COMPLEXITY OF DISSIMILARITY BASED PATTERN CLASSES

Robert P.W. Duin, Elżbieta Pękalska

Pattern Recognition Group, Department of Applied Physics
Delft University of Technology, The Netherlands
duin@ph.tn.tudelft.nl, ela@ph.tn.tudelft.nl

ABSTRACT

If the proper dissimilarity measures are provided, some sets of objects, e.g. curves or blobs, may be better described by using representation sets instead of features. The dissimilarity matrix of such a set is the base for further analysis. The question arises how from a given dissimilarity matrix can be judged whether the size of the training set is sufficient to describe the peculiarities of the set. In this paper the problem is defined, some approaches are discussed and illustrated by the analysis of dissimilarity matrices defined by the modified Hausdorff distances between sets of hand-written digits.

Keywords: problem complexity, representation set, dissimilarity representation, sample size.

1. INTRODUCTION

In pattern recognition objects are traditionally represented by features. Features preferably have to be defined on the basis of expert knowledge on the application domain. Each object thereby corresponds with a vector in a feature space. The dimensionality of this space equals the number of features. Alternatively, objects may be represented by their similarities or dissimilarities to a set of prototype objects, which we will call the *representation set* [1, 2, 3]. Consequently, objects can now be represented as vectors in a *representation space*. The dimensionality of this space equals the number of objects (prototypes) in the representation set. Like features, the dissimilarity measure has to be defined on the basis of expert knowledge.

Until now, pattern classification on the basis of (dis)similarities has been mainly performed by the nearest neighbor rule (1-NN rule): a new object is assigned to the class of the most similar object in the representation set; also the k-NN rule can be used. Important application areas are those for which no natural features can be defined, e.g. the recognition of curves or blobs using various approaches of deformable template matching [7, 8]. However, the representation space, as defined above, is not used as such. Recently, we have proposed [3, 4] to apply the traditional feature based classifiers like the Fisher discriminant in *this* space. We found that such linear functions of dissimilarities can be more efficient (need less prototypes) and/or more accurate than their direct use by the NN rules.

In this paper we will study some issues related to the use of dissimilarity representations. Initially, all given training objects are candidates for the representation set. An analysis thereby starts with an $n \times n$ matrix of dissimilarities. An ever returning question is whether the size of the training set, i.e. n, is sufficiently large, i.e. whether much can be gained by increasing the number of training examples. This question is related to the complexity of the classification problem [5, 12], a still ill-defined concept. Here, we will restrict ourselves to the more simple issue of the size of an unlabeled set of objects, possibly belonging to a single class. We will investigate the possibilities to judge whether the representation set is sufficiently sampled or not.

In the next section the basis of our approach will be defined, starting from an assumption on the continuity of dissimilarities. From that a number of statistics will be proposed that may be used for judging the size of the representation set. In the following section these criteria are investigated experimentally on artificial and real world datasets. Results are discussed in the final section and some open questions are defined.

2. APPROACHES

Assume that a dissimilarity measure $d(r,s)$ is given between two real world objects r and s. Let $d(r,s) = 0$ if and only if r and s are identical and $d(r,s) > 0$ if r and s differ. Define the representation set R to be a finite set of real world objects $\{r_1, r_2, ..., r_n\}$. A new object r is now represented by a row vector $\mathbf{d}$ defined by the dissimilarities of the object r to all objects in the representation set, i.e. $\mathbf{d} = [d(r,r_1)\ d(r,r_2)\ ....\ d(r,r_n)]$. The entire set itself is now represented by a dissimilarity matrix D (or D(R,R) indicating the dependence on the representation set R), consisting of the row vectors $\mathbf{d}_i$, $i=1,...,n$. D is therefore a set of n vectors in the n-dimensional representation space.

The research question now refers to a criterion defined on D judging how well the dataset is sampled. Phrased in other words: can we expect that new objects are understood in terms of the given ones or not? In the next subsections we will describe some possible statistics that might be used as such a criterion. They are all based on the so-called *compactness hypothesis* [5, 9] which states that real world objects that are similar are also close in their representation. Effectively, this puts a constraint on the dissimilarity measure. It has to be such that $d(r,s)$ is small if r and s are

very similar, i.e. it should be much smaller for similar objects than for objects that are very different.

For feature representations the above does not hold the other way around: two entirely different objects may have the same feature representation. This does not cause a problem if these feature values are improbable for all or for all but one of the classes. For a dissimilarity representation, however, the reverse of the compactness hypothesis also holds if it can be assumed that some continuity is valid for the dissimilarity measure $d(r,s)$. Since $d(r,s) = 0$ if and only if r and s are identical, this implies that they belong to the same class. This can be extended somewhat by assuming that all objects r' for which $d(r,r') < \delta\,(\delta > 0)$, are so similar to r (if δ is sufficiently small) that they belong to the same class as r. Consequently, the dissimilarities of r and r' to all objects in the representation set should be about the same, i.e. $d(r,r_i) \approx d(r',r_i)$, with $r_i \in R$, by which their representations $\mathbf{d}$ and $\mathbf{d}'$ are also almost the same, i.e. $\mathbf{d} \approx \mathbf{d}'$. We conclude that for dissimilarity representations that satisfy the above continuity, the reverse of the compactness hypothesis holds: objects that are similar in their representation are also similar in reality and belong thereby to the same class.

A representation set R can now be judged as to be sufficiently large if an arbitrary new object of the same class is not totally different from all given objects. This can be expected if R already contains many objects that are very similar, i.e. that they have a small dissimilarity to at least one other object.

