Clustering and Partitioning Methods: Applications, Problems and the Future

Petr Chmelař, Ivana Rudolfová, Lukáš Stryka
Department of Information Systems
Faculty of Information Technology
Brno University of Technology

The paper deals with applications of various clustering methods on many kind of data. Clustering is a non-supervised technique that tries to maximize the intraclass and minimize the interclass similarity. The task is in a pretty pickle, because there are no user interactions expected and all the intelligence is left upon the computer there (and computers are stupid). The only possibility how to influence the result is to guess some parameters in advance and then wait (a long) time to have some results that usually doesn't satisfy the user much. We present some practical applications of existing methods. The dataset we have tested is quite complex, it includes both highly structured and unstructured data from business, computer vision and viruses to human DNA samples. As each domain is different, the needs and requirements are contradictory. Thus, we have tested many methods. These include classical partitioning methods (k-Means, k-Medoids), hierarchical (BIRCH), density based (DBSCAN) and model-based approaches (GMM and ART). Although we have developed our own method (Voronoi Clustering), suitable for a huge number of samples and clusters, it still has disadvantages similar to the other methods. We think, the global goal in the clustering is to establish a semi-supervised method that will (once) process the data and present its preliminary results to the user in an interactive way, similarly to OLAP. The proposed analysis will be able to present hierarchies of clusters and modeled featureThe paper deals with applications of various clustering methods on many kind of data. Clustering is a non-supervised technique that tries to maximize the intraclass and minimize the interclass similarity. The task is in a pretty pickle, because there are no user interactions expected and all the intelligence is left upon the computer there (and computers are stupid). The only possibility how to influence the result is to guess some parameters in advance and then wait (a long) time to have some results that usually doesn't satisfy the user much. We present some practical applications of existing methods. The dataset we have tested is quite complex, it includes both highly structured and unstructured data from business, computer vision and viruses to human DNA samples. As each domain is different, the needs and requirements are contradictory. Thus, we have tested many methods. These include classical partitioning methods (k-Means, k-Medoids), hierarchical (BIRCH), density based (DBSCAN) and model-based approaches (GMM and ART). Although we have developed our own method (Voronoi Clustering), suitable for a huge number of samples and clusters, it still has disadvantages similar to the other methods. We think, the global goal in the clustering is to establish a semi-supervised method that will (once) process the data and present its preliminary results to the user in an interactive way, similarly to OLAP. The proposed analysis will be able to present hierarchies of clusters and modeled features in different dimensions, not only bind to Euclidean space.