On Combining Dissimilarity Representations

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
{ela,duin}@ph.tn.tudelft.nl

Abstract. For learning purposes, representations of real world objects can be built by using the concept of dissimilarity (distance). In such a case, an object is characterized in a relative way, i.e. by its dissimilarities to a set of the selected prototypes. Such dissimilarity representations are found to be more practical for some pattern recognition problems.

When experts cannot decide for a single dissimilarity measure, a number of them may be studied in parallel. We investigate two possibilities of combining either dissimilarity representations themselves or classifiers built on each of them separately. Our experiments conducted on a handwritten digit set demonstrate that when the dissimilarity representations are of different nature, a much better performance can be obtained by their combination than on individual representations.

1 Introduction

An alternative to the feature-based description is a representation based on dissimilarity relations between objects. In general, dissimilarities are built directly on raw or preprocessed measurements, e.g. based on template matching. The use of dissimilarities is especially of interest when features are difficult to obtain or when they have a little discriminative power. Such situations are encountered in practice when there is no straightforward manner to define features, when data is highly dimensional or when features consist of both, continuous and categorical measurements. The choice in favor of dissimilarity representations depends also on the application or the data itself. For instance, some particular characteristics of objects or measurements, like curves or shapes, may naturally lead to such representations, since they make recognition tasks more feasible.

To construct a decision rule on dissimilarities, the training set T of size n and the representation set R [2] of size r will be used. R consists of prototypes which are representatives of all classes present. In the learning process, a classifier is built on the $n \times r$ dissimilarity matrix $D(T, R)$, relating all training objects to all prototypes. The information on a set S of s new objects is provided in terms of their distances to R, i.e. as an $s \times r$ matrix $D(S, R)$.

A conventional way to discriminate between objects represented by dissimilarities is the nearest neighbor rule (NN) [1]. This method suffers, however, either from a potential loss of accuracy when a small set of prototypes is selected or

J. Kittler and F. Roli (Eds.): MCS 2001, LNCS 2096, pp. 359–368, 2001.

from its sensitivity to noise. To overcome these limitations, we have proposed an another approach. Our suggestion is to treat the dissimilarity representation $D\,(T, R)$ as a description of a space where each dimension corresponds to a distance to an object. $D\,(\boldsymbol{x}, R)$ can be, therefore, seen as a mapping of $\boldsymbol{x}$ onto an r-dimensional dissimilarity space. The advantage of such a representation is that any traditional decision rule operating on feature spaces may be used.

Most of the commonly-used dissimilarity measures, e.g. the Euclidean distance or the Hamming distance, are based on sums of differences between measurements. The choice of Bayesian classifiers [4], assuming normal distributions, is a natural consequence of the central limit theorem applied to them, when a large number of measurements is considered. The LNC (Linear Normal densities based Classifier) [4] is especially of interest because of its simplicity. Such a suggestion is strongly supported by our earlier experiments [7,8].

Selecting a good dissimilarity measure becomes an issue for the classification problem at hand. When considering a number of different possibilities, it may happen that there are no convincing arguments to prefer one measure over another. Therefore, the interesting question is whether combining dissimilarity representations might be beneficial. Two possibilities are here consider to study this problem. In the first one, the base classifiers (the LNC or the NN rule) are found on each dissimilarity representation separately and then combined into one decision rule. If the representations differ in character, a more powerful decision rule may be constructed by combining them. Secondly, instead of combining classifiers, representations are combined to create a new representation for which only one classifier has to be trained.

The paper is organized as follows. Section 2 gives some insight into the dissimilarity representations, classifiers and combining rules used. Section 3 describes the dataset and the experiments conducted. Results are discussed in section 4 and conclusions are summarized in section 5.

2 Combining Dissimilarity Representations

Assume that we are given the representation set R and p different dissimilarity representations $D^{(1)}(T, R)$, $D^{(2)}(T, R)$, ..., $D^{(p)}(T, R)$. Our idea is to combine good base classifiers, but on distinct representations. It is important to emphasize that the distance representations should have different character, otherwise they convey similar information and not much can be gained by their combination.

