Workshop on Representations for PR

The dissimilarity representation, a basis for domain based pattern recognition?

Robert P.W.Duin, Elżbieta Pękalska, Pavel Paclik, David Tax

Faculty of Electrical Engineering, Mathematics and Computer Science

Delft University of Technology, The Netherlands

Cambridge, August 2004

P.O. Box 5031, 2600GA Delft, The Netherlands.
Phone: +(31) 15 2786143, FAX: +(31) 15 2781843,
E-mail: r.p.w.duin@ewi.tudelft.nl

A Problem

Bring me an apple!!

??

Training

These are apples ...

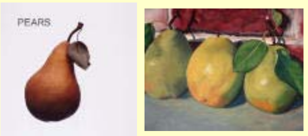
... and these are pears

!!

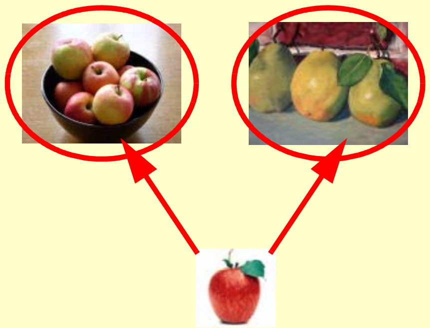
How to learn? Model based?

- Store all relevant properties of an apple

- Generalise over apple examples to obtain an ’apple class’ <u>model</u>



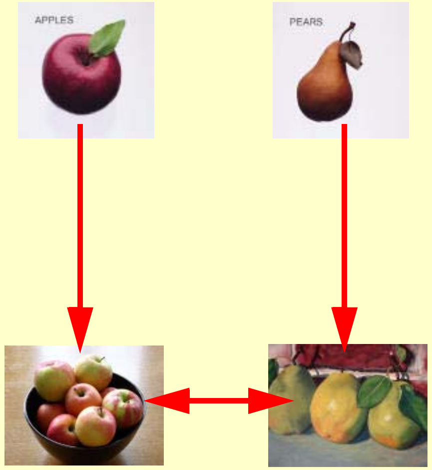
- Repeat for pears



- Find a way to compare ’apple-ness’ with ’pear-ness’

We only look at class differences at classification time!!

How to learn? Feature based?



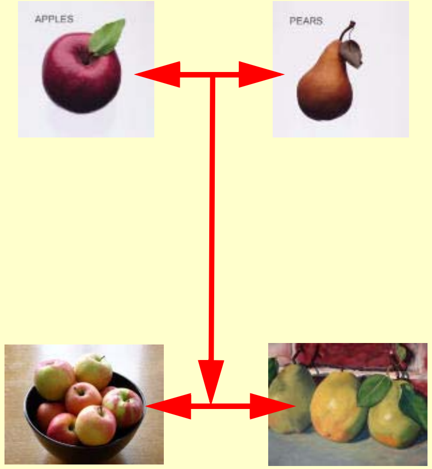
- Choose possible features to represent individual apples and pears

- Select relevant features for the difference of apples and pears

- Generalise over the examples

We only look at class differences at classification time!!

How to learn? Dissimilarity based?



- Use differences between apples and pears for representation

- Generalise over the examples

We only look at class differences during representation!!

The Pattern Recognition System

Model based Objects Measurements Models Class Differences

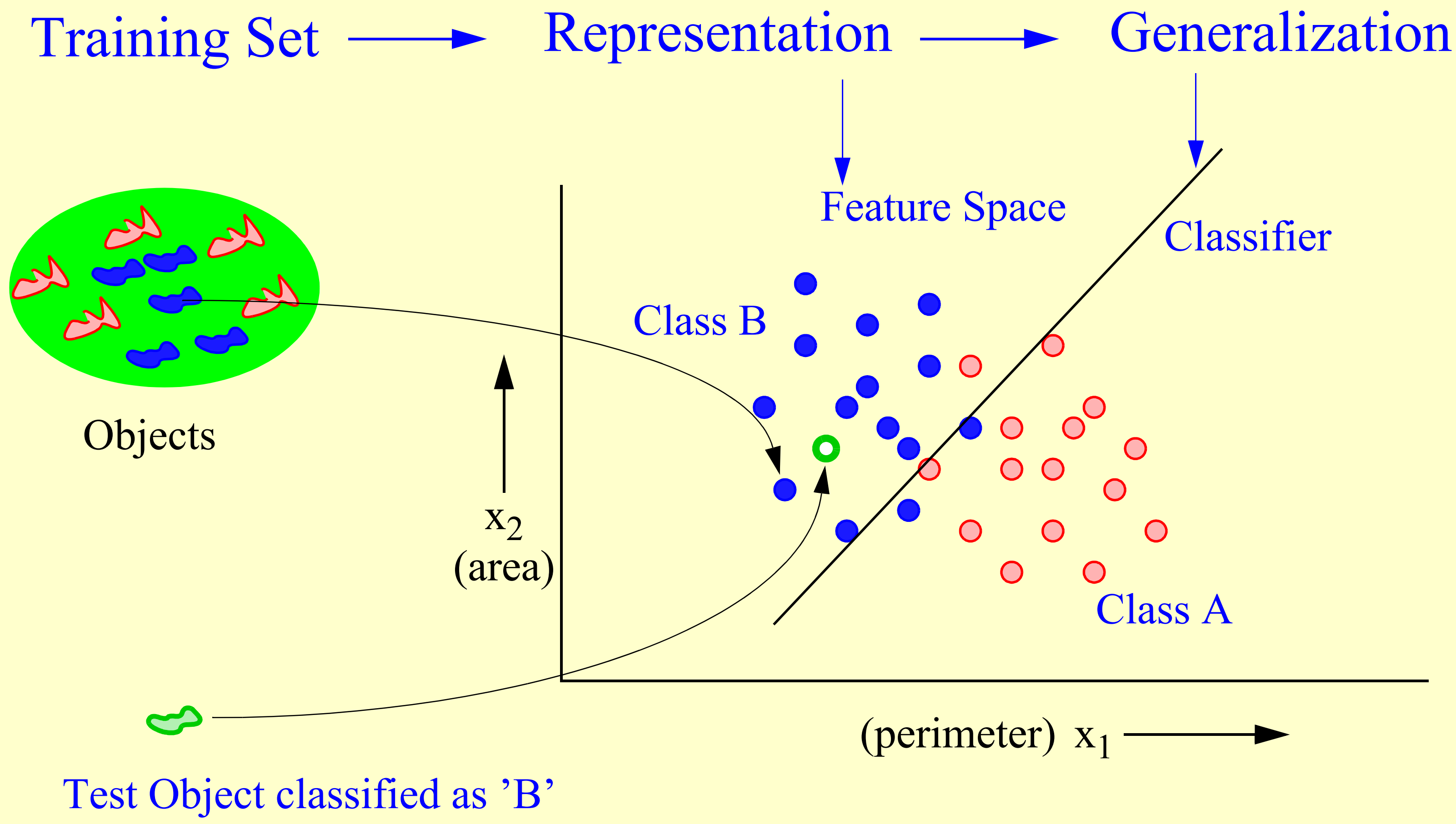
The Pattern Recognition System

Model based	Objects	Measurements	Models	Class Differences
Feature based	Objects	Features	Class Differences	Classifier

