Normalized k-means clustering of hyper-rectangles

M. Chavent

Laboratoire de Mathématiques Appliquées de Bordeaux, UMR CNRS 5466
Universités Bordeaux1 et 2, France

chavent@math.u-bordeaux1.fr
Introduction

Three different prototypes for points of $\mathbb{R}^p$ with $L_1$, $L_2$ or $L_\infty$ distances:

$$\sum_{i=1}^{n} d_2^2(x_i, y)$$

$$\sum_{i=1}^{n} d_1(x_i, y)$$

$$\max_{i=1...n} d_\infty(x_i, y)$$

k-means

Dynamical Clustering
Introduction

Three different measures of centrality $y^j$:

- The mean
- The median
- The middle
Introduction

Three different measures of centrality $y^j$:

- **The mean**
- **The median**
- **The middle**

Three different measures of dispersion $\sigma^j$:

- **The square deviation**
$$\sum_{i=1}^{n} |x_i^j - y^j|^2$$
- **The absolute deviation**
$$\sum_{i=1}^{n} |x_i^j - y^j|$$
- **The max deviation**
$$\max_{i=1\ldots n} |x_i^j - y^j|$$
Normalized $L_\alpha$ distance between two $\mathbb{R}^p$-points:

$$d_\alpha(x_1, x_2) = \left( \sum_{j=1}^{p} \frac{1}{\sigma^j} |x_1^j - x_2^j|^{\alpha} \right)^{\frac{1}{\alpha}}$$

Three normalized k-means algorithms:

<table>
<thead>
<tr>
<th>Prototype</th>
<th>Distance</th>
<th>Measure of dispersion</th>
</tr>
</thead>
<tbody>
<tr>
<td>The “mean” $\mathbb{R}^p$-point</td>
<td>$L_2$-Euclidean</td>
<td>The square deviation from the mean</td>
</tr>
<tr>
<td>The “median” $\mathbb{R}^p$-point</td>
<td>$L_1$-City-Block</td>
<td>The absolute deviation from the median</td>
</tr>
<tr>
<td>The “middle” $\mathbb{R}^p$-point</td>
<td>$L_\infty$-Max</td>
<td>The max deviation from the middle</td>
</tr>
</tbody>
</table>
Plan

Part 1  Interval data
Part 2  Comparing hyper-rectangles
Part 3  Define a class prototype
Part 3  Normalization
PART 1

Interval data
### French Guyana Fish Example

- Mercury contamination in some Amerindian populations in French Guyana

- Data table: 67 fish of 10 different species described by 5 quantitative variables based on the mercury concentration (μg/g) in five organs (gills, liver, intestine, stomach, kidney)

<table>
<thead>
<tr>
<th>Fish</th>
<th>liver</th>
<th>kidney</th>
<th>gills</th>
<th>intestine</th>
<th>stomach</th>
<th>species</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.116</td>
<td>0.352</td>
<td>-1.214</td>
<td>-1.147</td>
<td>NA</td>
<td>ageneiosus brevifii</td>
</tr>
<tr>
<td>2</td>
<td>-0.083</td>
<td>-0.457</td>
<td>-1.881</td>
<td>-1.171</td>
<td>-1.485</td>
<td>ageneiosus brevifii</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>8</td>
<td>1.416</td>
<td>0.684</td>
<td>-1.439</td>
<td>-1.554</td>
<td>-0.874</td>
<td>cynodon gibbus</td>
</tr>
<tr>
<td>9</td>
<td>0.115</td>
<td>-0.509</td>
<td>-1.910</td>
<td>NA</td>
<td>-1.610</td>
<td>cynodon gibbus</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>67</td>
<td>1.813</td>
<td>1.953</td>
<td>-2.251</td>
<td>0.390</td>
<td>-0.651</td>
<td>doras micopoeus</td>
</tr>
</tbody>
</table>

- How to cluster the 10 species
Classical or interval data representation
Classical or interval data representation
Classical or interval data representation
Classical or interval data representation
Classical or interval data representation

LeporinusFrederici
LeporinusFasciatus
DorasMicropoeus
PlatydorasCostatus
PotamotrygonHystrix
HopliasAimara
PseudoancistrusBarb
SemaprochilodusVari
CynodonGibbus
AgeneiosusBrevifili
Classical or interval data representation
Classical or interval data representation
Classical or interval data representation
SODAS Software

