Finding the number of clusters in a data set: Nb.clusters package

Malika Charrad\textsuperscript{1,}\textsuperscript{*}, Nadia Ghazzali\textsuperscript{1}, Veronique Boiteau\textsuperscript{1}, Azam Niknafs\textsuperscript{1}

\textsuperscript{1}. Universit Laval
Department of Mathematics and Statistics
\textsuperscript{*}Contact author: malika.charrad.1@ulaval.ca

Keywords: Number of clusters, Validity Indices, Cluster validity, Kmeans, Hierarchical clustering.

Clustering is the partitioning of a set of objects into groups (clusters) so that objects within a group are more similar to each others than objects in different groups.

Most of the clustering algorithms depend on certain assumptions in order to define the subgroups present in a data set. As a consequence, in most applications the resulting clustering scheme requires some sort of evaluation as regards its validity. In general terms, there are three approaches to investigate cluster validity. The first is based on external criteria, which consist in comparing the results of cluster analysis to externally known results, such as externally provided class labels. The second approach is based on internal criteria which use the information obtained from within the clustering process to evaluate how well the results of cluster analysis fit the data without reference to external information. The third approach of clustering validity is based on relative criteria. Here the basic idea is the evaluation of a clustering structure by comparing it with other clustering schemes, resulting by the same algorithm but with different parameters values, e.g. the number of clusters.

In the literature, a wide variety of indices have been proposed to find the optimal number of clusters in a partitioning of a data set during the clustering process. Although a vast number of references exist, few comparative studies have been performed on these indices [5]. Moreover, for most of indices proposed in the literature, programs are unavailable to test these indices and compare them.

The \texttt{R} package, \texttt{Nb.clusters}, has been developped specifically for that purpose. It implements 30 indices for cluster validation ready to apply on outputs produced by clustering algorithms, Hierarchical clustering and Kmeans, coming from the same package. Most of these indices are described in Milligan and Cooper study [5]. The \texttt{Nb.clusters} function allows to apply one or 30 indices simultaneously and proposes to user the best clustering scheme from the different results obtained by varying all combinations of number of clusters, distance measures ("euclidean", "maximum", "manhattan", "canberra", "binary", "minkowski"), and clustering methods ("ward", "single", "complete", "average", "mcquitty", "median", "centroid").

References


