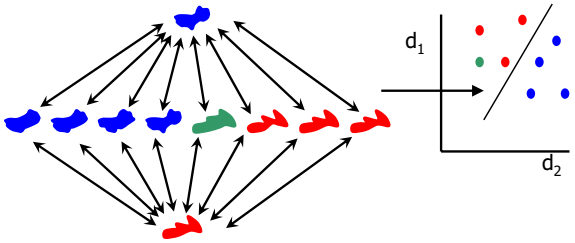
The dissimilarity representation for non-Euclidean pattern recognition, a tutorial

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

[1]Electrical Engineering, Mathematics and Computer Sciences,
Delft University of Technology, The Netherlands
http://rduin.nl/, r.duin@ieee.org

[2]School of Computer Science,
University of Manchester, United Kingdom
http://www.cs.man.ac.uk/∼pekalska/, pekalska@cs.man.ac.uk

November 2011



This tutorial presents an introduction to the studies undertaken by the authors and their collaborators between 1997 and 2011 on the topic of dissimilarity representations for pattern recognition. It emphasizes the significance of integrating non-Euclidean distance measures in representing structured objects.

Introduction

The dissimilarity representation is an alternative for the use of features in the recognition of real world objects like images, spectra and time-signal. Instead of an absolute characterization of objects by a set of features, the expert is asked to define a measure that estimates the dissimilarity between pairs of objects. As such a measure may also be defined for structural representations such as strings and graphs, the dissimilarity representation is potentially able to bridge structural and statistical pattern recognition.

The tutorial aims to give an introduction of the dissimilarity representation to students and researchers that need pattern recognition techniques in their applications. It will consist of three main parts and a discussion. The main parts are:

- Vectorial representations: features, pixels, dissimilarities. We will explain the problems of features: class overlap, the problems of pixels: overtraining and the potentials of dissimilarities.
- Handling dissimilarity data: the traditional nearest neighbor rule (or template matching) is compared to two alternatives: embedding and the dissimilarity space. This results into two entirely different vector spaces in which classifiers may be trained that may perform much better than the nearest neighbor approach.
- Problems with and significance of non-Euclidean data (related to indefinite kernels): in practice many dissimilarity measures used by application experts appear to be non-Euclidean. It will be explained why this is an essential pattern recognition problem. Possible solutions will be discussed.

In this tutorial many references are given for the sake of completeness. This may however confuse the starting student in this area. We recommend therefor to start with two recent papers that aim to give introductory reviews. Possibilities of the dissimilarity space are discussed in [13]. The significance of the use of non-Euclidean dissimilarity measures is discussed in [10].

1 Vector Representations

Automatic systems for the recognition of objects like images, videos, time signals, spectra, etcetera, can be designed by learning from a set of object examples labelled with the desired pattern class. Two main steps can be distinguished in this procedure:

Representation: In this step the individual objects are characterized by a simple mathematical entity such as a vector, string of symbols or a graph. A condition for this representation is that objects can easily be related in order to facilitate the following step.

Generalization: The representations of the object examples should enable the mathematical construction of models for object classes or class discriminants such that a good class estimate can be found for the representation of new, unseen and, thereby, unlabelled objects.

The topic of generalization has been intensively studied within the research areas such as statistical learning theory [45] statistical pattern recognition [21, 9, 27, 20], artificial neural networks [41] and machine learning [6, 2]. The most popular representations are based on Euclidean vector spaces, next to strings and graphs. More recently it has also been studied how to use vector sets for representing single objects; see e.g. [29]. Representations like strings of symbols and attributed graphs are sometimes preferred over vectors as they model the objects more accurately and offer more possibilities to include domain expert knowledge [4].

Representations in Euclidean vector spaces are well suited for generalization. Many tools are available to build (learn) models and discriminants from sets of object examples (training sets) that may be used to classify new objects into the right class. Traditionally, the Euclidean vector space is defined by a set of features. These should ideally characterize the patterns well and also be relevant for class differences at the same time. Such features have to be defined by experts exploiting their knowledge of the application.

A drawback of the use of features is that different objects may have the same representation as they differ by properties that were not expressed in the chosen feature set. This results in class overlap: in some areas in the feature space objects of different classes are represented by the same feature vectors. Consequently, they cannot be distinguished, which leads to an intrinsic classification error, usually called the Bayes error.

A more complete representation than features is just by sampling the objects. For images this is the pixel representation. It assumes that objects are sampled by the same number of pixels and that these pixels are aligned: the same pixel in different images have to describe objects on the same position. Pixels are less informative than features but are useful if no good features can be defined and training set sizes can be large so that still generalization is possible in the high dimensional spaces resulting from pixel representations. A vector representation based on pixels tears the objects in parts as information about the way the pixels constitute an image is lost: the spatial connectivity of the image is not represented in the pixel vector representation.

An alternative to the use of features and pixels is the dissimilarity representation based on direct pairwise object comparisons. If the entire objects are taken into account in the comparison, then only identical objects will have a dissimilarity zero (if

the dissimilarity measure has the property of 'identity of indiscernibles'). For such a representation class overlap does not exist if the objects can be unambiguously labelled: there are no real world objects in the application that belong to more than one class. Only identical objects have a zero-distance and they should have the same label as they are identical.

Another advantage of the dissimilarity representation is that it uses the expert knowledge in a different way. Instead of features, a dissimilarity measure has to be supplied. Of course, when the features are available, a distance measure between feature vectors may be used as a dissimilarity measure. But instead, also other measures, comparing the entire objects may be considered and are even preferred. In some applications, e.g. shape recognition, good features are much more difficult to define than a dissimilarity measure. Even 'bad' dissimilarity measures may be used (at the cost of large training sets) as long as only identical objects have a zero dissimilarity.

2 The dissimilarity representation

Dissimilarities have been used in pattern recognition for a long time. The idea of 'template matching' is based on them: objects are given the same class label if their difference is sufficiently small [8]. This is identical to the nearest neighbor rule used in vector spaces [9]. Also many procedures for cluster analysis make use of dissimilarities instead of feature spaces [44]. To some extent, the concept of dissimilarities is analogous to the use of kernels (and the potential functions as studied in the sixties [1]). The main difference is that kernels were originally defined in vector spaces to preferably fulfill Mercer's conditions [42, 43]. Kernel values can be interpreted as inner products between feature vectors and are, as such, similarities. Because of their properties they are very well suited for finding non-linear classifiers in vector spaces using Support Vector Machines (SVMs) [6].

Inspired by the use of kernels in the machine learning area and the use of dissimilarities in pattern recognition, authors of this tutorial started to experiment with building other classifiers than the ones based on template matching and the nearest neighbor rule for the dissimilarity representation [15, 18, 14, 33, 35], which they also discussed as generalized kernel approaches [40, 37]. Their target was to develop procedures for any type of dissimilarity matrix generated in pattern recognition applications.

The complete dissimilarity representation yields a square matrix with the dissimilarities between all pairs of objects. Traditionally, just the dissimilarities between the test objects and training objects are used. For every test object the nearest neighbors in the set of training objects are first found and used by the nearest neighbor classifier. This procedure does not use the relations between the training objects. The following two approaches construct a new vector space on the basis of the relations within the training set. The resulting vector space is used for training classifiers.

2.1 The dissimilarity space

In the first approach the dissimilarity matrix is considered as a set of row vectors, one for every object. They represent the objects in a vector space constructed by the dissimilarities to the other objects. Usually, this vector space is treated as a Euclidean

space and equipped with the standard inner product definition.

Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a training set. Given a dissimilarity function and/or dissimilarity data, we define a data-dependent mapping $D(\cdot, R) : \mathcal{X} \to \mathbb{R}^k$ from $\mathcal{X}$ to the so-called *dissimilarity space* (DS) [15, 23, 40]. The k-element set R consists of objects that are representative for the problem. This set is called the representation or prototype set and it may be a subset of $\mathcal{X}$. In the dissimilarity space each dimension $D(\cdot, p_i)$ describes a dissimilarity to a prototype p_i from R. In this paper, we initially choose $R := \mathcal{X}$. As a result, every object is described by an n-dimensional dissimilarity vector $D(x, \mathcal{X}) = [d(x, x_1) \ \ldots \ d(x, x_n)]^T$. The resulting vector space is endowed with the traditional inner product and the Euclidean metric.

Any dissimilarity measure ρ can be defined in the DS. One of them is the Euclidean distance:

$$\rho_{DS}(x, y) = (\sum_{i=1}^{n} [d(x, x_i) - d(y, x_i)]^2)^{1/2} \tag{1}$$

This is the distance computed on vectors defined by original dissimilarities. For a set of dissimilarity measures ρ it holds asymptotically that the nearest neighbor objects are unchanged by ρ_{DS}. This is however not necessarily true for finite data sets. It will be shown later that this can be an advantage.

The approaches discussed here are originally intended for dissimilarities directly computed between objects and not resulting from feature representations. It is, however, still possible to study dissimilarity representations derived from features and yields sometimes interesting results [36]. In Fig. 1 an example is presented that compares an optimized radial basis SVM with a Fisher linear discriminant computed in the dissimilarity space derived from the Euclidean distances in a feature space. The example shows a large variability of the nearest neighbor distances. As the radial basis kernel used by SVM is constant it cannot be optimal for all regions of the feature space. Fisher linear discriminant is computed in the dissimilarity space. Here the classes are linearly separable. Although the classifier is overtrained (the dissimilarity space is 100-dimensional and the training set has also 100 objects) it gives here perfect results. It should be realized that this example is specifically constructed to show the possibilities of the dissimilarity space.

In [35, 13] many examples are given that show the use of the dissimilarity space. Many classifiers perform in the dissimilarity space better than the direct use of the nearest neighbor rule, see also [11]. Even the nearest neighbor rule itself may in dissimilarity space outperform the nearest neighbor rule applied on the given dissimilarities. This shows that the total set of distances to the representation set can be informative.

2.2 Embedding the dissimilarity matrix

In the second approach, an attempt is made to embed the dissimilarity matrix in a Euclidean vector space such that the distances between the objects in this space are equal to the given dissimilarities. This can only be realized error free, of course, if the original set of dissimilarities are Euclidean themselves. If this is not the case, either an approximate procedure has to be followed or the objects should be embedded into a non-Euclidean vector space. This is a space in which the standard inner product definition and the related distance measure are changed, resulting in indefinite kernels.

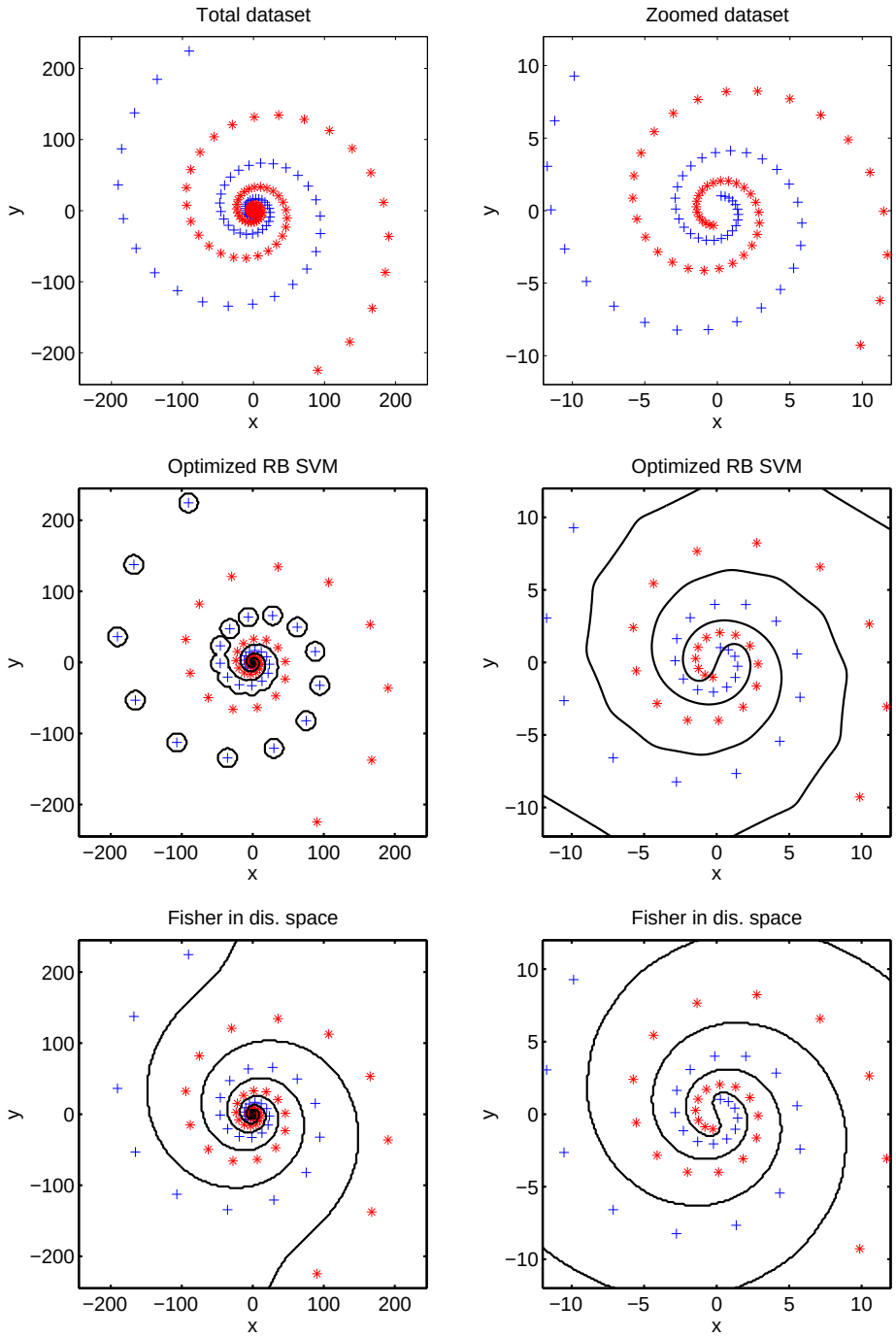


Figure 1: A spiral example with 100 objects per class. Left column shows the complete data sets, while the right column presents the zoom of the spiral center. 50 objects per class are used for training, systematically sampled. The middle row shows the training set and SVM with an optimized radial basis function; 17 out of 100 test objects are erroneously classified. The bottom row shows the Fisher Linear Discriminant (without regularization) computed in the dissimilarity space derived from the Euclidean distances. All test objects are correctly classified.

