ICON Challenge on Algorithm Selection

`http://challenge.icon-fet.eu`

Lars Kotthoff

July 30, 2015

1 Submissions

The challenge received a total of 8 submissions from 4 different groups of researchers comprising 15 people. Participants were based in 4 different countries on 2 continents. Table 1 gives an overview of all submissions.

All submissions were submitted for evaluation on all ASlib scenarios. Most systems used a presolver and specified a subset of features to use for each scenario.

2 Evaluation

The evaluation was performed as follows. For each scenario, 10 bootstrap samplings of the entire data were used to create 10 different train/test splits. No stratification was used. The training part was left unmodified. For the test part, algorithm performances were set to 0 and runstatus to "ok" for all algorithms and all instances – the ASlib specification requires algorithm performance data to be part of a scenario. A `cv.arff` file was generated for both training and

System name	Presolving/feature selection used?
ASAP_kNN	no
ASAP_RF	no
autofolio	yes
flexfolio	yes
sunny	no
sunny-presolv	yes
zilla	yes
zillafolio	yes

Table 1: Systems submitted to the challenge.

testing with 10 folds and the instances assigned to folds by the order in which they appeared in the original scenario.

For systems that specified a presolver, the instances that were solved by the presolver within the specified time were removed from the training set. If a subset of features was specified, only these features (and only the costs associated with these features) were left in both training and test set, with all other feature values removed.

Each system was trained on each train scenario and predicted on each test scenario. In total, 130 evaluations (10 for each of the 13 scenarios) per submitted system were performed. The total CPU time spent was 1965.03 hours.

The predictions were evaluated as follows. If a presolver was specified, it was "run" for the specified time. If the instance was solved within this time, the time to solve the instance was taken as the performance on that instance and the instance recorded as solved.

Otherwise, the time limit given for the presolving run was added to the time required to compute all features specified for the particular scenario and the solvers specified in the prediction schedule of the system were "run". For each instance, the predicted solvers were ordered by the `runID` specified. If a run was unable to solve an instance, the smaller of time the schedule specified to run it for and the time it actually took to run on the instance was added to the total. If a run solved the respective instance, the actual time required by the algorithm was added to the total and the instance recorded as solved. If the total time exceeded the time limit for the scenario, an instance was recorded as not solved.

Each system was evaluated in terms of mean PAR10 score, mean misclassification penalty, and mean number of instances solved for each of the 130 evaluations on each scenario and split.

To facilitate comparison of the different measures across the different scenarios, all measures were normalised by the performance of the virtual best (VBS) and the single best (SB) solver. The single best solver was determined as the solver with the smallest overall runtime across all instances. Equation 1 defines the normalisation of a score s.

$$s_{norm} = \frac{s - s_{VBS}}{s_{SB} - s_{VBS}} \tag{1}$$

This normalises the score to the interval 0 (VBS) to 1 (SB), with smaller values being better. The number denotes how much of the gap between single best and virtual best solver was left by the system.

To determine the overall winner, the mean across all of the normalised measurements was taken. For each submitted system, 390 scores were taken into account for this (13 scenarios times 10 splits times 3 measures).

3 Results

Table 2 shows the final ranking. The first and second placed entries are very close. All systems perform well on average, closing more than half of the gap between virtual and single best solver.

	System	Average total score
1	zilla	0.38953
2	autofolio	0.38980
3	zillafolio	0.39572
4	ASAP_RF	0.41169
5	ASAP_kNN	0.43386
6	flexfolio	0.45371
7	sunny-presolv	0.48956
8	sunny	0.48998

Table 2: Final ranking.

To assess how significant the difference are and how stable the ranking is, we took 1000 bootstrap samples from the scenario-split combinations and computed the scores and ranks on each of them. The mean average of the total score averages over the bootstrap samples and the confidence intervals are show in Table 3.

	System	Average total score	95% CI upper	95% CI lower
1	zilla	0.38902	0.39013	0.38791
2	autofolio	0.38946	0.39057	0.38836
3	zillafolio	0.39528	0.39647	0.39409
4	ASAP_RF	0.41121	0.41260	0.40981
5	ASAP_kNN	0.43338	0.43483	0.43194
6	flexfolio	0.45372	0.45529	0.45215
7	sunny-presolv	0.48966	0.49121	0.48810
8	sunny	0.48974	0.49131	0.48818

Table 3: Final ranking, bootstrapped.

