Training classifiers for non Euclidean data

Robert P.W. Duin, Delft University of Technology
(In cooperation with Elżbieta Pękalska, Univ. of Manchester)

Johns Hopkins University, 6 March 2008

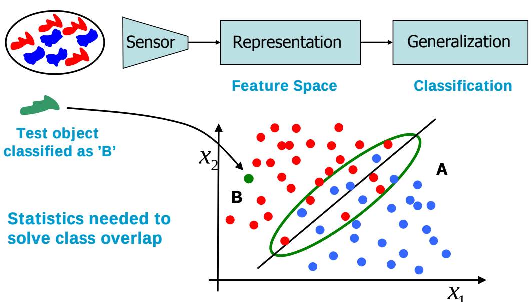
TUDelft

Topics

- Pattern recognition system
- Non-Euclidean data
- Indefinite representations
- Classifiers for indefinite representations
- Problems
- Conclusions

TUDelft

Pattern Recognition System

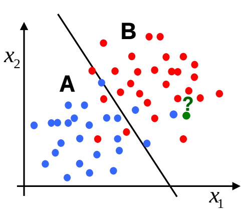


TUDelft

Recap of classification

TUDelft

Classification principles



How should this object be classified?

Class B, as it has the highest density for B.

Class A, as it is most close to an object of class A.

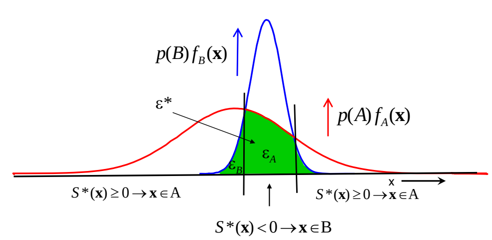
Class B, as it is on the B-side of the linear minimum error classifier.

Principles

- Probabilities (densities)
- Distances, dissimilarities
- Error minimization

TUDelft

Density based classifiers



Multi-dimensional may might be difficult to estimate

Metric independent, estimators may be metric dependent

TUDelft

Bayes decision rule, formal

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

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

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

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

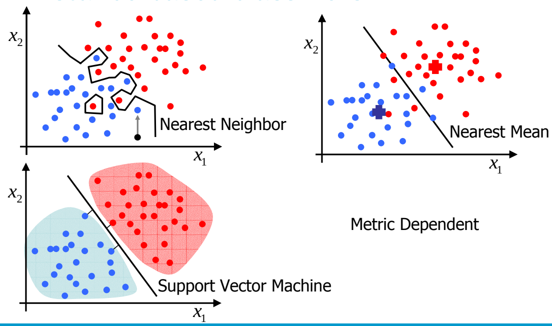
n-class problems: $\text{Class}(x) = \text{argmax}_{\omega}(p(x|\omega)\,p(\omega))$

Examples Density Based Classifiers

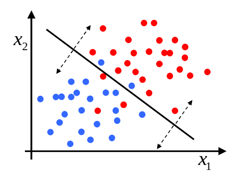
- QDA, Quadratic classifier based on normal distributions
- LDA, Linear classifier based on normal distributions
- MoG, Mixture of Gaussian classifier
- Parzen, Non-parametric density estimation
- k-NN, k-Nearest Neighbor rule
- Naïve Bayes classifier

Various regularizations are used
in case of degenerate vector spaces.

Distance based classifiers



Performance optimization



Examples
PERC Linear Perceptron
ANN Artificial Neural Network
Fisher's Linear Discriminant
Logistic Classifier
Support Vector Machine

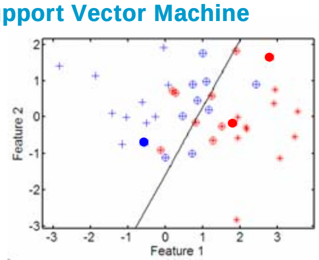
Optimize parameters of classifier
for some separability criterion