The Pattern Recognition System

Model based	Objects	Measurements	Models	Class Differences
Feature based	Objects	Features	Class Differences	Classifier
Dissimilarity based	Objects	Class Differences	Classifier	

Generalisation from Features



Feature representation → Object reduction → Class overlap → Probabilities

Classifiers for Feature Representations

Criterion : Min: $\varepsilon = \mathrm{Prob}(S(x) \rightarrow \omega \mid x \notin \omega)$

Neural Net Min: $\sum_i (S(x_i, w) - \lambda(x_i))^2$

Fisher: $S(x) = y = w \bullet x + w_0$, such that $L = \dfrac{(\bar{y}_1 - \bar{y}_2)^2}{\sigma_1^2 + \sigma_2^2}$ is minimum

Parzen: $S(x) = \dfrac{1}{n_1} \sum_{i \in \omega_1} \varphi(|x - x_i|, h_1) - \dfrac{1}{n_2} \sum_{i \in \omega_2} \varphi(|x - x_i|, h_2)$

SVC: Max: $\min_i \{ S(x_i)\lambda(x_i) \} + \sum_j \xi_j S(x_j)\lambda(x_j)$

Classifiers for Feature Representations

Probability Arguments

Criterion : Min: $\varepsilon = \mathrm{Prob}(S(x) \rightarrow \omega | x \notin \omega)$

Neural Net Min: $\sum_i (S(x_i, w) - \lambda(x_i))^2$

Fisher: $S(x) = y = w \bullet x + w_0$, such that $L = \frac{(\bar{y}_1 - \bar{y}_2)^2}{\sigma_1^2 + \sigma_2^2}$ is minimum

Parzen: $S(x) = \frac{1}{n_1} \sum_{i \in \omega_1} \varphi(|x - x_i|, h_1) - \frac{1}{n_2} \sum_{i \in \omega_2} \varphi(|x - x_i|, h_2)$

SVC: Max: $\min_i \{S(x_i)\lambda(x_i)\} + \sum_j \xi_j S(x_j)\lambda(x_j)$

Classifiers for Feature Representations

Probability Arguments

Distance Arguments

Criterion : Min: $\varepsilon = \mathrm{Prob}(S(x) \rightarrow \omega | x \notin \omega)$

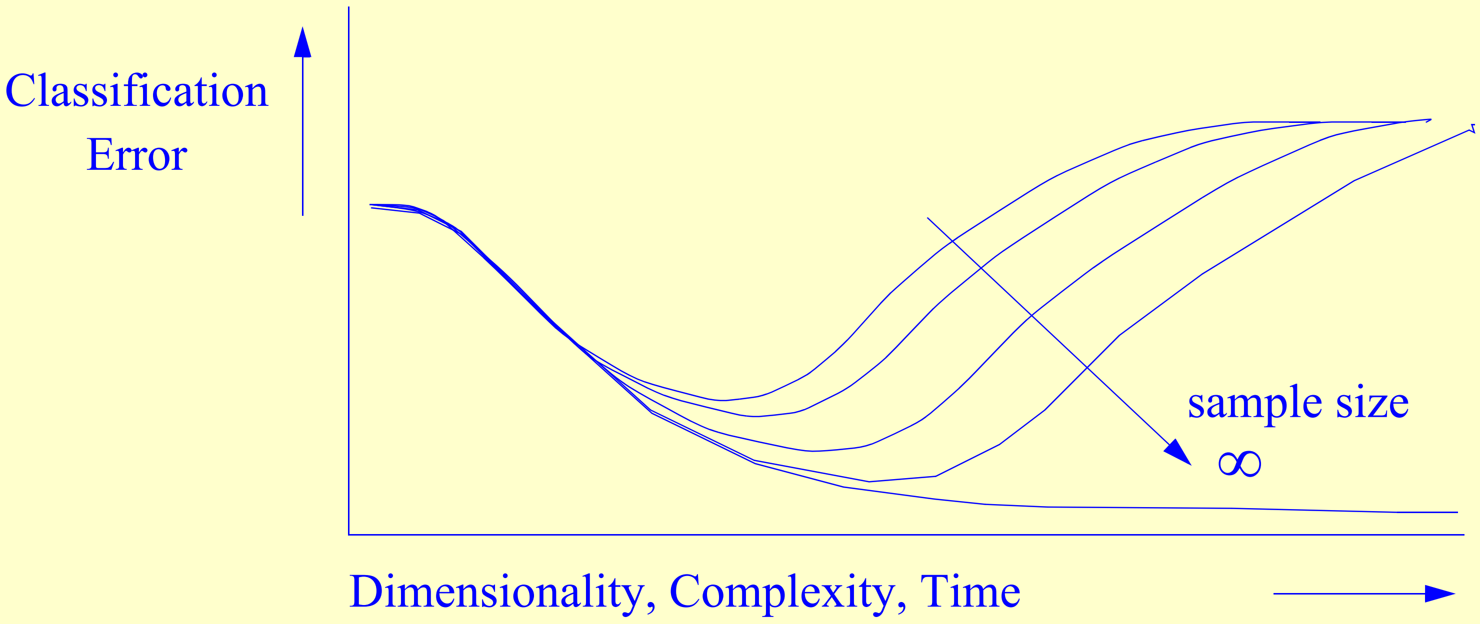
Neural Net Min: $\sum_i (S(x_i, w) - \lambda(x_i))^2$

Fisher: $S(x) = y = w \bullet x + w_0$, such that $L = \dfrac{(\bar{y}_1 - \bar{y}_2)^2}{\sigma_1^2 + \sigma_2^2}$ is minimum

Parzen: $S(x) = \dfrac{1}{n_1} \sum_{i \in \omega_1} \varphi(|x - x_i|, h_1) - \dfrac{1}{n_2} \sum_{i \in \omega_2} \varphi(|x - x_i|, h_2)$

