
Learning to Tune XGBoost with XGBoost

Johanna Sommer^{1,2}, Dimitrios Sarigiannis³, Thomas Parnell³

¹IBM Germany, ²Duale Hochschule Baden-Württemberg, ³IBM Research
johanna@mail-sommer.com, saridimi@gmail.com, tpa@zurich.ibm.com

Abstract

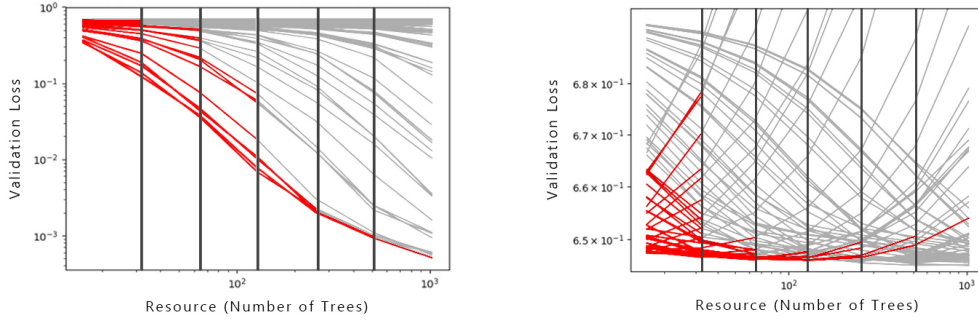
In this short paper we investigate whether meta-learning techniques can be used to more effectively tune the hyperparameters of machine learning models using successive halving (SH). We propose a novel variant of the SH algorithm (MeSH), that uses meta-regressors to determine which candidate configurations should be eliminated at each round. We apply MeSH to the problem of tuning the hyperparameters of a gradient-boosted decision tree model. By training and tuning our meta-regressors using existing tuning jobs from 95 datasets, we demonstrate that MeSH can often find a superior solution to both SH and random search.

1 Introduction

Modern machine learning models have a large number of hyperparameters that must be tuned. Gradient-boosted decision trees (GBDTs), perhaps the most popular classical machine learning model today, have at least five hyperparameters that must be tuned. Similarly, one can view the architecture of a deep neural network as a highly-structured hyperparameter space. This insight has led to the emerging field of neural architecture search, in which researchers attempt to let optimization algorithms discover neural architectures that out-perform those designed by their colleagues.

Model-free methods for hyperparameter optimization based on successive halving (SH) [10, 9] have recently gained a lot of traction in the aforementioned applications. Their popularity can be attributed to the ease with which they can be parallelized and/or distributed across a cluster of machines, as well as the lack of assumptions made about the nature of the underlying space relative to model-based approaches such as Bayesian optimization [5]. Essential to these methods is the notion of a resource [10]: normally either the number of iterations of the learning algorithm or the number of training examples. By randomly sampling a large number of hyperparameter configurations and evaluating them on a small resource, one can identify promising configurations quickly, and carry forward only the best for evaluation on a larger resource. Implicit to this algorithm is the idea that the relative ordering of configurations (in terms of validation loss) is consistent whether using a small or a large resource. However, when tuning GBDTs we often observe that this is far from reality, as illustrated in Figure 1.

Meta-learning [13] is an orthogonal field that focuses on *learning to learn*, or learning from experience. In the context of hyperparameter optimization this means: can we leverage our experience from previous tuning tasks on different datasets, to better tune the hyperparameters of a given model on a new, target dataset? Existing attempts to achieve this generally involve trying to warm-start hyperparameter optimization methods with a set of configurations that were known to perform well on *similar* datasets [4, 12, 3]. The task of identifying similar datasets is normally achieved by computing meta-features that characterize each dataset in an expressive way. Of particular interest in this paper is a class of meta-features, known as *landmarking meta-features* [11, 1]. Computing a landmarking meta-feature typically involves training a relatively inexpensive model (e.g. naive Bayes) and evaluating its performance on some hold-out data. Such meta-features are known to play a significant role in predicting the performance of an algorithm [8]. However, these meta-features are



(a) Ideal behaviour: the best configuration in the final round can be identified from the beginning and is not eliminated as the algorithm progresses.

(b) Problematic behaviour: promising configurations in the initial rounds do not correspond to the best configurations in the final round.