All the below statistics are based in one way or another on this observation. They will be illustrated by an artificial example based on an Euclidean distance matrix between n Gaussian distributed points in a k-dimensional space. We will vary both, n and k, between 5 and 500. If $n < k$ the points are in an n-1 dimensional subspace and we have certainly an undersampled, difficult problem. If $n \gg k$ the dataset may be judged as to be sufficiently sampled. Large values of k generate difficult (complex) problems as they demand large sample sizes n. The results we present on these problems are averages over 20 experiments, each time based on a new, randomly generated dataset.

2.1 Principal Component Analysis

A sufficiently large representation set will contain at least some objects that are very similar, i.e. their representations are very similar, which suggests that the rank of D will be less than its size, i.e. rank(D) $< n$. In practice, this will usually not be true exactly, since the objects are not completely similar. A more robust criterion will, therefore, be whether N_α, the number of eigenvectors of D for which the sum of the corresponding eigenvalues equals a fraction α (e.g. 0.99) of the total sum of eigenvalues, is small in comparison to n. As a criterion we propose:

$$J_{pca,\alpha} = N_\alpha/n \tag{1}$$

with N_α such that

$$\sum_{i}^{N_\alpha} \lambda_i \Big/ \sum_{i}^{n} \lambda_i = \alpha \tag{2}$$

In practice, there is usually no integer N_α for which (2) holds exactly. In our experiments, N_α, and thereby $J_{pca,\alpha}$ are therefore determined by interpolation. In fig. 1 the value of $J_{pca,0.99}$ is shown for the artificial Gaussian example, as a function of the size of the representation set for various dimensionalities k. From this graph it can be concluded that sets with more than 100 objects may be sufficiently large for small dimensionalities (e.g. $k = 5$ or 10), but that this is certainly too small for Gaussian datasets of larger dimensionality. These generate problems of a too high complexity for the given data size.
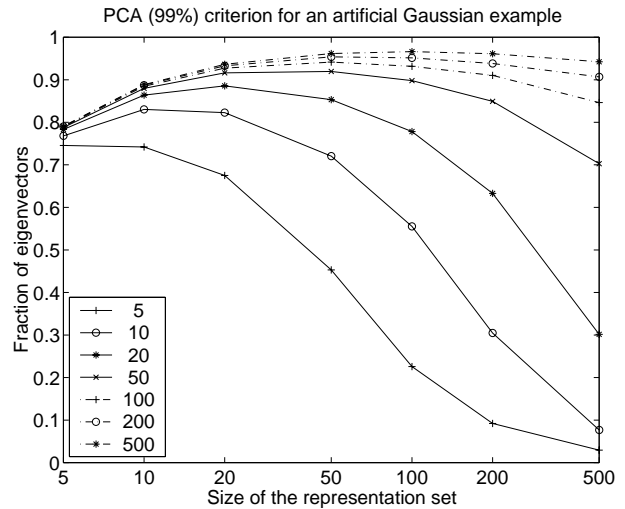

Fig. 1. The PCA (99%) criterion (1) for normal distribution based Euclidean distances for various sizes of the representation set and for various dimensionalities.

2.2 Skewness

A new object added to a set of objects that is still incomplete (not sufficiently sampled) will generate many large dissimilarities and just a few small ones. After it becomes saturated, however, for new objects there will be more and more very similar objects. As a result, the distribution of dissimilarities will peak for small values and show a long tail in the direction of large dissimilarities. Consequently, its skewness grows for increasing size of the representation set. The value to which it grows, however, is problem dependent. As a criterion we propose the skewness of the distribution of all single, non-diagonal values d in D:

$$E\left(\frac{d - E(d)}{\sqrt{E(d - E(d))^2}}\right)^3 \tag{3}$$

In fig. 2 the skewness of the artificial example is shown. For small representation sets this appears to be insufficient here, as can be concluded from the noisy behavior of the graphs in that area. For large representation sets the curves
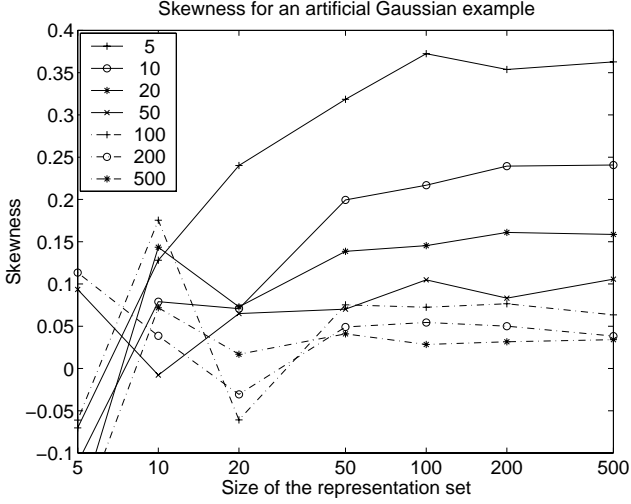
Fig. 2. The skewness criterion (3) for normal distribution based Euclidean distances for various sizes of the representation set and for various dimensionalities.

for different dimensionalities asymptotically increase to different values for the skewness. The final values may be reached earlier for the more simple problems in low dimensions ($k = 5$ or 10). This is, however, certainly not clearly observable. It has to be concluded that the skewness, as such, is not an informative, stable criterion.