It appears that an exact embedding is possible for every symmetric dissimilarity matrix with zeros on the diagonal. The resulting space is the so-called pseudo-Euclidean space.

Many of the dissimilarity measures used in the pattern recognition practice appear to be indefinite: they cannot be understood as distances in a Euclidean vector space, they are sometimes even not metric and they do not satisfy the Mercer conditions.

We will give some definitions.

A Pseudo-Euclidean Space (PES) $\mathcal{E} = \mathbb{R}^{(p,q)} = \mathbb{R}^p \oplus \mathbb{R}^q$ is a vector space with a non-degenerate indefinite inner product $\langle \cdot, \cdot \rangle_{\mathcal{E}}$ such that $\langle \cdot, \cdot \rangle_{\mathcal{E}}$ is positive definite on $\mathbb{R}^p$ and negative definite on $\mathbb{R}^q$ [22, 35]. The inner product in $\mathbb{R}^{(p,q)}$ is defined (wrt an orthonormal basis) as $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{E}} = \mathbf{x}^T \mathcal{J}_{pq} \mathbf{y}$, where $\mathcal{J}_{pq} = [I_{p \times p}\ 0;\ 0\ -I_{q \times q}]$ and I is the identity matrix. As a result, $d^2_{\mathcal{E}}(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \mathcal{J}_{pq} (\mathbf{x} - \mathbf{y})$. Obviously, a Euclidean space $\mathbb{R}^p$ is a special case of a pseudo-Euclidean space $\mathbb{R}^{(p,0)}$. An infinite-dimensional extension of a PES is a Kreĭn space. It is a vector space $\mathcal{K}$ equipped with an indefinite inner product $\langle \cdot, \cdot \rangle_{\mathcal{K}} \colon \mathcal{K} \times \mathcal{K} \to \mathbb{R}$ such that $\mathcal{K}$ admits an orthogonal decomposition as a direct sum, $\mathcal{K} = \mathcal{K}_+ \oplus \mathcal{K}_-$, where $(\mathcal{K}_+, \langle \cdot, \cdot \rangle_+)$ and $(\mathcal{K}_-, -\langle \cdot, \cdot \rangle_-)$ are separable Hilbert spaces with their corresponding positive and negative definite inner products.

A positive definite kernel function can be interpreted as a generalized inner product in some Hilbert space. This space becomes Euclidean when a kernel matrix is considered. In analogy, an arbitrary symmetric kernel matrix can be interpreted as a generalized inner product in a pseudo-Euclidean space. Such a PES is obviously data dependent and can be retrieved via an embedding procedure. Similarly, an arbitrary symmetric dissimilarity matrix with zero self-dissimilarities can be interpreted as a pseudo-Euclidean distance in a proper pseudo-Euclidean space. Since in practice we deal with finite data, dissimilarity matrices or kernel matrices can be seen as describing relations between vectors in the underlying pseudo-Euclidean spaces. These pseudo-Euclidean spaces can be either determined via an embedding procedure and directly used for generalization, or approached indirectly by the operations on the given indefinite kernel. Below it is explained how to find the embedded PES.

A symmetric dissimilarity matrix $D := D(\mathcal{X}, \mathcal{X})$ can be embedded in a Pseudo-Euclidean Space (PES) $\mathcal{E}$ by an isometric mapping [22, 35].

The embedding relies on the indefinite Gram matrix G, derived as $G := -\frac{1}{2} H D^{\star 2} H$, where $D^{\star 2} = (d^2_{ij})$ and $H = I - \frac{1}{n} \mathbf{1}\mathbf{1}^T$ is the centering matrix. H projects the data such that X has a zero mean vector. The eigendecomposition of G leads to $G = Q \Lambda Q^T = Q |\Lambda|^{\frac{1}{2}} [\mathcal{J}_{pq};\ 0] |\Lambda|^{\frac{1}{2}} Q^T$, where Λ is a diagonal matrix of eigenvalues, first decreasing p positive ones, then increasing q negative ones, followed by zeros. Q is the matrix of eigenvectors. Since $G = X \mathcal{J}_{pq} X^T$ by definition of a Gram matrix, $X \in \mathbb{R}^n$ is found as $X = Q_n \, |\Lambda_n|^{\frac{1}{2}}$, where Q_n consists of n eigenvectors ranked according to their eigenvalues Λ_n. Note that X has a zero mean and is uncorrelated, because the estimated pseudo-Euclidean covariance matrix $C = \frac{1}{n-1} X^T X \, \mathcal{J}_{pq} = \frac{1}{n-1} \Lambda_r$ is diagonal. The eigenvalues λ_i encode variances of the extracted features in $\mathbb{R}^{(p,q)}$.

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. If this space is a PES $\mathbb{R}^{(p,q)}$, $p+q = n$, the pseudo-Euclidean distance

is computed as:

$$\rho_{PES}(\mathbf{x},\mathbf{y}) = (\sum_{i=1}^{p}[x_i - y_i]^2 - \sum_{i=p+1}^{p+q}[x_i - y_i]^2)^{1/2}$$
$$= (\sum_{i=1}^{n}\delta(i,p)[x_i - y_i]^2)^{1/2},$$

where $\delta(i,p) = \text{sign}(p - i + 0.5)$. Since the complete pseudo-Euclidean embedding is perfect, $D(x,y) = \rho_{PES}(x,y)$ holds.

Other distance measures may also be defined between vectors in a PES, depending on how this space is interpreted. Two obvious choices are:

$$\rho_{PES+}(\mathbf{x},\mathbf{y}) = (\sum_{i=1}^{p}[x_i - y_i]^2)^{1/2}, \tag{2}$$

which neglects the axes corresponding to the negative dimensions (derived from negative eigenvalues in the embedding), and

$$\rho_{AES}(\mathbf{x},\mathbf{y}) = (\sum_{i=1}^{n}[x_i - y_i]^2)^{1/2}, \tag{3}$$

which treats the vector space $\mathbb{R}^n$ as Euclidean $\mathbb{R}^{p+q}$. This means that the negative subspace of PES is interpreted as a Euclidean subspace (i.e. the negative signs of eigenvalues are neglected in the embedding procedure).

To inspect the amount of non-Euclidean influence in the derived PES, we define the Negative Eigen-Fraction (NEF) as:

$$NEF = \sum_{j=p+1}^{p+q}|\lambda_j| / \sum_{i=1}^{p+q}|\lambda_i| \in [0,1] \tag{4}$$

Fig. 2 shows how NEF varies as a function of p of the Minkowski-p dissimilarity measure (k-dimensional spaces) for a two-dimensional example:

$$\rho_{Min_p}(\mathbf{x},\mathbf{y}) = (\sum_{i=1}^{k}[x_i - y_i]^p)^{1/p} \tag{5}$$

This dissimilarity measure is Euclidean for $p = 2$ and metric for $p > 1$. The measure is non-Euclidean for all $p \neq 2$. The value of NEF may vary considerably with a changing dimensionality. This phenomenon is illustrated in Fig. 3 for 100 points generated by a standard Gaussian distribution for various values of p. The one-dimensional dissimilarities obviously fit perfectly to a Euclidean space. For a vary high dimensionality, the sets of dissimilarities become again better embeddable in a Euclidean space.

2.3 Discussion on dissimilarity-based vector spaces

Here we make some remarks on the two procedures for deriving vector spaces from dissimilarity matrices, as discussed in previous subsection. On some aspects we will return at the end of this reports in relation to examples and experiments.

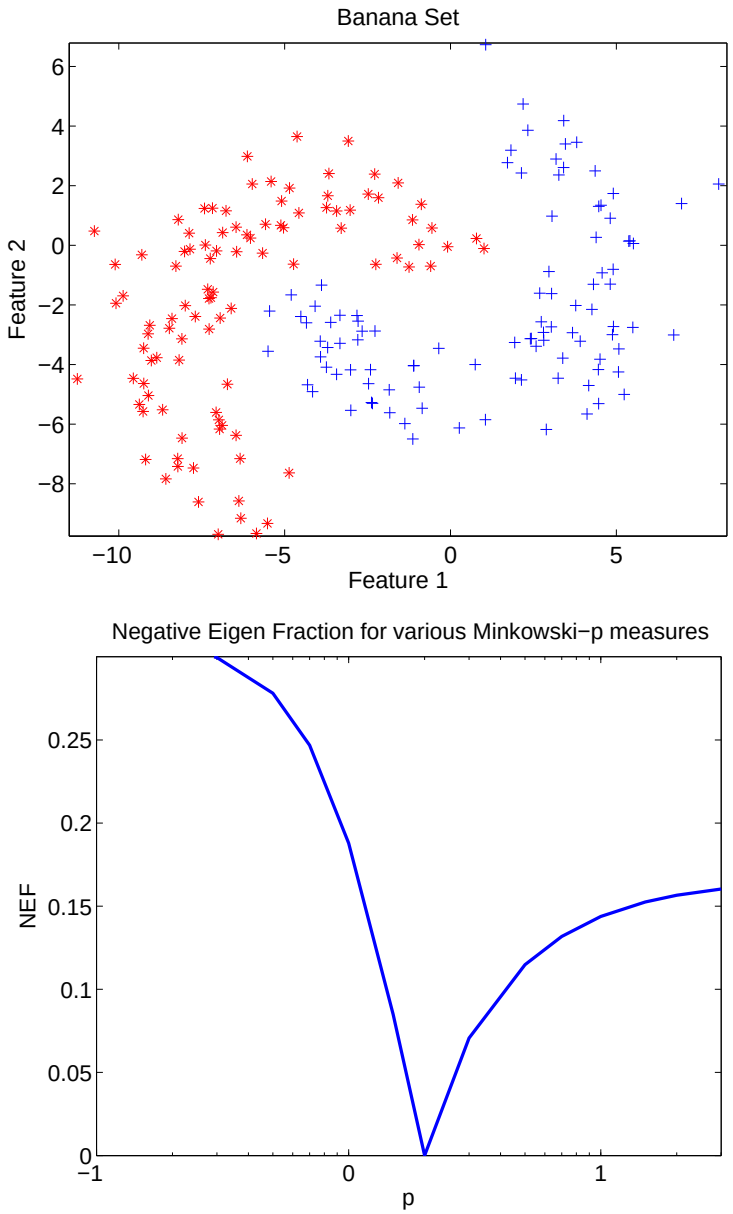


Figure 2: A two-dimensional data set (left) and the NEF as a function of p for various Minkowski-p dissimilarity measures.

The dissimilarity space in fact interprets the dissimilarities to particular prototypes (the representation set) as features. Their characteristics of dissimilarities is not used when a general classifier is applied. Special classifiers are needed to make use of that information. The good side of this 'disadvantage' is that the dissimilarity space can be used for any dissimilarity representation, including ones that are negative or asymmetric.

The embedding procedure is more restrictive. The dissimilarities are assumed to be symmetric and zero for the comparison with identical objects. Something like the pseudo-Euclidean embedding is needed in case of non-Euclidean data sets. The requirements of a proper metric or well-defined distances obeying the triangle inequality are not of use as they do not guarantee a Euclidean embedding. As we want to study

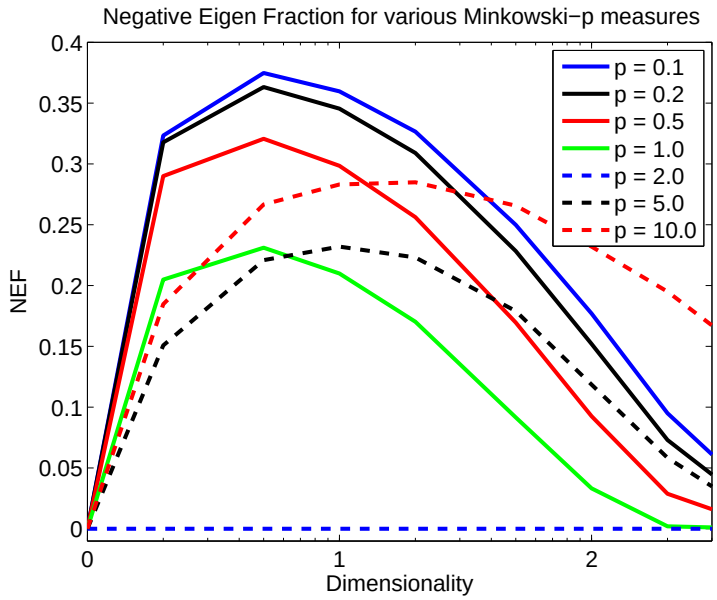


Figure 3: The Negative Eigen-Fraction for various Minkowski-p dissimilarity measures as a function of the dimensionality of a set of 100 points generated by a standard Gaussian distribution.

more general data sets we use the name of dissimilarities instead of distances.

A severe drawback of both procedures is that they initially result in vector spaces that have as many objects as dimensions. Specific classifiers or dimension reduction procedures are thereby needed. For the dissimilarity representation this is somewhat more feasible than for the feature representation: features can be very different, some might be very good, others might be useless, or only useful in relation with particular other features. This is not true for dissimilarities. The initial representation is just based on objects. They have similar characteristics. It is not useful to use two objects that are much alike. Systematic, or even random procedures that reduce the initial representation set (in fact prototype selection) can be very effective [34] for this reason.

3 Non-Euclidean dissimilarities

The work on the general dissimilarity matrices touches the gradually raising interest of the machine learning community in indefinite kernels: [24, 25, 26, 32, 31]. There is however some doubt whether the non-Euclidean aspects of the relations between pairwise comparison of objects are informative [39, 19, 30].

In this section preparations are discussed to study further the handling and possible informativeness of non-Euclidean dissimilarity matrices. From the observation that they arise often in the pattern recognition practice, it can be concluded that this is a significant issue [10]. We will therefore discuss the various circumstances under which such dissimilarity matrices arise and will try to characterize them, see also [12]. Next, we will discuss three ways to approach this problem:

1. Avoiding the non-Euclidean dissimilarities by adapting the measure.

2. Correcting dissimilarity matrices such that they become Euclidean and by this

traditional generalization procedures can be applied.

3. Leaving the data as they are and developing generalization procedures that can handle non-Euclidean dissimilarity data.

The purpose of our study is to find good generalization procedures for dissimilarity data that arise in practical pattern recognition applications. In between is the step of representation. In the previous section two procedures for deriving vector spaces are presented. One is general, but neglects the dissimilarity characteristic of the data. The other is specific but suffers from the possible non-Euclidean relations that are present in the data. In order to analyze possible transformations of the derived vector spaces, especially of the pseudo-Euclidean space, we will first summarize and categorize the ways in which non-Euclidean dissimilarity data can arise.

