Initializing Bayesian Hyperparameter Optimization via Meta-Learning

Matthias Feurer feurerm@cs.uni-freiburg.de
Jost Tobias Springenberg springj@cs.uni-freiburg.de
Frank Hutter fh@cs.uni-freiburg.de
Albert-Ludwigs-Universität Freiburg

Hyperparameters of machine learning algorithms should be optimized by automated methods, not by humans.
Bayesian Optimization is a powerful hyperparameter optimization tool.
In contrast to human domain experts, Bayesian Optimization does not use knowledge from previous runs on different datasets.
We employ meta-learning to obtain promising configurations to warmstart Bayesian Optimization.

SMBO with Meta-Learning

Find D_sim to D_train
\[ \text{ML Algorithm A} \]
\[ \text{Configuration Space A of A} \]
\[ \text{Dataset D_train} \]

Initialize Search with \( \lambda_D \)

Fit regression model on pairs of
(\( \lambda_A(D_{sim}) \), \( A_A(D_{sim}) \))

Select promising configuration \( \lambda \in \Lambda \)

Evaluate \( A_A(D_{sim}) \)

Standard Bayesian Optimization (black) together with meta-learning initialization (red).

MI-SMBO

• Meta-learning Initialized Sequential Model-based Bayesian Optimization
• Mimics human domain experts: uses configurations which are known to work well on similar datasets
• Similarity is defined by a distance between datasets based on metafeatures

Dataset Similarity

Similarity of datasets is defined by a distance function between dataset metafeatures. Some examples of metafeatures for the Iris dataset:

- # samples
- # categorical features
- # numerical features
- # classes
- # features
- # categorical features
- # training examples

We compared two distance functions:
- L_1 norm:
  \[ d_1(D_{sim}, D) = \sum |m_i^s - m_i^t| \]
- Spearman correlation coefficient between known model performances:
  \[ d_2(D_{sim}, D) = 1 - \text{corr}(f_{i,1}(\cdot), f_{t,1}(\cdot)) \]

Caveats:
- This only works for a fixed set of hyperparameters
- Cannot be computed for a new dataset \( D_{new} \)

Solution: compute \( d_i(D_{sim}, D) \) for all \( i \leq i \leq N \) and use regression to learn a mapping from \( (m_i^s, m_i^t) \) to \( d_i(D_{sim}, D) \). We used a random forest for this mapping.

Experiments

Setup

- Two experiments:
  1. Tuned the hyperparameters of an SVM (see paper)
  2. Combined algorithm selection and hyperparameter optimization (CASH) for scikit-learn: AutoSklearn

<table>
<thead>
<tr>
<th>Component</th>
<th>Hyperparameter</th>
<th># Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main</td>
<td>Classifier</td>
<td>3</td>
</tr>
<tr>
<td>Main</td>
<td>Programming</td>
<td>2</td>
</tr>
<tr>
<td>SVM</td>
<td>log(( \chi ))</td>
<td>21</td>
</tr>
<tr>
<td>SVM</td>
<td>log,()</td>
<td>18</td>
</tr>
<tr>
<td>LinearSVM</td>
<td>log(( \chi ))</td>
<td>24</td>
</tr>
<tr>
<td>LinearSVM</td>
<td>Penalty</td>
<td>1</td>
</tr>
<tr>
<td>BF</td>
<td>Metafeatures</td>
<td>5</td>
</tr>
<tr>
<td>BF</td>
<td>Min splits</td>
<td>10</td>
</tr>
<tr>
<td>BF</td>
<td>Criterion</td>
<td>1</td>
</tr>
<tr>
<td>XGB</td>
<td>max depth</td>
<td>2</td>
</tr>
</tbody>
</table>

- Validated our approach on 57 datasets from OpenML.org
  - Leave one dataset out: Ran MI-SMBO on one dataset and assumed knowledge of performance on all other 56
  - Precomputed a dense grid of 1623 hyperparameter configurations
  - Ran each optimization algorithm 10 times on each dataset
- Used 46 metafeatures from the literature
- Tried 40 different instantiations of MI-SMBO

Results

Average rank of different optimization algorithms. Since we ran each algorithm ten times on each dataset, we drew a bootstrap sample of 1000 joint runs and computed the average across these runs. We then further averaged these ranks across all 57 datasets.

Average rank of MI-SMBO with different number of initial configurations.

- Top: Percentage of datasets on which MI-SMBO performs statistically better than its competitors.
- Bottom: As above, but percentage of losses.

This plot shows that MI-SMBO improves over vanilla SMAC on 36% of the datasets, while it is worse on only 8%. We also observe that metalearning leads to a great performance boost in the beginning of SMBO.