The ranking is the same as the final ranking in Table 2, but the confidence intervals make clear that it is not stable – the confidence intervals of the systems ranked at the top overlap.

3.1 Winner – zilla

The winner of the ICON Challenge on Algorithm Selection is zilla by Chris Cameron, Alex Fréchette, Holger Hoos, Frank Hutter, and Kevin Leyton-Brown.

3.2 Honourable mention – ASAP_RF

ASAP_RF by François Gonard, Marc Schoenauer, and Michèle Sebag receives an honourable mention as a submission that has not been described in the literature before and showed respectable performance, beating all other approaches in some cases.

3.3 Alternative rank aggregations

An alternative (and probably fairer) way of determining the winner is to see the ranking of systems induced by each measure on each split of each scenario as a ballot (for a total of 260 ballots) and aggregate the ranks in those ballots. Here, we optimise the aggregated Spearman coefficient between candidate rankings and ballot rankings. That is, the final ranking has the optimal Spearman coefficient with respect to the ballots.

Table 4 shows the aggregated ranks. The ranking is the same except for the last positions, where sunny and sunny-presolv swap places.

	System
1	zilla
2	autofolio
3	zillafolio
4	ASAP_RF
5	ASAP_kNN
6	flexfolio
7	sunny
8	sunny-presolv

Table 4: Aggregated ranks.

There are significant changes however when averaging the performance across all measures, splits, and scenarios by median rather than mean. Table 5 shows this ranking. Zilla is now only in third position, beat by ASAP_RF and autofolio.

3.4 Detailed results

Tables 6 through 8 show the rankings by mean score across all splits and scenarios, but separately for each measure.

	System	Median total score
1	autofolio	0.30386
2	ASAP_RF	0.30740
3	zilla	0.31177
4	zillafolio	0.31906
5	ASAP_kNN	0.32238
6	flexfolio	0.33231
7	sunny	0.37875
8	sunny-presolv	0.41260

Table 5: Ranking by median.

	System	Mean PAR10 score
1	autofolio	0.34414
2	zillafolio	0.35194
3	zilla	0.36526
4	ASAP_RF	0.37047
5	ASAP_kNN	0.39599
6	flexfolio	0.40488
7	sunny	0.46654
8	sunny-presolv	0.46933

Table 6: Ranking by PAR10.

	System	Mean misclassification penalty
1	zilla	0.44725
2	autofolio	0.49968
3	zillafolio	0.50098
4	ASAP_RF	0.51172
5	ASAP_kNN	0.52580
6	sunny-presolv	0.53698
7	sunny	0.54565
8	flexfolio	0.57033

Table 7: Ranking by misclassification penalty.

	System	Mean number of instances solved
1	autofolio	0.32557
2	zillafolio	0.33425
3	ASAP_RF	0.35288
4	zilla	0.35608
5	ASAP_kNN	0.37979
6	flexfolio	0.38593
7	sunny	0.45775
8	sunny-presolv	0.46236

Table 8: Ranking by number of instances solved.

Table 9 shows the ranks for the different scenarios for all systems by mean across all measures and splits. Zilla ranks first most often, followed by ASAP_RF and flexfolio. However, zilla ranks last on four scenarios whereas ASAP_RF is never last.

scenario	ASAP_kNN	ASAP_RF	autofolio	flexfolio	sunny	sunny-presolv	zilla	zillafolio
ASP-POTASSCO	6	**1**	2	4	7	5	**8**	3
CSP-2010	4	**1**	7	3	6	**8**	2	5
MAXSAT12-PMS	2	4	7	**1**	3	6	5	**8**
PREMARSHALLING-ASTAR-2013	6	4	5	2	3	**1**	**8**	7
PROTEUS-2014	5	4	6	**1**	3	2	**8**	7
QBF-2011	3	**1**	6	2	4	5	**8**	7
SAT11-HAND	3	2	6	4	**8**	7	**1**	5
SAT11-INDU	5	4	**1**	7	6	**8**	2	3
SAT11-RAND	4	6	2	**8**	7	5	3	**1**
SAT12-ALL	4	5	3	6	7	**8**	**1**	2
SAT12-HAND	4	5	3	6	7	**8**	**1**	2
SAT12-INDU	5	6	2	4	**8**	7	**1**	3
SAT12-RAND	**8**	5	3	6	7	4	**1**	2

Table 9: Ranks by scenario.

