Algorithm Selector and Prescheduler in the ICON challenge

François Gonard, Marc Schoenauer, Michèle Sebag

Algorithm Selector and Prescheduler in the ICON challenge

François Gonard[1,2,3], Marc Schoenauer[2,3,1], and Michèle Sebag[3,2]

[1] Technological Research Institute SystemX**, 8 avenue de la Vauve 91127 Palaiseau cedex France
[2] INRIA
[3] LRI, CNRS UMR8623 and Université Paris-Sud, Bat. 660 Claude Shannon 91405 Orsay cedex France

Abstract. Algorithm portfolios are known to offer robust performances, efficiently overcoming the weakness of every single algorithm on some particular problem instances. Two complementary approaches to get the best out of an algorithm portfolio is to achieve algorithm selection (AS), and to define a scheduler, sequentially launching a few algorithms on a limited computational budget each. The presented *Algorithm Selector And Prescheduler* system relies on the joint optimization of a pre-scheduler and a *per instance* AS, selecting an algorithm well-suited to the problem instance at hand. ASAP has been thoroughly evaluated against the state-of-the-art during the ICON challenge for algorithm selection, receiving an honourable mention. Its evaluation on several combinatorial optimization benchmarks exposes surprisingly good results of the simple heuristics used; some extensions thereof are presented and discussed in the paper.

1 Introduction

In quite a few domains related to combinatorial optimization, such as satisfiability, constraint solving or operations research, it has been acknowledged for some decades that there exists no universal algorithm, dominating all other algorithms on all problem instances. This result, referred to as No Free Lunch theorem [20], has prompted the scientific community to design algorithm portfolios addressing the various types of difficulties involved in the problem instances, *i.e.*, such that at least one algorithm in the portfolio can efficiently handle any problem instance [9, 6]. Algorithm portfolios thus raise a new issue, that of selecting *a priori* an algorithm well suited to the application domain [12]. This issue, referred to as *Algorithm Selection* (AS) and first formalized by Rice [18], is key to the successful transfer of algorithms outside of research labs. It has been tackled by a number of authors in the last years (more in section 2).

Algorithm selection comes in different flavors, depending on whether the goal is to yield an optimal performance in expectation with respect to a given distribution of problem instances (global AS), or an optimal performance on a particular problem instance (*per instance* AS). Note that the measure of performance depends on the domain (e.g., time-to-solution in satisfiability, or time to reach the optimal solution up to a given precision in optimization[4]). This paper focuses on the *per-instance* setting, aimed at achieving peak performance on every problem instance.

In some domains, it is often the case that some problems can be solved in no time by some algorithms. It thus makes sense to allocate a part of the computational budget to a *pre-scheduler*, sequentially launching a few algorithms with a small computational budget each. The pre-scheduler is expected to solve "easy" instances in a first stage; in a second stage, AS is only launched on problem instances which have not been solved in the pre-scheduler phase. Note that the pre-scheduler enables to extract some additional information characterizing the problem at hand, which can be used together with the initial information about the problem instance, to support the AS phase.

This paper presents the *Algorithm Selector And Prescheduler* system (ASAP), aimed at algorithm selection in the domain of combinatorial optimization (section 3). The main contribution lies in the joint optimization of both a pre-scheduler and a per-instance algorithm selector. The extensive empirical validation of ASAP is conducted on the ICON challenge on algorithm selection [11]. This challenge leverages the Algorithm Selection library [1], aimed at the fair, comprehensive

** This research work has been funded by the French Program "Investissements d'Avenir".

[4] One often considers the joint problems of selecting an algorithm and the optimal hyper-parameters thereof, referred to as *Algorithm Configuration* (AC), as the choice of the hyper-parameter values governs the algorithm performance. AC is outside the scope of the paper and will not be further considered.

and reproducible benchmarking of AS approaches on 13 domains ranging from satisfiability to operations research (section 4).

The comparative empirical validation of ASAP demonstrates its good performances comparatively to state-of-art pre-schedulers and AS approaches (section 5), and its complementarity with respect to the prominent Zilla algorithms [22]. The paper concludes with a discussion of the limitations of the ASAP approach, and some perspectives for further research.

2 Related work

2.1 Algorithm selectors

The algorithm selection issue, aimed at selecting the algorithm best suited to the problem at hand, was first formalized by Rice [18] as follows. Given a problem space mapping each problem instance onto a description $\mathbf{x}$ thereof (usually $\mathbf{x}$ in $\mathbb{R}^d$) and the set $\mathcal{A}$ of algorithms in the portfolio, let us denote $\mathcal{G}(\mathbf{x}, a)$ a performance model, mapping each ($\mathbf{x}$, a) pair onto the performance of algorithm a onto problem instance $\mathbf{x}$. AS most naturally follows from such a performance model by selecting for each problem instance $\mathbf{x}$ the algorithm a with optimal $\mathcal{G}(\mathbf{x}, a)$.

$$AS(\mathbf{x}) = \arg\max_{a \in \mathcal{A}} \{\mathcal{G}(\mathbf{x}, a)\} \tag{1}$$

The performance model is most usually built by applying machine learning approaches onto a dataset reporting the algorithm performances on a comprehensive set of benchmark problem instances (with the exception of [5], using a multi-armed bandit approach). Such machine learning approaches range from k-nearest neighbors [15] to ridge regression [22], random forests [23], collaborative filtering [19, 14], or learning to rank approaches [16].