Figure 1: Example of how configurations are eliminated in the SH algorithm, applied to tuning XGBoost [2], for two different datasets. Vertical lines indicate the rounds at which half of the configurations are discarded, configurations that survive elimination are marked in red.

relatively expensive to compute, especially if one would like to generate landmarking meta-features using more complex models.

The critical insight behind this work is as follows: in the context of hyperparameter optimization using SH, such meta-features are essentially available at no cost. The SH algorithm is fundamentally based on evaluating configurations with an ever-increasing resource, and all previous evaluations can be regarded as landmarking meta-features. In the following, we explore how we may exploit this observation, to allow SH to learn from experience.

2 Improving Successive Halving via Meta-Learning

In this section we describe a new algorithm, MeSH, that aims to use **Meta-learning** to improve the behaviour of **Successive Halving**. The algorithm consists of an *online phase*, essentially a minor modification to the SH algorithm in which the decision regarding which configurations should be eliminated is based on the output of a set of predictive meta-models, and an *offline phase* during which the meta-models are trained and tuned. We will now describe the two phases of MeSH in more detail.

Online phase. The online phase of MeSH is presented in full in Algorithm 1. If we compare and contrast the algorithm with that proposed in [6, 7, 10, 9], we will notice two key differences. Firstly, we must provide a meta-model for each *round* of the algorithm. Secondly, when we eliminate configurations, instead of using the validation loss evaluated with the current resource, we use the output of the corresponding meta-model. For each configuration, this meta-model takes as an input a meta-feature vector consisting of (a) some dataset meta-features, (b) configuration meta-features (i.e., the setting of each hyperparameter) and (c) the validation loss for this configuration evaluated at the previous rounds. The meta-model then outputs, for each configuration, a prediction of how this configuration would perform (in terms of validation loss) if evaluated using the maximal resource. These predicted values are then sorted and only the top $1/\eta$ are carried over into the next round.

The online phase is graphically illustrated in Figure 2, for the case where $r_{min} = 16$, $r_{max} = 128$ and $\eta = 2$. The general idea is that as more and more landmarking features become available (for the later rounds), the more accurate the predictions of the meta-models. It is hoped that, by using the predicted value of the validation loss at the maximal resource, rather than using the measured validation loss at the current resource, that we can overcome the *crossover* problem that was illustrated in Figure 1(b). In terms of overheads relative to vanilla SH, since the landmarking meta-features are given for free, one must only pay the price of making predictions with the meta-model, which in general is negligible compared to the cost of training.

Algorithm 1 MeSH (Online Phase)

Require: Number of configurations n , eliminator factor η , min. resource r_{min} , max. resource r_{max}

1: $s_{max} \leftarrow \lfloor \log_{\eta}(r_{max}/r_{min}) \rfloor$

Require: Meta-models M_i for $i = 0, 1, \dots, s_{max}$, Dataset meta-features \mathcal{D}

Ensure: $n \geq \eta^{s_{max}}$

2: $T \leftarrow \text{sample_configurations}(n)$

3: **for** $i \in \{0, 1, \dots, s_{max}\}$ **do**

4: $n_i \leftarrow \lfloor n\eta^{-i} \rfloor$

5: $r_i \leftarrow r_{min}\eta^i$

6: $L_i \leftarrow \text{eval_and_return_val_loss}(\theta, r_i) : \theta \in T$

7: $L'_i \leftarrow \text{predict_final_val_loss}(M_i; \mathcal{D}, T, L_{i-1}, L_{i-2}, \dots, L_0)$

8: $T \leftarrow \text{top_k}(T, L'_i, n_i/\eta)$

9: **end for**

10: **return** Configuration with the lowest validation loss seen throughout the course of the algorithm.

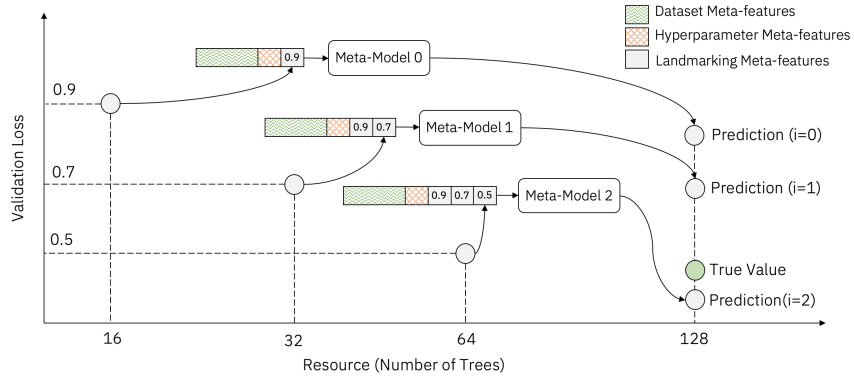


Figure 2: MeSH uses meta-regressors to predict the final validation loss.

