A Novel Data Representation based on Dissimilarity Increments

Helena Aidos and Ana Fred

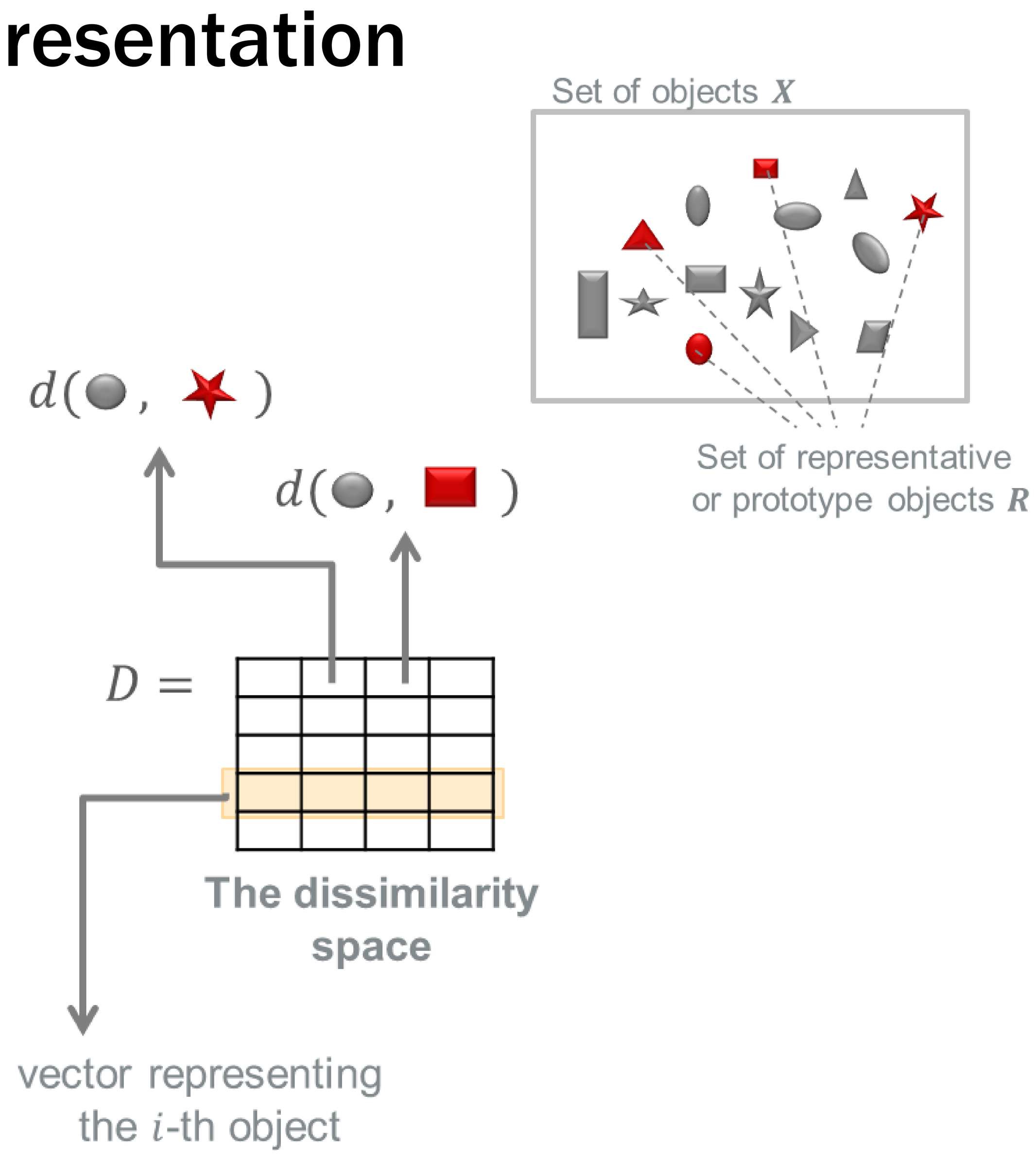
Instituto de Telecomunicações
Instituto Superior Técnico, Universidade de Lisboa, Portugal
{haidos, afred}@lx.it.pt

Motivation

- Typically, objects are represented by a set of features, which should characterize the objects and be relevant to discriminate among the classes.
- Problem: difficult to obtain a complete description of objects:
  - forces an overlap of the classes
  - leads to an inefficient learning process.
- Solution: Use a dissimilarity representation, which is based on comparisons between pairs of objects:
  - Solves the problem of class overlap, since only identical objects have a dissimilarity of zero.

Dissimilarity representation

- set of objects
- set of representative or prototype objects, such that
- Each object is described by a -dimensional dissimilarity vector
  where is a dissimilarity measure
  - is a row of the dissimilarity matrix , the **dissimilarity space**
- Define a vector space by , where the -th object is represented by the dissimilarity vector of the values.

