Optimally weighted ensembles of surrogate models for sequential parameter optimization

Echtenbruck, M.M.

Author: Echtenbruck, M.M.
Title: Optimally weighted ensembles of surrogate models for sequential parameter optimization
Issue Date: 2020-07-02

Chapter 6

Summary and Outlook

In this work, we specified a taxonomy of ensembles and discussed known methods for its strengths and weaknesses (e.g., restrictions to particular types of models or applications). The main contribution of this work aims to overcome these weaknesses and develop an ensemble strategy that works reliably and accurately on arbitrary objective functions and models and thus releases the user from the burden to select the right surrogate model. The proposed ensemble strategy is first developed and analyzed in a small setup of two base models and later extended for the use of a large heterogeneous set of base models. Finally, the method is adapted for and applied to SPO on a large set of objective functions of various characteristics.

Section 6.1 provides a summary of this work and discusses the conclusions drawn from the performed research. Section 6.2 gives an overview of possible future work on this topic.

6.1 Summary

A common method to perform an optimization on a function that can not be optimized analytically is to perform a search on this function by iteratively and strategically choosing and evaluating points of the function. However, in real-world optimization tasks, the budget in terms of number of function evaluations is often constrained by the time or cost for these function evaluations. In such circumstances, it is a common technique to learn a surrogate model, e.g., regression model, of the response function from available evaluations and to use this model to decide on the location of future evaluations. Sequential Parameter Optimization (SPO) is a well-known approach for solving black-box optimization problems with expensive function evaluations with the help of surrogate models.

For such optimization processes, the choice of the surrogate model can have a significant influence on the solution quality and performance of the optimizer. However, to make meaningful decisions on which surrogate model to select for a given problem, often expert knowledge about the objective function and the characteristics of the surrogate model likewise is required. However, in many situations, preliminary knowledge about the function, or all available models, is not available. Automated methods that learn all by itself which surrogate model type suits the problem best, could help to overcome this problem.

This thesis introduces new methods on how to manage multiple heterogeneous surrogate models for regression and optimization of expensive black-box optimization problems. The overarching goal is to release the user from the burden to select the right surrogate model and to create an ensemble building strategy that works reliably and accurately on arbitrary objective functions and models. The primary focus is on regression problems, and optimization processes that allow only for a comparatively small number of function evaluations since these are constraints that often come with real-world problems.

In Chapter 3 of this thesis a taxonomy of known methods to do this is introduced and specified. The model selection methods are classified into two types.
The 'Single Evaluation Model Selection' methods, in general, use a predefined strategy to select the most appropriate model. This strategy may also utilize data obtained from previous evaluations. However, such ad hoc rules ignore the rules of parsimony and do not, or only marginally, rely on the data to help select the best model.
Methods classified as 'Multi Evaluation Model Selection' tackle this problem by evaluating all available types of surrogate models. But under circumstances when there is more than one strong model in the set, it might be beneficial to combine inference output across several models.
This is what is done by the 'Model Combination' methods. However, the known strategies, in general, are restricted in one or the other way (e.g., to homogeneous models or particular applications).

Based on these studies of existing ensemble approaches and their strengths and weaknesses, a method is envisioned, that overcomes the deficiencies and combines the advantages of these approaches. To this end, the method should do an exhaustive preliminary evaluation of all models to gain the best insight into the models' performances, then trains all models on the data to enable the use of the complete knowledge of all models. Still, the method should follow the principle of parsimony and prefer a combination of predictions over a single prediction only if it is clearly beneficial for the overall accuracy. The same applies to the number of models used, it should not be a decision between a single model or a combination of all, but any number of models that seem to be best.

In Chapter 4 the insights from Chapter 3 are used to develop a new ensemble method. This approach is studied in a fundamental way, by first evaluating ensembles of only two surrogate models in detail. It is shown that the convex combination of models is beneficial in many cases since the convex combination of the predictions of two base models averages positive as well as negative prediction errors of the base models. The ensemble generated by the convex combination of models can compete with the base models and in some cases even outperforms them.

The insights gained in the studies of convex combinations of two base models make convex linear combinations of models an ideal choice for combining models. The advantages recognized in this study are:

- Due to the linear convex combination that is used for combination, the ensemble cannot perform worse than the weakest base model.
- The ensemble can perform better than the base models when compensating opposing prediction errors.
- A CCM is favored over a base model only if the overall fit of the ensemble model is actually better (in terms of RMSE) than the overall fit of both base models.
- The nature of the combination that is given by a weighted sum with a normalized positive weight space is intuitive and interpretable.
- The linear convex combination of predictions for a given set of weights is easy to compute.

In preparation for the implementation of the algorithm with a large set of heterogeneous models, the algorithm is then specified in a more general way using three base models. Furthermore, the exhaustive search used to find the optimal combination weights in a fixed grid is replaced by a more flexible (1+1)-ES. These adaptations are accompanied by experiments to ensure that the changes have no adverse influence on the performance of the method.

On this basis, in the following, the set of base models is extended to the intended size and experiments are carried out to ensure that the performance of the method does not suffer from this change. First experiments show that the (1+1)-ES, without further changes, is not able to handle the extended search space. Therefore, several approaches are discussed and implemented to overcome this difficulty. Finally, the chapter closes with comparing experiments of all approaches on a set of test functions based on physical models. It is shown that the approaches are at an advantage over the base models in some cases. Another essential insight gained in these experiments is that sometimes a model can make a beneficial contribution to the ensemble although no single, high-ranked model is able to do so.

