

Cluster analysis with k-means: what about the details ?

Maurice Roux
Université Paul Cézanne
Marseille, France

Background:

When using the k-means procedure (and its variants) there are several parameters to select beforehand, the main one being the number of clusters. The usual strategy to determine this number is to repeat the whole procedure with various cluster numbers and to select the one which leads to the best fit between the resulting partition and the initial data.

To evaluate this fit a number of indexes (internal criteria) have been proposed in the literature. In addition, for a fixed number of clusters it is recommended to restart "many" times the overall computations with new random initializations.

The present paper, based on both artificial and real life data, wants to help for the choice of a goodness-of-fit index and put forward some guidelines for the number of restarts.

Main results :

Three indexes do give consistent appreciations, namely Dunn's index, Kendall's tau and the contingency Khi-square based on the quadruples (pairs of pairs of objects). As for the second target parameter, it appears that the number of restarts is not a key parameter, since the "best" results are quickly reached after, say, a few tens of repeated random initial partitions.

Incidentally, after a multiple restart k-means it is very useful to run a correspondence analysis program applied to a consensus matrix over the objects. Such an analysis clearly detects those objects not included in any cluster which may be tagged as "unclassifiable". More over it confirms or invalidates the number of clusters.

When there exists a known partition of the data it may be tempting to use it as a reference to evaluate indexes and clustering methods. But an example in gene expression data shows this approach is questionable.

Conclusion:

The k-means clustering process is a very useful method, able to deal with very big data sets. It is even more efficient when a good quality index is used to establish the number of clusters. The present work is not really a benchmark but it emphasizes the difficulty of finding groups in real life data sets. The use of correspondence analysis with a consensus matrix greatly helps to discover "unclassifiable" observations which often confuse the clustering results.