Before becoming more specific, we like to emphasize how common non-Euclidean measures are. In [35] we already presented an extensive overview of such measures, but we encountered in many occasions that this fact is not sufficiently recognized.

Almost all probabilistic distance measures are non-Euclidean, including the Kolmogorov Variational Distance which is directly related to the classification error. This implies that when we want to build a classification system for a set of objects and each individual object is represented by a probability density function resulting from its invariants, the dissimilarity matrix resulting from the overlap between the object pdfs is non-Euclidean. Also the Mahalanobis class distance as well as the related Fisher criterion are non-Euclidean.

As a direct consequence of the above, many non-Euclidean distance measures are used in cluster analysis and in the analysis of spectra in chemometrics and hyperspectral image analysis. An energy spectrum can be considered as a pdf of energy contributions for different wavelengths. The popular absolute difference between two spectra is identical with the Minkowski-1 distance (related to the l_1-norm) between vector representations of the spectra.

In shape recognition, various dissimilarity measures are used based on the weighted edit distance as well as on variants of the Hausdorff distance. Usual parameters are optimized within an application w.r.t. the performance based on template matching and other nearest neighbor classifiers [5]. Most of them are still metric, some of them however are non-metric [7].

In the design and optimization of the dissimilarity measures it was in the past not an issue whether they were Euclidean. Just more recently, with the popularity of SVMs, it has became important to design kernels (similarity measures) which fulfill the Mercer conditions. This is equivalent to the possibility of Euclidean embedding. Next subsection discusses a number of reasons that give rise to violations of these conditions in applications, which lead to a set of non-Euclidean dissimilarities or indefinite kernels.

3.1 Non-intrinsic non-Euclidean dissimilarities

3.1.1 Numeric inaccuracies

A very simple reason why non-Euclidean dissimilarities arise is the numeric inaccuracies resulting from the use of computers with a finite word length. E.g., when we generate at random four points in an n-dimensional vector space and we follow the embedding

procedure discussed in section **??** the projected vectors will fit in a 3-dimensional Euclidean space. In the procedure three eigenvalues larger than zero are expected to be found. In case $n = 2$ one of these eigenvalues will be zero. In a numeric procedure, however, there is a probability of almost 50% that the smallest eigenvalue has a very small negative value due to numeric inaccuracies (resulting from iterative procedures of determining the eigenvalues).

For this reason it is advisable to neglect all very small positive as well as negative eigenvalues. As a consequence, the dimensionality of the embedded space will be smaller than its maximum value of n-1.

3.1.2 Overestimation of large distances

When dissimilarities are not directly computed in a vector space but derived on raw data such as images or objects detected in images instead, more complicated measures may be used. They may still rely on the concept that the distance between two objects is the length or cost of the shortest path that has to be followed to transform one object into the other. Examples of such transformations are the weighted edit distance [3] and deformable templates [28]. In the optimization procedure that minimizes the length of the path, a minimization procedure may be used based on approximating the costs from above. As a consequence, too large distances are found.

The detection of too large distances is not easy, except when they are so large that the triangle inequality has been violated. In that case $d(A, C) > d(A, B) + d(B, C)$, indicating that a lower cost is possible in the transformation of A to C via a detour over B. This violates the result of the cost minimization. See [16] for an example. Such violations can easily be detected and corrected. The result is however just the replacement of a non-metric measure by a metric one. A possible non-Euclidean set of dissimilarities resulting from relations between more than three objects may still exist.

3.1.3 Underestimation of small distances

The underestimation of small distances has the same result as the above discussed overestimation of large distances. Similar correction procedures may be applied and again they only correct the metric property but not the Euclidean one.

There may be different causes of underestimated small distances. They may arise as the consequence of neglecting different particular object properties in different pairwise comparisons. For instance, in consumer preference data, the ranking of the most interesting books by every reader individually yields (dis)similarities based on different books by different pairwise comparisons of books or readers. Unread books by both readers in a comparison are thereby not taken into account, resulting in a too small estimate, especially for the small dissimilarities. E.g., it is possible to estimate a dissimilarity of zero if the ranking of the books read by both readers is identical, while it may be larger if additional books are taken into account.

Phrased in more abstract terms, the underestimation of small distances occurs when object pairs have to be compared from different points of view, or suffering from different partial (information) occlusions.

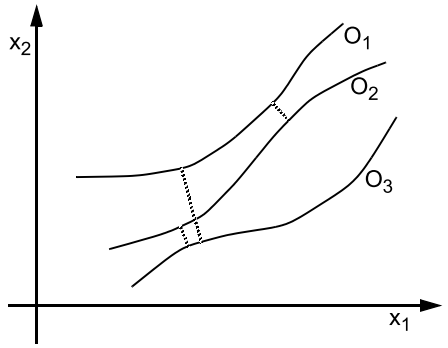


Figure 4: Vector space with the invariant trajectories for three objects O_1, O_2 and O_3. If the chosen dissimilarity measure is the minimal distance between these trajectories, triangle inequality can easily be violated, i.e. $d(O_1, O_2) + d(O_1, O_3) < d(O_1, O_3)$.

3.2 Intrinsic non-Euclidean dissimilarities

The causes discussed in the above may be judged as accidental. They result either from computational or observational problems. If better computers and observations were available, they would disappear. Now we will focuss on dissimilarity measures for which this will not happen. We will discuss three possibilities, without claiming completeness.

3.2.1 Non-Euclidean dissimilarities

As already indicated at the start of this section, there can be arguments from the application side to use another metric than the Euclidean one. An example is the Kolmogorov variational distance between pdfs as it is related to the classification error, or the l_1-distance between energy spectra as it is related to energy differences. Although the l_2-norm is very convenient for computational reasons or because it is rotation invariant in a Euclidean space, the l_1-norm may naturally arise from the demands in applications.

3.2.2 Invariants

A very fundamental reason is related to the occurrence of invariants. Frequently, one is not interested in the dissimilarity between two objects A and B, but between two families of objects $A(\theta)$ and $B(\theta)$ in which θ controls an invariant, e.g. rotation in case of shape recognition. One may define the dissimilarity between two objects A and B as the minimum difference between the two sets defined by all their invariants.

$$d^*(A, B) = \min_{\theta_A} \min_{\theta_B} (d(A(\theta_A), B(\theta_B))) \tag{6}$$

In general, this measure is non-metric: the triangle inequality may be violated as for different pairs of objects different values of θ may be found that minimize (6). An example is given in figure 3.2.2, which is taken from [37].

3.2.3 Sets of vectors

Finding relations between sets of vectors is an important issue in cluster analysis. Individual objects may be represented by single vectors, but in a
hierarchical clustering procedure the (dis)similarities between already grouped vectors are used to establish a new cluster level. Dissimilarity measures as used in the complete linkage and single linkage procedures are very common. The second, which is defined as the distance between the two most neighboring points of the two clusters being compared, is non-metric. It even holds for this distance measure that if $d(A, B) = 0$, then it does not follow that $A \equiv B$, because different clusters may just be touching.

For the single linkage dissimilarity measure it can be understood why the dissimilarity space may be useful. Given a set of such dissimilarities between clouds of vectors, it can be concluded that two clouds are similar if the entire sets of dissimilarities with all other clouds are about equal. If just their mutual dissimilarity is (close to) zero, they may still be very different. Fig. 5 illustrates this point.

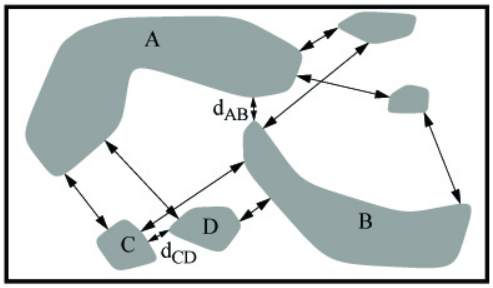


Figure 5: Single-linkage distance may be small for clusters which differ in position and shape.

The problem with the single linkage dissimilarity measure between two sets of vectors points to a more general problem in relating sets and even objects. In [29] an attempt has been made to define a proper Mercer kernel between two sets of vectors. Such sets are in this paper compared by the Hellinger distance derived from the Bhattacharyya's affinity between two pdfs $p_A(x)$ and $p_B(x)$ found for the two vector sets A and B:

$$d(A, B) = \left[\int \left(\sqrt{p_A(x)} - \sqrt{p_B(x)} \right)^2 \right]^{1/2} . \tag{7}$$

The authors state that by expressing $p(x)$ in any orthogonal basis of functions, the resulting kernel K is automatically positive definite. This is correct, but it should be realized that it has to be the same basis for all vector sets $A, B, ...$ to which the kernel is applied. If in a pairwise comparison of sets different bases are derived, the kernel will become indefinite. This may happen if the numbers of vectors per set are smaller than the dimensionality of the vector space. It will happen most likely if this vector space is already a Hilbert space, e.g. when the vectors are already derived from a kernelization step.

This also makes it clear that indefinite relations may arise in any pairwise comparison of real world objects if they are first represented in some joint space for the two objects, followed by a dissimilarity measure. These joint spaces may be different for different pairs! Consequently, the total set of dissimilarities can be non-Euclidean, even if a single comparison is defined as Euclidean, as in (7).

3.3 Other non-Euclidean measures

There may be other factors leading to non-Euclidean dissimilarity measures. After further inspection, they may simplify to one or both of the above. We now mention two possibilities:

- Dis/similarity judgements by human experts. In some applications, e.g. psychometrical experiments, subjects are asked to judge the similarity between various sets of observations. It is not clear on which ground such judgements are made, as also in the consumer preference data.

- Weighted combinations of different dis/similarity measures that focus on different aspects of objects, e.g.
$$d(x, y) = \sum_i \alpha_i d_i(x, y)$$
where α_i is a constant and $d_i(x, y)$ is a dissimilarity w.r.t. particular i-th characteristics. An example is to derive the dissimilarity between images as a weighted average of dissimilarities computed w.r.t. texture, color and response to particular shape detectors.

3.4 Example classifiers in pseudo-Euclidean spaces

In our recent studies on analyzing dissimilarity data [35, 37, 38, 26, 19, 17], we have given many examples for classifiers that can be trained in indefinite (pseudo-Euclidean) spaces, e.g.

- The nearest mean rule as means and distances to points are well defined.

- The nearest neighbor rule for the same reason.

- The Parzen classifier, as it can be expressed in distances to points.

- The linear and quadratic classifiers based on class covariances. In Euclidean spaces they are related to normal distributions. In the pseudo-Euclidean spaces they can still be computed, but the relation with densities is unclear.

- A kernelized version of the Fisher discriminant for indefinite kernels.

Problematic classifiers are the ones based on general probability density estimates, as they are not (yet) properly defined for pseudo-Euclidean spaces and classifiers that rely on a distance to a linear or nonlinear classification boundary, such as SVM. The SVM classifier may still be computed but convergence and uniqueness are not guaranteed [24].

In [26] two artificial examples are presented that illustrate the work and performance of classifiers built in pseudo-Euclidean spaces. In these examples the embedded PES has not been explicitly determined, but classifiers are considered that work on indefinite kernels instead: the indefinite kernel Fisher discriminant (IKFD), indefinite SVM (ISVM) and indefinite kernel nearest mean classifier (IKNMC).

In [19] a study on Euclidean corrections has been presented. Various transformations are studied that map data from the pseudo-Euclidean space to the Euclidean

space. Many examples have been found for which such corrections are counterproductive, suggesting that indefinite spaces can be informative. More subtle corrections have to be investigated further.

The above mentioned transformations are topology preserving. This does not hold for the construction of the dissimilarity space out of a dissimilarity representation. In that case, a new Euclidean space is postulated based on the relations of objects with all other objects. This may remove or diminish noise, or defects that arose in the construction of the original dissimilarities. Possible information of original indefinite relations will thereby only be maintained if it can be expressed in the totality of the relation of objects to all other objects in a Euclidean way.

4 Discussion

Two main causes of non-Euclidean behavior have been identified: non-intrinsic and intrinsic ones. The former are related to computational and computational problems. In case there are no other effects Euclidean representations can be expected asymptotically for increasing computational and observational resources. The latter, the intrinsic causes will remain to yield non-Euclidean dissimilarity matrices.

The question raises whether the correction and classification procedures should be different for these two cases. It may be argued that if it is to be expected that for some circumstances an Euclidean space is appropriate, that then an approximation of this space by some correction of the originally non-Euclidean dataset may approximate the desired representation well. In case of intrinsicly non-Euclidean problems approximative Euclidean spaces might be less effective.

Experiments reported in [19] and in [17] study correction procedures uing interpolations between the PES and several Euclidean spaces. Some of these change the dissimilarities in a monotonous way, by which the NN classification results don't change and thereby also don't improve. Such transformations are nevertheless important they show that for every classifier in the PES, so on the original representation, there exist an equivalent classifier in an Euclidean space. Nevertheless, from all experiments it can be concluded that for many cases the pseudo-Euclidean space can be transformed in a non-topology-preserving way into an Euclidean space in which better classifiers can be computed.

In case there exist an Euclidean space in which several classifiers obtain there best results, we may conclude that the corresponding problem is not intrinsic non-Euclidean. If this space has been found by a correction or transformation of a pseudo-Euclidean space this just suggests that sufficient knowledge lacks to construct such a representation directly from an appropriate set of features or Euclidean (dis)similarity measure. Non-Euclidean measures are thereby still of significant importance.

References

[1] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoér. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[3] H. Bunke and U. Bühler. Applications of approximate string matching to 2D shape recognition. *Pattern recognition*, 26(12):1797–1812, 1993.

[4] H. Bunke and A. Sanfeliu, editors. *Syntactic and Structural Pattern Recognition Theory and Applications*. World Scientific, 1990.

[5] H. Bunke and K. Shearer. A graph distance metric based on the maximal common subgraph. *Pattern Recognition Letters*, 19(3-4):255–259, 1998.

[6] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, UK, 2000.

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

[8] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1972.

[9] Richard Duda, Peter Hart, and David Stork. *Pattern Classification*. John Wiley and Sons, 2001. 0-471-05669-3.

[10] Robert P. W. Duin. Non-euclidean problems in pattern recognition related to human expert knowledge. In *ICEIS*, volume 73 of *Lecture Notes in Business Information Processing*, pages 15–28. Springer, 2011.

[11] Robert P. W. Duin, Marco Loog, Elzbieta Pekalska, and David M. J. Tax. Feature-based dissimilarity space classification. In *ICPR Contests*, volume 6388 of *Lecture Notes in Computer Science*, pages 46–55. Springer, 2010.

