Prototype selection for finding efficient representations of dissimilarity data

Elżbieta Pękalska 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
E-mail: {ela,duin}@ph.tn.tudelft.nl

Abstract

The nearest neighbor (NN) rule is a simple and intuitive method for solving classification problems. Originally, it uses distances to the complete training set. It performs well, however, it is sensitive to noisy objects, due to its operation on local neighborhoods only. A more global approach is possible by mapping the distance data onto a pseudo-Euclidean space, such that the distances are preserved as well as possible. Then, a classifier built in such a space can outperform the NN rule. However, again all objects from the training set are used for a projection of new data.

This paper addresses the issue of reducing the training set while possibly preserving the original structure of the mapped data. Some criteria are introduced and evaluated against two problems, polygon recognition and digit recognition. Our experiments show that the representation mismatch criterion is beneficial for the applications considered. Moreover, the linear classifier built in the pseudo-Euclidean space, determined by $20\% - 25\%$ of the training objects, outperforms the NN rule based on all of them.

1. Introduction

In the pattern recognition area, objects are conventionally represented by features. Such a representation may be inefficient for learning purposes or offer a little discrimination power. Also, features might be hard to define. An alternative representation can be built by using the concept of dissimilarity. An object is then characterized in a relative way, i.e. by its dissimilarities to a set of prototypes. Some particular characteristics of objects or measurements, like curves or shapes [8, 7, 4], may naturally lead to distance representations as they make recognition tasks more feasible. The use of dissimilarities, built directly on measurements, e.g. based on template matching [3], is of interest.

The nearest neighbor rule (NN) is traditionally applied to such representations. It performs well, but it suffers from high computational complexity, a potentially worse performance for a small set of prototypes and sensitivity to noise. To overcome such limitations and improve the recognition accuracy, a linear mapping of the dissimilarity data can be performed such that the distances are preserved. In such a space, a linear classifier (i.e. a more global decision rule than the NN method) can be considered. To project new objects, however, the distances to all training examples have to be computed. The goal of our work here is to investigate some criteria for the selection of a reduced representation set R, on which the mapping will be based only. The advantage of a small R is that only a small number of dissimilarities is needed for an evaluation of a new object, while the classifier may profit from the complete training set.

Our experiments on polygon and handwritten digit recognition problems will demonstrate that the tradeoff between the recognition accuracy and the computational effort is significantly improved by building a linear classifier in the projected space, based on a small R, instead of using the NN rule on the complete training set.

2. Linear projection of the dissimilarity data

For an Euclidean (metric) distance matrix, an isometric mapping onto an Euclidean space can be found [1]. However, non-metric distances often seem to arise when shapes or objects in images are compared by template matching or for other types of distances built in e.g. computer vision [3, 7]. For projection purposes, however, the symmetry condition is necessary. But, in general, for any symmetric distance matrix, an Euclidean space is not 'large enough' for an isometric mapping. It is, however, always possible [5] when a pseudo-Euclidean space is considered.

The pseudo-Euclidean space

A pseudo-Euclidean space $\mathcal{R}^{(p,q)}$ of the signature (p,q) [6, 5] 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 $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{p} x_i y_i - \sum_{j=p+1}^{p+q} x_j y_j = \boldsymbol{x}^T M \boldsymbol{y}$, $M = \begin{bmatrix} I_{p\times p} & 0 \\ 0 & -I_{q\times q} \end{bmatrix}$, ($I$ is the identity matrix). $d^2(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x}-\boldsymbol{y}||^2 = \langle \boldsymbol{x}-\boldsymbol{y}, \boldsymbol{x}-\boldsymbol{y} \rangle = (\boldsymbol{x}-\boldsymbol{y})^T M (\boldsymbol{x}-\boldsymbol{y})$, which can be positive, negative or zero. Note also that $\mathcal{R}^p = \mathcal{R}^{(p,0)}$.

