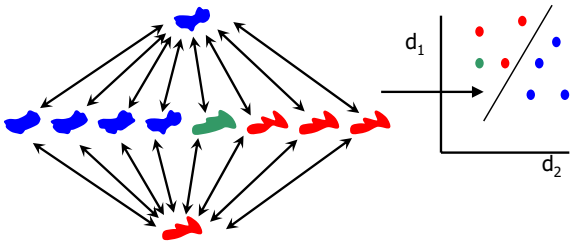
The dissimilarity representation for pattern recognition, a tutorial

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

[1]Electrical Engineering, Mathematics and Computer Sciences,
Delft University of Technology, The Netherlands

[2]School of Computer Science,
University of Manchester, United Kingdom

June 2009



This tutorial presents an introduction to the studies undertaken by the authors and their collaborators between 1997 and 2009 on the topic of dissimilarity representations for pattern recognition. Their research has been supported by the Dutch Organization for Scientific Research (NWO), by the EPSRC in the United Kingdom and by the FET programme within the EU FP7, under the SIMBAD project (contract 213250).

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 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.

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 [1] statistical pattern recognition [2, 3, 4, 5], artificial neural networks [6] and machine learning [7, 8]. 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. [9]. 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 [10].

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 [11]. This is identical to the nearest neighbor rule used in vector spaces [3]. Also many procedures for cluster analysis make use of dissimilarities instead of feature spaces [12]. To some extent, the concept of dissimilarities is analogous to the use of kernels (and the potential functions as studied in the sixties [13]). The main difference is that kernels were originally defined in vector spaces to preferably fulfill Mercer's conditions [14, 15]. 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) [7].

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 [16, 17, 18, 19, 20], which they also discussed as generalized kernel approaches [21, 22]. 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) [16, 23, 21]. 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 [24]. 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 [20] 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. 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.

