The dissimilarity representation, a basis for a domain-based pattern recognition?

Robert P.W. Duin, Elżbieta Pękalska, Pavel Paclík and David M.J. Tax

ICT group, Faculty of Electrical Engineering, Mathematics and Computer Sciences,
Delft University of Technology, The Netherlands

E-mail: {r.p.w.duin,e.pekalska,p.paclik,d.m.j.tax}@ewi.tudelft.nl

Abstract

Statistical pattern recognition traditionally relies on a feature representation. This approach can be powerful, if sufficient knowledge is available to select a small set of well-discriminating features. If there is a lack of such knowledge, a large set of possible features has to be collected and a large training set, representative in distribution for the given problem, is needed to build a reliable classifier. This is partially caused by the inherent difficulties in the feature based representation, when a (large) set of suboptimal features is used, in may result in a class overlap and strong feature dependency.

The dissimilarity representation aims at treating objects in their wholeness, avoiding the use of isolated features. If the dissimilarity measure is defined such that a zero value is only permitted for identical objects, class overlap may be avoided. Consequently, proper knowledge of class densities is not needed, which opens the possibility to a domain based classification in which the training set should be just representative for the domain of the classes.

In this paper, first, the basic ideas and some results of the dissimilarity representation are summarized. It is followed by a discussion on how this may be worked out for the domain based pattern recognition.

1. Issues of pattern recognition

In general, pattern recognition relies on the description of regularities in observations of classes of objects. *A class is a concept of a set of objects possessing similar characteristics or similar underlying factors.* This implies that the notion of 'similarity' is more fundamental than of a 'feature' or of a 'class', since it is the similarity which groups the objects together and, thereby, it should play a crucial role in the class constitution [9–11]. Such a proximity should possibly be modeled such that a class has an efficient and compact description.

Pattern recognition systems, aiming at the recognition of the patterns of real world objects and events, consist of two major steps [8]:

1. **Representation**. The individual objects have to be described such that their shared commonalities are extracted and encoded in some numerical, syntactical or symbolical ways. Since the objects cannot be handled in their totality (e.g. extracted features reduce the objects to a set of numbers), such a representation of objects is in fact a simplification. Yet, one hopes that it still captures the most prominent characteristics[1]. A suitable representation can be designed if a priori knowledge

[1]Note that in the process of deriving the final representation, some intermediate representations may naturally arise. For

of the problem or an application is incorporated there. An adequate representation can be optimized over a set of examples or/and a set of available tools. Also the possibility of comparing either pairs of objects or objects with already derived patterns of classes is significant for a good representation.

Examples of traditional representations are features and graphs (strings). The first, statistical, representations [2] are simple and easy to derive, but neglect the inter-relationship between the morphological subpatterns constructing the objects (if the original objects possess such a structure). The second, structural, representations [12] are more advanced, describe the morphological independences in the objects, but require a domain specific knowledge. It is also more difficult to relate the objects to each other.

2. **Learning and generalization**. In this step the representations of a set of examples (a training set) is transformed to a representation of the classes they belong to. This is the learning stage, where a dependence between the representations of objects and class memberships (class labels) is estimated. This is connected to an induction process. In the evaluation stage, referring to a deduction, new objects do not have to be related to all individual objects in the training set, but can be directly related to the derived concepts of classes. In a general form, however, the deduction step is simplified; it involves only the derivation of outcomes by following the principles defined in the learning stage.

   The object and class representations are not necessarily similar, but the generalization step should define the way in which they are related to each other. In the statistical approaches, a classifier (a discrimination function defined in a feature space) plays this role. There are many classifiers and there are many generalization procedures available [2]. Graphs are usually compared by user defined distance measures (that are sometimes optimized over the training set) in the nearest neighbor scenario. The generalization is also achieved by the identification, when syntactic grammars successfully parse the extracted sub-patterns of objects.

Additionally, an adaptation step can be distinguished in between the representation and the generalization [8]. This is based on a transformation of a given representation for the purpose of de-noising or enhancing some of its properties, or it is based on a modification of the tools used for the generalization. The goal is to optimize the overall performance of the system.

In applications, numerical features often come before the classifier or the generalization procedure is derived, so before the concept of proximity is taken into account. Using the notion of proximity (instead of features) as a primary concept renews the area of statistical learning in one of its foundations, i.e. the representation of objects [7, 13, 16, 18, 23, 24]. Proximity measures can capture both the statistical and structural information of patterns and, thereby, they form a natural bridge between these approaches. A number of researchers are conscious of the essential role that proximity plays for the class description [1, 9, 10, 13, 16, 17, 19, 23, 24]. Two main types of representations can be here considered: the ones which are fixed (specified measures) or optimized (over a set of parameters) and the ones which are learned. We will mostly focus on the former representations.

instance, one may start from raw measurements as collected by a CCD camera. These are further pre-processed to detect the shapes, then the contours are derived, which are finally encoded as Freeman strings.

The paper is organized as follows. In section 2, the dissimilarity representations and the dissimilarity measures are discussed. In section 3, basic learning approaches from such representations are summarized. In section 4, some classification examples are provided. Section 5 discusses further the possibilities of building domain-based classifiers. Section 6 is pertained to some issues of the learned dissimilarity representation. Section 7 presents a brief summary.

