Bachelor Thesis

Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Bachelor Thesis

Predictability of Classification Performance Measures with Meta-Learning

The goal of classification tasks is to assign a discrete class label to each object in a data set. For example, we could try to predict if a credit card transaction is fraudulent or not by considering transaction properties like timestamp, amount of money, recipient etc. Once we have learned a predictive model, we can use a variety of classification performance measures to assess how good our model is. For example, "accuracy" quantifies the overall number of correct predictions, and "recall" describes how many of the fraudulent transactions have been recognized.

Training a classification model usually takes some time, and additionally, there is a rich variety of model types as well as of hyperparameter settings. Hence, we are interested to estimate classification performance without actually training the model, just using some data set characteristics like percentage of missing values, average correlation of attributes with the class label etc. This problem is called meta-learning, as we build a meta-model to predict the performance of a base model.

Existing work on meta-learning usually considers only one particular measure of classification performance as prediction target. However, different types of performance measures used as meta-targets might result in different quality of meta-learning. As the meta-data sets as well as meta-models vary from study to study, one cannot obtain general insights on predictability of different classification performance measures by just comparing existing studies. There is the need of a systematic comparison based on one common meta-learning approach.

The goal of this thesis is to study the predictability of different classification performance measures. The following questions are particularly interesting:

- How well can we predict different classification performance measures with meta-learning?
- Do the most useful meta-features for the prediction depend on the classification performance measure used as target?
- How much does the choice of the meta-model matter when answering the previous questions?

The following steps are part of your thesis:

- Review literature about meta-learning classification performance.
- Design and implement experiments to meta-learn classification performance. You can use existing machine learning libraries, as well as existing meta-data sets.
- Evaluate your approach experimentally. You can use the server infrastructure of our chair.

During your work on this thesis, you will acquire practical knowledge about state-of-the-art machine learning libraries. You will get familiar with meta-learning as well as classification in general. You will gain experience in running and evaluating large scientific experiments.

You can write the thesis in English or German. Prior experience with classification in a programming language suitable for data science (e.g. Python, R) is beneficial, but not necessary if you are motivated to learn.

Contact

Jakob Bach jakob.bach@kit.edu +49 721 608-47339 Room: 351

Am Fasanengarten 5 76131 Karlsruhe Building: 50.34