A Family of Hierarchical Clustering Algorithms based on High-order Dissimilarities

Helena Aidos and Ana Fred

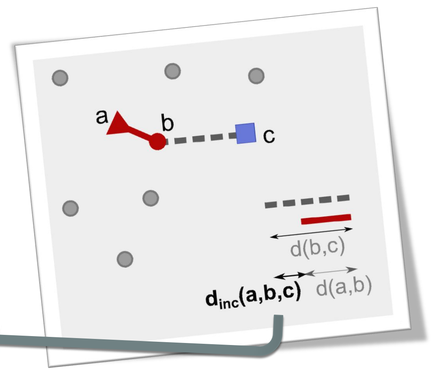
Instituto de Telecomunicações
Instituto Superior Técnico, Universidade de Lisboa, Portugal
{haidos, afred}@lx.it.pt

Dissimilarity increments: definition and distribution

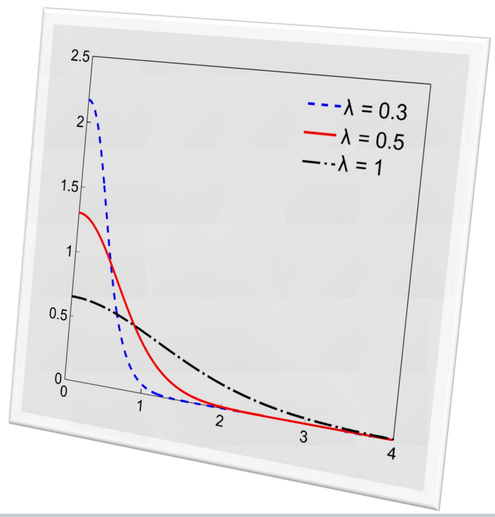
(x_i, x_j, x_k) – triplet of nearest neighbors
- x_j is the nearest neighbor of x_i
- x_k is the nearest neighbor of x_j (different from x_i)

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

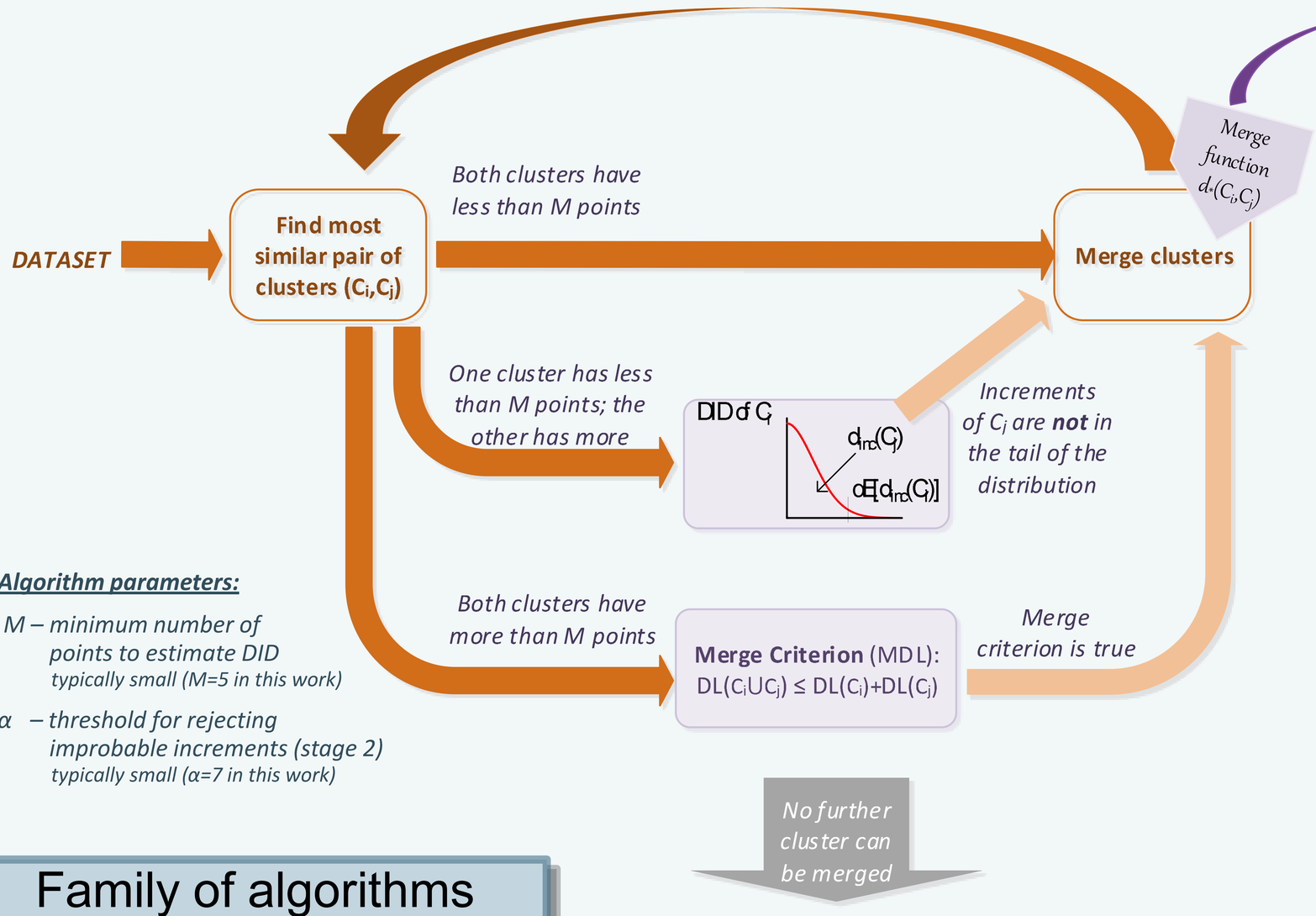
$$d_{inc}(x_i, x_j, x_k) = \left| d(x_i, x_j) - d(x_j, x_k) \right|$$