- The data table obtained with DB2SO method of SODAS software

<table>
<thead>
<tr>
<th>species</th>
<th>liver</th>
<th>kidney</th>
<th>gills</th>
<th>intestine</th>
<th>stomach</th>
</tr>
</thead>
<tbody>
<tr>
<td>ageneiosus brevifii</td>
<td>[-0.80:0.34]</td>
<td>[-1.50:0.35]</td>
<td>[-1.88:-1.21]</td>
<td>[-1.45:-0.48]</td>
<td>[-1.49:-1.05]</td>
</tr>
<tr>
<td>Cynodon Gibbus</td>
<td>[0.12:1.59]</td>
<td>[-0.51:1.18]</td>
<td>[-1.91:-1.44]</td>
<td>[-1.75:-0.68]</td>
<td>[-1.61:0.22]</td>
</tr>
<tr>
<td>Hoplias Aimara</td>
<td>[-0.44:0.90]</td>
<td>[-0.17:1.60]</td>
<td>[-1.98:-1.53]</td>
<td>[-2.17:-0.71]</td>
<td>[-2.36:-0.93]</td>
</tr>
<tr>
<td>Potamitrigon Hystrix</td>
<td>[0.66:2.01]</td>
<td>[0.77:2.15]</td>
<td>NA</td>
<td>[-0.50:0.23]</td>
<td>[-0.80:0.69]</td>
</tr>
<tr>
<td>Leporinus Fasciatus</td>
<td>[-0.98:-0.58]</td>
<td>[-0.32:0.35]</td>
<td>[-3.00:-2.63]</td>
<td>NA</td>
<td>[-2.11:-2.76]</td>
</tr>
<tr>
<td>Leporinus Frederici</td>
<td>[-0.82:-0.04]</td>
<td>[-0.95:-0.19]</td>
<td>[-3.27:-2.55]</td>
<td>[-1.74:-1.42]</td>
<td>[-2.03:-0.55]</td>
</tr>
<tr>
<td>Doras Micropoeus</td>
<td>[1.34:2.12]</td>
<td>[1.47:2.69]</td>
<td>[-2.38:-2.21]</td>
<td>[-1.99:0.39]</td>
<td>[-1.45:-0.24]</td>
</tr>
<tr>
<td>Platidoras Costatus</td>
<td>[0.41:2.42]</td>
<td>[-0.02:2.75]</td>
<td>[-2.90:-1.27]</td>
<td>[-1.22:0.38]</td>
<td>[-1.41:-0.49]</td>
</tr>
<tr>
<td>Pseudoancistrus Barbatus</td>
<td>[1.26:2.84]</td>
<td>[-0.99:0.99]</td>
<td>NA</td>
<td>[-0.31:0.68]</td>
<td>[-0.71:0.12]</td>
</tr>
<tr>
<td>Semaprochilodus Vari</td>
<td>[2.70:3.96]</td>
<td>[1.11:1.91]</td>
<td>[-1.79:-1.40]</td>
<td>[-0.91:0.52]</td>
<td>[-0.74:0.22]</td>
</tr>
</tbody>
</table>

- Each of the $n = 10$ species $i$ is an hyper-rectangle of $\mathbb{R}^p$ (here $p = 5$) noted:

$$x_i = \prod_{j=1}^{p} [a_i^j, b_i^j]$$
PART 2

Comparing hyper-rectangles
Several approaches

- Simple Euclidean distance or more generally Minkowsky distance \(\Rightarrow\) lower and upper bound are used independantly

<table>
<thead>
<tr>
<th>species</th>
<th>Axis1</th>
<th>Axis2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgeneiosusBrevi®li</td>
<td>[-1,957:-0,175]</td>
<td>[0,306:1,819]</td>
</tr>
<tr>
<td>CynodonGibbus</td>
<td>[-1,656:0,354]</td>
<td>[0,302:0,302]</td>
</tr>
<tr>
<td>DorasMicropoeus</td>
<td>[-2,967:-0,433]</td>
<td>[-0,397:1,337]</td>
</tr>
<tr>
<td>HopliasAimara</td>
<td>[-0,095:1,085]</td>
<td>[-0,554:0,158]</td>
</tr>
<tr>
<td>LeporinusFasciatus</td>
<td>[-2,491:-2,197]</td>
<td>[-2,031:0,236]</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>species</th>
<th>Axis1</th>
<th>Axis2</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1,957</td>
<td>-0,175</td>
<td>0,306</td>
</tr>
<tr>
<td>-1,656</td>
<td>0,354</td>
<td>0,302</td>
</tr>
<tr>
<td>-2,967</td>
<td>-0,433</td>
<td>-0,397</td>
</tr>
<tr>
<td>-0,095</td>
<td>1,085</td>
<td>-0,554</td>
</tr>
<tr>
<td>-2,491</td>
<td>-2,197</td>
<td>-2,031</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Several approaches