Two cases are here considered. In the first one, a single LNC is trained on each representation $D^{(i)}(T, R)$ separately and then all of them are combined in the end. In the second case, the NN rule is also included. The NN rule and the LNC differ in their decision-making process and their assignments. The NN method operates on dissimilarity information in a rank-based way, while the LNC approaches it in a feature-based way. Although the recognition accuracy of the NN method is often worse than of the LNC [8], still better results may be obtained when both types of classifiers are included in the combining procedure. Although many possibilities exist for combining classifiers [5], we limit ourselves to fixed rules operating on posterior probabilities. For the LNC, the posterior probabilities are based on normal density estimates, while for the NN method, they are estimated from distances to the nearest neighbor of each class [3].

Another approach to learning from many distinct dissimilarity representations is to combine them into a new one and then train e.g. a single LNC. As a result, a more powerful representation may be obtained, allowing for a better discrimination. The first method for creating a new representation relies on building an extended representation D_{ext}, in a matrix notation given by:

$$D_{ext}(T,R) = \left[D^{(1)}(T,R) \;\; D^{(2)}(T,R) \;\; \ldots \;\; D^{(p)}(T,R)\right] \tag{1}$$

It means that a single object is now characterized by pr dissimilarities coming from p various representations, but still computed to the same prototypes. The requirement of having the same prototypes is not crucial, however, for the sake of simplicity, the same representation sets are used here.

In the second method, all distances of different representations are first scaled so that they all take values in a similar range. Then, the final representation is created by computing their sum, as shown below:

$$D_{sum}(T,R) = \sum_{i=1}^{p} D^{(i)}_{max}(T,R), \tag{2}$$

where $D^{(i)}_{max}(T,R) = \alpha_i\, D^{(i)}(T,R)$ and α_i's scale all representations so that their maximum values become equal. (Note that now the representation sets should be identical to perform the sum operation.) The scaling procedure is necessary, otherwise the new representation will copy the character of a representation contributing the most to a sum, i.e. one with the largest distances. Scaling changes the orders of magnitude, but not the rankings, therefore all neighbor information is preserved. More sophisticated possibilities of scaling can be considered, as well, e.g. the weighted sum or the median from a sequence of dissimilarity values of different representations but relating a training object to the same prototype.

3 Dataset and Experiments

To illustrate our point, we investigate a 2-class classification problem between the NIST handwritten digits 3 and 8 [10]. The digits are represented as 128×128 binary images. Since no natural features arise from the application, constructing dissimilarities is an interesting possibility to deal with such a recognition problem. Three dissimilarity measures are considered: Hamming, modified-Hausdorff [6] and 'blurred', resulting in the representations: D_H, D_{MH} and D_B correspondingly. The Hamming distance counts the number of pixels which disagree. The modified-Hausdorff distance is found useful for template matching purposes [6]. It measures the difference between two sets (here two contours) $A = \{a_1, \ldots, a_g\}$ and $B = \{b_1, \ldots, b_h\}$ and is defined as $D_{MH}(A,B) = max(h_M\,(A,B), h_M\,(B,A))$, where $h_M\,(A,B) = \frac{1}{g}\sum_{a \in A} \min_{b \in B} ||a - b||$. To find

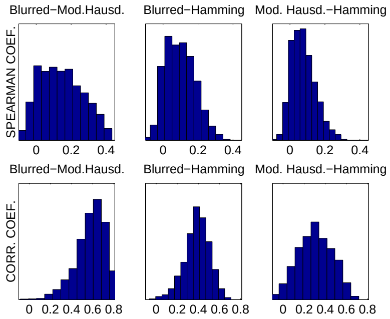


Fig. 1. Spearman coefficients (top) and traditional correlation coefficients (bottom) comparing dissimilarity representations.

D_B, images are first blurred with the Gaussian kernel and the standard deviation of 8 pixels. Then the Euclidean distance is computed between the blurred versions. The resulting distances are referred as to the 'blurred' distances.