2. Dissimilarity representations

The first question (step 1 in section 1) concerns the representation of objects. We base it on a proximity. Proximity representations can be divided into *relative* and *conceptual* representations [28]. In the relative representation, pairs of objects are related and compared by measuring the proximity between them. Consequently, each object is described by a vector of proximities to other objects [7, 29, 31]. They may be defined on a feature-based representation, by using the distances between feature vectors, but also by the structural description distances (matching costs) between graphs or other structural models, or directly on the raw data, e.g. by similarities between shapes in images.

Proximity representations can be extended to depict a relation of one entity to a number of them or of a partial concept to the whole concept or just two concepts. Such representations are called *conceptual* [28]. Examples are a resemblance of a particular mug to a class of mugs, a similarity of a language to a group of European languages, a growth and development of a child to a model development or an image query serving the purpose of retrieving similar images, in a process of redefining the query. Also, in the statistical sense, given C classes, the posterior probabilities of an object x (or in fact its feature-based representation) form a similarity conceptual representation $[P(x|\text{class } 1), \ldots, P(x|\text{class C})]$. Conceptual representations also naturally appear in the clustering process [2], in one-class classification problems [34] or in the context of classifier combining techniques [21]. Such representations are often used only marginally, in the sense that they are derived and directly used for making the final decision. However, when e.g. a trained combining rule is applied to such a conceptual representation, it is further used for the learning purpose [4].

We will now focus on the relative representations. In our approach, we choose to model a proximity as a dissimilarity. This is not essential, since the same reasoning and methodology can be applied to similarity representations (after suitable adaptations). It is done to limit the scope of discussion. The dissimilarity representation that has been studying by us for a number of years [26, 28, 29, 31] describes each object by its dissimilarities to a set of prototype objects, called the representation set R. Each object x (typically not a feature vector) is represented by a vector of the dissimilarities $D(x, R) = [d(x, p_1), d(x, p_2), ..., d(x, p_n)]$ to the objects $p_i \in R$. This is a numerical dissimilarity representation that can be further used in statistical learning. Below some issues concerning the measure itself are discussed.

2.1. Dissimilarity measures

An important issue of dissimilarity representations refers to the characteristics of informative measures. For a robust real-world object description, a measure should incorporate the necessary invariance, like translation, rotation, scale and illumination invariance. It should be robust to the measurement noise. Dissimilarities may be derived directly from the raw data such as images, spectra, or time signals, or from an intermediate representation given by features or by a graph. The dissimilarity measure has to be

user-specified and it is a way for the expert to integrate his knowledge of the application.

In general, we do not require the strict metric properties. We assume that a non-negative dissimilarity obeys the reflectivity condition, $d(x, x) = 0$. The definiteness, i.e. the fact that the zero dissimilarity imposes the equivalence of objects, $d(x, y) = 0 \Rightarrow x = y$, is sometimes required, as well. It does not mean, however, that any dissimilarity measure is permitted. Basically, the measure should be meaningful for the task, i.e. possessing a discriminating power to distinguish the classes as required by the user. Essentially, the measure should also be such that the compactness hypothesis [3, 5] holds, i.e. objects which are similar in reality should be close in their representations. This means that a small variation of an object should impose only a small change of a dissimilarity value, hence the natural variation of objects of the same class should be captured in a compact description. Ideally, the dissimilarity should be continuous in the factors that influence the measurement process (such as illumination, scale etc).

When, for instance, measurement noise is present in the sensory data, it might be necessary to improve the resulting dissimilarity measures. The noise can be reduced either in the pre-processing stage of the raw data or, if the measures are just given or directly result from an earlier analysis, by the use of some (non-)linear transformations. Such transformations may also serve the purpose of imposing a more compact class description, e.g. by making large distances be relatively smaller, or (if required) by imposing particular characteristics of distances, e.g. by making them metric; see also [28].

Note that there does not exist a general object dissimilarity (proximity) that can be universally measured or applied. A comparison of two objects is always with respect to a particular point of view, a particular context, basic characteristics, type of domain, attributes considered, etc. This means that both the background information, the existence of other classes or additional a priori knowledge influence the way objects are compared. Since different dissimilarity measures, between graphs and on the raw sensor data may reflect various aspects of the data characteristics, as well as, various kinds of expert knowledge, their combination can be beneficial. They can be considered either jointly or exclusively, or they might form a new dissimilarity representation. The possibility of a combination makes a dissimilarity representation even a more universal representation due to the increased flexibility. Now, a complex problem can be described by a number of representations between their different aspects or characteristics. For instance, a text document in a database can be represented (in intermediate stages) as a point in a feature space, where each feature corresponds to the frequency of the specified keyword, but also as a tree organization of a title, an introduction, body, conclusions and references, etc. Next, two different dissimilarity measures can be designed in the statistical and structural approaches, yielding two distinct dissimilarity representations, which can be further combined. Combining dissimilarity measures (or their transformed versions) is closely related to the area of combining classifiers [21].

3. Classification approaches for dissimilarity representations

Given the representation of objects, the next fundamental question refers to the learning paradigms (step 2 in section 1), especially these which deal either with non-metric or non-Euclidean measures. Basically, they take place in suitable spaces.

Assume a representation set R of n objects. A training set T of m objects is represented as the $m \times n$ dissimilarity matrix $D(T, R)$. Usually $R \subseteq T$. Such dissimilarity representations are interpreted in three principal approaches: the *pretopological* approach, where the dissimilarity values themselves denote the neighborhoods, the *embedding* approach, which builds on an embedded pseudo-Euclidean configuration

and the *dissimilarity space* approach, where the the new 'features' are defined by the dissimilarities to the representation objects. These approaches are explained below; see also [28] for details.