As expected, the efficiency of the machine learning approaches critically depends on the quality of the training data: the representativity of the problem instances used to train the performance model, and even more importantly, the description of the problem instances. Considerable care has been devoted to the definition of descriptive features in the SAT and Constraint domains [21].

2.2 Schedulers

Besides AS, an algorithm portfolio can also take advantage of parallel computer architectures, by launching several algorithms working independently or in cooperation on the considered problem instance (see e.g. [24], [10]). Schedulers embed the parallel solving strategies in a sequential computer architecture, by considering a sequence of κ (algorithm a_i, time-out τ_i) pairs, where the problem instance is successively tackled by algorithm a_i with a computational budget τ_i, until being solved. Notably, the famed restart strategy, launching a same algorithm with different random seeds or different initial conditions can be viewed as a particular case of scheduling strategy [6]. Likewise, AS can be viewed as a particular case of scheduler with $\kappa = 1$ and τ_1 set to the overall computational budget.

As shown by [23], schedulers and AS can be combined together along a multi-stage process, where a scheduler solves easy instances in a first stage, and remaining instances are handled by the AS and tackled by the selected algorithm in the next stage. [10] build *per-instance* schedules where the AS is one of the component algorithms incorporated.

3 Overview of ASAP

This section first discusses the rationale for the ASAP approach, before detailing the pre-scheduler and AS modules in ASAP.V1. Extensions thereof, forming ASAP.V2, are presented thereafter.

3.1 Analysis

A benchmark suite most generally involves easy and hard problem instances. The difficulty is that the hardness of a problem instance depends on the considered algorithm. As shown on Fig. 1 in the case of the SAT11-HAND dataset (section 4), while several algorithms might solve 20% of the

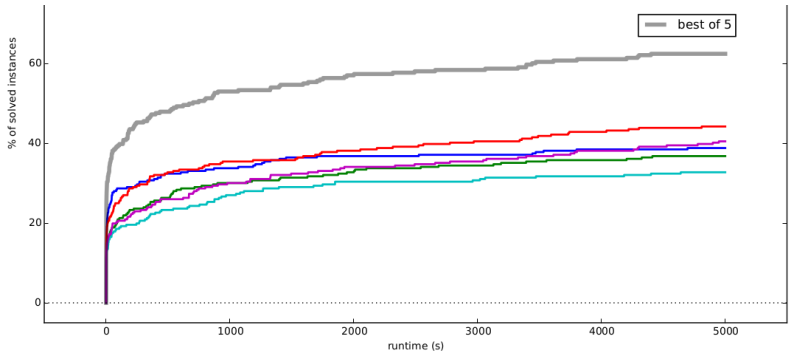

Fig. 1. Percentage of solved instances vs. runtime on the SAT11-HAND dataset, for 5 algorithms and the oracle (selecting the best algorithm out of 5 for each problem instance).

problem instances within seconds, the oracle (selecting the best one out of these algorithms for each problem instance) solves about 40% of the problem instances within seconds.

Accordingly, one might want to launch each one of these algorithms for a few seconds each on each problem instance: after this stage, referred to as pre-scheduler stage, circa 40% of the overall problem instances would be solved.

Definition 1 (Pre-scheduler). *Let $\mathcal{A}$ be a set of algorithms. A κ-component pre-scheduler, defined as a sequence of κ (algorithm a_i, time-out τ_i) pairs,*

$$((a_i, \tau_i)_{i=1}^{\kappa}) \;\; with (a_i, \tau_i) \in \mathcal{A} \times \mathbb{R}^+, \;\; \forall i \in 1, \ldots, \kappa$$

sequentially launches algorithm a_j on any problem instance $\mathbf{x}$ until either a_j solves $\mathbf{x}$, or time τ_j is reached, or a_j stops without solving $\mathbf{x}$. If $\mathbf{x}$ has been solved, the execution stops. Otherwise, j is incremented while $j \leq \kappa$.

Note that a pre-scheduler contributes to reduce the impact of the AS failures. An AS failure is manifested as a problem instance $\mathbf{x}$ for which the AS selects an inappropriate algorithm (requiring much computational resources to solve $\mathbf{x}$ or even failing to solve it), although there exists another algorithm which could have solved $\mathbf{x}$ in no time. Since a pre-scheduler increases the chance for each problem instance to be solved in no time, everything else being equal, it therefore mitigates the chances and impact of AS failures.