The **dissimilarity increments distribution (DID)** is a function of the mean value of the dissimilarity increments λ



PROPOSAL: A family of agglomerative hierarchical methods, integrating dissimilarity increments in traditional linkage algorithms



MERGE FUNCTION

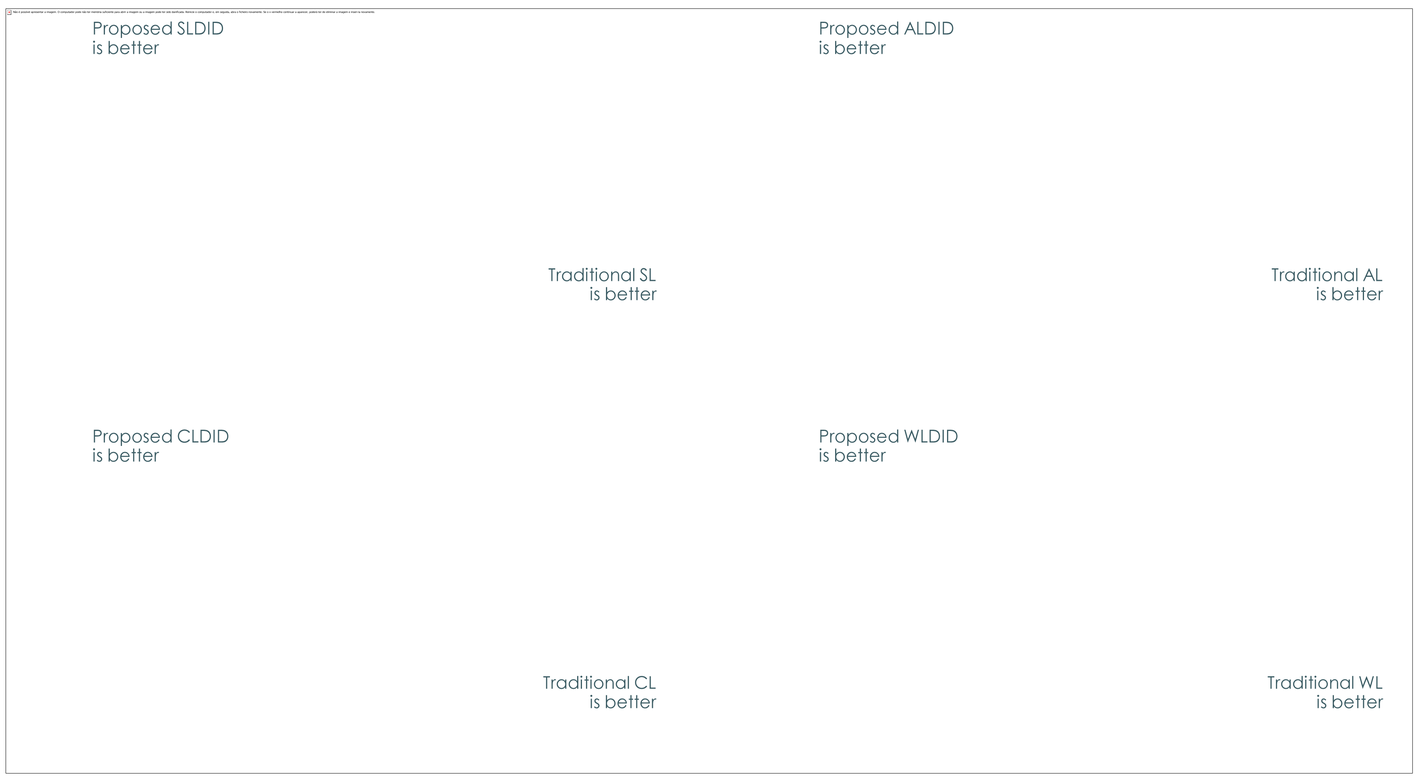
Consider the new formed cluster $C_b = C_i \cup C_j$, obtained by merging C_i and C_j, and C_a is one of the remaining clusters formed in previous steps. Lets consider $|C_i|$ the number of points in cluster C_i

- **SLDID**
$$d_s(C_a, C_b) = \min\{d(C_i, C_a), d(C_j, C_a)\}$$
- **ALDID**
$$d_A(C_a, C_b) = \frac{|C_i|\, d(C_i, C_a) + |C_j|\, d(C_j, C_a)}{|C_i| + |C_j|}$$
- **CLDID**
$$d_c(C_a, C_b) = \max\{d(C_i, C_a), d(C_j, C_a)\}$$
- **WLDID**
$$d_W(C_a, C_b) = \frac{|C_i| + |C_a|}{|C_i| + |C_j| + |C_a|} d(C_i, C_a) + \frac{|C_j| + |C_a|}{|C_i| + |C_j| + |C_a|} d(C_j, C_a) - \frac{|C_a|}{|C_i| + |C_j| + |C_a|} d(C_i, C_j)$$

Family of algorithms

➢ This family of hierarchical agglomerative algorithms is able to **automatically find the number of clusters** using a minimum description length criterion based on the dissimilarity increments distribution (DID)

➢ Each algorithm of the proposed family is able to find classes as unions of clusters, leading to the **identification of internal structures of classes**



DATASETS:
36 real-world datasets from the UCI Machine Learning Repository.

EVALUATION:
Percentage of correctly clustered points assuming that one class can be represented as the union of several clusters.

This work was supported by the Portuguese Foundation for Science and Technology grant PTDC/EEI-SII/2312/2012.