Since dissimilarity representations are interpreted in some vector spaces, consequently, the tools available for the feature representation may be used for learning from the dissimilarity representation as well. If the dissimilarities are computed in an already given feature space, then this is an additional tool for finding nonlinear classifiers in the original feature space that has to compete with an arsenal of classifiers in statistical pattern recognition that already exists. If the dissimilarities are, however, based on structural information, a novel way of building a generalization is possible by the integration of structural object description with statistical learning methodologies. Although the learning mostly relies on the statistical approaches adapted for dissimilarity representations, the added value of such a dissimilarity-based framework lies indeed not directly in the following methodology, but in the representation itself.

As we have discussed in the previous section, the representation can include the statistical and structural characteristics of the data. Note that any dissimilarity representation can now be handled (although the embedding requires a symmetric dissimilarity measure, any asymmetric measure can be expressed as a summation of two symmetric measures as $d_1(x,y) = \frac{1}{2}[d(x,y) + d(y,x)]$ and $d_2(x,y) = \frac{1}{2}[d(x,y) - d(y,x)]$, which allows to construct two representations and combine them or the learning methodologies applied there appropriately).

3.1. Pretopological approach: the nearest neighbor strategy

The usual way of classifying a new object x represented by $D(x,R)$ is by using the nearest neighbor rule. The object x is classified into the class of its nearest neighbor, that is the class of the representation object p_i given by $d(x,p_i) = \min_{p_j \in R} D(x,R)$. Also the k-nearest neighbor approaches are possible provided that R is randomly drawn from the universe of objects. Note that the available information stored in $D(T,R)$ is not used by the classification procedure (in fact, there is no training). If $R = T$ is used, the classification is the most reliable as it is based on the entire training set. However, various approaches exist to reduce T to a smaller set R to speed up the classification. Under some circumstances, the recognition may even be improved by this reduction. In general, $D(T,R)$ can be interpreted as a space of objects $\mathcal{X}$ with neighborhoods defined as the ε-dissimilarity balls, i.e. $B_\varepsilon(x) = \{y \in \mathcal{X} : d(x,y) < \varepsilon\}$. This may serve for building a pretopological space [22, 28, 33].

3.2. Embedding approach

Given any symmetric dissimilarity representation $D(R,R)$, a configuration X can be found in some vector space such that the distances between the vectors of X reflect the original ones. In general, a Euclidean space is not capable to accommodate such a distance-preserving mapping, but a pseudo-Euclidean space is [13]. It is a $(p+q)$-dimensional non-degenerate indefinite inner product space $\mathcal{E} := \mathbb{R}^{(p,q)}$ such that the inner product $\langle \cdot, \cdot \rangle_\mathcal{E}$ is positive definite on $\mathbb{R}^p$ and negative definite on $\mathbb{R}^q$. Therefore, $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_\mathcal{E} = \sum_{i=1}^{q} x_i y_i - \sum_{i=p+1}^{p+q} x_i y_i = \boldsymbol{x}^T \mathcal{J}_{pq} \boldsymbol{y}$,, where $\mathcal{J}_{pq} = \mathrm{diag}\,(I_{p\times p}; -I_{q\times q})$ and I is the identity matrix. Consequently, $d^2_\mathcal{E}(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}-\boldsymbol{y}, \boldsymbol{x}-\boldsymbol{y} \rangle_\mathcal{E} = d^2_{\mathbb{R}^p}(\boldsymbol{x}, \boldsymbol{y}) - d^2_{\mathbb{R}^q}(\boldsymbol{x}, \boldsymbol{y})$. Since $\mathcal{E}$ is a linear space, many inner product based properties can be appropriately extended from the Euclidean case [13, 28].

The inner product (Gram) matrix G of the underlying configuration X can be expressed by using the square dissimilarities $D^{*2} = (d^2_{ij})$ as $G = -\frac{1}{2} J D^{*2} J$, where $J = I - \frac{1}{r}\mathbf{1}\mathbf{1}^T$ [13, 28, 31]. So, X is determined by the eigendecomposion of $G = Q\Lambda Q^T = Q|\Lambda|^{1/2}\mathrm{diag}(\mathcal{J}_{p'q'}; 0)\,|\Lambda|^{1/2}Q^T$, where $|\Lambda|$ is a diagonal

matrix of first decreasing p' positive eigenvalues, then decreasing magnitudes of q' negative eigenvalues, followed by zeros. Q is a matrix of the corresponding eigenvectors. X is then uncorrelated [13, 31] and represented in $\mathbb{R}^k$, $k = p' + q'$, as $X = Q_k |\Lambda_k|^{1/2}$. Since only some eigenvalues are significantly large (in magnitude), the remaining ones can be disregarded as non-informative. By their removal, the data are not only de-noised, but the curse of dimensionality is also avoided. So, the reduced representation $X_{red} = Q_m |\Lambda_m|^{1/2}$, $m = p + q < k$, is determined by the largest p positive and the smallest q negative eigenvalues. New objects $D(T_t, R)$ are orthogonally projected onto $\mathbb{R}^m$; see [13, 28, 31] for details.

For a chosen R, the linear mapping onto an m-dimensional space is determined from $D(R, R)$ as described above. The remaining objects $D(T \setminus R, R)$ are then projected to the space and all of them are used for training.

