Is combining useful for 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
email: ela@ph.tn.tudelft.nl

Keywords: dissimilarity representation, combining classifiers, combining rules, linear normal density based classifier, nearest neighbor method

Abstract

For learning purposes, representations of real world objects can be built by using the concept of dissimilarity. 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. Now the question arises how to make use of all the information given. 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 different in 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 (distance) 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 is a set of prototypes which are representatives of all classes present. Here, R is chosen to be a subset of T ($R \subseteq T$), although, in general, R and T might be disjunct. 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 from its sensitivity to noise. To overcome these limitations, we have proposed another classifier [7] constructed on the same representation, but being a weighted combination of dissimilarities.

In practice, 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 seen as a mapping of $\boldsymbol{x}$ onto an r-dimensional dissimilarity space (note that the dimensionality of such a space is determined *only* by the size of R). 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. The LNC (Linear Normal densities based Classifier) [4] is especially of interest because of its simplicity. Such a suggestion is strongly supported by our ear-

lier experiments [7, 8], which demonstrate the good performance of the LNC on distance representations.

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 problem that we want to address here is whether combining dissimilarity representations might be beneficial. To study this question, two approaches are considered.

First, the base classifiers are found on each dissimilarity representation separately and then combined into one decision rule. If the representations have different characters, the resulting classifiers differ in their assignments. By combining them, a more powerful decision rule may be constructed. 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 $D^{(1)}(T,R)$, $D^{(2)}(T,R)$, $\ldots$, $D^{(p)}(T,R)$. Our idea is to combine good base classifiers, but on distinct representations. It is important to emphasize here that the distance representations should have different character, otherwise they convey similar classification information and not much can be gained by their combination.

Two cases are here considered. In the first one, a single LNC is trained for each representation $D^{(i)}(T,R)$ separately and then all of them are combined into one decision rule. 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, therefore they differ in their decision-making process and their assignments. Although the recognition accuracy of the NN rule is often worse than of the LNC, still better results may be obtained when both types of classifiers are included in the combining procedure. So, all individual LNC's and NN rules form a set of base classifiers to be combined.

Many possibilities exist for combining classifiers [5]. Here, we limit ourselves to fixed rules which operate to posterior probabilities, e.g. mean or product. 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 dissimilarities is to combine all the representations into a new one and then train e.g. the LNC. As a result, a more powerful representation may be obtained, allowing for a better discrimination. Two methods for creating a new representation are studied here. The first method relies on building an extended representation D_{ext}, which in matrix notation is given as:

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

It means that a single object is now characterized by pr dissimilarities coming from various representations (each representation describes it by r distances), but still computed to the same prototypes. The requirement of having the same prototypes is not crucial at all. Different representation sets are allowed, but for the sake of simplicity, we keep them the same here. (Note that although the distances coming from various representations may have different orders of magnitude, scaling is not essential since the sample covariance matrix used for the construction of the LNC takes care of that.)

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

$$\begin{aligned} D^{(i)}_{max}(T,R) &= \alpha_i \, D^{(i)}(T,R), \quad i = 1, \ldots, p \\ D_{sum}(T,R) &= \sum_{i=1}^{p} D^{(i)}_{max}(T,R), \end{aligned} \qquad (2)$$

where 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. Also instead of adding distinct dissimilarity representations, more sophisticated possibilities can be considered. An example is taking the weighted sum, the maximum 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

The NIST handwritten digit set [10] is used in the experimental study. To illustrate our point,

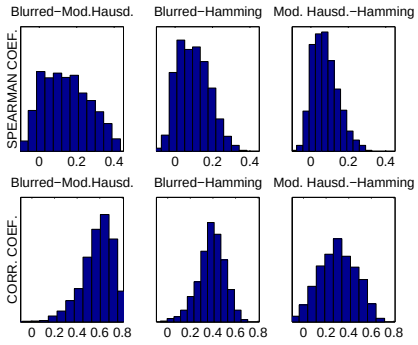


Figure 1: Spearman coefficients (top) and traditional correlation coefficients (bottom) comparing dissimilarity representations.

