Structural Inference of Sensor-Based Measurements

Robert P.W. Duin[1] and Elżbieta Pękalska[1,2]

[1] ICT group, Faculty of Electr. Eng., Mathematics and Computer Science
Delft University of Technology, The Netherlands
r.duin@ieee.org, e.m.pekalska@tudelft.nl
[2] School of Computer Science, University of Manchester, United Kingdom

Abstract. Statistical inference of sensor-based measurements is intensively studied in pattern recognition. It is usually based on feature representations of the objects to be recognized. Such representations, however, neglect the object structure. Structural pattern recognition, on the contrary, focusses on encoding the object structure. As general procedures are still weakly developed, such object descriptions are often application dependent. This hampers the usage of a general learning approach.

This paper aims to summarize the problems and possibilities of general structural inference approaches for the family of sensor-based measurements: images, spectra and time signals, assuming a continuity between measurement samples. In particular it will be discussed when probabilistic assumptions are needed, leading to a statistically-based inference of the structure, and when a pure, non-probabilistic structural inference scheme may be possible.

1 Introduction

Our ability to recognize patterns is based on the capacity to generalize. We are able to judge new, yet unseen observations given our experience with the previous ones that are similar in one way or another. Automatic pattern recognition studies the ways which make this ability explicit. We thereby learn more about it, which is of pure scientific interest, and we construct systems that may partially take over our pattern recognition tasks in real life: reading documents, judging microscope images for medical diagnosis, identifying people or inspecting industrial production.

In this paper we will reconsider the basic principles of generalization, especially in relation with sensor measurements like images (e.g. taken from some video or CCD camera), time signals (e.g. sound registered by a microphone), and spectra and histograms (e.g. the infra-red spectrum of a point on earth measured from a satellite). These classes of measurements are of particular interest since they can very often replace the real object in case of human recognition: we can read a document, identify a person, recognize an object presented on a monitor screen as well as by a direct observation. So we deal here with registered signals which contain sufficient information to enable human recognition in an almost natural way. This is an entirely different approach to study the weather patterns

D.-Y. Yeung et al. (Eds.): SSPR&SPR 2006, LNCS 4109, pp. 41–55, 2006.
© Springer-Verlag Berlin Heidelberg 2006

from a set of temperature and air pressure measurements than taken by a farmer who observes the clouds and the birds.

The interesting, common aspect of the above defined set of sensor measurements is that they have an observable structure, emerging from a relation between neighboring pixels or samples. In fact we do not perceive the pixel intensity values themselves, but we directly see a more global, meaningful structure. This structure, the unsampled continuous observation in space and/or time constitutes the basis of our recognition. Generalization is based on a direct observation of the similarity between the new and the previously observed structures.

There is an essential difference between human and automatic pattern recognition, which will be neglected here, as almost everywhere else. If a human observes a structure, he may directly relate this to a meaning (function or a concept). By assigning a word to it, the perceived structure is named, hence recognized. The word may be different in different languages. The meaning may be the same, but is richer than just the name as it makes a relation to the context (or other frame of reference) or the usage of the observed object. On the contrary, in automatic recognition it is often attempted to map the observations directly to class labels without recognizing the function or usage.

If we want to simulate or imitate the human ability of pattern recognition it should be based on object structures and the generalization based on similarities. This is entirely different from the most successful, mainline research in pattern recognition, which heavily relies on a feature-based description of objects instead of their structural representations. Moreover, generalization is also heavily based on statistics instead of similarities.

We will elaborate on this paradoxical situation and discuss fundamentally the possibilities of the structural approach to pattern recognition. This discussion is certainly not the first on this topic. In general, the science of pattern recognition has already been discussed for a long time, e.g. in a philosophical context by Sayre [1] or by Watanabe on several occasions, most extensively in his book on human and mechanical recognition [2]. The possibilities of a more structural approach to pattern recognition was one of the main concerns of Fu [3], but it was also clear that, thereby, the powerful tools of statistical approaches [4,5,6,7] should not be forgotten; see [8,9,10].

Learning from structural observations is the key question of the challenging and seminal research programme of Goldfarb [10,11,12]. He starts, however, from a given structural measurement, the result of a 'structural sensor' [13] and uses this to construct a very general, hierarchial and abstract structural description of objects and object classes in terms of primitives, the Evolving Transformation System (ETS) [11]. Goldfarb emphasizes that a good structural representation should be able to generate proper structures. We recognize that as a desirable, but very ambitious direction. Learning structures from examples in the ETS framework appears still to be very difficult, in spite of various attempts [14].

We think that starting from such a structural representation denies the quantitative character of the lowest level of senses and sensors. Thereby, we will again face the question how to come to structure, how to learn it from examples given

the numeric outcomes of a physical measurement process, that by its organization in time and space respects this structure. This question will not be solved here, as it is one of the most basic issues in science. However, we hope that a contribution is made towards the solution by our a summary of problems and possibilities in this area, presented from a specific point of view.