- Simple Euclidean distance or more generally Minkowsky distance ⇒ lower and upper bound are used independantly

<table>
<thead>
<tr>
<th>species</th>
<th>Axis1</th>
<th>Axis2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgeneiosusBrevifili</td>
<td>[-1,957:-0,175]</td>
<td>[0,306:1,819]</td>
</tr>
<tr>
<td>CynodonGibbus</td>
<td>[-1,656:0,354]</td>
<td>[0,302:0,302]</td>
</tr>
<tr>
<td>DorasMicropoeus</td>
<td>[-2,967:-0,433]</td>
<td>[-0,397:1,337]</td>
</tr>
<tr>
<td>HopliasAimara</td>
<td>[-0,095:1,085]</td>
<td>[-0,554:0,158]</td>
</tr>
<tr>
<td>LeporinusFasciatus</td>
<td>[-2,491:-2,197]</td>
<td>[-2,031:0,236]</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- Elaborated distances taking into account both position and span of the intervals ⇒ Explicit formulas for the optimum class prototype ?

ASMDA05, Brest – p.11/26
The Hausdorff distance

The Hausdorff distance between two sets $A, B \subset \mathbb{R}^p$ is:

$$d_{H,\alpha}(A, B) = \max(h(A, B), h(B, A))$$

with

$$h(A, B) = \sup_{u \in A} \inf_{v \in B} d_\alpha(u, v)$$

$\Rightarrow$ Depends on the distance $d_\alpha (L_1, L_2...L_\infty)$ chosen to compare two points of $\mathbb{R}^p$
Mathematical properties

Here $A$ and $B$ are two hyper-rectangles of $\mathbb{R}^p$ noted:

$$A = \prod_{j=1}^{p} A_j, \quad B = \prod_{j=1}^{p} B_j$$

where $A_j = [a_j, b_j]$ and $B_j = [\alpha_j, \beta_j]$ are intervals of $\mathbb{R}$.

- **Property 1.** In the one dimensional space we can drop the subscript $\alpha$ and:

$$d_H(A_j, B_j) = \max(|a_j - \alpha_j|, |b_j - \beta_j|)$$
Mathematical properties

Here $A$ and $B$ are two hyper-rectangles of $\mathbb{R}^p$ noted:

$$A = \prod_{j=1}^{p} A_j, \quad B = \prod_{j=1}^{p} B_j$$

where $A_j = [a_j, b_j]$ and $B_j = [\alpha_j, \beta_j]$ are intervals of $\mathbb{R}$.

- **Property 1.** In the one dimensional space we can drop the subscript $\alpha$ and:

$$d_H(A_j, B_j) = \max(|a_j - \alpha_j|, |b_j - \beta_j|)$$

- **Property 2.** With the $L_\infty$ distance, we have the following relation between the Hausdorff distance $d_{H,\infty}$ in $p$ dimensions and $d_H$ in one dimension:

$$d_{H,\infty}(A, B) = \max_{j=1,\ldots,p} d_H(A_j, B_j)$$
Two distances between hyper-rectangles

We are able to give an explicit formula of the optimum class prototype with:

- The $L_\infty$ Hausdorff distance:
  \[
  d_{H,\infty}(A, B) = \max_{j=1, \ldots, p} \max(\{a_j - \alpha_j, b_j - \beta_j\})
  \]
  \[
  d_H(A_j, B_j)
  \]

  ⇒ In the particular case of intervals reduced to single points, the $L_\infty$ Hausdorff distance is the well-known $L_\infty$ distance between $\mathbb{R}^P$ points
Two distances between hyper-rectangles

We are able to give an explicit formula of the optimum class prototype with:

- The $L_\infty$ Hausdorff distance:
  \[ d_{H,\infty}(A, B) = \max_{j=1,\ldots,p} \max \left( |a_j - \alpha_j|, |b_j - \beta_j| \right) \]
  \[ d_H(A_j, B_j) \]
  \[ \Rightarrow \text{In the particular case of intervals reduced to single points, the } L_\infty \text{ Hausdorff distance is the well-known } L_\infty \text{ distance between } \mathbb{R}^p \text{ points} \]

- The Hausdorff-based distance:
  \[ d(A, B) = \sum_{j=1}^{p} \max \left( |a_j - \alpha_j|, |b_j - \beta_j| \right) \]
  \[ d_H(A_j, B_j) \]
  \[ \Rightarrow \text{In the particular case of intervals reduced to single points, this distance is the well-known } L_1 \text{ City-Block distance between } \mathbb{R}^p \text{ points} \]
Example

<table>
<thead>
<tr>
<th>num</th>
<th>species</th>
<th>Axis1</th>
<th>Axis2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AgeneiosusBrevifili</td>
<td>[1,957:-0,175]</td>
<td>[0,306:1,819]</td>
</tr>
<tr>
<td>2</td>
<td>CynodonGibbus</td>
<td>[-1,656:0,354]</td>
<td>[0,302:0,302]</td>
</tr>
<tr>
<td>3</td>
<td>DorasMicropoeus</td>
<td>[-2,967:-0,433]</td>
<td>[-0,397:1,337]</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Comparing species 1 and 2 in one dimension:

\[
d_H(x_1^1, x_2^1) = \max(|-1,957 + 1,656|, |0,175 - 0,354|) = 0,529
\]

\[
d_H(x_1^2, x_2^2) = 1,517
\]
Example

<table>
<thead>
<tr>
<th>num</th>
<th>species</th>
<th>Axis1</th>
<th>Axis2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AgeneiosusBreviﬁli</td>
<td>$[1,957:-0,175]$</td>
<td>$[0,306:1,819]$</td>
</tr>
<tr>
<td>2</td>
<td>CynodonGibbus</td>
<td>$[-1,656:0,354]$</td>
<td>$[0,302:0,302]$</td>
</tr>
<tr>
<td>3</td>
<td>DorasMicropoeus</td>
<td>$[-2,967:-0,433]$</td>
<td>$[-0,397:1,337]$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Comparing species 1 and 2 in one dimension:

\[ d_H(x_1^1, x_2^1) = \max(| -1,957 + 1,656|, | -0,175 - 0,354|) = 0, 529 \]
\[ d_H(x_1^2, x_2^2) = 1, 517 \]

Comparing species 1 and 2 in two dimensions:

- with the $L_\infty$ Hausdorff distance:
  \[ d_{H,\infty}(x_1, x_2) = \max(0, 529, 1, 517) \]

- with the Hausdorff-based distance:
  \[ d(x_1, x_2) = 0, 529 + 1, 517 \]
PART 3

Define a class prototype
Define a class prototype

Adequacy criterion between

A distance between hyper-rectangles

Find an explicit formula for the prototype which optimizes
Define a class prototype
Define a class prototype

- Adequacy criterion $f$ between $y$ and $C = \{x_1, \ldots, x_5\}$
- A distance between hyper-rectangles $y$ and $x_i$
Define a class prototype

- Adequacy criterion $f$ between $y$ and $C = \{x_1, \ldots, x_5\}$
- A distance between hyper-rectangles $y$ and $x_i$

$\Rightarrow$ Find an explicit formula for the prototype $y$ which optimizes $f$
Distance between hyper-rectangles: \( d(x_i, y) = \sum_{j=1}^{p} d_H(x_i^j, y^j) \)

\[ \Rightarrow \text{Not an Hausdorff distance between \( \mathbb{R}^p \)-sets} \]
Distance between hyper-rectangles: \[ d(x_i, y) = \sum_{j=1}^{p} d_H(x_i^j, y^j) \]

⇒ Not an Hausdorff distance between \( \mathbb{R}^p \)-sets

Adequacy criterion: \[ f(y) = \sum_{i \in C} d(x_i, y) \] ("The star")
Distance between hyper-rectangles: \[ d(x_i, y) = \sum_{j=1}^{p} d_H(x_i^j, y^j) \]

\[ \Rightarrow \text{Not an Hausdorff distance between } \mathbb{R}^p\text{-sets} \]

