Meta-Learning: A Survey

Joaquin Vanschoren j.vanschoren@tue.nl
Eindhoven University of Technology
5600MB Eindhoven, The Netherlands

Abstract

Meta-learning, or *learning to learn*, is the science of systematically observing how different machine learning approaches perform on a wide range of learning tasks, and then learning from this experience, or *meta-data*, to learn new tasks much faster than otherwise possible. Not only does this dramatically speed up and improve the design of machine learning pipelines or neural architectures, it also allows us to replace hand-engineered algorithms with novel approaches learned in a data-driven way. In this chapter, we provide an overview of the state of the art in this fascinating and continuously evolving field.

1. Introduction

When we learn new skills, we rarely - if ever - start from scratch. We start from skills learned earlier in related tasks, reuse approaches that worked well before, and focus on what is likely worth trying based on experience (Lake et al., 2017). With every skill learned, learning new skills becomes easier, requiring fewer examples and less trial-and-error. In short, we *learn how to learn* across tasks. Likewise, when building machine learning models for a specific task, we often build on experience with related tasks, or use our (often implicit) understanding of the behavior of machine learning techniques to help make the right choices.

The challenge in meta-learning is to learn from prior experience in a systematic, data-driven way. First, we need to collect *meta-data* that describe prior learning tasks and previously learned models. They comprise the exact *algorithm configurations* used to train the models, including hyperparameter settings, pipeline compositions and/or network architectures, the resulting *model evaluations*, such as accuracy and training time, the learned model parameters, such as the trained weights of a neural net, as well as measurable properties of the task itself, also known as *meta-features*. Second, we need to *learn* from this prior meta-data, to extract and transfer knowledge that guides the search for optimal models for new tasks. This chapter presents a concise overview of different meta-learning approaches to do this effectively.

The term *meta-learning* covers any type of learning based on prior experience with other tasks. The more *similar* those previous tasks are, the more types of meta-data we can leverage, and defining task similarity will be a key overarching challenge. Perhaps needless to say, there is no free lunch (Wolpert and Macready, 1996; Giraud-Carrier and Provost, 2005). When a new task represents completely unrelated phenomena, or random noise, leveraging prior experience will not be effective. Luckily, in real-world tasks, there are plenty of opportunities to learn from prior experience.

In the remainder of this chapter, we categorize meta-learning techniques based on the type of meta-data they leverage, from the most general to the most task-specific. First, in Section 2, we discuss how to *learn purely from model evaluations*. These techniques can be used to recommend generally useful configurations and configuration search spaces, as well as transfer knowledge from *empirically similar* tasks. In Section 3, we discuss how we can *characterize* tasks to more explicitly express task similarity and build meta-models that learn the relationships between data characteristics and learning performance. Finally, Section 4 covers how we can *transfer trained model parameters* between tasks that are inherently similar, e.g. sharing the same input features, which enables transfer learning (Pan and Yang, 2010) and few-shot learning (Ravi and Larochelle, 2017).

Note that while *multi-task learning* (Caruana, 1997) (learning multiple related tasks simultaneously) and *ensemble learning* (Dietterich, 2000) (building multiple models on the same task), can often be meaningfully combined with meta-learning systems, they do not in themselves involve learning from prior experience on other tasks.

2. Learning from Model Evaluations

Consider that we have access to prior tasks $t_j \in T$, the set of all known tasks, as well as a set of learning algorithms, fully defined by their *configurations* $\theta_i \in \Theta$; here Θ represents a discrete, continuous, or mixed configuration space which can cover hyperparameter settings, pipeline components and/or network architecture components. $\mathbf{P}$ is the set of all prior scalar evaluations $P_{i,j} = P(\theta_i, t_j)$ of configuration θ_i on task t_j, according to a predefined evaluation measure, e.g. accuracy, and model evaluation technique, e.g. cross-validation. $\mathbf{P}_{new}$ is the set of known evaluations $P_{i,new}$ on a new task t_{new}. We now want to train a *meta-learner* L that predicts recommended configurations Θ^*_{new} for a new task t_{new}. The meta-learner is trained on meta-data $\mathbf{P} \cup \mathbf{P}_{new}$. $\mathbf{P}$ is usually gathered beforehand, or extracted from meta-data repositories (Vanschoren et al., 2014, 2012). $\mathbf{P}_{new}$ is learned by the meta-learning technique itself in an iterative fashion, sometimes *warm-started* with an initial $\mathbf{P}'_{new}$ generated by another method.

2.1 Task-Independent Recommendations

First, imagine not having access to any evaluations on t_{new}, hence $\mathbf{P}_{new} = \varnothing$. We can then still learn a function $f : \Theta \times T \rightarrow \{\theta^*_k\}$, $k = 1..K$, yielding a set of recommended configurations *independent* of t_{new}. These θ^*_k can then be evaluated on t_{new} to select the best one, or to warm-start further optimization approaches, such as those discussed in Section 2.3.

Such approaches often produce a ranking, i.e. an *ordered* set θ^*_k. This is typically done by discretizing Θ into a set of candidate configurations θ_i, also called a *portfolio*, evaluated on a large number of tasks t_j. We can then build a ranking per task, for instance using *success rates*, *AUC*, or *significant wins* (Brazdil et al., 2003a; Demšar, 2006; Leite et al., 2012). However, it is often desirable that equally good but faster algorithms are ranked higher, and multiple methods have been proposed to trade off accuracy and training time (Brazdil et al., 2003a; van Rijn et al., 2015). Next, we can aggregate these single-task rankings into a *global ranking*, for instance by computing the average rank (Lin, 2010; Abdulrahman et al., 2018) across all tasks. When there is insufficient data to build a global ranking, one can recommend *subsets of configurations* based on the best known configurations for each prior task (Todorovski and Dzeroski, 1999; Kalousis, 2002), or return *quasi-linear rankings* (Cook et al., 1996).

To find the best θ^* for a task t_{new}, never before seen, a simple anytime method is to select the top-K configurations (Brazdil et al., 2003a), going down the list and evaluating each configuration on t_{new} in turn. This evaluation can be halted after a predefined value for K, a time budget, or when a sufficiently accurate model is found. In time-constrained settings, it has been shown that multi-objective rankings (including training time) converge to near-optimal models much faster (Abdulrahman et al., 2018; van Rijn et al., 2015), and provide a strong baseline for algorithm comparisons (Abdulrahman et al., 2018; Leite et al., 2012).

A very different approach to the one above is to first fit a differentiable function $f_j(\theta_i) = P_{i,j}$ on all prior evaluations of a specific task t_j, and then use gradient descent to find an optimized configuration θ_j^* per prior task (Wistuba et al., 2015a). Assuming that some of the tasks t_j will be similar to t_{new}, those θ_j^* will be useful for warm-starting Bayesian optimization approaches.

2.2 Configuration Space Design

Prior evaluations can also be used to learn a better *configuration space* Θ^*. While again independent from t_{new}, this can radically speed up the search for optimal models, since only the more relevant regions of the configuration space are explored. This is critical when computational resources are limited, and proves to be an important factor in practical comparisons of AutoML systems (De Sa et al., 2017).

First, in the functional ANOVA (Hutter et al., 2014a) approach, hyperparameters are deemed important if they explain most of the variance in algorithm performance on a given task. van Rijn and Hutter (2018) evaluated this technique using 250,000 OpenML experiments with 3 algorithms across 100 datasets.

An alternative approach is to first *learn* an optimal hyperparameter default setting, and then define hyperparameter importance as the *performance gain* that can be achieved by tuning the hyperparameter instead of leaving it at that default value. Indeed, even though a hyperparameter may cause a lot of variance, it may also have one specific setting that always results in good performance. Probst et al. (2018) do this using about 500,000 OpenML experiments on 6 algorithms and 38 datasets. Default values are learned *jointly* for all hyperparameters of an algorithm by first training surrogate models for that algorithm for a large number of tasks. Next, many configurations are sampled, and the configuration that minimizes the average risk across all tasks is the recommended default configuration. Finally, the importance (or *tunability*) of each hyperparameter is estimated by observing how much improvement can still be gained by tuning it.

Weerts et al. (2018) learn defaults *independently* from other hyperparameters, and defined as the configurations that occur most frequently in the top-K configurations for every task. In the case that the optimal default value depends on meta-features (e.g. the number of training instances or features), simple functions are learned that include these meta-features. Next, a statistical test defines whether a hyperparameter can be safely left at this default, based on the *performance loss* observed when *not* tuning a hyperparameter (or a set of hyperparameters), while all other parameters are tuned. This was evaluated using 118,000 OpenML experiments with 2 algorithms (SVMs and Random Forests) across 59 datasets.

2.3 Configuration Transfer

If we want to provide recommendations for a specific task t_{new}, we need additional information on how similar t_{new} is to prior tasks t_j. One way to do this is to evaluate a number of recommended (or potentially random) configurations on t_{new}, yielding new evidence $\mathbf{P}_{new}$. If we then observe that the evaluations $P_{i,new}$ are similar to $P_{i,j}$, then t_j and t_{new} can be considered intrinsically similar, based on empirical evidence. We can include this knowledge to train a meta-learner that predicts a recommended set of configurations Θ^*_{new} for t_{new}. Moreover, every selected θ^*_{new} can be evaluated and included in $\mathbf{P}_{new}$, repeating the cycle and collecting more empirical evidence to learn which tasks are similar to each other.

2.3.1 Relative Landmarks

A first measure for task similarity considers the relative (pairwise) performance differences, also called *relative landmarks*, $RL_{a,b,j} = P_{a,j} - P_{b,j}$ between two configurations θ_a and θ_b on a particular task t_j (Fürnkranz and Petrak, 2001). *Active testing* (Leite et al., 2012) leverages these as follows: it warm-starts with the globally best configuration (see Section 2.1), calls it θ_{best}, and proceeds in a tournament-style fashion. In each round, it selects the 'competitor' θ_c that most convincingly outperforms θ_{best} on similar tasks. It deems tasks to be similar if the relative landmarks of all evaluated configurations are similar, i.e., if the configurations perform similarly on both t_j and t_{new} then the tasks are deemed similar. Next, it evaluates the competitor θ_c, yielding $P_{c,new}$, updates the task similarities, and repeats. A limitation of this method is that it can only consider configurations θ_i that were evaluated on many prior tasks.

