2

Object Representation, Sample Size, and Data Set Complexity

Robert P.W. Duin and Elżbieta Pękalska

Summary. The complexity of a pattern recognition problem is determined by its representation. It is argued and illustrated by examples that the sampling density of a given data set and the resulting complexity of a learning problem are inherently connected. A number of criteria are constructed to judge this complexity for the chosen dissimilarity representation. Some nonlinear transformations of the original representation are also investigated to illustrate that such changes may affect the resulting complexity. If the initial sampling density is originally insufficient, this may result in a data set of a lower complexity and with a satisfactory sampling. On the other hand, if the number of samples is originally abundant, the representation may become more complex.

2.1 Introduction

To solve a particular problem, one will be interested in its complexity to find a short path to the solution. The analyst will face an easy and straightforward task if the solution follows directly from the way the problem is stated. The problem will be judged as complex if one needs to use a large set of tools and has to select the best procedure by a trial-and-error approach or if one has to integrate several partial solutions. A possible way to proceed is to simplify the initial problem, e.g., by removing its most weakly determined aspects. This chapter focuses on these two issues: judging the complexity of a problem from the way it is presented, and discussing some ways to simplify it if the complexity is judged as too large.

The complexity of pattern recognition problems has recently raised some interest [16, 17]. It is hoped that its study may contribute to the selection of appropriate methods to solve a given problem. As the concept of problem complexity is still ill-defined, we will start to clarify our approach, building on some earlier work [10].

Pattern recognition problems may have some intrinsic overlap. This does not contribute to the problem complexity, as an existing intrinsic overlap cannot be removed by any means. The complexity of the problem lies in difficulties one encounters in the above sketched sense, while approaching a classification performance related to the intrinsic class overlap. Because problems are numerically encoded by data sets representing the classes of objects for which either pattern classes have to be learned or classifiers have to be determined, the complexity of the recognition problem is the complexity of the representation as one observes through some data set. Such representations heavily influence the complexity of the learning problem.

An important aspect of the representation is the nature of numerical encoding used for the characterization of objects, as, for example, features or proximities between pairs of objects, or proximities of objects to class models. Even if objects are first represented in a structural form, such as relational graphs or strings, we will assume that a numerical representation (e.g., by dissimilarities) is derived from such an intermediate description. In addition, the number of objects in the data set, i.e., the sample size, and the way the objects are sampled from the problem (at random or by some systematic procedure) influence the complexity. As the exploration or classification problems have to be solved using a data set based on some representation, the complexity of the problem is reflected by the data set and the representation.

This chapter focuses on the influence of sample size on the complexity of data sets used for learning pattern classes. These classes are characterized by dissimilarity representations [22, 23], which are primarily identified by sample sizes and not yet by the dimensionality of some space, as feature vector representations are. Since the given problem, the chosen representation and the derived data set are essentially connected, we will use the word *complexity* interchangeably with respect to these three concepts.

To analyze complexity in learning, one needs to understand better what complexity is. In general, complexity is defined as “the quality of being intricate and compounded” [34]. Loosely speaking, this means that an entity, a problem, a task, or a system is complex if it consists of a number of elements (components) related such that it is hard to separate them or to follow their interrelations. Intuitively, an entity is more complex if more components and more interdependencies can be distinguished. So, complexity can be characterized by the levels and the kinds of distinction and dependency. The former is related to the variability, i.e., the number of elements,and their size and shape, while the latter refers to the dependency between the components. It will be a key issue of this chapter to make clear that the set of examples used to solve the pattern recognition problem should be sufficiently large in order to meet the complexity of the representation.

Reductionism treats an entity by the sum of its components or a collection of parts. Holism, on the other hand, treats an entity as a whole, hence it does not account for distinguishable parts. The complexity can be seen as an interplay between reductionism and holism: it needs to see distinct elements, but also their interrelations, in order to realize that they cannot be separated without losing a part of their meaning; see also the development of the science of complexity as sketched by Waldrop [31]. In fact, reductionism and holism can be seen on different, organizational levels. For instance, to understand the complexity of an ant colony (see Hofstadter’s chapter on “Ant Fugue” [18]), one needs to observe the activities of individual ants as well as the colony as a whole. On the level of individuals, they may seem to move in random ways, yet on the level of specialized casts and the colony, clear patterns can be distinguished. These relate to a sequential (ants following other ants), parallel (groups of ants with a task), and simultaneous or emergent (the global movement) behavior of the colony. Therefore, complexity might be described by hierarchical systems, where the lowest, indivisible parts serve for building higher level structures with additional dependencies and abstraction (symbolism or meaning).

Complexity can also be placed between order and disorder (chaos). If all ants follow sequentially one another, then although the ant colony is composed of many individuals, its complexity is low because the pattern present there is simple and regular. In this sense, the colony possesses redundant information. A single ant and a direction of move will completely describe the entire colony. On the other hand, if individual ants move in different directions, but emerge into a number of groups with different tasks and following specified paths, the complexity of the ant colony becomes larger. Finally, if all ants move independently in random ways without any purpose and grouping behavior, no clear patterns can be identified. As a result, there is no complexity as it is just chaos. Therefore, complexity may be characterized by the surprise or unexpectedness on a low level that can be understood as following the structure observed from a higher point of view. In brief, following Waldrop's point of view [31], complexity arises at the edge of structure and chaos as it is pictorially illustrated in Figure 2.1.

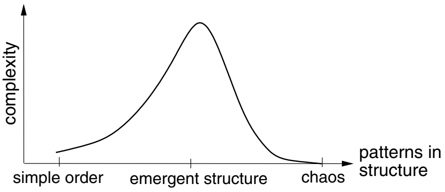


Fig. 2.1. Complexity vs. structure.

In pattern recognition one distinguishes the task of finding a classifier between some real-world classes of objects or phenomena. This task is defined on a high level. The classes may have some hidden structure that is partially reflected in the initial representation by which the problem is presented. For instance, this can be by features, dissimilarities, graphs, or other relations. Another part of the structure is implicitly available in the set of examples from which the pattern classifier has to be learned. The wholeness of the recognition problem is thereby available to us in its reduction to a set of examples by a chosen representation: the data set. The path from a pattern recognition problem to a data set determines the complexity we encounter if we try to solve the problem based on the given data set. The complexity of a pattern recognition problem (its intrinsic complexity) is simply not defined before a representation is chosen and a set of examples is collected. In the end, the data set depicts our problem.

The following example may illustrate this point. Imagine an automatic sorting of apples and pears on a moving conveyor. The complexity of this problem depends on a selection of a representative sample of apples and pears to learn from, initial measurements done by some sensors or other devices (images, spectral images, or simple characteristics such as weight, perimeter, or color) and the derived representation. In a chosen representation, the problem is complex if many examples are necessary to capture the variability and organization within the classes as well as the interrelations between the classes, leading to complicated decision functions. If one wishes to discriminate between apples and pears based on their weights only, such a problem will likely be simple. The reason is that a few suitably chosen examples will determine reliable thresholds on which such a decision relies, independently of whether this leads to frequent errors or not. On the other hand, if various Fourier coefficient and shape descriptors are computed on the images of apples and pears and treated as features, the resulting problem may become complex. Changes in light illumination or tilts of a camera may increase the variability of the (images of) apples and pears as perceived in their vector representations. This would require a large sample for a description. So, it is the representation that determines the complexity of the problem. We encounter this complexity through the data that are available.

Note, that the use of the data set as such is insufficient for solving the problem. It is just chaos if no additional background knowledge, such as the context, the way the examples are collected, or the way the numbers are measured, is given. This is very clearly shown by the "no free-lunch theorem" [33], which states that without additional knowledge, no learning algorithm is expected to be better than another. In particular, no learning algorithm outperforms a random assignment.

A very useful and often implicitly assumed type of knowledge used for a construction of the given data set is the compactness hypothesis [1, 8]. It states that similar real-world objects have similar representations. In practice, this hypothesis relies on some continuous mapping from an object to its (numerical) representation, because it is expected that a small change in an object will result in a small change in its representation. Still, the path from an object to its representation may be very nonlinear (and thereby attributing to the complexity of the problem), resulting in the violation of the reverse compactness hypothesis. This means that similar representations (e.g., feature vectors lying close in a feature vector space) may not necessarily refer to similar objects. This causes a class overlap (identical representations belong to essentially different objects as they differ in class membership) or complicates decision boundaries.

In a given data set of a limited cardinality the compactness might not be entirely realized if insufficient real-world objects are collected. Hence, it cannot be guaranteed that each object has at least one close companion. The complexity of the problem then demands a higher sampling density of (training) examples to make its characteristics apparent. As a result, the assumption needed for building classifiers on the data set is invalid and it is impossible to solve the pattern recognition problem with a sufficient accuracy. The data set resembles chaos (as patterns cannot be distinguished) and the structure of the problem cannot be determined.

The above discussion makes clear that complexity and sample size are interrelated. Complex problems (due to a complicated way they are represented by the data sets) need more samples. A question that arises now is: if the data set is insufficiently large, is it thereby less or more complex? We will return to this in the discussion section. In brief, the following issues are more explicitly studied by some examples:

- The influence of representation on the problem complexity
- The relation between the problem complexity and the necessary sample size
- The consequences of using too small sample sizes for solving complex problems

Our examples are based on a number of dissimilarity representations, which allow one to apply various modifications and transformations in a simple way. In section 2, the data sets and procedures are summarized. In section 3, various criteria are proposed and investigated to judge the sampling of single classes. Section 4 investigates and discusses the complexity issues in relation to classification. A final discussion is presented in section 2.5.

2.2 Data Sets

To limit the influence of dimensionality issues on the relations between the sample size and the complexity, we will focus on dissimilarity representations [22, 23, 26]. These are representations in which a collection of objects is encoded by their dissimilarities to a set of chosen examples, a so-called representation set. The reason we choose this name is twofold. First, the representation set is a set of examples that are not necessarily prototypical for the classes according to the usual understanding (on the contrary, some of them might be outliers). Second, this set serves for a construction of a representation space, in which both exploration and learning are performed. The representation set may be the training set itself, its randomly or selectively chosen subset or some other set. The representation set $R = \{p_1, p_2, \ldots, p_n\}$ of n examples, the (training) set $T = \{x_1, x_2, \ldots, x_N\}$ of N objects, and the dissimilarity measure d constitute together the representation $D(T, R)$. This is an $N \times n$ dissimilarity matrix, in which every entry $d(x_j, p_i)$ describes the difference between the object t_j and the representation object p_i.

Problems with various metric and nonmetric dissimilarity measures are chosen for the study. Six data sets are used in our experiments and are briefly summarized in Table 2.1. In addition to the given dissimilarity measures as listed in this table, two monotonic power transformations will be also investigated. Concerning the original representation $D = (d_{ij})$, the transformed representations are denoted as $D^{*2} = (d_{ij}^2)$ and $D^{*0.5} = (d_{ij}^{0.5})$, by taking the element-wise square or square root of the dissimilarities d_{ij}, respectively. Note that the metric properties of the measure d are preserved by a square root transformation, but not necessarily by a quadratic transformation [22]. By such modifications, it is expected that either large dissimilarities and, thereby, more global aspects of the data set are emphasized in D^{*2} or large dissimilarities are suppressed in $D^{*0.5}$, by which local aspects are strengthened. Remember that nondecreasing transformations like these do not affect the order of the given dissimilarities. Thereby, the nearest neighbor relations are preserved.

Digits-38. The data describe a set of scanned handwritten digits of the National Institute of Standards and Technology (NIST) data set [32], originally given as $128 \times$

Table 2.1. Data sets used in the experiments.

Data	Dissimilarity	Property	# classes	# objects per class
Digits-38	Euclidean	Euclidean	2	1000
Digits-all	Template-match	Nonmetric	10	200
Heart	Gower's	Euclidean	2	139/164
Polygon	Mod. Hausdorff	Nonmetric	2	2000
ProDom	Structural	Nonmetric	4	878/404/271/1051
Tumor-mucosa	$l_{0.8}$-distance	Nonmetric	2	132/856

128 binary images. Just two classes of digits, 3 and 8, are considered here. Each class consists of 1000 examples. The images are first smoothed by a Gaussian kernel with $\sigma = 8$ pixels and then the Euclidean distances between such blurred images are computed (summing up the squares of pixel-to-pixel gray value differences followed by the square root). The smoothing is done to make this distance representation more robust against tilting or shifting.

Digits-all. The data describe a set of scanned handwritten digits of the NIST data set [32], originally given as 128 × 128 binary images. The similarity measure, based on deformable template matching, as defined by Jain and Zongker [20], is used. Let $S = (s_{ij})$ denote the similarities. Since the similarity is asymmetric, the off-diagonal symmetric dissimilarities are computed as $d_{ij} = (s_{ii} + s_{jj} - s_{ij} - s_{ji})^{1/2}$ for $i \neq j$. D is significantly nonmetric [24].

