This thesis focuses on the algorithm selection problem, in which the goal is to recommend machine learning algorithms to a new dataset. The idea behind solving this issue is that algorithm performs similarly on similar datasets. The usual approach is to base the similarity measure on the fixed vector of metafeatures extracted out of each dataset. However, as the number of attributes among datasets varies, we may be loosing important information. Herein, we propose a family of algorithms able to handle even the non-propositional representations of datasets. Our methods use the idea of attribute assignment that builds the distance measure between datasets as a sum of distance given by the optimal assignment and an attribute distance measure. Furthermore, we prove that under certain conditions, we can guarantee the resulting dataset distance to be a metric. We carry out a series of metalearning experiments on the data extracted from the OpenML repository. We build up attribute distance using Genetic Algorithms, Genetic Programming and several regularization techniques such as multi-objectivization, coevolution, and bootstrapping. The experiment indicates that the resulting dataset distance can be successfully applied on the algorithm selection problem.
Although we use the proposed distance measures exclusively on metalearning, it is possible to use our methods even beyond this task. The algorithms can handle every situation where we have a notion of distance between elements of some set and are looking to define a distance on the power set of the original set.