1051-4651/02 $17.00 (c) 2002 IEEE

Linear embedding and adding new objects

Let T be the training set of n objects. Given a symmetric distance matrix $D(T,T) \in \mathcal{R}^{n \times n}$, a configuration $X \in \mathcal{R}^{n \times k}$ ($k < n$) in a pseudo-Euclidean space can be found, up to rotation and translation, such that the (square) distances are preserved. Without loss of generality, a linear embedding (isometric mapping) is constructed such that the origin coincides with the vector representation of the first object $\boldsymbol{x}_1$. X is then determined, based on the relation between distances and inner products [5, 10]. The matrix of inner products $B = \{b_{ij}\}$, for the vector representation $\{\boldsymbol{x}_1, .., \boldsymbol{x}_n\}$, can be expressed only by using the pseudo-Euclidean distances (note that by our assumption $||\boldsymbol{x}_j||^2 = d^2(\boldsymbol{x}_j, \boldsymbol{x}_1) = d^2(\boldsymbol{x}_j, \mathbf{0})$):

$$b_{ij} = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle = -\frac{1}{2}\left[d^2(\boldsymbol{x}_i, \boldsymbol{x}_j) - d^2(\boldsymbol{x}_i, \boldsymbol{x}_1) - d^2(\boldsymbol{x}_j, \boldsymbol{x}_1)\right]. \quad (1)$$

By the eigendecomposition of $B = XMX^T$, one obtains:

$$B = Q\Lambda Q^T = Q|\Lambda|^{\frac{1}{2}} \begin{bmatrix} M & \\ & 0 \end{bmatrix} |\Lambda|^{\frac{1}{2}} Q^T, \quad (2)$$

where $|\Lambda|$ is a diagonal matrix of decreasing p positive eigenvalues followed by decreasing absolute values of q negative eigenvalues and then 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). Based on (2), X is represented in the space $\mathcal{R}^k$ [5] as:

$$X = Q_k \, |\Lambda_k|^{\frac{1}{2}}. \quad (3)$$

New objects can be orthogonally projected onto the space $\mathcal{R}^{(p,q)}$. A configuration X^n is then sought, given the distance matrix $D_n^{(2)} \in \mathcal{R}^{s \times n}$, relating s new objects to the set T. Based on the matrix of inner products $B^n = \{b_{ij}^n\} \in \mathcal{R}^{s \times n}$, consisting of:

$$b_{ij}^n = -\frac{1}{2}\left[d^2(\boldsymbol{x}_i^n, \boldsymbol{x}_j) - d^2(\boldsymbol{x}_i, \boldsymbol{x}_1) - d^2(\boldsymbol{x}_j^n, \boldsymbol{x}_1)\right], \quad (4)$$

X^n, the solution to $B^n = X^n M X^T$, is given by [5, 10]:

$$X^n = B^n X |\Lambda_k|^{-1} M \;\text{ or }\; X^n = B^n B^{-1} X. \quad (5)$$

Reduction of dimensionality

In practice, since dissimilarities are noisy measurements, k, determined by the number of non-zero eigenvalues of B, is often close to n, but the intrinsic dimensionality of the data can be much smaller. A way to proceed is to express X w.r.t. the principle components (PCs), and then select the significant directions only. This means that the full representation should be first constructed so that the PCA projection can be applied. However, such a representation can be directly determined if the matrix of inner products B_{pc} is properly chosen. This is achieved when B_{pc} is defined as [5, 10]:

$$B_{pc} = -\frac{1}{2} J D^{(2)} J, \quad J = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T \in \mathcal{R}^{n \times n} \quad (6)$$

and $D^{(2)}$ is the matrix of square distances. J is the centering matrix, taking care that the final configuration has a zero mean. By the eigendecomposition (2) of B_{pc}, X can be found by (3). Since X is now an uncorrelated vector representation, the reduction of dimensionality is performed by neglecting directions corresponding to eigenvalues small in magnitude. The reduced representation, preserving the distances approximately, is then determined by p' positive and q' negative eigenvalues. Therefore, $X_{red} \in \mathcal{R}^{n \times m}$, $m < k$, is found as $X_{red} = Q_m |\Lambda_m|^{\frac{1}{2}}$, where $m = p' + q'$.

new objects X^n w.r.t. the principal axes is found by (5), where Λ_m is used and the matrix of inner products, B_{pc}^n relating new objects to the set T objects, is given as [5, 10]:

$$B_{pc}^n = -\frac{1}{2}(D_n^{(2)} J - U D^{(2)} J), \quad U = \frac{1}{s}\mathbf{1}\mathbf{1}^T \in \mathcal{R}^{s \times n}. \quad (7)$$

Classifiers

For a pseudo-Euclidean configuration X, 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 [5, 10]).