Heart. This data set comes from the University of California, Irvine (UCI) Machine Learning Repository [2]. The goal is to detect the presence of heart disease in patients. There are 303 examples, of which 139 correspond to diseased patients. Various measurements are performed; however, only 13 attributes are used by other researchers for the analysis, as provided in Blake and Merz [2]. These attributes are age, sex $(1/0)$, chest pain type (1-4), resting blood pressure, serum cholesterol, fasting blood sugar >120 mg/dl $(1/0)$, resting electrocardiograph results, maximum heart rate achieved, exercise-induced angina $(1/0)$, the slope of the peak exercise ST segment, ST depression induced by exercise relative to rest (1-3), number of major vessels colored by fluoroscopy (0-3), and heart condition (normal, fixed defect, and reversible defect). Hence, the data consist of mixed types: continuous, dichotomous, and categorical variables. There are also several missing values.

Gower's [14] dissimilarity is used for the representation. Assume m features and let x_{ik} be the kth feature value for the ith object. A similarity measure is defined as

$$s_{ij} = \frac{\sum_{k=1}^{m} w_k \, \delta_{ijk} \, s_{ijk}}{\sum_{k=1}^{m} w_k \, \delta_{ijk}}, \qquad (2.1)$$

where s_{ijk} is the similarity between the ith and jth objects based on the kth feature f_k only, and $\delta_{ijk} = 1$, if the objects can legitimately be compared, and zero otherwise, as, for example, in the case of missing values. For dichotomous variables, $\delta_{ijk} = 0$ if $x_{ik} = x_{jk} = 0$ and $\delta_{ijk} = 1$ otherwise. The strength of feature contributions is determined by the weights w_k, which are omitted here as all $w_k = 1$. The similarity

s_{ijk}, $i,j=1,\ldots,n$ and $k=1,\ldots,m$ becomes then $s_{ijk}=1-\frac{|x_{ik}-x_{jk}|}{r_k}$ if f_k is quantitative, $s_{ijk}=\mathcal{I}\left((x_{ik}=x_{jk})=1\right)$ if f_k is dichotomous, $s_{ijk}=\mathcal{I}\left(x_{ik}=x_{jk}\right)$ if f_k is categorical and $s_{ijk}=1-g(\frac{|x_{ik}-x_{jk}|}{r_k})$, where r_k is the range of f_k and g is a chosen monotonic transformation if f_k is ordinal. The Gower's dissimilarity between the ith and jth objects is defined as $d_{ij}=(1-s_{ij})^{1/2}$.

Polygon. The data consist of two classes of randomly generated polygons, convex quadrilaterals and irregular heptagons [22, 24]. Each class consists of 2000 examples. First, the polygons are scaled such that their total contour lengths are equal. Next, the modified Hausdorff distances [7] are computed between their corners. Let A and B be two polygons. The modified Hausdorff distance is defined as
$d_{MH}(A,B)=\max\{d_{avr}^{\triangleright}(A,B),d_{avr}^{\triangleright}(B,A)\}$,
where $d_{avr}^{\triangleright}(A,B)=\frac{1}{|A|}\sum_{a\in A}\min_{b\in B}d(a,b)$,
and it is evaluated at the polygon corners a and b. This measure is nonmetric [7, 22].

ProDom. ProDom is a comprehensive set of protein domain families [5]. A subset of 2604 protein domain sequences from the ProDom set [5] was selected by Roth et al. [28]. These examples are chosen based on a high similarity to at least one sequence contained in the first four folds of the Structural Classification of Proteins (SCOP) database. The pairwise structural alignments are computed by Roth using the FASTA software [12]. Each SCOP sequence belongs to a group as labeled by the experts [21]. We use the same set in our investigations. Originally, a structural symmetric similarity $S=(s_{ij})$ is derived first. Then, the nonmetric dissimilarities are obtained by $d_{ij}=(s_{ii}+s_{jj}-2s_{ij})^{1/2}$ for $i\neq j$.

Tumor-mucosa. The data consist of the autofluorescence spectra acquired from healthy and diseased mucosa in the oral cavity [29]. The spectra were collected from 97 volunteers with no clinically observable lesions of the oral mucosa and 137 patients having lesions in oral cavity. The measurements were taken using the excitation wavelength of 365 nm. After preprocessing [30], each spectrum consists of 199 bins. In total, 856 spectra representing healthy tissue and 132 spectra representing diseased tissue were obtained. The spectra are normalized to a unit area. Here, we choose the nonmetric $l_{0.8}$-distances (l_p-distance is $d_p(\boldsymbol{x},\boldsymbol{y})=[\sum_k(x_k-y_k)^p]^{1/p}$) between the first-order Gaussian-smoothed ($\sigma=3$ samples) derivatives of the spectra.[1] The zero-crossings of the derivatives indicate the peaks and valleys of the spectra, so they are informative. Moreover, the distances between smoothed derivatives contain some information of the order of bins. In this way, the property of a continuity of a spectrum is somewhat taken into account. This data set suffers from outliers, which are preserved here as we intend to illustrate their influence on the complexity.

[1] l_p-distances, $p\leq 1$, may be useful for problems characterized by the presence of a scattered and very heterogeneous class, such as the class of diseased people here. The effect of large absolute differences is diminished by $p<1$. Indeed, this measure was found advantageous in our earlier experiments [22].

2.3 Criteria for Sampling Density

Consider an $n \times n$ dissimilarity matrix $D(R, R)$, where $R = \{p_1, p_2, \ldots, p_n\}$ is a representation set. In general, R may be a subset of a larger learning set T, but we assume here that $R = T$. Every object p_i is then represented by a vector of dissimilarities $D(p_i, R)$, $i = 1, 2, \ldots, n$, to the objects from R. The research question to be addressed is whether n, the cardinality of R, is sufficiently large for capturing the variability in the data or, in other words, whether it is to be expected that only little new information can be gained by increasing the number of representation objects. This can be further rephrased as judging whether new objects can be expressed in terms of the ones already present in R or not. Given the dissimilarity representations, some criteria are proposed to judge its sampling sufficiency, and their usefulness is experimentally evaluated on the data sets introduced in section 2. We focus here on a set of unlabeled objects forming a single class.

Some possible statistics that can be used are based on the compactness hypothesis [1, 8, 9], which was introduced in section 2.1. As it states that similar objects are also similar (close) in their representation, it constrains the dissimilarity measure d in the following way: d has to be such that $d(x, y)$ is small if the objects x and y are very similar; i.e., it should be much smaller for similar objects than for objects that are very different.

Assume that the dissimilarity measure d is definite, i.e., $d(x, y) = 0$ iff the objects x and y are identical. If the objects are identical, they belong to the same class. This reasoning can be extended by assuming that all objects z for which $d(x, z) < \varepsilon$, and the positive ε is sufficiently small, are so similar to x that they belong to the same class as x. Consequently, the dissimilarities of x and z to the representation objects should be close (or positively correlated, in fact). This means that $d(x, p_i) \approx d(z, p_i)$, implying that the representations $d(x, R)$ and $d(z, R)$ are also close. We conclude that for dissimilarity representations that satisfy the above continuity, the reverse compactness hypothesis holds, as objects that are similar in their representations are also similar in reality. Consequently, they belong to the same class.

A representation set R can be judged as sufficiently large if an arbitrary new object of the same class is not significantly different from all other objects of that class in the data set. This can be expected if R already contains many objects that are very similar, i.e., if they have a small dissimilarity to at least one other object. All the criteria studied below are based, in one way or another, on this observation. In pathological cases, the data set may contain just an optimal set of objects, but if there are no additional objects to validate this, it has to be considered as being too small.

We will illustrate the performance of our criteria on an artificial example and present also the results for some real data sets. The artificial example is chosen to be the $l_{0.8}$-distance representation between n normally distributed points in a k-dimensional vector space $\mathbb{R}^k$. Both n and k vary between 5 and 500. If $n < k$, then the generated vectors lie in an $(n-1)$-dimensional subspace, resulting in an undersampled and difficult problem. If $n \gg k$, then the data set may be judged as sufficiently sampled. Large values of k lead to difficult (complex) problems as they demand a large data cardinality n. The results are averaged over 20 experiments, each time based on a new, randomly generated data set. The criteria are presented and discussed below.

2.3.1 Specification of the Criteria

Sampling criteria for dissimilarity representations are directly or indirectly addressed in three different ways: by the dissimilarity values as given; in dissimilarity vector spaces, in which every dimension is defined by a dissimilarity to a representation object and in embedded vector spaces, which are determined such that the original dissimilarities are preserved; see [22, 23, 25] for more details. Each criterion is introduced and illustrated by a separate figure, e.g., Figure 2.2 refers to the first criterion. The results for artificially generated Gaussian data sets with the dimensionality k varying from 5 to 500 represented by a Euclidean distance matrix D are always shown on the top. Then, the results of other statistics are presented as applied to the six real data sets.

Skewness. This is a statistic that evaluates the dissimilarity values directly. A new object added to a set of objects that is still insufficiently well sampled will generate many large dissimilarities and just a few small ones. As a result, for unsatisfactory sampled data, the distribution of dissimilarities will peak for small values and will show a long tail in the direction of large dissimilarities. After the set becomes "saturated," however, adding new objects will cause the appearance of more and more small dissimilarities. Consequently, the skewness will grow with the increase of $|R|$. The value to which it grows depends on the problem.

Let the variable d denote now the dissimilarity value between two arbitrary objects. In practice the off-diagonal values d_{ij} from the dissimilarity matrix $D = (d_{ij})$ are used for his purpose. As a criterion, the skewness of the distribution of the dissimilarities d is considered as

$$J_{sk} = E\left[\frac{d - E[d]}{\sqrt{E[d - E[d]]^2}}\right]^3, \tag{2.2}$$

where $E[\cdot]$ denotes the expectation. In Figure 2.2, top, the skewness of the Gaussian sets are shown. The cardinalities of small representation sets appear to be insufficient to represent the problem well, as it can be concluded from the noisy behavior of the graphs in that area. For large representation sets, the curves corresponding to the Gaussian samples of the chosen dimensionality "asymptotically" grow to some values of J_{sk}. The final values may be reached earlier for simpler problems in low dimensions, like $k = 5$ or 10. In general, the skewness curves for various k correspond to the expected pattern that the simplest problems (in low-dimensional spaces) reach the highest skewness values, whereas the most difficult problems are characterized by the smallest skewness values.

Mean rank. An element d_{ij} represents the dissimilarity between the objects p_i and p_j. The minimum of d_{ij} over all indices j points to the nearest neighbor of

p_i, say, p_z if $z = \text{argmin}_{j \neq i}(d_{ij})$. So, in the representation set R, p_z is judged as the most similar to p_i. We now propose that a representation $D(p_i, R)$ describes the object p_i well if the representation of p_z, i.e., $D(p_z, R)$, is close to $D(p_i, R)$ in the dissimilarity space $D(\cdot, R)$. This can be measured by ordering the neighbors of the vectors $D(p_i, R)$ and determining the rank number r_i^{NN} of $D(p_z, R)$ in the list of neighbors of $D(p_i, R)$. By this we compare the nearest neighbor as found in the original dissimilarities with the neighbors in the dissimilarity space. For a well-described representation, the mean relative rank

$$J_{mr} = \frac{1}{n} \sum_{i=1}^{n} r_i^{NN} - 1 \tag{2.3}$$

is expected to be close to 0. In Figure 2.3, top, the results for the Gaussian example are shown. It can be concluded that the sizes of the representation set R larger than 100 are sufficient for Gaussian samples in 5 or in 10 dimensions.

PCA (principal component analysis) dimensionality. A sufficiently large representation set R tends to contain some objects that are very similar to each other. This means that their representations, the vectors of dissimilarities to R, are very similar. This suggests that the rank of D should be smaller than $|R|$, i.e., rank$(D) < n$. In practice, this will not be true if the objects are not alike. A more robust criterion, therefore, may be based on the principal component analysis applied to the dissimilarity matrix D. Basically, the set is sufficiently sampled if n_α, the number of eigenvectors of D for which the sum of the corresponding eigenvalues equals a fixed fraction α, such as 0.95 of the total sum of eigenvalues (hence α is the explained fraction of the variance), is small in comparison to n. So, for well-represented sets, the ratio of n_α / n is expected to be smaller than some small constant (the faster the criterion curve drops with a growing R, the smaller intrinsic dimensionality of the dissimilarity space representation). Our criterion is then defined as

$$J_{\text{pca},\alpha} = \frac{n_\alpha}{n}, \tag{2.4}$$

with n_α such that $\alpha = \sum_{i=1}^{n_\alpha} \lambda_i / \sum_{i=1}^{n} \lambda_i$. There is usually no integer n_α for which the above holds exactly, so it would be found by interpolation. Note that this criterion relies on an intrinsic dimensionality[2] in a dissimilarity space $D(\cdot, R)$.

