ISA Project Proposal: Development of an R package for cluster evaluation

Name: Sine Vestergård Jensen

E-Mail: sine144@gmail.com

ECTS: 5

Time-frame: 01.09.2015 – 31.12.2015

The suggested ISA shall provide the candidate Sine Vestergård Jensen with insights into clustering and skills to be outlined in the following.

**Background**

The task of grouping a given set of objects such that objects in the same group (i.e., cluster) are more similar to each other than to those in other groups is called clustering. Clustering is a long standing problem in computer science; thus, there exist many different clustering tools, which all intend to find an “optimal” clustering in terms of optimizing different criterion. Furthermore, each clustering tool has at least one parameter that influences the resulting clustering.

One example of a clustering tool is K-Means, where K is the parameter defining the number of clusters in the result. Cluster validity indices, also called clustering quality measures, are an objective criteria for judging the quality of a clustering and/or compare a clustering against a gold standard. They also help to decide whether a clustering is feasible for a given dataset and they assist to detect an optimal parameter set resulting in the best possible clustering for a given tool. Generally, one discriminates between internal measures (based only on the clustering) and external measures (compare the clustering to a given gold standard). In course of this project, we will examine both internal and external indices. Similarly to clustering tools, the validity measures are all based on a certain assumption of a perfect clustering and thus some are more susceptible to noise and other properties of the dataset. In course of this ISA, the indices should also be systematically evaluated towards those influences.

**State of the art and aim**

Most cluster validity indices have been implemented in e.g. Java and other programming language. They all use different interfaces and require different packages to be installed. This is also true for the programming language R for which the implementation of several of the most popular cluster validity indices are spread over different R packages and follow different standards and call structures. The aim of this project is to provide researchers with one unified R package tailored for the needs of cluster evaluation. That includes the implementation of the most common validity indices following one common interface. This will provide a powerful evaluation tool minimizing the future implementation efforts when performing a cluster evaluation and allows for the systematic evaluation of these indices with respect to the aforementioned effects like noise or dataset properties. This might identify the biases of these indices towards certain data sets or towards certain clustering structures, e.g. singletons or large, spherical shaped clusters.

**Conclusion**

Until now, researchers have to use several packages if they want to assess the quality of their clustering, all of them following different call procedures and require the cluster information in specific formats. Furthermore, some indices haven't been implemented in R at all. This project will provide researchers with an easy way to evaluate their clustering with regard to various validity indices and furthermore provides an overview of the properties and biases of well-known validity indices. The applicant will acquire proficiencies in cluster analysis, i.e., the different clustering tools, the influence of parameters and cluster validity indices on the results. Furthermore, the project will provide the applicant with experience in R package development and with participation in a bigger software project.