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

Helena Aidos and Ana Fred

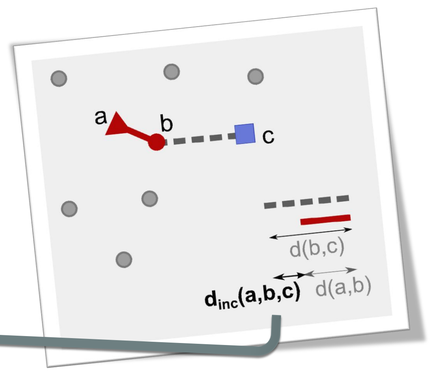
Instituto de Telecomunicações
Instituto Superior Técnico, Universidade de Lisboa, Portugal
{haidos, afred}@lx.it.pt

Dissimilarity increments: definition and distribution

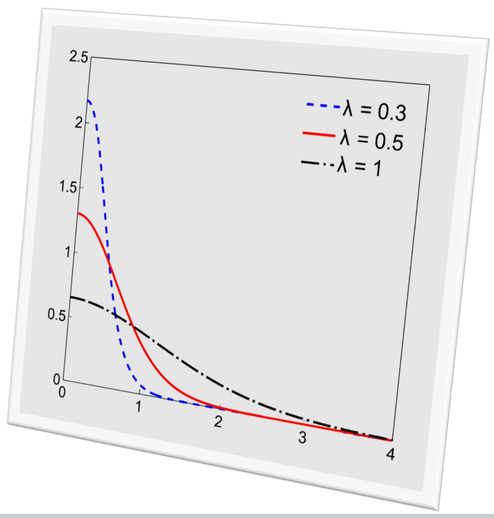
(, ,) – triplet of nearest neighbors
- is the nearest neighbor of
- is the nearest neighbor of (different from)

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

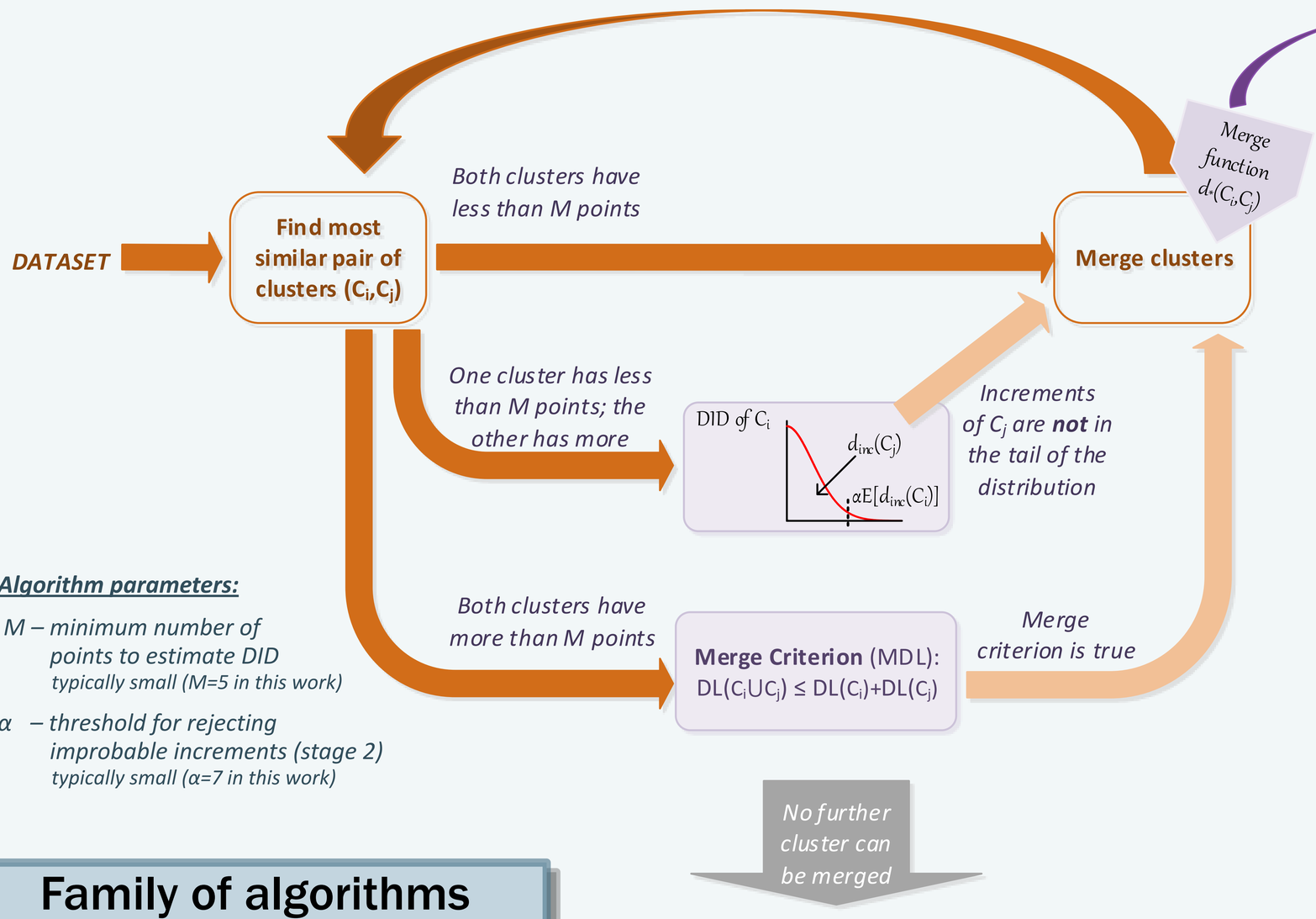
( ) | ( ) ( )|



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



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



Algorithm parameters:

M – minimum number of points to estimate DID
typically small (M=5 in this work)

α – threshold for rejecting improbable increments (stage 2)
typically small (α=7 in this work)

MERGE FUNCTION

Consider the new formed cluster = , obtained by merging and , and is one of the remaining clusters formed in previous steps. Lets consider | | the number of points in cluster

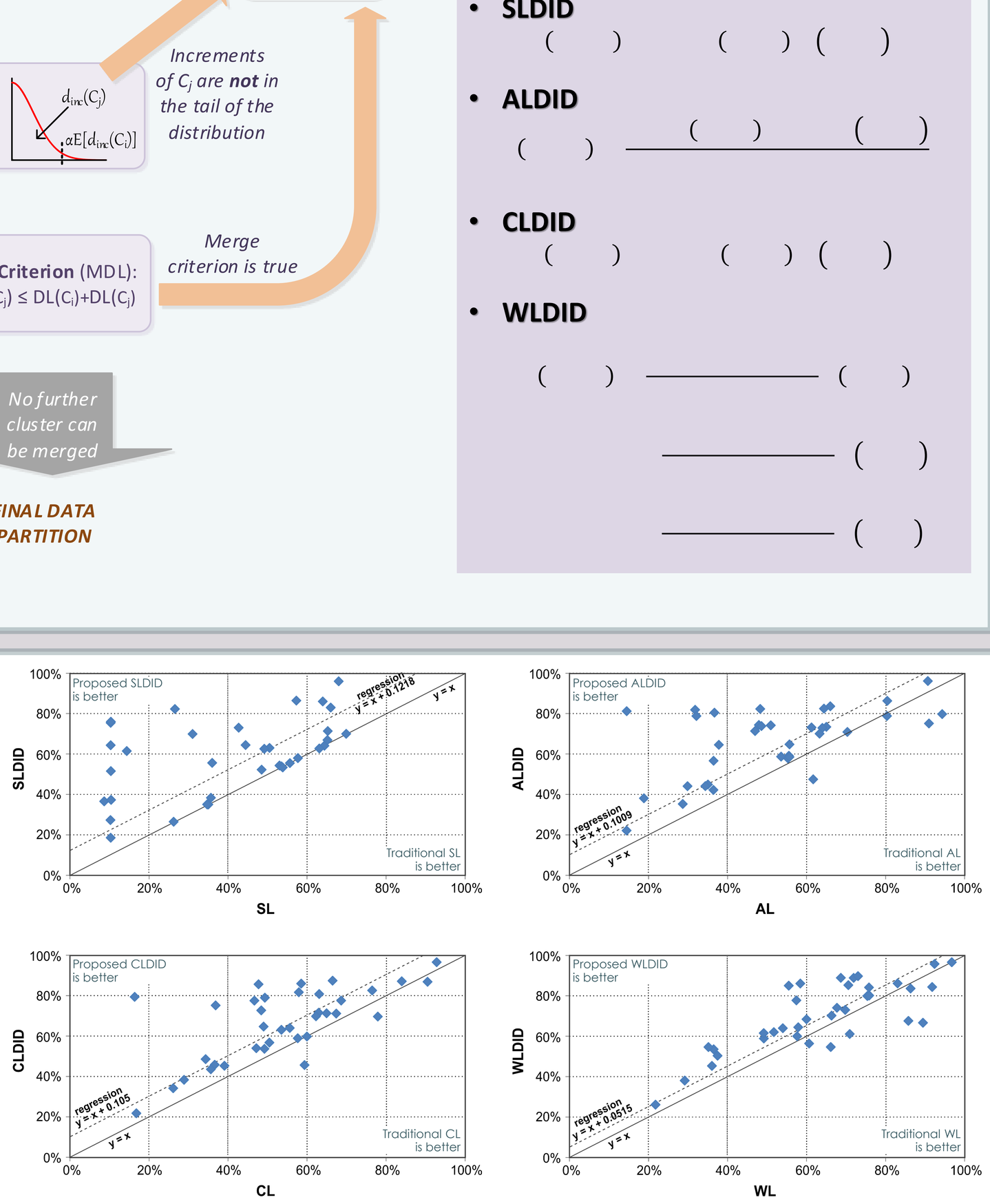
- **SLDID**
  ( ) ( ) ( )
- **ALDID**
  ( ) ( ) ( )
- **CLDID**
  ( ) ( ) ( )
- **WLDID**
  ( ) ( )
  ( )
  ( )

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**

Experimental results



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.