A Novel Data Representation based on Dissimilarity Increments

Helena Aidos and Ana Fred

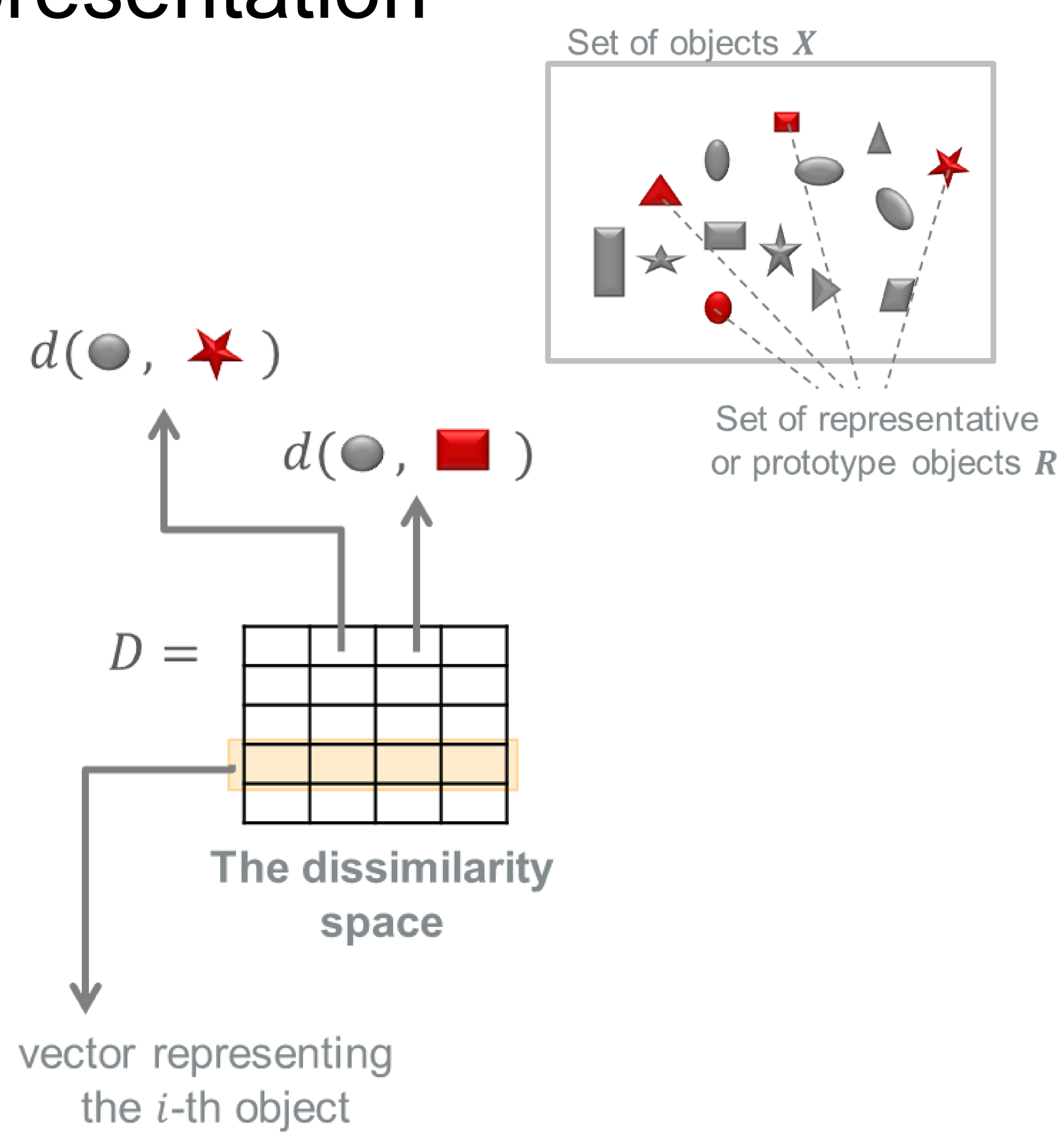
Instituto de Telecomunicações
Instituto Superior Técnico, Universidade de Lisboa, Portugal
{haidos, afred}@lx.it.pt

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

- $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ set of objects
- $R = \{\mathbf{e}_1, \dots, \mathbf{e}_r\}$ set of representative or prototype objects, such that $R \subseteq X$
- Each object $\mathbf{x}_i$ is described by a r-dimensional dissimilarity vector
$$D(\mathbf{x}_i, R) = [d(\mathbf{x}_i, \mathbf{e}_1) \dots d(\mathbf{x}_i, \mathbf{e}_r)]$$
where $d(\cdot,\cdot)$ is a dissimilarity measure
  - ➢ $D(\mathbf{x}_i, R)$ is a row of the $n \times r$ dissimilarity matrix D, the **dissimilarity space**
- Define a vector space Y by $Y = D$, where the i-th object is represented by the dissimilarity vector of the D_{ij} values.