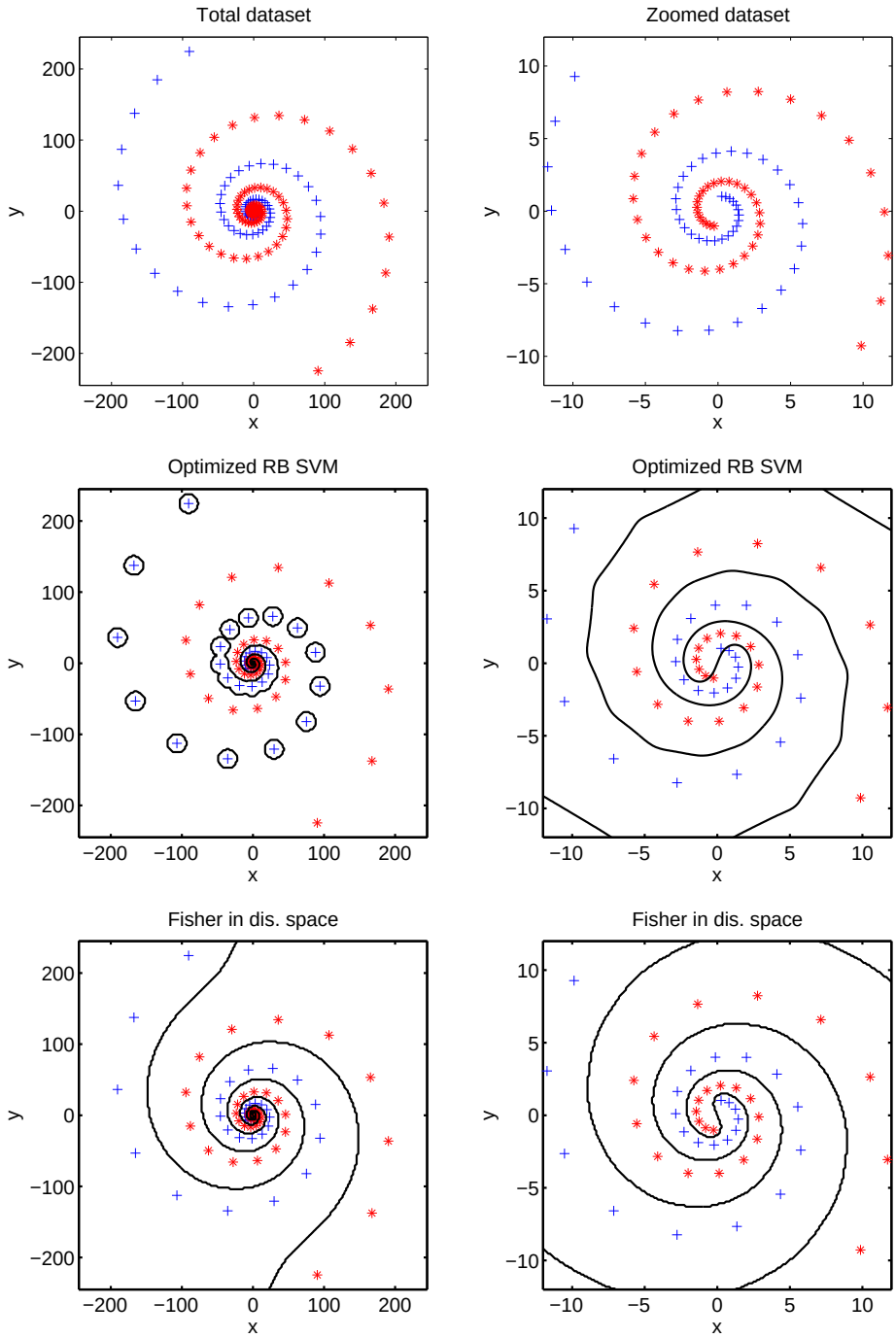


Figure 1: A spiral example with 100 objects per class. Left column shows the complete data sets, while the right column presents the zoom of the spiral center. 50 objects per class are used for training, systematically sampled. The middle row shows the training set and SVM with an optimized radial basis 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 $\mathbb{R}^q$ [25, 20]. 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 [25, 20].

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 a Non-Euclidean Coefficient (NEC) as:

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

Fig. 2 shows how NEC 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 NEC 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.

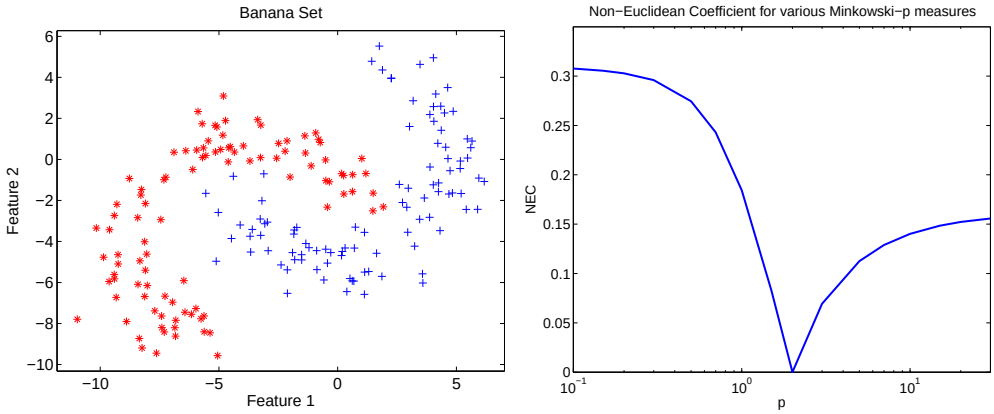


Figure 2: A two-dimensional data set (left) and the NEC as a function of p for various Minkowski-p dissimilarity measures.