After this discussion, the ASAP system involves two modules, a pre-scheduler and an AS. The pre-scheduler is meant to solve as many problem instances as possible in a first stage, and the AS takes care of the remaining problem instances. A primary decision concerns the division of labor between the two modules: how to split the available runtime between the two, and how many algorithms are involved in the pre-scheduler (parameter κ). It is clear that the number of problem instances solved by a module will increase with its computational budget, everything else being equal; the pre-scheduler and the AS modules are interdependent. For simplicity and tractability however, the maximal runtime allocated to the pre-scheduler is fixed to T_{ps}^{max} (10% of the overall computational budget in the experiments, section 5), and the number κ of algorithms in the pre-scheduler is set to 3. [10] and [13] use a most close setup for their fixed-split selection schedules, except they do not constrain the AS component to take the last part of the schedule.

Given T_{ps}^{max} and κ, ASAP tackles the optimization of the pre-scheduler and the AS modules. It is clear that both optimization problems remain inter-dependent: the AS should mostly focus on the problem instances which are not solved by the pre-scheduler, while the pre-scheduler should symmetrically focus on the problem instances which are most uncertain or badly identified by the AS. Formally, this interdependence is handled as follows:

- A performance model $\mathcal{G}(\mathbf{x}, a)$ is built for each algorithm over all training problem instances, defining AS_{init} (Eq. 1);
- A pre-scheduler is built to optimize the joint performance (pre-scheduler, AS_{init}) over all training problem instances;

– Another performance model $\mathcal{G}2(\mathbf{x}, a)$ is built over all training problem instances, using an additional boolean feature that indicates for each problem instance whether it was solved by the above pre-scheduler; let AS_{post} denote the AS based on performance model $\mathcal{G}2(\mathbf{x}, a)$.

ASAP finally is composed of the pre-scheduler followed by AS_{post}.

3.2 ASAP.V1 pre-scheduler

Let $(a_i, \tau_i)_{i=1}^{\kappa}$ denote a pre-scheduler, with overall computational budget $T_{ps} = \sum_{i=1}^{\kappa} \tau_i$, and let $\mathcal{F}((a_i, \tau_i)_{i=1}^{\kappa})$ denote the associated domain-dependent performance. ASAP.V1 considers for simplicity equal time-outs ($a_i = \frac{T_{ps}}{\kappa}, i = 1 \ldots \kappa$). The pre-scheduler is thus obtained by solving the following optimization problem:

$$\max_{T_{ps} \leq T_{ps}^{max}, a_1, \ldots a_\kappa} \left\{ \mathcal{F}\left((a_i, \frac{T_{ps}}{\kappa})_{i=1}^{\kappa} \right) \right\} \tag{2}$$

This mixed optimization problem is tackled in a hierarchical way, determining for each value of T_{ps} the optimal κ-uple of algorithms $a_1 \ldots a_\kappa$. Thanks to both small κ values ($\kappa = 3$ in the experiments) and small number of algorithms (≤ 31 in the ICON challenge, section 4), the optimal κ-uple is determined by exhaustive search conditionally to the T_{ps} value.

The ASAP.V1 pre-scheduler finally relies on the 1-dimensional optimization of the overall computational budget T_{ps} allocated to the pre-scheduler. In all generality, the optimization of T_{ps} is a multi-objective optimization problem, e.g. balancing the overall number of problems solved and the overall computational budget. Multi-objective optimization commonly proceeds by determining the so-called Pareto front, made of non-dominated solutions. In our case, the Pareto front depicts how the performance varies with the overall computational budget, as illustrated on Fig. 2, where the performance is set to the number of solved instances.

In multi-objective decision making [3, 2], the choice of a solution on the Pareto front is tackled using post-optimal techniques [4], including: i) compromise programming, where one wants to find the point the closest to an ideal target in the objective space; ii) aggregation of the objectives into a single one, e.g., using linear combination; or iii) marginal rate of return. The last heuristics consists of identifying the so-called "knees", that is, the points where any small improvement on a given criterion is obtained at the expense of a large decrease on another criterion, defining the so-called marginal rate of return. The vanilla marginal rate of return is however sensitive to strong local discontinuities; for instance, it would select point A in Fig. 2. Therefore, a variant taking into account the global shape of the curve, and measuring the marginal rate of improvement w.r.t. the extreme solutions on the Pareto front is used (e.g., selecting point K instead of point A in Fig.2).

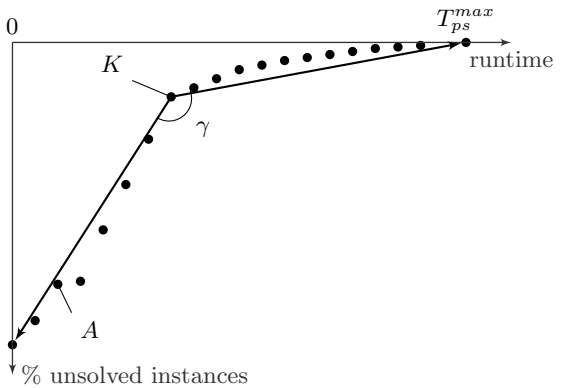

Fig. 2. Among a set of Pareto-optimal solutions, solution A has the best marginal rate of return; solution K, which maximizes the average rate of return w.r.t. the extreme solutions of the Pareto front (maximizing angle γ), is the knee selected in ASAP.