2.3 Nearest Neighbor relationships

An element d_{ij} in D represents the dissimilarity between the objects r_i and r_j. The minimum of d_{ij} over all values of j, points to the nearest neighbor of r_i: $\mathrm{NN}(r_i) = r_p(i)$ if $p=$ argmin_j (d_{ij}). So, r_p is the most similar object to r_i in the representation set R. We now state that a representation $\mathbf{d}_i$ of object r_i is good if the representation of $r_p(i)$, i.e. $\mathbf{d}_p$ is close to $\mathbf{d}_i$ in the representation space. This can be measured by ordering the neighbors of $\mathbf{d}_i$ in the representation space (using the Euclidean distance) and determining the
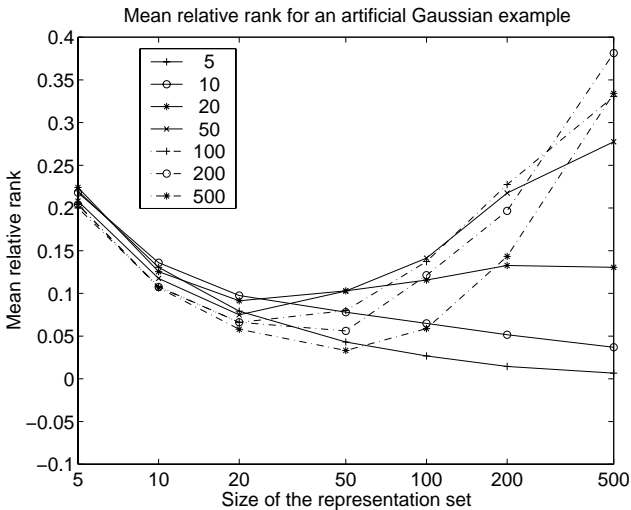

Fig. 3. The mean relative rank criterion (4) for normal distribution based Euclidean distances for various sizes of the representation set and for various dimensionalities.
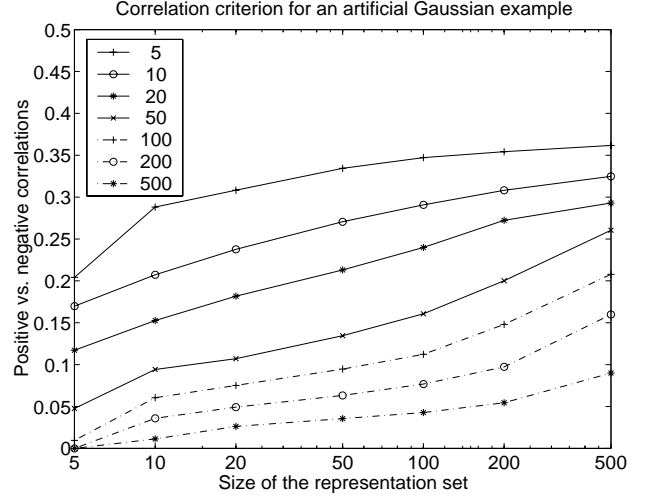
rank number m(i) of d_p in the list of neighbors or $\mathbf{d}_i$. For a good representation we expect that, the *mean relative rank*:

$$J_n = \frac{1}{n}\sum_{i}^{n} \mathrm{m}(i) - 1 \tag{4}$$

is close to 0. In fig. 3 the results for the artificial Gaussian example are shown. Like for the PCA criterion it can be concluded that sizes of the representation set larger than 100 are sufficient for distributions in 5 and 10 dimensions.

2.4 Correlations
We will also use the correlations between the objects in the representation space. Similar objects show similar dissimilarities to other objects and are thereby positively correlated. As a consequence, the average of positive correlations $\rho_+(\mathbf{d}_i, \mathbf{d}_j)$ to the average of absolute values of negative correlations $\rho_-(\mathbf{d}_i, \mathbf{d}_j)$:

$$J_\rho = \left(\frac{1}{N}\sum_{i,\,j\neq i}^{n} \rho_+(\mathbf{d}_i, \mathbf{d}_j)\right)\Big/\left(1 + \frac{1}{N}\sum_{i,\,j\neq i}^{n} |\rho_-(\mathbf{d}_i, \mathbf{d}_j)|\right) \tag{5}$$

where $N = 1/(n^2 - n)$, will increase for larger sample sizes. The constant 1 added in the denominator prevents J_ρ from becoming very large for the case of only small negative correlations. For a well-sampled representation set, J_ρ will be relatively large and it will increase only slightly when new objects are added (new objects should not significantly influence the averages of either positive or negative correlations). Fig. 4 shows that this criterion works well for the artificial Gaussian example: for less complex problems J_ρ reaches higher values and shows flatten behavior for the representation sets of at least 100 objects.

2.5 Intrinsic dimensionality
Another possibility to judge whether a representation set is sufficiently sampled is to estimate the intrinsic dimen-

sionality of the problem. Here, by the intrinsic dimensionality we understand the dimensionality of the underlying feature space found in such a way that the distances originally given are preserved. This can be achieved by a linear embedding, i.e. a distance preserving linear mapping, (provided that D is symmetric) onto an Euclidean or a pseudo-Euclidean space (see references [13] and [14] for details). The embedded data X is such that the squared (pseudo-) Euclidean distances are equal to the originally given squared dissimilarities, thus:

$$D^2_{(Ps)Eucl}(X, X) = D^2(R, R) \tag{6}$$

The representation X, consisting of $q <= n$ features, is determined to have uncorrelated features and be centered in the origin. If there are some features with large variances (i.e. taking large values), then the features with small variances are expected to reveal just noisy information [14]. Since only features with significant variances contribute to the distance values, the features with small variances can be skipped. (Note, that when all features have small variances, the intrinsic dimensionality is approximately n.) Let n_α be the number of features with significant variances for which the sum of the corresponding variance magnitudes equals a fraction α (e.g. 0.95) of the total sum. Of course, n_α may not be found exactly, so it is interpolated. Since n_α determines the intrinsic dimensionality, as a criterion we propose the following fraction:

$$J_{id, \alpha} = n_\alpha / n \tag{7}$$

For low intrinsic dimensionalities, smaller representation sets are needed to describe the data characteristics. Fig. 5 presents the behavior of our criterion as a function of the size of the representation set for various dimensionalities of the Gaussian data. The curves clearly reveal different intrinsic dimensionalities. If the representation set is sufficiently large, the fraction criterion should become relatively constant or decrease very slowly. We can then conclude that sets with more than 100 objects are satisfactorily sampled for originally low dimensionality, i.e. $k <= 20$. In other cases, the data is still too complex.