A Note on the Support Vector Machine

$$f(x) = w^T x + w_0$$

$$\min_w \|w\| + c \sum_{x_i \in \text{slacks}} \xi(x_i)$$

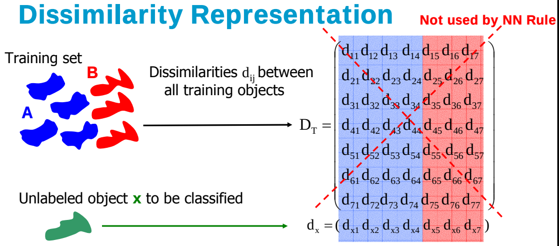
$$\text{s.t.} \quad y_i f(x_i) \geq 1 - \xi(x_i),$$
$$\xi(x_i) \geq 0$$



The SVM depends on **distances** of support vectors to the classifier (resulting in $\|w\|$) **as well as** on the **density** of the class overlap (resulting in the sum of slacks).

Dissimilarity Representation

Dissimilarity Representation



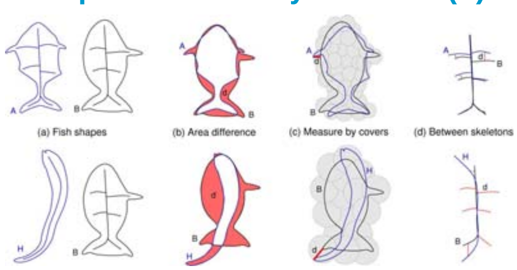
The traditional Nearest Neighbor rule (template matching) finds:
$\text{label}(\text{argmin}_{\text{trainset}}\{d_{xi}\})$,
without using D_T. Can we do any better?

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.

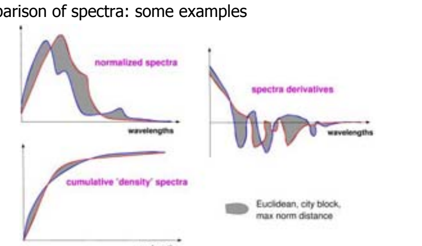
Examples Dissimilarity Measures (1)



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

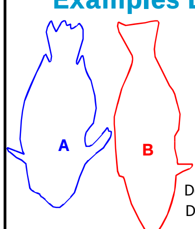
Examples Dissimilarity Measures (2)

Comparison of spectra: some examples

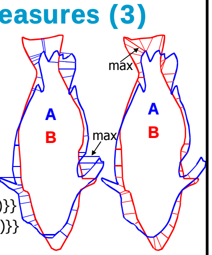


In real applications, the dissimilarity measure should be robust to noise and small aberrations in the (raw) measurements.

Examples Dissimilarity Measures (3)



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



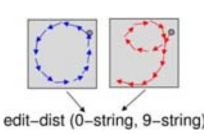
D(**A**,**B**) = max_a{min_b{d(a,b)}}
D(**B**,**A**) = max_b{min_a{d(b,a)}}

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)}}}

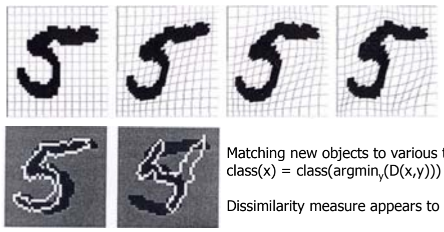
Examples Dissimilarity Measures (4)



weighted edit distance: non-Euclidean!

edit–dist (0–string, 9–string)

Examples Dissimilarity Measures (5)



Matching new objects to various templates:
class(x) = class(argmin$_y$(D(x,y)))

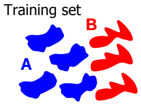
Dissimilarity measure appears to be non-metric.