[2] If a certain phenomenon can be described (or if it is generated) by m independent variables, then its intrinsic dimensionality is m. In practice, however, due to noise and imprecision in measurements or some other uncontrolled factors, such a phenomenon may seem to be generated by more variables. If all these factors are not too dominant such that they completely disturb the original phenomenon, one should be able to rediscover the proper number of significant variables. Hence, the intrinsic dimensionality is the minimum number of variables that explains the phenomenon in a satisfactory way. In pattern recognition, one usually discusses the intrinsic dimensionality with respect to a collection of data vectors in the feature space. Then, for classification, the intrinsic dimensionality can be defined as the minimum number of features needed to obtain a similar classification performance as by using all features. In a geometrical sense, the intrinsic dimensionality can be defined as the

In the experiments, in Figure 2.4, top, the value of $J_{\text{pca},0.95}$ is shown for the artificial Gaussian example as a function of $|R|$. The Gaussian data are studied as generated in spaces of a growing dimensionality k. It can be concluded that the data sets consisting of more than 100 objects may be sufficiently well sampled for small dimensionalities such as $k=5$ or $k=10$ as just a small fraction of the eigenvectors is needed (about 10% or less). On the other hand, the considered number of objects is too small for the Gaussian sets of a larger dimensionality. These generate problems of a too high complexity for the given data-set size.

Correlation. Correlations between objects in a dissimilarity space are also studied. Similar objects show similar dissimilarities to other objects and, thereby, are positively correlated. As a consequence, the ratio of the average of positive correlations $\rho_{+}(D(p_i,R),D(p_j,R))$ to the average of absolute values of negative correlations $\rho_{-}(D(p_i,R),D(p_j,R))$, given as

$$J_\rho = \frac{\frac{1}{n^2-n}\sum_{i,j\neq i}^{n}\rho_{+}(D(p_i,R),D(p_j,R))}{1+\frac{1}{n^2-n}\sum_{i,j\neq i}^{n}|\rho_{-}(D(p_i,R),D(p_j,R))|} \tag{2.5}$$

will increase for large sample sizes. The constant added in the denominator prevents J_ρ from becoming very large if only small negative correlations appear. For a well-sampled representation set, J_ρ will be large and it will increase only slightly when new objects are added (new objects should not significantly influence the averages of either positive or negative correlations). Figure 2.5, top, shows that this criterion works well for the artificial Gaussian example. For the lower dimensional data sets (apparently less complex) J_ρ reaches higher values and exhibits a flattening behavior for sets consisting of at least 100 objects.

Intrinsic embedding dimensionality. For the study of dissimilarity representations, one may perform dimensionality reduction of a dissimilarity space (as the PCA criterion, described above, does) or choose an embedding method. Consequently, the judgment about whether R is sufficiently sampled relies on the estimate of the intrinsic dimensionality of an underlying vector space determined such that the original dissimilarities are preserved. This can be achieved by a linear embedding of the original objects (provided that D is symmetric) into a (pseudo-)Euclidean space. A pseudo-Euclidean space[3] is needed if D does not exhibit the Euclidean behavior, as, for example, the l_1-distance or max-norm distance measures do [22, 23]. In this way, a vector space is found in spite of the fact that one starts from a dissimilarity matrix D. The representation X of $m\leq n$ dimensions is determined such that it is centered at the origin and the derived features are uncorrelated [13, 26].

dimension of a manifold that approximately (due to noise) embeds the data. In practice, the estimated intrinsic dimensionality of a sample depends on the chosen criterion. Thereby, it is relative for the task.

[3] A pseudo-Euclidean space $\mathcal{E} := \mathbb{R}^{(p,q)}$ is a $(p+q)$-dimensional nondegenerate indefinite inner product space such that the inner product $\langle\cdot,\cdot\rangle_{\mathcal{E}}$ is positive definite (pd) 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 $\mathcal{J}_{pq} = \text{diag}\,(I_{p\times p}; -I_{q\times q})$ and I is the identity matrix. Consequently, the square pseudo-Euclidean distance is $d^2_{\mathcal{E}}(\boldsymbol{x},\boldsymbol{y}) = \langle \boldsymbol{x}-\boldsymbol{y},\boldsymbol{x}-\boldsymbol{y}\rangle_{\mathcal{E}} = d^2_{\mathbb{R}^p}(\boldsymbol{x},\boldsymbol{y}) - d^2_{\mathbb{R}^q}(\boldsymbol{x},\boldsymbol{y})$.

The embedding relies on linear operations. The inner product (Gram) matrix G of the underlying configuration X is expressed by the square dissimilarities $D^{*2} = (d_{ij}^2)$ as $G = -\frac{1}{2}JD^{*2}J$, where $J = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is the centering matrix [13, 22, 26]. X is determined by the eigen-decomposition of $G = Q\Lambda Q^T = Q|\Lambda|^{1/2}\text{diag}(\mathcal{J}_{p'q'};0)\,|\Lambda|^{1/2}Q^T$, where $\mathcal{J}_{p'q'} = (I_{p'\times p'};-I_{q'\times q'})$ and I is the identity matrix, $|\Lambda|$ 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. The sought configuration is first represented in $\mathbb{R}^k$, $k = p'+q'$, as $Q_k|\Lambda_k|^{1/2}$. Because only some eigenvalues are large (in magnitude), the remaining ones can be disregarded as noninformative. This corresponds to the determination of intrinsic dimensionality. The final representation $X = Q_m\,|\Lambda_m|^{1/2}$, $m = p+q < k$, is defined by the largest p positive and the smallest q negative eigenvalues, since the features are uncorrelated.

This means that the number of dominant eigenvalues (describing the variances) should reveal the intrinsic dimensionality (small variances are expected to show just noise). (Note, however, that when all variances are similar, the intrinsic dimensionality is approximately n.) Let n_α^{emb} be the number of significant variances for which the sum of the corresponding magnitudes equals a specified fraction α, such as 0.95, of the total sum. Because n_α^{emb} determines the intrinsic dimensionality, the following criterion is proposed:

$$J_{emb,\alpha} = \frac{n_\alpha^{emb}}{n}. \tag{2.6}$$

For low intrinsic dimensionalities, smaller representation sets are needed to describe the data characteristics. Figure 2.6, top, presents the behavior of this criterion as a function of $|R|$ for the Gaussian data sets. The criterion curves clearly reveal different intrinsic embedding dimensionalities. If R is sufficiently large, then the intrinsic dimensionality estimate remains constant. Because the number of objects is growing, the criterion should then decrease and reach a relatively constant small value in the end (for very large sets). From the plot it can be concluded that data sets with more than 100 objects are satisfactorily sampled for Gaussian data of an originally low dimensionality such as $k \leq 20$. In other cases, the data set is too complex.

Compactness. As mentioned above, a symmetric distance matrix D can be embedded in a Euclidean or a pseudo-Euclidean space $\mathcal{E}$, depending on the Euclidean behavior of D. When the representation set is sufficiently large, the intrinsic embedding dimensionality is expected to remain constant during a further enlargement. Consequently, the mean of the data should remain approximately the same and the average distance to this mean should decrease (as new objects do not surprise anymore) or be constant. The larger the average distance, the less compact the class is, requiring more samples for its description. Therefore, a simple compactness criterion can be investigated. It is estimated in the leave-one-out approach as the average square distance to the mean vector in the embedded space $\mathcal{E}$:

$$J_{comp} = \frac{1}{n^2-n}\sum_{j=1}^{n}\sum_{i\neq j} d_{\mathcal{E}}^2(\boldsymbol{x}_i^{-j}, \boldsymbol{m}^{-j}), \tag{2.7}$$

where $\boldsymbol{x}_i^{-j}$ is a vector representation of the ith object in the pseudo-Euclidean space found by $D(R^{-j}, R^{-j})$ and R^{-j} is a representation set of all the objects, except the jth one. $\boldsymbol{m}^{-j}$ is the mean of such a configuration. This can be computed from the dissimilarities directly without the necessity of finding the embedded configuration; see [26]. Figure 2.7, top, shows the behavior of this criterion, clearly indicating a high degree of compactness of the low-dimensional Gaussian data. The case of $k = 500$ is judged as not having a very compact description.

Gaussian intrinsic dimensionality. If the data points come from a spherical normal distribution in an m-dimensional Euclidean space, then m can be estimated from the χ_m^2 distributed variable d^2 denoting the pairwise square Euclidean distances as $m = 2\frac{(E[d^2])^2}{E[d^4]-(E[d^2])^2}$, where $E[\cdot]$ denotes the expectation; see [22]. If the data points come from any other normal distribution, still some sort of an intrinsic dimensionality estimate can be found by the above formula. The judgement will be influenced by the largest variances in the data. Basically, the volume of the hyper-ellipsoidal normally distributed data is captured in the given distances. They are then treated as if computed from a spherically symmetric Gaussian distribution. Hence, the derived intrinsic dimensionality will reflect the dimensionality of a space to which the original data sample is made to fit isotropically (in simple words, one can imagine the original hyper-ellipsoidal Gaussian sample reshaped in space and "squeezed" in the dimensions to make it the largest hyper-spherical Gaussian sample, the dimensionality of the latter is then estimated). Since the above formula makes use of the distances only, it can be applied to any dissimilarity measure. The criterion is then defined as

$$J_{Gid} = 2\,\frac{(E[d^2])^2}{E[d^4] - (E[d^2])^2}, \tag{2.8}$$

where d^2 is realized by the off-diagonal square dissimilarity values d_{ij}^2.

Boundary descriptor. A class descriptor (a one-class classifier) in a dissimilarity space was proposed in Pekalska et al. [27]. It is designed as a hyperplane H: $\boldsymbol{w}^T D(x, R) = \rho$ in a dissimilarity space that bounds the target data from above (it assumed that d is bounded) and for which some particular distance to the origin is minimized. Non-negative dissimilarities impose both $\rho \geq 0$ and $w_i \geq 0$. This is achieved by minimizing $\rho/||\boldsymbol{w}||_1$, which is the max-norm distance of the hyperplane H to the origin in the dissimilarity space. Therefore, H can be determined by minimizing $\rho - ||\boldsymbol{w}||_1$. Normalizing such that $||\boldsymbol{w}||_1 = 1$ (to avoid any arbitrary scaling of $\boldsymbol{w}$), H is found by the optimization of ρ only. A (target) class is then characterized by a linear proximity function on dissimilarities with the weights $\boldsymbol{w}$ and the threshold ρ. It is defined as $\mathcal{I}(\sum_{w_j \neq 0} w_j D(x, p_j) \leq \rho)$, where $\mathcal{I}$ is the indentificator (characteristic) function (it takes the value of 1 if the condition is true and zero otherwise), w_j are found as the solution to a soft-margin linear programming formulation (the hard-margin case is then straightforward) with $\nu \in (0, 1]$ being the upper bound on the target rejection fraction in training [27]:

$$\begin{aligned}
&\text{Minimize } \rho + \tfrac{1}{\nu\, n} \textstyle\sum_{i=1}^{n} \xi_i \\
&\text{such that, } \boldsymbol{w}^T D(p_i, R) \le \rho + \xi_i,\ \ \textstyle\sum_j w_j = 1,\ w_j \ge 0, \\
&\qquad\qquad \rho \ge 0,\ \xi_i \ge 0, \quad i = 1, \ldots, n.
\end{aligned} \tag{2.9}$$

As a result, a sparse solution is obtained. This means that many weights w_i become zero and only some are positive. The objects $R_{so} \subseteq R$ for which the corresponding weights are positive are called *support objects* (SO). Our criterion then becomes the number of support objects:

$$J_{so} = |R_{so}|. \tag{2.10}$$

In the experiments we suffered from numerical problems for large representation set sizes. For that reason, the solutions were found for all but one of the dimensionalities, i.e., except for the case $|R|$ = 500.

2.3.2 Discussion on Sampling Density Experiments

While studying the results presented in Figures 2.2 to 2.8, one should recall that the height of the curve is a measure of the complexity and that a flat curve may indicate that the given data set is sufficiently sampled. For the skewness, mean rank and correlation statistics, it holds that lower values are related to a higher complexity. For the other criteria, it is the other way around: lower values are related to a lower complexity. An exception is the compactness, as defined here, since its behavior is scale dependent.

For all data sets and all criteria, it can be observed that the complexity of the original data set D (continuous lines) increases by the square root transformation (dashed lines) and decreases by the quadratic transformation (dotted lines). This implies that the $D^{*0.5}$ data sets tend to be undersampled in most cases. For the original data sets, this just holds for some of the classes of the Digits-all, the Heart, and the ProDom problems. The diseased class of the Tumor-mucosa problem shows a very irregular behavior, due to some large outliers. This is in fact useful as a number of very different outliers is a sign of undersampling. Most D^{*2} data sets may be judged as well sampled. Exceptions are the Heart data set and, again, the diseased class of the Tumor-mucosa problem.