Each of the distance measures uses the image information in a particular way: binary information, contours or blurring. From the process of the construction, it follows that our dissimilarity representations differ in properties. To prove, however, their different characteristics, the Spearman rank correlation coefficient is used to rank the distances computed to each prototype. Basically, we want to show that the rankings differ between representations. Therefore, for each pair of representations, the Spearman coefficients between the distance rankings to all prototypes are computed. Histograms of their distributions are presented in Fig. 1. All coefficients are between -0.05 and 0.4, where most of them are smaller than 0.2, which implies that the rankings differ significantly.

The traditional correlation coefficient is used to check whether the dissimilarity spaces of the individual representations (and, therefore, linear classifiers built there) are different. Such correlation values are higher than those given by the Spearman rates. It is to be expected, since now the exact distances are considered, which cannot completely vary from one representation to another, since the representations are descriptions of the same data and the same relations. On average, the correlations are found to be (see Fig. 1): 0.39 between the blurred and modified Hausdorff, 0.56 between the blurred and Hamming and 0.28 between the modified Hausdorff and Hamming. In the end, most coefficients are smaller than 0.6, thereby, they indicate only weak linear dependencies. Consequently, we can say that our dissimilarity representations differ in character.

The experiments are performed 25 times and the results are averaged. In a single experiment, the data, consisting of 1000 objects per class, is randomly split into two equally-sized sets: the design set L and the test set S. Both L and S contain 500 examples per class. The test set is kept constant, while L serves for obtaining the training sets T_1, T_2, T_3 and T_4 (being subsets of L) of the following sizes: 50, 100, 300 and $500\,(= L)$. For each training set, the experiments are conducted with varying size of the representation set R. Here, for simplicity, R is chosen to be a random subset of the training set.

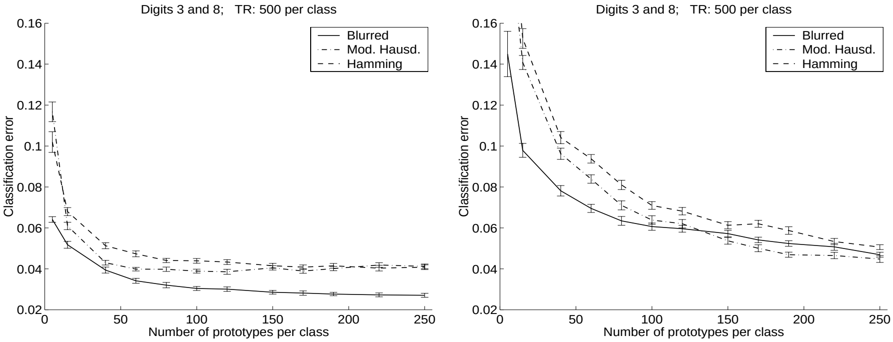


Fig. 2. Averaged classification error of the individual LNC's (left) and NN rules (right) as a function of the representation set size for the training set T_4.