SVC: Max: $\min_i \{S(x_i)\lambda(x_i)\} + \sum_j \xi_j S(x_j)\lambda(x_i)$

Peaking, Curse of Dimensionality, Overtraining



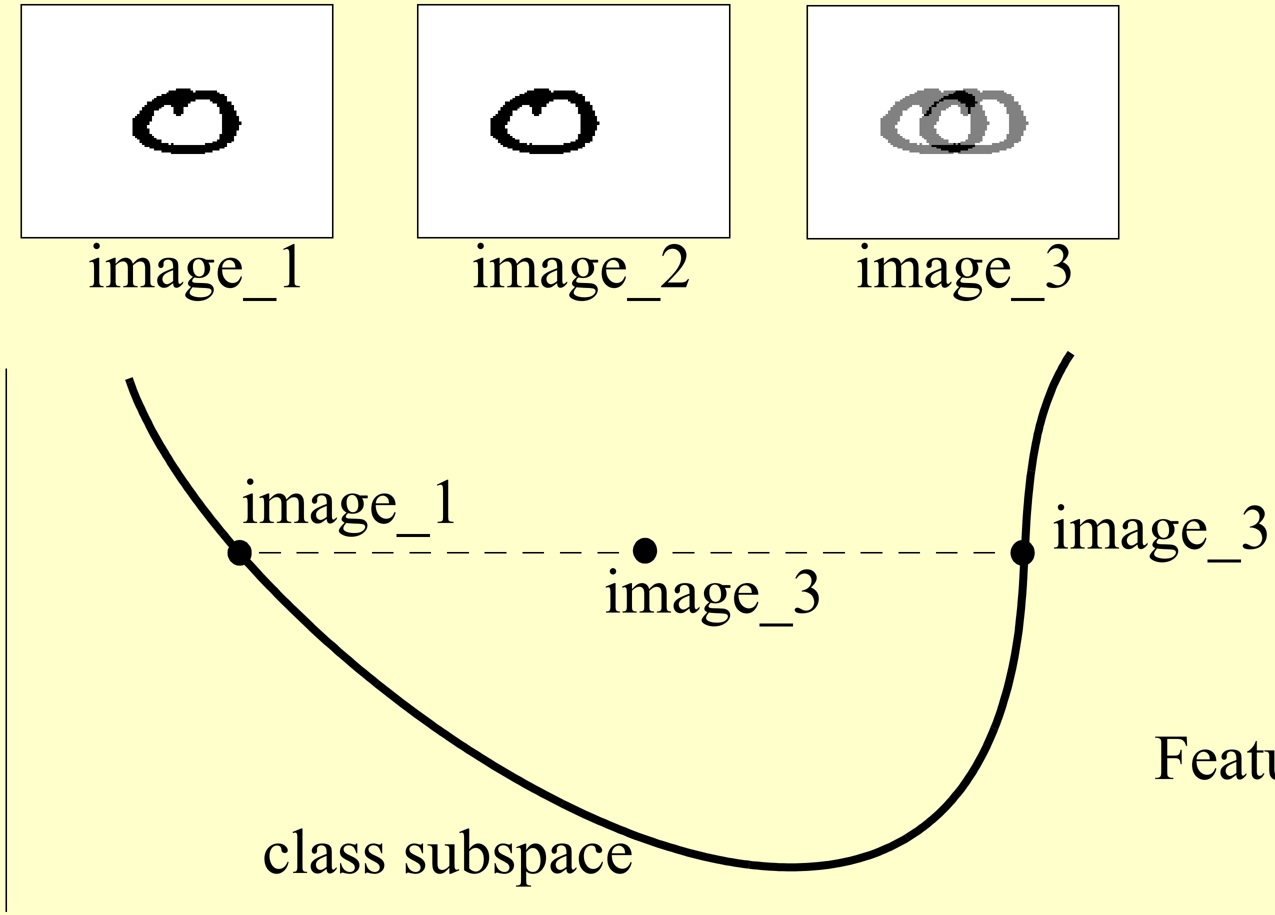
Asymptotically increasing classification error due to:

- Increasing Dimensionality — *Curse of Dimensionality*
- Increasing Complexity — *Peaking Phenomenon*
- Decreasing Regularization
- Increasing Computational Effort

} *Overtraining*

Problems with the Pixel_Feature Representation

Interpolation does not yield valid objects



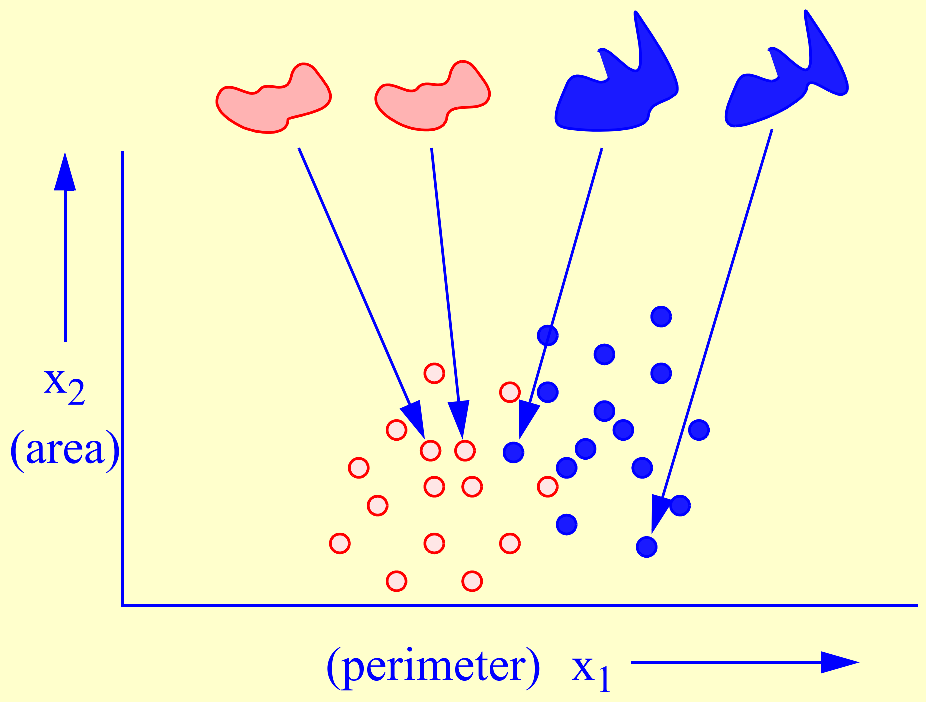
Feature Space

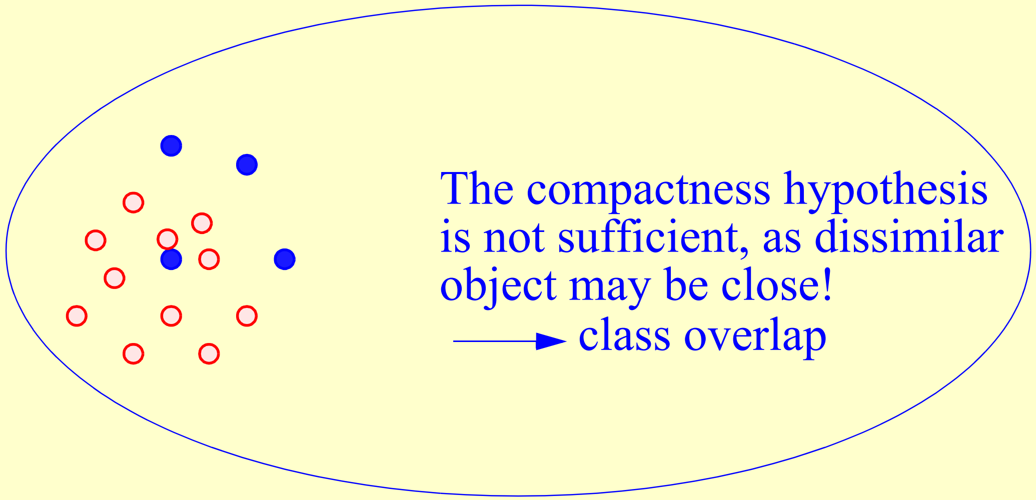
The Compactness Hypothesis

Representations of real world similar objects are close.