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

Classification of Dissimilarity Data

Alternatives for the Nearest Neighbor Rule

Training set



Dissimilarities d_{ij} between all training objects

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

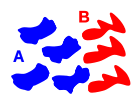
Unlabeled object **x** to be classified

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

1. Dissimilarity Space
2. Embedding

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

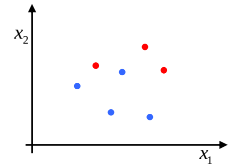
Embedding



Training set

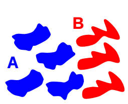
→ Dissimilarity matrix D → X

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

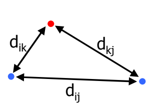


Position points in a vector space such that their Euclidean distances → D

Embedding of non-metric measurements

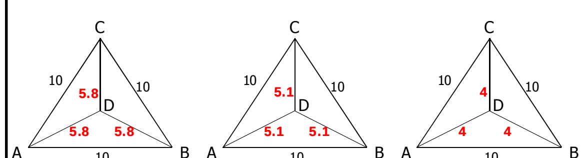


→ Dissimilarity matrix D → X



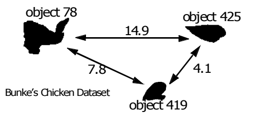
If the dissimilarity matrix cannot be explained from a vector space,
(e.g. for Hausdorff and Hamming distance of images)
or if $d_{ij} > d_{ik} + d_{kj}$ (triangle inequality not satisfied)
embedding in Euclidean space not possible
→ Pseudo-Euclidean embedding

Euclidean - Non Euclidean - Non Metric

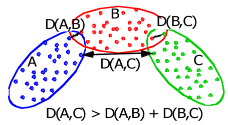


Non-metric distances

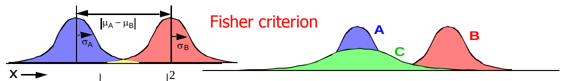
Weighted-edit distance for strings



Single-linkage clustering



Fisher criterion



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

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

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

Embedding of non-Euclidean Dissimilarities

(Pseudo-)Euclidean Embedding

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

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

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 $k \times k$ 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.

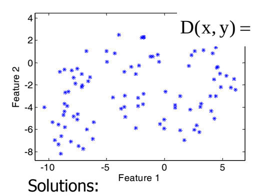
The inner-product matrix: $B_z = -\frac{1}{2}(D_z^{(2)} J - \frac{1}{n}\mathbf{11}^T D^{(2)} J)$

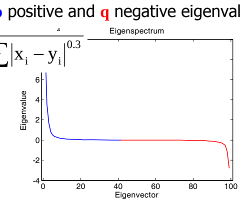
The embedded objects: $Z = B_z Q_k |\Lambda_k|^{-\frac{1}{2}} \Im_k$

Pseudo-Euclidean Embedding

If D is non-Euclidean then B has $\mathbf{p}$ positive and $\mathbf{q}$ 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

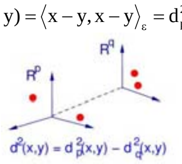
PES: Pseudo-Euclidean Space (Krein Space)

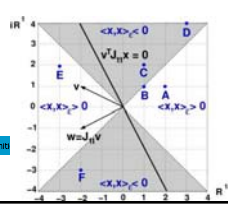
If D is non-Euclidean, B has p positive and q negative eigenvalues.

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:

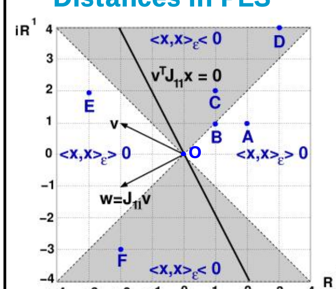
$$\langle x, y \rangle_{\varepsilon} = x^T \Im_{pq} y = \sum_{i=1}^{p} x_i y_i - \sum_{j=p+1}^{q} x_j y_j \qquad \Im_{pq} = \begin{bmatrix} I_{p \times p} & 0 \\ 0 & -I_{q \times q} \end{bmatrix}$$

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





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 $v^T J x = 0$.

Can it be used as a classifier?
Can we define a margin as in the SVM?