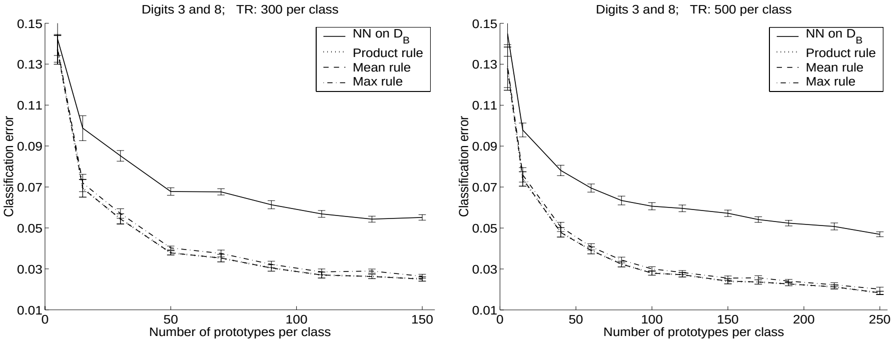
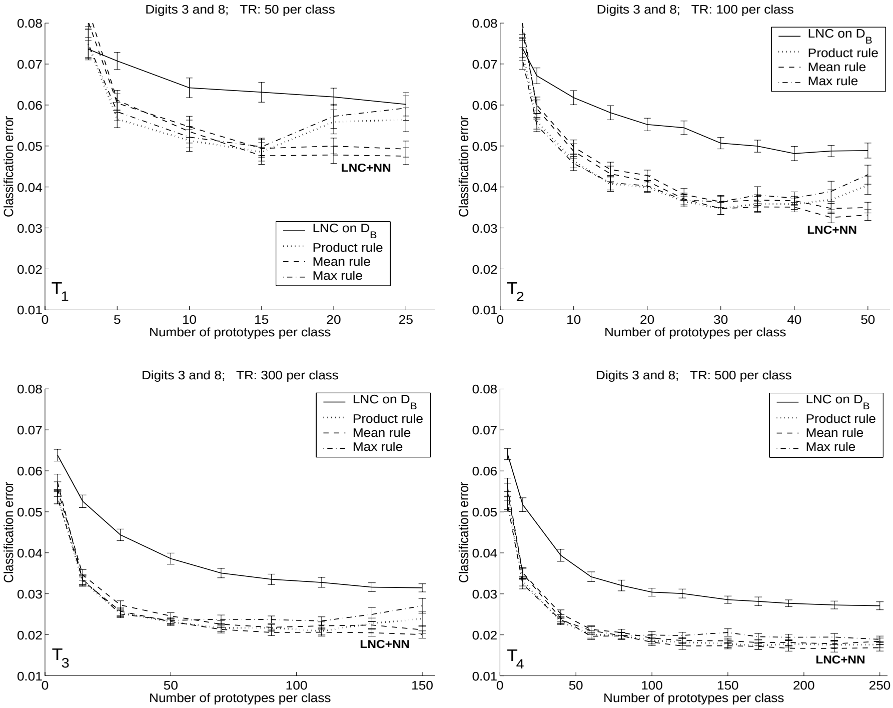


Fig. 3. Averaged classification error as a function of the representation set size for the individual NN rules trained on the sets T_3 (left) and T_4 (right).

4 Discussion

Considering single classifiers, it appears that the LNC consistently outperforms the NN rule for training sets: $T_1 - T_4$. Also, the LNC built on the blurred dissimilarities reaches a higher accuracy than for the other two representations. Since this behavior is repeated over all training sets, only the performance of the individual classifiers for the largest training set T_4 is presented in Fig. 2.

The results of combining either classifiers or representations for different training sets are presented in Fig. 3 – 6. These small, moderate and large training sets are considered to investigate the influence of the training size on our combining results. All plots in Fig. 3 – 6 show curves of averaged classification error (based on 25 runs) together with its standard deviation. Each error curve is a function of the representation set size, where the largest representation set considered is about half of the training set. Since our goal is to improve the performance of single classifiers by combining the information, all the results are presented with respect to the behavior of the LNC on the blurred representation D_B, as to the one that reaches the highest individual accuracy overall.


Digits 3 and 8; TR: 50 per class

Digits 3 and 8; TR: 100 per class

Digits 3 and 8; TR: 300 per class

Digits 3 and 8; TR: 500 per class

Fig. 4. Averaged classification error as a function of the representation set size for the individual LNC's combined by the product, mean or max rule or for both the LNC's and the NN methods combined by the mean operations.

Fig. 3 presents the generalization errors obtained for combining three individual NN methods by the mean, maximum and product rules. Operating on posterior probabilities is motivated by the intention of combining both the LNC and the NN method further on. Although the estimation of these probabilities is rather crude for the NN method, it still allows for an improvement of the combined rules. In all cases, the combination by the mean, max or product operation gives significantly better results than each individual NN rule. The larger, both training and representation sets, the more indicative gain in accuracy.

Fig. 4 shows the error curves obtained for three individual LNC's combined by the mean, maximum and product rules. For all training sets and small representation sets (in comparison to the training set size) considered, the product and maximum rules give slightly better results than the mean rule. However, for larger representation sets, the mean rule performs better. In addition, the error curve for the mean combiner of both the LNC and NN method is also shown. It can be observed, that incorporating the NN rule to the combiner, lowers somewhat the classification errors for larger representation sets. (This does not hold for small representation sets due to bad performance of each individual NN rule.)

