SPR 2002, LNCS 2396, pp. 488–497, 2002.

display is believed to allow for a better understanding of the data, since similar objects are represented by close points.

In the basic approach, MDS is realized by Sammon mapping [1,10]. This nonlinear, iterative projection minimizes an error function between original dissimilarities and Euclidean distances in a lower-dimensional space. For n objects, it requires computation of $\mathcal{O}(n^2)$ distances in each iteration step and the same memory storage. However, for a lower, m-dimensional representation, only mn variables should be determined, which suggests that a number of $\mathcal{O}(n^2)$ constraints on distances is redundant and, therefore, could be neglected. The leads to the idea that only distances to the, so-called, representation set R (a subset of all objects), could be preserved, for which a modified version of the Sammon mapping should be considered. A similar reduction of complexity can be applied to a linear projection of dissimilarity data, being an extension of Classical Scaling, i.e. the linear MDS technique [1].

In this paper, we compare the linear and nonlinear projection methods, reduced in complexity, for data visualization and classification. Our experiments show that for dissimilarity data of smaller intrinsic dimensionality, its lower-dimensional spatial representation allows for building a classifier that significantly outperforms the nearest neighbor (NN) rule, traditionally used to discriminate between objects represented by dissimilarities. The NN rule, based on local neighborhoods, suffers from sensitivity to noisy objects. The spatial representation of dissimilarities, reflecting the data structure, is defined in a more global way, and therefore, better results can be achieved.

The paper is organized as follows. Sections 2 and 3 give insight into linear and nonlinear projections of the dissimilarity data. Section 4 explains how the reduction of complexity is achieved. Section 5 describes the classification experiments conducted, presents some $2D$ projection maps and discusses the results. Conclusions are summarized in section 6.

2 Linear Projection of the Dissimilarity Data

Non-metric distances may arise when shapes or objects in images are compared e.g. by template matching [8,6]. For projection purposes, the symmetry condition is necessary, but for any symmetric distance matrix, an Euclidean space is not 'large enough' for a distance-preserving linear mapping onto the specified dimensionality. It is, however, always possible [4] for a pseudo-Euclidean space.

The Pseudo-Euclidean Space A pseudo-Euclidean space $\mathcal{R}^{(p,q)}$ of the signature (p,q) [5,4] is a real linear vector space of dimension $p+q$, composed of two Euclidean subspaces, $\mathcal{R}^p$ and $\mathcal{R}^q$, such that $\mathcal{R}^{(p,q)}=\mathcal{R}^p \oplus \mathcal{R}^q$ and the inner product $\langle \cdot, \cdot \rangle$ is positive definite on $\mathcal{R}^p$ and negative definite on $\mathcal{R}^q$. The inner product w.r.t. the orthonormal basis is defined as $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{p} x_i y_i - \sum_{j=p+1}^{p+q} x_j y_j = \boldsymbol{x}^T M \boldsymbol{y}$, $M = \begin{bmatrix} I_{p\times p} & 0 \\ 0 & -I_{q\times q} \end{bmatrix}$, where I is the identity matrix. Using the notion of

inner product, $d^2(\boldsymbol{x},\boldsymbol{y}) = ||\boldsymbol{x}-\boldsymbol{y}||^2 = \langle \boldsymbol{x}-\boldsymbol{y}, \boldsymbol{x}-\boldsymbol{y}\rangle = (\boldsymbol{x}-\boldsymbol{y})^T M(\boldsymbol{x}-\boldsymbol{y})$, can be positive, negative or zero. Note that an Euclidean space $\mathcal{R}^p$, is a pseudo-Euclidean space $\mathcal{R}^{(p,0)}$.