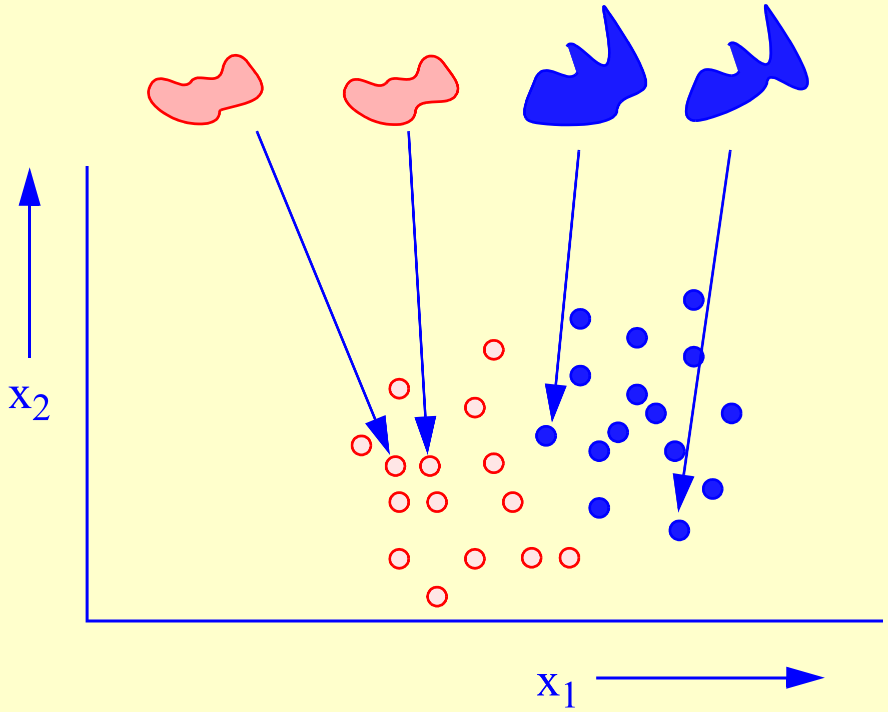
There is no ground for any generalization (induction) on representations
that do not obey this demand.

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





True Representations



Similar object should be close and dissimilar objects should be distant

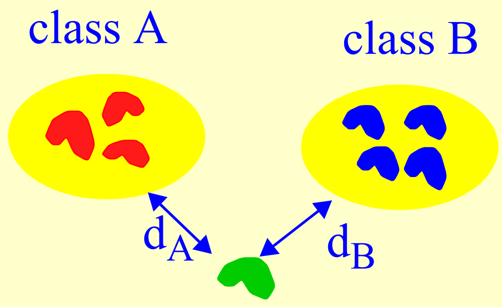
→ Dissimilarity representations based on measurement signals describing the ’whole’ object fulfill this.

TUDelft

Representation Principles

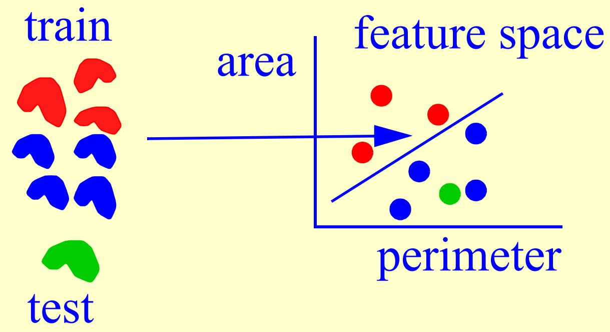
Model based, dissimilarities of objects and classes (Conceptual)

domains, structural, connectivity included



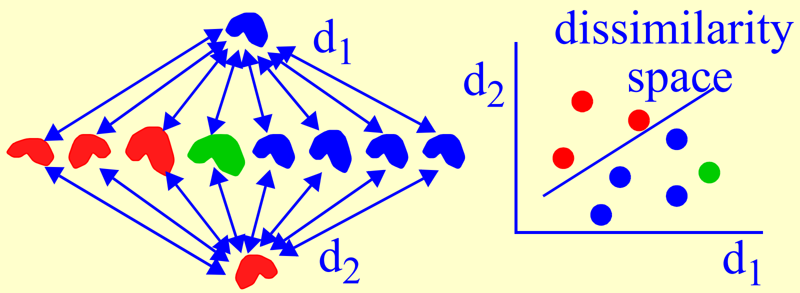
Feature based (Absolute)

distributions, connectivity neglected



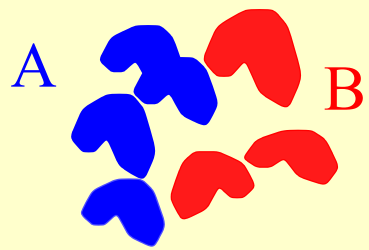
Dissimilarity based (Relative)

distributions or domains, connectivity possibly included

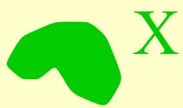


Dissimilarity Representation (DisRep)

Define dissimilarity measure d_{ij} between raw data of objects i and j



Given labeled training set T



Unlabeled object x to be classified

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

$$d_x = (d_1 \;\; d_2 \;\; d_3 \;\; d_4 \;\; d_5 \;\; d_6 \;\; d_7)$$

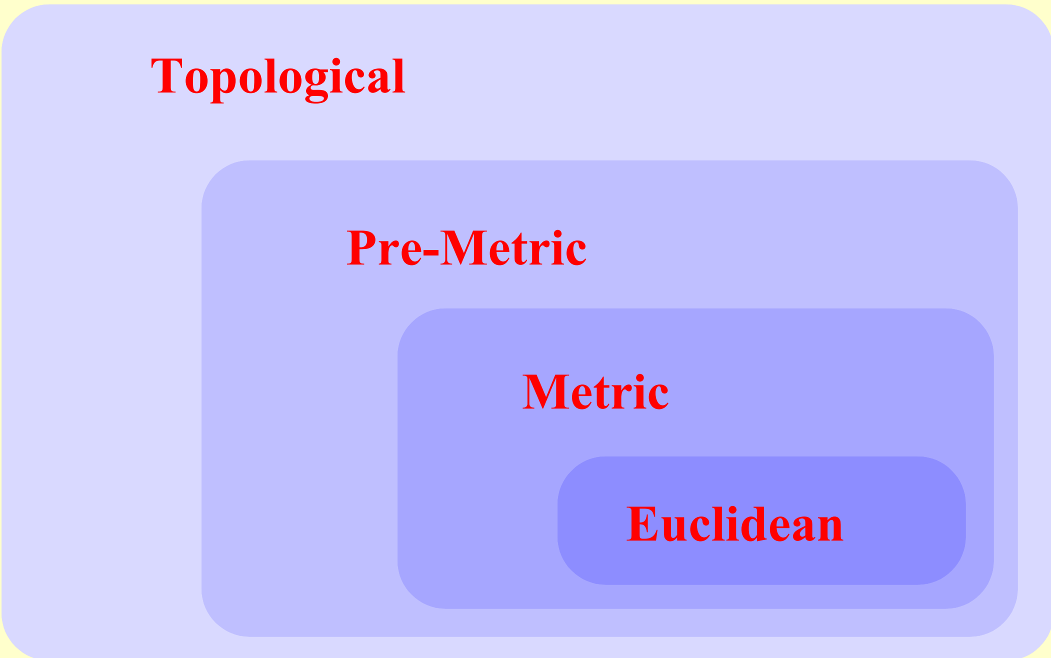
The traditional Nearest Neighbor rule (template matching) just finds:

label(argmin$_{\text{trainset}}$(di)),

without using DT. Can we do any better?

Dissimilarity Spaces - Examples

Pre-Topological ~~continuity~~



Mahalanobis ~~definiteness~~

Lp, p < 1
weighted edit-distance ~~triangle inequality~~

L1

L2, RMSE

Compactness always needed

Why Dissimilarity Spaces?