Fig. 5 presents the error curves of a single LNC operating on new dissimilarity representations constructed from the three given: D_B, D_{MH} and D_H. Two

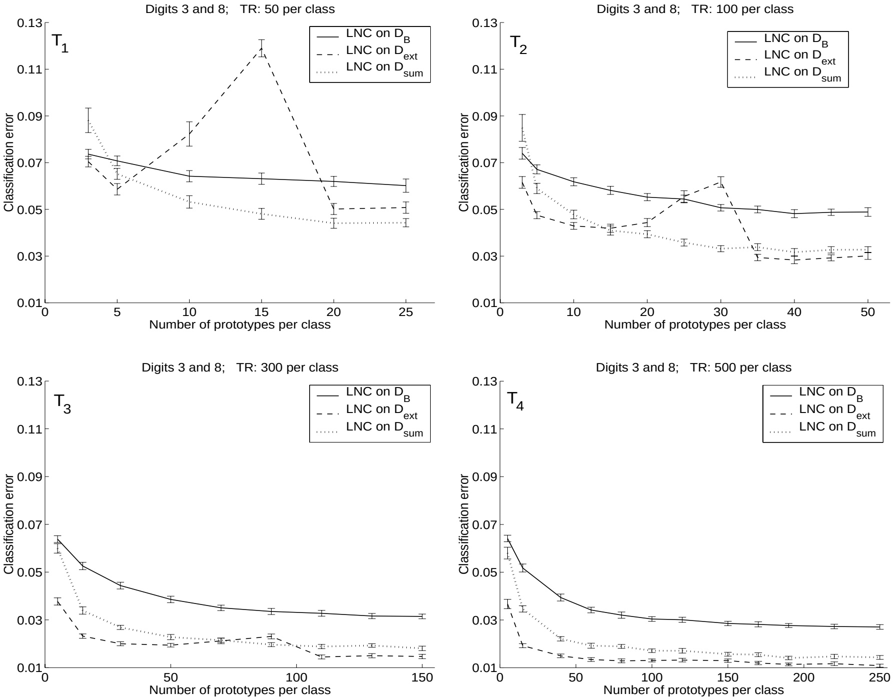


Fig. 5. Averaged classification error of the LNC as a function of the representation set size for the combined representations.

different cases are here considered: an extended representation D_{ext} (1) and the combined representation D_{sum} (2). The LNC on D_{sum} significantly outperforms the individual LNC's (it reaches higher accuracy than the best individual result on D_B), which is observed for all training sets. The LNC on D_{ext} can gain even better accuracy, however, the comparison between the representations D_{sum} and D_{ext} should be explained carefully. If the LNC is trained on D_{sum} using, say, r prototypes per class, then the representation D_{ext} is built from three such representations, each based on r prototypes, thereby the LNC operates in a $3r$-dimensional space. It means that for larger representations sets, the total number of dimensions exceeds the training size. The LNC is then not defined since the sample covariance matrix becomes singular and its inverse cannot be determined. In such cases, a fixed, relatively large regularization is used [4]. For moderate representation sizes (for which the dimensionality of D_{ext} approaches the number of training examples) the error curve of the LNC shows a peaking behavior (characteristic for this classifier). Therefore, worse performance is observed when number of prototypes is close to one third of the training size. For either small or larger representation sets, a very good performance is reached.

Fig. 6 presents the comparison between the mean combiner of individual classifiers and the LNC trained on the combined representation D_{sum}. For larger

