Robustness concepts for general clustering methods

C. Hennig[1]

[1] Department of Statistical Science, UCL, Gower St., London, WC1E 6BT

Keywords: dissolution point, isolation robustness, trimmed k-means, single linkage, complete linkage, average silhouette width, mixture model, BIC.

A unified theory is presented to assess the robustness of general clustering methods (GCM), i.e., methods mapping a data set on a collection of its subsets. The theory takes into account that robustness and stability in cluster analysis are not only data dependent, but even cluster dependent. That is, robustness is a function of the GCM and every single cluster of a data set.

The main principles are

- worst case assessment of the stability of a cluster under addition of g points,
- comparison of an original cluster with the most similar cluster in the induced clustering under addition of points by means of the Jaccard (1901) similarity between sets,
- the dissolution point, which is an adaptation of the breakdown point concept,
- isolation robustness: given a GCM, is it possible to dissolve, by addition of g points, an arbitrarily well separated cluster?

Results on hierarchical methods, k-means, medoids, trimmed k-means (Cuesta-Albertos et al., 1997), mixture models (with and without BIC-estimated number of components, noise component; Fraley and Raftery, 1998) and the average silhouette width (Kaufman and Rousseeuw, 1990) are given. The results indicate that the estimation of the number of clusters is essential for isolation robustness. Trimming of points can improve the dissolution robustness for methods with fixed k. This confirms the findings of Hennig (2004).

References

J.A. Cuesta-Albertos,A. Gordaliza, C. Matran (1997). Trimmed k-means: An Attempt to Robustify Quantizers. *Annals of Statistics*, 25, 553–576.

C. Fraley, A.E. Raftery (1998). How Many Clusters? Which Clustering Method? Answers Via Model Based Cluster Analysis. *Computer Journal*, 41, 578–588.

C. Hennig (2004). Breakdown points for ML estimators of location-scale mixtures. *Annals of Statistics*, 32, 1313–1340.

P. Jaccard (1901). Distribution de la flore alpine dans la Bassin de Dranses et dans quelques regions voisines. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37, 241–272.

L. Kaufman, P.J. Rousseeuw (1990). *Finding Groups in Data*. Wiley, New York.