It is interesting to observe the differences between various data sets, e.g., that the curves of the boundary descriptor sometimes start with a linear increase or that the correlation curve is usually an increasing function with some exceptions in the case of the Polygon data. The high increase of the PCA dimensionality criterion for the artificial Gaussian data set (Fig. 2.4) and for a large dimensionality k can nowhere be observed, with an exception of the Heart data set. A global comparison of all figures shows that the characteristics of high-dimensional Gaussian distributions cannot be found in real-world problems. This may indicate that various methods for data analysis and classification, based on the Gaussian assumption, need to be either improved before they can be used in practice or avoided.

In general, the flattened behavior of a criterion curve implies a sufficient sampling. All criteria, except for mean rank, are very sensitive to data modifications, indicating that quadratic transformations decrease the original data-set complexity, whereas square root transformation increase it. Concerning specific approaches, the following can be summarized:

- Skewness is informative to judge the distribution of dissimilarities. Negative skewness denotes a tail of small dissimilarities, whereas positive skewness describes a tail of large dissimilarities. Large positive values indicate outliers in the class (the Tumor-mucosa data), whereas large negative values indicate heterogeneous characteristics of the class (the Heart data) or a class of possible clusters having various spreads (the ProDom data). Skewness can be noisy for very small sample sizes.
- Mean rank is a criterion judging the consistency between the nearest neighbors directly applied to the given dissimilarities and the nearest neighbor in a dissimilarity space. For an increasing number of objects, this should approach zero. As original nearest neighbor relations do not change after nondecreasing transformations (although they are affected in a dissimilarity space), this criterion is not very indicative for such modifications. Except for the artificial Gaussian examples, the curves exhibit a similar behavior.
- PCA dimensionality describes the fraction of significant eigenvalues in a dissimilarity space of a growing dimensionality. If the data set is "saturated," then the criterion curve approaches a value close to zero since the intrinsic dimensionality should stay constant. If the criterion does not approach zero, the problem is characterized by many relatively similar eigenvalues, hence many similar intrinsic variables. In such cases, the problem is judged as complex, for instance for the Heart and the Digits-all problems.
- The correlation criterion indicates the amount of positive correlations versus negative correlations in a dissimilarity space. Positive values > 0.5 may suggest the presence of outliers in the data as observed in the case of the ProDom and Tumor-mucosa problems.
- Intrinsic embedding dimensionality is judged by the fraction of dominant dimensions determined by the number of dominant eigenvalues in a linear embedding. In contrast to the PCA dimensionality, it is not likely to observe the criterion curve approaching zero. Large dissimilarities determine the embedded space and considerably affect the presence of large eigenvalues. Therefore, the criterion curve may be close to zero if many eigenvalues tend to be so or if there are some notable outliers (as the diseased class of the Tumor-mucosa problem). In this case, a flat behavior of the curve may give evidence of an acceptable sampling. However, the larger the final value of the criterion curve, the more complex the class description (there is a larger variability in the class).
- Compactness indicates how compact a set of objects is as judged by the distances to the mean in an embedded space. In this case, the flattened behavior of the curve is not very indicative, as all our problems for small sample sizes would be judged as well sampled. What is more important is the value that the criterion curve attains – the smaller the value the more compact the description.
- Similarly to the criterion above, the smaller the final value to which the Gaussian intrinsic dimensionality criterion curve converges, the less complex the problem.

- The boundary descriptor indicates the number of boundary objects necessary to characterize the class. A large number of objects with respect to $|R|$ indicates a complex problem, as, for example, the Heart data set is. The criterion curves may be noisy for small samples, as observed for the ProDom and Tumor-mucosa cases, possibly indicating the presence of outliers.

In brief, the most indicative and insightful criteria are skewness, PCA dimensionality, correlation, and boundary description. Intrinsic embedding dimensionality may be also informative; however, a good understanding of the embedding procedure is needed to judge it well. The remaining criteria have less impact, but they still bring some additional information.

2.4 Classification Experiments

2.4.1 Introduction

Complexity should be studied with respect to a given task such as class description, clustering, or classification. Hence, the complexity of the data set should describe some of its characteristics or of an assumed model, relative to the chosen representation. In the previous section, some criteria for the complexity of unlabeled data (data geometry and class descriptions) were studied. This section is concerned with supervised learning.

As data-set complexity is a different issue than class overlap, its relation to classifier performance is not straightforward. We argued in the introduction that more complex problems may need more complex tools, or more training samples, which will be our focus here. Therefore, we will study the influence of data-set complexity on the classifier performance. The original representation will be transformed by the same power transformations as in section 2.3. As has been already observed, D^{*2} representations decrease, while $D^{*0.5}$ representations increase the data set complexity of the individual classes.

As we indicated in the chapter introduction, an intrinsic problem complexity, as such, does not exist. Its complexity is entirely determined by the representation and observed through the data set. If the data-set complexity is decreased by some transformation simplifying the problem, as a result simpler classifiers may be used. Note that no monotonic transformation of the data can either reduce or increase the intrinsic class overlap. Transformations are applied to enable one to train classifiers that reach a performance, which is closer to this intrinsic overlap. If the problem becomes less complex, smaller training sets probably will be sufficient. If it was originally abundant, the decreased complexity may yield a better classification performance. If the training set size was initially sufficient, the decreased complexity may decrease the performance (due to perceived higher class overlap). An increased problem complexity may open a way for constructing more complex classifiers. If the sample size permits, these classifiers will reach an increased performance. If the sample size is insufficient, such classifiers will be overtrained, resulting in a decrease of the performance.

In addition to these effects, there is a direct relation between data-set complexity and a desirable size of the representation set. Remember that this desirable size is indicated by the stability of the measures or the observed asymptotic behavior of the criteria identified to be useful in the preceding analysis. More complex problems need a larger representation set. The other way around also holds: a larger representation set used for the description may indicate more complex aspects of the problem.

The above effects will be illustrated by a set of classification experiments. Assume that a training set of N examples is provided. First, a suitable representation set $R \subset T$ has to be determined. We will proceed in two ways, starting from a full representation $D(T,T)$. The representation set will be chosen either as a condensed set found by the editing-and-condensing [condensed nearest neighbor (CNN)] procedure [6] or as the set of support objects determined in the process of constructing a sparse linear programming classifier (LPC). In the resulting dissimilarity space, a Fisher classifier on $D(T,R)$ is trained.

2.4.2 Classifiers

The following classifiers are used in our experiments:

1-Nearest neighbor rule (1-NN). This classifier operates directly on the dissimilarities computed for a test object. It assigns a test object to the class of the training object that is the most similar as judged by the smallest dissimilarity. Because no training is required, the values in $D(T,T)$ are not used for the construction of this rule.

k-Nearest neighbor rule (k-NN). Here, the test object is assigned to the most frequent class in the set of the k-nearest neighbors. The value of k is optimized over the original representation $D(T,T)$ using a leave-one-out procedure. In this way, the training set T is somewhat used in the learning process.

Editing and condensing (CNN). An editing and condensing algorithm is applied to the entire dissimilarity representation $D(T,T)$, resulting in a condensed set (CS) R_{CS}. Editing takes care that the noisy objects are first removed so that the prototypes can be chosen to guarantee a good performance of the 1-NN rule, which is used afterward.

Linear programming classifier (LPC). By training a properly formulated linear classifier $f(D(x,T)) = \sum_{j=1}^{N} w_j \, d(x,p_j) + w_0 = \boldsymbol{w}^T D(x,R) + w_0$ in a dissimilarity space $D(T,T)$, one may select objects from T necessary for the construction of the classifier. The separating hyperplane is obtained by solving a linear programming problem, where a sparse solution on R is imposed by minimizing the l_1-norm of the weight vector $\boldsymbol{w}$, $||\boldsymbol{w}||_1 = \sum_{j=1}^{r} |w_j|$; see [4, 11] on the sparseness issues. As a result, only some weights become nonzero. The corresponding objects define the representation set.

A flexible formulation of a classification problem is proposed in Graepel et al. [15]. The problem is to minimize $||\boldsymbol{w}||_1 - \mu \, \rho$, which means that the margin ρ becomes a variable of the optimization problem. To formulate such a minimization task properly, the absolute values $|w_j|$ should be eliminated from the objective function. Therefore, the weights w_j are expressed by nonnegative variables α_j and β_j as $w_j = \alpha_j - \beta_j$. (When the pairs (α_j, β_j) are determined, then at least one of them is zero.) Nonnegative slack variables ξ_i, accounting for possible classification errors are additionally introduced. Let $y_i = +1/-1$ indicate the class membership. By imposing $||\boldsymbol{w}||_1$ to be constant, the minimization problem for $x_i \in T$ then becomes

$$\begin{aligned} &\text{Minimize } \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \xi_i - \mu\,\rho \\ &\text{such that, } \textstyle\sum_{i=1}^{N}(\alpha_i + \beta_i) = 1 \\ &\qquad\qquad y_i\, f(D(x_i, T)) \geq 1 - \xi_i,\ i = 1, \ldots, N \\ &\qquad\qquad \xi_i,\ \alpha_i,\ \beta_i,\ \rho \geq 0. \end{aligned} \tag{2.11}$$

A sparse solution $\boldsymbol{w}$ is obtained, which means that important objects are selected (by nonzero weights) from the training set T, resulting in a representation set R_{so}. The solution depends on the choice of the parameter $\mu \in (0, 1)$, which is related to a possible class overlap [15]. To select it automatically, the following values are found (as rough estimates based on the 1-NN error computed over a number of representations $D(T, T)$ for various sizes of T). These are 0.2 for the Heart data, 0.1 for the Digits-all and Tumor-mucosa data, and 0.05 for the remaining sets.

The selection of objects described above is similar to the selection of features by linear programming in a standard classification task; see [3, 4] . The important point to realize is that we do not have control over the number of selected support objects. This can be somewhat influenced by varying the constant μ (hence influencing the trade-off between the classifier norm and the training classification errors).

Fisher classifier (FC). This linear classifier minimizes the mean square error on the training set $D(T, R)$ with respect to the desired labels $y_i = +1/-1$. It finds the minimal mean square error solution of $\sum_{j=1}^{N} w_j\, d(x_i, x_j) + w_0 = y_i$. Note that the common opinion that this classifier assumes Gaussian class densities is wrong. The truth is that in the case of Gaussian densities with equal covariance matrices, the corresponding Bayes classifier is found (in the case of equal class priors). The Fisher classifier, however, is neither based on a density assumption nor does it try to minimize the probability of misclassification in a Bayesian sense. It follows a mean square error approach. As a consequence, it does suffer from multimodality in class distributions.

Multiclass problems are solved for the LPC and the FC in a one-against-all-others strategy using the classifier conditional posterior probability estimates [10]. Objects are assigned to the class that receives the highest confidence as the “one” in this one-against-all-others scheme.

2.4.3 Discussion on the Classification Experiments

The classification results for the six data sets are presented in Figures 2.9 to 2.14. In each figure, the first plot shows the results of the LPC as a function of the training set size. The averaged classification errors for the three modifications of the dissimilarity measures are presented. For comparison also the results of the 1-NN, the k-NN, and the CNN rules are shown. Note that these are independent of the nondecreasing transformations. The CNN curves are often outside the shown interval.

The resulting reduced object sets selected by the CNN, the condensed set CS, are used as a representation set R. Then, the Fisher classifier FC is constructed on the dissimilarity representation $D(T,R)$. This will be denoted as FC-CS. The averaged errors of this classifier are, again, together with the results for the 1-NN, the k-NN, and the CNN rules (these are the same as in the first graph), shown in the second plot. All experiments are averaged over 30 repetitions in which independent training and test sets are generated from the original set of objects.

The third plot illustrates the sizes of the reduced training sets found by the LPC and the CNN. For most data sets, the CNN reduces the training set further than the LPC. The resulting sample sizes of the CNN set are approximately a linear function of the training size $|T|$. In all cases, the sets of support objects found by the LPC are for the D^{*2} representations smaller than for the original one, D, which are, in turn, smaller than for the $D^{*0.5}$ representations. This is in agreement with our expectation (see section 2.2) and with the results of section 2.3, that the data-set complexity of D^{*2} is lower and the data-set complexity of $D^{*0.5}$ is higher than it is of D.