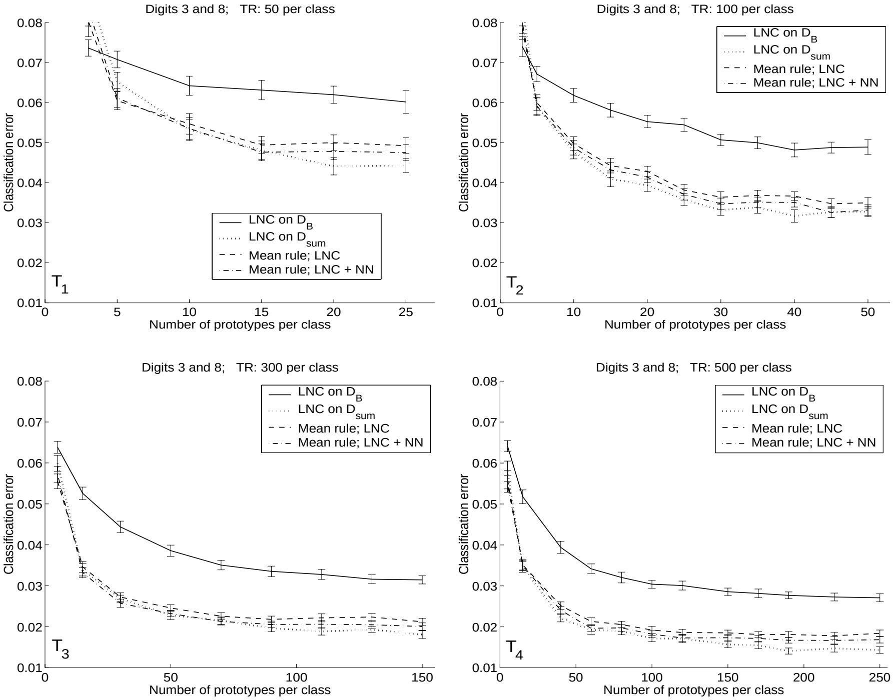


Fig. 6. Comparison between the accuracy of the combined classifiers found on each representation separately and one LNC on the combined representation D_{sum}. The classification error is as a function of the representation set size.

representation sets, the LNC trained on D_{sum} works somewhat better than the combined decision rule consisting of the LNC's and NN methods.

Summarizing, most of the combining rules perform significantly better than the individual classifiers. For small dissimilarity spaces, the representations tend to be independent and, therefore, the product rule based on the LNC's is expected to give better results than the mean rule [9] (here observed only slightly). For larger dissimilarity spaces, the posterior probabilities are not well estimated, and the product rule deteriorates; then the mean combiner is preferred. For the NN rule, the posterior probabilities are estimated from distances to the nearest neighbor and do not depend on the dimensionality of the problem. Therefore, both combiners perform about the same.

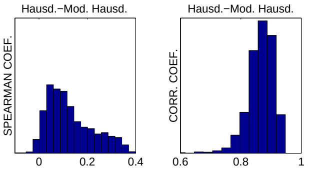


Fig. 8. Spearman and traditional correlation coefficients comparing D_{MH} and D_{HS}.

To illustrate the importance of dissimilarity representations of different nature, we present an example where the Hausdorff dissimilarity D_{HS} is used instead of the Hamming distance. Therefore, a triple $\{D_B, D_{MH}, D_{HS}\}$ is considered. The Hausdorff distance and the modified Hausdorff distance are similar,

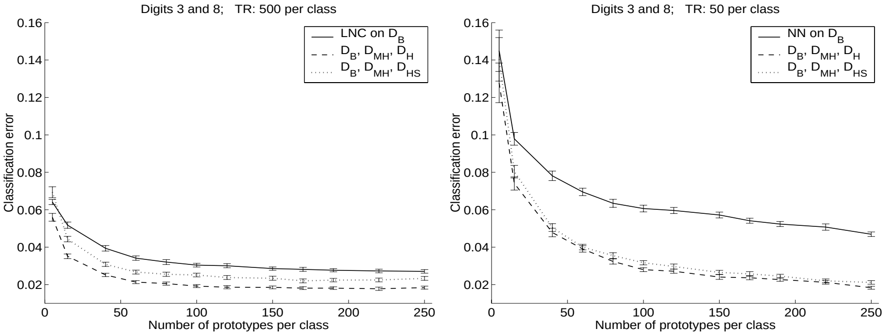


Fig. 7. Comparison of the classification error for the combined LNC's (left) and NN rules (right) and for two representation triples: $\{D_B, D_{MH}, D_H\}$ and $\{D_B, D_{MH}, D_{HS}\}$ and the training set T_4. Combining is done by applying the mean rule.