2.3.2 Surrogate Models

A more flexible way to transfer information is to build *surrogate models* $s_j(\theta_i) = P_{i,j}$ for all prior tasks t_j, trained using all available $\mathbf{P}$. One can then define task similarity in terms of the error between $s_j(\theta_i)$ and $P_{i,new}$: if the surrogate model for t_j can generate accurate predictions for t_{new}, then those tasks are intrinsically similar. This is usually done in combination with Bayesian optimization Rasmussen (2004) to determine the next θ_i.

Wistuba et al. (2018) train surrogate models based on Gaussian Processes (GPs) for every prior task, plus one for t_{new}, and combine them into a weighted, normalized sum, with the (new) mean μ defined as the weighted sum of the individual μ_j's (obtained from prior tasks t_j). The weights of the μ_j's are computed using the Nadaraya-Watson kernel-weighted average, where each task is represented as a vector of relative landmarks, and the Epanechnikov quadratic kernel (Nadaraya, 1964) is used to measure the similarity between the relative landmark vectors of t_j and t_{new}. The more similar t_j is to t_{new}, the larger the weight s_j, increasing the influence of the surrogate model for t_j.

Feurer et al. (2018a) propose to combine the predictive distributions of the individual Gaussian processes, which makes the combined model a Gaussian process again. The weights are computed following the agnostic Bayesian ensemble of Lacoste et al. (2014), which weights predictors according to an estimate of their generalization performance.

Meta-data can also be transferred in the acquisition function rather than the surrogate model (Wistuba et al., 2018). The surrogate model is only trained on $P_{i,new}$, but the next θ_i to evaluate is provided by an acquisition function which is the weighted average of the expected improvement (Jones et al., 1998) on $P_{i,new}$ and the predicted improvements on all prior $P_{i,j}$. The weights of the prior tasks can again be defined via the accuracy of the surrogate model or via relative landmarks. The weight of the expected improvement component is gradually increased with every iteration as more evidence $P_{i,new}$ is collected.

2.3.3 Warm-Started Multi-task Learning

Another approach to relate prior tasks t_j is to learn a joint task representation using $\mathbf{P}$. Perrone et al. (2017) train task-specific Bayesian linear regression (Bishop, 2006) surrogate models $s_j(\theta_i)$ and combine them in a feedforward Neural Network $NN(\theta_i)$ which learns a joint task representation that can accurately predict $P_{i,new}$. The surrogate models are pre-trained on OpenML meta-data to provide a warm-start for optimizing $NN(\theta_i)$ in a multi-task learning setting. Earlier work on multi-task learning (Swersky et al., 2013) assumed that we already have a set of 'similar' source tasks t_j. It transfers information between these t_j and t_{new} by building a joint GP model for Bayesian optimization that learns and exploits the exact relationship between the tasks. Learning a joint GP tends to be less scalable than building one GP per task, though. Springenberg et al. (2016) also assume that the tasks are related and similar, but learns the relationship between tasks during the optimization process using Bayesian Neural Networks. As such, their method is somewhat of a hybrid of the previous two approaches. Golovin et al. (2017) assume a sequence order (e.g., time) across tasks. It builds a stack of GP regressors, one per task, training each GP on the residuals relative to the regressor below it. Hence, each task uses the tasks before it as its priors.

2.3.4 Other Techniques

Multi-armed bandits (Robbins, 1985) provide yet another approach to find the source tasks t_j most related to t_{new} (Ramachandran et al., 2018a). In this analogy, each t_j is one arm, and the (stochastic) reward for selecting (pulling) a particular prior task (arm) is defined in terms of the error in the predictions of a GP-based Bayesian optimizer that models the prior evaluations of t_j as noisy measurements and combines them with the existing evaluations on t_{new}. The cubic scaling of the GP makes this approach less scalable, though.

Another way to define task similarity is to take the existing evaluations $P_{i,j}$, use Thompson Sampling (Thompson, 1933) to obtain the optima distribution ρ^j_{max}, and then measure the KL-divergence (Kullback and Leibler, 1951) between ρ^j_{max} and ρ^{new}_{max} (Ramachandran et al., 2018b). These distributions are then merged into a mixture distribution based on the similarities and used to build an acquisition function that predicts the next most promising configuration to evaluate. It is so far only evaluated to tune 2 SVM hyperparameters using 5 tasks.

Finally, a complementary way to leverage $\mathbf{P}$ is to recommend which configurations should *not* be used. After training surrogate models per task, we can look up which t_j are most similar to t_{new}, and then use $s_j(\theta_i)$ to discover regions of Θ where performance is predicted to be poor. Excluding these regions can speed up the search for better-performing ones. Wistuba et al. (2015b) do this using a task similarity measure based on the Kendall tau rank correlation coefficient (Kendall, 1938) between the ranks obtained by ranking configurations θ_i using $P_{i,j}$ and $P_{i,new}$, respectively.

2.4 Learning Curves

We can also extract meta-data about the training process itself, such as how fast model performance improves as more training data is added. If we divide the training in steps s_t, usually adding a fixed number of training examples every step, we can measure the performance $P(\theta_i, t_j, s_t) = P_{i,j,t}$ of configuration θ_i on task t_j after step s_t, yielding a *learning curve* across the time steps s_t. Learning curves are used extensively to speed up hyperparameter optimization on a given task (Kohavi and John, 1995; Provost et al., 1999; Swersky et al., 2014; Chandrashekaran and Lane, 2017). In meta-learning, however, learning curve information is transferred across tasks.

While evaluating a configuration on new task t_{new}, we can halt the training after a certain number of iterations $r < t$, and use the partially observed learning curve to predict how well the configuration will perform on the full dataset based on prior experience with other tasks, and decide whether to continue the training or not. This can significantly speed up the search for good configurations.

One approach is to assume that similar tasks yield similar learning curves. First, define a distance between tasks based on how similar the partial learning curves are: $dist(t_a, t_b) = f(P_{i,a,t}, P_{i,b,t})$ with $t = 1, ..., r$. Next, find the k most similar tasks $t_{1..k}$ and use their complete learning curves to predict how well the configuration will perform on the new complete dataset. Task similarity can be measured by comparing the shapes of the partial curves across all configurations tried, and the prediction is made by adapting the 'nearest' complete curve(s) to the new partial curve (Leite and Brazdil, 2005, 2007). This approach was also successful in combination with active testing (Leite and Brazdil, 2010), and can be sped up further by using multi-objective evaluation measures that include training time (van Rijn et al., 2015).

Interestingly, while several methods aim to predict learning curves during neural architecture search (Elsken et al., 2018), as of yet none of this work leverages learning curves previously observed on other tasks.

3. Learning from Task Properties

Another rich source of meta-data are characterizations (meta-features) of the task at hand. Each task $t_j \in T$ is described with a vector $m(t_j) = (m_{j,1}, ..., m_{j,K})$ of K meta-features $m_{j,k} \in M$, the set of all known meta-features. This can be used to define a task similarity measure based on, for instance, the Euclidean distance between $m(t_i)$ and $m(t_j)$, so that we can transfer information from the most similar tasks to the new task t_{new}. Moreover, together with prior evaluations $\mathbf{P}$, we can train a *meta-learner* L to predict the performance $P_{i,new}$ of configurations θ_i on a new task t_{new}.

3.1 Meta-Features

Table 1 provides a concise overview of the most commonly used meta-features, together with a short rationale for why they are indicative of model performance. Where possible, we also show the formulas to compute them. More complete surveys can be found in the literature (Rivolli et al., 2018; Vanschoren, 2010; Mantovani, 2018; Reif et al., 2014; Castiello et al., 2005).