The first two plots can be considered as learning curves (note, however, that the determined representation set R increases with a growing training set T). The dissimilarity-based classifiers, the LPC and the FC-CS, perform globally better than the nearest neighbor rules, which is in agreement with our earlier findings; see [22, 23, 25]. The LPC and the FC-CS are comparable. The LPC is often better than the FC-CS for smaller sample sizes, whereas the FC-CS is sometimes somewhat better than the LPC for larger sample sizes. This might be understood from the fact that the LPC, like the support vector machine, focuses on the decision boundary, whereas the FC uses the information of all objects in the training set. Where this is profitable, the FC will reach a higher accuracy.

Learning curves usually show a monotonic decreasing behavior. For simple data sets they will decrease fast, whereas for complex data sets they will decrease slowly. The complexity is understood here in relation to single class descriptions and to the intricateness of the decision boundary between the classes (hence their geometrical position in a dissimilarity space). The asymptotic behavior will be similar if a more complex representation does not reveal any additional details that are useful for the class separation. If it does, however, a more complex representation will show a higher asymptotic accuracy, provided that the classifier is able to use the extra information.

Following this reasoning, it is to be expected that the learning curves for D^{*2} representations decrease fast, but may have worse asymptotic values. This appears to be true with a few exceptions. For the Tumor-mucosa problem (Fig. 2.15), the expectation is definitely wrong. This is caused by the outliers as the quadratic transformation strengthens their influence. The global behavior, expected from this transformation, is overshadowed by a few outliers that are not representative for the problem. A second exception can be observed in the Digits-all results (see Fig. 2.11), especially for the FC. In this multiclass problem the use of the FC suffers from the multimodality caused by the one-against-all-others strategy.

The learning curves for the $D^{*0.5}$ data sets change in most cases, as expected, slower than those for the original data sets. The FC-CS for the Digits-all case (Fig. 2.11), is again an exception. In some cases, these two learning curves are almost on top of each other; in some other cases, they are very different, as for the FC-CS and the ProDom data set (Fig. 2.14). This may indicate that the data set complexity increased by the square root transformation is really significant.

There are a few situations for which crossing points of the learning curves can be observed after which a more complex representation ($D^{*0.5}$ or D) enables the classifiers to reach a higher performance than a simpler one (D or D^{*2}, respectively) due to a sufficient sample size. Examples are the LPC classification of the Digits-all data (Fig. 2.11) and the Polygon data (Fig. 2.13).

Finally, we observe that for the undersampled Heart data set (see section 2.3), the k-NN does relatively very well. This is the only case where the dissimilarity-based classifiers LPC and FC-CS perform worse than the straightforward use of the nearest neighbor rule.

2.5 Discussion

A real-world pattern recognition problem may have an inherent complexity: objects of different classes may be similar; classes may consist of dissimilar subgroups; and essential class differences may be hidden, distributed over various attributes or may be context dependent. All that matters, however, is how the problem is represented using object models, features, or dissimilarity measures. The problem has to be solved from a given representation, and its complexity should be judged from that. It is the representation that is explicitly available, and it may be such that seemingly simple problems are shown as complex or the other way around.

In this chapter we argued that the complexity of a recognition problem is determined by the given representation and observed through a data set and may be judged from a sample size analysis. If for a given representation, a problem is sampled sufficiently well, then it is simpler than for a representation for which it appears to be too low. In section 2.3, a number of tools are presented to judge the sample size for a given unlabeled dissimilarity representation. It has been shown that these tools are consistent with modifications of the representation that make it either more or less complex. All the considered criteria are useful when judged as complementary to each other. The most indicative ones, however, are skewness, PCA dimensionality, correlation, embedding intrinsic dimensionality, and boundary descriptor.

In section 2.4, the same observations concerning the power transformations have been confirmed by classification experiments. By putting emphasis on remote objects (hence considering D^{*2} representations), a problem becomes simpler as local class differences become less apparent. As a result, this simpler problem will have a higher class overlap, but may be solved by a simpler classifier. By emphasizing small distances between objects (hence considering $D^{*0.5}$ representations), on the contrary, local class distances may be used better. The problem may now be solved by a more complex classifier, requiring more samples, but resulting in a lower error rate.

It can be understood from this study that data-set complexity is related to sampling density if the data set has to be used for generalization like the training of classifiers. A more complex data set needs a higher sampling density, and, consequently, better classifiers may be found. If the training set is not sufficiently large, representations having a lower complexity may perform better. This conclusion is consistent with the earlier insights in the cause of the peaking phenomenon and the curse of dimensionality [19]. The concepts of representation complexity and data-set complexity, however, are more general than the dimensionality of a feature space.

In conclusion, we see a perspective for using the sample density to build a criterion judging the complexity of a representation as given by a data set. If sufficient samples are available, the representation may be changed such that local details become highlighted. If not, then the representation should be simplified by emphasizing its more global aspects.

Acknowledgments

This work is supported by the Dutch Technology Foundation (STW), grant RRN 5699, as well as the Dutch Organization for Scientific Research (NWO). The authors thank Douglas Zongker, Anil Jain, and Volker Roth for some of the data sets used in this study.

References

[1] A.G. Arkadev, E.M. Braverman. *Computers , Pattern Recognition*. Washington, DC: Thompson, 1966.

[2] C.L. Blake, C.J. Merz. UCI repository of machine learning databases. University of California, Irvine, Department of Information and Computer Sciences, 1998.

[3] P.S. Bradley, O.L. Mangasarian. Feature selection via concave minimization and support vector machines. In *International Conference on Machine Learning*, pages 82–90. San Francisco: Morgan Kaufmann, 1998.

[4] P.S. Bradley, O.L. Mangasarian, W.N. Street. Feature selection via mathematical programming. *INFORMS Journal on Computing*, 10, 209–217, 1998.

[5] F. Corpet, F. Servant, J. Gouzy, D. Kahn. Prodom and prodom-cg: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Research*, 28, 267–269, 2000.

[6] P.A. Devijver, J. Kittler. *Pattern Recognition, A Statistical Approach*. Englewood Cliffs, NJ: Prentice Hall, 1982.

[7] M.P. Dubuisson, A.K. Jain. Modified Hausdorff distance for object matching. In *International Conference on Pattern Recognition*, volume 1, pages 566–568, 1994.

[8] R.P.W. Duin. Compactness and complexity of pattern recognition problems. In *International Symposium on Pattern Recognition 'In Memoriam Pierre Devijver'*, pages 124–128. Brussels: Royal Military Academy, 1999.

[9] R.P.W. Duin, E. Pękalska. Complexity of dissimilarity based pattern classes. In *Scandinavian Conference on Image Analysis.* pages 663–670, Bergen, Norway, 2001.

[10] R.P.W. Duin, D.M.J. Tax. Classifier conditional posterior probabilities. In A. Amin, D. Dori, P. Pudil, H. Freeman, eds. *Advances in Pattern Recognition, LNCS*, volume 1451, pages 611–619. New York: Springer Verlag, 1998.

[11] R.P.W. Duin, D.M.J. Tax. Combining support vector and mathematical programming methods for classification. In B. Schoelkopf, C. Burges, A. Smola, eds. *Advances in Kernel Methods — Support Vector Machines*, pages 307–326. Cambridge, MA: MIT Press, 1999.

[12] Fasta. http://www.ebi.ac.uk/fasta/index.html.

[13] L. Goldfarb. A new approach to pattern recognition. In L.N. Kanal, A. Rosenfeld, eds. *Progress in Pattern Recognition*, volume 2, pages 241–402. Amsterdam: Elsevier Science Publishers BV, 1985.

[14] J.C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27, 25–33, 1971.

[15] T. Graepel, B. Schölkopf, et al. Classification on proximity data with LP-machines. In *International Conference on Artificial Neural Networks*, pages 304–309, 1999.

[16] T.K. Ho, M. Basu. Measuring the complexity of classification problems. In *International Conference on Pattern Recognition*, volume 2, pages 43–47, Barcelona, Spain, 2000.

[17] T.K. Ho, M. Basu. Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3), 289–300, 2002.

[18] D. Hofstadter. *Gödel, Escher, Bach — an Eternal Golden Braid.* New York: Basic Books, 1979.

[19] A.K. Jain, B. Chandrasekaran. Dimensionality and sample size considerations in pattern recognition practice. In P.R. Krishnaiah, L.N. Kanal, eds. *Handbook of Statistics*, volume 2, pages 835–855. Amsterdam: North-Holland, 1987.

[20] A.K. Jain, 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] A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Jornal of Molecular Biology*, 247, 536–540, 1995.

[22] E. Pękalska. *Dissimilarity representations in pattern recognition. Concepts, theory and applications*. Ph.D. thesis, Delft University of Technology, Delft, The Netherlands, January 2005.

[23] E. Pękalska, R.P.W. Duin. Dissimilarity representations allow for building good classifiers. *Pattern Recognition Letters*, 23(8), 943–956, 2002.

[24] E. Pękalska, R.P.W. Duin. On not making dissimilarities euclidean. In T. Caelli, A. Amin, R.P.W. Duin, M. Kamel, de D. Ridder, eds. *Joint IAPR International Workshops on SSPR and SPR, LNCS*, pages 1143–1151. New York: Springer-Verlag, 2004.

[25] E. Pękalska, R.P.W. Duin, and P. Paclík. Prototype selection for dissimilarity-based classifiers. *Pattern Recognition*, 39(2), 189–208, 2006.

[26] E. Pękalska, P. Paclík, R.P.W. Duin. A generalized kernel approach to dissimilarity based classification. *Journal of Machine Learning Research*, 2, 175–211, 2001.

[27] E. Pękalska, D.M.J. Tax, R.P.W. Duin. One-class LP classifier for dissimilarity representations. In S. Thrun S. Becker, K. Obermayer, eds. *Advances in Neural Information Processing Systems 15*, pages 761–768. Cambridge, MA: MIT Press, 2003.

[28] V. Roth, J. Laub, J.M. Buhmann, K.-R. Müller. Going metric: Denoising pairwise data. In *Advances in Neural Information Processing Systems*, pages 841–856. Cambridge, MA: MIT Press, 2003.

[29] M. Skurichina, R.P.W. Duin. Combining different normalizations in lesion diagnostics. In O. Kaynak, E. Alpaydin, E. Oja, L. Xu, eds. *Artificial Neural Networks and Information Processing, Supplementary Proceedings ICANN/ICONIP*, pages 227–230, Istanbul, Turkey, 2003.

[30] D.C.G. de Veld, M. Skurichina, M.J.H. Witjes, et al. Autofluorescence characteristics of healthy oral mucosa at different anatomical sites. *Lasers in Surgery and Medicine*, 23, 367–376, 2003.

[31] M.M. Waldrop. *Complexity, the Emerging Science at the Edge of Order and Chaos*. New York: Simon & Schuster, 1992.

[32] C.L. Wilson, M.D. Garris. Handprinted character database 3. Technical report, National Institute of Standards and Technology, February 1992.

[33] D. Wolpert. *The Mathematics of Generalization*. New York: Addison-Wesley, 1995.

[34] Wordnet dictionary. http://www.cogsci.princeton.edu/ wn/.