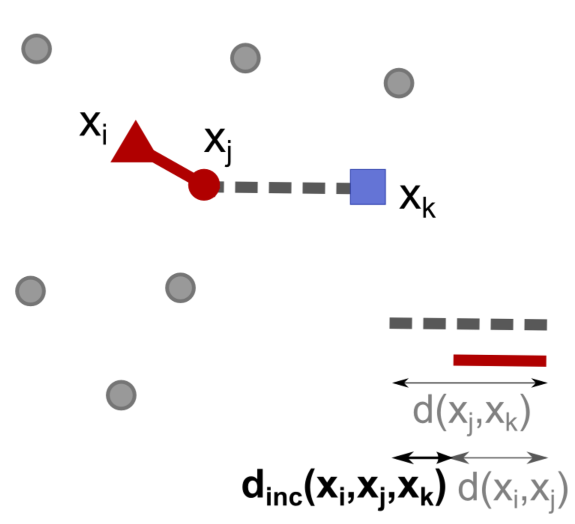


PROPOSAL: A novel dissimilarity representation of data, based on a second-order dissimilarity measure.

Second-order dissimilarity measure: the dissimilarity increments

Given some dissimilarity measure, , between patterns,

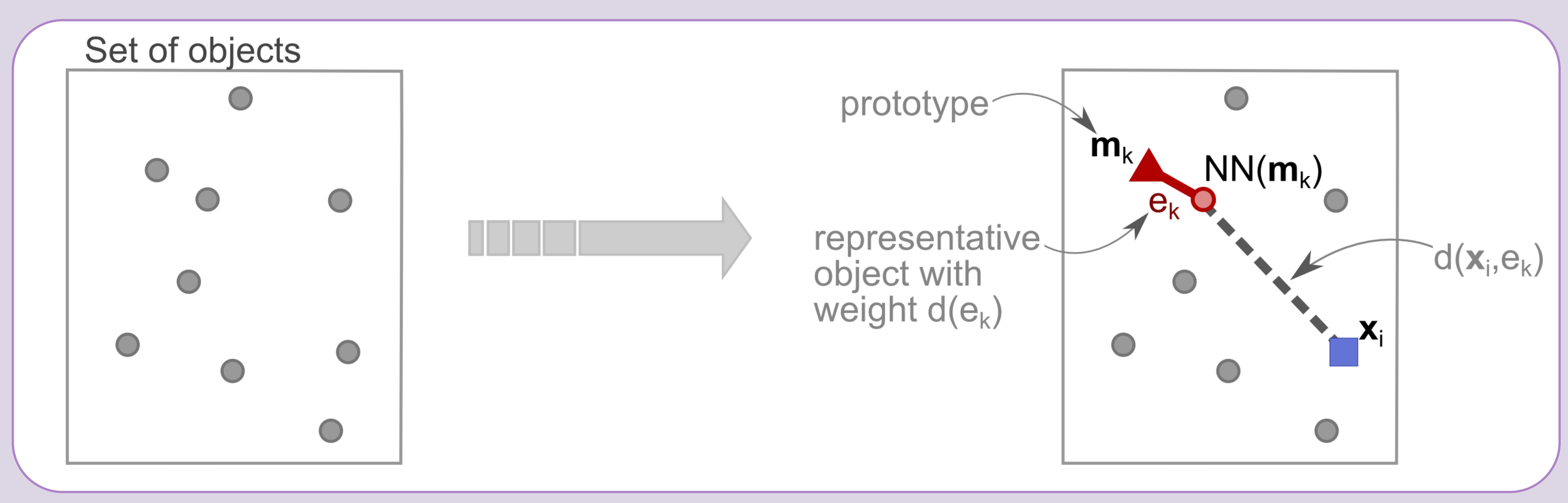
– triplet of nearest neighbor
- is the nearest neighbor of
- is the nearest neighbor of (different from)



The **dissimilarity increments** between neighboring patterns is defined as

Dissimilarity increments space

- set of prototype objects, with an edge between a prototype and its nearest neighbor
- weight of edge
- Distance between any object and the representative object is



The -th element of the Dinc space is defined as

Characterization

- Dissimilarity spaces have **higher discriminant power** of features in separating the classes.
- Dissimilarity spaces have **less overlap between the classes**, which may facilitate the learner to separate the samples of different classes.
- Even if the classes are more separable, they are nonlinearly separable by 1-NN classifier.