Name	Formula	Rationale	Variants
Nr instances	n	Speed, Scalability (Michie et al., 1994)	p/n, $log(n)$, log(n/p)
Nr features	p	Curse of dimensionality (Michie et al., 1994)	$log(p)$, % categorical
Nr classes	c	Complexity, imbalance (Michie et al., 1994)	ratio min/maj class
Nr missing values	m	Imputation effects (Kalousis, 2002)	% missing
Nr outliers	o	Data noisiness (Rousseeuw and Hubert, 2011)	o/n
Skewness	$\frac{E(X-\mu_X)^3}{\sigma_X^3}$	Feature normality (Michie et al., 1994)	min,max,μ,σ,q_1, q_3
Kurtosis	$\frac{E(X-\mu_X)^4}{\sigma_X^4}$	Feature normality (Michie et al., 1994)	min,max,μ,σ,q_1, q_3
Correlation	$\rho_{X_1 X_2}$	Feature interdependence (Michie et al., 1994)	min,max,μ,σ,ρ_{XY}
Covariance	$cov_{X_1 X_2}$	Feature interdependence (Michie et al., 1994)	min,max,μ,σ,cov_{XY}
Concentration	$\tau_{X_1 X_2}$	Feature interdependence (Kalousis and Hilario, 2001)	min,max,μ,σ,τ_{XY}
Sparsity	sparsity(X)	Degree of discreteness (Salama et al., 2013)	min,max,μ,σ
Gravity	gravity(X)	Inter-class dispersion (Ali and Smith-Miles, 2006a)	
ANOVA p-value	$p_{val_{\mathtt{X}_1 X_2}}$	Feature redundancy (Kalousis, 2002)	$p_{val_{XY}}$(Soares et al., 2004)
Coeff. of variation	$\frac{\sigma_Y}{\mu_Y}$	Variation in target (Soares et al., 2004)	
PCA ρ_{λ_1}	$\sqrt{\frac{\lambda_1}{1+\lambda_1}}$	Variance in first PC (Michie et al., 1994)	$\frac{\lambda_1}{\sum_i \lambda_i}$(Michie et al., 1994)
PCA skewness		Skewness of first PC (Feurer et al., 2014)	PCA kurtosis
PCA 95%	$\frac{dim_{95\%var}}{p}$	Intrinsic dimensionality (Bardenet et al., 2013)	
Class probability	$P(\mathtt{C})$	Class distribution (Michie et al., 1994)	min,max,μ,σ
Class entropy	$H(\mathtt{C})$	Class imbalance (Michie et al., 1994)	
Norm. entropy	$\frac{H(\mathtt{X})}{log_2 n}$	Feature informativeness (Castiello et al., 2005)	min,max,μ,σ
Mutual inform.	$MI(\mathtt{C},\mathtt{X})$	Feature importance (Michie et al., 1994)	min,max,μ,σ
Uncertainty coeff.	$\frac{MI(\mathtt{C},\mathtt{X})}{H(\mathtt{C})}$	Feature importance (Agresti, 2002)	min,max,μ,σ
Equiv. nr. feats	$\frac{H(C)}{MI(C,X)}$	Intrinsic dimensionality (Michie et al., 1994)	
Noise-signal ratio	$\frac{H(X)-MI(C,X)}{MI(C,X)}$	Noisiness of data (Michie et al., 1994)	
Fisher's discrimin.	$\frac{(\mu_{c1}-\mu_{c2})^2}{\sigma_{c1}^2-\sigma_{c2}^2}$	Separability classes c_1, c_2 (Ho and Basu, 2002)	See Ho:2002
Volume of overlap		Class distribution overlap (Ho and Basu, 2002)	See Ho and Basu (2002)
Concept variation		Task complexity (Vilalta and Drissi, 2002)	See Vilalta (1999)
Data consistency		Data quality (Köpf and Iglezakis, 2002)	See Köpf and Iglezakis (2002)
Nr nodes, leaves	$\lvert\eta\rvert, \lvert\psi\rvert$	Concept complexity (Peng et al., 2002)	Tree depth
Branch length		Concept complexity (Peng et al., 2002)	min,max,μ,σ
Nodes per feature	$\lvert\eta_X\rvert$	Feature importance (Peng et al., 2002)	min,max,μ,σ
Leaves per class	$\frac{\lvert\psi_c\rvert}{\lvert\psi\rvert}$	Class complexity (Filchenkov and Pendryak, 2015)	min,max,μ,σ
Leaves agreement	$\frac{n_{\psi_i}}{n}$	Class separability (Bensusan et al., 2000)	min,max,μ,σ
Information gain		Feature importance (Bensusan et al., 2000)	min,max,μ,σ, gini
Landmarker(1NN)	$P(\theta_{1NN}, t_j)$	Data sparsity (Pfahringer et al., 2000)	See Pfahringer et al. (2000)
Landmarker(Tree)	$P(\theta_{Tree}, t_j)$	Data separability (Pfahringer et al., 2000)	Stump,RandomTree
Landmarker(Lin)	$P(\theta_{Lin}, t_j)$	Linear separability (Pfahringer et al., 2000)	Lin.Disciminant
Landmarker(NB)	$P(\theta_{NB}, t_j)$	Feature independence (Pfahringer et al., 2000)	See Ler et al. (2005)
Relative LM	$P_{a,j} - P_{b,j}$	Probing performance (Fürnkranz and Petrak, 2001)	
Subsample LM	$P(\theta_i, t_j, s_t)$	Probing performance (Soares et al., 2001)	

Table 1: Overview of commonly used meta-features. Groups from top to bottom: simple, statistical, information-theoretic, complexity, model-based, and landmarkers. Continuous features X and target Y have mean μ_X, stdev σ_X, variance σ_X^2. Categorical features $\mathtt{X}$ and class $\mathtt{C}$ have categorical values π_i, conditional probabilities $\pi_{i|j}$, joint probabilities $\pi_{i,j}$, marginal probabilities $\pi_{i+} = \sum_j \pi_{ij}$, entropy $H(\mathtt{X}) = -\sum_i \pi_{i+} log_2(\pi_{i+})$.

To build a meta-feature vector $m(t_j)$, one needs to select and further process these meta-features. Studies on OpenML meta-data have shown that the optimal set of meta-features depends on the application (Bilalli et al., 2017). Many meta-features are computed on single features, or combinations of features, and need to be aggregated by summary statistics (min,max,μ,σ,quartiles,...) or histograms (Kalousis and Hilario, 2001). One needs to systematically extract and aggregate them (Pinto et al., 2016). When computing task similarity, it is also important to normalize all meta-features (Bardenet et al., 2013), perform feature selection (Todorovski et al., 2000), or employ dimensionality reduction techniques (e.g. PCA) (Bilalli et al., 2017). When learning meta-models, one can also use relational meta-learners (Todorovski and Dzeroski, 1999) or case-based reasoning methods (Lindner and Studer, 1999; Hilario and Kalousis, 2001; Kalousis and Hilario, 2003).

Beyond these general-purpose meta-features, many more specific ones were formulated. For streaming data one can use streaming landmarks (van Rijn et al., 2018, 2014), for time series data one can compute autocorrelation coefficients or the slope of regression models (Arinze, 1994; Prudêncio and Ludermir, 2004; dos Santos et al., 2004), and for unsupervised problems one can cluster the data in different ways and extract properties of these clusters (Soares et al., 2009). In many applications, domain-specific information can be leveraged as well (Smith-Miles, 2009; Olier et al., 2018).

3.2 Learning Meta-Features

Instead of manually defining meta-features, we can also *learn* a joint representation for groups of tasks. One approach is to build meta-models that generate a landmark-like meta-feature representation M' given other task meta-features M and trained on performance meta-data $\mathbf{P}$, or $f : M \mapsto M'$. Sun and Pfahringer (2013) do this by evaluating a predefined set of configurations θ_i on all prior tasks t_j, and generating a binary metafeature $m_{j,a,b} \in M'$ for every pairwise combination of configurations θ_a and θ_b, indicating whether θ_a outperformed θ_b or not, thus $m'(t_j) = (m_{j,a,b}, m_{j,a,c}, m_{j,b,c}, ...)$. To compute $m_{new,a,b}$, *meta-rules* are learned for every pairwise combination (a,b), each predicting whether θ_a will outperform θ_b on task t_j, given its other meta-features $m(t_j)$.

We can also learn a joint representation based entirely on the available $\mathbf{P}$ meta-data, i.e. $f : \mathbf{P} \times \Theta \mapsto M'$. We previously discussed how to do this with feed-forward neural nets (Perrone et al., 2017) in Section 2.3. If the tasks share the same input space, e.g., they are images of the same resolution, one can also use Siamese networks to learn a meta-feature representation (Kim et al., 2017). These are trained by feeding the data of two different tasks to two twin networks, and using the differences between the predicted and observed performance $P_{i,new}$ as the error signal. Since the model parameters between both networks are tied in a Siamese network, two very similar tasks are mapped to the same regions in the latent meta-feature space. They can be used for warm starting Bayesian hyperparameter optimization (Kim et al., 2017) and neural architecture search (Afif, 2018).

3.3 Warm-Starting Optimization from Similar Tasks

Meta-features are a very natural way to estimate task similarity and initialize optimization procedures based on promising configurations on similar tasks. This is akin to how human experts start a manual search for good models, given experience on related tasks.

Starting a *genetic search* algorithm in regions of the search space with promising solutions can significantly speed up convergence to a good solution. Gomes et al. (Gomes et al., 2012) recommend initial configurations by finding the k most similar prior tasks t_j based on the L1 distance between vectors $m(t_j)$ and $m(t_{new})$, where each $m(t_j)$ includes 17 simple and statistical meta-features. For each of the k most similar tasks, the best configuration is evaluated on t_{new}, and used to initialize a genetic search algorithm (Particle Swarm Optimization), as well as Tabu Search. Reif et al. (2012) follow a very similar approach, using 15 simple, statistical, and landmarking meta-features. They use a forward selection technique to find the most useful meta-features, and warm-start a standard genetic algorithm (GAlib) with a modified Gaussian mutation operation. Variants of active testing (see Sect. 2.3) that use meta-features were also tried (Miranda and Prudêncio, 2013; Leite et al., 2012), but did not perform better than the approaches based on relative landmarks.

Also model-based optimization approaches can benefit greatly from an initial set of promising configurations. SCoT (Bardenet et al., 2013) trains a single surrogate ranking model $f : M \times \Theta \rightarrow R$, predicting the rank of θ_i on task t_j. M contains 4 meta-features (3 simple ones and one based on PCA). The surrogate model is trained on all the rankings, including those on t_{new}. Ranking is used because the scale of evaluation values can differ greatly between tasks. A GP regression converts the ranks to probabilities to do Bayesian optimization, and each new $P_{i,new}$ is used to retrain the surrogate model after every step.

Schilling et al. (2015) use a modified multilayer perceptron as a surrogate model, of the form $s_j(\theta_i, m(t_j), b(t_j)) = P_{i,j}$ where $m(t_j)$ are the meta-features and $b(t_j)$ is a vector of j binary indications which are 1 if the meta-instance is from t_j and 0 otherwise. The multi-layer perceptron uses a modified activation function based on factorization machines (Rendle, 2010) in the first layer, aimed at learning a latent representation for each task to model task similarities. Since this model cannot represent uncertainties, an ensemble of 100 multilayer perceptrons is trained to get predictive means and simulate variances.

Training a single surrogate model on all prior meta-data is often less scalable. Yogatama and Mann (2014) also build a single Bayesian surrogate model, but only include tasks similar to t_{new}, where task similarity is defined as the Euclidean distance between meta-feature vectors consisting of 3 simple meta-features. The $P_{i,j}$ values are standardized to overcome the problem of different scales for each t_j. The surrogate model learns a Gaussian process with a specific kernel combination on all instances.

Feurer et al. (2014) offer a simpler, more scalable method that warm-starts Bayesian optimization by sorting all prior tasks t_j similar to Gomes et al. (2012), but including 46 simple, statistical, and landmarking meta-features, as well as $H(\mathtt{C})$. The t best configurations on the d most similar tasks are used to warm-start the surrogate model. They search over many more hyperparameters than earlier work, including preprocessing steps. This warm-starting approach was also used very effectively, and combined with ensembling, in autosklearn (Feurer et al., 2015).