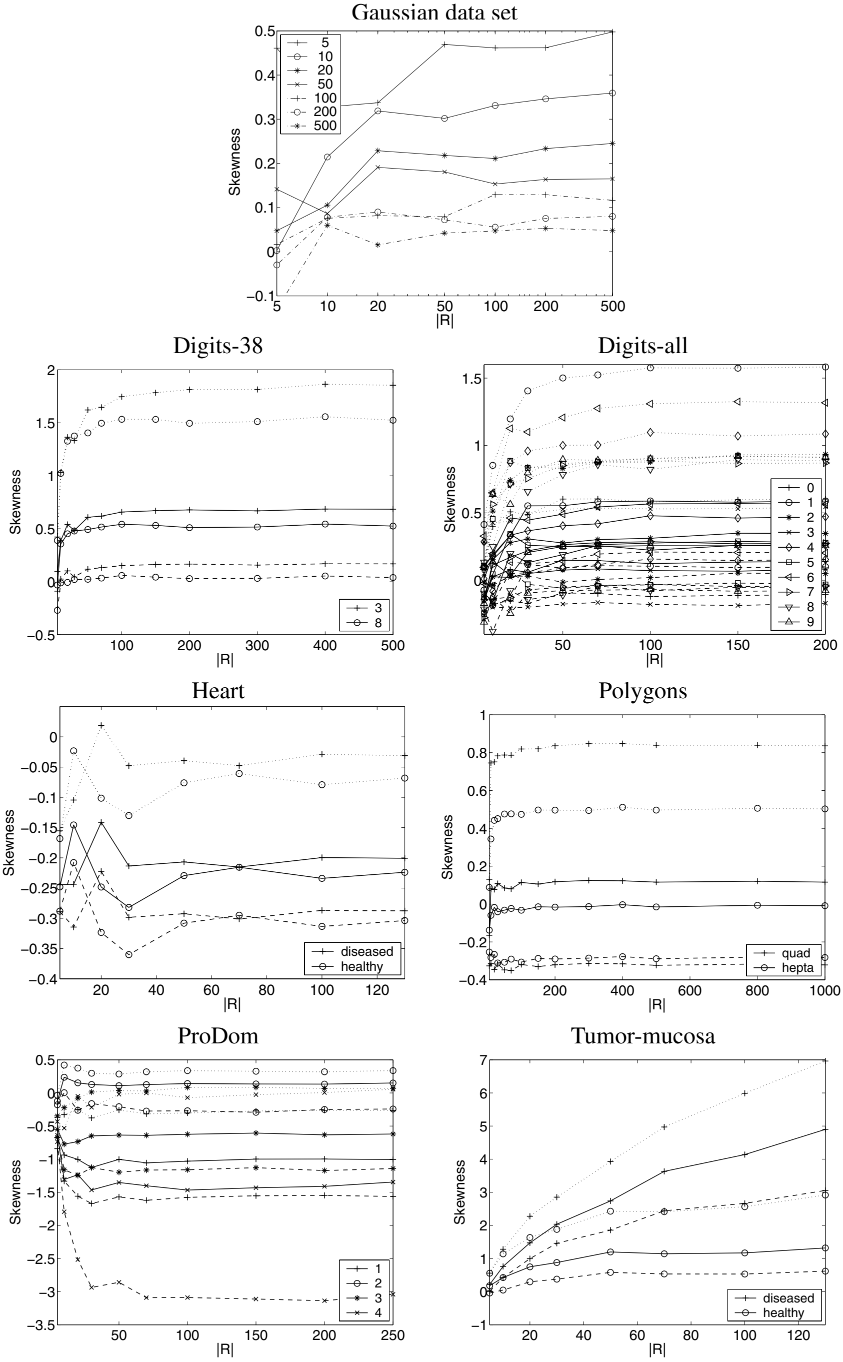


Fig. 2.2. Skewness criterion applied to dissimilarity representations $D^{*p}(R,R)$, $p=0.5,1,2$, per class. Continuous curves refer to the original representation, while the dashed and dotted curves correspond to D^{*05} and D^{*2} representations, respectively. Note scale differences.

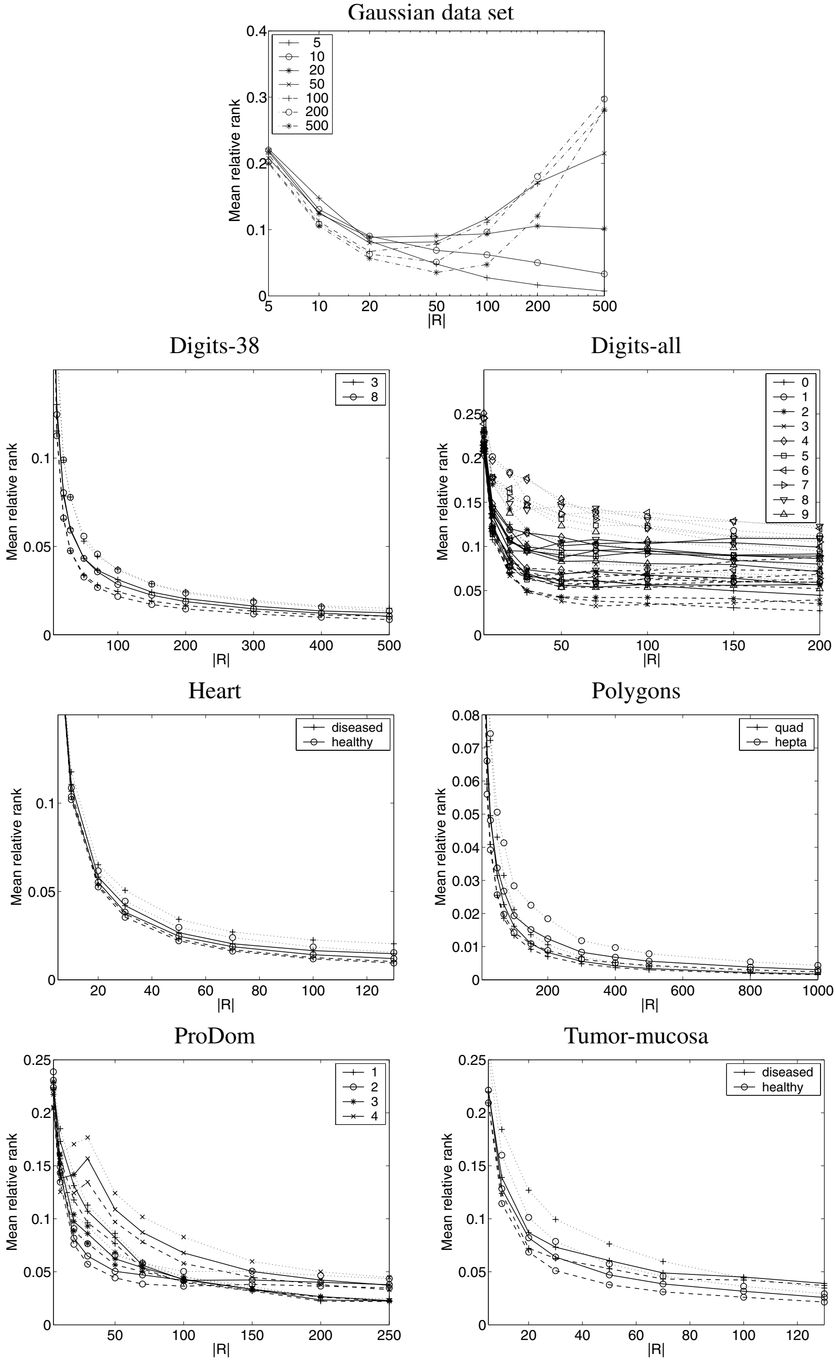


Fig. 2.3. Mean rank criterion applied to dissimilarity representations $D^{*p}(R,R)$, $p = 0.5, 1, 2$, per class. Continuous curves refer to the original representation, while the dashed and dotted curves correspond to D^{*05} and D^{*2} representations, respectively. Note scale differences.

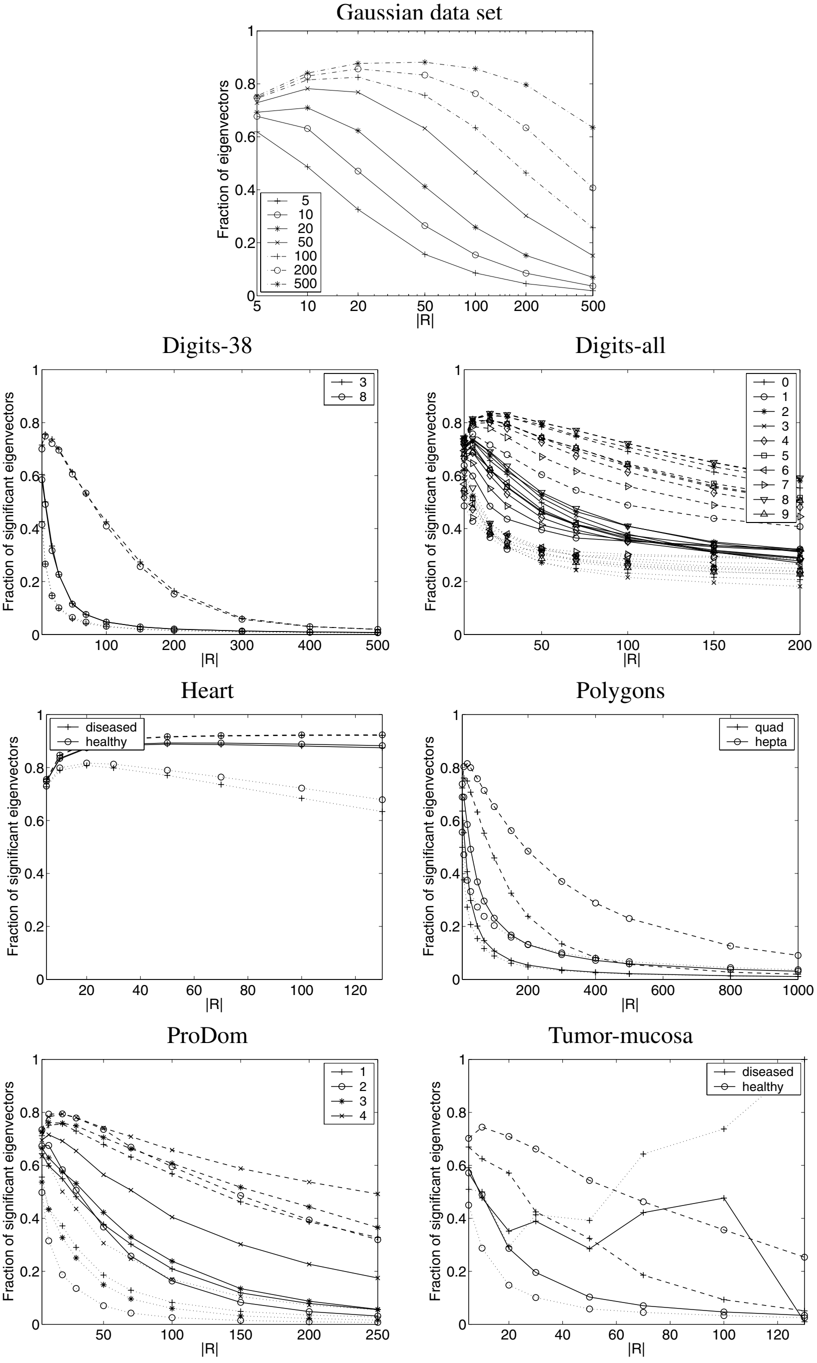


Fig. 2.4. PCA dimensionality criterion applied to dissimilarity representations $D^{*p}(R,R)$, $p = 0.5, 1, 2$, per class. Continuous curves refer to the original representation, while the dashed and dotted curves correspond to D^{*05} and D^{*2} representations, respectively.

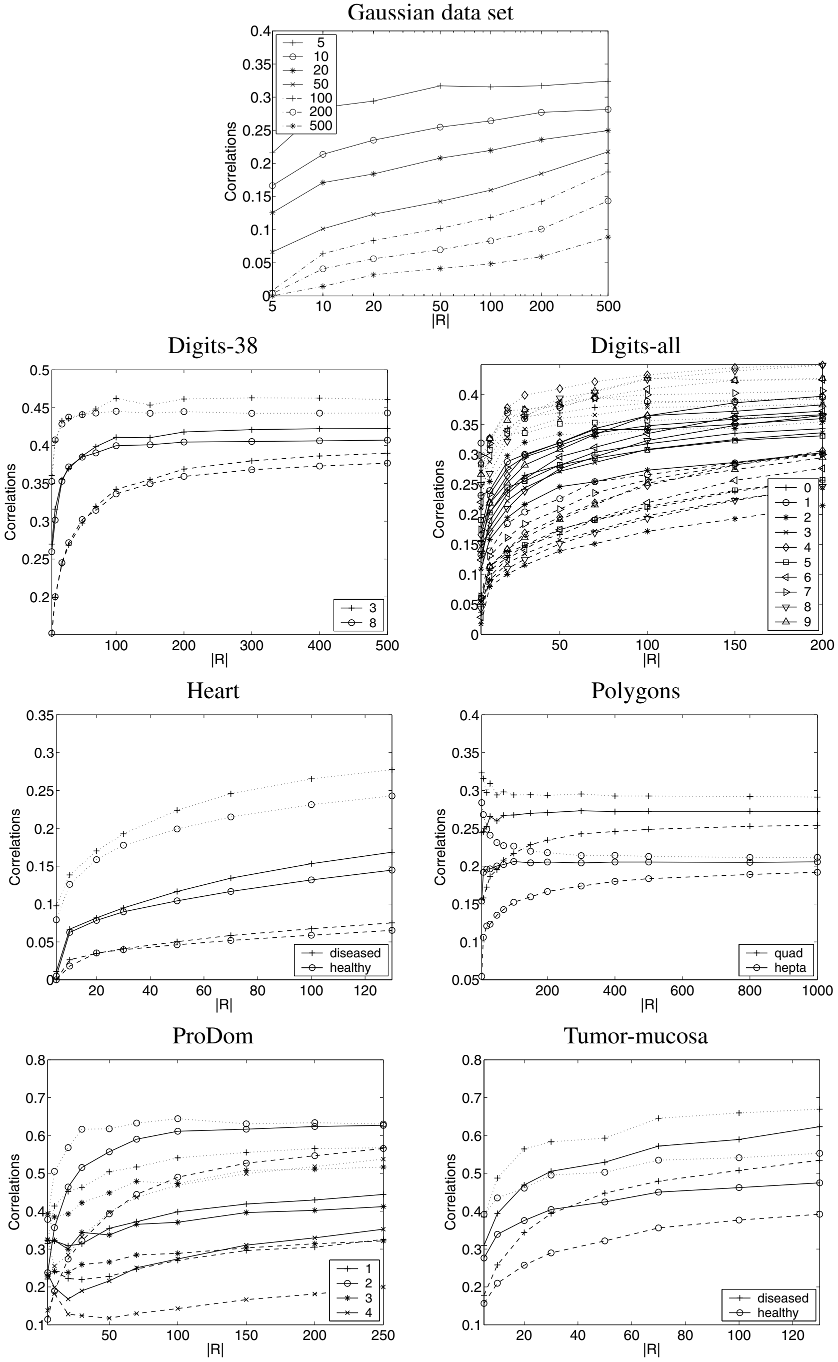


Fig. 2.5. Correlation criterion applied to dissimilarity representations $D^{*p}(R,R)$, $p = 0.5, 1, 2$, per class. Continuous curves refer to the original representation, while the dashed and dotted curves correspond to D^{*05} and D^{*2} representations, respectively. Note scale differences.