3. Selection of the representation set

Reduction of dimensionality is useful for data representation, since both noise and non-significant information are neglected. Still, the reduced X_{red} is determined by all n training objects. For an m-dimensional pseudo-Euclidean space, only $m+1$ objects can define it: one object will serve as the origin and m objects will correspond to the basis vectors. The task is now as follows: given X_{red} w.r.t. to the principal axes, choose the representation set R of $m+1$ objects such that the projection defined by R, gives a configuration which is close to X_{red} (according to a criterion).

A set R, spanning the reduced space $\mathcal{R}^m = \mathcal{R}^{(p',q')}$ such that $\mathcal{R}^m$ is defined by m leading principal axes, might not, however, exist. To avoid an intractable search over all possible subsets, an error measure between the approximated and reduced/original configurations can be defined to be minimized, e.g. in a greedy approach (see [11]). The origin will always be fixed to the vector representation of the object r_0 which is the closest to the mean (i.e. the origin). Such an object is easily detected as the one whose average distance to T is the smallest [5, 10]. Having determined r_0, the whole configuration X_{red} is then shifted to the new origin. The basis objects, are then successively added, in each step minimizing a criterion, until m objects are found.

To assure that the chosen objects are linearly independent and to make our selection a feasible process, in the step j, only K objects $Z^j = \{z_1^j, ..., z_K^j\}$ with the largest (in magnitude) projections on the j-th principal axis are preselected to be tested against the specified criterion. K is assigned to the 10% of the training size. This holds for all criteria introduced below.

Basis reconstruction error (BRE)

Let $\mathcal{R}^{j-1}$ be a subspace of the reduced space $\mathcal{R}^{(p',q')}$ defined by the basis objects $R_{j-1} = \{r_1, ..., r_{j-1}\}$. In the step j, a potential basis set $\{R_{j-1}, z\}$, where $z \in Z^j$, is considered. An object z is chosen to be r_j as the one for which the smallest average reconstruction error of the basis vector representations is achieved.

Projection error (PE)
Let V be a j-dimensional subspace of the space $\mathcal{R}^k = \mathcal{R}^{(p,q)}$. Then, based on the properties of inner products and the embedding [5] (see section 2), the distance between a vector $\boldsymbol{x}_i \in \mathcal{R}^k$ and its projection $\boldsymbol{x}_i^V$ onto V can be expressed as:

$$||\boldsymbol{x}_i - \boldsymbol{x}_i^V||^2 = ||\boldsymbol{x}_i||^2 - ||\boldsymbol{x}_i^V||^2 = d^2(\boldsymbol{x}_i, \boldsymbol{x}_1) - \boldsymbol{b}_i^n \, B^{-1} \, (\boldsymbol{b}_i^n)^T, \quad (8)$$

where $\boldsymbol{b}_i^n$ is the i-th row of B^n and both B and B^n refer now to the representation in $\mathcal{R}^j$ defined by pairwise dissimilarities between $j+1$ objects (i.e. origin and the basis).

Having chosen the basis $R_{j-1} = \{r_1, ..., r_{j-1}\}$, in the step j, an object $z \in Z^j$ is selected to be r_j such that the average projection error (8) of all training objects onto the space $\mathcal{R}^j$, defined by the objects $\{r_0, R_{j-1}, z\}$ is the smallest.

Representation mismatch error (RME)
In the step j, for each object $z \in Z^j$, all training data is projected w.r.t. the potential basis set $\{R_{j-1}, z\}$, resulting in X_{pot}^j and compared to the j significant axes of the shifted representation X_{red} by computing the average square pseudo-Euclidean distance between corresponding vectors of the two configurations. The object z, providing the smallest mismatch, is then selected as r_j.

The results obtained by the above criteria, selecting $R = \{r_0, R_m\}$, are judged by three measures. The first one is the mean square error E_{mse}^D on the square distances:

$$E_{mse}^D = \frac{1}{n(n-1)} \sum_{i<j} \left(d_{ij}^2 - (d_{ij}^{apr})^2 \right)^2,$$

where d_{ij} are the original distances (i.e. of the full representation X) and d_{ij}^{apr} are the pseudo-Euclidean distances of either the reduced representation or of the approximated representation obtained by the projection based on the objects from R only. This measure indicates how much distortion was introduced to the distances by using the set R.