Inner product based classifiers can appropriately be redefined in a pseudo-Euclidean space. A linear classifier $f(\boldsymbol{x}) = \boldsymbol{v}^T \mathcal{J}_{pq} \boldsymbol{x} + v_0$ is e.g. constructed by addressing it as $f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + v_0$, where $\boldsymbol{w} = \mathcal{J}_{pq} \boldsymbol{v}$; see also [13, 28, 31].

3.3. Dissimilarity space approach

In a dissimilarity space, each dimension corresponds to a dissimilarity $D(\cdot, p_i)$ to an object p_i. Hence, the dimensions convey a homogeneous type of information. The property that dissimilarities should be small for similar objects (belonging to the same class) and large for distinct objects, gives a possibility for a discrimination. Thereby, $D(\cdot, p_i)$ can be interpreted as an attribute. This reasoning justifies the usage of traditional classifiers, e.g. linear ones, built in dissimilarity spaces. They can outperform the k-NN rule since they become more global in their decisions by making use of a larger training set T, while maintaining a small R. By using weighted combinations of dissimilarities, such classifiers suppress the influence of noisy examples [29, 31].

4. Examples of dissimilarity-based classification results

In this section, a few previously published results on the use of the proposed dissimilarity-based framework are briefly presented. Some other examples can be found e.g. in [5–7, 26–29, 31, 32]. The first case [6] considers a ten-class digit recognition problem studied by Jain and Zongker [20], where a similarity measure based on a deformable template approach has been developed. In total 2000 objects are available; 200 digits per class. The data are highly non-Euclidean and non-metric. A randomly selected set S consisting of 500 examples was used for testing and 1500 examples were used for the design set L. A representation set R is chosen at random from L for growing cardinalities until it becomes identical to L. The classifiers are built on $D(R, R)$ (hence $T = R$) and tested on $D(S, R)$. Figure 1 shows the generalization error as a function of $|R|$ for the following classifiers: a regularized linear normal density based classifier in a dissimilarity space (RNLC), a linear programming classifier in a dissimilarity space (LPC), a regularized quadratic normal density based classifier in a dissimilarity space (SRQC), a linear normal density based classifier in a 100-dimensional embedded space (NLC), the 3-NN, and the 5-NN rules (as the best within the k-NN rules); see also [28]. We used fixed, not optimized regularization parameters. From Figure 1 it can be observed that the classifiers constructed in the dissimilarity and embedded spaces achieve better results than the k-NN method.

The second example deals with the prototype selection issue. The selection of a representation set R

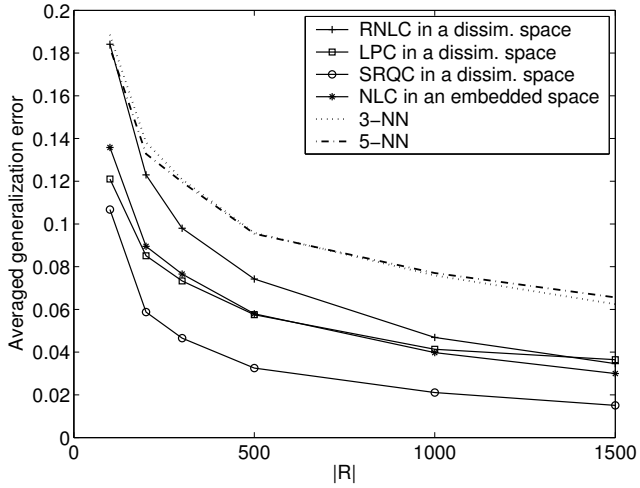


Figure 1. Classifiers in the dissimilarity and embedded spaces compared to the nearest neighbor rule for a ten-class digit classification problem.

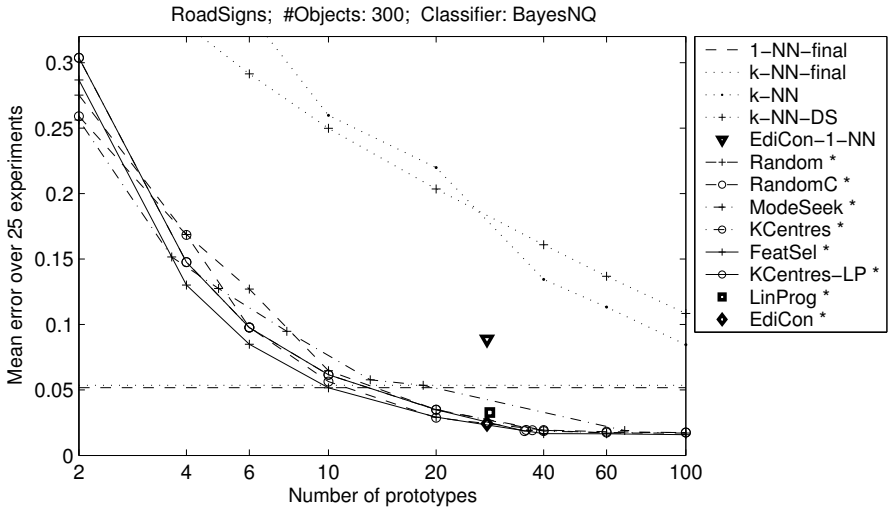
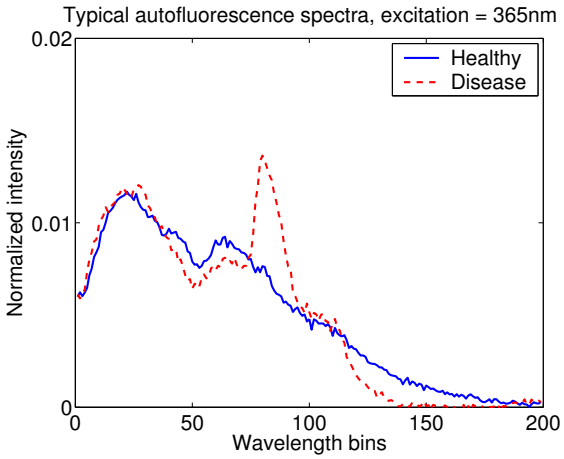