however, the latter violates the triangle inequality. Therefore, in the modified Hausdorff representation the dissimilarity rankings are changed with respect to the Hausdorff one. However, the dissimilarity spaces D_{MH} and D_{HS} are rather similar. In Fig. 8 histograms of both the Spearman and traditional correlation coefficients for these two representations are plotted. The Spearman values are similar to those obtained before (compare Fig. 1), but the traditional correlations become much higher, on average 0.91, indicating high dependence between those two dissimilarity spaces. It means that although by combining the individual NN rules for D_B, D_{MH} and D_{HS} an essential improvement may be gained, it does not necessarily hold for combining the LNC's. Fig. 7 presents the comparison between the performances of such classifiers combined by the mean rule for the training set T_4. It can be clearly observed that when D_{HS} is used instead of D_H, the performance of the combined LNC's deteriorates. Still, the combined NN rules are behaving only somewhat worse than for the triple $\{D_B, D_{MH}, D_H\}$.

When the Hausdorff representation is added to the original three, the performances of the combined individual classifiers or the LNC on D_{sum} are slightly better or not at all. The only significant improvement is observed for the extended representation D_{ext}.

5 Conclusions

Combining a number of distance representations may be of interest when there is no clear preference for a particular one. It can be beneficial when the dissimilarity representations emphasize different data characteristics. This is illustrated by a 2-class recognition problem between the digits 3 and 8 for three dissimilarity representations: Hamming, modified Hausdorff and blurred.

We have analyzed two possibilities of combining such information, either by combining classifiers or by combining representations themselves. In the first approach, individual classifiers are found for each representation separately and then they are combined into one rule. Our experiments show that the mean combining rule works well, especially for larger representation sets (with respect to the training size). In comparison to the best results of individual classifiers, the mean combiner based on three LNC's (built on each representation separately) or even better, the mean combiner based on three LNC's and three NN methods, performs significantly better.

In the second approach, dissimilarity representations are combined into a new one on which a single LNC is built. They are first scaled so that their maximal values are equal and then summed up, resulting in the representation D_{sum} (see (2)). We have also investigated scaling, e.g. by making the means identical or the maximum values for each prototype equal. They gave worse results and, therefore, are not reported here. The LNC on D_{sum} significantly improves the results of each individual LNC. It appears that the combined representation, built in this way, has a more discriminative power. As a reference, the extended representation D_{ext} is also considered (see (1)). The LNC on such a representation reaches even better results than on D_{sum}, provided that the number of all prototypes is either small or large in comparison with the training set size.

In conclusion, when dissimilarity representations differ in character, combining either individual classifiers or by creating a new representation can be beneficial. In our experiments, we have shown that when distinct representations are combined into D_{sum}, as a result, a representation which allows for a better discrimination can be obtained. This not only improves the classifier, but also it is of interest because of the computational aspect.

Acknowledgments. This work is supported by the Dutch Organization for Scientific Research (NWO).

References

1. R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, Inc., 1973.
2. R.P.W. Duin. Classifiers for dissimilarity-based pattern recognition. In *15th Int. Conf. on Pattern Recognition*, volume 2, pages 1–7, Barcelona (Spain), 2000.
3. R.P.W. Duin and D.M.J. Tax. Classifier conditional posterior probabilities. In *Advances in Pattern Recognition, Lecture Notes in Computer Science*, volume 1451, pages 611–619, Sydney, 1998. Proc. Joint IAPR Int. Workshops SSPR and SPR.
4. K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Acad. Press, 1990.
5. Duin, R.P.W. Kittler, J., Hatef M. and Matas, J. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.
6. 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.
7. E. Pekalska and R.P.W. Duin. Classifiers for dissimilarity-based pattern recognition. In *15th ICPR*, volume 2, pages 12–16, Barcelona, 2000.
8. E.Pekalska and R.P.W. Duin. Automatic pattern recognition by similarity representations. *Electronic Letters*, 37(3):159–160, 2001.
9. Duin, R.P.W. Tax, D.M.J. and Kittler, J. Combining multiple classifiers by averaging or by multiplying? *Pattern Recognition*, 33(9):1475–1485, 2000.
10. C.L. Wilson and M.D. Garris. Handprinted character database 3. Technical report, National Institute of Standards and Technology, February 1992.