Our viewpoint, which will be explained in the next sections, is that the feature vector representation directly reduces the object representation. This causes a class overlap that can only be solved by a statistical approach. An indirectly reducing approach based on similarities between objects and proximities of their representations, may avoid, or at least postpone such a reduction. As a consequence, classes do not overlap intrinsically, by which a statistical class description can be avoided. A topological- or domain-based description of classes may become possible, in which the structural aspects of objects and object classes might be preserved. This discussion partially summarizes our previous work on the dissimilarity approach [15], proximities [16], open issues [17] and the science of pattern recognition [18].

2 Generalization Principles

The goal of pattern recognition may be phrased as the derivation of a general truth (e.g. the existence of a specified pattern) from a limited, not exhaustive set of examples. We may say that we thereby generalize from this set of examples, as the establishment of a general truth gives the possibility to derive non-observed properties of objects, similar to those of observed examples.

Another way to phrase the meaning of generalization is to state that the truth is *inferred* from the observations. Several types of inference can be distinguished:

Logical inference. This is the original meaning of inference: a truth is derived from some facts, by logical reasoning, e.g.

1. Socrates is a man.
2. All man are mortal.
3. Consequently, Socrates is mortal.

It is essential that the conclusion was derived before the death of Socrates. It was already known without having observed it.

Grammatical inference. This refers to the grammar of an artificial language of symbols, which describes the "sentences" that are permitted from a set of observed sequences of such symbols. Such grammars may be inferred from a set of examples.

Statistical inference. Like above, there are observations and a general, accepted or assumed, rule of a statistical (probabilistic) nature. When such a rule is applied to the observations, more becomes known than just the directly collected facts.

Structural inference. This is frequently used in the sense that structure is derived from observations and some general law. E.g. in some economical publications, "structural inference" deals with finding the structure of a statistical model (such as the set of dependencies) by statistical means [19]. On

the contrary, "structural inference" can also be understood as using structural (instead of statistical) properties to infer unobserved object properties.

Empirical inference. This term is frequently used by Vapnik, e.g. in his recent revised edition of the book on structural risk minimization [20]. It means that unnecessary statistical models are avoided if some value, parameter, or class membership has to be inferred from observational data. It is, however, still based on a statistical approach, in the sense that probabilities and expectations play a role. The specific approach of empirical inference avoids the estimation of statistical functions and models where possible: do not estimate an entire probability density function if just a decision is needed.

It should be noted that in logical, statistical and empirical inferences object properties are inferred by logical, statistical and empirical means, respectively. In the terms of "grammatical inference" and "structural inference", the adjective does not refer to the means but to the goal: finding a grammar or a structure. The means are in these cases usually either logical or statistical. Consequently, the basic tools for inference are primarily logic and statistics. They correspond to knowledge and observations. As logic cannot directly be applied to sensor data, statistical inference is the main way for generalization in this case.

We will discuss whether in addition to logic and statistics, also structure can be used as a basic means for inference. This would imply that given the structure of a set of objects and, for instance, the corresponding class labels, the class label of an unlabeled object can be inferred. As we want to learn from sensor data, this structure should not be defined by an expert, but should directly be given from the measurements, e.g. the chain code of an observed contour.

Consider the following example. A professor in archeology wants to teach a group of students the differences in the styles of A and B of some classical vases. He presents 20 examples for each style and asks the students to determine a rule. The first student observes that the vases in group A have either ears or are red, while those of group B may also have ears, but only if they are blue (a color that never occurs for A). Moreover, there is a single red vase in group B without ears, but with a sharp spout. In group A only some vases with ears have a spout. The rule he presents is: `if (ears` $\wedge$ `not_blue)` $\vee$ `(red` $\wedge$ `no_ears` $\wedge$ `no_spout) then A else B.` The second student measures the sizes (weight and height) of all vases, plots them on a 2D scatter plot and finds a straight line that separates the vases with just two errors. The third student manually inspects the vases from all sides and concludes that the lower part is ball-shaped in group A and egg-shaped in group B. His rule is thereby: `if ball-shaped then A, if egg-shaped then B.`

The professor asked the first student why he did not use characteristic paintings on the vases for their discrimination. The student answered that they were not needed as the groups could have perfectly been identified by the given properties. They may, however, be needed if more vases appear. So, this rule works for the given set of examples, but does it generalize?

The second solution did not seem attractive to the professor as some measurement equipment is needed and, moreover, two errors are made! The student

responded that these two errors showed in fact that his statistical approach was likely better than the logical approach of the first student, as it was more general (less overtrained). This remark was not appreciated by the professor: very strange to prove the quality of a solution by the fact that errors are made!

The third student seemed to have a suitable solution. Moreover, the shape property was in line with other characteristics of the related cultures. Although it was clear what was meant by the ball-ness and the egg-ness of the vase shapes, the question remained whether this could be decided by an arbitrary assistant. The student had a perfect answer. He drew the shapes of two vases, one from each group, on a glass window in front of the table with vases. To classify a given vase, he asked the professor to look through each of the two images to this vase and to walk to and from the window to adjust the size until a match occurs.