Figure 2. The averaged classification error of the BayesNQ* and the k-NN classifiers in dissimilarity spaces and the direct k-NN as a function of the selected prototypes.

out of a training set T for the construction of classifiers in a dissimilarity space serves a similar goal as the selection of prototypes to be used by the NN rule. Basically, the decision function should be of a moderate complexity and the computational effort for the evaluation of new incoming objects should be small. There is, however, an important difference with respect to the demands. Once selected, the set of prototypes defines the NN rule *independently* of the remaining part of the training set. This means that the remaining objects, i.e. objects from $T \setminus R$ are disregarded. On the other hand, the selection of the representation set in a dissimilarity space approach is less crucial, as it will define a dissimilarity space in which the entire training set T is still used for training of a classifier. For this reason, even a randomly selected representation set may work well.

In [30] several procedures are compared to select a suitable R of a specified cardinality. In one of the



Dissim. Rep.	AUC100 (st.dev.)	# SO
Single dissimilarity representations		
D_1	72.3 (0.7)	2.5
D_1^{der}	72.0 (0.7)	2.8
D_1^{2der}	78.2 (0.6)	2.7
D_{BH}	68.1 (0.8)	3.1
D_{geo}	75.1 (0.6)	2.1
Combined dissimilarity representations		
Mean	93.1 (0.5)	4.9
Prod	93.6 (0.4)	4.6
Min	85.0 (0.6)	15.3
Max	84.1 (0.9)	7.2

Figure 3. Typical examples of two auto-fluorescence spectra in the oral cavity (left) and the one-class classification results for the combined measures (right). SO stands for support objects.

experiments, the data describing a two-class problem of the discrimination between the road signs and non-signs have been used. 300 road sign images (highly multi-modal) and 300 non-road sign images acquired under general illumination are considered [25]. The latter images are identified by a sign detector using a circular template based on local edge orientations. Since the circular template was used to detect the boards, this a priori knowledge was used to remove the pixels in the background. Dissimilarities are derived from the normalized cross-correlations in an template matching fashion. In a repeated series of 25 experiments, 600 objects are randomly split into the training and testing sets of equal sizes. Several procedures are used to for the selection of R, of the specified sizes, for the construction of a dissimilarity space; see Figure 2. In this space, a normal density based quadratic classifier (BayesNQ) is trained. The averaged generalization errors are shown for the following selection procedures: random selection (Random), random selection with equal number of objects per class (RandomC), a mode seeking clustering procedure (ModeSeek), an equivalent formulation of the k-means procedure for the dissimilarities (KCentres), forward feature selection (FeatSel), a linear programming procedure based on the entire training set T (LinProg), a linear programming procedure minimizing the used dimensions based on the KCentres results (KCentres-LP), the same linear programming procedure and an editing-and-condensing procedure (EdiCon); see [30] for details. For comparison also the results of some nearest neighbor based classifiers are presented: the final results (based on the entire training set) of the 1-NN and k-NN rules, the k-NN result based on the representation sets selected by the RandomC procedure, the k-NN result in the dissimilarity space itself (KNN-DS) and the 1-NN error of the above mention EdiCon procedure.

Figure 2 shows that the dissimilarity based approach yields much better results than the traditional nearest neighbor rule. Moreover, for small representation sets of about 10 to 15 objects in total the classification performance is the same as for the nearest neighbor rule based on all 300 objects.

Another example is based on the one-class classifiers trained by a properly formulated linear programming description (LPDD) in the dissimilarity space [28, 32]. Since a sparse solution is offered, effectively, the description relies on the selected objects, called the support objects. The data consist of autofluorescence spectra acquired from healthy (target) and diseased (outlier) mucosa in the oral cavity; [36]. The measurements were taken at 11 different anatomical locations. Some typical examples are shown in Figure 3. In total, 856 spectra of the healthy people and 132 spectra of the diseased patients

are available. The classifier is trained on the healthy class only, using a random subset of 514 examples (60%) as the training set T. The remaining 40%, 337 examples are used for testing. Based on a test set, a ROC curve (a true positive ratio as a function of a false positive ratio) is derived. Since the costs of misclassifications are not known, a general criterion, the Area-Under the Curve (AUC) is used as a performance measure. The AUC equal to 1 stands for a perfect recognition of both healthy and diseased cases. This procedure is repeated 30 times and the results are averaged.

Five dissimilarity representations are used for the normalized spectra (to a unit area). The first three representations are based on the l_1 (city block) distances computed between the smoothed spectra themselves (D_1) and their first and the second order Gaussian-smoothed derivatives (D_1^{der} and D_1^{2der}, respectively) [26, 27]. D_{SAM} is based on the spectral angular mapper, $d_{SAM}(\boldsymbol{x}, \boldsymbol{y}) = \arccos(\boldsymbol{x}^T \boldsymbol{y})$. D_{BH} is based on the Bhattacharyya distance, a divergence measure between two probability distributions. This measure is applicable, since the normalized spectra, say, s_i, can be considered as unidimensional histogram-like distributions. They are constant on disjoint intervals $I_1, \ldots, I_N$, such that $s_i(x) = \sum_{z=1}^{N} h_z^i \mathcal{I}(x \in I_z)$, where $h_z^i \geq 0$. The Bhattacharyya distance is then: $d_{BH}(s_i, s_j) = -\log\left[\sum_{z=1}^{N} (h_z^i h_z^j)^{1/2} |I_z|\right]$. So, all the dissimilarity representations emphasize different aspects of the spectra.