Offline phase. The approach above can only prove useful if the meta-models are indeed accurate at predicting the final-round validation loss. In order to achieve this, in the offline phase of MeSH, we train and tune the meta-models using a large amount of data. Firstly, in order to construct the required meta-datasets, one must conduct a large experiment. This experiment involves sampling hyperparameter configurations, uniformly at random, for a large number of datasets and evaluating their validation loss given resource r_i for $i = 0, 1, \dots, s_{max}$. For the i -th meta-model, one then constructs a meta-dataset for which the meta-feature vectors consists of (a) dataset meta-features, (b) hyperparameter configuration meta-features and (c) all validation losses evaluated with resource r_j for $j = 0, 1, \dots, (i-1)$. The meta-model is then trained to minimize the mean-squared error between the predicted values and the target, given by the validation loss for the corresponding configuration evaluated at r_{max} . In order to tune the hyperparameters of the meta-model itself, one can use standard cross-validation techniques.

3 Experimental Results

Hyperparameter tuning for XGBoost. In this section we consider the problem of tuning the hyperparameters of an XGBoost model. Namely, we wish to tune: `lambda`, `colsample_bytree`, `max_depth` and `learning_rate` and `num_boost_rounds`. We will compare three solutions: random search (RS), SH and MeSH. We take `num_boost_rounds` to be the resource (as defined in SH and MeSH), and compare schemes under the constraint that the total resource utilized (budget) is the same. For SH and MeSH, we take $n = 64$, $\eta = 2$, $r_{min} = 16$ and $r_{max} = 1024$, resulting in 7 rounds in both cases. For RS, the scheme of equivalent budget corresponds to 7 configurations evaluated using the maximal resource¹. In all cases we use logistic loss as the training and validation loss function.

¹We leverage the early-stopping functionality of XGBoost so that the training can stop early (i.e., for a smaller number of boosting rounds) if the validation score starts to increase.

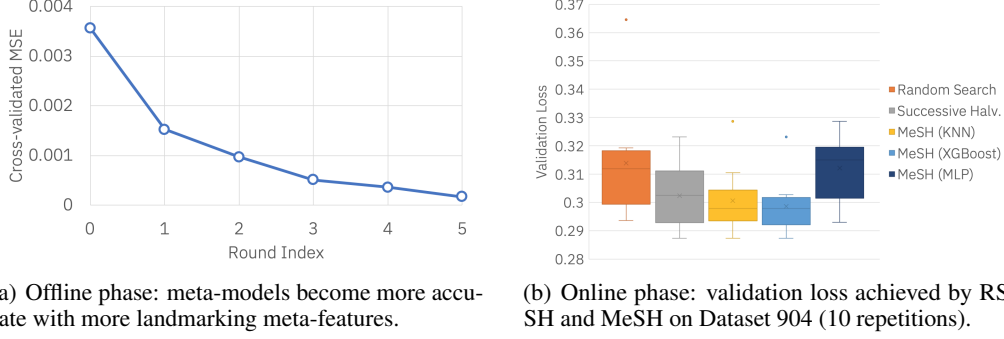


Figure 3: Performance of MeSH during offline phase and online phase.

Tuning Algorithm	Dataset 799		Dataset 904		Dataset 813		Dataset 897		Dataset 930	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Random Search	0.2209	0.015	0.3139	0.020	0.2073	0.015	0.0295	0.003	0.4903	0.005
Successive Halving	0.2140	0.008	0.3024	0.011	0.1978	0.011	0.0262	0.002	0.4907	0.004
MeSH (KNN)	0.2171	0.007	0.3007	0.012	0.2034	0.013	0.0264	0.003	0.4939	0.006
MeSH (MLP)	0.2173	0.008	0.3122	0.010	0.2076	0.010	0.0264	0.002	0.4901	0.004
MeSH (XGBoost)	0.2140	0.008	0.2986	0.011	0.1975	0.010	0.0262	0.002	0.4881	0.004

Table 1: Full validation loss results for the 5 test datasets (10 repetitions).