[12] Robert P. W. Duin and Elzbieta Pekalska. Non-euclidean dissimilarities: Causes and informativeness. In *SSPR/SPR*, volume 6218 of *Lecture Notes in Computer Science*, pages 324–333. Springer, 2010.

[13] Robert P. W. Duin and Elzbieta Pekalska. The dissimilarity space: between structural and statistical pattern recognition. *Pattern Recognition Letters*, 2011.

[14] R.P.W. Duin. Relational discriminant analysis and its large sample size problem. In *ICPR*, pages Vol I: 445–449, 1998.

[15] R.P.W. Duin, D. de Ridder, and D.M.J. Tax. Experiments with object based discriminant functions; a featureless approach to pattern recognition. *Pattern Recognition Letters*, 18(11-13):1159–1166, 1997.

[16] R.P.W. Duin and E. Pękalska. Structural inference of sensor-based measurements. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 41–55, 2006.

[17] R.P.W. Duin and E. Pękalska. On refining dissimilarity matrices for an improved nn learning. In *ICPR*, pages 1–4, 2008.

[18] R.P.W. Duin, E. Pękalska, and D. de Ridder. Relational discriminant analysis. *Pattern Recognition Letters*, 20(11-13):1175–1181, 1999.

[19] R.P.W. Duin, E. Pękalska, A. Harol, W.-J. Lee, and H. Bunke. On euclidean corrections for non-euclidean dissimilarities. In *SSPR/SPR*, pages 551–561, 2008.

[20] R.P.W. Duin and D.M.J. Tax. Statistical pattern recognition. In C. H. Chen and P. S. P. Wang, editors, *Handbook of Pattern Recognition and Computer Vision, Third Edition*, pages 3–24. World Scientific, 2005.

[21] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, London, 2^{nd} edition, 1990.

[22] L. Goldfarb. A new approach to pattern recognition. In L.N. Kanal and A. Rosenfeld, editors, *Progress in Pattern Recognition*, volume 2, pages 241–402. Elsevier, 1985.

[23] T. Graepel, R. Herbrich, P. Bollmann-Sdorra, and K. Obermayer. Classification on pairwise proximity data. In *Advances in Neural Information System Processing 11*, pages 438–444, 1999.

[24] B. Haasdonk. Feature space interpretation of SVMs with indefinite kernels. *IEEE TPAMI*, 25(5):482–492, 2005.

[25] B. Haasdonk and H. Burkhardt. Invariant kernel functions for pattern analysis and machine learning. *Machine Learning*, 68(1):35–61, 2007.

[26] B. Haasdonk and E. Pękalska. Indefinite kernel fisher discriminant. In *ICPR*, pages 1–4, 2008.

[27] Anil K. Jain, Robert P. W. Duin, and Jianchang Mao. Statistical pattern recognition: A review. *IEEE Trans. Pattern Anal. Mach. Intell*, 22(1):4–37, 2000.

[28] Anil K. Jain and Douglas E. Zongker. Representation and recognition of handwritten digits using deformable templates. *IEEE Trans. Pattern Anal. Mach. Intell*, 19(12), 1997.

[29] Risi Imre Kondor and Tony Jebara. A kernel between sets of vectors. In *ICML*, pages 361–368, 2003.

[30] Julian Laub, Volker Roth, Joachim M. Buhmann, and Klaus-Robert Müller. On the information and representation of non-euclidean pairwise data. *Pattern Recognition*, 39(10):1815–1826, 2006.

[31] Cheng Soon Ong. Kernels: Regularization and optimization, 2005.

[32] C.S. Ong, S. Mary, X.and Canu, and Smola A.J. Learning with non-positive kernels. In *Int. Conference on Machine Learning*, pages 639–646, Brisbane, Australia, 2004.

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

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

[35] E. Pękalska and R.P.W Duin. *The Dissimilarity Representation for Pattern Recognition. Foundations and Applications*. World Scientific, Singapore, 2005.

[36] E. Pękalska and R.P.W. Duin. Dissimilarity-based classification for vectorial representations. In *ICPR (3)*, pages 137–140, 2006.

[37] E. Pękalska and R.P.W. Duin. Beyond traditional kernels: Classification in two dissimilarity-based representation spaces. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 38(6):729–744, Nov. 2008.

[38] E. Pękalska and B. Haasdonk. Kernel discriminant analysis with positive definite and indefinite kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, accepted.

[39] E. Pękalska, A. Harol, R.P.W. Duin, B. Spillmann, and H. Bunke. Non-euclidean or non-metric measures can be informative. In *SSPR/SPR*, pages 871–880, 2006.

[40] E. Pękalska, P. Paclík, and R.P.W. Duin. A Generalized Kernel Approach to Dissimilarity Based Classification. *J. of Machine Learning Research*, 2(2):175–211, 2002.

[41] Robert J. Schalkoff. *Artificial Neural Networks*. McGraw-Hill Higher Education, 1997.

[42] B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.

[43] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge University Press, UK, 2004.

[44] S. Theodoridis and K. Koutroumbas. *Pattern Recognition, 4th Edition*. Academic Press, 2008.

[45] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., 1998.

The Dissimilarity Representation for Non-Euclidean Pattern Recognition: Introduction and Examples

Tutorial
Pucon, Chile, 14 November 2011

Robert P.W. Duin, Delft University of Technology

(In cooperation with Elżbieta Pękalska, Univ. of Manchester)

Pattern Recognition Lab
Delft University of Technology
The Netherlands

//PRLab.tudelft.nl/~duin

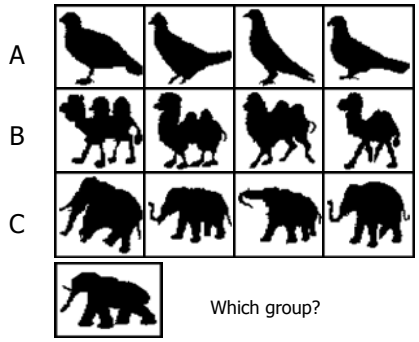
TUDelft

Contents

TUDelft

Pattern Recognition Problems

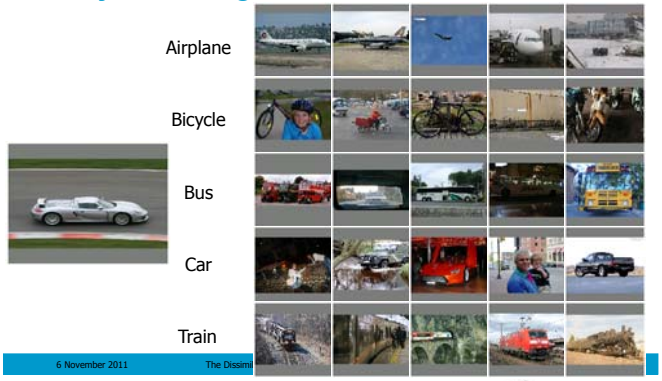
Blob Recognition



Which group?

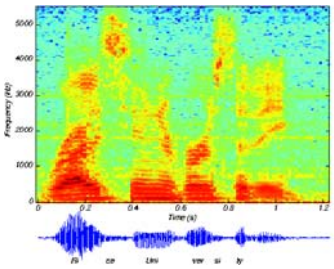
TUDelft

Object Recognition



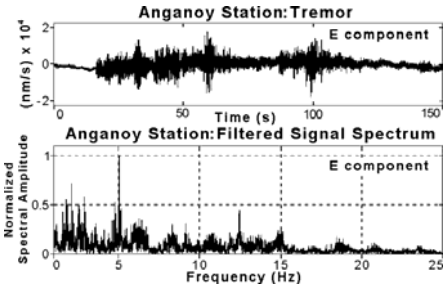
TUDelft

Pattern Recognition: Speech



TUDelft

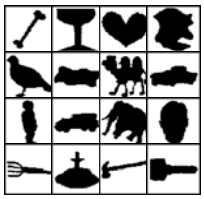
Pattern Recognition: Seismics

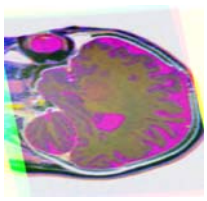


Earthquakes

TUDelft

Pattern Recognition Problems





To which class belongs an image

To which class (segment) belongs every pixel?

Where is an object of interest (detection); What is it (classification)?

TUDelft

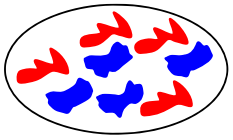
Pattern Recognition: Shape Recognition

Pattern Recognition is very often Shape Recognition:

- Images: B/W, grey value, color, 2D, 3D, 4D
- Time Signals
- Spectra

TUDelft

Pattern Recognition: Shapes



Examples of objects for different classes

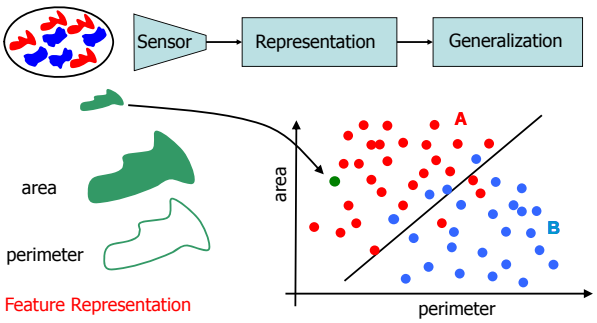
Object of unknown class to be classified

A ? B

TUDelft

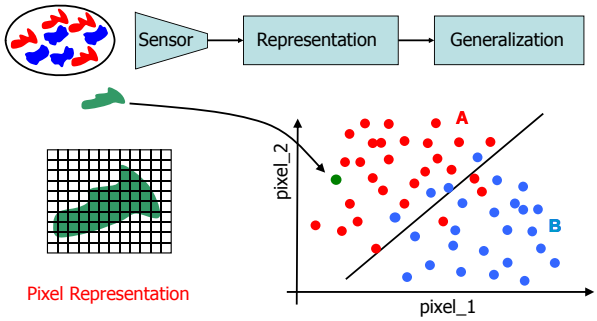
Vector Representation

Pattern Recognition System



TUDelft

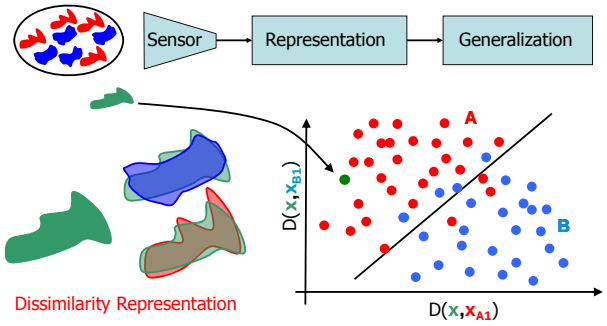
Pattern Recognition System



Pixel Representation

TUDelft

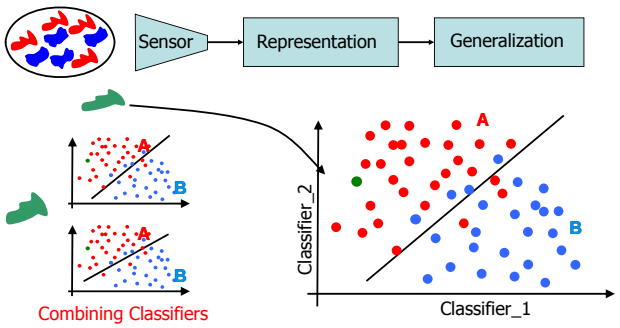
Pattern Recognition System



Dissimilarity Representation

TUDelft

Pattern Recognition System



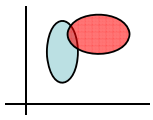
Combining Classifiers

TUDelft

Good Representations

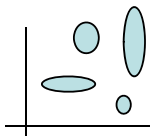
- Class specific

  Different classes should be represented in different positions in the representation space.



- Compact

  Every class should be represented in a small set of finite domains.

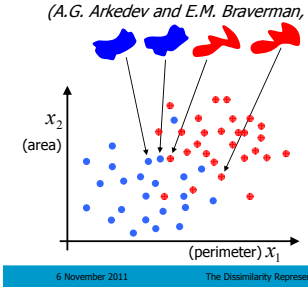


TUDelft

Compactness

Representations of real world similar objects are close.
There is no ground for any generalization (induction) on representations that do not obey this demand.

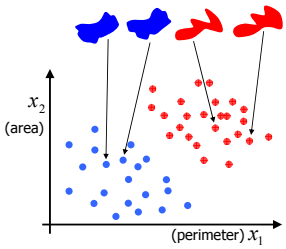
(A.G. Arkedev and E.M. Braverman, Computers and Pattern Recognition, 1966.)



The compactness hypothesis is not sufficient for perfect classification as dissimilar objects may be close.
→ class overlap
→ probabilities

TUDelft

True Representations



Similar objects are close
and
Dissimilar objects are distant.

→ no probabilities needed, domains are sufficient!

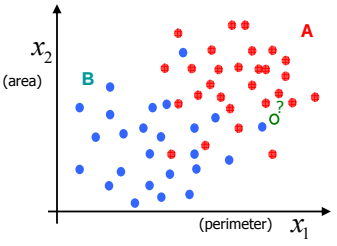
TUDelft

Distances and Densities

? to be classified as

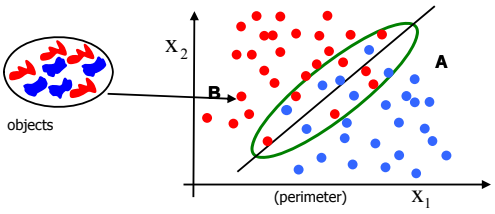
B – because it is most close to an object B

A – because the local density of A is larger.



TUDelft

Features Reduce



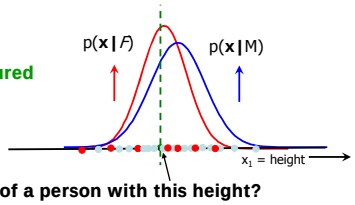
Due to reduction essentially different objects are represented identically.

→ The feature space representation needs a statistical, probabilistic generalization

TUDelft

Probabilistic Generalization

x = height measured



What is the gender of a person with this height?

Best guess is to choose the most 'probable' class (→ small error).

⇒ Good for overlapping classes.

⇒ Assumes the existence of a probabilistic class distribution and a representative set of examples.