Many (exotic) dissimilarity measures are used in pattern recognition

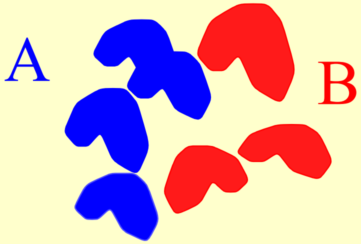
- they may solve the connectivity problem (e.g. pixel based features)
- they may offer a way to integrate structural and statistical approaches
  e.g. by graph distances.

Prospect of zero-error classifiers by avoiding class overlap

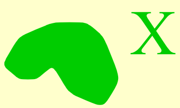
Better rules than the nearest neighbour classifier appear possible

(more accurate, faster)

DisRep Approach: NN Rule, Pre-topological Space



Given labeled training set T



Unlabeled object x to be classified

$d_x = (d_1 \;\; d_2 \;\; d_3 \;\; d_4 \;\; d_5 \;\; d_6 \;\; d_7)$

$\text{class}(x) = \text{label}(\text{argmin}(d_i))$

- Computationally expensive
- Locally sensitive
- Consistent: if size(T) --> ∞ then error --> 0

Dissimilarity Representation Approach: Dissimilarity Space

Dissimilarities

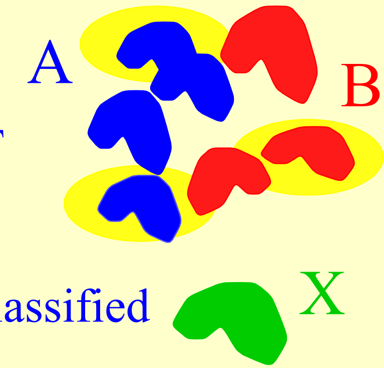
$$
\begin{array}{cc}
 & \begin{array}{ccccccc} & r_1 & & r_2 & & & r_3 \end{array} \\
D_T = & \begin{pmatrix}
d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} & d_{17} \\
d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} & d_{27} \\
d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} & d_{37} \\
d_{41} & d_{42} & d_{43} & d_{44} & d_{45} & d_{46} & d_{47} \\
d_{51} & d_{52} & d_{53} & d_{54} & d_{55} & d_{56} & d_{57} \\
d_{61} & d_{62} & d_{63} & d_{64} & d_{65} & d_{66} & d_{67} \\
d_{71} & d_{72} & d_{73} & d_{74} & d_{75} & d_{76} & d_{77}
\end{pmatrix}
\end{array}
$$

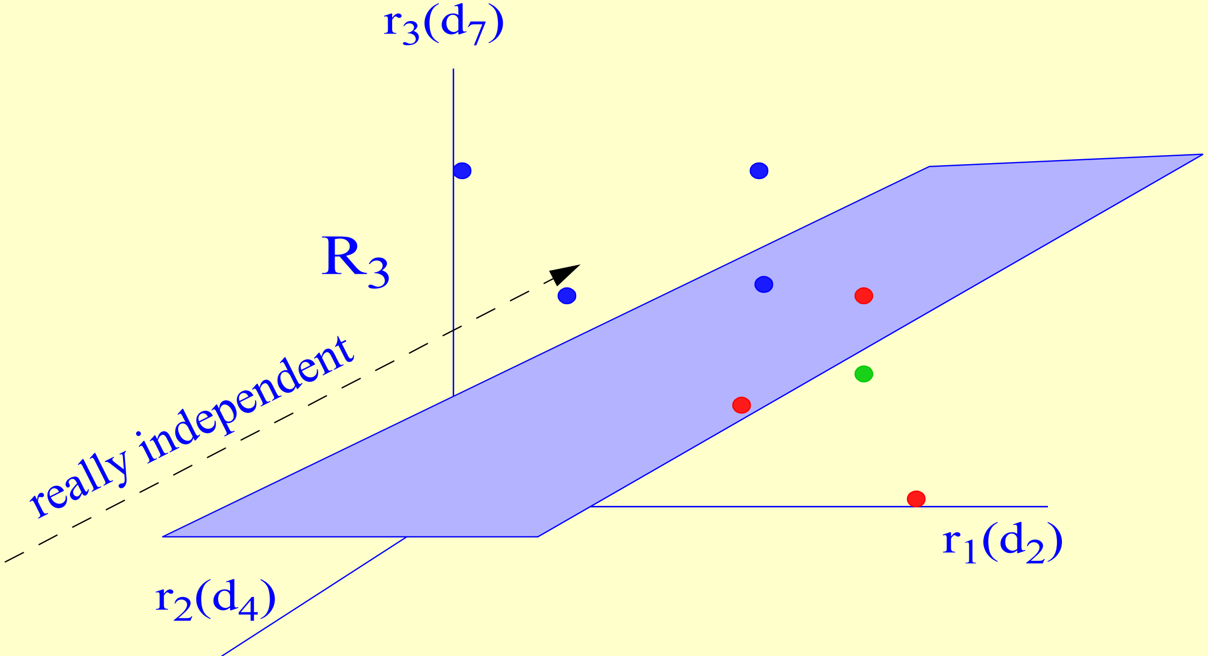
$$d_x = (\, d_1 \;\; d_2 \;\; d_3 \;\; d_4 \;\; d_5 \;\; d_6 \;\; d_7 \,)$$

Selection of 3 objects for representation

Given labeled training set T

Unlabeled object x to be classified

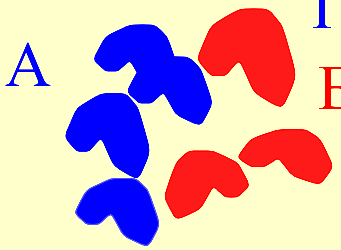




Classification in a 3D representation space

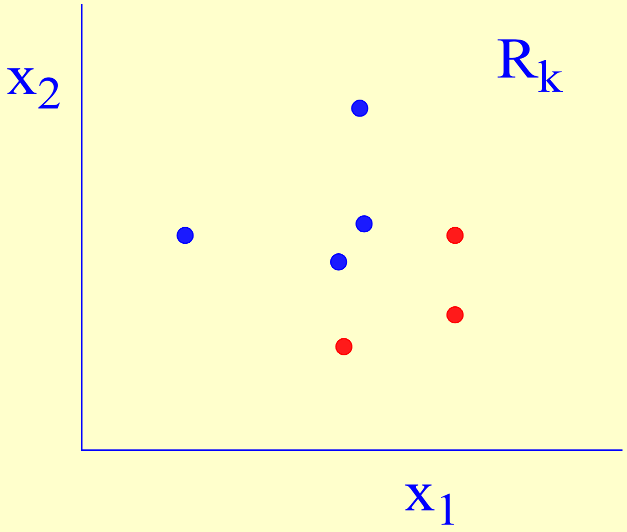
Dissimilarity Representation Approach: Embedding

Training set



→ Dissimilarity matrix D → X?

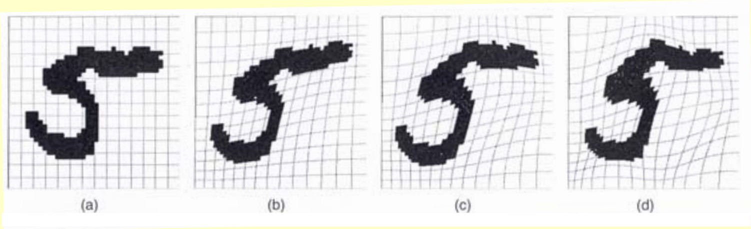
Is there a feature space X for which Dist(X,X) = D?