PE Space ←→ Kernels

$$K(x,y) = -\tfrac{1}{2} J D(x,y)^{(2)} J \qquad J = I - \tfrac{1}{m}\mathbf{11}$$

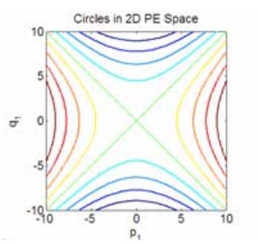
may by considered as a kernel. If

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

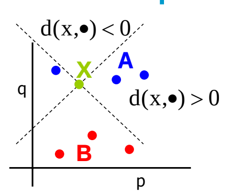
- The **kernel trick** may be used: operations defined on inner products in kernel space can be operated directly on K(x,y) **without embedding**!
- True for **Mercer kernels** (all eigenvalues ≥ 0).
- Difficult for **indefinite kernels**.
- Studying classifiers in **PE space** is studying the indefinite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

Classifiers in Pseudo Euclidean Space

Distance based classifiers in PE Space



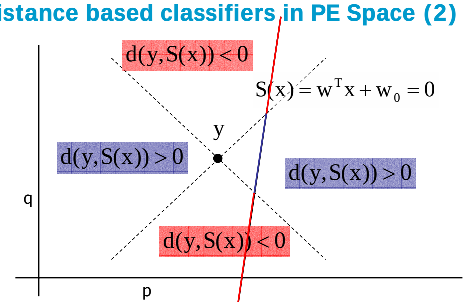
Metric in PE Space.
Equidistant points to the origin.



X assigned to B

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

Distance based classifiers in PE Space (2)



Some points on $S(x) = 0$ have very negative distances to y.
What is the distance between $S(x)$ and y?

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.*

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:

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

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.

Performance optimization in PES

Needs specific definitions

Euclidean corrections for non-Euclidean dissimilarities

Correction Procedures PES ←→ ES

Make the PE space (more) Euclidean:

- Reduction of the q-space contribution

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

  $d_\varepsilon^2(x,y) = d_p^2(x,y) - \alpha d_q^2(x,y) \qquad \alpha \in [-1,1]$
- Enlarging dissimilarities

  $d_\varepsilon^2(x,y) \Leftarrow d_\varepsilon^2(x,y) + c, x \neq y$
- Relaxing dissimilarity measure

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

Example: Chickenpieces (H. Bunke, Bern)

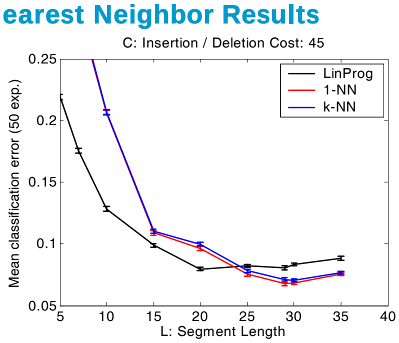
446 binary images, varying size, e.g.: 100 x 130

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.

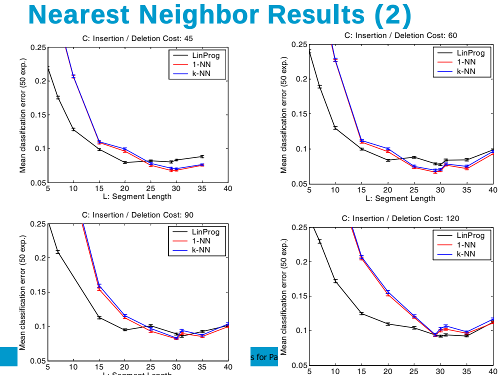
Shape classification by weighted-edit distances (Bunke)

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

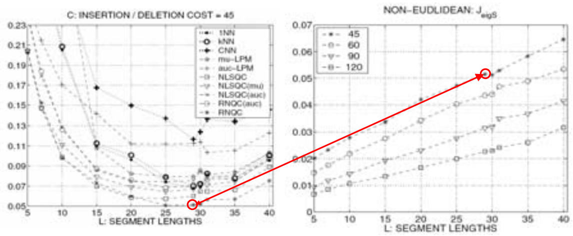
Nearest Neighbor Results



Nearest Neighbor Results (2)