3.3 ASAP.V1 algorithm selector

As detailed in section 3.1, the AS relies on the performance model learned from the training problem instances. Two learning algorithms are considered in this paper: random forests and k-nearest neighbors. One hyper-parameter was adapted for each ML approach (all other hyper-parameters being set to their default value, using the Python scikit-learn library [17]), based on a few preliminary experiments: 35 trees are used for the RandomForest algorithm and the number of neighbors is set to $k = 3$ for the k-nearest neighbors. In the latter case, the predicted value associated to problem instance $\mathbf{x}$ is set to the weighted sum of the performance of its nearest neighbors, weighted by their relative distance to $\mathbf{x}$:

$$\widehat{\mathcal{F}}(\mathbf{x}, a) = \frac{\sum_i ||\mathbf{x} - \mathbf{x}_i|| \mathcal{F}(a, \mathbf{x}_i)}{\sum_i ||\mathbf{x} - \mathbf{x}_i||}$$

where $\mathbf{x}_i$ ranges over the 3 nearest neighbors of $\mathbf{x}$. Features were normalized (zero mean, unit variance) before selecting the neighbors.

A main difficulty comes from the descriptive features forming the representation of problem instances. Typically, the feature values are missing for some groups of features, for quite a few problem instances, due to diverse causes (computation exceeded time limit, exceeded memory, presolved the instance, crashed, other, unknown). The lack of feature value is handled by i) replacing the missing value by the feature average value; ii) adding to the set of descriptive features 7 additional boolean features per group of initial features, indicating whether the feature group values are available or the reason why they are missing otherwise[5] .

3.4 ASAP.V2

Several extensions of ASAP.V1 have been considered after the closing of the ICON challenge, aimed at exploring a richer pre-scheduler-AS search space while preventing the risk of overfitting induced by a larger search space.

We investigated the use of different time-outs for each algorithm in the pre-scheduler, while keeping the set of algorithms $(a_1, \ldots, a_\kappa)$ and the overall computational budget T_{ps}. The sequential optimization strategy (section 3.2), deterministically selecting T_{ps} as the solution with maximal average return rate, exhaustively determining the κ-uple of algorithms conditionally to T_{ps}, is thus extended to optimize the $(\tau_1, \ldots \tau_{\kappa-1})$ vector conditionally to $\sum_{i=1}^{\kappa-1} \leq T_{ps}$, using a prominent continuous black-box optimizer, specifically the Covariance-Matrix Adaptation-Evolution Strategy (CMA-ES) [7].

This extended search space is first investigated by considering the **raw optimization criterion** $\mathcal{F}_{raw}\left((a_i, \tau_i)_{i=1}^{\kappa}\right)$ defined in section 3.2, that is, the cumulative performance of ASAP over all training problem instances. However a richer search space entails some risk of overfitting, where the higher performance on data used to optimize ASAP (training data) is obtained at the expense of a lower performance on test data. Generally speaking, the datasets used to train an AS are small ones.

A **penalized optimization criterion** is thus considered:

$$\mathcal{F}_{L2}\left((a_i, \tau_i)_{i=1}^{\kappa}\right) = \mathcal{F}\left((a_i, \tau_i)_{i=1}^{\kappa}\right) + w \sum_{i=1}^{\kappa} \left(\tau_i - \frac{T_{ps}}{\kappa}\right)^2$$

which penalizes the L_2 distance between the (τ_i) vector and the uniform time outs $(\tau_i = \frac{T_{ps}}{\kappa})$. The rationale for this penalization is to prevent brittle improvements on the training set due to opportunistic adjustments of the τ_is, at the expense of stable performances on further instances. The penalization weight w is adjusted using a nested CV process.

A **randomized optimization criterion** is also considered. By construction, the fitness function aggregates the performances of all training problem instances. As the training problem instances sample the problem domain, this fitness defines a noisy optimization problem. Sophisticated

[5] The increase in the overall number of features is handled by an embedded feature selection mechanism, removing all features with negligible importance criterion ($< 10^{-5}$ in the experiments) in a separately learned 10-trees random forest regression model.

approaches have been proposed to address this noisy optimization issue in non-convex optimization-based machine learning settings (see e.g. [8]). Another approach is proposed here, based on the bootstrap principle: in each CMA-ES generation, the set of n problem instances used to compute the performance is uniformly drawn with replacement from the n-size training set. In this manner, each optimization generation considers a slightly different optimization objective noted $\mathcal{F}_{rand}$, thereby discarding the insignificant improvements.

Finally, a **probabilistic optimization criterion** is considered, handling the ASAP performance on a single problem instance as a random variable with a triangle-shape distribution (Fig. 3) centered on the actual performance $p(\mathbf{x})$, with support in $[p(\mathbf{x}) - \theta, p(\mathbf{x}) + \theta]$, and taking the expectation thereof. The merit of this triangular probability distribution function is to allow for an analytical computation of the overall fitness expectation, noted $\mathcal{F}_{dfp}$.