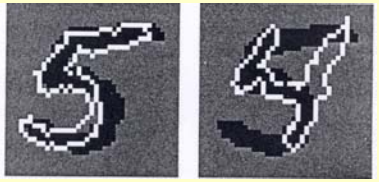


Euclidean distances → D

If D is non-Euclidean, embedding results in a pseudo-Euclidean Space (*Goldfarb, Pekalska*)

Digit Classification Example

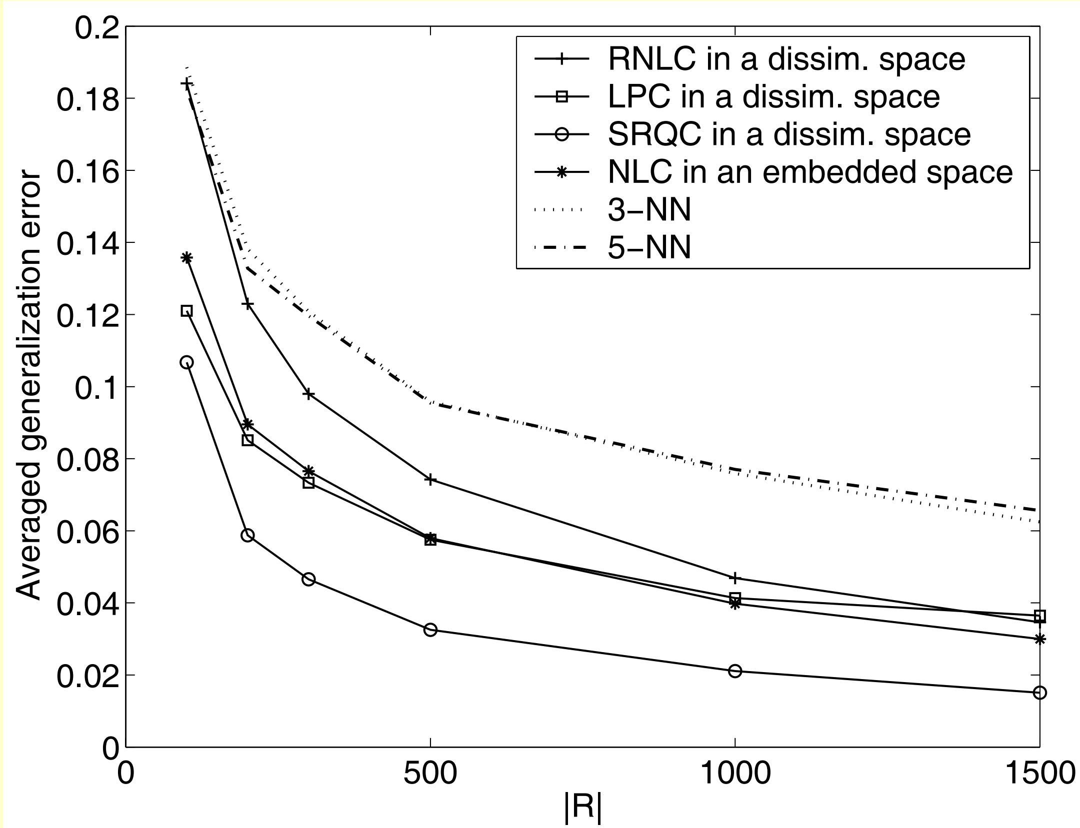




Matching new objects x to various templates y

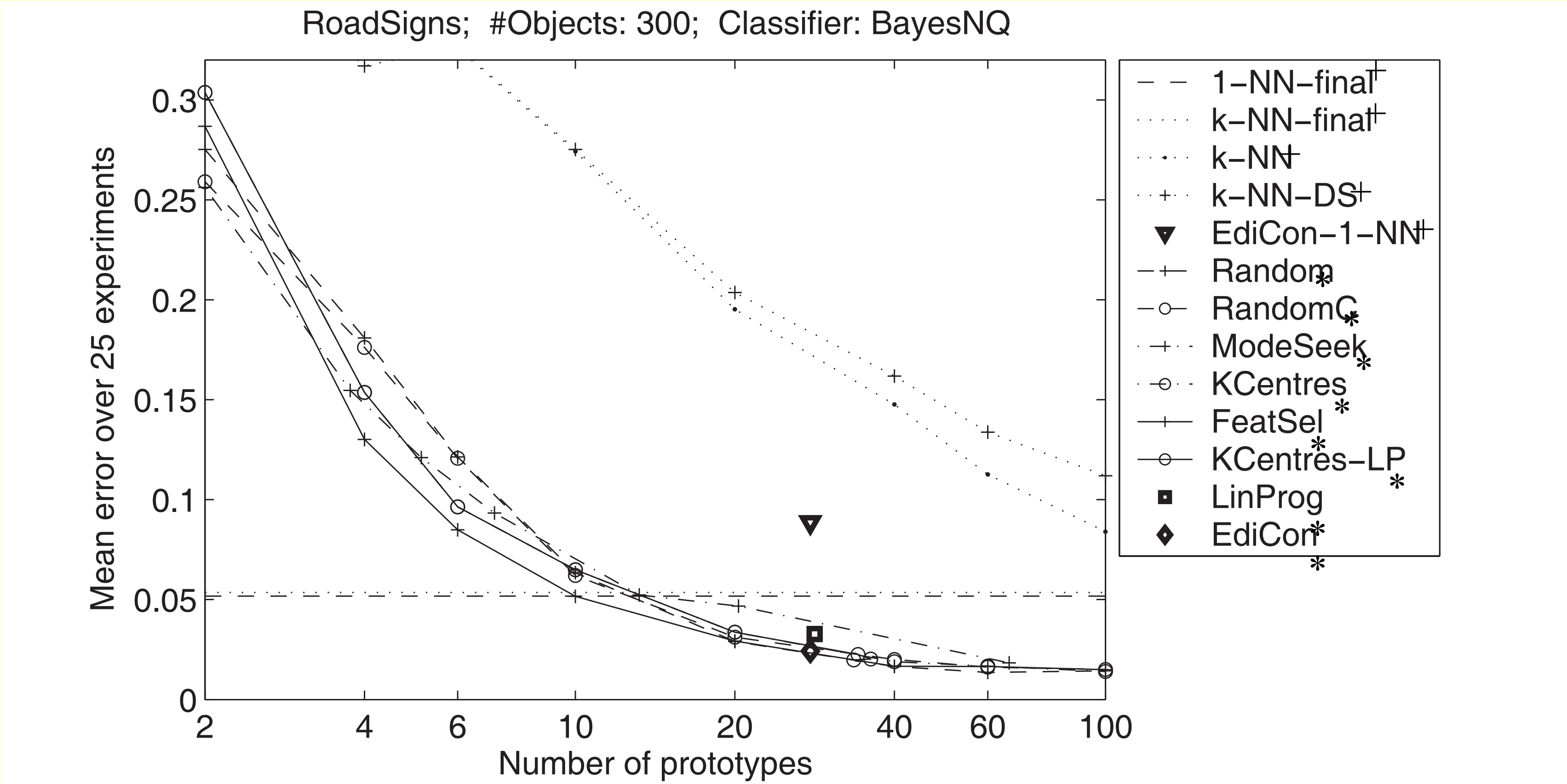
$$\text{class}(x) = \text{class}(\text{argmin}_y(D(x, y)))$$

A.K. Jain, D. Zongker, PAMI, vol. 19, no. 12, 1997.