In Chapter 5 the developed ensemble strategy is introduced in more detail and adapted for the application in SPO. These adaptations are needed, e.g., to adjust the models continuously in the presence of dynamically changing and non-uniform data sets. These adaptations are:

- Periodically rebuilding of models and temporarily suspending models, to take dynamical updates into account.
- Local density weighted cross-validation, to deal with non-uniform point distributions.
- Adaptation of the (1+1)-ES weight optimization method, to deal with large ensemble sets.

The dynamically adapting ensemble strategy is then extensively tested for its performance and computation time in sequential optimization processes on various objective functions. First, instances of the method using different settings for the rebuild and the suspension interval are compared and the impact of these settings on the performance and computation time is analyzed. In both, its performance as well as its computation time, the method is also compared to the base models and to two strong ensemble competitors that share different parts of the characteristics of the proposed ensemble method.

The results show that the dynamically adapting ensemble strategy performs reliably and, in terms of accuracy, can compete with the base models as well as the competitors. It is shown that the proposed ensemble strategy method has a clear advantage over approaches that are restricted to the selection of a single best base model in fixed intervals. Given that the best base model is not known beforehand, the dynamically adapting ensemble strategy would be the best choice. The analysis of the impact of the rebuild interval and the suspension interval on the performance of the ensemble strategy showed that a regular adaptation of the ensemble setup is preferable, though this update should be restricted to a smaller subset of stronger base models that may be updated in a less frequent interval. A well-considered choice for the length of these intervals also has a distinct impact on the calculation time of the ensemble strategy.

In summary, and with the main research question in mind, it can be said that the primary goal to develop a strategy that works as reliably and as accurately as possible on arbitrary objective functions, and that uses arbitrary types of surrogate models is accomplished. The ensemble method proposed in this thesis selects or combines surrogate models from a set of heterogeneous surrogate models to achieve prediction results that can compete with or even improve the predictions of single models. Though it is knowingly accepted that this could happen at the expense of the ensembles computation time, this drawback can be lessened using well-considered settings for the rebuild interval and the suspension time. However, with a focus on expensive real-world applications, the computation times needed is expected to be negligible.

6.2 Outlook

First and foremost the approach would lend itself to be implemented in other sequential optimization packages, such as SUMO [55]. Furthermore, it could be beneficial to use the proposed ensemble method in algorithm configuration frameworks, such as SMAC [15, 112] and IRACE [56].

Some questions remained unanswered and require further investigation; other questions were not addressed yet. Questions that were not addressed in this work concern the ways of combining and evaluating models. We chose to use convex linear combinations of the models' predictions since it is both easy to calculate also for several heterogeneous models and comprehensive in terms of meaningfulness. For the evaluation of the models, we used the RMSE, which we adapted to a weighted variant for the application in SPO. Both, the evaluation method, as well as the way to combine the models, may be further investigated to determine the best way to do this. The question of how to evaluate and find the best model is accompanied by the general questions if better surrogates always result in improved performance.
The first of these questions concerns the characteristics of the search space for the model weights. Some of the results in Chapter 4 suggest the assumption that the function that is searched might be convex. If this assumption could be confirmed

this would allow for more direct search strategies to be applied and thus enhance the proposed ensemble strategy.
In this work, we chose an (1+1)-ES. Building on the assumption that the regarded search space is supposedly but not surely unimodal and convex this would be a reliable and robust search algorithm that performs well on a search landscape as assumed, but can also handle more difficult search spaces. However, other search strategies may be tested to improve the overall performance.
The recommendation for the values of τ and λ are based on a comparatively small number of experiments. Still, these values have a high impact on the overall performance and computation time of the proposed ensemble method. Further investigation may be done to allow for a more precise and reliable recommendation.
Clustering of the observed data points is an interesting idea and might be a beneficial addition to the proposed method. Though for this work, we did not consider this since the main focus was on real-world applications which are often restricted to a smaller number of function evaluations that complicate reasonable clustering of the data points.

The calculation time of the method may be negotiable for expensive real-world applications but remains a drawback for less expensive applications. However, the method has the right prerequisites to be further sped up through parallelization. First and foremost, the calculation time needed for the cross-validation step, which makes the majority of the calculation time, could be immensely reduced by parallelization. Additional calculation time may also be saved by parallelization of the optimization step by applying parallel optimization methods as introduced by Mostaghim et al. [113].

Surrogate models are not only used in Sequential Optimization but also in the selection of Evolutionary Algorithms. Also, for such surrogate-model assisted evolutionary algorithms the approach could be beneficially implemented. For an overview, see Chugh et al. [114].

Finally, the method could be applied to multi-objective optimization problems. First approaches in this direction were presented by Hussein et al. [115, 116]. Since the proposed approach showed good results in single-objective optimization problems, it is expected to show comparable good results in multi-objective optimization problems. Today, optimization with (much) more than 3 objective functions is often referred to as many objective optimization. Also for these problems first surrogate model based approaches have been proposed by Chugh et al. [117] and it could be investigated how these could benefit from mixed models.

Some of the questions and ideas formulated in Section 6.2 are based on questions and ideas of the bookchapter *'Open Issues in Surrogate-Assisted Optimization'* by Stork et al. [118].