Universiteit Leiden
The Netherlands

Massively collaborative machine learning

Rijn, J.N. van

English Summary

Many scientists are focussed on building models. Models are a schematical representation of a concept (often simplified). These models are based on observed data from the past, and make predictions about the future (yet unseen data). Subconsciously, we work with many models. We nearly process all information to a model. The weather is a classic Dutch example. When we observe dark clouds in the morning, we are inclined to think that it will be a rainy day. Conversely, when there is a clear blue sky, we will expect a sunny day. In this case, the model that we are using is: "If there are clouds, chances are that it is going to rain; if there is a clear sky, we might have a nice day." Somehow, we are capable of turning our experiences from the past into such models, and use these to make predictions about the future. This is called 'learning'.

There are many techniques that enable computers to learn as well. The field of research that develops such techniques is called Machine Learning. We encounter Machine Learning on a daily basis; some examples are:

- social media: the computer decides which advertisements should be shown to a specific user
- the stock market: the computer decides which stocks should be bought or sold
- spam filters: the computer determines whether an incoming message is spam

Many of these subjects are considered hot topics, hence many research is devoted to develop computer programs capable of building models (algorithms). Many of such algorithms exist, and these often consist of various options that subtly influence performance (parameters). Furthermore, there is mathematical proof that there exists no single algorithm that works well on every dataset. This complicates the task of selecting the right algorithm for a given task.

The field of meta-learning aims to resolve these problems. The purpose is to determine what kind of algorithms work well on which datasets. This is often done experimentally. A common approach is to execute many algorithms on many datasets, and based on the results learn which combinations work well (hence the name: meta-learning). Although this works well in practise, time is a limiting factor. Many of these experiments are very time-consuming, which unintentionally limits the scale of many studies.

In order to solve this problem, we have developed OpenML. This is an online database on which researches can share experimental results amongst each other, potentially scaling up the size of meta-learning studies. Having earlier experimental results freely accessible and reusable for others, it is no longer required to conduct time expensive experiments. Rather, researchers can answer such experimental questions by a simple database look-up.

This thesis addresses how OpenML can be used to answer fundamental meta-learning questions such as 'which algorithm should be preferred on a certain kind of data?', 'what is the effect of a given parameter on the performance of an algorithm?' and 'how does a newly developed algorithm perform compared to existing algorithms?'