K-MEANS CLUSTERING – NEW VARIATIONS

Andrzej Sokołowski[1], Małgorzata Markowska[1] and Maciej Laburda[2]

[1] Krakow University of Economics, Poland
(e-mail: sokolows@uek.krakow.pl)

[2] Wroclaw University of Economics and Business, Poland,

(e-mail: malgorzata.markowska@ue.wroc.pl)

3 Krakow University of Economics, Poland

(e-mail: s223060@student.uek.krakow.pl)

ABSTRACT: k-means is one of the most popular methods in cluster analysis. It can handle the large set of data since there is no need to store the distance matrix in the memory, and the algorithm converges very quickly to the situation when no object should be relocated (each one is closer to the mean of its "own" cluster, that to the other one). Two main drawbacks of the method are that the number of clusters should be defined properly and that the final partition tends to be formed by spherical clusters. In the literature, there are many variations, improvements, and new versions of k-means based on the original model.
In this contribution we discuss two new ideas. The first one can be called *n%-neighbors k-means*. When we have to decide whether an object should be relocated to another cluster, we consider only some percentage of the total set of objects, only points closest to the one which is considering at the moment. So partial means should be calculated and considered. It is possible that some distance clusters will not be taken into account if their members are not included in n% nearest neighbors of this point. The second new proposition can be called *local standardization k-means.* Standarization is performed separately for each cluster, using its mean and standard deviation, excluding point which is considered for relocation. Than this point is "standardized" using means and standard deviations of consecutive clusters and distances are calculated.
Simulation analysis is the main tool to evaluate the quality of the proposed approaches.

KEYWORDS: k-means, nearest neighbors, standardization.

References

MACQUEEN, J. 1967. Some methods for classification and analysis of multivariate observations. In: *Fifth Berkeley Symposium on Mathematics. Statistics and Probability.* University of California Press, 281-297

BOCK, H.-H. 2008. Origins and extentions of the k-means algorithm in cluster analysis. *Electronic Journal for History of Probability and Statistics*, **4** (2), 1-18

JAIN, A.K. 2009. Data Clustering: 50 years beyond K-means. *Pattern Recognition Letters*, **31**(8), 651-666