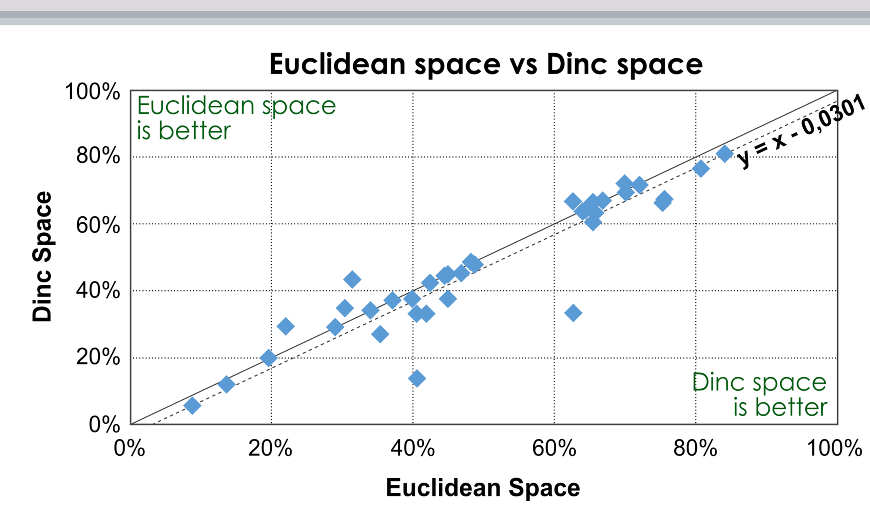
Assume that $R = X$, meaning that all objects of X are used as prototypes.

Euclidean dissimilarity space

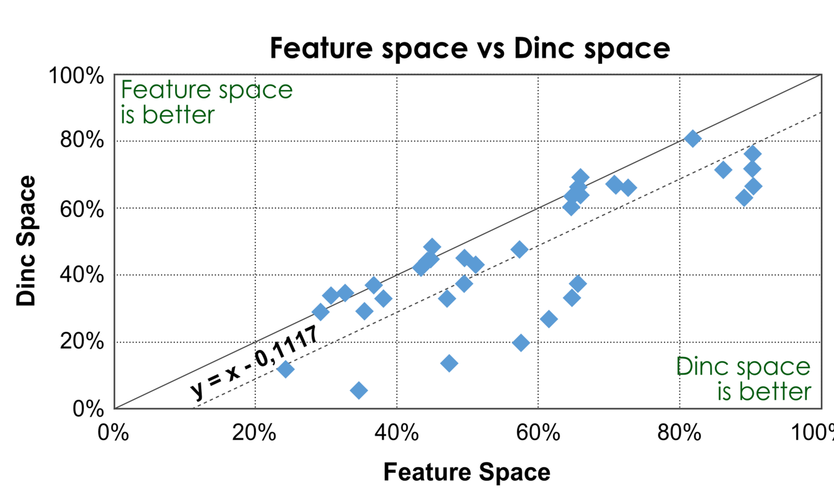
Each element, , of the dissimilarity matrix , is the Euclidean distance between -th and -th objects.

Datasets: 36 real-world datasets from the UCI Machine Learning repository
Evaluation: Error rates of median-link, when the true number of clusters is known

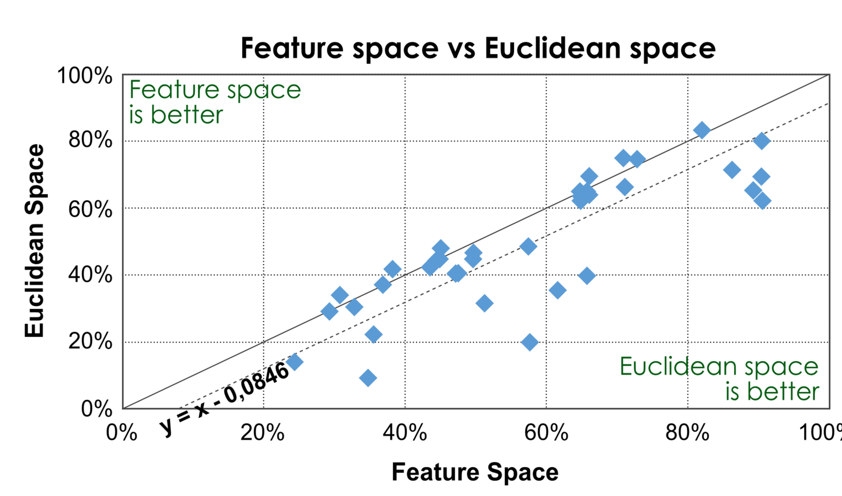
Experimental results



- 8 datasets for Euclidean space
  - Best on average 4.0% than Dinc space
- 18 datasets for Dinc space
  - Best on average 7.1% than Euclidean space



- 28 datasets for Dinc space
  - Best on average 13.6% than Feature space
- 6 datasets for Feature space
  - Best on average 2.2% than Dinc space



- 25 datasets for Euclidean space
  - Best on average 11.8% than Feature space
- 9 datasets for Feature space
  - Best on average 2.6% than Euclidean space

This work was supported by the Portuguese Foundation for Science and Technology scholarship number SFRH/BPD/103127/2014, and grant PTDC/EEI-SII/2312/2012.