We hope that this example makes clear that logic, statistics and structure can be used to infer a property like a class label. Much more has to be explained about how to derive the above decision rules by automatic means. In this paper, we will skip the logical approach as it has little to do with the sensory data we are interested in.

3 Feature Representation

We will first shortly summarize the feature representation and some of its advantages and drawbacks. In particular, it will be argued how this representation necessarily demands a statistical approach. Hence, this has far reaching consequences concerning how learning data should be collected. Features are object properties that are suitable for their recognition. They are either directly measured or derived from the raw sensor data. The feature representation represents objects as vectors in a (Euclidean) feature space. Usually, but not always, the feature representation is based on a significant reduction. Real world objects cannot usually be reconstructed from their features. Some examples are:

- Pieces of fruit represented by their color, maximum length and weight.
- Handwritten digits represented by a small set of moments.
- Handwritten digits represented by the pixels (in fact, their intensities) in images showing the digits.

This last example is special. Using pixel values as features leads to pixel representations of the original digits that are reductions: minor digit details may not be captured by the given pixel resolution. If we treat, however, the digital picture of a digit as an object, the pixel representation is complete: it represents the object in its entirety. This is not strange as in handling mail and money transfers, data typists often have to recognize text presented on monitor screens. So the human recognition is based on the same data as used for the feature (pixels) representation.

Note that different objects may have identical representations, if they are mapped on the same vector in the feature space. This is possible if the feature representation reduces the information on objects, which is the main cause for class overlap, in which objects belonging to different classes are identically represented.

The most common and most natural way to solve the problem of class overlap is by using probability density functions. Objects in the overlap area are assigned to the class that is the most probable (or likely) for the observed feature vector. This not only leads to the fully Bayesian approaches, based on the estimation of class densities and using or estimating class prior probabilities, but also to procedures like decision trees, neural networks and support vector machines that use geometrical means to determine a decision boundary between classes such that some error criterion is minimized.

In order to find a probability density function in the feature space, or in order to estimate the expected classification performance for any decision function that is considered in the process of classifier design, a set of objects has to be available that is representative for the distribution of the future objects to be classified later by the final classifier. This last demand is very heavy. It requires that the designer of a pattern recognition system knows exactly the circumstances under which it will be applied. Moreover, he has to have the possibility to sample the objects to be classified. There are, however, many applications in which it is difficult or impossible. Even in the simple problem of handwritten digit recognition it may happen that writing habits change over time or are location dependent. In an application like the classification of geological data for mining purposes, one likes to learn from existing mining sites how to detect new ones. Class distributions, however, change heavily over the earth.

Another problem related to class overlap is that densities are difficult to estimate for more complete and, thereby, in some sense better representations, as they tend to use more features. Consequently, they have to be determined in high-dimensional vector spaces. Also the geometrical procedures suffer from this, as the geometrical variability in such spaces is larger. This results in the paradox of the feature representation: more complete feature representations need larger training sets or will deteriorate in performance [21].

There is a fundamental question of how to handle the statistical problem of overlapping classes in case no prior information is available about the possible class distributions. If there is no preference, the No-Free-Lunch-Theorem [22] states that all classifiers perform similarly to a random class assignment if we look over a set of problems on average. It is necessary to restrict the set of problems significantly, e.g. to compact problems in which similar objects have similar representations. It is, however, still an open issue how to do this [23]. As long as the set of pattern recognition problems is based on an unrealistic set, studies on the expected performance of pattern recognition systems will yield unrealistic results. An example is the Vapnik-Chervonenkis error bound based on the structural risk minimization [20]. Although a beautiful theoretical result is obtained, the prescribed training set sizes for obtaining a desired (test) performance are far from being realistic. The support vector machine (SVM), which is based on structural risk minimization, is a powerful classifier for relatively small training sets and classes that have a small overlap. As a general solution for overlapping classes, as they arise in the feature space, it is not suitable. We will point this out below.

We will now introduce the idea of domain-based classifiers [24]. They construct decision boundaries between classes that are described just by the domains they cover in the feature space (or in any representation space) and do not depend on (the estimates of) probability distributions. They are, thereby, insensitive to ill-sampled training sets, which may even be selectively sampled by an expert. Such classifiers may be beneficial for non-overlapping, or slightly overlapping classes and are optimized for distances instead of densities. Consequently, they are sensitive to outliers. Therefore, outliers should be removed in the firs step. This is possible as the training set can be sampled in a selective way. Domain-based classification may be characterized as taking care of the structure of the classes in the feature space instead of their probability density functions.

If Vapnik's concept of structural risk minimization [20] is used for optimizing a separation function between two sets of vectors in a vector space, the resulting classifier is the maximum margin classifier. In case no linear classifier exists to make a perfect separation, a kernel approach may be used to construct a non-linear separation function. Thanks to the reproducing property of kernels, the SVM becomes then a maximum margin hyperplane in a Hilbert space induced by the specified kernel [25]. The margin is only determined by support vectors. These are the boundary objects, i.e. the objects closest to the decision boundary $f(\mathbf{x};\boldsymbol{\theta})$ [26,25]. As such, the SVM is independent of class density models. Multiple copies of the same object added to the training set do not contribute to the construction of the SVM as they do for classifiers based on some probabilistic model. Moreover, the SVM is also not affected by adding or removing objects of the same class that lie further away from the decision boundary. This decision function is, thereby, a truly domain-based classifier, as it optimizes the separation of class domains and class density functions.