New combined representations can be now derived from the others. The fixed combining rules we used are the same ones as often studied with respect to combining classifiers, but are used to combine the dissimilarities themselves, reflecting the analogy between the classifier outputs and the dissimilarities. Given K dissimilarity representations, the possibilities for deriving the combined representation are [28, 32]:

$$\begin{aligned}
D_{\text{avr}}(t_i, p_j) &= \tfrac{1}{K} \textstyle\sum_{\tau=1}^{K} \alpha_\tau \, D^{(\tau)}(t_i, p_j) \\
D_{\text{prod}}(t_i, p_j) &= \textstyle\sum_{\tau=1}^{K} \log\left(1 + \alpha_\tau \, D^{(\tau)}(t_i, p_j)\right) \\
D_{\text{min}}(t_i, p_j) &= \min_\tau \{\alpha_\tau \, D^{(\tau)}(t_i, p_j)\} \\
D_{\text{max}}(t_i, p_j) &= \max_\tau \{\alpha_\tau \, D^{(\tau)}(t_i, p_j)\}
\end{aligned}$$

The nonnegative weights α_τ may be additionally used to emphasize the importance of some measures. Ideally, they should be learned for the problem at hand. Here, they are assumed to be equal. In Figure 3, on the right, the detection results are summarized for the individual dissimilarity representations, as well as for some combined ones. Results show that the use of the combined representations improves the results significantly for all the studied combiners. The number of support objects, the actual size of the needed representation set, is surprisingly small. On average just a few spectra examples of the healthy people out of the original representation set of 514 spectra are needed for the computation of dissimilarities.

5. Towards domain-based classification

We will now discuss the issue that was initially touched in [6]. An inherent problem of the feature representation is that it is based on a reduction of the objects. Entirely different objects may yield the same feature vector. This is a cause of class overlap: if objects are represented by shape and color, then an apple and a pear may be mapped to the same point in feature space in this case. A minimum error classifier, thereby, needs to be based on the class distributions. Such distributions can only be

estimated from a training set of objects that is representative for the *distribution* of the universe of all objects. If the set of features is small, the class overlap may be large. If the set of features is enlarged, the estimation of the class densities becomes more difficult, demanding larger and larger training sets for proper estimations.

If the dissimilarity measure is zero if and only if the corresponding objects are identical, and when the real objects can be unambiguously labeled, then the class overlap may be avoided. The reasons is that no two objects in the representation set can have the same representation unless they are identical. This thinking opens a possibility of constructing the zero-error classifiers [6] in the dissimilarity representation, which should make use of the property of non-overlapping classes and define the decision function in the 'gap' between them. This is based on a set of assumptions:

1. The real, physical classes of objects are separable, i.e. there is no physical object that is a member of more than one class.

2. The raw measurements of the objects are made such that this separability is maintained[2].

3. The dissimilarity measure $d(x, y)$ between the objects x and y represented by their raw measurements (e.g. scanned images) is such that $d(x, x) = 0$ and $d(x, y) \geq \delta > 0$ if x and y belong to different classes. This assumption states that there exists some 'gap' between the classes of the δ-size: objects of different classes have a dissimilarity of at least δ.

4. The raw measurements of objects x and y are continuous functions of the parameters $\boldsymbol{\theta}$ that allow for their generation (e.g. lighting conditions, small rotations or sensor deviations). Hence, the dissimilarity $d(x(\boldsymbol{\theta}), y(\boldsymbol{\theta}))$ is continuous in $\boldsymbol{\theta}$. The noise is such that for any two measurements x and x' of the same physical object $d(x, x') < \delta$ holds.

5. The digitalization of the measurements and, thereby, the computer representation of the objects is such that the minimum class gap is preserved.

In fact, the existence of the gap implies that the 1-NN rule will constitute such a zero-error classifier. It may demand, however a very large training set. As the classifiers in the dissimilarity and embedded spaces appear to be much more efficient than the 1-NN by requiring a smaller number of prototypes for the construction, the question arises whether these classifiers may also have an asymptotic zero-error. We admit that technical difficulties may arise in practice, as it may demand a high sensor resolution to guarantee that all object differences are always registered and incorporated to the measure. Moreover, as the assumption initially just holds for the representation set, it has to be verified for additional training objects, as well as after a reduction of the representation set. This may introduce computational problems. The consequence of the fact that one deals with non-overlapping classes may be large, as the above sketched necessity to reach proper density estimates in feature spaces does not hold anymore. The sampling of the classes should now just be sufficient to find the separation boundary between the classes.

This yields the following research questions:

[2]One way to inspect this is to let the objects be labeled by humans based on the measurements (e.g. a video screen that displays the object image to be used for a further processing). The objects (e.g. characters) should be still labeled correctly after the scanning and display.