Linear Projection and Generalization to New Objects Let T consists of n objects. Given a symmetric distance matrix $D(T,T) \in \mathcal{R}^{n\times n}$, a configuration $X_{red} \in \mathcal{R}^{n\times m}$ $(m < n)$ in a pseudo-Euclidean space can be found, up to rotation and translation, such that the distances are preserved as well as possible. Without loss of generality, a linear mapping is constructed such that the origin coincides with the mean. X is then determined, based on the relation between distances and inner products. The matrix of inner products B can be expressed only by using the square distances $D^{(2)}$ [4,9]:

$$B = -\frac{1}{2} J D^{(2)} J, \quad J = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T \in \mathcal{R}^{n\times n}, \tag{1}$$

where J takes care that the final configuration has a zero mean. By the eigendecomposition of $B = XMX^T$, one obtains: $B = Q\Lambda Q^T = Q|\Lambda|^{\frac{1}{2}} \begin{bmatrix} M \\ 0 \end{bmatrix} |\Lambda|^{\frac{1}{2}} Q^T$, where $|\Lambda|$ is a diagonal matrix of first, decreasing p positive eigenvalues, then decreasing absolute values of q negative eigenvalues, and finally zeros. Q is the matrix of corresponding eigenvectors and $M \in \mathcal{R}^{k\times k}$, $k = p+q$, is defined as before (or it is equal to $I_{k\times k}$ if $\mathcal{R}^k$ is Euclidean). X is then represented in the space $\mathcal{R}^k$ as $X = Q_k\,|\Lambda_k|^{\frac{1}{2}}$ [4]. Note that X is an uncorrelated representation, i.e. given w.r.t. the principal axes. The reduced representation $X_{red} \in \mathcal{R}^{n\times m}$, $m < k$, is, therefore, determined by largest p' positive and smallest q' negative eigenvalues, i.e. $m = p'+q'$, and it is found as [4,9]:

$$X_{red} = Q_m\,|\Lambda_m|^{\frac{1}{2}}, \tag{2}$$

New objects can be orthogonally projected onto the space $\mathcal{R}^m$. Given the matrix of square distances $D_n^{(2)} \in \mathcal{R}^{s\times n}$, relating s new objects to the set T, a configuration X_{red}^n is then sought. Based on the matrix of inner products $B^n \in \mathcal{R}^{s\times n}$:

$$B^n = -\frac{1}{2}(D_n^{(2)} J - U D^{(2)} J), \quad U = \frac{1}{s}\mathbf{1}\mathbf{1}^T \in \mathcal{R}^{s\times n}, \tag{3}$$

$$X_{red}^n = B^n X_{red} |\Lambda_m|^{-1} M_m \quad \text{or} \quad X_{red}^n = B^n B^{-1} X_{red}. \tag{4}$$

Classifiers For a pseudo-Euclidean configuration, a linear classifier $f(\boldsymbol{x}) = \langle \boldsymbol{v}, \boldsymbol{x}\rangle + v_0 = \boldsymbol{v}^T M\boldsymbol{x} + v_0$ can be constructed by addressing it as in the Euclidean case, i.e. $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x}\rangle_{Eucl} + v_0 = \boldsymbol{w}^T\boldsymbol{x} + v_0$, where $\boldsymbol{w} = M\boldsymbol{v}$; see [4,9].

3 Nonlinear Projection of the Dissimilarity Data

Sammon mapping [10,1] is the basic MDS technique used. It is a nonlinear projection onto an Euclidean space, such that the distances are preserved. For

this purpose, an error function, called *stress*, is defined, which measures the difference between the original dissimilarities and Euclidean distances of the configuration X (consisting of n objects) in an m-dimensional space. Let D be the given dissimilarity matrix and $\tilde{D}$ be the distance matrix for the projected configuration X. A variant of the Sammon stress is here considered [3,10]:

$$S = \frac{1}{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij}^2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (d_{ij} - \tilde{d}_{ij})^2 \tag{5}$$

and it is chosen since it emphasizes neither large nor small distances. To find a Sammon representation, one starts from an initial configuration of points for which all the pairwise distances are computed and the stress value is calculated. Next, the points are adjusted such that the stress will decrease. This is done in an iterative manner, until a configuration corresponding to a (local) minimum of S is found. Here, the scaled conjugate gradients algorithm is used to search for the minimum of S. It is important to emphasize that the minimum found depends on the initialization. In this paper, the principal component projection of the dissimilarity data is used to initialize the optimization procedure.