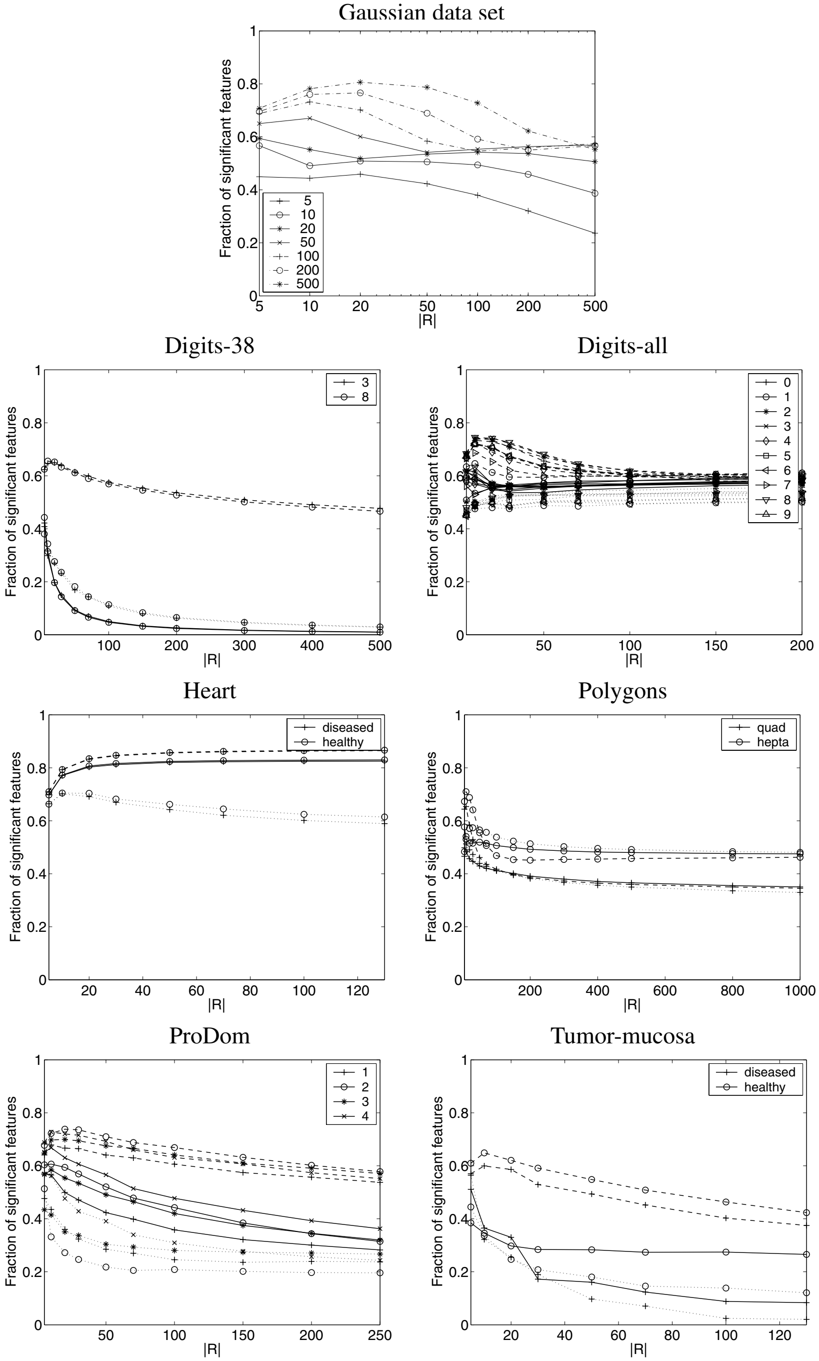


Fig. 2.6. Intrinsic embedding dimensionality criterion applied to dissimilarity representations $D^{*p}(R,R)$, $p=0.5,1,2$, per class. Continuous curves refer to the original representation, while the dashed and dotted curves correspond to D^{*05} and D^{*2} representations, respectively.

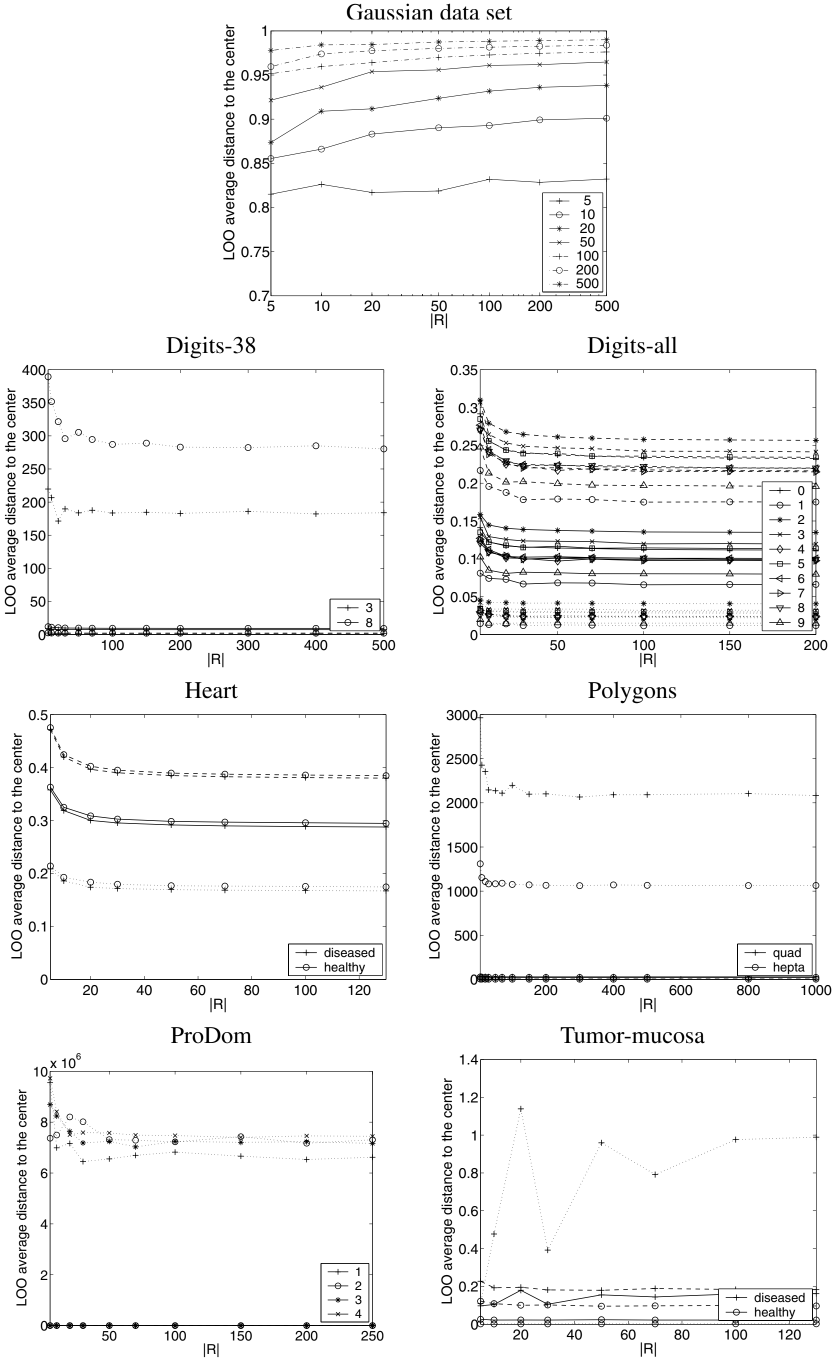


Fig. 2.7. Compactness criterion applied to dissimilarity representations $D^{*p}(R, R)$, $p = 0.5, 1, 2$, per class. Continuous curves refer to the original representation, while the dashed and dotted curves correspond to D^{*05} and D^{*2} representations, respectively. Note scale differences.

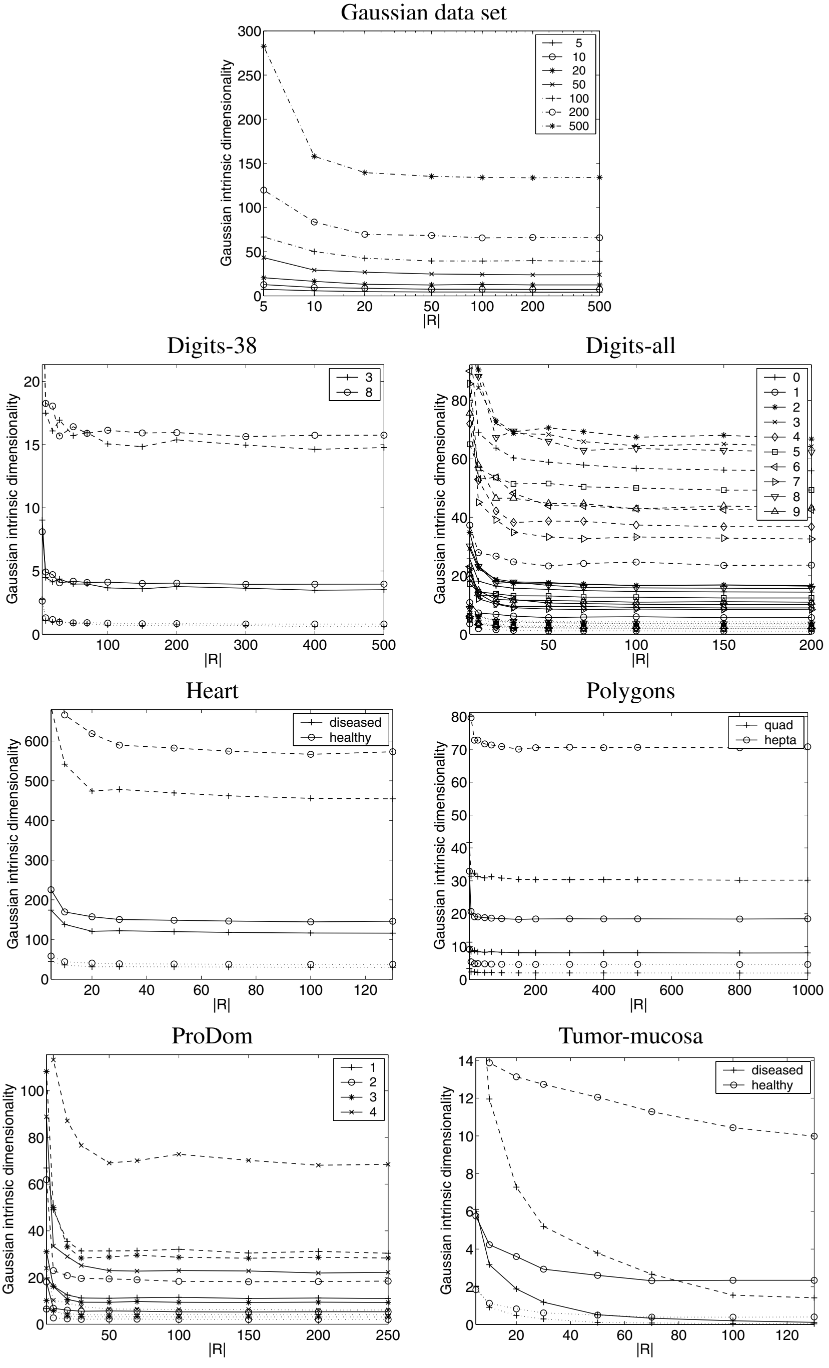


Fig. 2.8. Gaussian intrinsic dimensionality criterion applied to dissimilarity representations $D^{*p}(R, R)$, $p = 0.5, 1, 2$, per class. Continuous curves refer to the original representation, while the dashed and dotted curves correspond to D^{*05} and D^{*2} representations, respectively. Note scale differences.