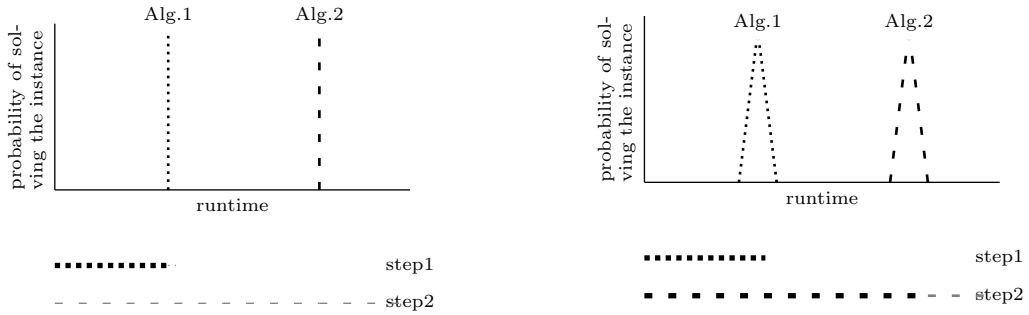

Fig. 3. Schedule execution difference between punctual and triangular pdf. On the left (punctual pdf), schedule stops during step 1. On the right (triangular pdf, part of the instance is not solved during step 1 and the schedule executes until step 2 solves the rest.

4 Experimental setting: The ICON challenge

4.1 ASlib data format

Due to the difficulty of comparing the many algorithm selection systems and the high entry ticket to the AS field, a joint effort was undertaken to build the Algorithm Selection Library (ASlib), providing comprehensive resources to facilitate the design, sharing and comparison of AS systems [1]. ASlib (version 1.0.1) involves 13 datasets, also called scenarios (Table 1), gathered from recent challenges and surveys in the operations research, artificial intelligence and optimization fields. The interested reader is referred to [1] for a more comprehensive presentation.

Table 1. ASlib datasets (V1.0.1)

dataset	# instances	# algorithms	# features
ASP-POTASSCO	1294	11	138
CSP-2010	2024	2	86
MAXSAT12-PMS	876	6	37
PREMARSHALLING-ASTAR-2013	527	4	16
PROTEUS-2014	4021	22	198
QBF-2011	1368	5	46
SAT11-HAND	296	15	115
SAT11-INDU	300	18	115
SAT11-RAND	600	9	115
SAT12-ALL	1614	31	115
SAT12-HAND	767	31	115
SAT12-INDU	1167	31	115
SAT12-RAND	1362	31	115

Each dataset includes i) the performance and computation status of each algorithm on each problem instance; ii) the description of each problem instance, as a vector of the expert-designed feature values (as said, this description considerably facilitates the comparison of the AS systems); iii) the computational status of each such feature (e.g. indicating whether the feature could be computed, or if it failed due to insufficient computational or memory resources). Last but not least, each dataset is equi-partitioned into 10 subsets, to enforce the reproducibility of the 10 fold CV assessment of every AS algorithm.

4.2 The ICON Challenge on Algorithm Selection

The ICON Challenge on Algorithm Selection, within the ASlib framework, was carried on between February and July 2015 to evaluate AS systems in a fair, comprehensive and reproducible manner[6]. Each submitted system was assessed on the 13 ASlib datasets [1] with respect to three measures: i) number of problem instances solved; ii) extra runtime compared with the virtual best solver (VBS, also called oracle); and iii) Penalized Average Time-10 (PAR10) which is the cumulative runtime needed to solve all problem instances (set to ten times the overall computational budget whenever the problem instance is unsolved).

As the whole datasets were available to the community from the start, the evaluation was based on hidden splits between training and test set. Each submitted system provides a dataset-dependent, instance-dependent schedule of algorithms, optionally preceded by a dataset-dependent presolver (single algorithm running on all instances during a given runtime before the per-instance schedule runs). Each system can also, in a dataset-dependent manner, specify the groups of features to be used (in order to save the time needed to compute useless features).

Two baselines are considered: the oracle, selecting the best algorithm for each problem instance; and the single best (SB) algorithm, with best average performance over all problem instances in the dataset. The baselines are used to normalize every system performance over all datasets, associating performance 0 to the oracle (respectively performance 1 to the single best), supporting the aggregation of the system results over all datasets.

5 Experimental validation

5.1 Comparative results

Table 2 reports the results of all submitted systems on all datasets (the statistical significance tests are reported in Fig. 5). The general trend is that zilla algorithms dominate all other algorithms on the SAT datasets, as expected since they have consistently dominated the SAT contests in the last decade. On non-SAT problems however, zilla algorithms are dominated by ASAP_RF.V1.

The robustness of the ASAP approach is demonstrated as they never rank last; they however perform slightly worse than the single best on some datasets. The rescaled performances of ASAP_RF.V1 is compared to zilla and autofolio (Fig. 4, on the left), demonstrating that ASAP_RF.V1 offers a balanced performance, significantly lower than for zilla and autofolio on the SAT problems, but significantly higher on the other datasets; in this respect it can be viewed as a low-risk system.