For nonlinear classifiers defined on nonlinear kernels, the SVM has, however, a similar drawback as the nonlinear neural network. The distances to the decision boundary are computed in the output Hilbert space defined by the kernel and not in the input space. A second problem is that the soft-margin formulation [26], the traditional solution to overlapping classes, is not domain-based. Consider a two-class problem with the labels $y \in \{-1, +1\}$, where $y(\mathbf{x})$ denotes the true label of $\mathbf{x}$. Assume a training set $X = \{\mathbf{x}_i, y(\mathbf{x}_i)\}_{i=1}^n$. The optimization problem for a linear classifier $f(\mathbf{x}) = \mathbf{w}^\mathsf{T}\mathbf{x} + w_0$ is rewritten into:

$$\begin{array}{ll} \min_{\mathbf{W}} & ||\mathbf{w}||^2 + C \sum_{\mathbf{x}_i \in X} \xi(\mathbf{x}_i), \\ s.t. & y(\mathbf{x}_i) f(\mathbf{x}_i) \geq 1 - \xi(\mathbf{x}_i), \\ & \xi(\mathbf{x}_i) \geq 0, \end{array} \tag{1}$$

where $\xi(\mathbf{x}_i)$ are slack variables accounting for possible errors and C is a trade-off parameter. $\sum_{\mathbf{x}_i \in X} \xi(\mathbf{x}_i)$ is an upper bound of the misclassification error on the training set, hence it is responsible for minimizing *a sum of error contributions.* Adding a copy of an erroneously assigned object will affect this sum and, thereby, will influence the sought optimum $\mathbf{w}$. The result is, thereby, based on a mixture of approaches. It is dependent on the distribution of objects (hence statistics) as well as on their domains (hence geometry).

A proper domain-based solution should minimize the class overlap in terms of distances and not in terms of probability densities. Hence, a suitable version of the SVM should be derived for the case of overlapping domains, resulting in the *negative margin SVM* [24]. This means that the distance of the furthest away misclassified object should be minimized. As the signed distance is negative, the negative margin is obtained. In the probabilistic approach, this classifier is unpopular as it will be sensitive to outliers. As explained above, outliers are neglected in domain-based classification, as they have to be removed beforehand.

Our conclusion is that the use of features yields a reduced representation. This leads to class overlap for which a probabilistic approach is needed. It relies on a heavy assumption that data are drawn independently from a fixed (but unknown) probability distribution. As a result, one demands training sets that are representative for the probability density functions. An approach based on distances and class structures may be formulated, but conflicts with the use of densities if classes overlap.

4 Proximity Representation

Similarity or dissimilarity measures can be used to represent objects by their proximities to other examples instead of representing them by a preselected set of features. If such measurements are derived from original objects, or from raw sensor data describing the objects fully (e.g. images, time signals and spectra that are as good as the real objects for the human observer), then the reduction in representation, which causes class overlap in the case of features, is circumvented. For example, we may demand that the dissimilarity of an object to itself is zero and that it can only be zero if it is related to an identical object. If it can be assumed that identical objects belong to the same class, classes do not overlap. (This is not always the case, e.g. a handwritten '7' may be identical to a handwritten '1').

In principle, such proximity representations may avoid class overlap. Hence, they may offer a possibility to use the structure of the classes in the representation, i.e. their domains, for building classifiers. This needs a special, not yet well studied variant of the proximity representation. Before a further explanation, we will first summarize two variants that have been worked out well. This summary is an adapted version of what has been published as [16]. See also [15].

Assume we are given a representation set R, i.e. a set of real-world objects that can be used for building the representation. $R = \{p_1, p_2, \ldots, p_n\}$ is, thereby, a set of prototype examples. We also consider a proximity measure d, which should incorporate the necessary invariance (such as scale or rotation invariance) for the given problem. Without loss of generality, let d denote dissimilarity. An object x is then represented as a vector of dissimilarities computed between x and the prototypes from R, i.e. $d(x, R) = [d(x, p_1), d(x, p_2), \ldots, d(x, p_n)]^{\mathsf{T}}$. If we are also given an additional labeled training set $T = \{t_1, t_2, \ldots, t_N\}$ of N real-world objects, our proximity representation becomes an $N \times n$ dissimilarity matrix $D(T, R)$, where $D(t_i, R)$ is now a row vector. Usually R is selected out of T (by

various prototype selection procedures) in a way to guarantee a good tradeoff between the recognition accuracy and the computational complexity. R and T may also be different sets.