2.6 Compactness

As mentioned in the previous section, given a symmetric distance matrix D, a configuration X is found in a (pseudo) Euclidean space, such that the distances are preserved. When the representation set is sufficiently large, it is to be expected that the intrinsic dimensionality will remain constant. The constant intrinsic dimensionality can be only observed when the number of points is significantly large. Consequently, the centroid of the data should remain approximately the same and the average distance to this mean should decrease or be constant. The larger the average distance, the less compact the class is, requiring more samples for its description. Therefore, we propose a compactness criterion as the leave-one-out estimator of the average distance to the mean vector in an embedded space:
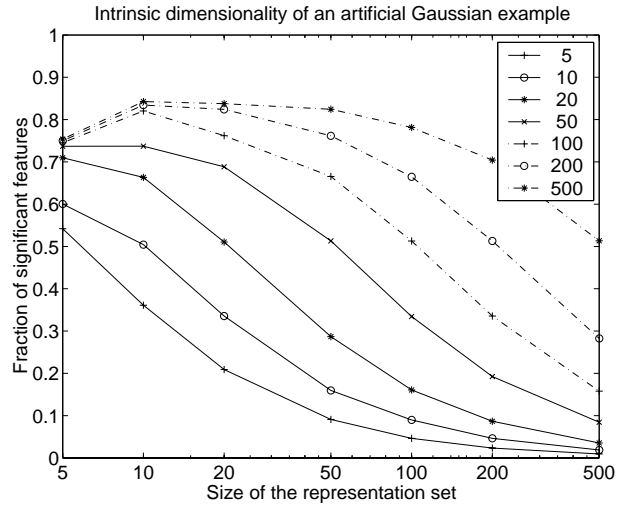

Fig. 5. The intrinsic dimensionality criterion (7) for normal distribution based Euclidean distances for various sizes of the representation set and for various dimensionalities.

$$C_{L\text{-}1\text{-}O} = \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{i \neq j} dist^2(x_i^{(j)}, m^{(j)}) \tag{8}$$

where $x_i^{(j)}$ is a vector representation of the i-th object in the (pseudo-) Euclidean space determined by all objects but the j-th object and $m^{(j)}$ is the mean of such a configuration. Fig. 6 shows the behavior of this criterion, clearly indicating the high compactness of originally low-dimensional Gaussian data. The case of $k=500$ is judged as a not very compact decription.
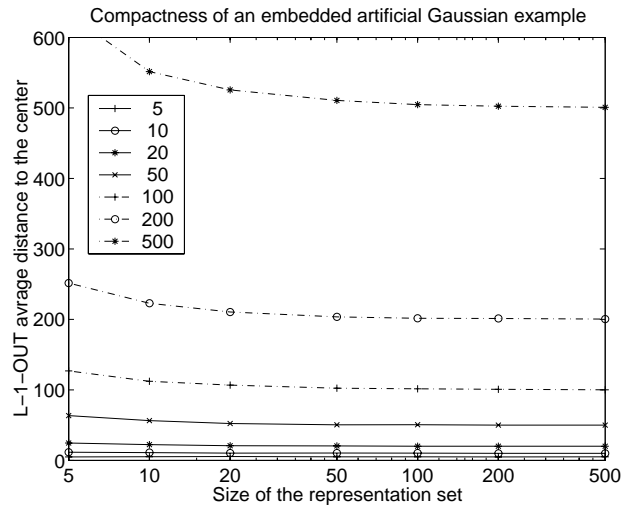

Fig. 6. The compactness criterion (8) for normal distribution based Euclidean distances for various sizes of the representation set and a for various dimensionalities.

3. EXPERIMENTS WITH REAL DATA

Representation sets can be applied for shape recognition. A training set of shapes may constitute a representation set for which an appropriate dissimilarity measure is used. New shapes are recognized on the basis of their dissimilar-

ity representation. Here, we will study the representation sets for 4 classes (0, 1, 2, 3) of handwritten digits from the NIST database [11]. For each class a set of $n = 200$ objects is taken. We used subsampled characters of the 128x128 size. As the dissimilarity measure the Modified Hausdorff Distance D_M [10] is used. We studied three variants of this distance measure: D_M^5, D_M and $D_M^{0.2}$. These power transformations do not change the order of the dissimilarities, but they change the representation space, and thereby the criterion values, in a non-linear way. This will give some impression to what extent the final definition of a distance measure may influence the resulting representation.

In the figures 7-13 the results are shown for the six criteria introduced in the previous section. The following observations can be made:

- The four characters '0'-'3' show slightly different behavior. In general, the set of '1's is the most simple one and the set of '3'-s is the most difficult one.

- It makes a significant difference if the dissimilarities are nonlinearly transformed by using some powers. Note that a power of 0.2 has some normalizing effect as it removes the tails of the distribution of dissimilarities, see fig. 14.

- The PCA criterion (1) indicates that the sample size for $D_M^{0.2}$ is far from being sufficient, it even shows some yet unexplained peaking phenomenon. For D_M the set of '1's sufficiently sampled according to (1) and for D_M^5 all four character sets are large enough, see fig. 7.

- The skewness criterion (3) is noisy and not very informative, see fig. 8.

- The mean relative rank (4) shows, fig. 9, that the $D_M^{0.2}$ set builds a good representation space in which distances correspond very well to the original dissimilarities. This can be explained by the linearizing effect of the small power $\ll 1$. At the same time the difference in complexity between the character sets has vanished.

- The strongly nonlinear D_M^5 set appears to be difficult according to the mean relative rank (5), see fig. 9.

- The correlation criterion (6) shows interesting results, see fig. 11. It indicates that the set of '1'-s is sufficiently sampled and that the other sets still need more observations.