Adequacy criterion: \[ f(y) = \sum_{i \in C} d(x_i, y) \] (’The star’)

Explicit formula of the minimizer \( \hat{y} = \prod_{j=1}^{p} [\hat{\mu}^j - \hat{\lambda}^j, \hat{\mu}^j + \hat{\lambda}^j] \):

\[
\hat{\mu}^j = \text{median}\{m_i^j \mid i \in C\} \\
\hat{\lambda}^j = \text{median}\{l_i^j \mid i \in C\}
\]

with \( m_i^j \) and \( l_i^j \) the midpoints and the half-lengths of the intervals \( x_i^j \)
An example
An example
An example
An example
An example
An example
An example
An example
An example
An example
$L_\infty$ Hausdorff distance: $d_{H,\infty}(x_i, y) = \max_{j=1...p} d_H(x_i^j, y^j)$

⇒ "Real" Hausdorff distance between $\mathbb{R}^p$-sets
$L_\infty$ Hausdorff distance: $d_{H,\infty}(x_i, y) = \max_{j=1...p} d_H(x_i^j, y^j)$

$\Rightarrow$ "Real" Hausdorff distance between $\mathbb{R}^p$-sets

Adequacy criterion: $f(y) = \max_{i \in C} d_{H,\infty}(x_i, y)$ ('The radius')
**L∞ Hausdorff distance:** \( d_{H,\infty}(x_i, y) = \max_{j=1...p} d_H(x_i^j, y^j) \)

\( \Rightarrow \) "Real" Hausdorff distance between \( \mathbb{R}^p \)-sets

**Adequacy criterion:** \( f(y) = \max_{i \in C} d_{H,\infty}(x_i, y) \) ('The radius')

**Explicit formula of a minimizer** \( \hat{y} = \prod_{j=1}^{p} [\hat{\alpha}^j, \hat{\beta}^j] \):

\[
\hat{\alpha}^j = \frac{\max_{i \in C} a_i^j + \min_{i \in C} a_i^j}{2}
\]

\[
\hat{\beta}^j = \frac{\max_{i \in C} b_i^j + \min_{i \in C} b_i^j}{2}
\]

with \( a_i^j \) and \( b_i^j \) the lower and upper bounds of the intervals \( x_i^j \)
An example
An example
An example
An example
An example
An example
An example
An example
An example
PART 4

Normalization
Two “central” intervals $\hat{y}^j$:

- The “median” interval
- The “middle” interval
Measure of centrality and dispersion

Two “central” intervals $\hat{y}^j$:

- $\hat{\mu}^j$ to $\hat{\lambda}^j$ for The “median” interval
- $\hat{\alpha}^j$ to $\hat{\beta}^j$ for The “middle” interval

Two measures of dispersion $\sigma^j$ from a “central” interval:

The “star”:

$$\sigma^j = \sum_{i=1}^{n} \max(|a_i^j - \hat{\mu}^j + \hat{\lambda}^j|, |b_i^j - \hat{\mu}^j - \hat{\lambda}^j|)$$

The “radius”:

$$\sigma^j = \max_{i=1\ldots n} \max(|a_i^j - \hat{\alpha}^j|, |b_i^j - \hat{\beta}^j|)$$
Normalized distance or data table

Initial data table $(x_i^j)_{n \times p}$

\[ x_i^j = [a_i^j, b_i^j] \]

\[ d_H(x_i^j, x_i'^j) \]
Normalized distance or data table

Initial data table \((x^j_i)_{n \times p}\)

\[ x^j_i = [a^j_i, b^j_i] \]

\[ d_H(x^j_i, x^j_i') \]

\[ \Downarrow \]

Normalized distance

\[ \sum_{j=1}^{p} \frac{1}{\sigma^j} d_H(x^j_1, x^j_2) \]

\[ \max_{j=1 \ldots p} \frac{1}{\sigma^j} d_H(x^j_1, x^j_2) \]
Normalized distance or data table

Initial data table \((x_i^j)_{n \times p}\)

\[ x_i^j = [a_i^j, b_i^j] \]

\[ d_H(x_i^j, x_i'^{j}) \]

\[ \downarrow \]

Normalized data table \((z_i^j)_{n \times p}\)

\[ z_i^j = \left[ \frac{a_i^j}{\sigma^j}, \frac{b_i^j}{\sigma^j} \right] \]

\[ d_H(z_i^j, z_i'^{j}) \]