we investigate a 2-class classification problem between the digits 3 and 8. The digits are given as binary images with the resolution of 128×128 pixels. 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 three representations: D_H, D_{MH} and D_B correspondingly. The Hamming distance simply counts the number of pixels which disagree, i.e. have different binary values. The Hausdorff distance is often used for image comparison. Here, we will use a variant, namely the modified-Hausdorff distance since it is found more useful for template matching purposes [6]. We apply it on the contours of digits. The modified-Hausdorff distance measures the difference between two sets $A = \{a_1, \ldots, a_g\}$ and $B = \{b_1, \ldots, b_h\}$ (here two contours) 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 the last representation 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. We will refer to the resulting distances as to the 'blurred' distances.

Each of the distance measures uses the image information in a particular way, so from the process of their construction, it follows that our dissimilarity representations differ in properties. To prove, however, their different characteristics, the Spearman rank correlation coefficient r_S is used. For two rankings $\boldsymbol{p}^{(1)}$ and $\boldsymbol{p}^{(2)}$ it is defined as:

$$r_S = 1 - 6 \frac{\sum_{i=1}^{N} \left(\boldsymbol{p}_i^{(1)} - \boldsymbol{p}_i^{(2)} \right)^2}{N(N^2 - 1)}$$

Now we rank the distances computed to each prototype. Basically, we want to show that the rankings differ between representations., i.e. they are not linearly correlated. 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 the upper row of 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 significantly differ. Different rankings influence the variation in assignments of the NN rule the most. To check whether the dissimilarity spaces of the individual representations are different, the traditional correlation coefficient is used.

The traditional 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. Some coefficients are very small, some are larger. On average, the correlations are found to be (see bottom row of Fig. 1): 0.39 between the blurred and modified Hausdorff representations, 0.56 between the blurred and Hamming representations and 0.28 between the modified Hausdorff and Hamming representations. In the end, most coefficients are smaller than 0.6, thereby, they all indicate only weak linear dependencies. In summary, we can say that our dissimilarity representations differ in character.

The experiments are performed 25 times. 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.

4 Discussion

Considering single classifiers, it appears that the LNC consistently outperforms the NN rule for four training sets: $T_1 - T_4$. Also, in all cases, the LNC 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 are presented in Fig. 3 - 6. All figures show the same type of plots but for different training sets. These small, moderate and large training sets are considered in order to investigate the influence of the training size on our combining results.

All plots in Fig. 3 - 6 show curves of averaged clas-

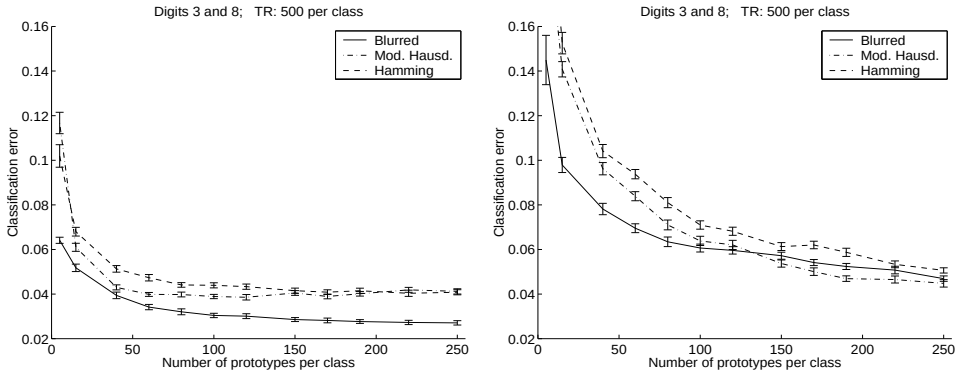
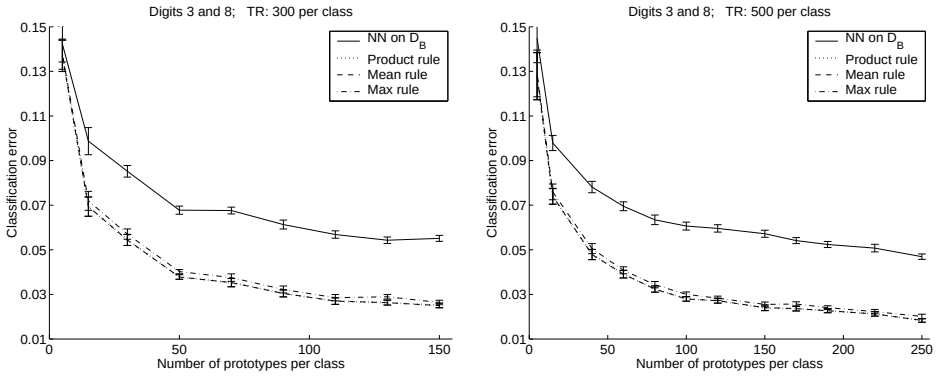