The second measure, S, is the ratio of the average between-class square distance B to the average within-class square distance W. It gives an indication on the class separability. Given N classes, S is defined as [11]:

$$S = \frac{1}{(N-1)(N-2)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{\sqrt{|B(C_i, C_j)|}}{\sqrt{|W(C_i)|} + \sqrt{|W(C_j)|}}$$

The third measure is the classification error on the independent test set. Since, in the end, our purpose is the classification task, it is not so crucial that the distances are well preserved when the classification results are good.

4. Experiments

Two datasets are used in our study. The first data consists of randomly generated polygons (see Fig.1): 4-edge convex polygons and 7-edge (non-)convex polygons. The polygons are first scaled and then the modified Hausdorff distance [3] is computed. The second data describes images of the

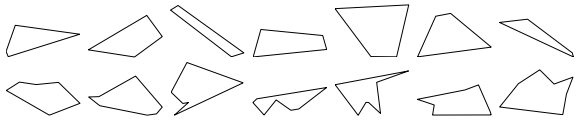

Figure 1. Examples of the polygon data.

NIST digits [12], with the symmetric dissimilarity (made by averaging suitable values) between two images based on deformable template matching, as defined in [8]. For such data considered, no input feature space is given.

The experiments are performed $100/10$ times for polygon/digit data and the results are averaged. In each run, both datasets are randomly split into equally sized the training and testing sets with $50/100$ objects per class for the polygon/digit data. In each experiment, first the reduced m-dimensional representation X_{red} is found and then the set R of $m+1$ objects is chosen according to a specified criterion. Next, the approximated space, defined by objects from R is determined (i.e. the mapping is based on $D(R, R)$ only). The Fisher linear classifier (FLC) is then trained both in the reduced and approximated spaces, where the new data is also projected and the generalization error is computed.

Figures 2 and 3 present the results (note differences in scale). The E_{mse}^D measure shows how much the distances of the reduced or approximated spaces differ from the original ones. The S measure gives an indication on the average separability between all pairs of classes (the original separability is the averaged S measure for the original distances). The classification error is given for the FLC. All the evaluation measures are shown in dependence of the reduced dimensionality m. On average, for m up to 10 (polygons) and for m up to 7 (digits), the reduced space is Euclidean. For larger m, it becomes strictly pseudo-Euclidean.

5. Discussion and conclusions

Analyzing the evaluation measures for reduced spaces determined by *all* n objects (solid lines in Fig. 2 and 3), we can see that the intrinsic dimensionality is about $12-15$ for the polygon data and $80-100$ for the digit data, since the measures stabilize their values. In such cases, efficient approximated spaces defined by $m+1$ objects can be built.

For a classification task, all the criteria selecting R, introduced here, allow the FLC for reaching about the same accuracy in the approximated spaces (see the rightmost plots of Fig. 2-3). They weaken their performance w.r.t. the reduced spaces defined by *all* objects, but their achievements are based on less then 25% of the data! It is important to emphasize that for $m = 13/80$ (polygons/digits), the FLC, built in a space defined by $m+1$ objects, performs as well as the 1-NN based on all of them. For $m = 24/180$ (polygons/digits), the FLC significantly outperforms both the 1-NN (an error of $0.13/0.089$ for the polygon/digit data) and the best k-NN, for $k = 1-15$, (an error of $0.11/0.081$ for the polygon/digit data). The condensed NN has also been considered [2], which, on average, uses

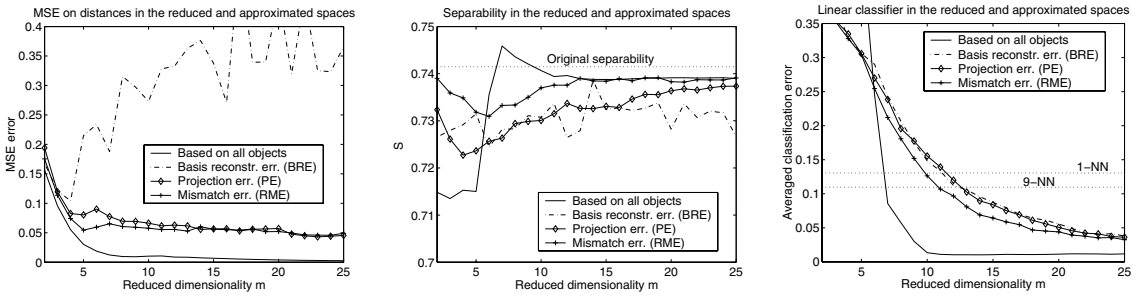

Figure 2. Polygon data; the MSE on distances, separability measure and classification error.