Figures 1 through 3 give a more detailed overview of the performance of the systems on the different scenarios. The colour of each boxplot denotes the system, the mean performance of which is shown in the legend (this corresponds to the number in the respective table above). The boxplot shows the variation of performance across the 10 different splits for each scenario. The solid black line denotes the performance of the single best solver; anything above is worse.

Two of the SAT scenarios are hard for all systems in the sense that the performance they deliver on at least one of the splits is worse than the performance of the single best solver. For most other scenarios, using any algorithm selection system gives a significant performance improvement compared to the single best solver though.

3.5 Time required to run

The time required to train the models and make the predictions varied significantly across systems and scenarios, with some completing in minutes and others requiring hours. Figure 4 presents a summary.

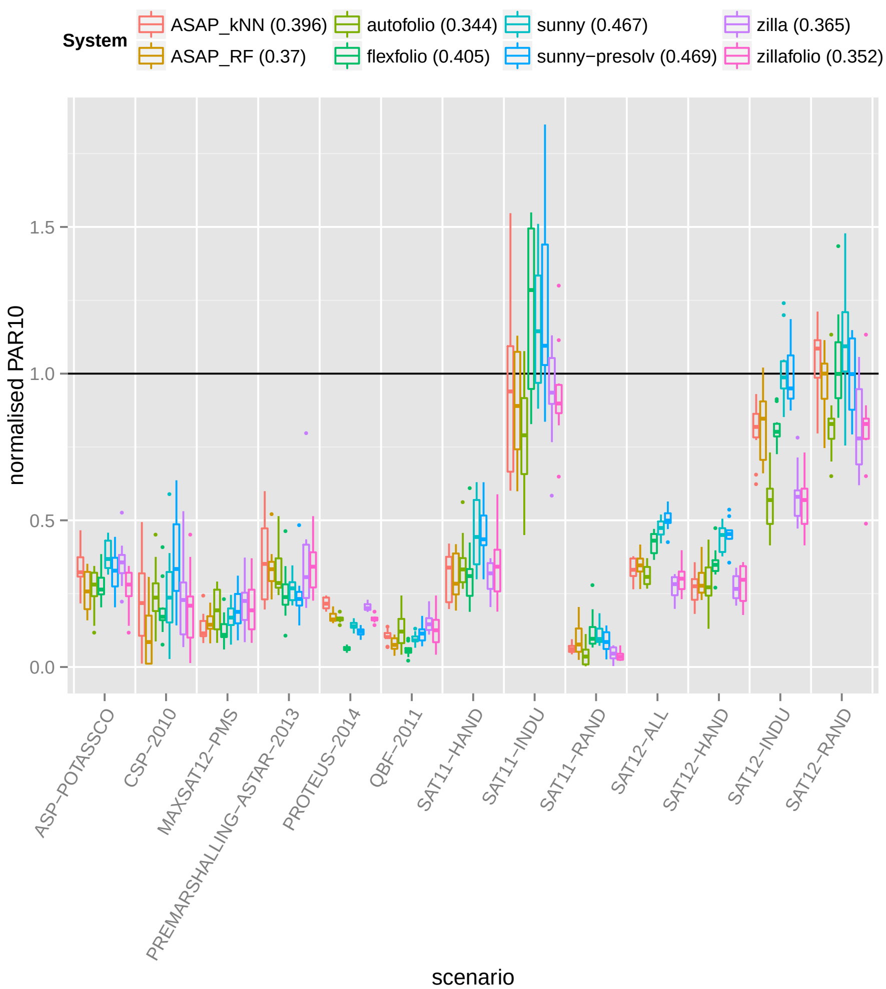

Figure 1: PAR10 scores.