5.2 Sensitivity analysis

The sensitivity analysis conducted after the closing of the challenge compares ASAP.V2 (with different time-outs in the pre-scheduler) and ASAP.V1, and examines the impact of the different optimization criteria, aimed at avoiding overfitting: the raw fitness, the L2-penalized fitness, the randomized fitness and the probabilistic fitness (section 3.4).

The impact of the hyper-parameters used in the AS (number of trees set to 35, 100, 200, 300 and 500 trees in the Random Forest) is also investigated.

Table 3 summarizes the experimental results that each ASAP.V2 configuration would have obtained in the ICON challenge framework, together with the actual submissions results, including

[6] The codes of all submitted systems and the results are publicly available, http://challenge.icon-fet.eu/challengeas

Table 2. Normalized performances of submitted systems, aggregated across all folds and all measures (the lower, the better). Ranks of zilla (challenge winner) and ASAP_RF.V1 (honourable mention) are given in parenthesis. Numbers were computed from the challenge outputs.

	ASAP_RF.V1	ASAP_kNN.V1	autofolio	flexfolio	sunny	sunny-presolv	zilla	zillafolio
ASP-POTASSCO	0.294 (2)	0.359	0.299	0.314	0.37	0.336	0.319 (5)	**0.283**
CSP-2010	**0.146** (1)	0.247	0.288	0.223	0.263	0.406	0.2 (3)	0.157
MAXSAT12-PMS	0.168 (4)	0.159	0.45	**0.149**	0.166	0.224	0.201 (5)	0.233
PREMARSHALLING-ASTAR-2013	0.349 (4)	0.369	0.359	0.307	0.325	**0.296**	0.374 (7)	0.385
PROTEUS-2014	0.16 (4)	0.177	0.222	**0.056**	0.134	0.103	0.245 (8)	0.223
QBF-2011	0.097 (2)	**0.091**	0.169	0.096	0.142	0.162	0.191 (7)	0.194
SAT11-HAND	0.341 (4)	0.318	0.342	0.342	0.466	0.464	0.328 (3)	**0.302**
SAT11-INDU	1.036 (5)	0.957	**0.875**	1.144	1.13	1.236	0.905 (2)	0.966
SAT11-RAND	0.104 (6)	0.09	**0.046**	0.226	0.116	0.088	0.053 (2)	0.067
SAT12-ALL	0.392 (5)	0.383	0.306	0.502	0.509	0.532	**0.273** (1)	0.322
SAT12-HAND	0.334 (5)	0.31	**0.256**	0.434	0.45	0.467	0.272 (2)	0.296
SAT12-INDU	0.955 (6)	0.919	0.604	0.884	1.074	1.018	0.618 (3)	**0.594**
SAT12-RAND	1.032 (5)	1.122	0.862	1.073	1.126	0.97	**0.779** (1)	0.79

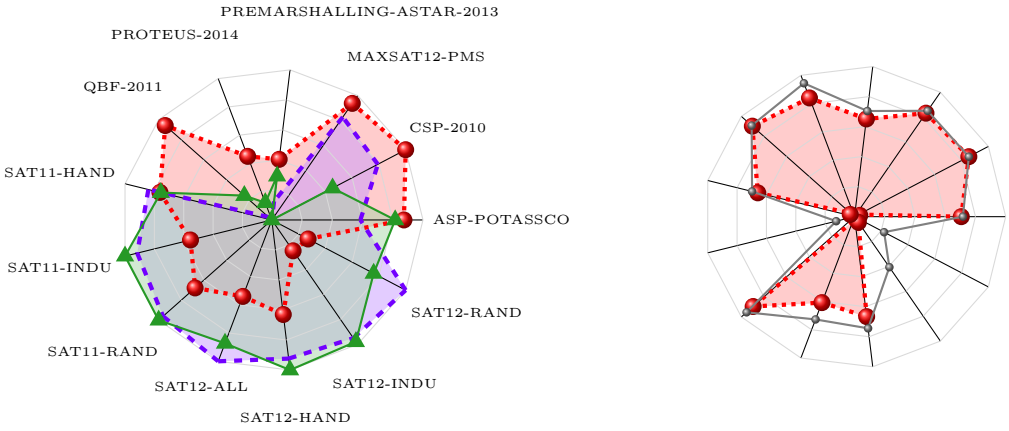

Fig. 4. On the left: per-dataset performances of ASAP_RF.V1 (balls, dotted line), zilla (no marker, dashed line) and autofolio (triangles, solid line) scaled to the range of performance of all submitted systems. As a comparison, the per-dataset best submitted system (small balls, solid line) and ASAP_RF scores before rescaling are depicted on the right.

systems that were not competing in the challenge: llama-regr and llama-regrPairs from the organizers, and autofolio-48 which is identical to autofolio but with 48h time for training (12h was the time limit authorized in the challenge) [11].

The significance analysis, using a Wilcoxon signed-rank test, is reported in Fig. 5. A first result is that all ASAP.V2 variants improve on ASAP.V1 with significance level 1%. A second result is that ASAP.V2 with the probabilistic optimization criterion is not statistically significantly different from zilla, autofolio and zillafolio.