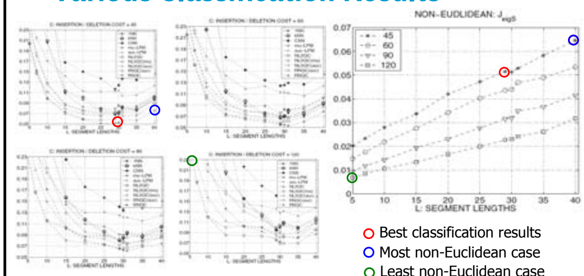


Classification Results for Various Dissimilarity Measures



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

Various Classification Results



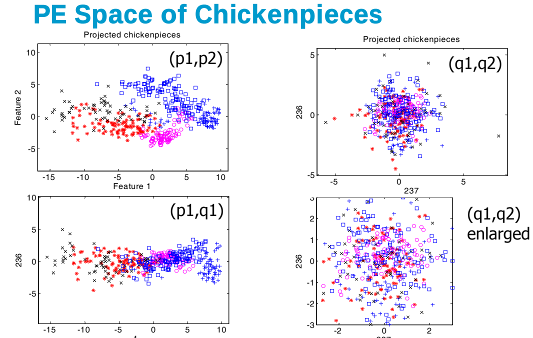
- Best classification results
- Most non-Euclidean case
- Least non-Euclidean case

Pekalska, Harol, Duin, Bunke, Spillman, SSSPR 2006

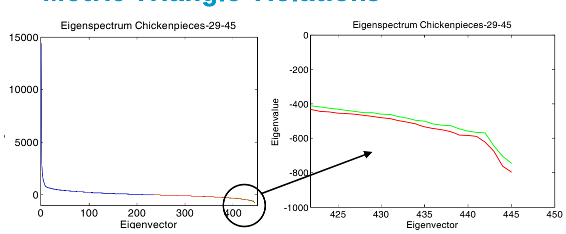
Is non-Euclidean behavior here essential for good classification?

- Non-Euclidean or non-metric measures can be informative (*Pekalska, Harol, Duin, Bunke, Spillman, SSSPR 2006).*
- More Euclidean measures behave worse.
- However: Is it possible to transform the non-Euclidean data such that a (more) Euclidean embedding can be found that obtains similar or better classification results?

PE Space of Chickenpieces



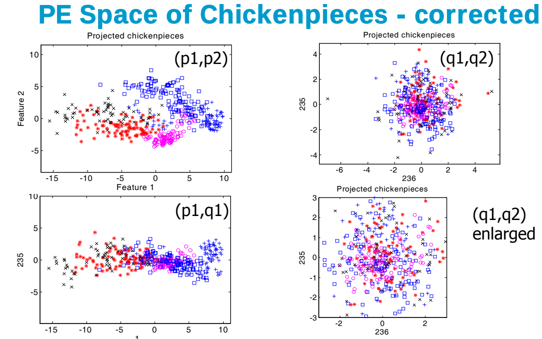
Metric Triangle Violations



8772 triangle violations
5458 corrections to make D metric

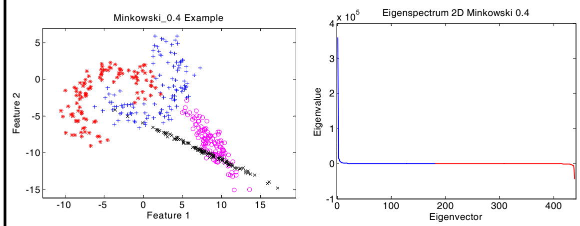
Corrections have a insignificant influence

PE Space of Chickenpieces - corrected



Metric corrections are minor and do not improve classification results

Minkowski 0.4 Example

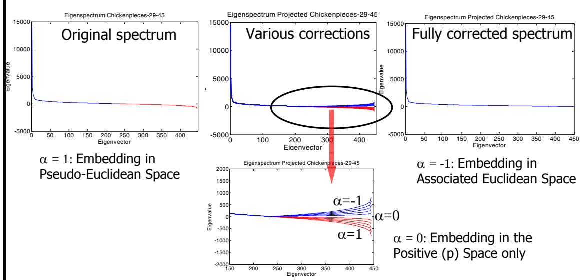


440 points in 2D, 4 classes

$$d(x,y) = (\sum_i |x_i - y_i|^{0.4})^{1/0.4}$$

Signature: (p,q): (181, 258)

q-Space Reduction, PES → AES



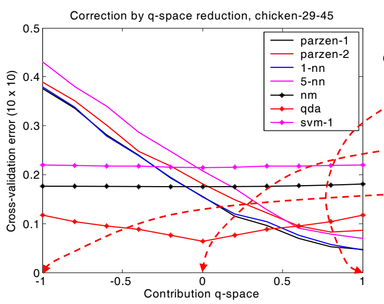
α = 1: Embedding in Pseudo-Euclidean Space