1. *How can it be tested whether the zero-error assumption is appropriate and not violated by sensor noise or by undersampling of the objects?*
2. *How can a given dissimilarity representation be improved, e.g. by combining with other dissimilarity representations, such that the zero-error assumption can be better sustained?*
3. *What are the proper classifiers for non-overlapping classes?*

   A possible classifier is the support vector machine with no slack variables. The problem is, however, far from simple as the choice for the appropriate kernel is not straightforward. Even if the classes do not overlap, they may show a virtual overlap if the classifier model does not match the shape of the true classification boundary. A solution may be found by studying a series of radial basis kernels with a decreasing width. For some width a zero error classifier should exist. In the experiments presented in [6], however, based on a comparable linear programming technique, no zero test results could be obtained for what was thought to be a simple artificial example.

   A second possibility to make use of the non-overlapping classes may be offered by the use of one-class classifiers, as classes may be treated independently. If density estimates are not needed, a domain based classification boundary may do [35]. An intrinsic problem here is, however, that a large representation set may be required to realize the non-overlap, but that the definition of a domain description in a high-dimensional space is not straightforward.
4. *What are techniques for one-class domain description?*

   A problem that exists in both approaches is the size of the representation set. The assumption of non-class overlap holds for sufficiently large representation sets. It is not clear when this point is reached. Moreover, depending on the classifier, not all objects are needed, but just specific ones. A selection technique is thereby essential to prevent the use of large sets.
5. *What are good prototype selection techniques for the construction of proper representation sets?*

   In the selection of prototypes for the construction of representation sets, it is important to take into account that many dissimilarity measures used in pattern recognition are non-Euclidean or even non-metric. Consequently, the use of metric properties is not possible. The use of dissimilarity spaces is not effected hereby, but the embedding approach results in pseudo-Euclidean spaces. The construction of classifiers in the dissimilarity space, however, does not use the fact that they deal with dissimilarities. They may show better results than the embedding approach an even better result may be reached if the dissimilarity character of the representation is preserved.
6. *How to make an optimal use of the dissimilarity character of the representation in case of non-metric dissimilarities?*

In [28] it is argued that the non-metric measures result in a pretopological space. By an appropriate closure operation a topology may be constructed that constitutes a base for generalization. Much work has still to be done in this direction.

6. Learned dissimilarity representations: the future

We realize that the developed framework for dissimilarity representations is only a first step in the direction of unifying both statistical and structural approaches, the problem of constructing an informative

representation and proper learning methodologies. For dissimilarity representations, the measure itself is assumed to be given. To some extent, it can be optimized with respect to a set of objects, but rather in a limited way, like the specification of some parameters. The next step is to investigate how dissimilarity measures can be *learned* from a set of examples. For this purpose, a *learned* representation can be considered, primarily based on the structure present in real objects. Some proposals in this direction have been made by Goldfarb and his colleagues; see e.g. [14–17].

Two possibilities can be now considered: to learn a relative representation or to learn a conceptual representation. The first focuses on defining a dissimilarity measure and a set of prototypes to which other objects will refer. Such a representation is used further for learning. The conceptual representation describes a dissimilarity of an object to a class. Such a dissimilarity is related to the costs (weights of transformations) of the object generation from a set of primitives (basic descriptors) in the presence of the objects within the class, as well as the ones outside the class. This is an attempt of a truly inductive way of learning [17], where not only the essential transformations and the weights are learned, but primitives as well. Such a formulation is close to the one-class classification [34, 35]. How to learn such measures is open for a future research.

7. Summary

We have introduced the dissimilarity representations as the ones that aim at treating objects in their totality. They allow to integrate both statistical and structural approaches by quantitative comparisons between the intermediate (structural) descriptions derived from (classes of) objects. The numerical dissimilarity representations describing each object by a set of dissimilarities to a selected prototypes is the basis for learning. These are the relative representations. Also, conceptual representations relating objects and concepts (classes, classifiers) may be used in the generalization step or for learning the measure.

Instead of using the traditional classifiers developed in statistical pattern recognition we may try to make use of the essential differences between dissimilarities and features (in terms of their different meanings and characteristics). This area has not been exploited, yet. It has some resemblance to the approach where a combined classifier is trained on the outputs of classifiers which in fact reflect (dis)similarities of an object to the specified classes.

In general, our results are encouraging to develop such representations and the related methodology further on.

Acknowledgments

This work was supported by the Dutch Technology Foundation (STW), grant RRN 5699, as well as the Dutch Organization for Scientific Research (NWO).

References

[1] H. Bunke, S. Günter, and X. Jiang. Towards bridging the gap between statistical and structural pattern recognition: Two new concepts in graph matching. In *International Conference on Advances in Pattern Recognition: Springer LNCS 2013*, pages 1–11, 2001.

[2] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2nd edition, 2001.