- The intrinsic dimensionality (7) of the set of '1'-s is relatively low and much smaller than for other data sets; see fig. 12. The largest intrinsic dimensionality has the set of '0's. Remarkably, the dimensionality of the combined set seems to be relatively low, indicating that all classes share some descriptions.

- The set of '0'-s is the most compact class according to the criterion (8), see fig. 12. The sets of '1'-s and '2'-s are much less compact, indicating possible subclasses or elongated distributions. Not surprisingly, (8) judges the combined set of all characters as more complicated.

Finally, we performed some classification experiments in the representation space and compared these with the nearest neighbor classifier directly applied to the dissimilarities, see fig. 13. The results of this classifier do not depend
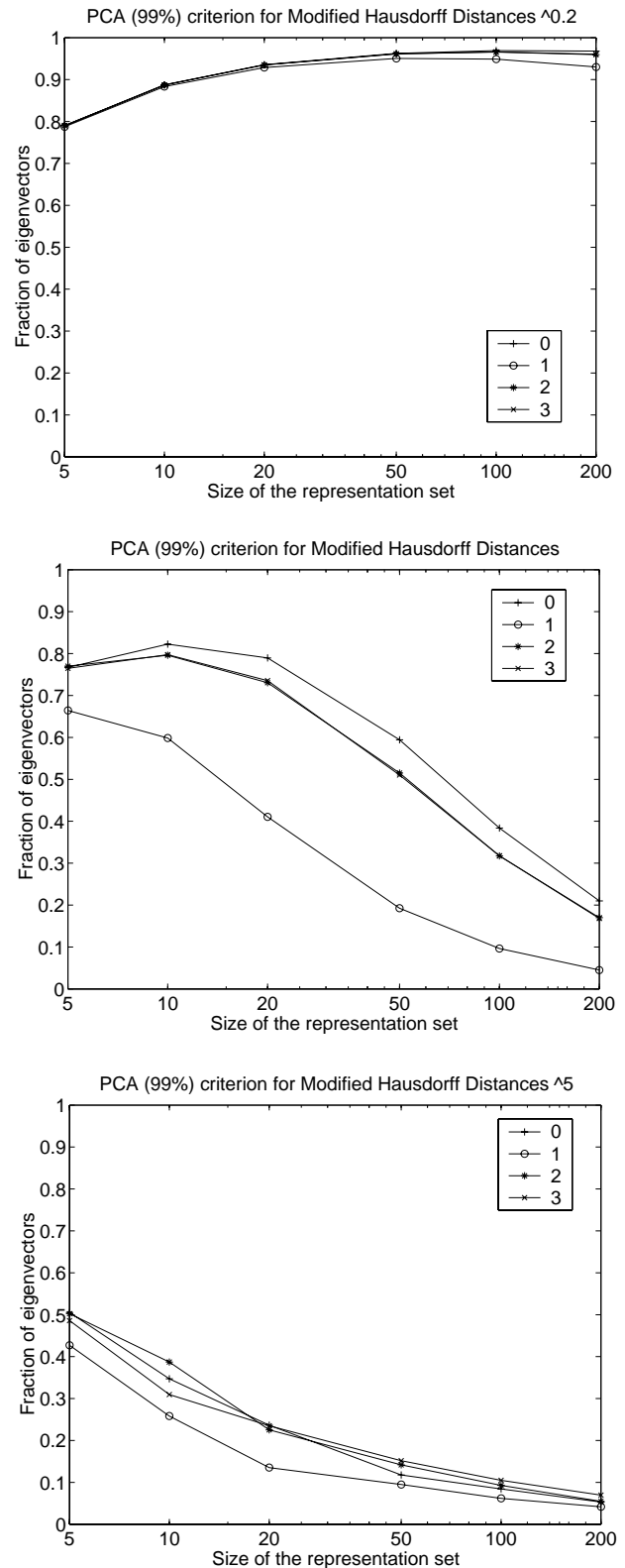

Fig. 7. The PCA criterion (1) as a function of the size of the representation set for four sets of handwritten digits represented by their Modified Hausdorff Distances D_M. From top to bottom three different transformations of the distances are used: $D_M^{0.2}$, D_M and D_M^5
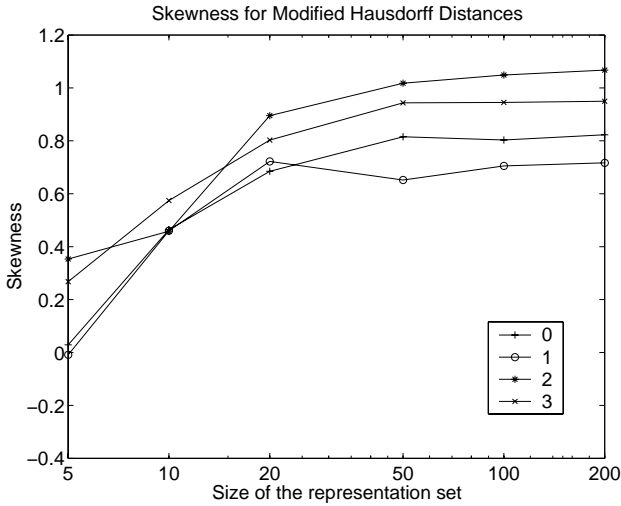
Fig. 8. The skewness criterion (3) as a function of the size of the representation set for four sets of handwritten digits represented by their Modified Hausdorff Distances D_M.

on the nonlinear transformations of the modified Hausdorff dissimilarities, as they are monotonic. The representation space was built using 50 objects per class. These objects, combined with another 50 objects (per class) were used for training in the representation space three classifiers: again a 1-NN classifier, a linear classifier assuming normal distributions with equal covariance matrices and a quadratic classifier based also on normal distributions. The independent test set consists of 100 objects per class. Results are averaged over 20 experiments, each time based on another randomly drawn distribution of the objects over the three sets used for representation, training and testing.