Normalized distance

\[ \sum_{j=1}^{p} \frac{1}{\sigma^j} d_H(x_1^j, x_2^j) \]

\[ \max_{j=1 \ldots p} \frac{1}{\sigma^j} d_H(x_1^j, x_2^j) \]
Normalized distance or data table

Initial data table \((x_i^j)^{n \times p}\)

\[ x_i^j = [a_i^j, b_i^j] \]

\[ d_H(x_i^j, x_i'^j) \]

\[ \Downarrow \]

Normalized distance

\[ \sum_{j=1}^{p} \frac{1}{\sigma^j} d_H(x_1^j, x_2^j) \]

\[ \max_{j=1 \ldots p} \frac{1}{\sigma^j} d_H(x_1^j, x_2^j) \]

Normalized data table \((z_i^j)^{n \times p}\)

\[ z_i^j = [\frac{a_i^j}{\sigma^j}, \frac{b_i^j}{\sigma^j}] \]

\[ d_H(z_i^j, z_i'^j) \]

\[ \Downarrow \]

Initial distance

\[ \sum_{j=1}^{p} d_H(z_1^j, z_2^j) \]

\[ \max_{j=1 \ldots p} d_H(z_1^j, z_2^j) \]
Normalized distance or data table

Initial data table \((x_i^j)_{n \times p}\)

\[
x_i^j = [a_i^j, b_i^j]
\]

\[
\frac{1}{\sigma_j} \; d_H(x_i^j, x_i^j')
\]

\[\Downarrow\]

Normalized distance

\[
\sum_{j=1}^{p} \frac{1}{\sigma_j} d_H(x_1^j, x_2^j)
\]

\[
\max_{j=1 \ldots p} \frac{1}{\sigma_j} d_H(x_1^j, x_2^j)
\]

Normalized data table \((z_i^j)_{n \times p}\)

\[
z_i^j = [a_i^j, b_i^j]
\]

\[
d_H(z_i^j, z_i^j')
\]

\[\Downarrow\]

Initial distance

\[
\sum_{j=1}^{p} d_H(z_1^j, z_2^j)
\]

\[
\max_{j=1 \ldots p} d_H(z_1^j, z_2^j)
\]
Normalized distance or data table

Initial data table \((x_i^j)_{n \times p}\)

\[
x_i^j = [a_i^j, b_i^j]
\]

\[
\frac{1}{\sigma^j} d_H(x_i^j, x_i^j') = \downarrow
\]

Normalized distance

\[
\sum_{j=1}^{p} \frac{1}{\sigma^j} d_H(x_1^j, x_2^j) = \sum_{j=1}^{p} d_H(z_1^j, z_2^j)
\]

\[
\max_{j=1 \ldots p} \frac{1}{\sigma^j} d_H(x_1^j, x_2^j) = \max_{j=1 \ldots p} d_H(z_1^j, z_2^j)
\]

Normalized data table \((z_i^j)_{n \times p}\)

\[
z_i^j = [\frac{a_i^j}{\sigma^j}, \frac{b_i^j}{\sigma^j}]
\]

\[
\downarrow
\]

Initial data table \((z_i^j)_{n \times p}\)

\[
\]

\[
\]
Conclusion

- General approach for the “normalization” of k-means (dynamical clustering) algorithms

- Two normalized k-means methods for hyper-rectangles clustering:

<table>
<thead>
<tr>
<th>Prototype</th>
<th>Distance</th>
<th>Measure of dispersion</th>
</tr>
</thead>
<tbody>
<tr>
<td>The “median” hyper-rectangle</td>
<td>Hausdorff-based</td>
<td>The “star” from the “median” hyper-rectangle</td>
</tr>
<tr>
<td>The “middle” hyper-rectangle</td>
<td>$L_{\infty}$-Hausdorff</td>
<td>The “radius” deviation from the “middle” hyper-rectangle</td>
</tr>
</tbody>
</table>

- Explicit formula for prototypes with a real $L_1$ or $L_2$ Hausdorff distance between hyper-rectangles?
Normalized k-means clustering of hyper-rectangles

M. Chavent

Laboratoire de Mathématiques Appliquées de Bordeaux, UMR CNRS 5466
Universités Bordeaux1 et 2, France

chavent@math.u-bordeaux1.fr