The k-NN rule can directly be applied to such proximity data. Although it has good asymptotic properties for metric distances, its performance deteriorates for small training (here: representation) sets. Alternative learning strategies represent proximity information in suitable representation vector spaces, in which traditional statistical algorithms can be defined. So, they become more beneficial. Such vector spaces are usually determined by some local or global embedding procedures. Two approaches to be discussed here rely on a linear isometric embedding in a pseudo-Euclidean space (where necessarily $R \subseteq T$) and the use of proximity spaces; see [16,15].

Pseudo-Euclidean linear embedding. Given a symmetric dissimilarity matrix $D(R,R)$, a vectorial representation X can be found such that the distances are preserved. It is usually not possible to determine such an isometric embedding into a Euclidean space, but it is possible into a pseudo-Euclidean space $\mathcal{E} = \mathbb{R}^{(p,q)}$. It is a $(p+q)$-dimensional non-degenerate indefinite inner product space such that the inner product $\langle \cdot, \cdot \rangle_{\mathcal{E}}$ is positive definite on $\mathbb{R}^p$ and negative definite on $\mathbb{R}^q$ [10]. Then, $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{E}} = \mathbf{x}^{\mathsf{T}} \mathcal{J}_{pq} \mathbf{y}$, where $\mathcal{J}_{pq} = \mathrm{diag}\,(I_{p\times p}; -I_{q\times q})$ and I is the identity matrix. Consequently, $d^2_{\mathcal{E}}(\mathbf{x},\mathbf{y}) = \langle \mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{y} \rangle_{\mathcal{E}} = d^2_{\mathbb{R}^p}(\mathbf{x},\mathbf{y}) - d^2_{\mathbb{R}^q}(\mathbf{x},\mathbf{y})$, hence $d^2_{\mathcal{E}}$ is a difference of square Euclidean distances found in the two subspaces, $\mathbb{R}^p$ and $\mathbb{R}^q$. Since $\mathcal{E}$ is a linear space, many properties related to inner products can be extended from the Euclidean case [10,15].

The (indefinite) Gram matrix G of X can be expressed by the square distances $D^{\star 2} = (d^2_{ij})$ as $G = -\frac{1}{2} J D^{\star 2} J$, where $J = I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathsf{T}}$ [10,27,15]. Hence, X can be determined by the eigendecomposion of G, such that $G = Q \Lambda Q^{\mathsf{T}} = Q|\Lambda|^{1/2} \mathrm{diag}(\mathcal{J}_{p'q'}; 0)\, |\Lambda|^{1/2} Q^{\mathsf{T}}$. $|\Lambda|$ 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 uncorrelated and represented in $\mathbb{R}^k$, $k = p' + q'$, as $X = Q_k |\Lambda_k|^{1/2}$ [10,27]. Since only some eigenvalues are significant (in magnitude), the remaining ones can be disregarded as non-informative. The reduced representation $X_r = Q_m\, |\Lambda_m|^{1/2}$, $m = p+q < k$, is determined by the largest p positive and the smallest q negative eigenvalues. New objects $D(T_{test}, R)$ are orthogonally projected onto $\mathbb{R}^m$; see [10,27,15]. Classifiers based on inner products can appropriately be defined in $\mathcal{E}$. A linear classifier $f(\mathbf{x}) = \mathbf{v}^{\mathsf{T}} \mathcal{J}_{pq} \mathbf{x} + v_0$ is e.g. constructed by addressing it as $f(\mathbf{x}) = \mathbf{w}^{\mathsf{T}} \mathbf{x} + v_0$, where $\mathbf{w} = \mathcal{J}_{pq} \mathbf{v}$ in the associated Euclidean space $\mathbb{R}^{(p+q)}$ [10,27,15].

Proximity spaces. Here, the dissimilarity matrix $D(X,R)$ is interpreted as a data-dependent mapping $D(\cdot, R)$: $X \to \mathbb{R}^n$ from some initial representation X to a vector space defined by the set R. This is the *dissimilarity space* (or a similarity space, if similarities are used), in which each dimension $D(\cdot, p_i)$ corresponds to a dissimilarity to a prototype $p_i \in R$. The property that dissimilarities should be small for similar objects (belonging to the same class) and large for distinct objects, gives them a discriminative power. Hence, the vectors $D(\cdot, p_i)$

can be interpreted as 'features' and traditional statistical classifiers can be defined [28,15]. Although the classifiers are trained on $D(\cdot, R)$, the weights are still optimized on the complete set T. Thereby, they can outperform the k-NN rule as they become more global in their decisions.

Normal density-based classifiers perform well in dissimilarity spaces [27,28,15]. This especially holds for summation-based dissimilarity measures, summing over a number of components with similar variances. Such dissimilarities are approximately normally distributed thanks to the central limit theorem (or they approximate the χ^2 distribution if some variances are dominant) [15]. For instance, for a two-class problem, a quadratic normal density based classifier is given by $f(D(x,R)) = \sum_{i=1}^{2} \frac{(-1)^i}{2}(D(x,R)-\mathbf{m}_i)^\mathsf{T} S_i^{-1}\,(D(x,R)-\mathbf{m}_i) + \log\frac{p_1}{p_2} + \frac{1}{2}\log\frac{|S_1|}{|S_2|}$, where $\mathbf{m}_i$ are the mean vectors and S_i are the class covariance matrices, all estimated in the dissimilarity space $D(\cdot, R)$. p_i are the class prior probabilities. By replacing S_1 and S_2 by the average covariance matrix, a linear classifier is obtained.