TUDelft

Bayes decision rule, formal

Bayes:

$$p(A|x) > p(B|x) \rightarrow \text{A else B}$$

$$\frac{p(x|A)\,p(A)}{p(x)} > \frac{p(x|B)\,p(B)}{p(x)} \rightarrow \text{A else B}$$

$$p(x|A)\,p(A) > p(x|B)\,p(B) \rightarrow \text{A else B}$$

2-class problems: $S(x) = p(x|A)\,p(A) - p(x|B)\,p(B) > 0 \rightarrow$ A else B

n-class problems: $\text{Class}(x) = \text{argmax}_{\omega}(p(x|\omega)\,p(\omega))$

TUDelft

Density estimation

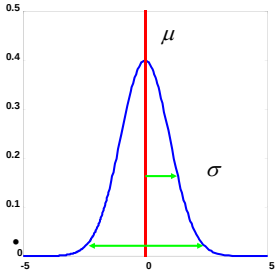
- The density is defined on the whole feature space.
- Around object x, the density is defined as:

$$p(x) = \frac{dP(x)}{dx} = \left(\frac{\text{fraction of objects}}{\text{volume}}\right)$$

- Given n measured objects, e.g. person's height (m) how can we estimate $p(x)$?

TUDelft

The Gaussian distribution (3)

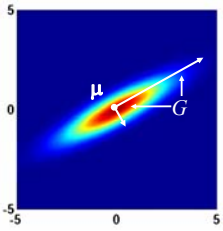


- Normal distribution = Gaussian distribution
- Standard normal distribution: $\mu = 0,\ \sigma^2 = 1$
- 95% of data between $[\,\mu - 2\sigma,\ \mu + 2\sigma\,]$ (in 1D!)

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

TUDelft

Multivariate Gaussians



$$G = \begin{bmatrix} 3 & 1\tfrac{1}{2} \\ 1\tfrac{1}{2} & 2 \end{bmatrix}$$

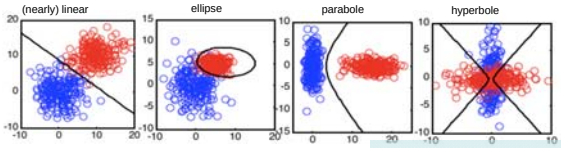
- k - dimensional density:

$$p(x) = \frac{1}{\sqrt{2\pi^k \det(G)}} \exp\left(-\frac{1}{2}(x-\mu)^T G^{-1}(x-\mu)\right)$$

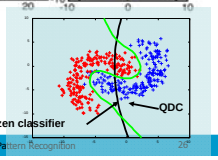
TUDelft

Quadratic discriminant functions

$$R(x) = -\frac{1}{2}(x-\hat{\mu}_A)^T\hat{\Sigma}_A^{-1}(x-\hat{\mu}_A) + \frac{1}{2}(x-\hat{\mu}_B)^T\hat{\Sigma}_B^{-1}(x-\hat{\mu}_B) + \text{const}$$

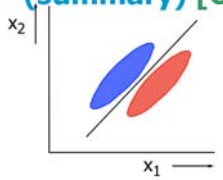


QDC assumes that classes are normally distributed. Wrong decision boundaries are estimated if this does not hold.



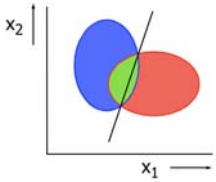
TUDelft

Linear discriminant function (summary) [G]



Normal distributions with equal covariance matrices Σ are optimally separated by a linear classifier

$$R(x) = (\mu_A - \mu_B)^T\Sigma^{-1}x + \text{const}$$



Optimal classifier for normal distributions with unequal covariance matrices Σ_A and Σ_B can be approximated by:

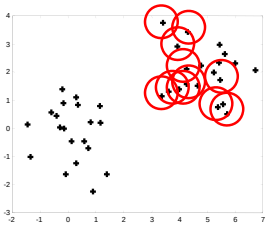
$$R(x) = (\mu_A - \mu_B)^T(p(A)\Sigma_A + p(B)\Sigma_B)^{-1}x + \text{const}$$

TUDelft

Parzen density estimation (1)

- Fix volume of bin, vary positions of bins, add contribution of each bin
- Define 'bin'-shape (kernel):

$$K(\mathbf{r}) > 0$$
$$\int K(\mathbf{r})\,d\mathbf{r} = 1$$



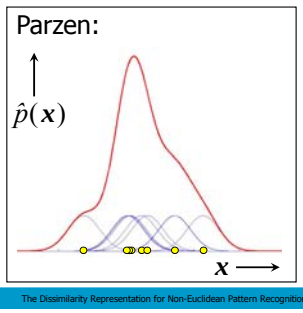
- For test object $\mathbf{z}$ sum all bins

$$p(\mathbf{z}) = \frac{1}{hn}\sum_i K\left(\frac{\mathbf{z}-\mathbf{x}_i}{h}\right)$$

TUDelft

Parzen density estimation (2)

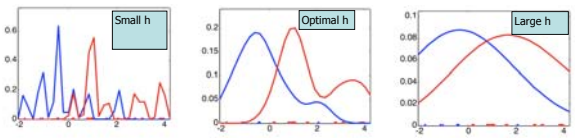
- With Gaussian kernel: $K(x) = \frac{1}{h\sqrt{2\pi}}\exp\left(-\frac{x^2}{2h^2}\right)$



TUDelft

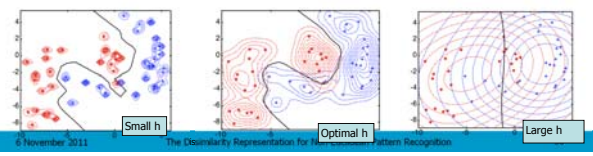
Parzen: density estimates vs the smoothing parameter

1D



Increasing smoothing parameter h

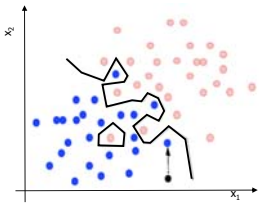
2D



TUDelft

Nearest neighbor rule (1-NN rule)

Assign a new object to the class of the nearest neighbor in the training set.



1-NN rule:

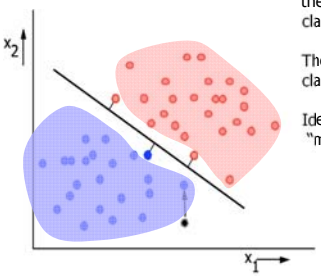
- Often relies on the Euclidean distance. Other distance measures can be used.
- Insensitive to prior probabilities!
- Scaling dependent. Features should be scaled properly.

There are no errors on the training set. The classifier is overtrained.

TUDelft

Support vector machine (SVM)

1995-2005



For linearly-separable classes find the few objects that determine the classifier. These are **support vectors.**

They have the same distance to the classifier: **the margin.**

Identical to "maximum-margin classifier"

$$S(x) = \sum_i \alpha_i (x_i^T x)$$

$$S(x) = w^T x, \min(w^T w)$$

TUDelft

Pixel Representation

Measuring Human Relevant Information

A

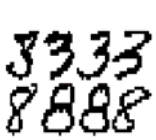
B

Nearest neighbours sorted:

B A A A B A A B A B

TUDelft

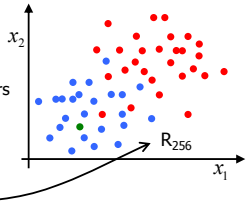
Pixel Representation



16 x 16

Pixels

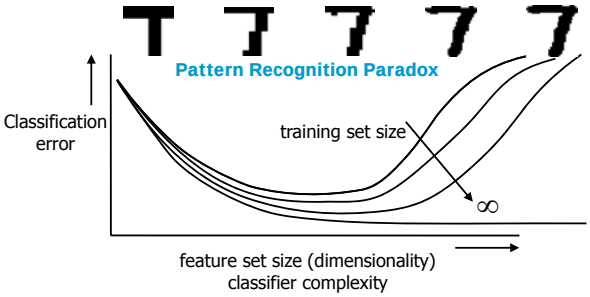
Features
Shapes
Moments
Fourier descriptors
Faces
Morphology



Pixels are more general, initially complete representation
Large datasets are available → good results for OCR

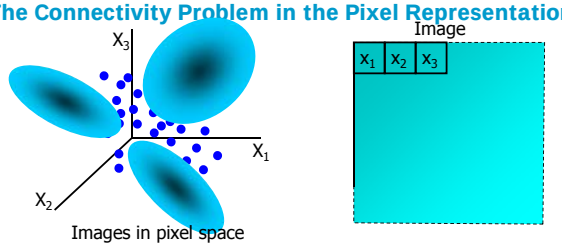
TUDelft

Peaking Phenomenon, Overtraining Curse of Dimensionality, Rao's Paradox



TUDelft

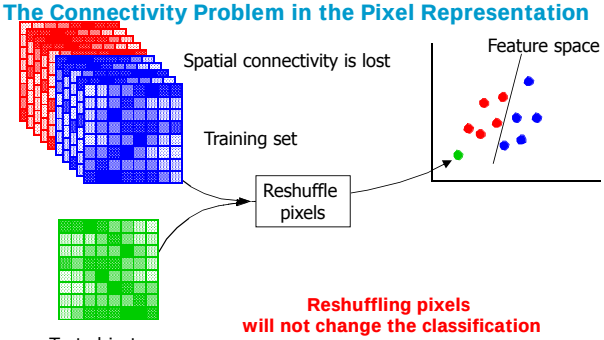
The Connectivity Problem in the Pixel Representation



Dependent (connected) measurements are represented independently.
The dependency has to be refound from the data.

TUDelft

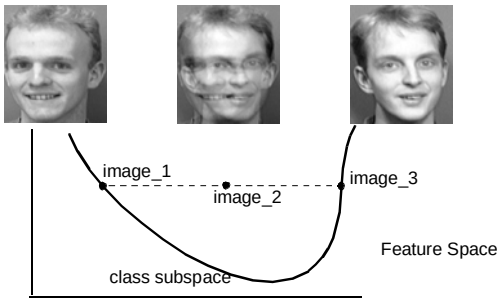
The Connectivity Problem in the Pixel Representation



Reshuffling pixels will not change the classification

TUDelft

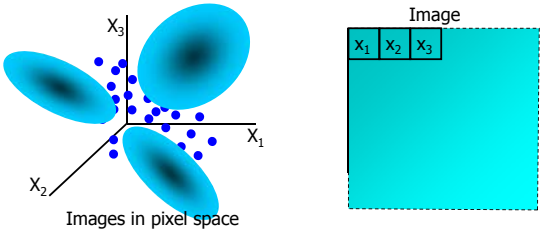
The Connectivity Problem in the Pixel Representation



Interpolation does not yield valid objects

TUDelft

The Connectivity Problem in the Pixel Representation



Dependent (connected) measurements are represented independently.
The dependency has to be refound from the data.

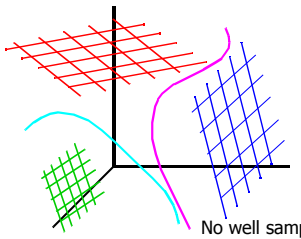
TUDelft

Reasons for selecting a pixel (sample) representation

- No good features can be found
- Sufficient training samples are available
- Direct, fast classification of the image
  (linear classifier == convolution)

TUDelft

Domains instead of Densities



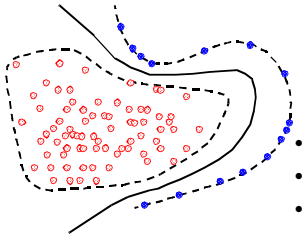
No well sampled training sets are needed.

Statistical classifiers have still to be developed.

Class structure ←→ Object invariants

TUDelft

Domain-based Classification

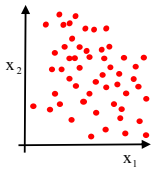


- Do not trust class densities.
- Estimate for each class a domain.
- Outlier sensitive.
- Distances instead of densities

How to construct domain based classifiers?

TUDelft

Wrong Intuition of High Dimensional Spaces



2D-intuition does not work for high dimensional spaces

All points are boundary points
1000 normally distr. points in R20: 95% on the convex hull.

Points tend to have equal distances
Squared Euclidean distances of points in R1024 are distributed as $N(1024, 32\sqrt{2})$, so distances are all equal within 10%.

Class overlap is not visible
1000 points of two 5% overlapping classes in R50 can be linearly separable

Moreover:
do real world measurements with n > 100 really exist?
⇒ Subspace approaches

TUDelft

Reasons for selecting a pixel (sample) representation

- No good features can be found
- Sufficient training samples are available
- Direct, fast classification of the image (linear classifier == convolution)

TUDelft

Vector Representations

- Features: reduce → class overlap
- Pixels: dimensionality problems
- Dissimilarities?

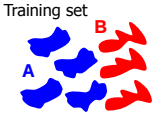
TUDelft

Dissimilarity Representation

Dissimilarity Representation

Not used by NN Rule

Training set



Dissimilarities d_{ij} between all training objects

$$D_T = \begin{pmatrix} d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} & d_{17} \\ d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} & d_{27} \\ d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} & d_{37} \\ d_{41} & d_{42} & d_{43} & d_{44} & d_{45} & d_{46} & d_{47} \\ d_{51} & d_{52} & d_{53} & d_{54} & d_{55} & d_{56} & d_{57} \\ d_{61} & d_{62} & d_{63} & d_{64} & d_{65} & d_{66} & d_{67} \\ d_{71} & d_{72} & d_{73} & d_{74} & d_{75} & d_{76} & d_{77} \end{pmatrix}$$

Unlabeled object **x** to be classified

$$d_x = (d_{x1}\ d_{x2}\ d_{x3}\ d_{x4}\ d_{x5}\ d_{x6}\ d_{x7})$$

The traditional Nearest Neighbor rule (template matching) finds:
$\text{label}(\text{argmin}_{\text{trainset}}\{d_{xi}\})$,
without using D_T. Can we do any better?

TUDelft

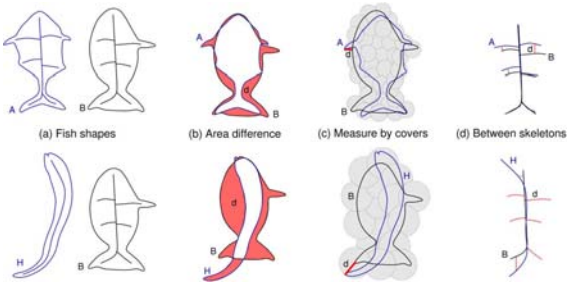
Dissimilarities – Possible Assumptions

Metric

1. Positivity: $d_{ij} \geq 0$
2. Reflexivity: $d_{ii} = 0$
3. Definiteness: $d_{ij} = 0$ objects i and j are identical
4. Symmetry: $d_{ij} = d_{ji}$
5. Triangle inequality: $d_{ij} < d_{ik} + d_{kj}$
6. Compactness: if the objects i and j are very similar then $d_{ij} < \delta$.
7. True representation: if $d_{ij} < \delta$ then the objects i and j are very similar.
8. Continuity of d.