α = -1: Embedding in Associated Euclidean Space

α = 0: Embedding in the Positive (p) Space only

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

q-Space Reduction, PES → AES



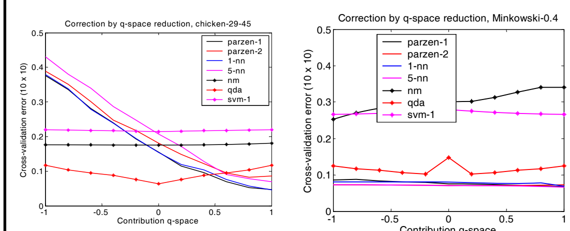
$$d_\varepsilon^2(x,y) = d_p^2(x,y) - \alpha d_q^2(x,y)$$

α = 1: Embedding in Pseudo-Euclidean Space

α = 0: Embedding in the Positive (p) Space only

α = -1: Embedding in Associated Euclidean Space

q-Space Reduction: Examples Compared

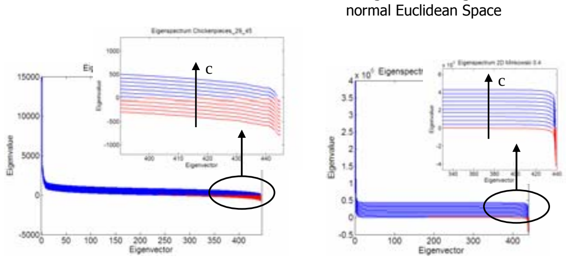


Growing Distances

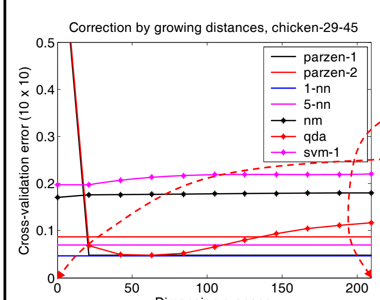
$$d_\varepsilon^2(x,y) \Leftarrow d_\varepsilon^2(x,y) + c, x \neq y$$

c = 0: Original embedding in Pseudo-Euclidean Space

c large: Embedding in normal Euclidean Space



Growing Distances

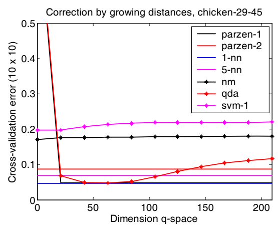


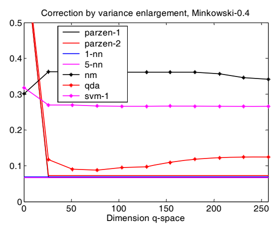
$$d_\varepsilon^2(x,y) \Leftarrow d_\varepsilon^2(x,y) + c, x \neq y$$

c = 0: Original embedding in Pseudo-Euclidean Space

c large: Embedding in normal Euclidean Space

Growing Distances



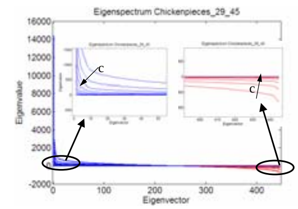


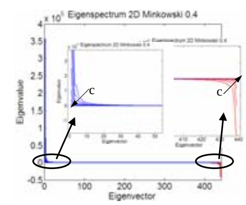
Dissimilarity Measure Relaxation

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

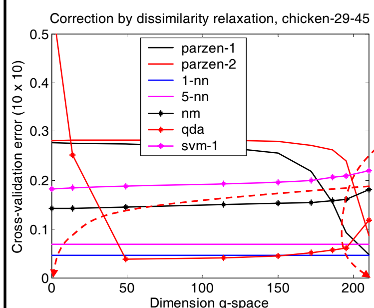
$c = 1$: Original embedding in Pseudo-Euclidean Space