Figure 2: Averaged classification error of the individual LNC's (left) and NN rules (right) as a function of the representation set size for the training set T_4.



(a) Training set T_3 (b) Training set T_4

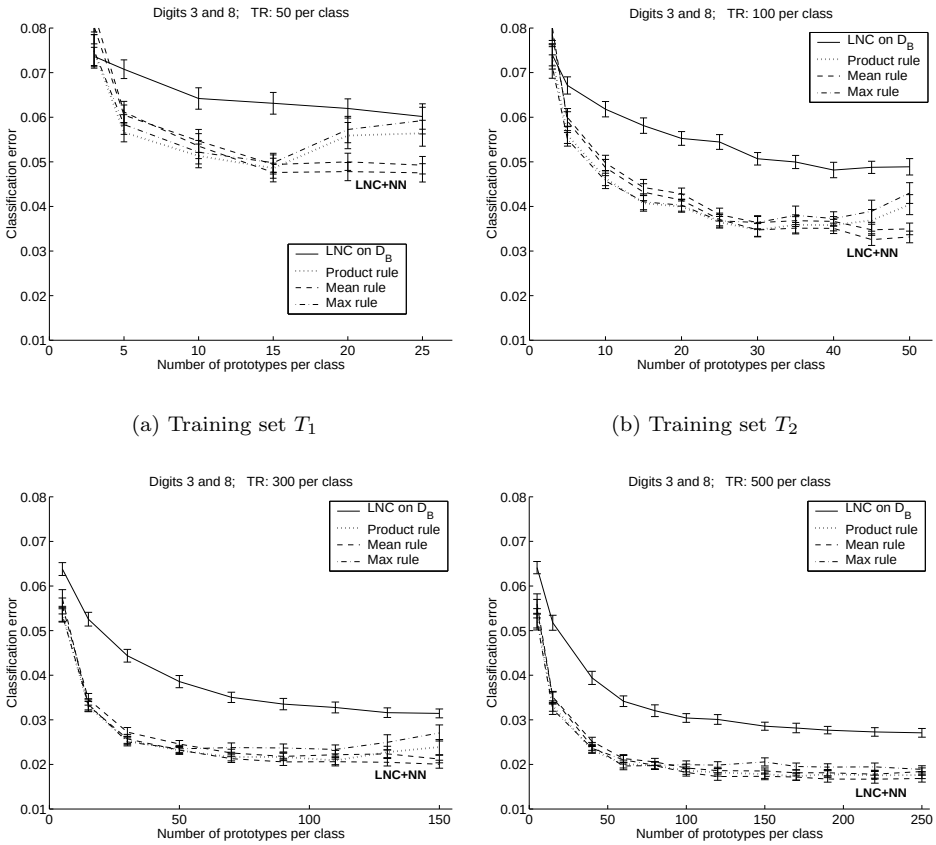
Figure 3: Averaged classification error as a function of the representation set size for the individual NN classifiers combined by the product, mean or max operation.

sification error (based on 25 runs) together with its standard deviation. Each error curve is a function of the representation set size (i.e. the number of prototypes). 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 in the relation to the behavior of the LNC on the blurred distance representation D_B, as to the one that reaches the highest individual accuracy overall.