TUDelft

Examples Dissimilarity Measures (1)

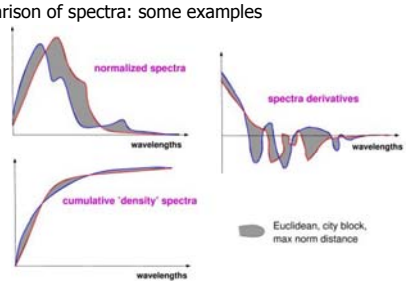


The measure should be descriptive. If there is no preference, a number of measures can be combined.

TUDelft

Examples Dissimilarity Measures (2)

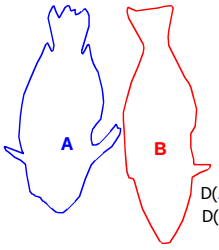
Comparison of spectra: some examples



In real applications, the dissimilarity measure should be robust to noise and small aberrations in the (raw) measurements.

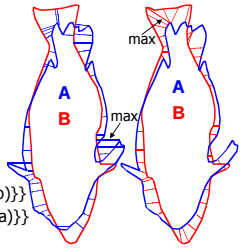
TUDelft

Examples Dissimilarity Measures (3)



Dist(**A**,**B**):
a ∈ **A**, points of **A**
b ∈ **B**, points of **B**
d(a,b): Euclidean distance

D(**A**,**B**) = max_a{min_b{d(a,b)}}
D(**B**,**A**) = max_b{min_a{d(b,a)}}



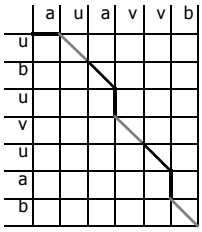
Hausdorff Distance (metric):
DH = max{max_a{min_b{d(a,b)}} , max_b{min_a{d(b,a)}}}

D(**A**,**B**) ≠ D(**B**,**A**)

Modified Hausdorff Distance (non-metric):
DM = max{mean_a{min_b{d(a,b)}},mean_b{min_a{d(b,a)}}}

TUDelft

Examples Dissimilarity Measures (4)



$X = (x_1, x_2, \ldots, x_k)$ $Y = (y_1, y_2, \ldots, y_n)$

$D_E(X,Y)$: Σ edit operations $X \Rightarrow Y$ (insertions, deletions, substitutions)

DE(snert ,meer) = 3:
snert ⇒ seert ⇒ seer ⇒ meer

DE(ner ,meer) = 2:
ner ⇒ mer ⇒ meer

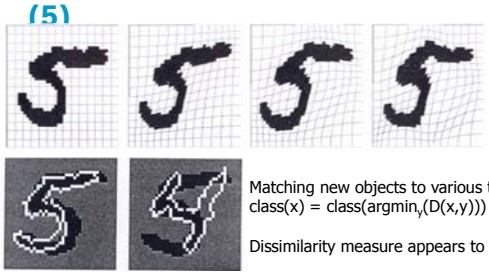
Possibly weighted

Triangle inequality ⇒ computationally feasible

D(aa,bb) < D(abcdef,bcdd)

TUDelft

Examples Dissimilarity Measures (5)



Matching new objects to various templates:
class(x) = class(argmin_y(D(x,y)))

Dissimilarity measure appears to be non-metric.

A.K. Jain, D. Zongker, Representation and recognition of handwritten digit using deformable templates, IEEE-PAMI, vol. 19, no. 12, 1997, 1386-1391.

TUDelft

Prospect of Dissimilarity based Representations: Zero

Let us assume that we deal with true representations:
$d_{ab} < \delta$ if and only if the objects a and b are very similar.

If δ is sufficiently small then a and b belong to the same class, as b is just a minor distortion of a (assuming true representations).

However, as Prob(b) > 0, there will be such an object for sufficiently large training sets ⇒ zero classification error possible!

⇒ Dissimilarity representation can be a true representation

TUDelft

Why a Dissimilarity Representation?

Many (exotic) dissimilarity measures are used in pattern recognition
- they may solve the connectivity problem (e.g. pixel based features)
- they may offer a way to integrate structural and statistical approaches e.g. by graph distances.

Prospect of zero-error classifiers by avoiding class overlap

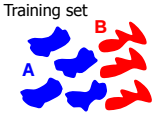
Better rules than the nearest neighbor classifier appear possible (more accurate, faster)

TUDelft

Classification of Dissimilarity Data

Alternatives for the Nearest Neighbor Rule

Training set



Dissimilarities d_{ij} between all training objects

$$D_T = \begin{pmatrix} d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} & d_{17} \\ d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} & d_{27} \\ d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} & d_{37} \\ d_{41} & d_{42} & d_{43} & d_{44} & d_{45} & d_{46} & d_{47} \\ d_{51} & d_{52} & d_{53} & d_{54} & d_{55} & d_{56} & d_{57} \\ d_{61} & d_{62} & d_{63} & d_{64} & d_{65} & d_{66} & d_{67} \\ d_{71} & d_{72} & d_{73} & d_{74} & d_{75} & d_{76} & d_{77} \end{pmatrix}$$

Unlabeled object **x** to be classified

$$d_x = (d_{x1}\ d_{x2}\ d_{x3}\ d_{x4}\ d_{x5}\ d_{x6}\ d_{x7})$$

1. Dissimilarity Space
2. Embedding

Pekalska, The dissimilarity representation for PR. World Scientific, 2005.

TUDelft

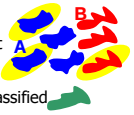
The Dissimilarity Space

Alternative 1: Dissimilarity Space

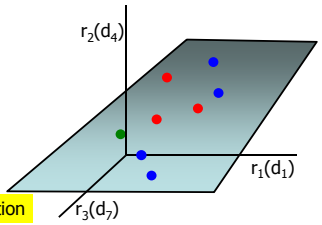
Dissimilarities

$$D_T = \begin{pmatrix} d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} & d_{17} \\ d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} & d_{27} \\ d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} & d_{37} \\ d_{41} & d_{42} & d_{43} & d_{44} & d_{45} & d_{46} & d_{47} \\ d_{51} & d_{52} & d_{53} & d_{54} & d_{55} & d_{56} & d_{57} \\ d_{61} & d_{62} & d_{63} & d_{64} & d_{65} & d_{66} & d_{67} \\ d_{71} & d_{72} & d_{73} & d_{74} & d_{75} & d_{76} & d_{77} \end{pmatrix} \quad d_x = (d_{x1}\ d_{x2}\ d_{x3}\ d_{x4}\ d_{x5}\ d_{x6}\ d_{x7})$$

Given labeled training set



Unlabeled object to be classified



Selection of 3 objects for representation

TUDelft

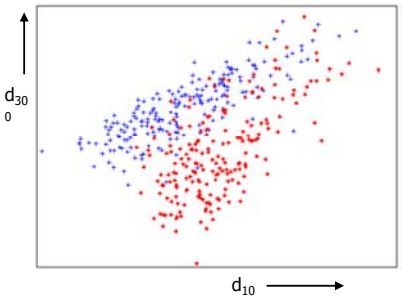
Example Dissimilarity Space: NIST Digits 3 and 8

Example of raw data

TUDelft

Example Dissimilarity Space: NIST Digits 3 and 8

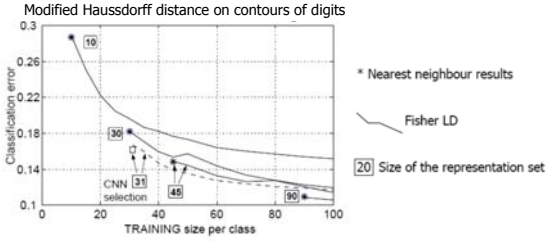
NIST digits: Hamming distances of 2 x 200 digits



TUDelft

Dissimilarity Space Classification ←→ Nearest Neighbor Rule

Modified Haussdorff distance on contours of digits

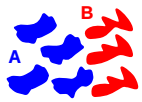


Dissimilarity based classification outperforms the nearest neighbor rule.

TUDelft

Embedding of (non-Euclidean) Dissimilarities

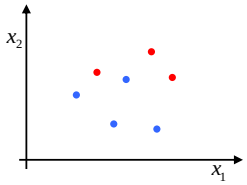
Alternative 2: Embedding



→ Dissimilarity matrix D → X

Training set

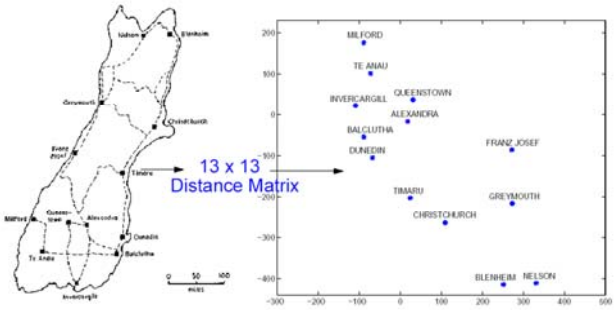
Is there a feature space for which Dist(X,X) = D ?



Position points in a vector space such that their Euclidean distances → D

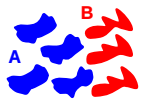
TUDelft

Embedding

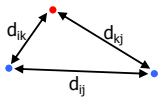


TUDelft

Embedding of non-metric measurements



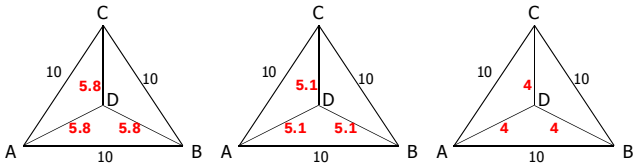
→ Dissimilarity matrix D → X



If the dissimilarity matrix cannot be explained from a vector space,
(e.g. for Hausdorff and Hamming distance of images)
or if $d_{ij} > d_{ik} + d_{kj}$ (triangle inequality not satisfied)
embedding in Euclidean space not possible
→ Pseudo-Euclidean embedding

TUDelft

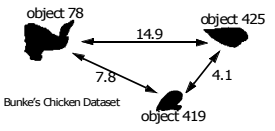
Euclidean - Non Euclidean - Non Metric



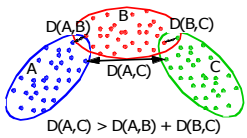
TUDelft

Non-metric distances

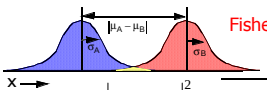
Weighted-edit distance for strings



Single-linkage clustering

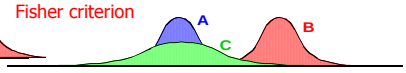


D(A,C) > D(A,B) + D(B,C)



$$J(A,B) = \frac{|\mu_A - \mu_B|^2}{\sigma_A^2 + \sigma_B^2}$$

Fisher criterion



$J(A,C) = 0 \quad J(A,B) = \text{large}$

$J(C,B) = \text{small} \neq J(A,B)$

TUDelft

(Pseudo-)Euclidean Embedding

m×m D is a given, imperfect dissimilarity matrix of training objects.

Construct inner-product matrix: $B = -\frac{1}{2} J D^{(2)} J \quad J = I - \frac{1}{m}\mathbf{11}$

Eigenvalue Decomposition , $B = Q \Lambda Q^T$

Select k eigenvectors: $X = Q_k \Lambda_k^{\frac{1}{2}}$ (problem: $\Lambda_k < 0$)

Let $\Im_k$ be a k x k diag. matrix, $\Im_k(i,i) = \text{sign}(\Lambda_k(i,i))$

$\Lambda_k(i,i) < 0 \rightarrow$ Pseudo-Euclidean

n×m D_z is the dissimilarity matrix between new objects and the training set.

The inner-product matrix: $B_z = -\frac{1}{2}(D_z^{(2)} J - \frac{1}{n}\mathbf{11}^T D^{(2)} J)$

The embedded objects: $Z = B_z Q_k |\Lambda_k|^{-\frac{1}{2}} \Im_k$

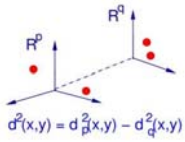
TUDelft

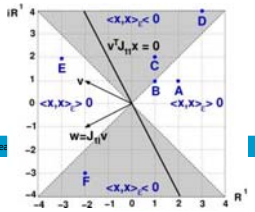
PES: Pseudo-Euclidean Space (Krein Space)

If D is non-Euclidean, B has p positive and q negative eigenvalues.
A pseudo-Euclidean space $\mathcal{E}$ with signature (p,q), k =p+q, is a non-degenerate inner product space $\Re_k = \Re_p \oplus \Re_q$ such that:

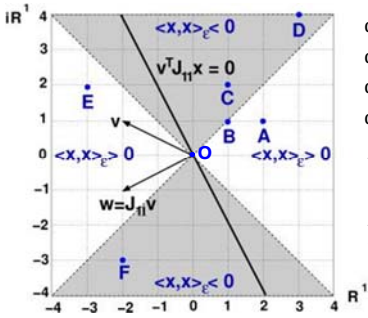
$$\langle x, y \rangle_\varepsilon = x^T \Im_{pq} y = \sum_{i=1}^{p} x_i y_i - \sum_{j=p+1}^{q} x_j y_j \qquad \Im_{pq} = \begin{bmatrix} I_{p\times p} & 0 \\ 0 & -I_{q\times q} \end{bmatrix}$$

$$d_\varepsilon^2(x, y) = \langle x - y, x - y \rangle_\varepsilon = d_p^2(x, y) - d_q^2(x, y)$$





Distances in PES



$d^2(O, A) > 0$
$d^2(O, E) > 0$
$d^2(O, B) = 0$
$d^2(O, D) < 0$

All points in the grey area are closer to O than O itself !?

Any point has a negative square distance to some points on the line $v^T Jx=0$.
Can it be used as a classifier?
Can we define a margin as in the SVM?