4 Reduction of Complexity

For our (non-)linear projection, although X has the dimensionality m, it is still determined by n objects. In general, such a space can be defined by $m+1$ linearly independent objects. If they were lying one in the origin and the others on the axes, they would determine our space exactly. Since this is unlikely to happen, the space retrieved will be an approximation of the original one. When more objects are used, the space becomes more filled and, therefore, better defined. The question now arises how to select the representation set $R \subseteq T$ of the size $r > m$, on which the (non-)linear mapping could be based. Following [2], we choose objects, lying in the areas of higher density, i.e. with relatively many close neighbors. For a dissimilarity representation $D(T, T)$, a natural way to proceed is the K-centers algorithm. It looks for K center objects, i.e. examples that minimize the maximum of the distances over all objects to their nearest neighbors, i.e. it minimizes the error $E_{K-cent} = \max_i (\min_k d_{ik})$. It uses a forward search strategy, starting from a random initialization. (Note that the K-means [3] cannot be used since no potential feature representation is assumed.)

For a chosen R, the linear mapping onto m-dimensional space is defined by formulas (1)–(2) based on $D(R, R)$. The remaining objects $D(T \backslash R, R)$ can then be added by the use of (3) and (4). In this way, the complexity is reduced from $\mathcal{O}(mn^2)$ (computing m eigenvectors and eigenvalues) to $\mathcal{O}(mr^2) + \mathcal{O}(nr)$.

In case of the Sammon mapping, a modified version should be defined, which generalizes to new objects. Following [2], first the Sammon mapping of $D(R, R)$ onto the space $\mathcal{R}^m$ is performed, yielding the configuration X_R.The remaining objects can be mapped to this space, while preserving the dissimilarities to the

set R, i.e. $D' = D(T \backslash R, R)$. This can be done via an iterative minimization procedure of the modified stress S_M, using the found representation X_R:

$$S_M = \frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{r}(d'_{ij})^2} \sum_{i=1}^{n}\sum_{j=1}^{r}(d'_{ij} - \tilde{d}'_{ij})^2 \tag{6}$$

This procedure allows for adding objects to an existing map, which can now be used for classification purposes. Its complexity reduces from $\mathcal{O}(mn^2)$, computing $\mathcal{O}(n^2)$ distances in the $\mathcal{R}^m$ space, to $\mathcal{O}(nmr+nr^2)$ in each iteration step.

5 Experiments

Two datasets are used in our study. The first data consists of randomly generated polygons (see Figure 1): 4-edge convex polygons and 7-edge convex and non-convex polygons. The polygons are first scaled and then the modified Hausdorff distance [8] is computed. The second data describes the NIST digits [11], represented by 128×128 binary images. Here, the symmetric dissimilarity, based on deformable template matching, as defined by Zongker and Jain [7], is used.

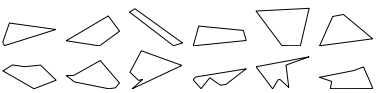


Fig. 1. Examples of the polygons

The experiments are performed 50/10 times for the polygon/digit data and the results are averaged. In each run, both datasets are randomly split into equally sized the training and testing sets. Each class is represented by 50/100 objects (i.e. $n = 100/1000$) for the polygon/digit data. In each experiment, first the dimensionality m of the projection is established. In case of the linear mapping, one may predict the intrinsic dimensionality based on the number of significant eigenvalues [4,9] (similarly to the principal component analysis [3]). However, this might be different for Sammon mapping. Therefore, a few distinct dimensionalities are used. For the dimensionality m, representation sets of the size r, varying from $m+1$ to n are considered. Each set R is selected by the K-centers algorithm, except for R equal to the training set T (i.e. $r = n$). Next, an approximated space, defined by objects from R is determined (i.e. the (non-)linear mapping is based on $D(R, R)$). The remaining $T \backslash R$ objects are then mapped to this space, as described in section 4 and the Fisher linear classifier (FLC) is trained on *all* n objects (a quadratic classifier has also been used, but the linear one performs better). The test data is then projected to the space and the classification error is found. For a new object, only r distances have to be computed and the complexity of the testing stage becomes $\mathcal{O}(mr)$ for the linear projection and $\mathcal{O}(\max(mr, r^2))$ in each iteration step for Sammon mapping.

