Clustering methods and their modification dedicated to exploration of experimental data

PhD student: MSc Klaudia Drab
Doctoral supervisor: Dr. Sc. Michał Daszykowski

Advanced technology allows chemists analysis of different types of samples. Instrumental methods are well suited for the compositional analysis of drugs, biological fluids, food samples, cosmetics, etc. However, with the increasing number of samples and parameters being analyzed, we receive multidimensional data, where useful chemical information can be hidden. Therefore, in order to analysis this kind of data, the mathematical tools are needed. The set of mathematical and statistical methods used in data exploration, data analysis, and modelling are part of relatively new science - chemometrics. Especially interesting are exploration methods, such as Cluster Analysis. They uncover groups of samples by means of clustering with respect to their specific chemical features. The similarity of samples are evaluated using distance between samples in the space of parameters. Samples which are close in this space are considerably similar. Distance between samples (in the space of measured parameters) is calculated by distance/similarity measures. The most common distance measures are Euclidean distance, Mahalanobis distance and the Pearson correlation coefficient.

The increasing complexity of data requires modification of basic clustering algorithms and distance measures. Therefore, the first aim of this thesis was to modify of the DBSCAN algorithm. The algorithm becomes unstable when detecting border objects of adjacent clusters. In order to solve this issue new version of algorithm was introduced. In revised DBSCAN the way of processing objects was changed. Firstly, groups contain only core objects were formed. After the detection of all clusters, border objects are assigned to their closest group. The distances between border objects and groups of similar core objects are calculated based on Euclidean distance. The second aim of this thesis was to develop a new similarity measure (s_{ij}) for comparing two-way chromatographic fingerprints. New framework is alternative to the alignment methods. It is robust against peak shifts and additionally allows dealing with peak coelution issue. The last part of this PhD thesis considered uncertainty associated with experimental data. So far, error associated with data was not taken into account during the step of exploration or data analysis. Recently, there is a new trend in developing of new algorithms for clustering data in presence of errors. Using the conception presented in [1], the uncertainty associated with every single object was introduced to DBSCAN algorithm. This approach greatly improves effectiveness of the DBSCAN algorithm.

[1] M. Kumar, N. R. Patel, Clustering data with measurement errors, Comput. Stat. Data Anal. 51 (2007) 6084–6101.