Abstract

Thesis title: *Adaptive hierarchical clustering algorithms based on data aggregation methods*
Author: Anna Cena
Supervisor: Marek Gągolewski

Cluster analysis aims at determining an input data set's partition in such a way that the observations within each group are as similar (with respect to a given criterion) as possible to each other, while diversifying those from different groups. In other words, we would like to identify a hidden *structure* of the given data set. Cluster analysis methods are widely applied in computational biology, social sciences, image recognition, signal processing, and so forth. It is worth noting that, since no information on the correct grouping is given a priori, clustering algorithms often rely on different pairwise "similarity" or "distance" measures, neighborhood relations, or local density estimates.

Hierarchical methods, which are among the most often used clustering approaches, construct a family of nested partitions such that each of its members is a well-defined grouping itself. Therefore, one may not only cut the resulting hierarchy at the desired level, but also gain insight into the whole grouping process.

Moreover, most hierarchical methods do not require any sophisticated assumptions on the input data space: typically only a pairwise distance function is needed. Such a similarity measure may already reflect the results of feature engineering/weighting/selection at the data pre-processing stage. However, one must specify a particular extension of the chosen pairwise distance onto the set of the whole point groups, i.e., the so-called linkage method. Different linkages allow for capturing various aspects of the cluster merge process, for example, some can be more sensitive to the local points' density, whereas other may focus more on the relations between all the inputs. In this thesis we thoroughly investigated the linkage schemes proposed by Yager in 2000 and then re-introduced by Nasıbov and Kandemır-Cavas in 2011, which are generated by the weighted ordered averages – the OWA operators, generalizing the single, complete, and average linkages. We assessed their influence on the resulting partitions and studied their performance on a broad array of benchmark datasets. We also extended this setting to a three-phase aggregation process. This brought us to the conclusion that local density estimates tend to perform better than the other approaches.

Unfortunately, many hierarchical methods are computationally demanding. The single linkage (minimum-based) scheme is one of the few notable exceptions, as it may be constructed based on the minimum spanning tree of the underlying weighted pairwise distance graph. However, the original algorithm is known to be very sensitive to outliers. Also, it tends to output highly unbalanced partitions. The Genie algorithm, which we proposed recently, aims at reducing its drawbacks while still allowing for its effective implementation. This method introduces a correction for the evenness of cluster sizes' distribution: if inequity of cluster sizes raises above a given threshold g, then a forced merge of some of the smallest clusters is conveyed. We have assessed the quality of the resulting partitions on a wide set of benchmark data. What is more, we investigated the effects of applying the Genie-like correction on other OWA-based linkage schemes. It turns out that such a modification of the cluster merge process significantly improves

the overall quality of the tested algorithms. Since the Genie algorithm depends on the choice of an inequity measure and its threshold g, the effects of their choice on the grouping process was thoroughly studied. Based on the obtained results, we were able to recommend some practically useful default settings of the parameters.

In this thesis, we have also proposed a new hierarchical agglomerative method which acts on a minimum spanning tree and is based on the partial information on the data structure obtained by applying the Genie algorithm. The initial, partial grouping is determined in an adaptive manner by computing the intersection of the partitions generated by the Genie method with a wide range of the inequity measure's threshold. Moreover, each element of the nested family of partitions is generated according to the information criterion or by minimizing a certain nearest neighbor-based dissimilarity with specific constraints. The proposed method was compared with other state-of-the-art algorithms, like other hierarchical methods, the K-means, HDBSCAN*, ITM, spectral clustering, Birch, among others. The new method outperformed the listed ones. Moreover, it is worth noting that it does not require a user to specify any underlying parameters, which makes it easy to use in practice.