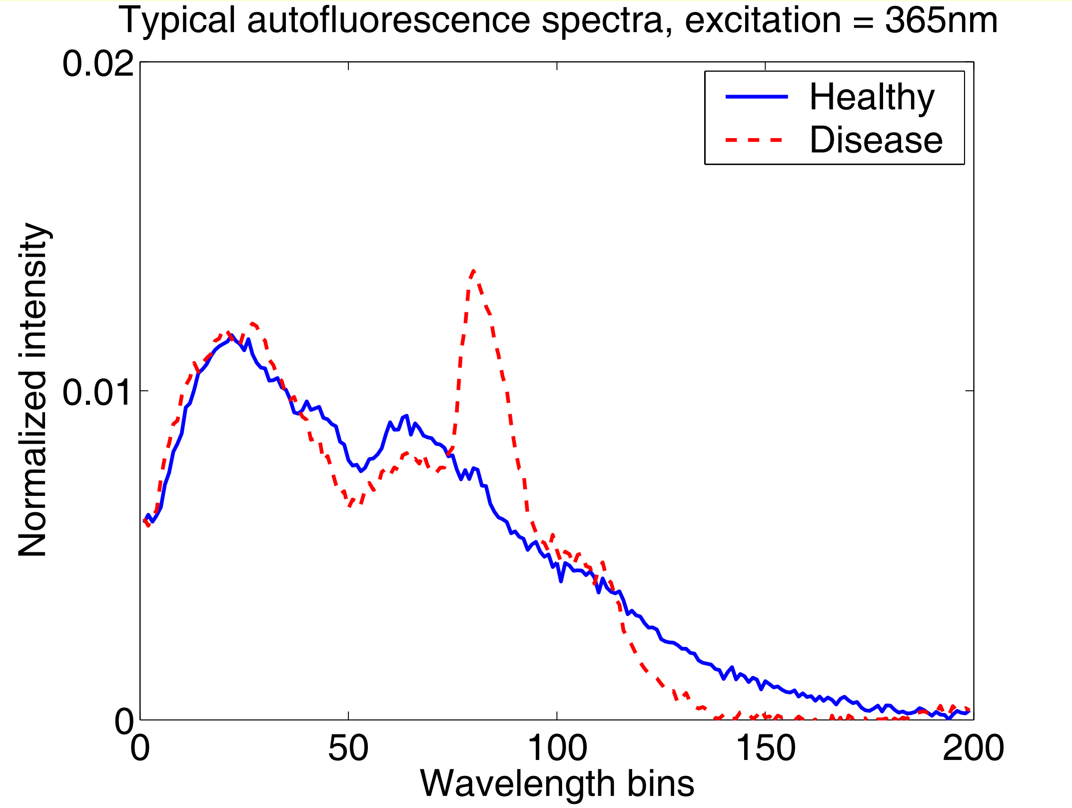
Dissimilarity based classifiers compared to the nearest neighbor rule for a 10-class digit classification problem.

Prototype Selection for Road Sign Recognition



The averaged error (over 25 experiments) of the quadratic Bayes classifier (*) in dissimilarity spaces of various dimensionalities, based on a series of selection procedures. For comparison a number of nearest neighbor results (+) are presented.

Combining Dissimilarity Measures for Spectra Recognition

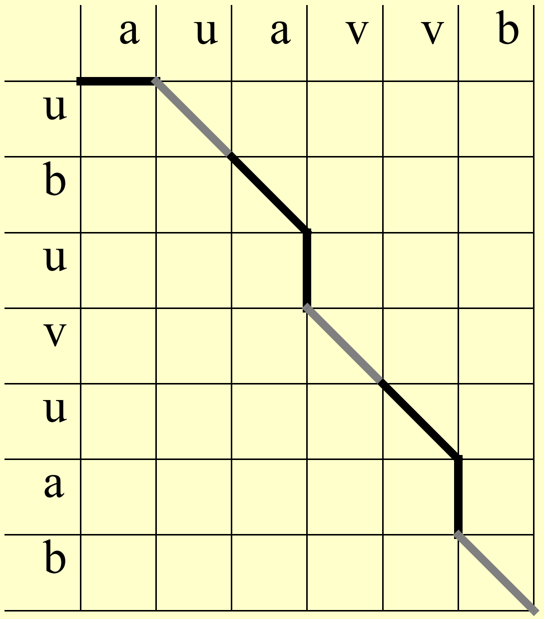


. Dissimilarity combining experiment

	Dis. Rep.	AUC (st. dev)	#SO
Dis Measure	$D^{(1)}$	72.3 (0.7)	2.5
	$D^{(2)}$	72.0 (0.7)	2.8
	$D^{(3)}$	78.2 (0.6)	2.7
	$D^{(4)}$	68.1 (0.8)	3.1
	$D^{(5)}$	75.1 (0.6)	2.1
Comb. Dis.Meas.	Mean	93.1 (0.5)	4.9
	Prod	93.6 (0.4)	4.6
	Min	85.0 (0.6)	15.3
	Max	84.1 (0.9)	7.2

Typical examples of two auto-fluorescence spectra in the oral cavity

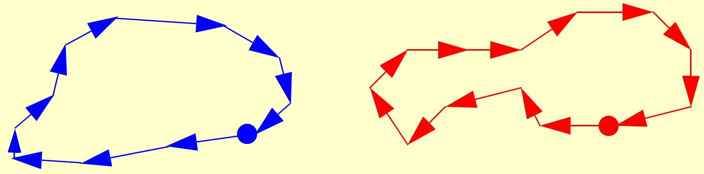
Representation by Orders Sets (Strings); Edit Distance



Possibly weighted.

Triangle inequality -->
computational feasible.

Length normalisation problem:
D(aa,bb) < D(abcdef,bcdd)



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

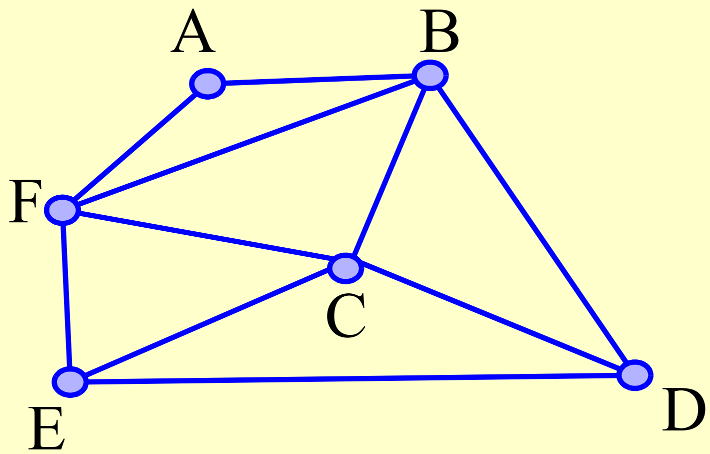
$D_E(X,Y)$: Σ edit operations X --> Y
(insertions, deletions, substitutions)