The results show that the representation space built by the $D_M^{0.2}$ set is the best for building classifiers. This corresponds with the observations made by the mean relative rank criterion. The normal based classifiers perform well, supporting the supposition that the transformation normalizes the distributions. Finally, it is remarkable that even the 1-NN rule in the representation space performs here better (for small sizes of the representation set) than the original 1-NN rule. Recall, however, that the latter just uses the representation set and no additional training objects.

4. DISCUSSION AND CONCLUSIONS

We considered six statistics that may be used for examining whether a representation set contains sufficient objects to describe the problem. This is still work in progress. The problem itself is ill-defined as it is application dependent what 'sufficient' is for a single class. For that reason we showed here some classification results for comparison. Still, one might imagine that classes are well sampled but in such a complicated way positioned with respect to each other that for most classifiers the problem is difficult. As a consequence the size of the training set should be judged from an evaluation of the classification result using a test
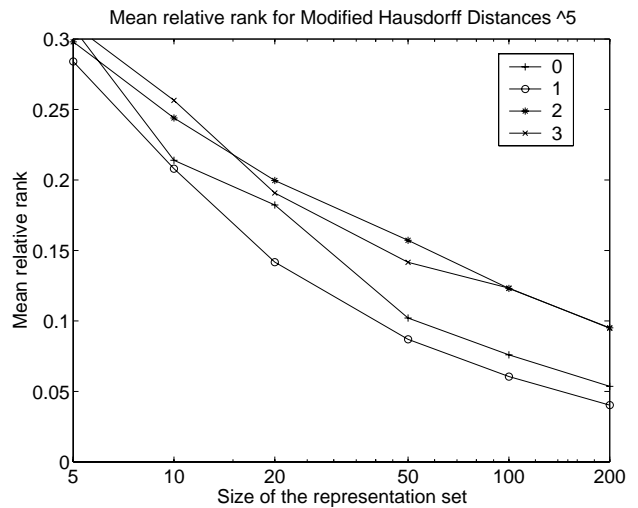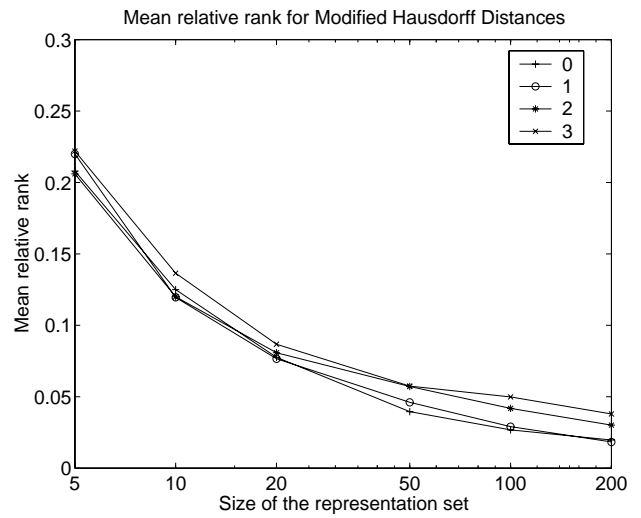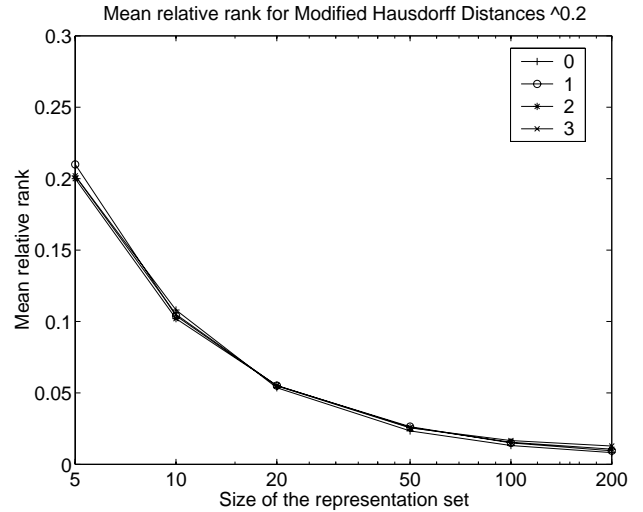

Fig. 9. The mean relative rank (3) as a function of the size of the representation set for four sets of handwritten digits represented by their Modified Hausdorff Distances D_M. From top to bottom three different transformations of the distances are used: $D_M^{0.2}$ D_M and D_M^5 .
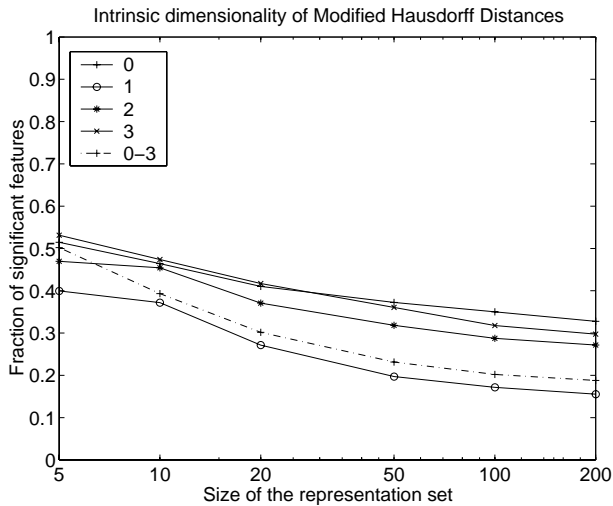
Fig. 10. The intrinsic dimensionality (7) as a function of the size of the representation set for four sets of handwritten digits represented by their Modified Hausdorff Distances. Also the results for the combined set are shown.
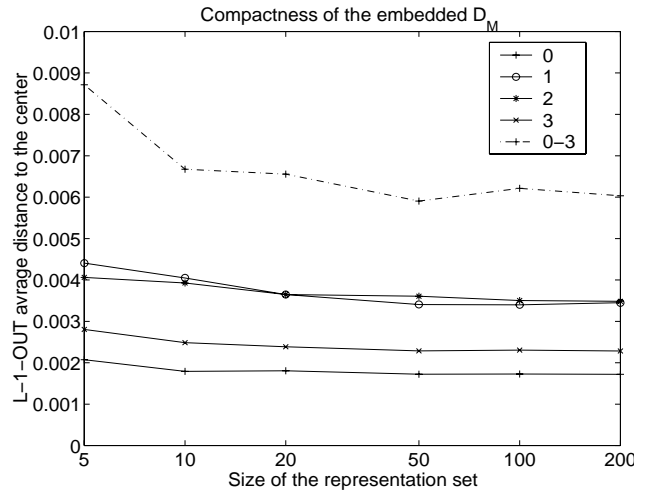