Fig. 3 presents the generalization errors obtained for combining three individual NN methods, each found for one dissimilarity representation. The combining rules are the mean, maximum and product applied to posterior probabilities. Operating on posterior probabilities is motivated by the intention of combining both the LNC and NN rule 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 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 when three individual LNC's are combined on posterior probabilities by the mean, maximum and product rules. For all training sets, when small representation sets (in comparison to the training set size) are considered, the product and maximum rules give somewhat better results than the mean rule. However, for larger representation sets, the mean rule is better. In addition, the error curve for both the LNC and NN method combined for all representations by the mean rule is also shown. We can see that including the NN rule to the combining procedure, lowers somewhat the classification errors for larger representation sets (this does not happen 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 the combined dissimilarity representations constructed from the three given: D_B, D_{MH} and D_H. Two different cases are here considered: an



(a) Training set T_1

(b) Training set T_2

(c) Training set T_3

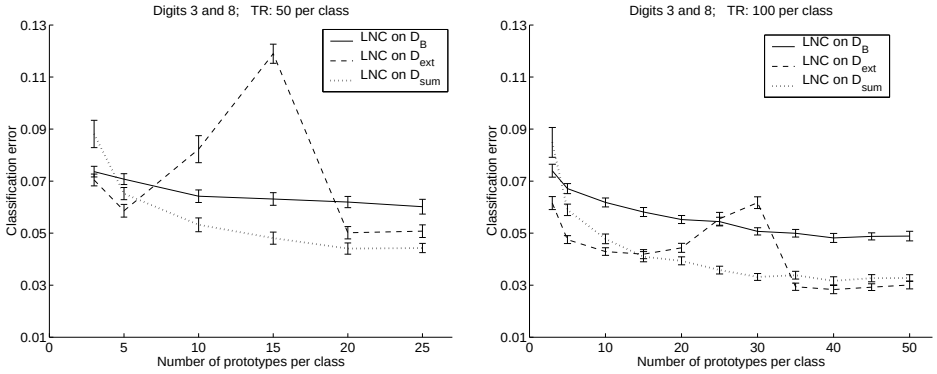
(d) Training set T_4

Figure 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 the LNC's and the NN methods combined by the mean.

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 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 (1%) 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.

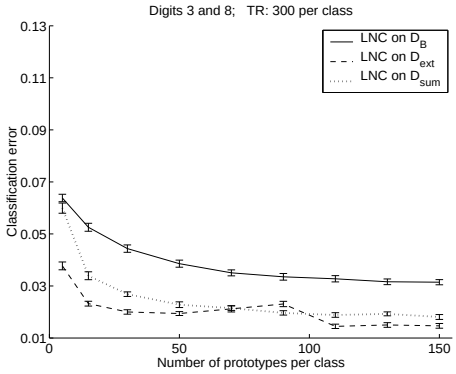
Fig. 6 presents the comparison between the combinations (by the mean rule) of individual classifiers found on different dissimilarity representations and the LNC trained on the combined representation D_{sum}. For all training set sizes, adding the NN rule to the process of combining classifiers improves the results for larger representation sets (this does not happen for small representation sets due to bad performance of each individual NN rule). For larger number of prototypes, the LNC trained on the representation D_{sum} works slightly better than the combined decision rule consisting of the LNC's and NN classifiers. It can be observed once again that adding the NN rule to the set of base classifiers improves the performance of the combiner.

Summarizing, the accuracy of individual classifiers and combined classifiers increases with the increasing size of both the training set and representation set. For all training sets, most of the combining procedures perform significantly better than the in-

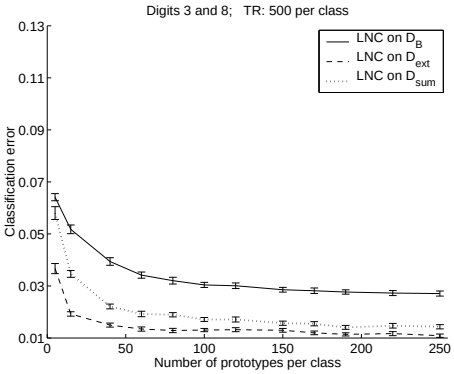


(a) Training set T_1

(b) Training set T_2



(c) Training set T_3



(d) Training set T_4

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

dividual classifiers. For the LNC and smaller representation sets the product rule gives slightly better results than the mean rule, while for larger representation sets it behaves much worse, worse than the individual classifiers. However, for the combination of the NN methods, both combining operations give nearly the same results. This phenomenon can be explained as follows. For small dissimilarity spaces (i.e. small number of prototypes) such representations tend to be independent and therefore the product rule based on the LNC's is expected to give better results [9]. For larger dissimilarity spaces, the posterior probabilities (based on normal density estimates) 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 rules perform the same.

