replication.SDA(symbolicDA)

Modification of replication analysis for cluster validation of symbolic data

Formally, replicating clusters proceeds as follows (see Breckenridge [2000], 262-263; Milligan [1996], 368-369; Gordon [1999], 184):

1. The data are randomly divided into two samples: sample *A* (primary set) and sample *B* (replication set). Data must exist on the same set of variables in both samples.

2. Sample *A* is clustered into *u* clusters $A=\{A_1,...,A_u\}$ via optimization method: `SClust`, `DClust`, `pam` or hierarchical method: `single`, `complete`, `average`, `mcquitty`, `median`, `centroid`, `Ward`, `diana` (see e.g. Bock, Diday [2000]; Diday, Noirhomme-Fraiture [2008]). Once clusters have been identified, the medoids of *u* clusters are selected.

3. Sample *B* is clustered, using the same clustering procedure, into *u* clusters $B=\{B_1,...,B_u\}$.

4. The distances between sample *B* objects and the medoids of sample *A* are determined (see e.g. Bock, Diday [2000]; Diday, Noirhomme-Fraiture [2008]). Each element of sample *B* is assigned to the nearest medoid determined in step 2 (this provides a partition of sample *B* into no more than *u* clusters). This produces a clustering of sample *B* based on characteristics of sample *A*: $B^*=\{B_1^*,...,B_u^*\}$.

5. The steps 1-4 are repeated *S* times (*S* – the number of simulations).

6. A measure of agreement (average value of adjusted Rand index) between two clusterings *B* and **B* are computed for *S* simulations. The level of agreement between the two partitions reflects the stability of the clustering in the data. Adjusted Rand index takes values of $(-\infty,1]$ and its value close to 1 means relatively stable division of objects set into *u* clusters.

References

Breckenridge, J.N. (2000), *Validating cluster analysis: consistent replication and symmetry*, "Multivariate Behavioral Research", 35 (2), pp. 261-285.

Gordon, A.D. (1999), *Classification*, Chapman and Hall/CRC, London.

Hubert, L., Arabie, P. (1985), *Comparing Partitions*, "Journal of Classification", no. 1, pp. 193-218.

Milligan, G.W. (1996), *Clustering Validation: Results and Implications for Applied Analyses*, In P. Arabie, L.J. Hubert, G. de Soete (Eds.), \emph{Clustering and classification}, World Scientific, Singapore, pp. 341-375.

Bock H.H., Diday E. (Eds.) (2000), *Analysis of Symbolic Data. Explanatory methods for extracting statistical information from complex data*, Springer-Verlag, Berlin.

Diday E., Noirhomme-Fraiture M. (Eds.) (2008), *Symbolic Data Analysis with SODAS Software*, John Wiley & Sons, Chichester.