Smart Sampling Strategies for Rule Discovery Optimization

Why

Rule discovery on knowledge bases refers to the process of mining novel, relevant, and potentially interesting patterns that can give insight into the data, and which can be used for various different downstream tasks, such as data validation. A well-studied approach to rule discovery is based on *frequent pattern mining*, which assumes that patterns that occur more often in the data are more interesting than those which occur only sporadically. Many rule discovery methods exploit this assumption, and produce a set of rules that each capture an often occurring pattern in the data. An example of such a rule might be that any bridge that is made of wood and that is older than fifty years should be subject to inspection every three months. Unfortunately, discovering such rules accurately is costly and time consuming, since every axiom in the data has to be evaluated. A possible solution to this is to use sampling, thereby giving up a bit of accuracy while reducing the time complexity.

What

The goal of this project is to optimize an off-the-shelf rule discovery pipeline with the introduction of smart sampling strategies: rather than computing the rules over the entire knowledge base, a proper subset is taken from which the rules are discovered instead. Different sampling strategies will have to be tested and evaluated, and the resulting trade-offs between accuracy and time complexity should be analysed. This may be further extended to parallelization strategies.

How

This project will follow the follow steps:

- Brief literature review about rule discovery and sampling strategies
- Implementing, testing, and analysis of various sampling strategies
- (optional) Implementing and testing different parallelization strategies
- Experimenting with and evaluating sampling strategies in a data validation task

Who

This project aims at students who are interested in logic and optimization. A strong familiarity with programming (Python) and first-order logic is required. Familiarity with statistics, data mining, and/or machine learning is a plus.

Supervisors

Xander Wilcke & TBA