The two learning frameworks of pseudo-Euclidean embedding and dissimilarity spaces appear to be successful in many problems with various kinds of dissimilarity measures. They can be more accurate and more efficient than the nearest neighbor rule, traditionally applied to dissimilarity data. Thereby, they provide beneficial approaches to learning from structural object descriptions for which it is more easy to define dissimilarity measures between objects than to find a good set of features. As long as these approaches are based on a fixed representation set, however, class overlap may still arise as two different objects may have the same set of distances to the representation set. Moreover, most classifiers used in the representation spaces are determined based on the traditional principle of minimizing the overlap. They do not make a specific use of principles related to object distances or class domains. So, what is still lacking are procedures that use class distances to construct a structural description of classes. The domain-based classifiers, introduced in Section 3, may offer that in future provided that the representation set is so large that the class overlap is (almost) avoided. A more fundamental approach is described below.

Topological spaces. The topological foundation of proximity representations is discussed in [15]. It is argued that if the dissimilarity measure itself is unknown, but the dissimilarity values are given, the topology cannot, as usual, be based on the traditional idempotent closures. An attempt has been made to use neighborhoods instead. This has not resulted yet in a useful generalization over finite training sets.

Topological approaches will aim to describe the class structures from local neighborhood relations between objects. The inherent difficulty is that many of the dissimilarity measures used in structural pattern recognition, like the normalized edit distance, are non-Euclidean, and even sometimes non-metric. It has been shown in a number of studies that straightforward Euclidean corrections are counter productive in some applications. This suggests that the non-Euclidean aspects may be informative. Consequently, a non-Euclidean topology would be needed. This area is still underdeveloped.

A better approach may rely on two additional sources of information that are additionally available. These are the definition of the dissimilarity measure and the assumption of class compactness. They may together tell us what is really local or how to handle the non-Euclidean phenomena of the data. This should result in a topological specification of the class structure as learned from the training set.

5 Structural Representation

In the previous section we arrived at *a structure of a class* (or a concept), i.e. the structural or topological relation of the set of all objects belonging to a particular class. This structure is influenced by the chosen representation, but is in fact determined by the class of objects. It reflects, for instance, the set of continuous transformations of the handwritten digits '7' that generate exclusively all other forms that can be considered as variants of a handwritten '7'. This basically reflects the concept used by experts to assign the class label. Note, however, that this rather abstract structure of the concept should be clearly distinguished from the structure of individual objects that are the manifestations of that concept.

The *structure of objects*, as presented somewhere in sensory data of images, such as time signals and spectra, is directly related to shape. The shape is a one- or multi-dimensional set of connected boundary points that may be locally characterized by curvature and described more globally by morphology and topology. Note that the object structure is related to an outside border of objects, the place where the object ends. If the object is a black blob in a two-dimensional image (e.g. a handwritten digit) then the structure is expressed by the contour, a one-dimensional closed line. If the grey-value pixel intensities inside the blob are relevant, then we deal with a three-dimensional blob on a two-dimensional surface. (As caves cannot exist in this structure it is sometimes referred to as a 2.5-dimensional object).

It is important to realize that the sensor measurements are characterized by a sampling structure (units), such as pixels or time samples. This sampling structure, however, has nothing to do with the object structure. In fact, it disturbs it. In principle, objects (patterns describing real objects) can lie anywhere in an image or in a time frame. They can also be rotated in an image and appear in various scales. Additionally, we may also vary the sampling frequency. If we analyze the object structure for a given sampling, then the object is "nailed" to some grid. Similar objects may be nailed in an entirely different way to this grid. How to construct structural descriptions of objects that are independent of the sampling grid on which the objects are originally presented is an important topic of structural pattern recognition.

The problem of structural inference, however, is not the issue of representation itself. It is the question how we can establish the membership of an object to a given set of examples based on their structure. Why is it more likely that a new object X belongs to a set A than a set B? A few possible answers are presented below.

1. X is an example of A, because the object in $A \cup B$ that is most similar to X belongs to A. This decision may depend on the accidental availability of particular objects. Moreover, similarity should appropriately be defined.
2. X is an example of A, because the object from $A \cup B$ that is most easily transformed to X belongs to A. In this case similarity relies on the effort of transformation. This may be more appropriate if structures need to be compared. The decision, however, still depends on a single object. The entire sets or classes simply store examples that may be used when other objects have to be classified.
3. X is an example of A, because it can more easily be generated by transforming the prototype of set A than by transforming the prototype of set B. The *prototype* of a set may be defined as the (hypothetical) object that can most easily be transformed into any of the objects of the set. In this assignment rule (as well as in the rule above) the definition of transformation is universal, i.e. independent of the considered class.
4. X is an example of A, because it can more easily be transformed from a (hypothetical) prototype object by the transformations T_A that are used to generate the set A than by the transformations T_B that are used to generate the set B. Note that we now allow that the sets are generated from possibly the same prototype, but by using different transformations. These are derived (learnt) from the sets of examples. The transformations T_A and T_B may be learnt from a training set.