Fig. 12. The compactness criterion (8) for four sets of handwritten digits represented by their Modified Hausdorff Distances D_M. Also the results for the combined set are shown.
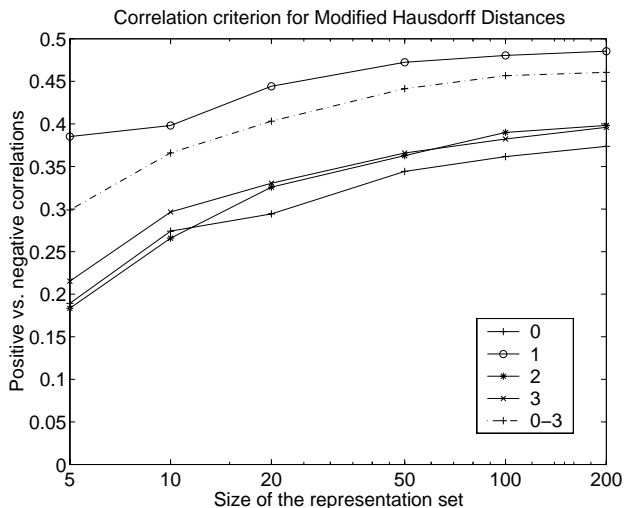

Fig. 11. The correlation criterion (5) as a function of the size of the representation set for four sets of handwritten digits represented by their Modified Hausdorff Distances. Also the results for the combined set are shown.
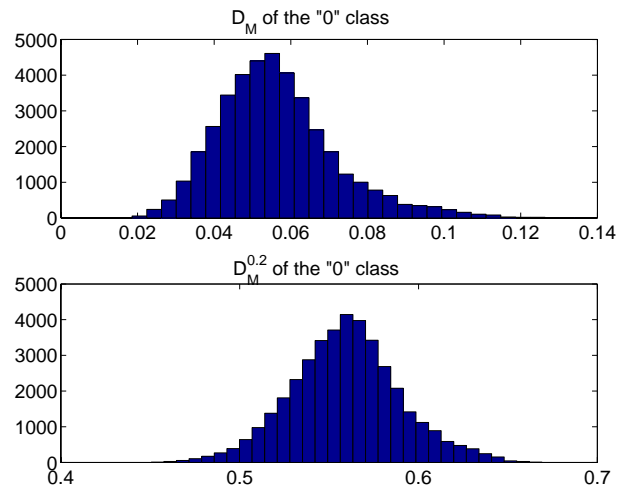

Fig. 14. Histograms for the set of handwritten '0' represented by their modified Hausdorff distances. Two different transformations are used: D_M (top) and $D_M^{0.2}$ (bottom).

set. The attempt of the presented study is to find out whether it is possible to judge from a single dissimilarity matrix its sampling density.

With respect to the criteria under study the following may be concluded. The PCA criterion works for distances computed for Gaussian distributed data. In the real world example in which a completely other dissimilarity measure was used, this criterion appeared to be much less useful.

The skewness is a noise sensitive statistic. Moreover, it did not appear to be very useful in our experiments. Still, we suspect that the distribution of the dissimilarity values may be indicative for the complexity of the problem in one way or another.

The nearest neighbor relationships on which the mean relative rank criterion is used appeared to be useful in both, the artificial problem, as well as for the real data.

Both the estimation of intrinsic dimensionality and compactness of the data description are found to be informative, especially when treated as complement information. They give an indication of the problem complexity.

The correlation criterion performed very well, both in the artificial problem as well as on the real data.