A third and most surprising result is that the difference between the challenge-winner zilla and most of ASAP.V2 variants is not statistically significant.

[8] As the CSP-2010 dataset gives the choice between only two algorithms, the pre-scheduler consists of a single algorithm running for the whole time devoted to the pre-scheduler. For that particular case, all ASAP_RF.V2 variants with the same selector hyperparameter are identical.

Table 3. Optimized pre-scheduler performances[8] aggregated across all datasets, all splits and all measures (the lower, the better). The hyperparameters for f_{L^2} and f_{dfp} were chosen after preliminary experiments using the cross validation provided with ASlib. For each configuration of the selector, the best-evaluated fitness function appears in bold

fitness function (if relevant)	f_{L^2}	f_{dfp}	f_{rand}	f_{raw}	none
ASAP_RF.V2 35	0.416	0.414	0.412	**0.410**	0.414
ASAP_RF.V2 100	0.404	**0.398**	0.405	0.402	0.414
ASAP_RF.V2 200	0.404	0.402	0.402	**0.399**	0.405
ASAP_RF.V2 300	0.399	0.399	0.402	**0.393**	0.405
ASAP_RF.V2 500	0.398	**0.394**	0.398	0.398	0.401
ASAP_RF.V1	0.416				
ASAP_kNN.V1	0.423				
autofolio	0.391				
flexfolio	0.442				
sunny	0.482				
sunny-presolv	0.485				
zilla	0.366				
zillafolio	0.37				
autofolio-48	0.375				
llama-regrPairs	0.395				
llama-regr	0.425				

equivalent to the means over the columns of table 2

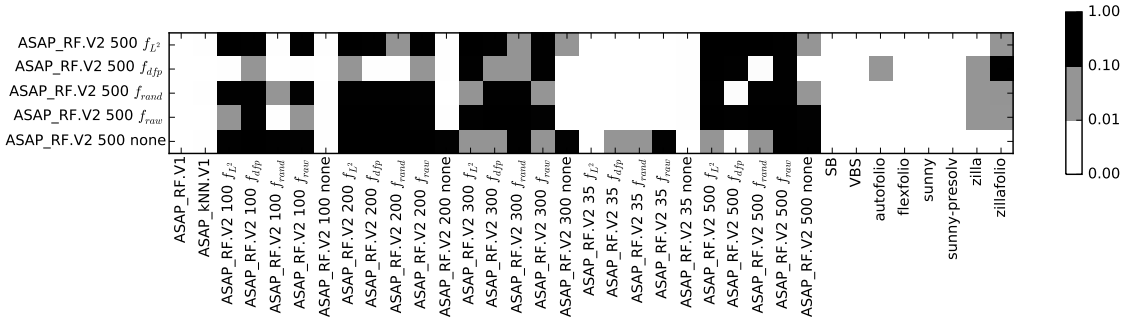

Fig. 5. Wilcoxon signed-rank test p-value between ASAP.V2 variants with 500 trees and every other systems including the single best algorithm (SB) and the virtual best solver (VBS). In particular, the different fitness functions do not differ significantly from each other in most cases.

Fig. 6 details per dataset the performance improvement between ASAP.V2 (500 trees, f_{L^2} version) and ASAP.V2 (500 trees, f_{dfp} version) and on the other hand ASAP.V1_RF (35 trees). Note that ASAP.V2 outperforms the per-dataset best submission to the challenge for 3 datasets.

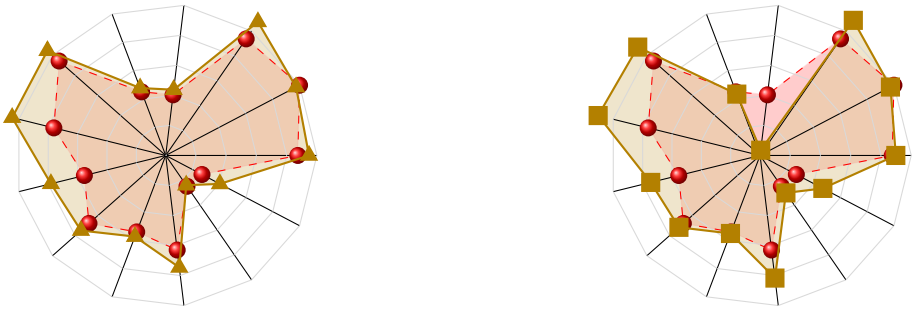

Fig. 6. Per-dataset performances of ASAP_RF.V2 with 500 trees, optimized using f_{L^2} (on the left with triangles) and using f_{dfp} (on the right with squares) scaled to the range of performance of the challenge submitted systems. As a comparison, ASAP_RF.V1 appears with the balls and dashed line.

