Efficient tuning of automated machine learning pipelines

Nguyen, D.A.

English Summary

AutoML has attracted community attention due to its success in shortening the machine learning development cycle for real-world applications. Optimization plays a crucial role in AutoML frameworks by helping to identify a fine-tuned ML pipeline that suits a given practical problem. Several state-of-the-art optimization approaches, including Bayesian optimization, Bandit learning, and Racing procedures, have been proposed to enhance the performance of AutoML. However, the existing studies often formulate the AutoML optimization as a Hyperparameter Optimization (HPO) problem using the Combined Algorithm Selection and Hyperparameter Optimization (CASH) approach. This limited perspective can restrict their effectiveness in addressing the underlying problem effectively.

In this thesis, we comprehensively address this limitation by *formulating AutoML optimization*, covering the entire ML pipeline synthesis, as discussed in Chapter 1. Specifically, we introduce a novel class of hyperparameter, called "*algorithm choice*", which enables the modeling of algorithm selection. This class incorporates a unique attribute that allows for the hierarchical organization of algorithms based on their technical similarities. Notably, to our best knowledge, this is the first attempt to visualize the relationship between algorithms.

Next, a comprehensive investigation to provide a holistic understanding of AutoML and its optimization is provided in Chapters 2 and 3. Chapter 2 comprehensively investigates AutoML and its integral concepts and aspects. It delves into the foundations of AutoML, exploring the motivations behind its development and its significance in shortening the machine learning development cycle for real-world applications. Chapter 3 focuses specifically on AutoML optimization. This chapter delves into various optimization techniques and algorithms that are employed to improve the performance and efficiency of AutoML frameworks. It explores approaches such as Grid search, Random search, Bayesian optimization, Bandit learning, and Racing procedures, providing an in-depth analysis of their principles, strengths, and limitations in the context of AutoML optimization.

Another point of concern is benchmarking to evaluate the robustness and general applicability of optimization approaches empirically. Chapter 4 presents two sets of benchmark experiments. These benchmarks consist of 117 agreed-on datasets that may require processing through an ML pipeline of multiple operators such as encoding, normalization, resampling, and classification. The chosen datasets may contain categorical data, incomplete instances, or have an imbalanced distribution. Furthermore, the benchmarks comprise standardized search spaces along with an experimental methodology and setup for benchmarking purposes.

In Chapter 5, we conducted a detailed study on the effectiveness of AutoML optimization in handling class imbalance problems. Our findings suggest that BO is a highly efficient approach to tackle this problem. This finding enabled us to focus on improving BO to solve AutoML optimization problems confidently. Consequently, we were able to create two effective optimization algorithms based on BO in Chapter 7 and Chapter 8, as well as successfully implement the AutoML optimization approach to a real-world application in Chapter 6.

Another key contribution is the development of a performance metric specifically designed to address the dual challenges of class imbalance and unequal importance problems. This metric, discussed in Chapter 6, offers a comprehensive evaluation framework for assessing the performance of ML models in such scenarios.

In addition to the mentioned contributions, this thesis presents two novel AutoML optimization approaches. In Chapter 7 of this thesis, we presented a combinational sampling method that efficiently solves AutoML optimization problems through Bayesian optimization. Our approach is designed to maximize the coverage of ML algorithm samples in the search space, which results in a robust surrogate model for BO. We conducted experiments on 117 benchmark datasets and found that using our sampling approach significantly improves BO performance when compared to BO without our approach. In addition to the proposed sampling approach, we have also introduced an optimization software known as BO4ML in this chapter. This software is implemented based on the proposed sampling approach and can be used for optimization. Moving on to Chapter 8, we proposed an improved AutoML optimization technique known as DACOpt, which is an efficient contesting procedure. This innovative method aims to allocate tuning resources for optimizers over search areas more effectively. Specifically, DACOpt rewards resources to areas that show promise while terminating the optimizing process on areas that are less likely to yield desired results. Our experiments on

117 benchmark datasets indicate that our proposed approach offers a significant performance advantage over BO for AutoML optimization problems.