The results of our experiments on the polygon/digit data are presented in Figure 2. For the polygon data, the best performance of the FLC is achieved when the dimensionality of the projected space is 15 for Sammon mapping or 20 for the linear mapping. For the set R consisting of only 20 training objects, the FLC built in both linear and nonlinear projected spaces (i.e. using distances to

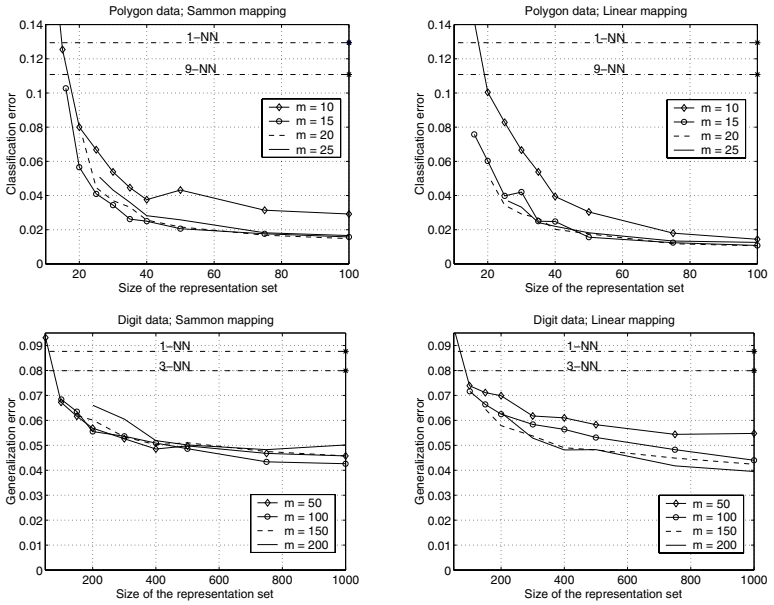


Fig. 2. The NN rule on dissimilarities (marked by '*') and the FLC on the spatial representations for the polygon data (top) and the digit data (bottom)

the set R only), outperforms the 1-NN rule and the best 9-NN rule, both based on 100 objects. This shows that by making use of the structure information present in the data, a less noise-sensitive decision rule than, the NN method, can be constructed. When R contains $30-40\%$ of the data, the error of nearly 0.02 is reached, which is close to the error of $0.015-0.018$ gained when $R{=}T$.

For the digit data, the best accuracy is found when $m = 100$ or $m = 200$ for the Sammon mapping or the linear projection, respectively. For the set R, consisting of 10% of the training objects only, the FLC built in both nonlinear and linear 50-dimensional spaces, outperforms the 1-NN rule and the best 3-NN rule, both based on all 1000 objects. When $r = 400$ objects are chosen to the set R, an error of 0.05 can be reached; when $R{=}T$, an error of 0.04 is achieved.

In Figure 3, one can also observe that for both data, the stress S changes only slightly when R is larger than half of the training set. For the linear mapping, the stress values are not shown for $r{=}m{+}1$, since some of the pseudo-Euclidean distances are negative and S becomes complex. For larger r, the imaginary part of S becomes nearly zero and can, therefore, be neglected. The stress is, of course, relatively large for the linear mapping, but this does not disturb a good classification performance. Apparently, the variance present in the data, revealed by the linear projection, is good enough for discrimination purposes, since major differences in classes are captured.

In summary, in terms of the stress, a nonlinear configuration preserves the data structure better than the linear one. The nonlinear mapping requires less