4.1 Other considerations

In order to illustrate the importance of dissimilarity representations of a different nature, we present

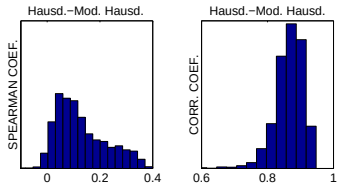
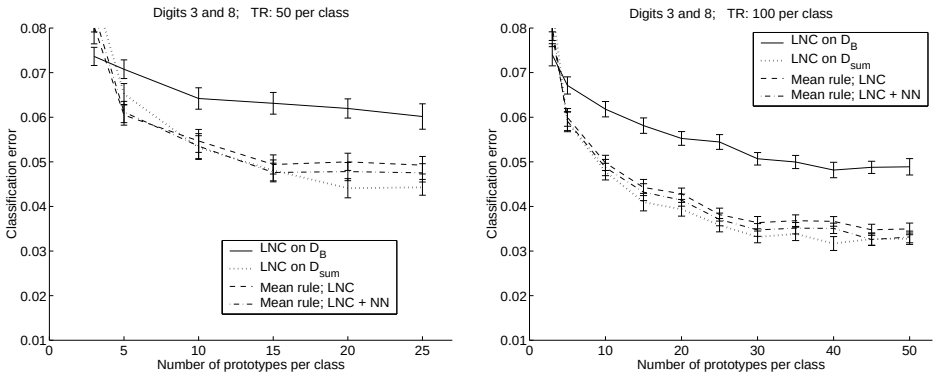
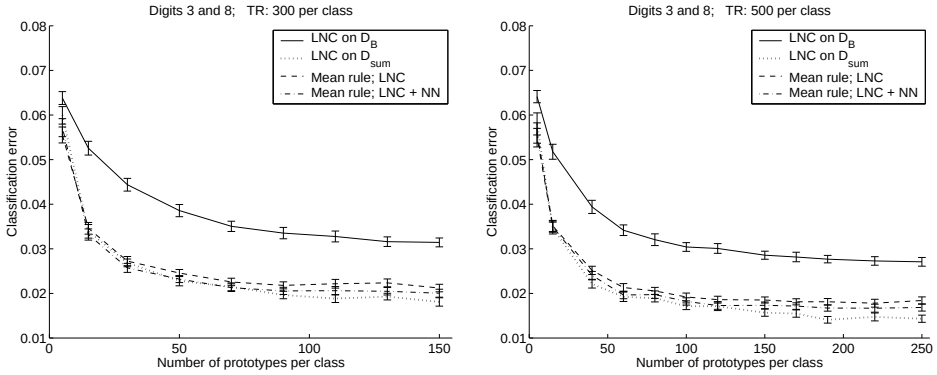


Figure 7: Histograms of Spearman and traditional correlation coefficients comparing D_{MH} and D_{HS}.

below an example where the Hausdorff dissimilarity D_{HS} is used instead of Hamming. Therefore, a triple $\{D_B, D_{MH}, D_{HS}\}$ is considered for experiments. The modified Hausdorff representation D_{MH} is only a modification of the Hausdorff distance. It changes the dissimilarity rankings (it is expected since the modified Hausdorff measure violates the triangle inequality), but the dissimilarity spaces are rather similar. In Fig. 7 histograms of both the Spearman and traditional correlation coefficients for these two representations are plotted. The Spearman values do not differ much from the same values for other pairs of representations considered before



(a) Training set T_1 (b) Training set T_2



(c) Training set T_3 (d) Training set T_4

Figure 6: Comparison between the accuracy of the combined classifiers found on each representation separately and a single LNC on the combined representation D_{sum}.