Finally, one can also use *collaborative filtering* to recommend promising configurations (Stern et al., 2010). By analogy, the tasks t_j (users) provide ratings ($P_{i,j}$) for the configurations θ_i (items), and matrix factorization techniques are used to predict unknown $P_{i,j}$ values and recommend the best configurations for any task. An important issue here is the cold start problem, since the matrix factorization requires at least some evaluations on t_{new}. Yang et al. (2018) use a D-optimal experiment design to sample an initial set of evaluations

$P_{i,new}$. They predict both the predictive performance and runtime, to recommend a set of warm-start configurations that are both accurate and fast. Misir and Sebag (2013) and Mısır and Sebag (2017) leverage meta-features to solve the cold start problem. Fusi et al. (2017) also use meta-features, following the same procedure as Feurer et al. (2015), and use a probabilistic matrix factorization approach that allows them to perform Bayesian optimization to further optimize their pipeline configurations θ_i. This approach also yields useful latent embeddings of both the tasks and configurations.

3.4 Meta-Models

We can also *learn* the complex relationship between a task's meta-features and the utility of specific configurations by building a meta-model L that recommends the most useful configurations Θ^*_{new} given the meta-features M of the new task t_{new}. There exists a rich body of earlier work (Brazdil et al., 2009; Lemke et al., 2015; Giraud-Carrier, 2008; Luo, 2016) on building meta-models for algorithm selection (Bensusan and Giraud-Carrier, 2000; Pfahringer et al., 2000; Kalousis, 2002; Bischl et al., 2016) and hyperparameter recommendation (Kuba et al., 2002; Soares et al., 2004; Ali and Smith-Miles, 2006b; Nisioti et al., 2018). Experiments showed that boosted and bagged trees often yielded the best predictions, although much depends on the exact meta-features used (Kalousis and Hilario, 2001; Köpf and Iglezakis, 2002).

3.4.1 Ranking

Meta-models can also generate a *ranking* of the top-K most promising configurations. One approach is to build a k-nearest neighbor (kNN) meta-model to predict which tasks are similar, and then rank the best configurations on these similar tasks (Brazdil et al., 2003b; dos Santos et al., 2004). This is similar to the work discussed in Section 3.3, but without ties to a follow-up optimization approach. Meta-models specifically meant for ranking, such as predictive clustering trees (Todorovski et al., 2002) and label ranking trees (Cheng et al., 2009) were also shown to work well. Approximate Ranking Trees Forests (ART Forests) (Sun and Pfahringer, 2013), ensembles of fast ranking trees, prove to be especially effective, since they have 'built-in' meta-feature selection, work well even if few prior tasks are available, and the ensembling makes the method more robust. *autoBagging* (Pinto et al., 2017) ranks Bagging workflows including four different Bagging hyperparameters, using an XGBoost-based ranker, trained on 140 OpenML datasets and 146 meta-features. Lorena et al. (2018) recommend SVM configurations for regression problems using a kNN meta-model and a new set of meta-features based on data complexity.

3.4.2 Performance Prediction

Meta-models can also directly predict the performance, e.g. accuracy or training time, of a configuration on a given task, given its meta-features. This allows us to estimate whether a configuration will be interesting enough to evaluate in any optimization procedure. Early work used linear regression or rule-base regressors to predict the performance of a discrete set of configurations and then rank them accordingly (Bensusan and Kalousis, 2001; Köpf et al., 2000). Guerra et al. (Guerra et al., 2008) train an SVM meta-regressor per classification algorithm to predict its accuracy, under default settings, on a new task t_{new}

given its meta-features. Reif et al. (Reif et al., 2014) train a similar meta-regressor on more meta-data to predict its *optimized* performance. Davis et al. (Davis and Giraud-Carrier, 2018) use a MultiLayer Perceptron based meta-learner instead, predicting the performance of a specific algorithm configuration.

Instead of predicting predictive performance, a meta-regressor can also be trained to predict algorithm training/prediction time, for instance, using an SVM regressor trained on meta-features (Reif et al., 2011), itself tuned via genetic algorithms (Priya et al., 2012). Yang et al. (2018) predict configuration runtime using polynomial regression, based only on the number of instances and features. Hutter et al. (2014b) provide a general treatise on predicting algorithm runtime in various domains.

Most of these meta-models generate promising configurations, but don't actually tune these configurations to t_{new} themselves. Instead, the predictions can be used to warm-start or guide any other optimization technique, which allows for all kinds of combinations of meta-models and optimization techniques. Indeed, some of the work discussed in Section 3.3 can be seen as using a distance-based meta-model to warm-start Bayesian optimization (Feurer et al., 2014; Fusi et al., 2017) or evolutionary algorithms (Gomes et al., 2012; Reif et al., 2012). In principle, other meta-models could be used here as well.

Instead of learning the relationship between a task's meta-features and configuration performance, one can also build surrogate models predicting the performance of configurations on specific tasks(Eggensperger et al., 2018). One can then learn how to combine these per-task predictions to warm-start or guide optimization techniques on a new task t_{new} (Feurer et al., 2018a; Perrone et al., 2017; Springenberg et al., 2016; Wistuba et al., 2018), as discussed in Section 2.3. While meta-features could also be used to combine per-task predictions based on task similarity, it is ultimately more effective to gather new observations $P_{i,new}$, since these allow to refine the task similarity estimates with every new observation (Feurer et al., 2018b; Wistuba et al., 2018; Leite et al., 2012).

3.5 Pipeline Synthesis

When creating entire machine learning pipelines (Serban et al., 2013), the number of configuration options grows dramatically, making it even more important to leverage prior experience. One can control the search space by imposing a fixed structure on the pipeline, fully described by a set of hyperparameters. One can then use the most promising pipelines on similar tasks to warm-start a Bayesian optimization (Feurer et al., 2015; Fusi et al., 2017).

Other approaches give recommendations for certain pipeline steps (Post et al., 2016; Strang et al., 2018), and can be leveraged in larger pipeline construction approaches, such as planning (Nguyen et al., 2014; Kietz et al., 2012; Gil et al., 2018; Wever et al., 2018) or evolutionary techniques (Olson et al., 2016; Sun et al., 2013). Nguyen et al. (2014) construct new pipelines using a beam search focussed on components recommended by a meta-learner, and is itself trained on examples of successful prior pipelines. Bilalli et al. (2018) predict which pre-processing techniques are recommended for a given classification algorithm. They build a meta-model per target classification algorithm that, given the t_{new} meta-features, predicts which preprocessing technique should be included in the pipeline.

Similarly, Schoenfeld et al. (2018) build meta-models predicting when a preprocessing algorithm will improve a particular classifier's accuracy or runtime.

AlphaD3M (Drori et al., 2018) uses a *self-play* reinforcement learning approach in which the current state is represented by the current pipeline, and actions include the addition, deletion, or replacement of pipeline components. A Monte Carlo Tree Search (MCTS) generates pipelines, which are evaluated to train a recurrent neural network (LSTM) that can predict pipeline performance, in turn producing the action probabilities for the MCTS in the next round. The state description also includes meta-features of the current task, allowing the neural network to learn across tasks.

3.6 To Tune or Not to Tune?

To reduce the number of configuration parameters to be optimized, and to save valuable optimization time in time-constrained settings, meta-models have also been proposed to predict whether or not it is worth tuning a given algorithm *given the meta-features of the task at hand* (Ridd and Giraud-Carrier, 2014) and how much improvement we can expect from tuning a specific algorithm versus the additional time investment (Sanders and Giraud-Carrier, 2017). More focused studies on specific learning algorithms yielded meta-models predicting when it is necessary to tune SVMs (Mantovani et al., 2015a), what are good default hyperparameters for SVMs given the task (including interpretable meta-models) (Mantovani et al., 2015b), and how to tune decision trees (Mantovani et al., 2016).

4. Learning from Prior Models

The final type of meta-data we can learn from are prior machine learning models themselves, i.e., their structure and learned model parameters. In short, we want to train a *meta-learner* L that learns how to train a (base-) learner l_{new} for a new task t_{new}, given similar tasks $t_j \in T$ and the corresponding optimized models $l_j \in \mathcal{L}$, where $\mathcal{L}$ is the space of all possible models. The learner l_j is typically defined by its model parameters $W = \{w_k\}$, $k = 1..K$ and/or its configuration $\theta_i \in \Theta$.

4.1 Transfer Learning

In *transfer learning* (Thrun and Pratt, 1998), we take models trained on one or more *source* tasks t_j, and use them as starting points for creating a model on a similar *target* task t_{new}. This can be done by forcing the target model to be structurally or otherwise similar to the source model(s). This is a generally applicable idea, and transfer learning approaches have been proposed for kernel methods (Evgeniou et al., 2005; Evgeniou and Pontil, 2004), parametric Bayesian models (Rosenstein et al., 2005; Raina et al., 2006; Bakker and Heskes, 2003), Bayesian networks (Niculescu-Mizil and Caruana, 2005), clustering (Thrun, 1998) and reinforcement learning (Hengst, 2002; Dietterich et al., 2002). Neural networks, however, are exceptionally suitable for transfer learning because both the structure and the model parameters of the source models can be used as a good initialization for the target model, yielding a *pre-trained* model which can then be further fine-tuned using the available training data on t_{new} (Thrun and Mitchell, 1995; Baxter, 1996; Bengio, 2012; Caruana, 1995). In some cases, the source network may need to be modified before transferring it

(Sharkey and Sharkey, 1993). We will focus on neural networks in the remainder of this section.

Especially large image datasets, such as ImageNet (Krizhevsky et al., 2012), have been shown to yield pre-trained models that transfer exceptionally well to other tasks (Donahue et al., 2014; Sharif Razavian et al., 2014). However, it has also been shown that this approach doesn't work well when the target task is not so similar (Yosinski et al., 2014). Rather than hoping that a pre-trained model 'accidentally' transfers well to a new problem, we can purposefully imbue meta-learners with an inductive bias (learned from many similar tasks) that allows them to learn new tasks much faster, as we will discuss below.