Constructing the meta-datasets. In order to generate robust meta-datasets we ran a large experiment using 95 binary classification datasets from the OpenML platform [14]. For each dataset, we ran an experiment for 40 CPU-hours, randomly sampling from the hyperparameter space, and recording the validation loss when trained with resource $r_i = 16, 32, \dots, 1024$. We then create the meta-feature vectors for the meta-dataset, comprising the landmarking meta-features extracted from the aforementioned experiment, together with the hyperparameter meta-features and the meta-features that describe the dataset. For the latter, we use a subset of the meta-features suggested in [13]. The resulting meta-dataset has around 300 thousand examples and up to 26 meta-features.

Training and tuning the meta-models. We consider three candidate meta-regressors: K-nearest neighbors (KNN), a multi-layer perceptron (MLP) and XGBoost itself. When training and tuning the meta-models, it is critical that the datasets that we wish evaluate in the online phase are not contained within the meta-dataset. We thus identify 5 test datasets. For each test dataset, we remove all of the corresponding examples from the meta-dataset, and perform 3-fold cross-validation on the remaining examples to tune the hyperparameters of the meta-regressors.² In Figure 3(a) we plot the validation MSE obtained during the offline tuning phase as a function of the round index. As expected, we see that as more landmarking meta-features become available, the meta-regressors are able to predict the validation loss at 1024 boosting rounds very effectively.

Evaluation of the online phase. In Figure 3(b) we present the best validation loss found for one of the test datasets. Our findings are twofold: (1) MeSH is able to find a significantly better solution than both RS and SH, (2) XGBoost is the best choice of meta-regressor. In Table 1, we present the full set of results for all 5 test datasets. We observe that: (a) in 3 of the 5 cases, MeSH with XGBoost is able to outperform both RS and SH, and (b) in the 2 remaining cases, MeSH is no worse than SH.

4 Conclusion and Future Work

We have proposed a novel modification of the SH algorithm (MeSH) that uses meta-regressors, trained and tuned on previous hyperparameter tuning jobs, in order to more effectively determine which configurations should be eliminated. We have applied this new algorithm to the problem of tuning the hyperparameters of XGBoost and have shown it is often capable of finding a better solution than both SH and RS. As a next step, we would like to see how MeSH performs when applied to more challenging problems like neural architecture search.

²Due to time restrictions, for the MLP meta-regressor we did not perform cross-validation and simply trained the meta-model with the default parameters.

References

- [1] Hilan Bensusan and Christophe Giraud-Carrier. Discovering task neighbourhoods through landmark learning performances. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 325–330. Springer, 2000.
- [2] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pages 785–794, New York, NY, USA, 2016. ACM.
- [3] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems* 28, pages 2962–2970. Curran Associates, Inc., 2015.
- [4] Matthias Feurer, Jost Tobias Springenberg, and Frank Hutter. Initializing bayesian hyperparameter optimization via meta-learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI’15, pages 1128–1135. AAAI Press, 2015.
- [5] F. Hutter, H. H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *Proceedings of the conference on Learning and Intelligent Optimization (LION 5)*, pages 507–523, January 2011.
- [6] Kevin Jamieson and Robert Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. pages 1–6, 03 2014.
- [7] Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 240–248, Cadiz, Spain, 09–11 May 2016. PMLR.
- [8] Alexandros Kalousis. Algorithm selection via meta-learning. 2002.
- [9] Liam Li, Kevin G. Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. Massively parallel hyperparameter tuning. *CoRR*, abs/1810.05934, 2018.
- [10] Lisha Li, Kevin G. Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Efficient hyperparameter optimization and infinitely many armed bandits. *CoRR*, abs/1603.06560, 2016.
- [11] Bernhard Pfahringer, Hilan Bensusan, and Christophe G. Giraud-Carrier. Meta-learning by landmarking various learning algorithms. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML ’00, pages 743–750, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.
- [12] Chris Thornton, Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’13, pages 847–855, New York, NY, USA, 2013. ACM.
- [13] Joaquin Vanschoren. Meta-learning: A survey. *CoRR*, abs/1810.03548, 2018.
- [14] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. Openml: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013.