c large (e.g. 10): Embedding in normal Euclidean Space





Dissimilarity Measure Relaxation

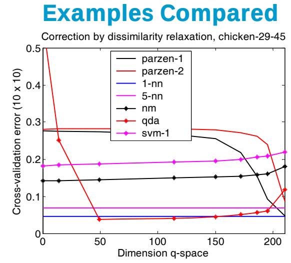


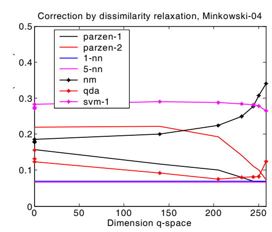
$$d_\varepsilon(x,y) \Leftarrow d_\varepsilon(x,y)^{1/c}, c \geq 1$$

$c = 1$: Original embedding in Pseudo-Euclidean Space

c large (e.g. 10): Embedding in normal Euclidean Space

Dissimilarity Measure Relaxation: Examples Compared



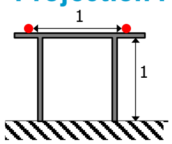


Conclusions on classifiers in PE Space

- Locally sensitive rules like 1-NN and Parzen_small do very well.
- Euclidean corrections do not improve computable classifications (may be other classifiers in Euclidean spaces perform better).
- QDA seems to be a possible globally sensitive rule, but is usually outperformed by 1-NN and Parzen_small.

Test Data projected in PE Space

Projection Problems of PE Space



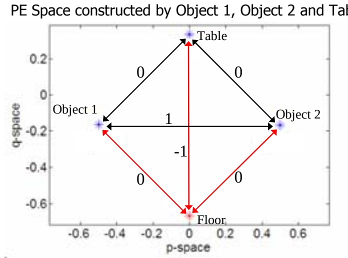
True distances:

$$\begin{array}{r} \text{Object1} \\ \text{Object2} \\ \text{Table} \\ \text{Floor} \end{array} : \left[\begin{array}{ccc|c} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ \hline 1 & 1 & 0 & 0 \end{array}\right]$$

After projection:

$\text{Floor}: [0 \quad 0 \quad -1]$

PE Space constructed by Object 1, Object 2 and Table.



Projections of new objects in PE space may be entirely wrong

February 2008 Indefinite Representations for Pattern Recognition - Robert P.W. Duin 61

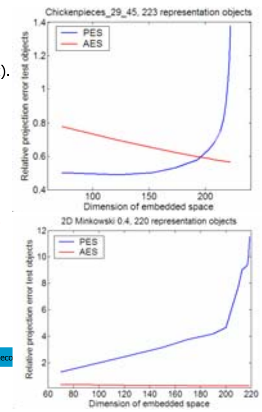
TU Delft

Experiment

- Take a non-Euclidean dissimilarity matrix.
- Use 50% for embedding (representation set).
- Project remaining 50% (test set).
- Compute all distances between these sets in the projection space
- Compare them with given distances
- Repeat for Euclidean distances derived from the Associated Euclidean Space (AES)

Linear projection of new objects in a PES may be entirely wrong.

Due to the occurrence of negative square distances linear procedures are not sufficient.



February 2008 Indefinite Representations for Pattern Reco

Avoiding Large Projection Errors

Do not use a full space for embedding.

Or, project training set and test set simultaneously. (Transductive Learning).

Consequences for indefinite kernel approaches ???

February 2008 Indefinite Representations for Pattern Recognition - Robert P.W. Duin 63

TU Delft

Conclusions on informativeness of non-metric dissimilarity measures

- Non-metric measures can be informative.
- After Euclidean correction performances do not significant improve.
- Traditional classification approaches based on distances or densities should be redesigned to construct global generalising classifiers in PE space.
- Projection of new data in PE space is problematic.

February 2008 Indefinite Representations for Pattern Recognition - Robert P.W. Duin 64

TU Delft