4.2 Meta-Learning in Neural Networks

An early meta-learning approach is to create recurrent neural networks (RNNs) able to modify their own weights (Schmidhuber, 1992, 1993). During training, they use their own weights as additional input data and observe their own errors to learn how to modify these weights in response to the new task at hand. The updating of the weights is defined in a parametric form that is differentiable end-to-end and can jointly optimize both the network and training algorithm using gradient descent, yet is also very difficult to train. Later work used reinforcement learning across tasks to adapt the search strategy (Schmidhuber et al., 1997) or the learning rate for gradient descent (Daniel et al., 2016) to the task at hand.

Inspired by the feeling that backpropagation is an unlikely learning mechanism for our own brains, Bengio et al. (1995) replace backpropagation with simple biologically-inspired parametric rules (or evolved rules (Chalmers, 1991)) to update the synaptic weights. The parameters are optimized, e.g. using gradient descent or evolution, across a set of input tasks. Runarsson and Jonsson (2000) replaced these parametric rules with a single layer neural network. Santoro et al. (2016b) instead use a memory-augmented neural network to learn how to store and retrieve 'memories' of prior classification tasks. Hochreiter et al. (2001) use LSTMs (Hochreiter and Schmidhuber, 1997) as a meta-learner to train multilayer perceptrons.

Andrychowicz et al. (2016) also replace the optimizer, e.g. stochastic gradient descent, with an LSTM trained on multiple prior tasks. The loss of the meta-learner (optimizer) is defined as the sum of the losses of the base-learners (optimizees), and optimized using gradient descent. At every step, the meta-learner chooses the weight update estimated to reduce the optimizee's loss the most, based on the learned model weights $\{w_k\}$ of the previous step as well as the current performance gradient. Later work generalizes this approach by training an optimizer on synthetic functions, using gradient descent (Chen et al., 2016). This allows meta-learners to optimize optimizees even if these do not have access to gradients.

In parallel, Li and Malik (2016) proposed a framework for learning optimization algorithms from a reinforcement learning perspective. It represents any particular optimization algorithm as a policy, and then learns this policy via guided policy search. Follow-up work (Li and Malik, 2017) shows how to leverage this approach to learn optimization algorithms for (shallow) neural networks.

The field of *neural architecture search* includes many other methods that build a model of neural network performance for a specific task, for instance using Bayesian optimization or reinforcement learning. See Elsken et al. (2018) for an in-depth discussion. However, most of these methods do not (yet) generalize across tasks and are therefore not discussed here.

4.3 Few-Shot Learning

A particularly challenging meta-learning problem is to train an accurate deep learning model using only a few training examples, given prior experience with very similar tasks for which we have large training sets available. This is called *few-shot learning*. Humans have an innate ability to do this, and we wish to build machine learning agents that can do the same (Lake et al., 2017). A particular example of this is 'K-shot N-way' classification, in which we are given many examples (e.g., images) of certain classes (e.g., objects), and want to learn a classifier l_{new} able to classify N new classes using only K examples of each.

Using prior experience, we can, for instance, learn a common feature representation of all the tasks, start training l_{new} with a better model parameter initialization W_{init} and acquire an inductive bias that helps guide the optimization of the model parameters, so that l_{new} can be trained much faster than otherwise possible.

Earlier work on *one-shot* learning is largely based on hand-engineered features (Fei-Fei et al., 2006; Fei-Fei, 2006; Fink, 2005; Bart and Ullman, 2005). With meta-learning, however, we hope to learn a common feature representation for all tasks in an end-to-end fashion.

Vinyals et al. (2016) state that, to learn from very little data, one should look to non-parameteric models (such as k-nearest neighbors), which use a memory component rather than learning many model parameters. Their meta-learner is a Matching Network that apply the idea of a memory component in a neural net. It learns a common representation for the labelled examples, and *matches* each new test instance to the memorized examples using cosine similarity. The network is trained on minibatches with only a few examples of a specific task each.

Snell et al. (2017) propose Prototypical Networks, which map examples to a p-dimensional vector space such that examples of a given output class are close together. It then calculates a prototype (mean vector) for every class. New test instances are mapped to the same vector space and a distance metric is used to create a softmax over all possible classes. Ren et al. (2018) extend this approach to semi-supervised learning.

Ravi and Larochelle (2017) use an LSTM-based meta-learner to learn an update rule for training a neural network learner. With every new example, the learner returns the current gradient and loss to the LSTM meta-learner, which then updates the model parameters $\{w_k\}$ of the learner. The meta-learner is trained across all prior tasks.

Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017), on the other hand, does not try to learn an update rule, but instead learns a model parameter initialization W_{init} that generalizes better to similar tasks. Starting from a random $\{w_k\}$, it iteratively selects a batch of prior tasks, and for each it trains the learner on K examples to compute the gradient and loss (on a test set). It then backpropagates the *meta-gradient* to update the weights $\{w_k\}$ in the direction in which they would have been easier to update. In other words, after each iteration, the weights $\{w_k\}$ become a better W_{init} to start finetuning any of the tasks. Finn and Levine (2017) show that MAML is able to approximate any learning algorithm when using a sufficiently deep ReLU network and certain losses. They also conclude that the MAML initializations are more resilient to overfitting on small samples, and generalize more widely than meta-learning approaches based on LSTMs. Grant et al. (2018) present a novel derivation of and extension to MAML, illustrating that this algorithm can be understood as inference for the parameters of a prior distribution in a hierarchical Bayesian model.

REPTILE (Nichol et al., 2018) is an approximation of MAML that executes stochastic gradient descent for K iterations on a given task, and then gradually moves the initialization weights in the direction of the weights obtained after the K iterations. The intuition is that every task likely has more than one set of optimal weights $\{w_i^*\}$, and the goal is to find a W_{init} that is close to at least one of those $\{w_i^*\}$ for every task.

Finally, we can also derive a meta-learner from a black-box neural network. Santoro et al. (2016a) propose Memory-Augmented Neural Networks (MANNs), which train a Neural Turing Machine (NTM) (Graves et al., 2014), a neural network with augmented memory capabilities, as a meta-learner. This meta-learner can then memorize information about previous tasks and leverage that to learn a learner l_{new}. SNAIL (Mishra et al., 2018) is a generic meta-learner architecture consisting of interleaved temporal convolution and causal attention layers. The convolutional networks learn a common feature vector for the training instances (images) to aggregate information from past experiences. The causal attention layers learn which pieces of information to pick out from the gathered experience to generalize to new tasks.

Overall, the intersection of deep learning and meta-learning proves to be particular fertile ground for groundbreaking new ideas, and we expect this field to become more important over time.

4.4 Beyond Supervised Learning

Meta-learning is certainly not limited to (semi-)supervised tasks, and has been successfully applied to solve tasks as varied as reinforcement learning, active learning, density estimation and item recommendation. The base-learner may be unsupervised while the meta-learner is supervised, but other combinations are certainly possible as well.

Duan et al. (2016) propose an end-to-end reinforcement learning (RL) approach consisting of a task-specific *fast* RL algorithm which is guided by a general-purpose *slow* meta-RL algorithm. The tasks are interrelated Markov Decision Processes (MDPs). The meta-RL algorithm is modeled as an RNN, which receives the observations, actions, rewards and termination flags. The activations of the RNN store the state of the fast RL learner, and the RNN's weights are learned by observing the performance of fast learners across tasks.

In parallel, Wang et al. (2016) also proposed to use a deep RL algorithm to train an RNN, receiving the actions and rewards of the previous interval in order to learn a base-level RL algorithm for specific tasks. Rather than using relatively unstructured tasks such as random MDPs, they focus on structured task distributions (e.g., dependent bandits) in which the meta-RL algorithm can exploit the inherent task structure.

Pang et al. (2018) offer a meta-learning approach to active learning (AL). The base-learner can be any binary classifier, and the meta-learner is a deep RL network consisting of a deep neural network that learns a representation of the AL problem across tasks, and a policy network that learns the optimal policy, parameterized as weights in the network. The meta-learner receives the current state (the unlabeled point set and base classifier state) and reward (the performance of the base classifier), and emits a query probability, i.e. which points in the unlabeled set to query next.

Reed et al. (2017) propose a few-shot approach for density estimation (DE). The goal is to learn a probability distribution over a small number of images of a certain concept (e.g., a handwritten letter) that can be used to generate images of that concept, or compute the probability that an image shows that concept. The approach uses autoregressive image models which factorize the joint distribution into per-pixel factors, usually conditioned on (many) examples of the target concept. Instead, a MAML-based few-shot learner is used, trained on examples of many other (similar) concepts.

Finally, Vartak et al. (2017) address the cold-start problem in matrix factorization. They propose a deep neural network architecture that learns a (base) neural network whose biases are adjusted based on task information. While the structure and weights of the neural net recommenders remain fixed, the meta-learner learns how to adjust the biases based on each user's item history.

All these recent new developments illustrate that it is often fruitful to look at problems through a *meta-learning lens* and find new, data-driven approaches to replace hand-engineered base-learners.

5. Conclusion

Meta-learning opportunities present themselves in many different ways, and can be embraced using a wide spectrum of learning techniques. Every time we try to learn a certain task, whether successful or not, we gain useful experience that we can leverage to learn new tasks. We should never have to start entirely from scratch. Instead, we should systematically collect our 'learning exhaust' and learn from it to build AutoML systems that continuously improve over time, helping us tackle new learning problems ever more efficiently. The more new tasks we encounter, and the more similar those new tasks are, the more we can tap into prior experience, to the point that most of the required learning has already been done beforehand. The ability of computer systems to store virtually infinite amounts of prior learning experiences (in the form of meta-data) opens up a wide range of opportunities to use that experience in completely new ways, and we are only starting to learn how to learn from prior experience effectively. Yet, this is a worthy goal: learning how to learn any task empowers us far beyond knowing how to learn specific tasks.

Acknowledgments

The author would like to thank Pavel Brazdil, Matthias Feurer, Frank Hutter, Raghu Rajan, and Jan van Rijn for many invaluable discussions and feedback on the manuscript.

References

S. Abdulrahman, P. Brazdil, J. van Rijn, and J. Vanschoren. Speeding up Algorithm Selection using Average Ranking and Active Testing by Introducing Runtime. *Machine Learning*, 107:79–108, 2018.