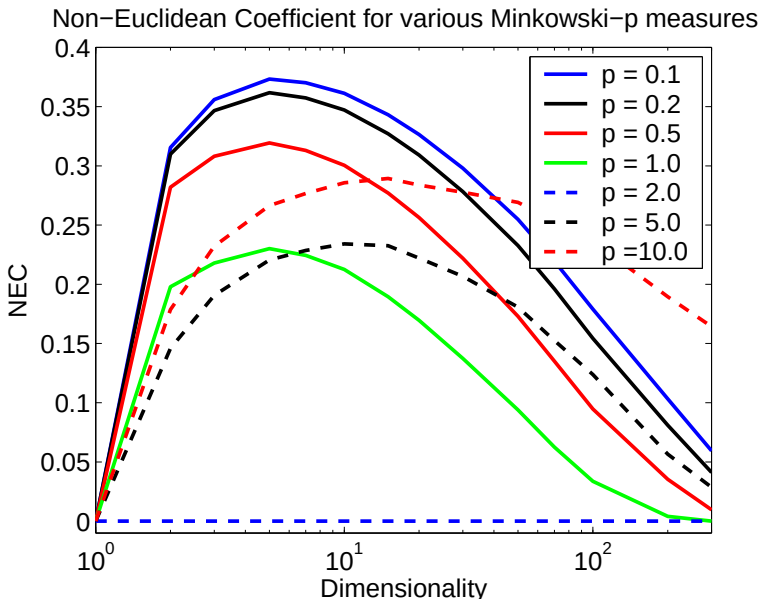


Figure 3: The Non-Euclidean Coefficient for various Minkowski-p dissimilarity measures as a function of the dimensionality of a set of 100 points generated by a standard Gaussian distribution.

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 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 [26] 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: [27, 28, 29, 30, 31]. There is however some doubt whether the non-Euclidean aspects of the relations between pairwise comparison of objects are informative [32, 33, 34].

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. We will therefore discuss the various circumstances under which such dissimilarity matrices arise and will try to characterize them. 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 [20] 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 hyper-spectral 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 [35]. Most of them are still metric, some of them however are non-metric [36].

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 [37] and deformable templates [38]. 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 [39] 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.

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

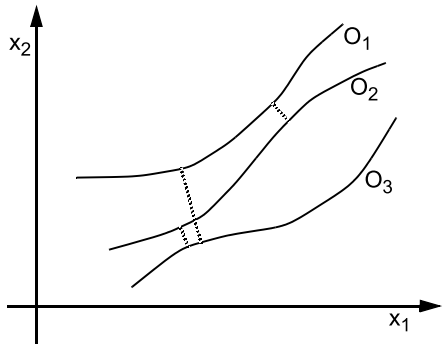


Figure 4: Vector space with the invariant trajectories for three objects O_1, O_2 and O_3. If the chosen dissimilarity measure is the minimal distance between these trajectories, triangle inequality can easily be violated, i.e. $d(O_1, O_2) + d(O_1, O_3) < d(O_1, O_3)$.

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 [22].

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.

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 [9] 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:

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

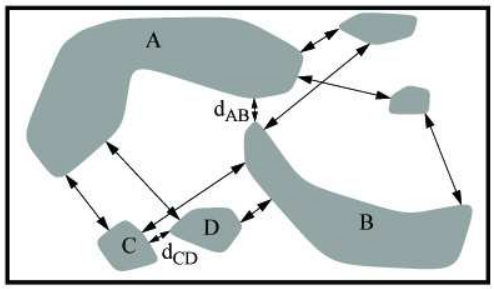


Figure 5: Single-linkage distance may be small for clusters which differ in position and shape.

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 [20, 22, 40, 29, 33, 41], 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 [27].

In [29] 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 [33] 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 [33] and in [41] 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] Vapnik, V.: Statistical Learning Theory. John Wiley & Sons, Inc. (1998)

[2] Fukunaga, K.: Introduction to Statistical Pattern Recognition. 2^{nd} edn. Academic Press, London (1990)

[3] Duda, R., Hart, P., Stork, D.: Pattern Classification. John Wiley and Sons (2001) 0-471-05669-3.

