Pattern Recognition as a Human Centered non-Euclidean Problem

ICEIS, Funchal, Madeira, 9 June 2010

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


Human Judgments ...

... often don't fit in an Euclidean world.

Learning about the world

Human knowledge grows in the debate between
- those who see the patterns, and
- those who know the universal laws

Automatic Pattern Recognition

Can we replace our recognition of real world objects by a formal system, also when our mental system is non-Euclidean? How to train? How to apply?

Blob Recognition

Which group?
Object Recognition

Airplane
Bicycle
Bus
Car
Train

Pattern Recognition: Speech

Experts confused by details

The diagnostic classification accuracy of a group of doctors. Sample size is 100.

Computers confused by details

Character recognition classification performance as a function of the number of n-tuples used.

Pattern Recognition: Shapes

Examples of objects for different classes

Object of unknown class to be classified

Pattern Recognition System

Feature Representation

area
perimeter

Representation

Generalization

Sensor
Measuring Human Relevant Information

Nearest neighbours sorted:

Pixel Representation

- Features:
  - Shapes
  - Moments
  - Fourier descriptors
  - Faces
  - Morphology

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

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

Pattern Recognition Paradox

Classification error vs. training set size

The Connectivity Problem in the Pixel Representation

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

Representations

- Features: details lost
- Pixels: no connectivity
- Dissimilarities: shape dependent
Examples Dissimilarity Measures

- Hausdorff Distance (metric): 
  \[ DH = \max(\max_a(\min_b(d(a,b))), \max_b(\min_a(d(a,b)))) \]

- Modified Hausdorff Distance (non-metric): 
  \[ DM = \max(\text{mean}_a(\min_b(d(a,b))), \text{mean}_b(\min_a(d(b,a)))) \]

### Dissimilarity Representation

- Training set
- Dissimilarities \( d_{ij} \) between all training objects
- Unlabeled object \( x \) to be classified

The traditional Nearest Neighbor rule (template matching) finds:

\[ \text{label}(\arg\min_{\text{train set}} d(x_i)) \]

without using \( D_{ij} \). Can we do any better?

### Pattern Recognition System

- Sensor → Representation → Generalization

### Alternatives for the Nearest Neighbor Rule

1. Dissimilarity Space
2. Embedding


Alternative 1: Dissimilarity Space

Dissimilarities

Given labeled training set

Unlabeled object to be classified

Selection of 3 objects for representation

Dissimilarity Space Classification ↔ Nearest Neighbor Rule

Dissimilarity based classification outperforms the nearest neighbor rule.

Embedding

→ Dissimilarity matrix \( D \) → \( X \)

Training set

Is there a feature space for which \( \text{Dist}(X,X) = D \) ?

Position points in a vector space such that their Euclidean distances \( D \)

Euclidean - Non Euclidean - Non Metric

Non-metric distances

Weighted-edit distance for strings

Single-linkage clustering

Fisher criterion

\[ J(A,B) = \frac{\mu_A - \mu_B}{\sigma_A + \sigma_B} \]

\[ J(A,C) = 0 \]

\[ J(A,B) = \text{large} \]

\[ J(C,B) = \text{small} \neq J(A,B) \]
(Pseudo-)Euclidean Embedding

Let $D$ be a given, imperfect dissimilarity matrix of training objects. Construct inner-product matrix: $B = -\frac{1}{2}JD - \frac{1}{2}11$  
Eigenvalue Decomposition, $B = QAQ^T$
Select k eigenvectors: $X = Q_kA_k$  (problem: $A_k < 0$)
Let $S$ be a k x k diag matrix, $S_{i,i} = \text{sign}(A_k(i,i))$

$A_k(i,i) < 0 \rightarrow$ Pseudo-Euclidean
The inner-product matrix: $B = -\frac{1}{2}(D - \frac{1}{2}11^T)D^T$
The embedded objects: $Z = BQ_kA_k^T S_k$

PES: Pseudo-Euclidean Space (Krein Space)

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

$$\langle x, y \rangle = x^T Q_n y = \sum_{i} x_i y_i - \sum_{p} x_p y_p$$

$$d_i^2(x,y) = \langle x - y, x - y \rangle = d_i^2(x,y) - d_i^2(y,x)$$

Dissimilarity based classification procedure compared

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

Three Approaches Compared for the Zongker Data

Dissimilarity Space equivalent to Embedding better than Nearest Neighbour Rule

Non-Euclidean Representations

- Why do we have them?
- Are they essential?
- Can we build classifiers for them? (to some extend)
- Can we transform them into Euclidean representations? (Yes, but at the cost of performance loss)
Computational Noise

Even for Euclidean distance matrices zero eigenvalues may show negative, e.g:
- \( X = N(50,20) \) : 50 points in 20 dimensions
- \( D = \text{Dist}(X) \): 50 x 50 distance matrix
- Expected: 49-20 = 29 zero eigenvalues
- Found: 15 negative eigenvalues

Lack of information

1890:
Crossing the Jostedalsbreen was impossible. Travelling around (200 km) lasted 5 days. Until the shared point \( X \) was found.
People could visit each other in 8 hours.

\( D(Y,J) = 5 \text{ days} \)
\( D(Y,X) = 4 \text{ hours} \)
\( D(X,J) = 4 \text{ hours} \)

Graph Matching \( \rightarrow \) Dissimilarities

Representation by Connected Graphs

Graph (Nodes, Edges, Attributes)
Distance (Graph_1, Graph_2)

Intrinsicly Non-Euclidean Dissimilarity Measures

Single Linkage

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

Single-linkage clustering

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

Boundary distances

A set of boundary distances may characterize sets of datapoints:
Distances \( \rightarrow \) features

Intrinsicly Non-Euclidean Dissimilarity Measures

Mahalanobis

Pairwise comparison between differently shaped data distributions
Different pairs \( \rightarrow \) different comparison frameworks
\( \rightarrow \) non-Euclidean
Intrinsic Non-Euclidean Dissimilarity Measures

Invariants

Object space

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

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

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