speccl (clusterSim)

A spectral clustering algorithm

Ng, Jordan and Weiss [2002]; Walesiak [2011; 2012]

1. Form the data matrix $\mathbf{X}_{n\times m}$ ($i,k=1,\ldots,n$ – the number of object, $j=1,\ldots,m$ – the number of variable).

2. Form the affinity matrix $\mathbf{A}=[A_{ik}]$, where $A_{ii}=0$ and $A_{ik}=\exp(-\sigma\cdot d_{ik})$, where: σ – scale parameter (see algorithm below), d_{ik} – distance measure ("sEuclidean" – squared Euclidean distance, "euclidean" – Euclidean distance, "manhattan" – city block distance, "maximum" – Chebyshev distance, "canberra" – Lance & Williams Canberra distance, "BC" – Bray-Curtis distance measure for ratio data, "GDM1" – GDM distance for metric data, "GDM2" – GDM distance for ordinal data, "SM" – Sokal-Michener distance measure for nominal variables).

3. Construct the matrix $\mathbf{L}=\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ ($\mathbf{D}$ – diagonal matrix whose (*i*, *i*)-element is the sum of *i*-th row of matrix $\mathbf{A}=[A_{ik}]$.

4. Find the u (u – number of clusters) largest eigenvectors of $\mathbf{L}$. Form the new data matrix $\mathbf{E}=[e_{ij}]_{n\times u}$ by stacking the eigenvectors in columns.

5. Normalization step: $y_{ij}=e_{ij}\Big/\sqrt{\sum_{j=1}^{u}e_{ij}^{\,2}}$ ($i=1,\ldots,n$ – the number of object, $j=1,\ldots,u$ – the number of variable, u – number of clusters). Each row of matrix $\mathbf{Y}=[y_{ij}]_{n\times u}$ has unit length.

6. Cluster objects of matrix $\mathbf{Y}$ into *u* clusters using *k*-means method.

Algorithm for searching optimal value of σ parameter

Walesiak and Dudek [2009]

Bootstrapping sample $\mathbf{X}'$ is chosen from data matrix $\mathbf{X}$ (containing n' objects, where $\frac{1}{2}n\le n'\le\frac{3}{4}n$).

Step 0. σ parameter belongs to interval S_0: sigma.interval="default" – from zero to square root of sum of all distances in lower triangle of distance matrix for "sEuclidean" and from zero to sum of all distances in lower triangle of distance matrix for other distances; sigma.interval=1000 – from zero to value given by researcher, e.g. 1000.

Step 1. The interval S_k (k – iteration number; at the beginning $S_k=S_0$) is divided into intervals of equal length: $p_r^k=[\underline{p_r^k};\overline{p_r^k}]$, $r=1,\ldots,R$ (R – the number of intervals in each iteration: default $R=10$).

Step 2. For each interval p_r^k we calculate its centre: $\sigma_r^k=\frac{\underline{p_r^k}+\overline{p_r^k}}{2}$. Spectral clustering of data set $\mathbf{X}'$ is performed on a fixed number of clusters *u* for all values σ_r^k.

Step 3. Chosen is such value of σ_r^k for which within-cluster sum of squared distances is minimal.

Step 4. With selected interval go to step 1 and continue the procedure until the default number of iterations is reached (default: three iterations).

References

Karatzoglou, A. (2006), *Kernel methods. Software, algorithms and applications*, Dissertation, Wien, Technical University.

Ng, A., Jordan, M., Weiss, Y. (2002), *On spectral clustering: analysis and an algorithm*, In: T. Di-

etterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14. MIT Press, 849-856.

Walesiak, M. (2011), *Uogólniona miara odległości w statystycznej analizie wielowymiarowej z wykorzystaniem programu R* [*The Generalized Distance Measure in multivariate statistical analysis with R]*, Wydawnictwo UE, Wroclaw.

Walesiak, M. (2012), *Klasyfikacja spektralna a skale pomiaru zmiennych [Spectral clustering and measurement scales of variables]*, "Przegląd Statystyczny" (Statistical Review) no. 1, 13-31.

Walesiak, M., Dudek, A. (2009), *Odległość GDM dla danych porządkowych a klasyfikacja spektralna [GDM distance for ordinal data with spectral clustering]*, Prace Naukowe UE we Wrocławiu no. 84, 9-19.