I. Nur Afif. Warm-starting deep learning model construction using meta-learning. Master's thesis, TU Eindhoven, 2018.

A. Agresti. *Categorical Data Analysis.* Wiley Interscience, 2002.

Shawkat Ali and Kate A. Smith-Miles. On learning algorithm selection for classification. *Applied Soft Computing*, 6(2):119 – 138, 2006a.

Shawkat Ali and Kate A. Smith-Miles. Metalearning approach to automatic kernel selection for support vector machines. *Neurocomput.*, 70(1):173–186, 2006b.

Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, pages 3981–3989, 2016.

B Arinze. Selecting appropriate forecasting models using rule induction. *Omega*, 22(6): 647–658, 1994.

B. Bakker and T. Heskes. Task Clustering and Gating for Bayesian Multitask Learning. *Journal of Machine Learning Research*, 4:83–999, 2003.

Rémi Bardenet, Mátyás Brendel, Balázs Kégl, and Michele Sebag. Collaborative hyperparameter tuning. In *Proceedings of ICML 2013*, pages 199–207, 2013.

Evgeniy Bart and Shimon Ullman. Cross-generalization: Learning novel classes from a single example by feature replacement. In *CVPR*, pages 672–679, 2005.

J. Baxter. Learning Internal Representations. In *Advances in Neural Information Processing Systems, NIPS*, 1996.

Samy Bengio, Yoshua Bengio, and Jocelyn Cloutier. On the search for new learning rules for anns. *Neural Processing Letters*, 2(4):26–30, 1995.

Y. Bengio. Deep learning of representations for unsupervised and transfer learning. In *ICML Unsupervised and Transfer Learning*, pages 17–36, 2012.

H Bensusan and A Kalousis. Estimating the predictive accuracy of a classifier. *Lecture Notes in Computer Science*, 2167:25–36, 2001.

Hilan Bensusan and Christophe Giraud-Carrier. Discovering task neighbourhoods through landmark learning performances. In *PKDD*, pages 325–330, 2000.

Hilan Bensusan, Christophe Giraud-Carrier, and Claire Kennedy. A higher-order approach to meta-learning. In *ILP*, pages 33 – 42, 2000.

Besim Bilalli, Alberto Abelló, and Tomàs Aluja-Banet. On the predictive power of meta-features in OpenML. *International Journal of Applied Mathematics and Computer Science*, 27(4):697 – 712, 2017.

Besim Bilalli, Alberto Abelló, Tomàs Aluja-Banet, and Robert Wrembel. Intelligent assistance for data pre-processing. *Computer Standards & Interf.*, 57:101 – 109, 2018.

B. Bischl, P. Kerschke, L. Kotthoff, M. Lindauer, Y. Malitsky, A. Fréchette, H. Hoos, F. Hutter, K. Leyton-Brown, K. Tierney, and J. Vanschoren. ASLib: A benchmark library for algorithm selection. *Artificial Intelligence*, 237:41–58, 2016.

Christopher M Bishop. Pattern recognition and machine learning. *Springer*, 2006.

P. Brazdil, C. Soares, and J. Pinto da Costa. Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results. *Machine Learning*, 50(3):251–277, 2003a.

Pavel Brazdil, Christophe Giraud-Carrier, Carlos Soares, and Ricardo Vilalta. *Metalearning: Applications to Data Mining*. Springer-Verlag Berlin Heidelberg, 2009.

Pavel B. Brazdil, Carlos Soares, and Joaquim Pinto Da Coasta. Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results. *Machine Learning*, 50(3):251–277, 2003b.

R. Caruana. Learning many related tasks at the same time with backpropagation. *Neural Information Processing Systems*, pages 657–664, 1995.

R. Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, 1997.

Ciro Castiello, Giovanna Castellano, and Anna Maria Fanelli. Meta-data: Characterization of input features for meta-learning. In *2nd International Conference on Modeling Decisions for Artificial Intelligence (MDAI)*, pages 457 – 468, 2005.

David J Chalmers. The evolution of learning: An experiment in genetic connectionism. In *Connectionist Models*, pages 81–90. Elsevier, 1991.

Akshay Chandrashekaran and Ian R Lane. Speeding up hyper-parameter optimization by extrapolation of learning curves using previous builds. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 477–492, 2017.

Yutian Chen, Matthew W Hoffman, Sergio Gómez Colmenarejo, Misha Denil, Timothy P Lillicrap, Matt Botvinick, and Nando de Freitas. Learning to learn without gradient descent by gradient descent. *arXiv preprint arXiv:1611.03824*, 2016.

Weiwei Cheng, Jens Hühn, and Eyke Hüllermeier. Decision tree and instance-based learning for label ranking. In *ICML*, pages 161–168, 2009.

W. D. Cook, M. Kress, and L. W. Seiford. A general framework for distance-based consensus in ordinal ranking models. *European Journal of Operational Research*, 96(2):392–397, 1996.

Christian Daniel, Jonathan Taylor, and Sebastian Nowozin. Learning step size controllers for robust neural network training. In *AAAI*, pages 1519–1525, 2016.

C. Davis and C. Giraud-Carrier. Annotative experts for hyperparameter selection. In *AutoML Workshop at ICML 2018*, 2018.

Alex De Sa, Walter Pinto, Luiz Otavio Oliveira, and Gisele Pappa. RECIPE: A grammar-based framework for automatically evolving classification pipelines. In *European Conference on Genetic Programming*, pages 246–261, 2017.

J. Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

T Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15, 2000.

T. Dietterich, D. Busquets, R. Lopez de Mantaras, and C. Sierra. Action Refinement in Reinforcement Learning by Probability Smoothing. In *19th International Conference on Machine Learning*, pages 107–114, 2002.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, pages 647–655, 2014.

P dos Santos, T Ludermir, and R Prudêncio. Selection of time series forecasting models based on performance information. *4th International Conference on Hybrid Intelligent Systems*, pages 366–371, 2004.

Iddo Drori, Yamuna Krishnamurthy, Remi Rampin, Raoni de Paula Lourenco, Jorge Piazentin Ono, Kyunghyun Cho, Claudio Silva, and Juliana Freire. AlphaD3M: Machine learning pipeline synthesis. In *AutoML Workshop at ICML*, 2018.

Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

K. Eggensperger, M. Lindauer, H.H. Hoos, F. Hutter, and K. Leyton-Brown. Efficient Benchmarking of Algorithm Configuration Procedures via Model-Based Surrogates . *Machine Learning*, 107:15–41, 2018.

Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *arXiv preprint arXiv:1808.05377*, 2018.

T. Evgeniou and M. Pontil. Regularized multi-task learning. In *Tenth Conference on Knowledge Discovery and Data Mining*, 2004.

T. Evgeniou, C. Micchelli, and M. Pontil. Learning Multiple Tasks with Kernel Methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

Li Fei-Fei. Knowledge transfer in learning to recognize visual objects classes. In *Intern. Conf. on Development and Learning*, page Art. 51, 2006.

Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *Pattern analysis and machine intelligence*, 28(4):594–611, 2006.

M Feurer, B Letham, and E Bakshy. Scalable meta-learning for Bayesian optimization. *arXiv*, 1802.02219, 2018a.

Matthias Feurer, Jost Tobias Springenberg, and Frank Hutter. Using meta-learning to initialize Bayesian optimization of hypxerparameters. In *International Conference on Meta-learning and Algorithm Selection*, pages 3 – 10, 2014.

Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems 28*, pages 2944–2952, 2015.

Matthias Feurer, Benjamin Letham, and Eytan Bakshy. Scalable meta-learning for bayesian optimization using ranking-weighted gaussian process ensembles. In *AutoML Workshop at ICML 2018*, 2018b.

Andray Filchenkov and Arseniy Pendryak. Dataset metafeature description for recommending feature selection. In *ISMW FRUCT*, pages 11–18, 2015.

Michael Fink. Object classification from a single example utilizing class relevance metrics. In *Neural information processing syst.*, pages 449–456, 2005.

Chelsea Finn and Sergey Levine. Meta-learning and universality. *arXiv 1710.11622*, 2017.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135, 2017.

J Fürnkranz and J Petrak. An evaluation of landmarking variants. *ECML/PKDD 2001 Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning*, pages 57–68, 2001.

Nicolo Fusi, Rishit Sheth, and Huseyn Melih Elibol. Probabilistic matrix factorization for automated machine learning. *arXiv preprint arXiv:1705.05355*, 2017.

Yolanda Gil, Ke-Thia Yao, Varun Ratnakar, Daniel Garijo, Greg Ver Steeg, Pedro Szekely, Rob Brekelmans, Mayank Kejriwal, Fanghao Luo, and I-Hui Huang. P4ML: A phased performance-based pipeline planner for automated machine learning. In *AutoML Workshop at ICML 2018*, 2018.

Christophe Giraud-Carrier. Metalearning-a tutorial. In *Tutorial at the International Conference on Machine Learning and Applications*, pages 1–45, 2008.

Christophe Giraud-Carrier and Foster Provost. Toward a justification of meta-learning: Is the no free lunch theorem a show-stopper. In *Proceedings of the ICML-2005 Workshop on Meta-learning*, pages 12–19, 2005.

D. Golovin, B. Solnik, S. Moitra, G. Kochanski, J. Karro, and D. Sculley. Google vizier: A service for black-box optimization. In *ICDM*, pages 1487–1495, 2017.

Taciana AF Gomes, Ricardo BC Prudêncio, Carlos Soares, André LD Rossi, and André Carvalho. Combining meta-learning and search techniques to select parameters for support vector machines. *Neurocomputing*, 75(1):3–13, 2012.

Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. *arXiv preprint arXiv:1801.08930*, 2018.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

Silvio B Guerra, Ricardo BC Prudêncio, and Teresa B Ludermir. Predicting the performance of learning algorithms using support vector machines as meta-regressors. In *ICANN*, pages 523–532, 2008.

B. Hengst. Discovering Hierarchy in Reinforcement Learning with HEXQ. In *International Conference on Machine Learning*, pages 243–250, 2002.

M Hilario and A Kalousis. Fusion of meta-knowledge and meta-data for case-based model selection. *Lecture Notes in Computer Science*, 2168:180–191, 2001.