Our four favorite criteria, the mean relative rank, the correlation, intrinsic dimensionality and the compactness (explored together) should be tested further on artificial datasets and in real applications. Other criteria using label information may be considered as well in relation with classification problems.
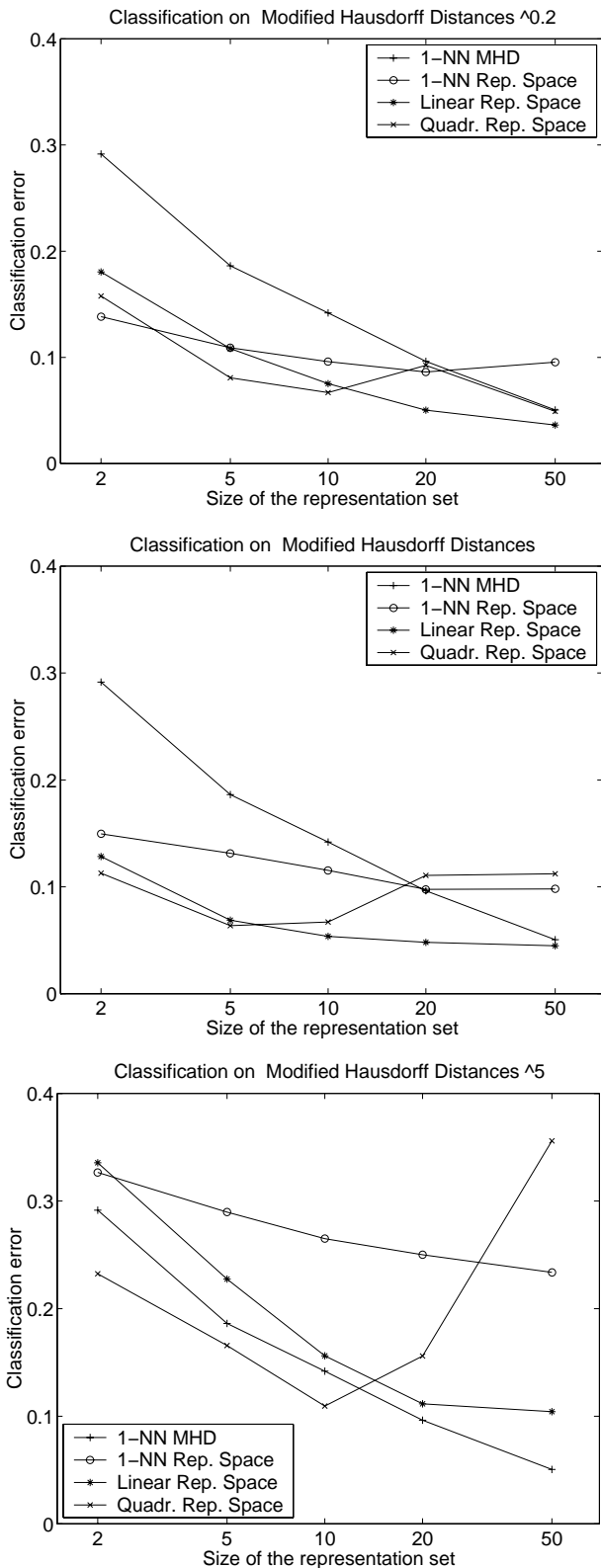
Fig. 13. Classification results as a function of the size of the representation set for four sets of handwritten digits represented by their Modified Hausdorff Distances D_M. From top to bottom three different transformations of the distances are used: $D_M^{0.2}$ D_M and D_M^5

REFERENCES

[1] R.P.W. Duin, Classifiers in Almost Empty Spaces, in: A. Sanfeliu, J.J. Villanueva, M. Vanrell, R. Alquezar, A.K. Jain, J. Kittler (eds.), *ICPR15, Proc. 15th Int. Conference on Pattern Recognition* (Barcelona, Spain, Sep.3-7), vol. 2, IEEE Computer Society Press, Los Alamitos, 2000, 1-7.

[2] E. Pekalska and R.P.W. Duin, Classifiers for dissimilarity-based pattern recognition, in: A. Sanfeliu, J.J. Villanueva, M. Vanrell, R. Alquezar, A.K. Jain, J. Kittler (eds.), *ICPR15, Proc. 15th Int. Conference on Pattern Recognition* (Barcelona, Spain, Sep.3-7), vol. 2, IEEE Computer Society Press, Los Alamitos, 2000, 12-16.

[3] R.P.W. Duin, D. de Ridder, and D.M.J. Tax, Experiments with object based discriminant functions; a featureless approach to pattern recognition, *Pattern Recognition Letters*, vol. 18, no. 11-13, 1997, 1159-1166.

[4] E. Pekalska and R.P.W. Duin, *Automatic pattern recognition by similarity representations*, Electronics Letters, vol. 37, no. 3, 2001, 159-160.

[5] R.P.W. Duin, Compactness and Complexity of Pattern Recognition Problems, in: C. Perneel (eds.), *Proc. Int. Symposium on Pattern Recognition "In Memoriam Pierre Devijver"* (Brussels, B, Feb.12), Royal Military Academy, Brussels, 1999, 124-128.

[6] P.J. Verveer and R.P.W. Duin, An evaluation of intrinsic dimensionality estimators, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 1, 1995, 81-86.

[7] A.K. Jain, Y. Zhong, and S. Lakshmanan, Object matching using deformable templates, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 3, 1996, 267-278.

[8] A.K. Jain and D. Zongker, Representation and recognition of handwritten digits using deformable templates, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 12, 1997, 1386-1391.

[9] A.G. Arkedev and E.M. Braverman, *Computers and Pattern Recognition*, Thompson, Washington, D.C., 1966.

[10] M.P. Dubuisson and A.K. Jain, A modified Hausdorff distance for object matching, *Proceedings, 12th IAPR International Conference on Pattern Recognition*, Conference A (Jerusalem, Israel, October 9-13, 1994), IEEE Computer Society Press, Los Alamitos, CA, 1994, 566-568.

[11] C.L. Wilson, M.D. Marris, *Handprinted character database 2*, april 1990. National Institute of Standards and Technology; Advanced Systems division.

[12] T.K. Ho, M. Basu, Measuring the complexity of classification problems, in: A. Sanfeliu, J.J. Villanueva, M. Vanrell, R. Alquezar, A.K. Jain, J. Kittler (eds.), *ICPR15, Proc. 15th Int. Conference on Pattern Recognition* (Barcelona, Spain, Sep.3-7), vol. 2, IEEE Computer Society Press, Los Alamitos, 43-47.

[13] I. Borg and P. Groenen, *Modern Multidimensional Scaling, Theory and applications.* 1997. Springer-Verlag.

[14] L. Goldfarb. Progress in Pattern Recognition, 1985, vol. 2, chapter 9: A New Approach to Pattern Recognition, 241-402. Elsevier Science Publishers.