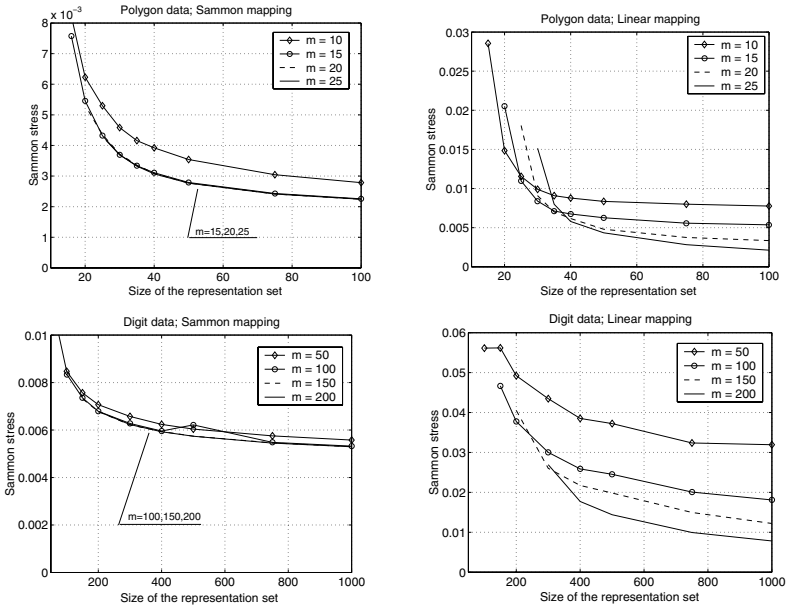


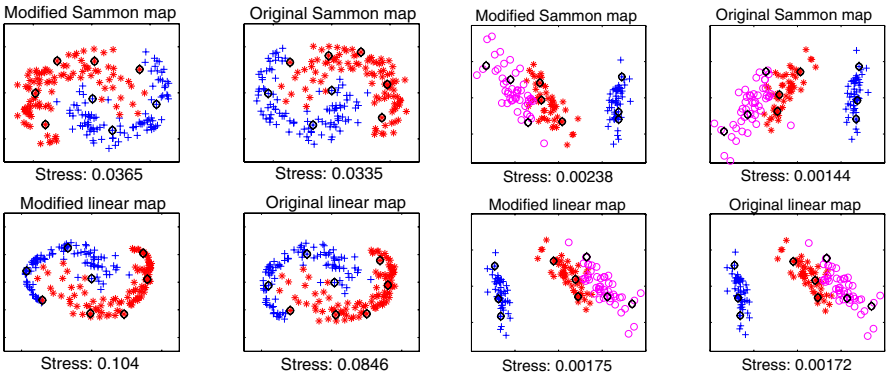
Fig. 3. Sammon stress for the spatial representations of the polygon data (top) and the digit data (bottom)

dimensions for about the same performance of the FLC than the linear mapping, although for the latter, a bit higher accuracy can overall be reached.

Visualization From all the linear mappings of the fixed dimensionality, our linear projection preserves the distances in the best way [1,4]. Since it is constructed to explain the maximum of the (generalized) variance in the data, some details in the structure might remain unrevealed. When data lies in a nonlinear subspace, Sammon mapping is preferred since it provides an additional information.

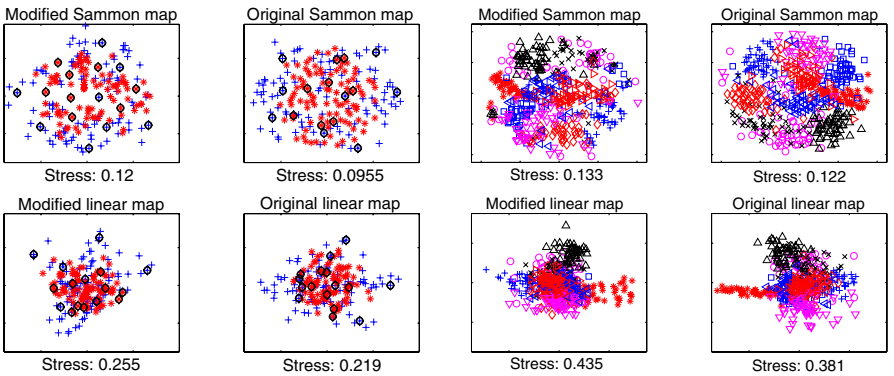
The difference between the original (non-)linear $2D$ maps and the maps based on smaller representation sets can be observed in Figure 4, where the results for four datasets are shown. The first two examples are illustrative: banana dataset is an artificial $2D$ dataset for which the theoretical, nearly-Euclidean distance is found; for the $4D$ Iris dataset, the Euclidean distance is considered. The last two datasets refer to data from our classification experiments.

Each subfigure presents plots for the linear and nonlinear projections. Those plots show the difference between the original (non-)linear maps and the maps, constructed while preserving the dissimilarities to the set R only. From Figure 4, one can observe that the (non-)linear maps, based on a smaller R resemble well the original maps, based on all objects. The Sammon stress computed for those configurations reveals the loss up to 20%. This is reasonable, given that the



(a) Banana data; r = 8

(b) Iris data; r = 9



(c) Polygon data; r = 16

(d) Digit data; r = 50

Fig. 4. Linear and nonlinear 2D maps; set R marked in black 'o' when feasible

chosen R consists of less than 10% of all objects, which means that around 90% of distances are not taken into account during the mapping process.