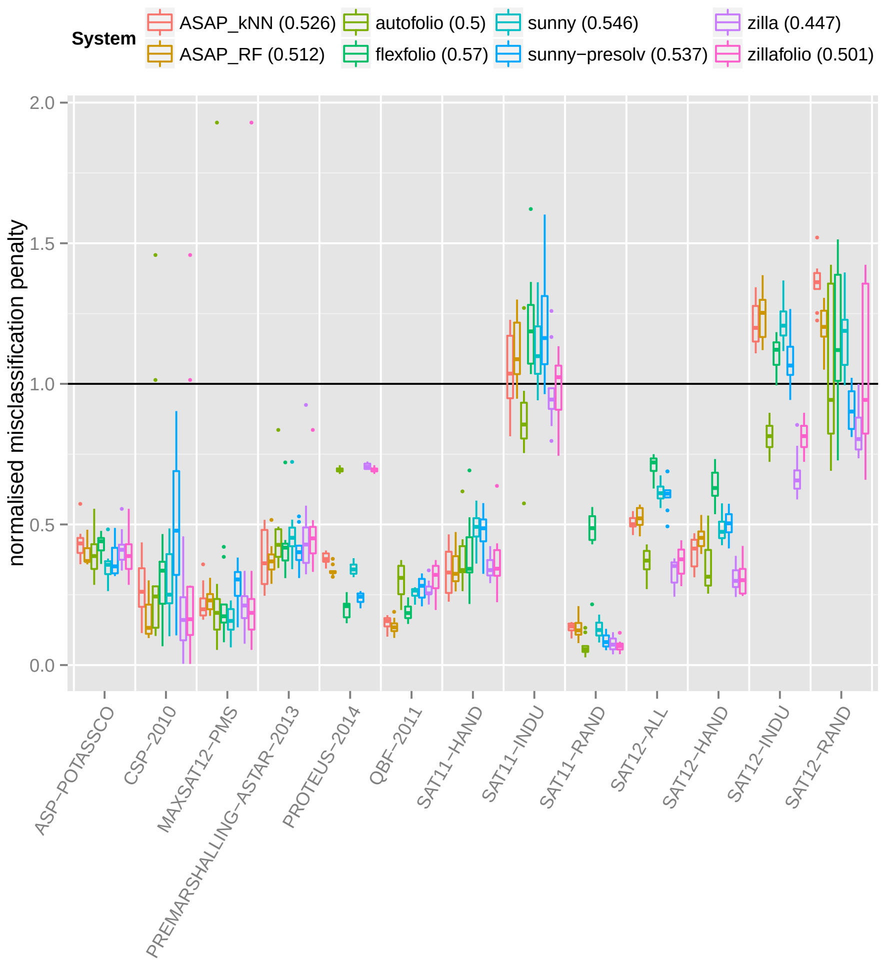

Figure 2: Misclassification penalty.