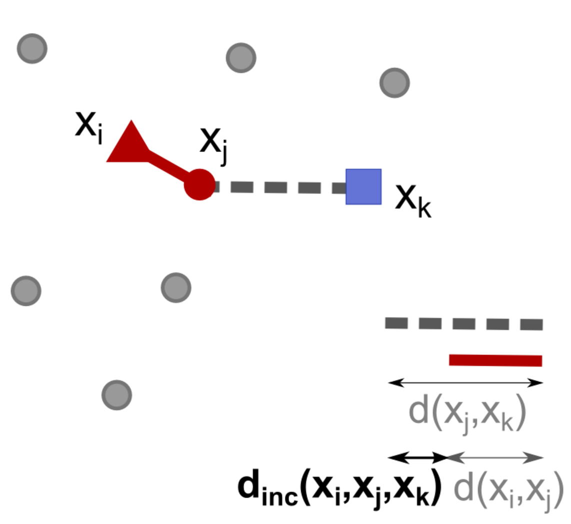


PROPOSAL: A novel dissimilarity representation of data, based on a second-order dissimilarity measure.

Second-order dissimilarity measure: the dissimilarity increments

Given some dissimilarity measure, $d(\cdot,\cdot)$, between patterns,

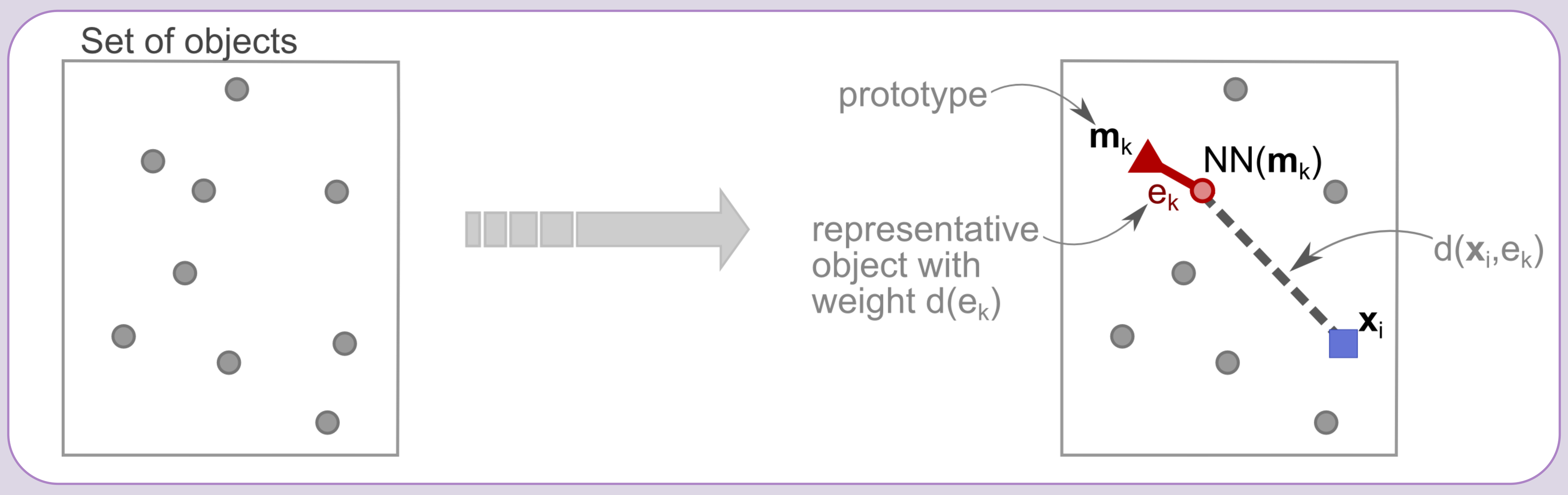
$(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$ – triplet of nearest neighbor
- $\mathbf{x}_j$ is the nearest neighbor of $\mathbf{x}_i$
- $\mathbf{x}_k$ is the nearest neighbor of $\mathbf{x}_j$ (different from $\mathbf{x}_i$)



The **dissimilarity increments** between neighboring patterns is defined as
$$d_{inc}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) = |d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{x}_j, \mathbf{x}_k)|$$

Dissimilarity increments space

- $R = \{\mathbf{e}_1, \dots, \mathbf{e}_r\}$ set of prototype objects, with $\mathbf{e}_j$ an edge between a prototype $\mathbf{m}_j$ and its nearest neighbor $\mathbf{x}_{\mathbf{m}_j} = NN(\mathbf{m}_j)$
- $d(\mathbf{e}_j) = d\left(\mathbf{m}_j, \mathbf{x}_{\mathbf{m}_j}\right)$ weight of edge $\mathbf{e}_j$
- Distance between any object $\mathbf{x}_i$ and the representative object $\mathbf{e}_j$ is
$$d(\mathbf{x}_i, \mathbf{e}_j) = \min\left\{d(\mathbf{x}_i, \mathbf{m}_j), d\left(\mathbf{x}_i, \mathbf{x}_{\mathbf{m}_j}\right)\right\}$$



The (i,j)-th element of the Dinc space is defined as
$$D(\mathbf{x}_i, \mathbf{e}_j) = |d(\mathbf{x}_i, \mathbf{e}_j) - d(\mathbf{e}_j)|$$

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.