TUDelft

PE Space ←→ Kernels

$$K(x, y) = -\tfrac{1}{2} J D(x, y)^{(2)} J \qquad J = I - \tfrac{1}{m}\mathbf{1}\mathbf{1}$$

may by considered as a kernel. If

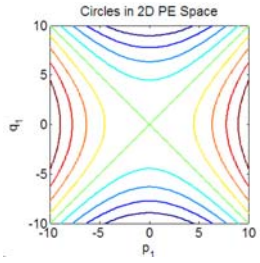
$$K(x, y) = < L(x), L(y) >$$

- The **kernel trick** may be used: operations defined on inner products in kernel space can be operated directly on K(x,y) **without embedding**!
- True for **Mercer kernels** (all eigenvalues ≥ 0).
- Difficult for **indefinite kernels**.
- Studying classifiers in **PE space** is studying the indefinite **kernel space**.
- Dissimilarities are more informative than kernels (due to normalization).

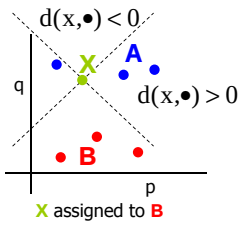
TUDelft

Classifiers in Pseudo-Euclidean Space

Distance based classifiers in PE Space



Metric in PE Space.
Equidistant points to the origin.



X assigned to B

Nearest Neighbour and
Nearest Mean can be properly defined.
SVM ? What is the distance to a line?

TUDelft

SVM in PE Space

- SVM on indefinite kernels may not converge as Mercer's conditions are not fulfilled.
- However, if it converges the solution is proper:

  $|\, w^T \Im w \,|$

  is minimized.
- *See also: B. Haasdonk, Feature Space Interpretation of SVMs with Indefinite Kernels, IEEE PAMI, 24, 482-492, 2005.*

TUDelft

Densities in PE Space

- Densities can be defined in a vector space on de basis of volumes, without the need of a metric.
- Density estimates however, often need a metric. E.g. the Parzen estimator:

$$\hat{f}(x) = \frac{1}{n}\sum_{y_i} c \exp(-\frac{d(x,y_i)^2}{2h^2})$$

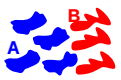
needs a distance definition $d(x,y)$.

- There is no problem, however, in case for all objects $d(x,y) > 0$.
- How can Gaussian densities be defined?
- Note that QDA in PES is identical to the QDA in AES as the signature cancels. The relation with a Gaussian distribution, however, is lost.

TUDelft

Dissimilarity based classifiers compared

Dissimilarity based classification procedured compared

Training set 

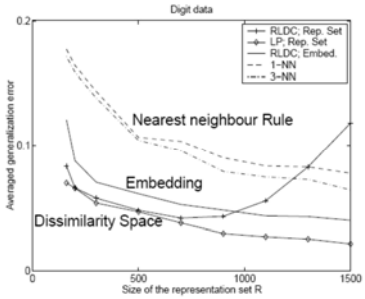
→ Dissimilarity matrix D

Test object x  → Dissimilarities d_x with training set

1. Nearest Neighbour Rule
2. Reduce training set to representation set
⇒ dissimilarity space
3. Embedding:Select large $\Lambda_{ii} > 0 \Rightarrow$ Euclidean space
Select large $|\Lambda_{ii}| > 0 \rightarrow$ pseudo-Euclidean space
} discriminant function

TUDelft

Three Approaches Compared for the Zongker Data



Dissimilarity Space equivalent to Embedding better than Nearest Neighbour Rule

TUDelft

Polygon Data

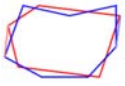
Convex Pentagons 

Heptagons 

no class overlap
zero error

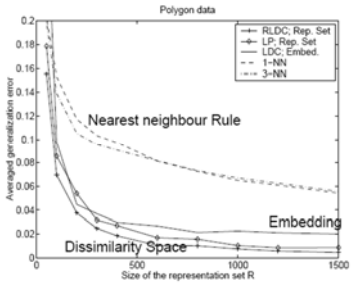
Minimum edge length: 0.1 of maximum edge length

Distance measures: Hausdorff $D = \max \{ \max_i(\min_j(d_{ij})) , \max_j(\min_i(d_{ij})) \}$.
Modified Hausdorff $D = \max \{\text{mean}_i(\min_j(d_{ij})), \text{mean}_j(\min_i(d_{ij})) \}$. (no metric!)
d_{ij} = distance between vertex i of polygon_1 and vertex j of polygon_2.
Polygons are scaled and centered.

Find the largest of the smallest vertex distances 

TUDelft

Dissimilarity Based Classification of Polygons



Zero error difficult to reach!

TUDelft

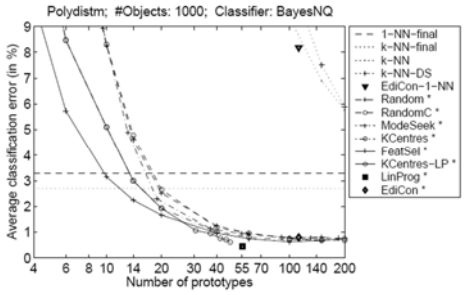
Prototype Selection

Assume D(T,R) are the distances between a training set T and a representation set R.
A classifier can be trained:
- on the distances directly
- in dissimilarity spaces
- in embedded spaces defined by D(R,R)

Selection of prototypes $R \subset T$:
- Random
- k-centres, mode seeking or some clustering procedure
- Feature selection methods
- Editting and condensing methods
- Sparse linear programming methods (L_1-norm SVM)

TUDelft

Prototype Selection: Polygon Dataset



The classification performance of the quadratic Bayes Normal classifier and the k-NN in dissimilarity spaces and the direct k-NN, as a function of the number of selected prototypes. Note that for 10-20 prototypes already better results are obtained than by using 1000 objects in the NN rules.

Dissimilarity Representation

- Based on a pairwise comparison of objects
- Alternative to features for using expert knowledge
- Various ways of construction vector spaces, useful for traditional classifiers.
- May show good performances compared to nearest neigbour rule

TUDelft

Non-Euclidean Representations

Causes, Corrections, Informativeness

Non-Euclidean Representations: Causes

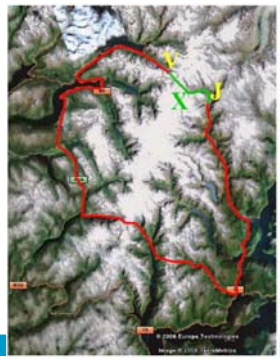
Computational Noise

Even for Euclidean distance matrices zero eigenvalues may show negative, e.g:

- X = N(50,20) : 50 points in 20 dimensions
- D = Dist(X): 50 x 50 distance matrix
- Expected: 49-20 = 29 zero eigenvalues
- Found: 15 negative eigenvalues

TUDelft

Lack of information



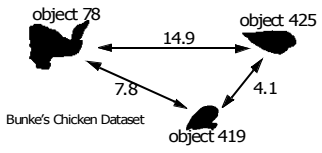
1800:
Crossing the Jostedalsbreen was impossible.
Travelling around (200 km) lasted 5 days.
Untill the shared point X was found.
People could visit each other in 8 hours.

D(V,J) = 5 days
D(V,X) = 4 hours
D(X,J) = 4 hours

TUDelft

Computational Problems

Large distances are overestimated due to computational problems

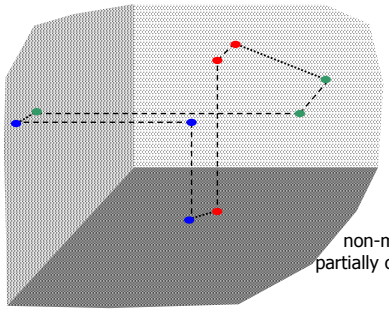


Weighted edit distance for strings

TUDelft

Projections - Occlusions

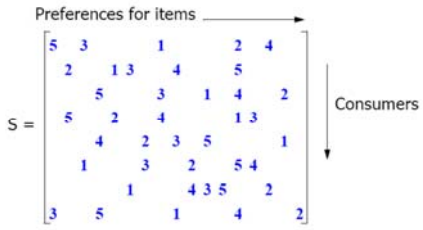
Small distances are underestimated



non-metric data due to partially observed projections

TUDelft

Projections - Occlusions

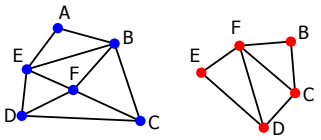


Example: consumer preferences for recommendation systems

TUDelft

Graph Matching → Dissimilarities

Representation by Connected Graphs



Graph (Nodes, Edges, Attributes)

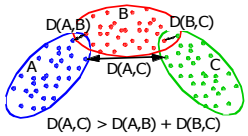
Distance (Graph_1, Graph_2)

TUDelft

Intrinsicly Non-Euclidean Dissimilarity Measures Single Linkage

Distance(Table,Book) = 0
Distance(Table,Cup) = 0
Distance(Book,Cup) = 1

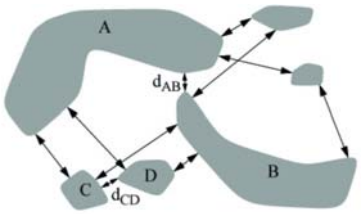
Single-linkage clustering



D(A,C) > D(A,B) + D(B,C)

TUDelft

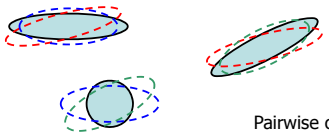
Boundary distances



A set of boundary distances may characterize sets of datapoints:
Distances → features

TUDelft

Intrinsicly Non-Euclidean Dissimilarity Measures Mahalanobis

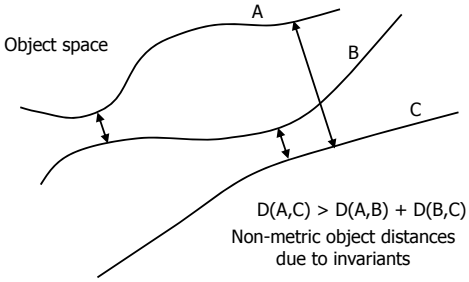


Pairwise comparison between
differently shaped data distributions

Different pairs → different comparison frameworks
→ non-Euclidean

TUDelft

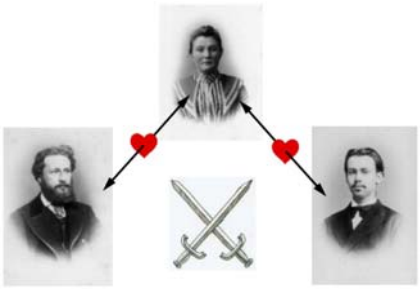
Intrinsicly Non-Euclidean Dissimilarity Measures Invariants



D(A,C) > D(A,B) + D(B,C)
Non-metric object distances
due to invariants

TUDelft

Intrinsicly Non-Euclidean Dissimilarity Measures



Non-Euclidean human relations

TUDelft

Objects may have an 'inner life'

In dissimilarity measures the 'inner life' of objects may be taken into account (e.g. invariants).

→ Objects cannot be represented anymore as points

→ Non-Euclidean dissimilarities

TUDelft

Causes of Non-Euclidean Dissimilarities

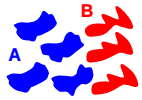
- **Computational / Observational Limitations**
  - numerical accuracy problems
  - overestimated large distances (too difficult to compute)
  - underestimated small distances (one-sided view of objects)
- **Essential non-Euclidean distance definitions**
  - the human distance concept differs from the mathematical one
  - no global framework
  - invariants

TUDelft

Euclidean corrections for non-Euclidean dissimilarities

SSSPR 2008
R.P.W. Duin, E. Pekalska, A. Harol, W.J. Lee and H. Bunke

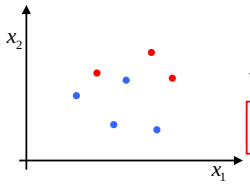
Alternative 2: Embedding



Training set

→ Dissimilarity matrix D → X

Is there a feature space for which Dist(X,X) = D ?



Position points in a vector space such that their Euclidean distances → D

However, many popular dissimilarity measures are not Euclidean!!

TUDelft

(Pseudo-)Euclidean Embedding

$m \times m$ D is a given, imperfect dissimilarity matrix of training objects.

Construct inner-product matrix: $B = -\frac{1}{2} J D^{(2)} J \quad J = I - \frac{1}{m}\mathbf{1}\mathbf{1}$

Eigenvalue Decomposition , $B = Q \Lambda Q^T$

Select k eigenvectors: $X = Q_k \Lambda_k^{\frac{1}{2}}$ (problem: $\Lambda_k < 0$)

Let $\Im_k$ be a $k \times k$ diag. matrix, $\Im_k(i,i) = \text{sign}(\Lambda_k(i,i))$

$\Lambda_k(i,i) < 0 \rightarrow$ Pseudo-Euclidean

$n \times m$ D_z is the dissimilarity matrix between new objects and the training set.

The inner-product matrix: $B_z = -\frac{1}{2}(D_z^{(2)} J - \frac{1}{n}\mathbf{1}\mathbf{1}^T D^{(2)} J)$

The embedded objects: $Z = B_z Q_k |\Lambda_k|^{-\frac{1}{2}} \Im_k$

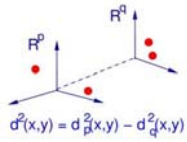
TUDelft

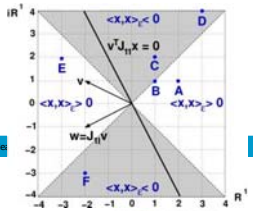
PES: Pseudo-Euclidean Space (Krein Space)

If D is non-Euclidean, B has p positive and q negative eigenvalues.
A pseudo-Euclidean space $\mathcal{E}$ with signature (p,q), k =p+q, is a non-degenerate inner product space $\Re_k = \Re_p \oplus \Re_q$ such that:

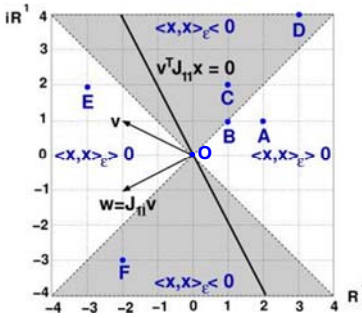
$$\langle x, y \rangle_\varepsilon = x^T \Im_{pq} y = \sum_{i=1}^{p} x_i y_i - \sum_{j=p+1}^{q} x_j y_j \qquad \Im_{pq} = \begin{bmatrix} I_{p\times p} & 0 \\ 0 & -I_{q\times q} \end{bmatrix}$$

$$d_\varepsilon^2(x, y) = \langle x - y, x - y \rangle_\varepsilon = d_p^2(x, y) - d_q^2(x, y)$$





Distances in PES



$d^2(O, A) > 0$
$d^2(O, E) > 0$
$d^2(O, B) = 0$
$d^2(O, D) < 0$

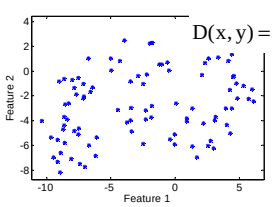
All points in the grey area are closer to O than O itself !?