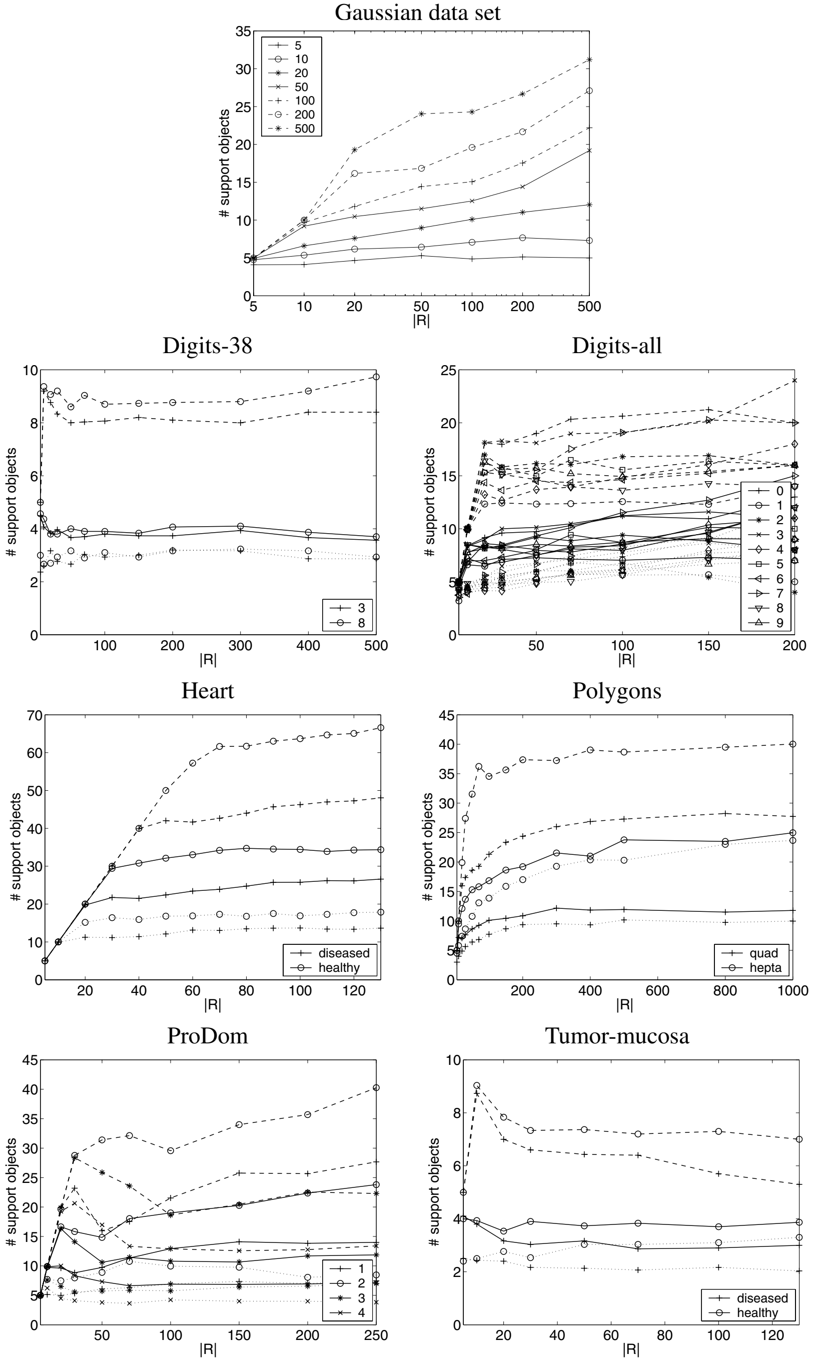
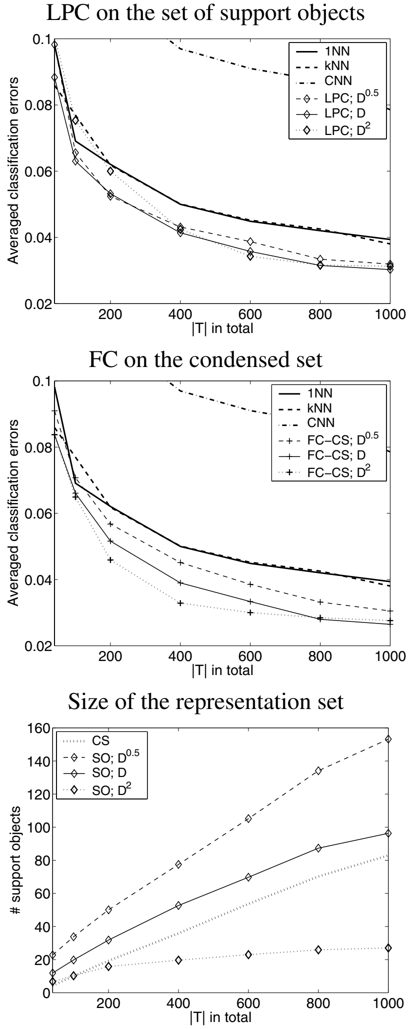


Fig. 2.9. Boundary descriptor criterion applied to dissimilarity representations $D^{*p}(R,R)$, $p=0.5, 1, 2$, per class. Continuous curves refer to the original representation, while the dashed and dotted curves correspond to D^{*05} and D^{*2} representations, respectively. Note scale differences.

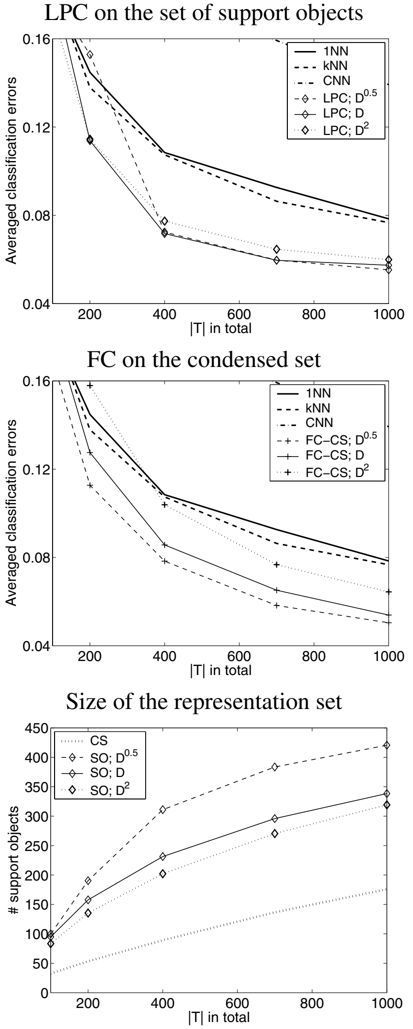


LPC on the set of support objects

FC on the condensed set

Size of the representation set

Fig. 2.10. Results of the classification experiments on the **Digits-38 data**.

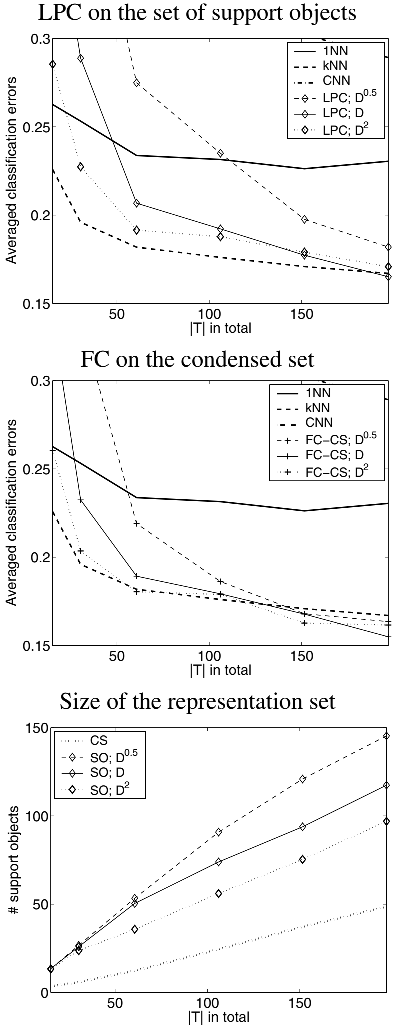


LPC on the set of support objects

FC on the condensed set

Size of the representation set

Fig. 2.11. Results of the classification experiments on the **Digits-all data**.

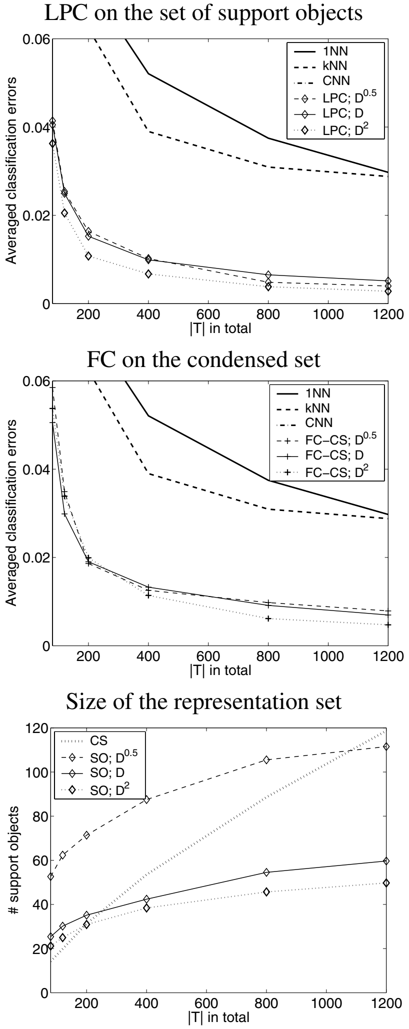


LPC on the set of support objects

FC on the condensed set

Size of the representation set

Fig. 2.12. Results of the classification experiments on the **Heart data**.



LPC on the set of support objects

FC on the condensed set

Size of the representation set

Fig. 2.13. Results of the classification experiments on the **Polygon data**.

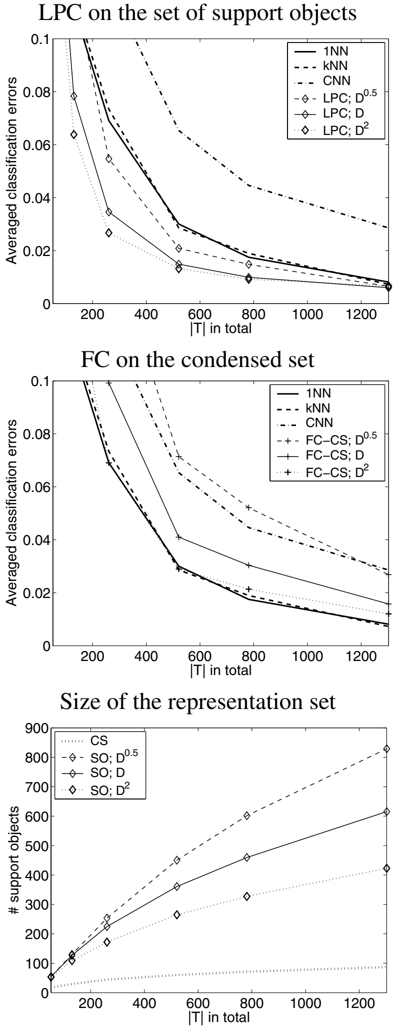


Fig. 2.14. Results of the classification experiments on the **ProDom data**.

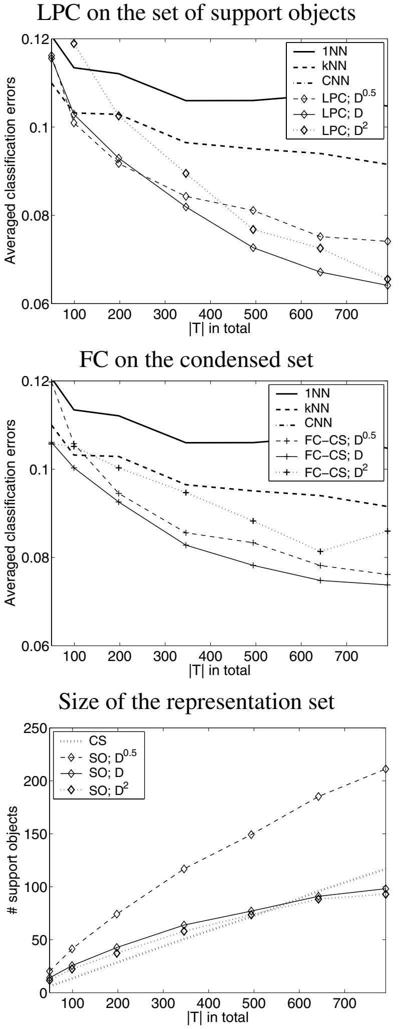


Fig. 2.15. Results of the classification experiments on the **Tumor-mucosa data**.