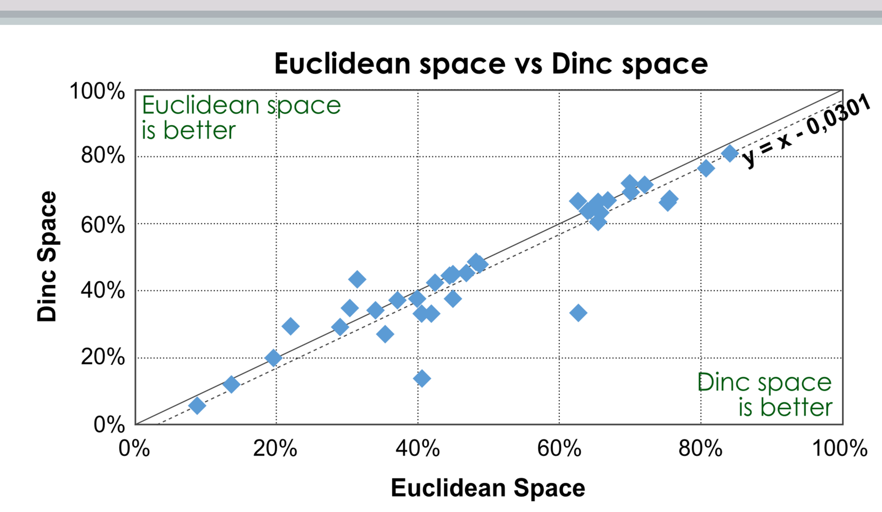
Assume that $R = X$, meaning that all objects of X are used as prototypes.

Euclidean dissimilarity space

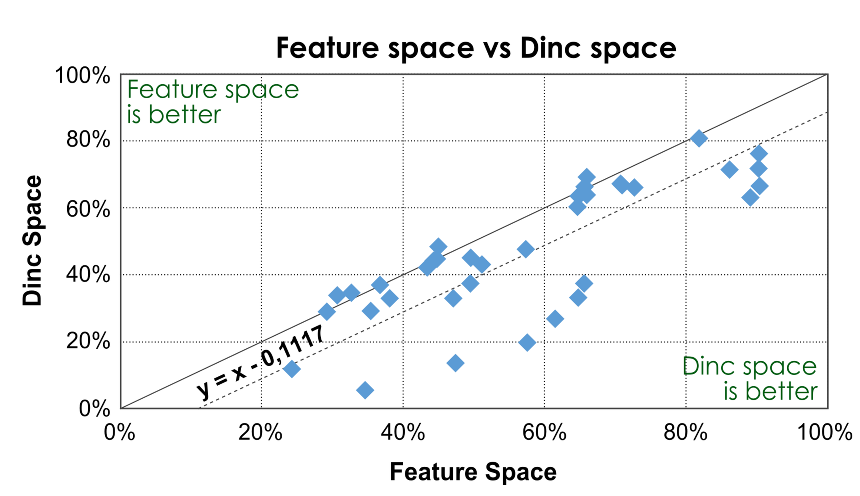
Each element, D_{ij}, of the dissimilarity matrix D, is the Euclidean distance between i-th and j-th objects.

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

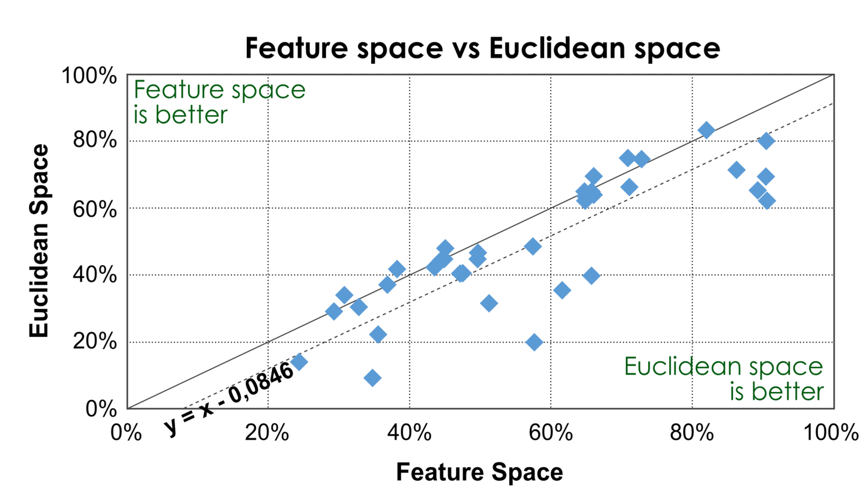
Experimental results



- 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.