There is a strong resemblance with the statistical class descriptions: classes may differ by their means as well as by the shape of their distributions. A very important difference, however, between structural and statistical inference is that for an additional example that is identical to a previous one changes the class distribution, but not the (minimal) set of necessary transformations.

This set of assignment rules can easily be modified or enlarged. We like to emphasize, however, that the natural way of comparing objects, i.e. by accounting for their similarity, may be defined as the effort of transforming one structure into another. Moreover, the set of possible transformations may differ from class to class. In addition, classes may have the same or different prototypes. E.g. a sphere can be considered as a basic prototype both for apples as well as for pears. In general, classes may differ by their prototypes and/or by their set of transformations.

What has been called easiness in transformation can be captured by a measurable cost, which is an example of a similarity measure. It is, thereby, related to the proximity approaches, described above. Proximity representations are naturally suitable for structural inference. What is different, however, is the use of statistical classifiers in embedded and proximity spaces. In particular, the embedding approach has to be redefined for structural inference as it makes use of averages and the minimization of an expected error, both statistical concepts. Also the use of statistical classifiers in these spaces conflicts with structural inference. In fact, they should be replaced by domain-based classifiers. The discussed topological approach, on the other hand, fits to the concept of structural inference.

The idea that transformations may be class-dependent has not been worked out by us in the proximity-based approach. There is, however, not a fundamental

objection against the possibility to attribute set of objects, or even individual objects in the training set with their own proximity measure. This will very likely lead to non-Euclidean data, but we have shown ways how to handle them. What is not studied is how to optimize proximity measures (structure transformations) over the training data. A possibility might be to normalize for differences in class structure by adapting the proximity measures that determined these structures.

There is, however, an important aspect of learning from structures that cannot currently be covered by domain-based classifiers built for a proximity representation. Structures can be considered as assemblies of more primitive structures, similarly as a house is built from bricks. These primitives may have a finite size, or may also be infinitesimally small. The corresponding transformations from one structure into another become thereby continuous. In particular, we are interested in such transformations as they may constitute the compactness of classes on which a realistic set of pattern recognition problems can be defined. It may be economical to allow for locally-defined functions in order to derive (or learn) transformations between objects. For instance, while comparing dogs and wolves, or while describing these groups separately, other transformations may be of interest for the description of ears then for the tails. Such a decomposition of transformations is not possible in the current proximity framework, as it starts with relations between entire objects. A further research is needed.

The automatic detection of parts of objects where different transformations may be useful for the discrimination (or a continuous varying transformation over the object) seems very challenging, as the characteristics inside an object are ill-defined as long as classes are not fully established during training. Some attempts in this direction have been made by Paclík [29,30] when he tries to learn the proximity measure from a training set.

In summary, we see three ways to link structural object descriptions to the proximity representation:

- Finding or generating prototypical objects that can easily be transformed into the given training set. They will be used in the representation set.
- Determining specific proximity measures for individual objects or for groups of objects.
- Learning locally dependent (inside the object) proximity measures.

6 Discussion and Conclusions

In this paper, we presented a discussion of the possibilities of structural inference as opposed to statistical inference. By using the structural properties of objects and classes of a given set of examples, knowledge such as class labels is inferred for new objects. Structural and statistical inference are based on different assumptions with respect to the set of examples needed for training and for the object representation. In a statistical approach, the training set has to be representative for the class distributions as the classifiers have to assign objects to the most probable class. In a structural approach, classes may be assumed to be separable. As a consequence, domain-based classifiers may be used [18,24]. Such classifiers, which are mainly still under development, do not need training sets

that are representative for the class distributions, but which are representative for the class domains. This is greatly advantageous as these domains are usually stable with respect to changes in the context of application. Training sets may thereby be collected by a selective, instead of unselective sampling.

The below table summarizes the main differences between representations based on features (F), proximities (P) and structures (S) for the statistical and structural inference.

	Statistical inference	Structural inference
F	Features reduce; statistical inference is almost obligatory.	The structural information is lost by representing the aspects of objects by vectors and/or due to the reduction.
P	Proximity representations can be derived by comparing pairs of objects (e.g. initially described by features or structures). Statistical classifiers are built in proximity spaces or in (pseudo-Euclidean) embedded spaces.	Transformations between the structures of objects may be used to build proximity representations. Classes of objects should be separated by domain-based classifiers.
S	Statistical learning is only possible if a representation vector space is built (by features or proximities), in which density functions can be defined.	Transformations might be learnt by using a domain-based approach that transforms one object into another in an economical way.