[4] Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: A review. IEEE Trans. Pattern Anal. Mach. Intell **22** (2000) 4–37

[5] Duin, R., Tax, D.: Statistical pattern recognition. In Chen, C.H., Wang, P.S.P., eds.: Handbook of Pattern Recognition and Computer Vision, Third Edition. World Scientific (2005) 3–24

[6] Schalkoff, R.J.: Artificial Neural Networks. McGraw-Hill Higher Education (1997)

[7] Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines. Cambridge University Press, UK (2000)

[8] Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)

[9] Kondor, R.I., Jebara, T.: A kernel between sets of vectors. In: ICML. (2003) 361–368

[10] Bunke, H., Sanfeliu, A., eds.: Syntactic and Structural Pattern Recognition Theory and Applications. World Scientific (1990)

[11] Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley, New York (1972)

[12] Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 4th Edition. Academic Press (2008)

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

[14] Schölkopf, B., Smola, A.: Learning with Kernels. MIT Press, Cambridge (2002)

[15] Shawe-Taylor, J., Cristianini, N.: Kernel methods for pattern analysis. Cambridge University Press, UK (2004)

[16] Duin, R., de Ridder, D., Tax, D.: Experiments with object based discriminant functions; a featureless approach to pattern recognition. Pattern Recognition Letters **18** (1997) 1159–1166

[17] Duin, R., Pękalska, E., de Ridder, D.: Relational discriminant analysis. Pattern Recognition Letters **20** (1999) 1175–1181

[18] Duin, R.: Relational discriminant analysis and its large sample size problem. In: ICPR. (1998) Vol I: 445–449

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

[20] Pękalska, E., Duin, R.: The Dissimilarity Representation for Pattern Recognition. Foundations and Applications. World Scientific, Singapore (2005)

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

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

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

[24] Pękalska, E., Duin, R.: Dissimilarity-based classification for vectorial representations. In: ICPR (3). (2006) 137–140

[25] Goldfarb, L.: A new approach to pattern recognition. In Kanal, L., Rosenfeld, A., eds.: Progress in Pattern Recognition. Volume 2. Elsevier (1985) 241–402

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

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

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

[29] Haasdonk, B., Pękalska, E.: Indefinite kernel fisher discriminant. In: ICPR. (2008) 1–4

[30] Ong, C., Mary, X.and Canu, S., A.J., S.: Learning with non-positive kernels. In: Int. Conference on Machine Learning, Brisbane, Australia (2004) 639–646

[31] Ong, C.S.: Kernels: Regularization and optimization (2005)

[32] Pękalska, E., Harol, A., Duin, R., Spillmann, B., Bunke, H.: Non-euclidean or non-metric measures can be informative. In: SSPR/SPR. (2006) 871–880

[33] Duin, R., Pękalska, E., Harol, A., Lee, W.J., Bunke, H.: On euclidean corrections for non-euclidean dissimilarities. In: SSPR/SPR. (2008) 551–561

[34] Laub, J., Roth, V., Buhmann, J.M., Müller, K.R.: On the information and representation of non-euclidean pairwise data. Pattern Recognition **39** (2006) 1815–1826

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

[36] Dubuisson, M., Jain, A.: Modified Hausdorff distance for object matching. In: Int. Conference on Pattern Recognition. Volume 1. (1994) 566–568

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

[38] Jain, A.K., Zongker, D.E.: Representation and recognition of handwritten digits using deformable templates. IEEE Trans. Pattern Anal. Mach. Intell **19** (1997)

[39] Duin, R., Pękalska, E.: Structural inference of sensor-based measurements. In: Structural, Syntactic, and Statistical Pattern Recognition. (2006) 41–55

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

[41] Duin, R., Pękalska, E.: On refining dissimilarity matrices for an improved nn learning. In: ICPR. (2008) 1–4