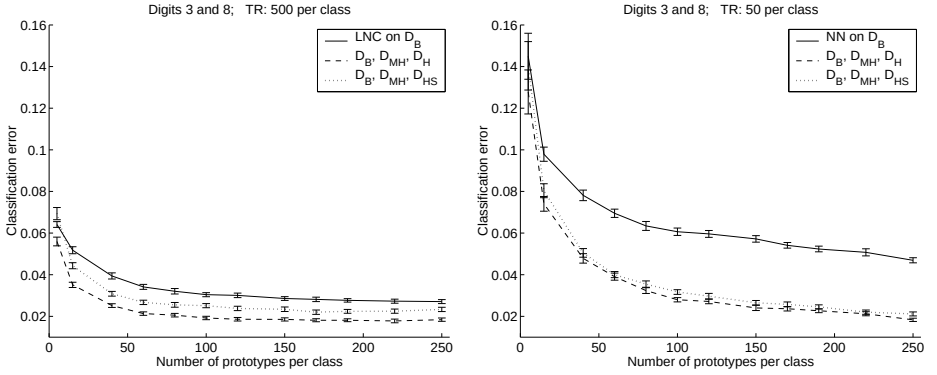


Figure 8: Comparison between the averaged classification error for the combined LNC's (left) and NN rules (right) and for two representation triples: $\{D_B, D_{MH}, D_H\}$ and $\{D_B, D_{MH}, D_{HS}\}$ and the training set T_4.

(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. 8 presents the comparison between the performances of such classifiers combined by the mean rule for two training sets: T_1 and T_4. It can be clearly observed that when D_{HS} is used instead of D_H, the combined LNC's performs much worse. For small training set T_1 it may even achieve a worse accuracy than that of the best individual LNC, i.e. the LNC on D_B. Still, the combined NN rules are behaving only somewhat worse than for the triple $\{D_B, D_{MH}, D_H\}$.

When the Hausdorff representation was added to the original three, the performances of the combined classifiers (now trained on four representations) or the LNC on D_{sum} were very slightly better or not at all. The only significant improvement was observed for the extended representation D_{ext}. Those two examples explain that having distinct representations is crucial for useful combining.

5 Conclusions

Using a number of distance representations may be of interest when there is no clear preference for a particular one. Here, one example of combining information from a few distinct representations is investigated, i.e. a 2-class problem between the handwritten digits 3 and 8 is studied for three dissimilarity representations: Hamming, modified Hausdorff and blurred.

We analyzed two possibilities of combining such representations, either by combining classifiers or by combining representations themselves. First, individual classifiers are found for all representations separately and then they are combined into one rule. Since dissimilarities can be approached in two different ways (feature-based and rank-based), two types of classifiers can be used: the LNC (linear one) and the NN rule. They differ in their construction and the decision process, therefore combining them may be of interest. In comparison to the best results achieved on dissimilarity representations, the mean combiner based on three LNC's (built on each representation separately) or the mean combiner based on three LNC's and three NN methods, perform significantly better.

In the second approach, dissimilarity representations are combined into a new one for which a single LNC can be applied. Our proposal is to scale them first so that their values lie in a similar range and then to sum them up, resulting in the representation D_{sum} (see (2)). Here, scaling is done by making the maximum values equal. We have also investigated another ways of scaling, like making the means identical or the maximum values for each prototype equal. They gave worse results and therefore are not reported in this paper. The LNC on D_{sum} significantly improves the results of each individual LNC. Combining representations in this way allows to get one, which has a more discriminative power.

As a reference, the extended representation D_{ext} is also considered. It is created from three representations so that each object is now characterized by $3r$ distances (see (1)). The LNC on such representation reaches even better results than on D_{sum}, provided that the number of all prototypes is either small or large in comparison to 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 showed 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 it is also of interest because of the computational aspect.

6 Acknowledgments

This work was partly 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 (Australia), 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. Pękalska and R.P.W. Duin. Classifiers for dissimilarity-based pattern recognition. In *15th Int. Conf. on Pattern Recognition*, volume 2, pages 12–16, Barcelona (Spain), 2000.

[8] E. Pękalska 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.