Tin Kam Ho and Mitra Basu. Complexity measures of supervised classification problems. *Pattern Analysis and Machine Intellig.*, 24(3):289–300, 2002.

S. Hochreiter, A.S. Younger, and P.R. Conwell. Learning to learn using gradient descent. In *Lecture Notes on Computer Science, 2130*, pages 87–94, 2001.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

F. Hutter, H. Hoos, and K. Leyton-Brown. An Efficient Approach for Assessing Hyperparameter Importance. In *Proceedings of ICML*, 2014a.

F. Hutter, L. Xu, H. Hoos, and K. Leyton-Brown. Algorithm runtime prediction: Methods & evaluation. *Artificial Intelligence*, 206:79–111, 2014b.

Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

A. Kalousis. *Algorithm Selection via Meta-Learning.* PhD thesis, University of Geneva, Department of Computer Science, 2002.

A Kalousis and M Hilario. Representational issues in meta-learning. *Proceedings of ICML 2003*, pages 313–320, 2003.

Alexandros Kalousis and Melanie Hilario. Model selection via meta-learning: a comparative study. *Intl Journ. on Artificial Intelligence Tools*, 10(4):525–554, 2001.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Jörg-Uwe Kietz, Floarea Serban, Abraham Bernstein, and Simon Fischer. Designing KDD-workflows via HTN-planning for intelligent discovery assistance. In *5th Planning to Learn Workshop at ECAI 2012*, 2012.

J. Kim, S. Kim, and S. Choi. Learning to warm-start Bayesian hyperparameter optimization. *arXiv preprint arXiv:1710.06219*, 2017.

Ron Kohavi and George H John. Automatic parameter selection by minimizing estimated error. In *Proceedings of the International Conference Machine Learning*, pages 304–312, 1995.

C Köpf and I Iglezakis. Combination of task description strategies and case base properties for meta-learning. *ECML/PKDD Workshop on Integration and Collaboration Aspects of Data Mining*, pages 65–76, 2002.

C. Köpf, C. Taylor, and J. Keller. Meta-analysis: From data characterization for meta-learning to meta-regression. In *PKDD Workshop on Data Mining, Decision Support, Meta-Learning and ILP.*, pages 15–26, 2000.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

P. Kuba, P. Brazdil, C. Soares, and A. Woznica. Exploiting sampling and meta-learning for parameter setting support vector machines. In *Proceedings of IBERAMIA 2002*, pages 217–225, 2002.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Alexandre Lacoste, Mario Marchand, François Laviolette, and Hugo Larochelle. Agnostic Bayesian learning of ensembles. In *ICML*, pages 611–619, 2014.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Beh. and Brain Sc.*, 40, 2017.

R Leite and P Brazdil. Predicting relative performance of classifiers from samples. *Proceedings of ICML*, pages 497–504, 2005.

R Leite and P Brazdil. An iterative process for building learning curves and predicting relative performance of classifiers. *Lecture Notes in Computer Science*, 4874:87–98, 2007.

R. Leite, P. Brazdil, and J. Vanschoren. Selecting Classification Algorithms with Active Testing. *Lecture Notes in Artif. Intel.*, 10934:117–131, 2012.

Rui Leite and Pavel Brazdil. Active testing strategy to predict the best classification algorithm via sampling and metalearning. In *ECAI 2010*, pages 309–314, 2010.

C. Lemke, M. Budka, and B. Gabrys. Metalearning: a survey of trends and technologies. *Artificial intelligence review*, 44(1):117–130, 2015.

Daren Ler, Irena Koprinska, and Sanjay Chawla. Utilizing regression-based landmarkers within a meta-learning framework for algorithm selection. *Technical Report 569. University of Sydney*, pages 44–51, 2005.

Ke Li and Jitendra Malik. Learning to optimize. *arXiv preprint arXiv:1606.01885*, 2016.

Ke Li and Jitendra Malik. Learning to optimize neural nets. *arXiv preprint arXiv:1703.00441*, 2017.

S. Lin. Rank aggregation methods. *WIREs Computational Statistics*, 2:555–570, 2010.

G. Lindner and R. Studer. AST: Support for algorithm selection with a CBR approach. In *ICML Workshop on Recent Advances in Meta-Learning and Future Work*, pages 38–47. J. Stefan Institute, 1999.

Ana Carolina Lorena, Aron I. Maciel, Péricles B. C. de Miranda, Ivan G. Costa, and Ricardo B. C. Prudêncio. Data complexity meta-features for regression problems. *Machine Learning*, 107(1):209–246, 2018. doi: 10.1007/s10994-017-5681-1.

Gang Luo. A review of automatic selection methods for machine learning algorithms and hyper-parameter values. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 5(1):18, 2016.

Rafael G Mantovani, André LD Rossi, Joaquin Vanschoren, Bernd Bischl, and André CPLF Carvalho. To tune or not to tune: recommending when to adjust SVM hyper-parameters via meta-learning. In *Proceedings of IJCNN*, pages 1–8, 2015a.

Rafael G Mantovani, Tomáš Horváth, Ricardo Cerri, Joaquin Vanschoren, and André CPLF de Carvalho. Hyper-parameter tuning of a decision tree induction algorithm. In *Brazilian Conference on Intelligent Systems*, pages 37–42, 2016.

Rafael Gomes Mantovani, André LD Rossi, Joaquin Vanschoren, and André Carlos Carvalho. Meta-learning recommendation of default hyper-parameter values for SVMs in classifications tasks. In *ECML PKDD Workshop on Meta-Learning and Algorithm Selection*, 2015b.

R.G. Mantovani. *Use of meta-learning for hyperparameter tuning of classification problems.* PhD thesis, University of Sao Carlos, Brazil, 2018.

Donald Michie, David J. Spiegelhalter, Charles C. Taylor, and John Campbell. *Machine Learning, Neural and Statistical Classification.* Ellis Horwood, 1994.

P.B.C. Miranda and R.B.C. Prudêncio. Active testing for SVM parameter selection. In *Proceedings of IJCNN*, pages 1–8, 2013.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *Proceedings of ICLR*, 2018.

Mustafa Misir and Michèle Sebag. Algorithm Selection as a Collaborative Filtering Problem. Research report, INRIA, 2013.

Mustafa Mısır and Michèle Sebag. Alors: An algorithm recommender system. *Artificial Intelligence*, 244:291–314, 2017.

Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9 (1):141–142, 1964.

Phong Nguyen, Melanie Hilario, and Alexandros Kalousis. Using meta-mining to support data mining workflow planning and optimization. *Journal of Artificial Intelligence Research*, 51:605–644, 2014.

A. Nichol, J. Achiam, and J. Schulman. On first-order meta-learning algorithms. *arXiv*, 1803.02999v2, 2018.

A. Niculescu-Mizil and R. Caruana. Learning the Structure of Related Tasks. In *Proceedings of NIPS Workshop on Inductive Transfer*, 2005.

E. Nisioti, K. Chatzidimitriou, and A Symeonidis. Predicting hyperparameters from meta-features in binary classification problems. In *AutoML Workshop at ICML*, 2018.

I. Olier, N. Sadawi, G.R. Bickerton, J. Vanschoren, C. Grosan, L. Soldatova, and R.D. King. Meta-QSAR: learning how to learn QSARs. *Machine Learning*, 107:285–311, 2018.

Randal S Olson, Nathan Bartley, Ryan J Urbanowicz, and Jason H Moore. Evaluation of a tree-based pipeline optimization tool for automating data science. In *Proceedings of GECCO*, pages 485–492, 2016.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

K Pang, M. Dong, Y. Wu, and T. Hospedales. Meta-learning transferable active learning policies by deep reinforcement learning. In *AutoML Workshop at ICML*, 2018.

Y Peng, P Flach, C Soares, and P Brazdil. Improved dataset characterisation for meta-learning. *Lecture Notes in Com. Sc.*, 2534:141–152, 2002.

Valerio Perrone, Rodolphe Jenatton, Matthias Seeger, and Cedric Archambeau. Multiple adaptive Bayesian linear regression for scalable Bayesian optimization with warm start. *arXiv preprint arXiv:1712.02902*, 2017.

Bernhard Pfahringer, Hilan Bensusan, and Christophe G. Giraud-Carrier. Meta-learning by landmarking various learning algorithms. In *17th International Conference on Machine Learning (ICML)*, pages 743 – 750, 2000.

Fábio Pinto, Carlos Soares, and João Mendes-Moreira. Towards automatic generation of metafeatures. In *Proceedings of PAKDD*, pages 215–226, 2016.

Fábio Pinto, Vítor Cerqueira, Carlos Soares, and João Mendes-Moreira. autoBagging: Learning to rank bagging workflows with metalearning. *arXiv*, 1706.09367, 2017.

Martijn J. Post, Peter van der Putten, and Jan N. van Rijn. Does Feature Selection Improve Classification? A Large Scale Experiment in OpenML. In *Advances in Intelligent Data Analysis XV*, pages 158–170, 2016.

Rattan Priya, Bruno F. De Souza, Andre Rossi, and Andre Carvalho. Using genetic algorithms to improve prediction of execution times of ML tasks. In *Lecture Notes in Comp. Science*, volume 7208, pages 196–207, 2012.

P. Probst, B. Bischl, and A.-L. Boulesteix. Tunability: Importance of hyperparameters of machine learning algorithms. *ArXiv 1802.09596*, 2018.

Foster Provost, David Jensen, and Tim Oates. Efficient progressive sampling. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 23–32, 1999.

R Prudêncio and T Ludermir. Meta-learning approaches to selecting time series models. *Neurocomputing*, 61:121–137, 2004.

R. Raina, A. Y. Ng, and D. Koller. Transfer Learning by Constructing Informative Priors. In *Proceedings of ICML*, 2006.

Anil Ramachandran, Sunil Gupta, Santu Rana, and Svetha Venkatesh. Selecting optimal source for transfer learning in Bayesian optimisation. In *Proceedings of PRICAI*, pages 42–56, 2018a.