6 Discussion and Conclusions

The presented mappings of finding a faithful spatial configuration do not make use of class labels. So, the class separability could potentially be enhanced by using such information. This remains an open issue for further research.

To reduce noise in the data, in a mapping process, the distances are preserved approximately. By this, the class separability may be somewhat improved, although, in general, it is reflected in a similar way as given by all the dissimilarity relations. The advantage of building e.g. a linear classifier in such a projected space over the k-NN is that the data information is used in a more complex

and comprehensive way, based on relations between a number of objects both in the mapping process and in the classifier construction. Since the k-NN rule is locally noise sensitive, for dissimilarity data, noisy in local neighbourhoods, our approach can be beneficial. It is important to emphasize, however, that the generality of our approach holds for data of a lower intrinsic dimensionality.

A number of conclusions can be drawn from our study. First of all, the modified Sammon algorithm allows for adding new data to the existing map. Secondly, the (non-)linear mapping onto m dimensions, based on the set R of the size r, reduces its complexity both in the training and testing stage. For an evaluation of a novel object, only r dissimilarities have to be computed, and for the linear mapping $\mathcal{O}(mr)$ operations are needed, while for the Sammon mapping, $\mathcal{O}(\max(mr, r^2))$ operations are necessary in each iteration step.

Thirdly, the projections considered, allow for obtaining a spatial configuration of the dissimilarity data, which can be beneficial for the classification task. Our experiments with dissimilarity representations of the polygon and digit data show that such spaces offer a possibility to build decision rules that significantly outperform the NN method. Based on the set R consisting of 45% of the training objects, the FLC, constructed in a projected space defined by the dissimilarities to R only, reaches an error of 0.02/0.05, while the best NN rule makes an error of 0.11/0.088 and makes use of all objects.

Next, the 2D spatial representations of dissimilarity data, obtained by the linear and modified Sammon projections, resemble the original maps. A similar structure is revealed in the data when R consists of 10% of objects, chosen by the K-centers algorithm, as well as of all of them. These approaches are especially useful when dealing with large datasets. In general, Sammon maps provide an extra insight into the data and can be preferred for visualization. Our experience shows also that the use of the K-centers is not crucial; what is important is the choice of significantly *different* objects to represent the variability in the data.

Finally, the FLC built on the linear configuration yields about the same (somewhat better) classification results as the FLC on the modified-Sammon representation, but in a space of a larger dimensionality than for the nonlinear case. However, since no iterations are involved for an evaluation of novel examples, the linear projection can be recommended for the classification task.

Acknowledgments

This work is supported by the Dutch Organization for Scientific Research (NWO). The authors thank prof. Anil Jain for the NIST dissimilarity data.

References

1. I. Borg and P. Groenen. *Modern Multidimensional Scaling.* Springer-Verlag, New York, 1997. 488, 489, 490, 494
2. D. Cho and D. J. Miller. A Low-complexity Multidimensional Scaling Method Based on Clustering. *concept paper*, 2002. 491

3. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2nd edition, 2001. 491, 492
4. L. Goldfarb. A new approach to pattern recognition. In L.N. Kanal and A. Rosenfeld, editors, *Progress in Pattern Recognition*, volume 2, pages 241–402. Elsevier Science Publishers B.V., 1985. 489, 490, 492, 494
5. W. Greub. *Linear Algebra*. Springer-Verlag, 1975. 489
6. D. W. Jacobs, D. Weinshall, and Y. Gdalyahu. Classification with Non-Metric Distances: Image Retrievaland Class Representation. *IEEE Trans. on PAMI*, 22(6):583–600, 2000. 488, 489
7. A. K. Jain and D. Zongker. Representation and recognition of handwritten digits using deformable templates. *IEEE Trans. on PAMI*, 19(12):1386–1391, 1997. 492
8. 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. 488, 489, 492
9. E. Pekalska, P. Paclík, and R. P. W. Duin. A Generalized Kernel Approach to Dissimilarity Based Classification. *J. of Mach. Learn. Research*, 2:175–211, 2001. 490, 492
10. J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transaction on Computers*, C-18:401–409, 1969. 488, 489, 490, 491
11. C. L. Wilson and M. D. Garris. Handprinted character database 3. Technical report, National Institute of Standards and Technology, February 1992. 492