`cluster.Sim` (clusterSim)

Paths characteristics in determination of optimal clustering procedure for a data set

<table>
<tr><th rowspan="2">No.</th><th rowspan="2">Steps in a typical cluster analysis</th><th colspan="9">Path’s number</th></tr>
<tr><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th><th>8</th><th>9</th></tr>
<tr><td>I</td><td>Selection of objects and variables</td><td colspan="9">data matrix $[x_{ij}]$</td></tr>
<tr><td rowspan="3">II</td><td>Measurement scale of variables</td><td>ratio</td><td>ratio</td><td>interval or mixed[1]</td><td>ordinal[2]</td><td>multi-state nominal[3]</td><td>binary</td><td>ratio</td><td>interval or mixed[1]</td><td>ratio</td><td>interval or mixed[1]</td></tr>
<tr><td>Selection of normalization formula[4]</td><td><code>n6</code> – <code>n11</code></td><td><code>n1</code> – <code>n5</code></td><td><code>n1</code> – <code>n5</code></td><td>N.A.</td><td colspan="2">N.A.</td><td colspan="2">without normalization</td><td><code>n6-n11</code> / <code>n1-n5</code></td><td><code>n1-n5</code></td></tr>
<tr><td>Transformed measurement scale of variables</td><td>ratio</td><td>interval</td><td>interval</td><td>ordinal</td><td>multi-state nominal</td><td>binary</td><td>ratio</td><td>interval or mixed[1]</td><td>ratio / interval</td><td>interval</td></tr>
<tr><td>III</td><td>Selection of distance measure[5]</td><td><code>d1</code> – <code>d7</code></td><td><code>d1</code> – <code>d5</code></td><td><code>d1</code> – <code>d5</code></td><td><code>d8</code></td><td><code>d9</code></td><td><code>b1</code> – <code>b10</code></td><td><code>d1</code> – <code>d7</code></td><td><code>d1</code> – <code>d5</code></td><td colspan="2">N.A.</td></tr>
<tr><td>IV</td><td>Selection of clustering method[6]</td><td colspan="7"><code>m1</code> – <code>m8</code></td><td colspan="2"><code>m9</code></td></tr>
<tr><td rowspan="3">V</td><td>Maximal number of possible variants</td><td>[(6 x 7 x 5)+ (6 x 1 x 3)] + [(5 x 5 x 5) +(5 x 1 x 3)] = 368</td><td>(5 x 5 x 5) + (5 x 1 x 3) = 140</td><td>1 x 5 = 5</td><td>1 x 5 = 5</td><td>10 x 5 = 50</td><td>(7 x 5) + (1 x 3) = 38</td><td>(5 x 5) + (1 x 3) = 28</td><td>11</td><td>5</td></tr>
<tr><td>Number of all classifications</td><td colspan="9">$\mathtt{LK} = (\mathtt{maxClusterNo} - \mathtt{minClusterNo} + 1) \cdot \mathtt{LW}_p$, where
<code>minClusterNo</code> minimal number of clusters,
<code>maxClusterNo</code> maximal number of clusters,
$\mathtt{LW}_p$ – number of variants for p-th path.</td></tr>
<tr><td>Internal cluster quality index</td><td colspan="2">1. Calinski & Harabasz (<code>G1</code>)
2. Baker & Hubert (<code>G2</code>)
3. Hubert & Levine (<code>G3</code>)
4. Silhouette (<code>S</code>)
5. Krzanowski & Lai (<code>KL</code>)</td><td colspan="3">1. N.A.
2. <code>G2</code>
3. <code>G3</code>
4. <code>S</code>
5. N.A.</td><td colspan="2">1. <code>G1</code>
2. <code>G2</code>
3. <code>G3</code>
4. <code>S</code>
5. <code>KL</code></td><td colspan="2">1. <code>G1</code>
2. N.A.
3. N.A.
4. N.A.
5. <code>KL</code></td></tr>
</table>

[1] Ratio & interval.

[2] We can use ratio, interval or mixed data (ratio, interval, ordinal), however these data are treated as ordinal because in the construction of the GDM2 distance measure only such relations as: “equal to”, “higher than”, “lower than” are taken into account.

[3] We can use ratio, interval, ordinal or mixed data (ratio, interval, ordinal, nominal), however these data are treated as nominal because in the construction of the Sokal & Michener distance measure only such relations as: “equal to”, “not equal to” are taken into account.

[4] `n1` – (x-mean)/sd, `n2` – (x-Me)/MAD, `n3` – (x-mean)/range, `n4` – (x-min)/range, `n5` – (x-mean)/max[abs(x-mean)], `n6` – (x/sd), `n7` – (x/range), `n8` – (x/max), `n9` – (x/mean), `n10` – (x/sum), `n11` – x/sqrt(SSQ).

[5] `d1` – Manhattan, `d2` – Euclidean, `d3` – Chebychev (max), `d4` – squared Euclidean, `d5` – GDM1, `d6` – Canberra, `d7` – Bray-Curtis; `d8` – GDM2, `d9` – Sokal & Michener; `b1` – `b10` (available in R `dist.binary` procedure): `b1` = Jaccard; `b2` = Sokal & Michener; `b3` = Sokal & Sneath (1); `b4` = Rogers & Tanimoto; `b5` = Czekanowski; `b6` = Gower & Legendre (1); `b7` = Ochiai; `b8` = Sokal & Sneath (2); `b9` = Phi of Pearson; `b10` = Gower & Legendre (2).

[6] `m1` – single link, `m2` – complete link, `m3` – average link, `m4` – McQuitty, `m5` – *k*-medoids (PAM), `m6` – Ward, `m7` – centroid, `m8` – median, `m9` – *k*-means. For clustering methods `m6` – `m8` squared Euclidean distance is used only.

N.A. – Not Applicable.

Source: Walesiak, M., Dudek, A. (2006), *Symulacyjna optymalizacja wyboru procedury klasyfikacyjnej dla danego typu danych – oprogramowanie komputerowe i wyniki badan*, Prace Naukowe AE we Wroclawiu no. 1126, 120-129.