D_E(snert ,meer) = 3:
snert --> seert --> seer --> meer

D_E(ner ,meer) = 2:
ner --> mer --> meer

See Marzal & Vidal, IEEE PAMI-15, 1993, 926-932

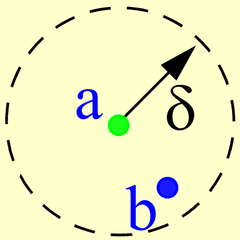
Representation by Connected Sets (Graphs)



Graph (Nodes, Connections, Attributes)

Distance (Graph_1 , Graph_2)

The Prospect of Dissimilarity based Representations: Zero Error



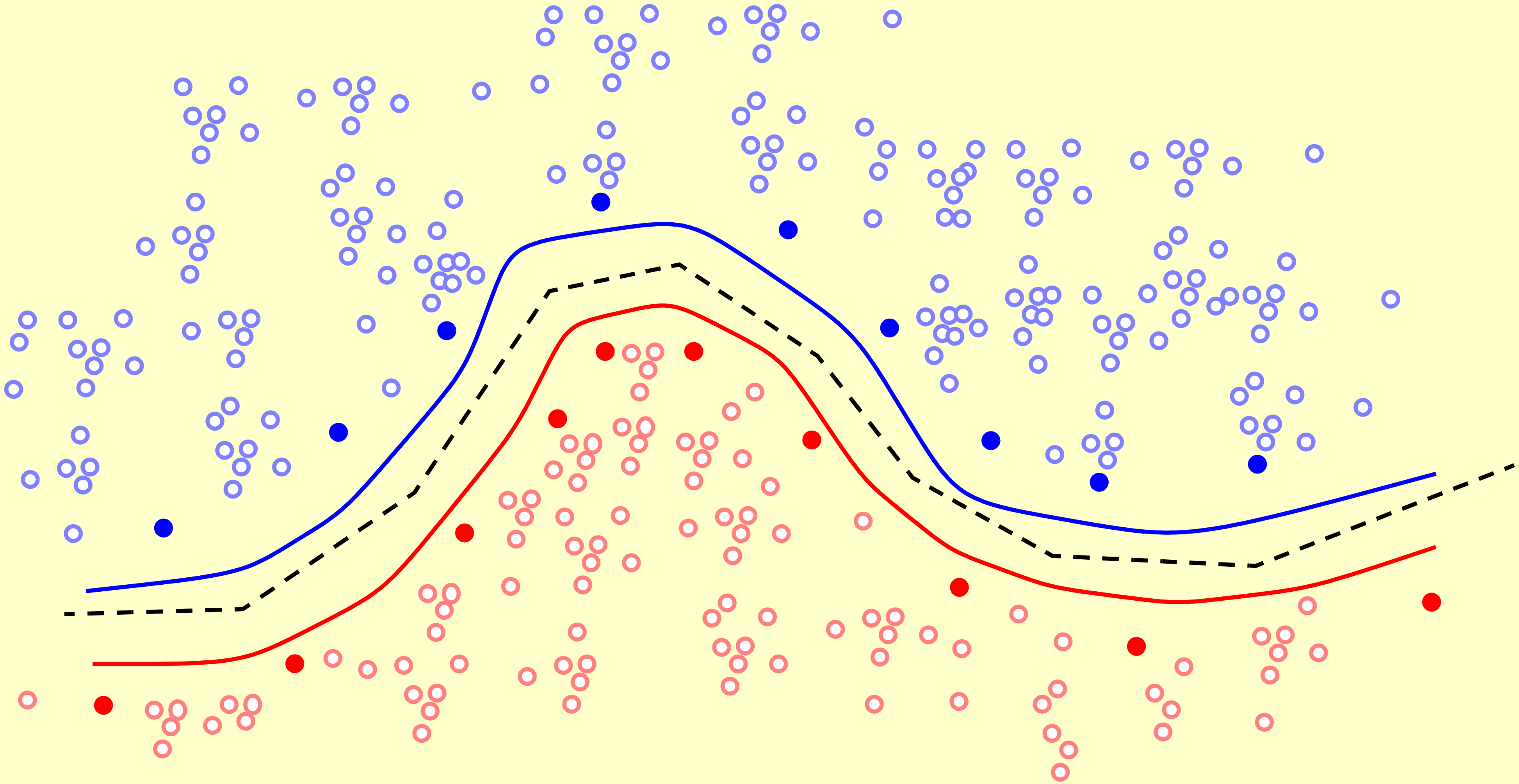
Let us assume that we deal with true representations:

$d_{ab} < \delta$ if and only if the objects a and b are very similar.

If δ is sufficiently small than 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!

Zero-Error Classification



Domain Based Classification

Objects from different classes have non-zero distance (assumption)

d(apple1,apple2) $\geq$ 0, d(apple, pear) > 0

$\rightarrow$ Classes don't overlap

$\rightarrow$ Probabilistic approaches are not needed

$\rightarrow$ No need for stochastic sampling

$\rightarrow$ Good for ill-defined, ill sampled problems, or problems with unknown priors.

Towards Domain Classifiers: Remove Probabilistic Contributions

Probability Arguments

Distance Arguments

Criterion : Min: $\varepsilon = \text{Prob}(S(x) \rightarrow \omega \mid x \notin \omega)$

Neural Net Min: $\sum_i (S(x_i, w) - \lambda(x_i))^2$

Fisher: $S(x) = y = w \bullet x + w_0$, such that $L = \dfrac{(\bar{y}_1 - \bar{y}_2)^2}{\sigma_1^2 + \sigma_2^2}$ is minimum

Parzen: $S(x) = \dfrac{1}{n_1} \sum_{i \in \omega_1} \varphi(|x - x_i|, h_1) - \dfrac{1}{n_2} \sum_{i \in \omega_2} \varphi(|x - x_i|, h_2)$

SVC: Max: $\min_i \{S(x_i)\lambda(x_i)\} + \sum_j \xi_j S(x_j)\lambda(x_i)$

Domain Based Classifiers

Criterion : Max: $\delta = \min_x(S(x)\lambda(x))$

Distance Arguments

Neural Net Min: $\max_i((S(x_i, w) - \lambda(x_i))^2)$

Fisher: $S(x) = y = w \bullet x + w_0$, such that $L = \dfrac{(\bar{y}_1 - \bar{y}_2)^2}{d_1^2 + d_2^2}$ is minimum

Parzen: $S(x) = \max_i(\varphi(|x - x_i|, h_1)) - \max_i(\varphi(|x - x_i|, h_2))$

SVC: Max: $\min_i\{S(x_i)\lambda(x_i)\}$

Can Dissimilarity Measures Be Learned?

Dissimilarity based Objects Class Differences Classifier

If the representation is based on dissimilarities of raw measurements, an optimization of the dissimilarity measure has to be based on raw data.

Should dissimilarities or similarities be used?

Ways to Adjust or Constitute Dissimilarity Measures

1. Combining different dissimilarity measures (compare combining classifiers)
2. Combining dissimilarities (similarities) for different object parts
3. Monotonic transformations of given dissimilarities
4. Transforming non-Euclidean distances to Euclidean distances*

**See Pekalska, SSSPR2004, On not making dissimilarities Euclidean*

Conclusions

Dissimilarity based pattern recognition uses class differences during representation.

In many applications it is a good alternative for the feature based approach.

It thereby may combine structural approaches with learning from examples.

For some applications class overlaps may be avoided and domain based classifiers become of interest.

Learning and improving dissimilarity measures have to be studied further.