Scalable Gaussian process-based transfer surrogates for hyperparameter optimization

MI2 DataLab Summer Research Seminar 2023
Mateusz Grzyb supported by Katarzyna Woźnica

Introduction

Paper

- Proposes a meta-learning extension of the SMBO HPO method
- Published in 2017 in the Springer Machine Learning journal
- Has 100 citations according to the Google Scholar service
- Contains a quite nice mini-survey on some other HPO methods

Mach Learn (2018) 107:43–78
https://doi.org/10.1007/s10994-017-5684-y

CrossMark

Scalable Gaussian process-based transfer surrogates for hyperparameter optimization

Martin Wistuba[1] · Nicolas Schilling[1] ·
Lars Schmidt-Thieme[1]

Received: 7 November 2016 / Accepted: 4 October 2017 / Published online: 22 December 2017

Abstract Algorithm selection as well as hyperparameter optimization are tedious task that have to be dealt with when applying machine learning to real-world problems. Sequential model-based optimization (SMBO), based on so-called "surrogate models", has been employed to allow for faster and more direct hyperparameter optimization. A surrogate model is a machine learning regression model which is trained on the meta-level instances in order to predict the performance of an algorithm on a specific data set given the hyperparameter settings and data set descriptors. Gaussian processes, for example, make good surrogate models as they provide probability distributions over labels. Recent work on SMBO also

MI

Authors

Martin Wistuba

Nicolas Schilling

Lars Schmidt-Thieme

Information Systems and Machine Learning Lab, University of Hildesheim, Germany

Prerequisites

HPO problem definition

Sequential Model-based Optimization (SMBO)

Meta-learning system

1. A meta-learning system must include a learning subsystem which adapts with experience.
2. Experience is gained by exploiting meta-knowledge extracted
   a. in a previous learning episode on a single data set, and/or
   b. from different domains or problems.

“Our contributions to SMBO lead to a system that fulfills all of these requirements. Our system adapts with experience by updating the surrogate model which represents the meta-knowledge. Furthermore, we exploit meta-knowledge extracted on the new data set and from previous problems.”

Gaussian processes

Contribution

Scalable Hyperparameter Optimization with Products of Gaussian Process Experts

Nicolas Schilling, Martin Wistuba and Lars Schmidt-Thieme

Information Systems and Machine Learning Lab
Universitätsplatz 1, 31141 Hildesheim, Germany
{schilling,wistuba,schmidt-thieme}@ismll.uni-hildesheim.de

Abstract. In machine learning, hyperparameter optimization is a challenging but necessary task that is usually approached in a computationally expensive manner such as grid-search. Out of this reason, surrogate based black-box optimization techniques such as sequential model-based optimization have been proposed which allow for a faster hyperparameter optimization. Recent research proposes to also integrate hyperparameter performances on past data sets to allow for a faster and more efficient hyperparameter optimization.
In this paper, we use products of Gaussian process experts as surrogate models for hyperparameter optimization. Naturally, Gaussian processes are a decent choice as they offer good prediction accuracy as well as estimations about their uncertainty. Additionally, their hyperparameters can be tuned very effectively. However, in the light of large meta data sets, learning a single Gaussian process is not feasible as it involves inversion of a large kernel matrix. This directly limits their usefulness for hyperparameter optimization if large scale hyperparameter performances on past data sets are given.
By using products of Gaussian process experts the scalability issues can be circumvened, however, this usually comes with the price of having less predictive accuracy. In our experiments, we show empirically that products of experts nevertheless perform very well compared to a variety of published surrogate models. Thus, we propose a surrogate model that performs as well as the current state of the art, is scalable to large scale meta knowledge, does not include hyperparameters itself and finally is even very easy to parallelize.

Keywords: Hyperparameter Optimization, Sequential Model-Based Optimization, Product of Experts

1 Introduction

Two-Stage Transfer Surrogate Model for Automatic Hyperparameter Optimization

Martin Wistuba, Nicolas Schilling, and Lars Schmidt-Thieme

Information Systems and Machine Learning Lab
University of Hildesheim
Universitätsplatz 1, 31141 Hildesheim, Germany
{wistuba,schilling,schmidt-thieme}@ismll.uni-hildesheim.de

Abstract. The choice of hyperparameters and the selection of algorithms is a crucial part in machine learning. Bayesian optimization methods have been used very successfully to tune hyperparameters automatically, in many cases even being able to outperform the human expert. Recently, these techniques have been massively improved by using meta-knowledge. The idea is to use knowledge of the performance of an algorithm on given other data sets to automatically accelerate the hyperparameter optimization for a new data set.
In this work we present a model that transfers this knowledge in two stages. At the first stage, the function that maps hyperparameter configurations to hold-out validation performances is approximated for previously seen data sets. At the second stage, these approximations are combined to rank the hyperparameter configurations for a new data set. In extensive experiments on the problem of hyperparameter optimization as well as the problem of combined algorithm selection and hyperparameter optimization, we are outperforming the state of the art methods.

Keywords: hyperparameter optimization, meta-learning, transfer learning

1 Introduction

The tuning of hyperparameters is an omnipresent problem in the machine learning community. In comparison to model parameters, which are estimated by a learning algorithm, hyperparameters are parameters that have to be specified before the execution of the algorithm. Typical examples for hyperparameters are the trade-off parameter C of a support vector machine or the number of layers and nodes in a neural network. Unfortunately, the choice of the hyperparameters is crucial and decides whether the performance of an algorithm is state

MI

Scalable Gaussian Process Transfer (SGPT)

Weights - Product of experts

Weights - Kernel regression

Transfer Acquisition Function (TAF)

Experiments

Meta-data sets

1. SVM
    - 50 random data sets from the UCI repository
    - Linear SVM, RBF SVM, and polynomial SVM
    - Kernel choice, C and gamma/d - hyperparameters
    - 288 hyperparameter configurations in total
2. WEKA
    - algorithm + hyperparameter configuration selection
    - 19 algorithms, 59 data sets (?), 21,871 configurations
    - roughly 1.3 million individual experiments in total

Competing strategies

1. Random Search (Random)
2. Independent Gaussian Procces (I-GP)
3. Independent Random Forest (I-RF)
4. Initialization for I-GP and I-RF (I-GP (init) and I-RF (init))
5. Surrogate Collaborative Tuning (SCoT)
6. Gaussian Process with Multi-Kernel Learning (MKL-GP)
7. Factorized Multilayer Perceptron (FMLP)
8. Scalable Gaussian Process Transfer Surrogate (SGPT-{PoE, M, R})
9. Transfer Acquisition Function (TAF-{PoE, M, R})

Evaluation metrics

Results

Conclusion

Conclusion

- Utilizing meta-data can be beneficial…
- …given that smart algorithms are used.
- Scalability of a method plays a major role.
- SGPT works well, but it has some problems…
  - different performance scales
  - constant meta-data influence
- …which TAF seems to solve and it works better.
- Number of iterations does not tell the full story.