Supervised learning of clustering method performances on biological data

Background

The Computational Biology research group was established in October 2012 at the Department of Mathematics and Computer Science (IMADA) at the University of Southern Denmark (SDU). One of our goals is to develop computational methods to analyze large amounts of biological data. Often these data sets come without labels and are therefore analyzed with unsupervised learning methods as a first step to get an overview and indication of which follow up studies may make sense. Clustering methods are one popular class of such unsupervised learning approaches. Many different clustering methods exist which are based on different assumptions about the data and with different definitions of “good clusterings”. The same holds for validity indices, which judge the quality of a clustering and determine how significant the identified grouping is. Not each method and validity index is equally well suited for each type of data. Due to the large number of available methods and cluster validity indices, the scientist has a difficult time to identify suited methods and validity indices for a cluster study. Thus, it is of high importance to be able to identify preferable methods and indices prior to carrying out the cluster study. To this end we developed ClustEval, an integrated automatized framework which can analyze data sets, cluster them with many methods and evaluate many validity indices.

This project is about extending ClustEval with a supervised learning method, which aims at predicting the performance of clustering methods (best achieved cluster validity) on a new dataset based on intrinsic properties of the input data set and previous performances of that method: “Can the performance of methods be predicted using few data properties?”. Also, this supervised learning approach can be used to predict relationships between data set properties and the cluster validity indices: “Does this index always penalize/promote clusterings of data sets which have a certain property?”.

Aims

(1) Familiarize yourself with several supervised learning methods; understand their pros and cons. At least the following prediction methods should be evaluated: Linear Regression, SVM, Random Forest.

(2) Evaluate these methods in predicting method performance based on data properties and previous performances. Models should be trained on existing data sets and validated on training sets (e.g. consisting of new synthetic and real-world data sets).

(3) The prediction models should be validated using appropriate validation techniques (e.g. k-fold cross validation; bootstrapping)

Conclusion

Given the importance of clustering as a first step to determine promising follow-up studies on new data sets and the ever growing complexity due to the large number of clustering methods and validity indices, their reduction prior to cluster studies is of high interest for scientists. This project will provide the applicant with experience with machine learning techniques and working on larger software projects. The project will focus on providing an integrated solution, and ease access of analysis and results to other researches.