index.KL(clusterSim)

Krzanowski and Lai index

$$KL(u) = \left|\frac{DIFF_u}{DIFF_{u+1}}\right|,$$
$$DIFF_u = (u-1)^{2/m} tr\boldsymbol{W}_{u-1} - u^{2/m} tr\boldsymbol{W}_u,$$

where: $\boldsymbol{X} = \{x_{ij}\}$, $i = 1, \dots, n$; $j = 1, \dots, m$ – data matrix,

n – number of objects,

m – number of variables,

$\boldsymbol{W}_u = \sum_r \sum_{i \in C_r} (\boldsymbol{x}_{ri} - \bar{\boldsymbol{x}}_r)(\boldsymbol{x}_{ri} - \bar{\boldsymbol{x}}_r)^T$ – within-group dispersion matrix for data clustered into u clusters,

$\boldsymbol{x}_{ri}$ – m-dimensional vector of observations of the i-th object in cluster r,

$\bar{\boldsymbol{x}}_r$ – centroid or medoid of cluster r,

$r = 1, \dots, u$ – cluster number,

u – number of clusters ($u = 2, \dots, n-2$),

C_r – the indices of objects in cluster r.

The value of u, which maximizes $KL(u)$, is regarded as specifying the number of clusters.

References

Krzanowski , W.J., Lai, Y.T. (1988), *A criterion for determining the number of groups in a data set using sum of squares clustering*, “Biometrics”, 44, no. 1, 23-34.

Tibshirani R., Walther G., Hastie T. (2001), *Estimating the number of clusters in a data set via the gap statistic*, „Journal of the Royal Statistical Society”, ser. B, vol. 63, part 2, 411-423.