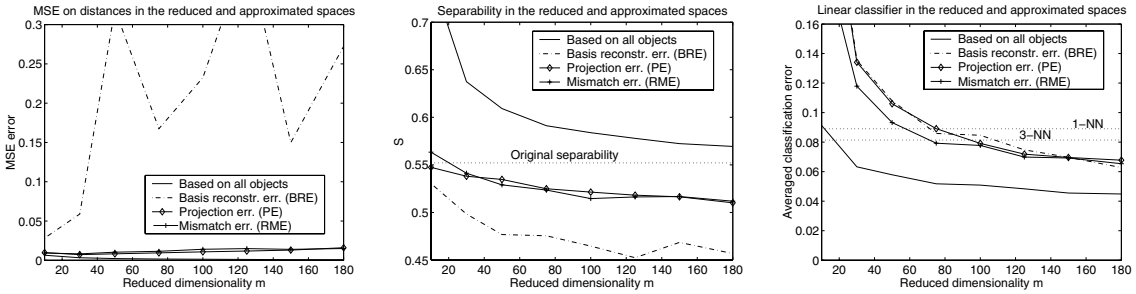

Figure 3. Digit data; the MSE on distances, separability measure and classification error.

$29/233$ (polygons/digits) prototypes and gives an error of $0.19/0.13$ (polygons/digits), while the FLC, constructed in the approximated space defined by the same number of prototypes, reaches an error of $0.038/0.062$.

Concerning two other evaluation measures, the BRE criterion is rather poor. It simply does not preserve the dissimilarity structure. The remaining two criteria behave about the same, however, the RME might be slightly preferable.

Important conclusions can be drawn from our study on dissimilarity data embedded in pseudo-Euclidean spaces. First of all, the FLC, built in the reduced space defined by all objects, can significantly outperform the NN rule.

Secondly, a set R of, in our case, at most 25% of the training objects can be selected, on which the approximated space is defined. In such a space, the FLC can reach much higher accuracy than the NN method based on *all* objects. The condensed NN, using $23-29\%$ of all the training objects, deteriorates much w.r.t. the 1-NN, by which our results become even more appealing.

Thirdly, the PE and RME criteria for the selection of R, define approximated spaces, where not only the FLC performs well, but the dissimilarity structure is mostly preserved. In the selection process, those criteria are judged w.r.t. all training objects, while the BRE criterion - w.r.t. the basis objects only. As observed (Fig. 2-3), this is not sufficient for preserving the structure, although the classification performance can still be good.

Next, it is possible to use more than $m+1$ objects in order to define an approximated space better; see [9]. Finally, no label information was used for the selection process. We expect that for classification purposes, even better accuracy can be reached, when such information is incorporated. This remains an issue for further research.

Acknowledgments

This work is partly 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.

[2] P. Devijver and J. Kittler. *Pattern recognition: A statistical approach*. Prentice/Hall, London, 1982.

[3] M. P. Dubuisson and A. K. Jain. Modified Hausdorff distance for object matching. In *12th Int. Conf. on Pattern Recognition*, volume 1, pages 566–568, 1994.

[4] S. Edelman. *Representation and Recognition in Vision*. MIT Press, Cambridge, 1999.

[5] L. Goldfarb. A new approach to pattern recognition. In L. Kanal and A. Rosenfeld, editors, *Progress in Pattern Recognition*, volume 2, pages 241–402. Elsevier Science Publishers B.V., 1985.

[6] W. Greub. *Linear Algebra*. Springer-Verlag, 1975.

[7] D. 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. Jain and D. Zongker. Representation and recognition of handwritten digits using deformable templates. *IEEE Trans. on PAMI*, 19(12):1386–1391, 1997.

[9] E. Pękalska and R. Duin. Spatial representation of dissimilarity data via lower-complexity linear and nonlinear mappings. In *International Workshop on SPR + SSPR; accepted*, 2002.

[10] E. Pękalska, P. Paclík, and R. Duin. A Generalized Kernel Approach to Dissimilarity Based Classification. *Journal of Machine Learning Research*, 2:175–211, 2001.

[11] R. Verma. A metric approach to isolated word recognition. Master's thesis, Dept. of Computer Science, University of Toronto, 1991.

[12] C. Wilson and M. Garris. Handprinted character database 3. Technical report, NIST, February 1992.