Determining the number of clusters in a data set via repeated data clustering in two clusters

Jerzy Korzeniewski
University of Lodz, Lodz, Poland jurkor@wp.pl

In the paper a new algorithm of determining the number of clusters in a data set of objects described with continuous variables is presented. The idea consists in repeated division of data set (or the clusters resulting from the previous step) into two clusters. We accept the division (raising the number of clusters by one) if a division quality measure exceeds the threshold and we forget it (reverting to the data set structure from before the division) if the measure falls below the threshold. One can apply different division quality measures but a new one based on the Rand index is proposed. The performance of the new index is compared with that of the leading indices constructed so far i.e. Calinski-Harabasz index and the gap index.

Keywords: cluster analysis, number of clusters, Rand index, gap statistic, Calinski-Harabasz index.