6 Discussion

A new hybrid algorithm selection approach, the ASAP system was presented in this paper, combining a pre-scheduler and a per-instance algorithm selector. ASAP.V1 introduced a selector learned conditionally to a predetermined schedule so that it focuses on instances that were not solved by the pre-scheduler . ASAP.V2 completes the loop as it re-adapts the schedule to the new AS. The main message is that the scheduler and the AS must be optimized jointly to reflect the division of labor achieved by these two components

ASAP.V1, thoroughly evaluated in the ICON challenge on algorithm selection (ranked 4th) received an honourable mention, due to its novelty and good performance comparatively to the famed and long-known Zilla algorithms.

The ASAP.V2 extension achieved significantly better results along the same challenge setting. It must be emphasized that these results must be comforted by additional experiments on fresh data. A main lesson learned is the importance of the regularization, as the amount of available data does not permit to consider richer AS search spaces without incurring a high risk of overfitting. The probabilistic performance criterion successfully contributed to a more stable optimization problem. Further research work will be devoted to extending this criterion.

References

1. Bischl, B. et al.: ASlib: A Benchmark Library for Algorithm Selection. arXiv:1506.02465v1 (2015)
2. Branke, J., Deb, K., Dierolf, H., Osswald, M.: Finding knees in multi-objective optimization. In: Parallel Problem Solving from Nature-PPSN VIII. pp. 722–731. Springer (2004)
3. Das, I.: On characterizing the "knee" of the pareto curve based on normal-boundary intersection. Structural Optimization 18(2-3), 107–115 (1999)
4. Deb, K.: Multi-objective evolutionary algorithms: Introducing bias among pareto-optimal solutions. In: Advances in evolutionary computing, pp. 263–292. Springer
5. Gagliolo, M., Schmidhuber, J.: Algorithm portfolio selection as a bandit problem with unbounded losses. Annals of Mathematics and Artificial Intelligence 61(2), 49–86 (2011)
6. Gomes, C.P., Selman, B.: Algorithm portfolios. Artificial Intelligence 126(1), 43–62 (2001)
7. Hansen, N., Müller, S.D., Koumoutsakos, P.: Reducing the time complexity of the derandomized Evolution Strategy with Covariance Matrix Adaptation. Evolutionary Computation 11(1), 1–18 (2003)
8. Heidrich-Meisner, V., Igel, C.: Hoeffding and Bernstein races for selecting policies in evolutionary direct policy search. In: Danyluk, A.P. et al. (ed.) ICML. pp. 401–408. ACM Intl Conf. Proc. 382 (2009)
9. Huberman, B.A., Lukose, R.M., Hogg, T.: An economics approach to hard computational problems. Science 275(5296), 51–54 (1997)
10. Kadioglu, S., Malitsky, Y., Sabharwal, A., Samulowitz, H., Sellmann, M.: Algorithm selection and scheduling. In: Proc. 17th CP, pp. 454–469. LNCS 6876, Springer (2011)
11. Kotthoff, L.: ICON challenge on algorithm selection. CoRR abs/1511.04326 (2015)
12. Leyton-Brown, K., Nudelman, E., Andrew, G., McFadden, J., Shoham, Y.: A portfolio approach to algorithm selection. In: Proc. IJCAI. pp. 1542–1543 (2003)
13. Malitsky, Y., Sabharwal, A., Samulowitz, H., Sellmann, M.: Algorithm portfolios based on cost-sensitive hierarchical clustering. In: Proc. 23rd IJCAI. pp. 608–614. AAAI Press (2013)
14. Mısır, M., Sebag, M.: Algorithm selection as a collaborative filtering problem. Tech. rep., INRIA (2013)
15. O´Mahony, E., Hebrard, E., Holland, A., Nugent, C., O´Sullivan, B.: Using case-based reasoning in an algorithm portfolio for constraint solving. In: Proc. ICAICS. pp. 210–216 (2008)
16. Oentaryo, R.J., Handoko, S.D., Lau, H.C.: Algorithm selection via ranking. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI) (2015)
17. Pedregosa, F. et al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
18. Rice, J.R.: The algorithm selection problem. Advances in Computers 15, 65–118 (1976)
19. Stern, D., Herbrich, R., Graepel, T., Samulowitz, H., Pulina, L., Tacchella, A.: Collaborative expert portfolio management. In: Proc. 24th AAAI. pp. 179–184 (2010)
20. Wolpert, D., Macready, W.: No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation 1(1), 67–82 (1997)
21. Xu, L., Hutter, F., Hoos, H., Leyton-Brown, K.: Features for SAT (2012), university of British Columbia
22. Xu, L., Hutter, F., Hoos, H.H., Leyton-Brown, K.: Satzilla: portfolio-based algorithm selection for sat. Journal of Artificial Intelligence Research pp. 565–606 (2008)
23. Xu, L., Hutter, F., Shen, J., Hoos, H.H., Leyton-Brown, K.: Satzilla2012: improved algorithm selection based on cost-sensitive classification models. Balint et al.(Balint et al., 2012a) pp. 57–58 (2012)
24. Yun, X., Epstein, S.L.: Learning algorithm portfolios for parallel execution. In: Hamadi, Y., Schoenauer, M. (eds.) Proc LION 6, pp. 323–338. LNCS 7219, Springer (2012)