Anil Ramachandran, Sunil Gupta, Santu Rana, and Svetha Venkatesh. Information-theoretic transfer learning framework for Bayesian optimisation. In *Proceedings of ECMLPKDD*, 2018b.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *Proceedings of ICLR*, 2017.

Scott Reed, Yutian Chen, Thomas Paine, Aäron van den Oord, SM Eslami, Danilo Rezende, Oriol Vinyals, and Nando de Freitas. Few-shot autoregressive density estimation: Towards learning to learn distributions. *arXiv preprint arXiv:1710.10304*, 2017.

Matthias Reif, Faisal Shafait, and Andreas Dengel. Prediction of classifier training time including parameter optimization. In *Proc. of GfKI*, pages 260 – 271, 2011.

Matthias Reif, Faisal Shafait, and Andreas Dengel. Meta-learning for evolutionary parameter optimization of classifiers. *Machine learning*, 87(3):357–380, 2012.

Matthias Reif, Faisal Shafait, Markus Goldstein, Thomas Breuel, and Andreas Dengel. Automatic classifier selection for non-experts. *Pattern Analysis and Applications*, 17(1): 83 – 96, 2014.

Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv 1803.00676*, 2018.

S Rendle. Factorization machines. In *ICDM 2015*, pages 995–1000, 2010.

Parker Ridd and Christophe Giraud-Carrier. Using metalearning to predict when parameter optimization is likely to improve classification accuracy. In *ECAI Workshop on Metalearning and Algorithm Selection*, pages 18–23, 2014.

A. Rivolli, L.P.F. Garcia, C. Soares, J. Vanschoren, and A.C.P.L.F. de Carvalho. Towards reproducible empirical research in meta-learning. *arXiv preprint*, 1808.10406, 2018.

Herbert Robbins. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*, pages 169–177. Springer, 1985.

M. T. Rosenstein, Z. Marx, and L. P. Kaelbling. To Transfer or Not To Transfer. In *NIPS Workshop on transfer learning*, 2005.

Peter J. Rousseeuw and Mia Hubert. Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):73 – 79, 2011.

Thomas Philip Runarsson and Magnus Thor Jonsson. Evolution and design of distributed learning rules. In *IEEE Symposium on Combinations of Evolutionary Computation and Neural Networks*, pages 59–63, 2000.

Mostafa A. Salama, Aboul Ella Hassanien, and Kenneth Revett. Employment of neural network and rough set in meta-learning. *Memetic Comp.*, 5(3):165–177, 2013.

S. Sanders and C. Giraud-Carrier. Informing the use of hyperparameter optimization through metalearning. In *Proc. ICDM*, pages 1051–1056, 2017.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850, 2016a.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. One-shot learning with memory-augmented neural networks. *arXiv preprint arXiv:1605.06065*, 2016b.

N. Schilling, M. Wistuba, L. Drumond, and L. Schmidt-Thieme. Hyperparameter optimization with factorized multilayer perceptrons. In *Proceedings of ECML PKDD*, pages 87–103, 2015.

Jürgen Schmidhuber. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Comp.*, 4(1):131–139, 1992.

Jürgen Schmidhuber. A neural network that embeds its own meta-levels. In *Proceedings of ICNN*, pages 407–412, 1993.

Jürgen Schmidhuber, Jieyu Zhao, and Marco Wiering. Shifting inductive bias with success-story algorithm, adaptive levin search, and incremental self-improvement. *Machine Learning*, 28(1):105–130, 1997.

B. Schoenfeld, C. Giraud-Carrier, M. Poggeman, J. Christensen, and K. Seppi. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *AutoML Workshop at ICML*, 2018.

F. Serban, J. Vanschoren, J.U. Kietz, and A.A Bernstein. A survey of intelligent assistants for data analysis. *ACM Computing Surveys*, 45(3):Art.31, 2013.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of CVPR 2014*, pages 806–813, 2014.

N. E. Sharkey and A. J. C. Sharkey. Adaptive Generalization. *Artificial Intelligence Review*, 7:313–328, 1993.

Kate A. Smith-Miles. Cross-disciplinary perspectives on meta-learning for algorithm selection. *ACM Computing Surveys*, 41(1):1 – 25, 2009.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Neural Information Processing Systems*, pages 4077–4087, 2017.

C Soares, J Petrak, and P Brazdil. Sampling based relative landmarks: Systematically testdriving algorithms before choosing. *Lecture Notes in Computer Science*, 3201:250–261, 2001.

C. Soares, P. Brazdil, and P. Kuba. A meta-learning method to select the kernel width in support vector regression. *Mach. Learn.*, 54:195–209, 2004.

C Soares, T Ludermir, and F De Carvalho. An analysis of meta-learning techniques for ranking clustering algorithms applied to artificial data. *Lecture Notes in Computer Science*, 5768:131–140, 2009.

J. Springenberg, A. Klein, S. Falkner, and Frank Hutter. Bayesian optimization with robust Bayesian neural networks. In *Advances in Neural Information Processing Systems*, 2016.

David H Stern, Horst Samulowitz, Ralf Herbrich, Thore Graepel, Luca Pulina, and Armando Tacchella. Collaborative expert portfolio management. In *Proceedings of AAAI*, pages 179–184, 2010.

Benjamin Strang, Peter van der Putten, Jan N. van Rijn, and Frank Hutter. Don't Rule Out Simple Models Prematurely. In *Adv. in Intelligent Data Analysis*, 2018.

Q. Sun, B. Pfahringer, and M. Mayo. Towards a Framework for Designing Full Model Selection and Optimization Systems. In *International Workshop on Multiple Classifier Systems*, pages 259–270, 2013.

Quan Sun and Bernhard Pfahringer. Pairwise meta-rules for better meta-learning-based algorithm ranking. *Machine Learning*, 93(1):141–161, 2013.

Kevin Swersky, Jasper Snoek, and Ryan P Adams. Multi-task Bayesian optimization. In *Adv. in neural information processing systems*, pages 2004–2012, 2013.

Kevin Swersky, Jasper Snoek, and Ryan Prescott Adams. Freeze-thaw bayesian optimization. *arXiv preprint arXiv:1406.3896*, 2014.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

S. Thrun. Lifelong Learning Algorithms. In *Learning to Learn*, chapter 8, pages 181–209. Kluwer Academic Publishers, MA, 1998.

S. Thrun and T. Mitchell. Learning One More Thing. In *Proceedings of IJCAI*, pages 1217–1223, 1995.

S. Thrun and L. Pratt. Learning to Learn: Introduction and Overview. In *Learning to Learn*, pages 3–17. Kluwer, 1998.

L Todorovski and S Dzeroski. Experiments in meta-level learning with ILP. *Lecture Notes in Computer Science*, 1704:98–106, 1999.

L Todorovski, P Brazdil, and C Soares. Report on the experiments with feature selection in meta-level learning. *PKDD 2000 Workshop on Data mining, Decision support, Meta-learning and ILP*, pages 27–39, 2000.

L. Todorovski, H. Blockeel, and S. Džeroski. Ranking with predictive clustering trees. *Lecture Notes in Artificial Intelligence*, 2430:444–455, 2002.

J. van Rijn, S. Abdulrahman, P. Brazdil, and J. Vanschoren. Fast Algorithm Selection Using Learning Curves. In *Proceedings of IDA*, 2015.

J. van Rijn, G. Holmes, B. Pfahringer, and J. Vanschoren. The Online Performance Estimation Framework. Heterogeneous Ensemble Learning for Data Streams. *Machine Learning*, 107:149–176, 2018.

J. N. van Rijn and Frank Hutter. Hyperparameter importance across datasets. In *Proceedings of KDD*, pages 2367–2376, 2018.

Jan N van Rijn, Geoffrey Holmes, Bernhard Pfahringer, and Joaquin Vanschoren. Algorithm selection on data streams. In *Discovery Science*, pages 325–336, 2014.

J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo. OpenML: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.

Joaquin Vanschoren. *Understanding Machine Learning Performance with Experiment Databases.* PhD thesis, Leuven Univeristy, 2010.

Joaquin Vanschoren, Hendrik Blockeel, Bernhard Pfahringer, and Geoffrey Holmes. Experiment databases. *Machine Learning*, 87(2):127–158, 2012.

Manasi Vartak, Arvind Thiagarajan, Conrado Miranda, Jeshua Bratman, and Hugo Larochelle. A meta-learning perspective on cold-start recommendations for items. In *Advances in Neural Information Processing Systems*, pages 6904–6914, 2017.

R Vilalta. Understanding accuracy performance through concept characterization and algorithm analysis. *ICML Workshop on Recent Advances in Meta-Learning and Future Work*, 1999.

R Vilalta and Y Drissi. A characterization of difficult problems in classification. *Proceedings of ICMLA*, 2002.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016.

Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

H. Weerts, M. Meuller, and J. Vanschoren. Importance of tuning hyperparameters of machine learning algorithms. Technical report, TU Eindhoven, 2018.

Marcel Wever, Felix Mohr, and Eyke Hüllermeier. Ml-plan for unlimited-length machine learning pipelines. In *AutoML Workshop at ICML 2018*, 2018.

M. Wistuba, N. Schilling, and L. Schmidt-Thieme. Learning hyperparameter optimization initializations. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, 2015a.

M. Wistuba, N. Schilling, and L. Schmidt-Thieme. Hyperparameter search space pruning, a new component for sequential model-based hyperparameter optimization. In *ECML PKDD 2015*, pages 104–119, 2015b.

Martin Wistuba, Nicolas Schilling, and Lars Schmidt-Thieme. Scalable Gaussian process-based transfer surrogates for hyperparameter optimization. *Machine Learning*, 107(1): 43–78, 2018.

D.H. Wolpert and W.G. Macready. No free lunch theorems for search. Technical Report SFI-TR-95-02-010, The Santa Fe Institute, 1996.

C. Yang, Y. Akimoto, D.W Kim, and M. Udell. Oboe: Collaborative filtering for automl initialization. *arXiv preprint arXiv:1808.03233*, 2018.

Dani Yogatama and Gideon Mann. Efficient transfer learning method for automatic hyperparameter tuning. In *AI and Statistics*, pages 1077–1085, 2014.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.