[3] R.P.W. Duin. Compactness and complexity of pattern recognition problems. In *International Symposium on Pattern Recognition 'In Memoriam Pierre Devijver'*, pages 124–128, Royal Military Academy, Brussels, 1999.
[4] R.P.W. Duin. The combining classifier: To train or not to train? In *International Conference on Pattern Recognition*, volume II, pages 765–770, Quebec City, Canada, 2002.
[5] R.P.W. Duin and E. Pękalska. Complexity of dissimilarity based pattern classes. In *Scandinavian Conference on Image Analysis*, Bergen, Norway, 2001.
[6] R.P.W. Duin and E. Pękalska. Possibilities of zero-error recognition by dissimilarity representations, an invated talks. In J.M. Inesta and L. Mico, editors, *Pattern Recognition in Information Systems*, Allicante, Spain, 2002.
[7] R.P.W. Duin, D. de Ridder, and D.M.J. Tax. Featureless pattern classification. *Kybernetika*, 34(4):399–404, 1998.
[8] R.P.W. Duin, F. Roli, and D. de Ridder. A note on core research issues for statistical pattern recognition. *Pattern Recognition Letters*, 23(4):493–499, 2002.
[9] S. Edelman. *Representation and Recognition in Vision*. MIT Press, Cambridge, 1999.
[10] S. Edelman, S. Cutzu, and S. Duvdevani-Bar. Representation is representation of similarities. *Behavioral and Brain Sciences*, 21:449–498, 1998.
[11] S. Edelman and S. Duvdevani-Bar. Similarity, connectionism, and the problem of representation in vision. *Neural Computation*, 9:701–720, 1997.
[12] K.S. Fu. *Syntactic Pattern Recognition and Applications*. Pretice-Hall, 1982.
[13] L. Goldfarb. A new approach to pattern recognition. In L.N. Kanal and A. Rosenfeld, editors, *Progress in Pattern Recognition*, volume 2, pages 241–402. Elsevier Science Publishers BV, 1985.
[14] L. Goldfarb. What is distance and why do we need the metric model for pattern learning? *Pattern Recognition*, 25(4):431–438, 1992.
[15] L. Goldfarb, D. Gay, O. Golubitsky, and D. Korkin. What is a structural representation? second version. Technical Report TR04-165, University of New Brunswick, Fredericton, Canada, 2004.
[16] L. Goldfarb and O. Golubitsky. What is a structural measurement process? Technical Report TR01-147, University of New Brunswick, Fredericton, Canada, 2001.
[17] L. Goldfarb, O. Golubitsky, and D. Korkin. What is a structural representation? Technical Report TR00-137, University of New Brunswick, Fredericton, Canada, 2000.
[18] A. Guérin-Dugué and G. Celeux. Discriminant analysis on dissimilarity data: A new fast gaussian like algorithm. In *Workshop on Artificial Intelligence and Statistics*, pages 635–644, Florida, USA, 2001.
[19] D.W. Jacobs, D. Weinshall, and Y. Gdalyahu. Classification with Non-Metric Distances: Image Retrieval and Class Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):583–600, 2000.
[20] A.K. Jain and D. Zongker. Representation and recognition of handwritten digits using deformable templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(12):1386–1391, 1997.
[21] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.
[22] F. Lebourgeois and H. Emptoz. Pretopological approach for supervised learning. In *International Conference on Pattern Recognition*, pages 256–260, Los Alamitos, CA, 1996. IAPR, IEEE Computer Society Press.
[23] V.V. Mottl, S.D. Dvoenko, O.S. Seredin, C.A. Kulikowski, and I.B. Muchnik. Featureless pattern recognition in an imaginary Hilbert space and its application to protein fold classification. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 322–336, Leipzig, 2001.
[24] V.V. Mottl, S.D. Dvoenko, O.S. Seredin, C.A. Kulikowski, and I.B. Muchnik. Featureless pattern recognition in an imaginary hilbert space. In R. Kasturi, D. Laurendeau, and C. Suen, editors, *International Conference on Pattern Recognition*, Quebec City, Canada, 2002.
[25] P. Paclík, J. Novovičová, P. Somol, and P. Pudil. Road sign classification using laplace kernel classifier. *Pattern Recognition Letters*, 21(13-14):1165–1173, 2000.
[26] P. Paclík and R.P.W. R.P.W. Duin. Classifying spectral data using relational representation. In *Spectral Imaging Workshop*, Graz, Austria, 2003.
[27] P. Paclík and R.P.W. R.P.W. Duin. Dissimilarity-based classification of spectra: computational issues. *Real Time Imaging*, 9(4):237–244, 2003.
[28] E. Pękalska. *Dissimilarity representations in pattern recognition. Concepts, theory and applications*. PhD thesis, Delft University of Technology, Delft, The Netherlands, January 2005.
[29] E. Pękalska and R.P.W. Duin. Dissimilarity representations allow for building good classifiers. *Pattern Recognition Letters*, 23(8):943–956, 2002.

[30] E. Pękalska, R.P.W. Duin, and P. Paclík. Prototype selection for dissimilarity-based classifiers. *Pattern Recognition*, submitted, 2004.
[31] E. Pękalska, P. Paclík, and R.P.W. Duin. A Generalized Kernel Approach to Dissimilarity Based Classification. *Journal of Machine Learning Research*, 2:175–211, 2001.
[32] E. Pękalska, M. Skurichina, and R.P.W. Duin. Combining Dissimilarity Representations in One-class Classifier Problems. In F. Roli, J. Kittler, and T. Windeatt, editors, *Multiple Classifier Systems, LNCS*, volume 3077, pages 122–133. Springer Verlag, 2004.
[33] B.M.R. Stadler, P.F. Stadler, G.P. Wagner, and W. Fontana. The topology of the possible: Formal spaces underlying patterns of evolutionary change. *Journal of Theoretical Biology*, 213(2):241–274, 2001.
[34] D.M.J. Tax. *One-class classification*. PhD thesis, Delft University of Technology, The Netherlands, 2001.
[35] D.M.J. Tax and R.P.W. Duin. Support vector data description. *Machine Learning*, 54(1):45–56, 2004.
[36] D.C.G. de Veld, M. Skurichina, M.J.H. Witjes, and et.al. Autofluorescence characteristics of healthy oral mucosa at different anatomical sites. *Lasers in Surgery and Medicine*, 23:367–376, 2003.