Any point has a negative square distance to some points on the line $v^T Jx=0$.
Can it be used as a classifier?
Can we define a margin as in the SVM?

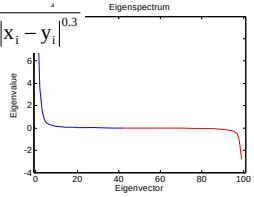
TUDelft

Pseudo-Euclidean Embedding

If D is non-Euclidean then B has **p** positive and **q** negative eigenvalues



$$D(x, y) = \sqrt[0.3]{\sum |x_i - y_i|^{0.3}}$$



Solutions:

- Remove all eigenvectors with small and negative eigenvalues
- or, take absolute values of eigenvalues and proceed
- or, construct a pseudo-Euclidean space

TUDelft

Correction Procedures PES ←→ ES

- **PES**: Pseudo Euclidean Space

  $$d_\varepsilon^2(x, y) = d_p^2(x, y) - d_q^2(x, y)$$

- **PES+**: Positive contributions only

  $$d_\varepsilon^2(x, y) = d_p^2(x, y)$$

- **AES**: Treat entire space as Euclidean

  $$d_\varepsilon^2(x, y) = d_p^2(x, y) + d_q^2(x, y)$$

TUDelft

Correction Procedures PES ←→ ES (2)

- **DEC**: Enlarging dissimilarities

  $$d_\varepsilon^2(x, y) \Leftarrow d_\varepsilon^2(x, y) + c, x \neq y$$

  use smallest c such that $D \succ 0$

- **Relax**: Relaxing dissimilarity measure

  $$d_\varepsilon(x, y) \Leftarrow d_\varepsilon(x, y)^{1/c}, c \geq 1$$

  use smallest c such that $D \succ 0$

TUDelft

Correction Procedures PES ←→ ES (3)

- **Laplace**:
  Adaptation of Laplace correction from spectral graph theory:

  $$d_\varepsilon^2(x, y) \Leftarrow \text{Norm}(d_\varepsilon^2(x, y))$$
  $$d_\varepsilon^2(x, y) \Leftarrow 1 - \delta(x, y) + d_\varepsilon^2(x, y)$$
  $$d_\varepsilon^2(x, y) \Leftarrow d_\varepsilon^2(x, y) + c, x \neq y \quad D \succ 0$$

- (unpublished result: a 'minimum' Laplace correction can be obtained by normalizing the dissimilarity matrix, followed by the DEC correction)

TUDelft

Example: Chickenpieces (H. Bunke, Bern)

BACK

BREAST

DRUMSTICK

THIGH-AND-BACK

WING

446 binary images, varying size, e.g.: 100 x 130

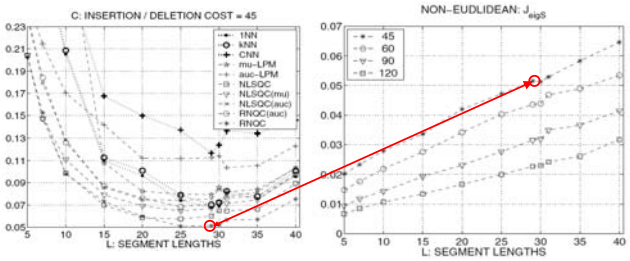
Andreu, G., Crespo, A., Valiente, J.M.: Selecting the toroidal self-organizing feature maps (TSOFM) best organized to object recogn. In: ICNN. (1997) 1341–1346.

Shape classification by weighted-edit distances (Bunke)

*Bunke, H., Buhler, U.: Applications of approximate string matching to 2D shape recognition. Pattern recognition **26** (1993) 1797–1812*

TUDelft

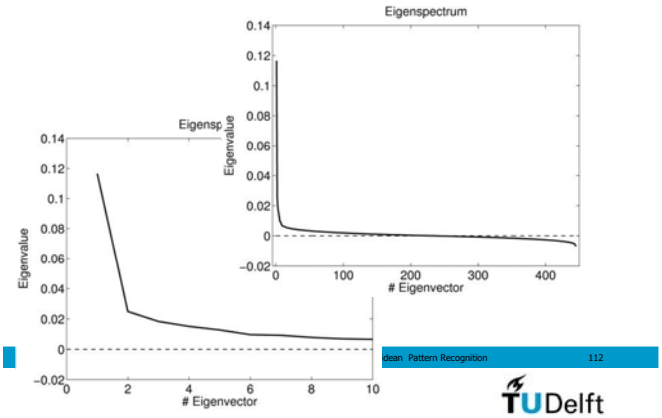
Classification Results for Various Dissimilarity Measures



Best classification result is for a very non-Euclidean dissimilarity measure !

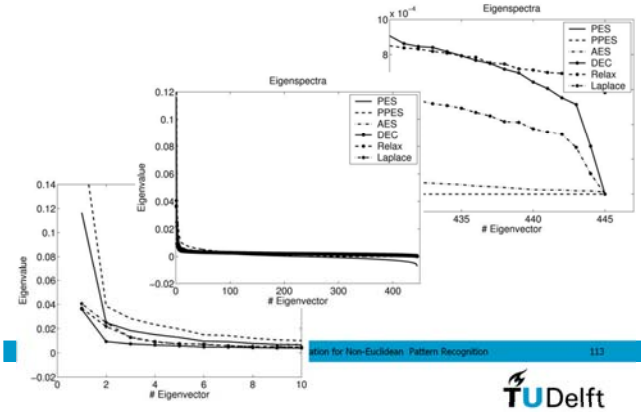
TUDelft

Eigenspectrum original data (PES)



TUDelft

Eigenspectra corrected data



TUDelft

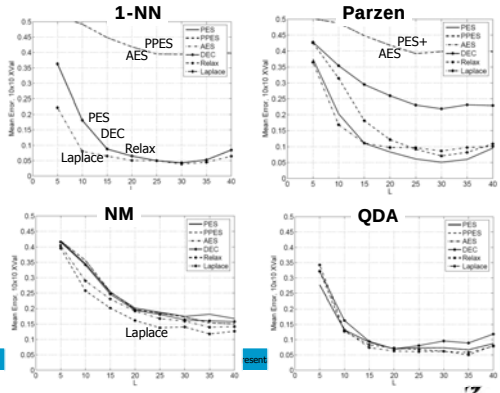
Classifiers:

- **1-NN**: 1-Nearest Neighbor (local distances)
- **Parzen**: non-parametric density estimation based on given dissimilarities (local densities)
- **NM**: Nearest Mean Classifier (global distances)
- **QDA**: Quadratic Discriminant Analysis (global densities)

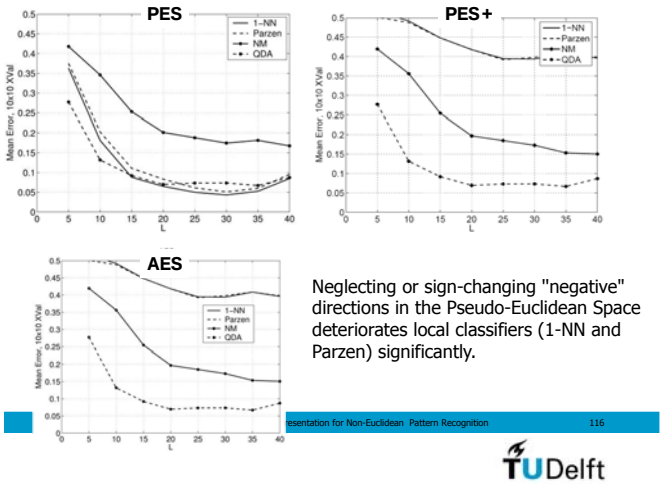
All four classifiers can be computed in ES as well as in PES

TUDelft

Results

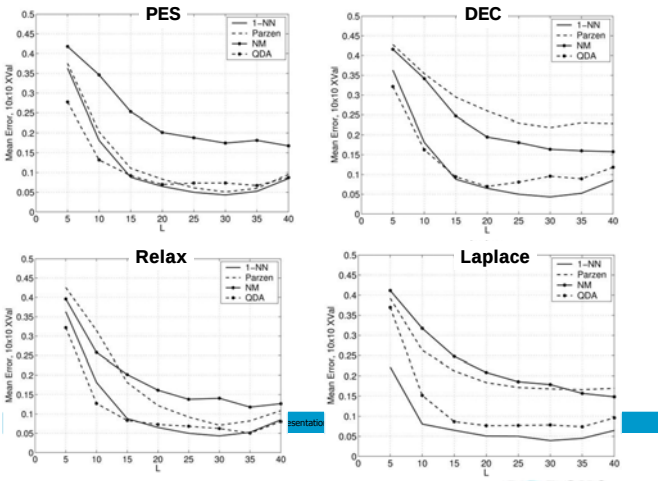


TUDelft



Neglecting or sign-changing "negative" directions in the Pseudo-Euclidean Space deteriorates local classifiers (1-NN and Parzen) significantly.

TUDelft



Conclusion w.r.t. Euclidean corrections

- Globally sensitive classifiers are hardly affected by corrections.
- 1-NN is insensitive to monotonic corrections, but really suffers from crisp corrections in the PES.
- Parzen is always disturbed by corrections, so they damage the local structure in the data.
- Corrections should be studied in relation with the classifier.

TUDelft

Non-Euclidean Representations: Informativeness

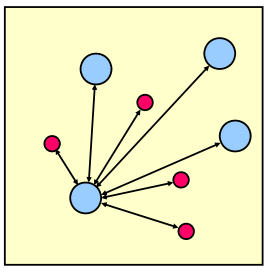
Negative Euclidean Fraction

$$\mathrm{NEF} = \frac{\sum_{\lambda_i<0} |\lambda_i|}{\sum_{\forall \lambda_i} |\lambda_i|}$$

$$0 \le \mathrm{NEF} \le 1$$

TUDelft

Artificial Example ‚Ball Distances



- Generate sets of balls (classes) uniformly, in a (hyper)cube; not intersecting.
- Balls of the same class have the same size.
- Compute all distances between the ball surfaces.

-> Dissimilarity matrix D

TUDelft

Balls3D

Classifier	PE Sp	Ass Sp	Pos Sp	Neg Sp	Cor Sp
1-NN	47.4 (2.0)	47.4 (2.0)	47.4 (2.0)	44.2 (1.5)	47.4 (2.0)
Parzen	45.7 (1.7)	45.5 (1.6)	45.6 (1.7)	35.5 (1.7)	45.7 (1.7)
NM	47.5 (2.0)	47.7 (2.0)	47.6 (1.9)	49.6 (0.2)	48.1 (1.8)
SVM-1	50.7 (2.2)	50.0 (2.7)	50.0 (2.5)	62.1 (1.7)	50.1 (2.0)

Classifier	PE Dis Sp	Ass Dis Sp	Pos Dis Sp	Neg Dis Sp	Cor Dis Sp
1-NN	49.8 (2.2)	49.8 (2.2)	49.8 (2.2)	5.1 (0.8)	49.7 (2.2)
Parzen	47.9 (2.2)	47.9 (2.2)	47.9 (2.2)	4.6 (0.5)	47.9 (2.2)
NM	49.8 (2.2)	49.8 (2.2)	49.8 (2.2)	5.0 (0.8)	49.9 (2.2)
SVM-1	50.2 (1.6)	50.8 (1.7)	50.7 (1.7)	1.9 (0.5)	49.8 (1.5)

10 x (2-fold crossvalidation of 50 objects per class)

TUDelft

Representation Strategies

Avoiding the PE space

Dissimilarity Space: $X = D$

Correcting

Associated space	$X = \{[Xp, Xq], \varnothing\}$	$\tilde{d}_{ij}^2 = d_p^2(x_i, x_j) + d_q^2(x_i, x_j)$
Positive space	$X = X_p$	$\tilde{d}_{ij}^2 = d_p^2(x_i, x_j)$
Negative space	$X = X_q$	$\tilde{d}_{ij}^2 = d_q^2(x_i, x_j)$
Additive Correction	$\tilde{d}_{ij}^2 = d_{ij}^2 + c, i \neq j$	$X = \text{Embedding}(\tilde{D})$

As it is

Pseudo Euclidean Space	$X = \{Xp, Xq\}$	$d_{ij}^2 = d_p^2(x_i, x_j) - d_q^2(x_i, x_j)$

Classifiers to be developed further

TUDelft

	size	classes	Non-Metric	NEF	Rand Err	Original, D	Positive, D_p	Negative, D_q
Chickenpieces45	446	5	0	0.156	0.791	**0.022**	0.132	0.175
Chickenpieces60	446	5	0	0.162	0.791	**0.020**	0.067	0.173
Chickenpieces90	446	5	0	0.152	0.791	**0.022**	0.052	0.148
Chickenpieces120	446	5	0	0.130	0.791	**0.034**	0.108	0.148
FlowCyto	612	3	1e-5	0.244	0.598	0.103	**0.100**	0.327
WoodyPlants50	791	14	5e-4	0.229	0.928	**0.075**	0.076	0.442
CatCortex	65	4	2e-3	0.208	0.738	**0.046**	0.077	0.662
Protein	213	4	0	0.001	0.718	0.0[illegible]	[illegible]	[illegible]
Balls3D	200	2	3e-4	0.001	0.500	**0.470**	0.495	0.000
GaussM1	500	2	0	0.262	0.500	**0.202**	**0.202**	0.228
GaussM02	500	2	5e-4	0.393	0.500	0.204	**0.174**	0.252
CoilYork	288	4	8e-8	0.258	0.750	**0.267**	0.313	0.618
CoilDelftSame	288	4	0	0.027	0.750	**0.413**	0.417	0.597
CoilDelftDiff	288	4	8e-8	0.128	0.750	**0.3[illegible]**	[illegible]	[illegible]
NewsGroups	600	4	4e-5	0.202	0.733	**0.198**	0.213	0.[illegible]
BrainMRI	124	2	5e-5	0.112	0.499	0.226	**0.218**	0.556
Pedestrians	689	3	4e-8	0.111	0.348	**0.010**	0.015	0.030

Informative

Extremely Informative

Not Informative

Conclusions

- Pseudo Euclidean Space (PES) is sometimes informative (corrections are not helpful).
- The corresponding problems may be intrinsic non-Euclidean
- Classifiers for non-Euclidean data have to be studied further

TUDelft