This paper summarizes the possibilities of structural inference. In particular, the possibilities of the proximity representation are emphasized, provided that domain-based learning procedures follow. More advanced approaches, making a better usage of the structure of individual objects have to be studied further. They may be based on the generation of prototypes or on trained, possibly local transformations, which will separate object classes better. Such transformations can be used to define proximity measures, which will be further used to construct a proximity representation. Representations may have to be directly built on the topology derived from object neighborhoods. These neighborhoods are constructed by relating transformations to proximities. The corresponding dissimilarity measures will be non-Euclidean, in general. Consequently, non-Euclidean topology has to be studied to proceed in this direction fundamentally.

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

References

1. Sayre, K.: Recognition, a study in the philosophy of artificial intelligence. University of Notre Dame Press (1965)
2. Watanabe, S.: Pattern Recogn. Human and Mechanical. Academic Press (1974)
3. Fu, K.: Syntactic Pattern Recognition and Applications. Pretice-Hall (1982)
4. Fukunaga, K.: Introduction to Statistical Pattern Recogn. Academic Press (1990)
5. Duda, R., Hart, P., Stork, D.: Pattern Classification. John Wiley & Sons, Inc. (2001)
6. Webb, A.: Statistical Pattern Recognition. John Wiley & Sons, Ltd. (2002)

7. Jain, A., Duin, R.P.W., Mao, J.: Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(1) (2000) 4–37
8. Bunke, H., Günter, S., Jiang, X.: Towards bridging the gap between statistical and structural pattern recognition: Two new concepts in graph matching. In: International Conference on Advances in Pattern Recognition. (2001) 1–11
9. Fu, K.: A step towards unification of syntactic and statistical pattern recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **8** (1986)
10. Goldfarb, L.: A new approach to pattern recognition. In Kanal, L., Rosenfeld, A., eds.: Progress in Pattern Recognition. Volume 2. Elsevier Science Publishers BV (1985) 241–402
11. Goldfarb, L., Gay, D.: What is a structural representation? Fifth variation. Technical Report TR05-175, University of New Brunswick, Fredericton, Canada (2005)
12. Goldfarb, L.: What is distance and why do we need the metric model for pattern learning? Pattern Recognition **25**(4) (1992) 431–438
13. Goldfarb, L., Golubitsky, O.: What is a structural measurement process? Technical Report TR01-147, University of New Brunswick, Fredericton, Canada (2001)
14. Gutkin, A., Gya, D., Goldfarb, L., Webster, M.: On the articulatory representation of speech within the evolving transformation system formalism. In Goldfarb, L., ed.: Pattern representation and the future of pattern recognition, ICPR 2004 Workshop Proceedings, Cambridge, United Kingdom (2004) 57–76
15. Pękalska, E., Duin, R.P.W.: The Dissimilarity Representation for Pattern Recognition. Foundations and Applications. World Scientific, Singapore (2005)
16. Pękalska, E., Duin, R.P.W.: Learning with general proximity measures. In: Pattern Recognition in Information Systems. Volume 6. (2006)
17. Duin, R.P.W., Pękalska, E.: Open issues in pattern recognition. In: Computer Recognition Systems. Springer, Berlin (2005) 27–42
18. Duin, R.P.W., Pękalska, E.: The science of pattern recognition. Achievements and perspectives. (2006, submitted)
19. Dawid, A., Stone, M., Zidek, J.: Marginalization paradoxes in Bayesian and structural inference. J. Royal Stat. Soc., B **35** (1973) 180–223
20. Vapnik, V.: Estimation of Dependences based on Empirical Data, 2nd ed. Springer Verlag (2006)
21. Jain, A.K., Chandrasekaran, B.: Dimensionality and sample size considerations in pattern recognition practice. In Krishnaiah, P.R., Kanal, L.N., eds.: Handbook of Statistics. Volume 2. North-Holland, Amsterdam (1987) 835–855
22. Wolpert, D.: The Mathematics of Generalization. Addison-Wesley (1995)
23. Duin, R.P.W., Roli, F., de Ridder, D.: A note on core research issues for statistical pattern recognition. Pattern Recognition Letters **23**(4) (2002) 493–499
24. Duin, R.P.W., Pękalska, E.: Domain-based classification. Technical report, TU Delft (2005) `http://ict.ewi.tudelft.nl/~{}duin/papers/Domain_class_05.pdf`.
25. Vapnik, V.: Statistical Learning Theory. John Wiley & Sons, Inc. (1998)
26. Cristianini, N., Shawe-Taylor, J.: Support Vector Machines and other kernel-based learning methods. Cambridge University Press, UK (2000)
27. Pękalska, E., Paclík, P., Duin, R.P.W.: A Generalized Kernel Approach to Dissim. Based Classification. Journal of Machine Learning Research **2**(2) (2002) 175–211
28. Pękalska, E., Duin, R.P.W., Paclík, P.: Prototype selection for dissimilarity-based classifiers. Pattern Recognition **39**(2) (2006) 189–208
29. Paclík, P., Novovicova, J., Duin, R.P.W.: Building road sign classifiers using a trainable similarity measure. Journal of Intelligent Transportations Systems (2006)
30. Paclík, P., Novovicova, J., Duin, R.P.W.: A trainable similarity measure for image classification. In: 17th Int. Conf. on Pattern Recognition. (2006)