speccl (clusterSim)

A spectral clustering algorithm

Ng, Jordan and Weiss [2002]; Walesiak [2010; 2011]

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 – *kernel width* (see algorithm below), d_{ik} – distance (e.g. squared Euclidean distance, GDM1 distance for metric data, GDM2 distance for ordinal data, Sokal-Michener distance measure for nominal variables, Bray-Curtis distance measure for ratio data or others distances included in functions `dist`, `dist.GDM` and `dist.binary`).

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 $S_0=[0;D]$ (D – sum of all distances d_{ik} in distance matrix).

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 sum of within-clusters 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. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14. MIT Press, 849-856.

Walesiak, M. (2006), *Uogólniona miara odległości w statystycznej analizie wielowymiarowej* [*The Generalized Distance Measure in multivariate statistical analysis]*, Wydawnictwo AE, Wro-

claw.
Walesiak, M. (2010), *Klasyfikacja spektralna z wykorzystaniem odległości GDM*, In: K. Jajuga, M. Walesiak (Eds.), *Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia 17, Prace Naukowe UE we Wrocławiu no. 107, 161-171.
Walesiak, M. (2011), *Klasyfikacja spektralna a skale pomiaru zmiennych [Spectral clustering and measurement scales of variables]*, Prace Naukowe UE we Wrocławiu (in preparation).
Walesiak, M., Dudek, A. (2009), *Odległość GDM dla danych porządkowych a klasyfikacja spektralna*, Prace Naukowe UE we Wrocławiu no. 84, 9-19.