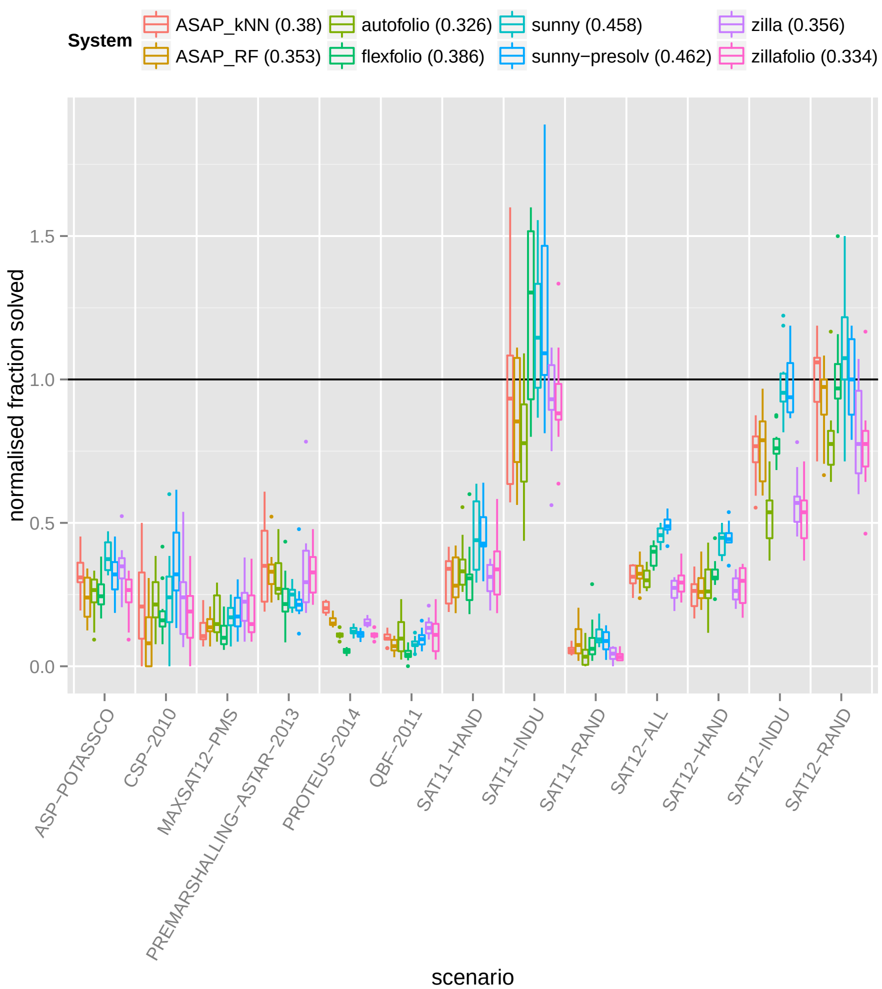

Figure 3: Instances solved.

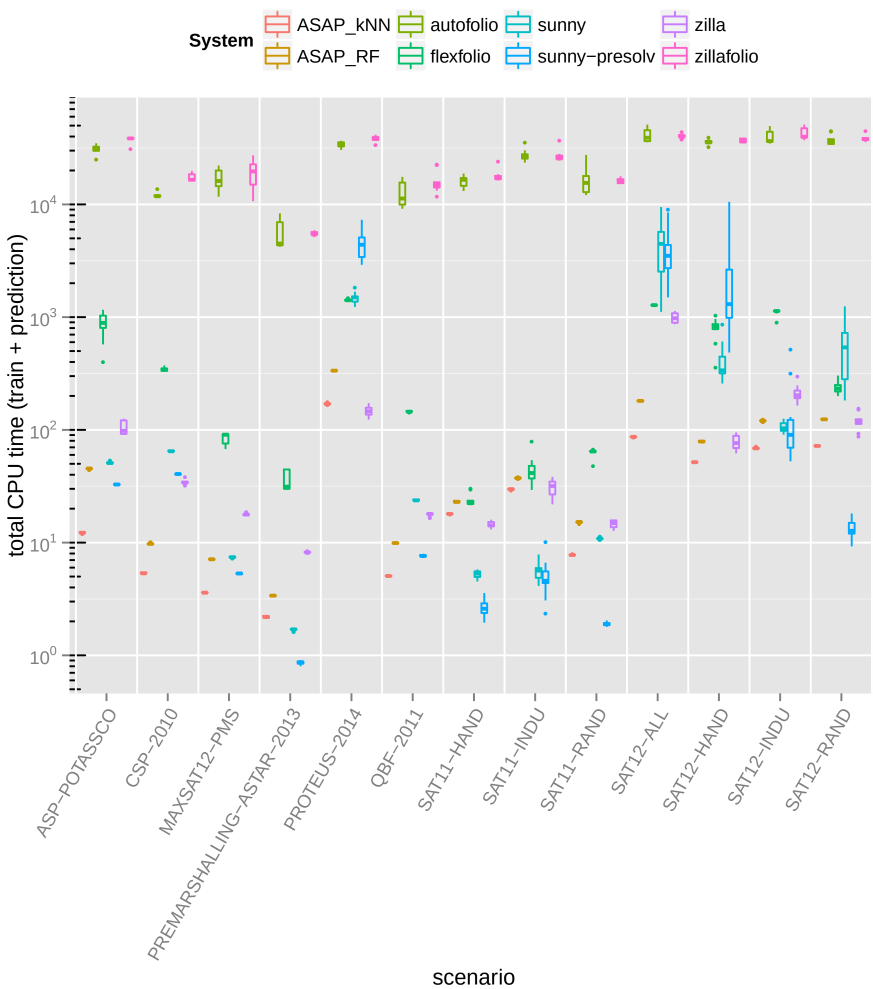

Figure 4: Train + prediction time.

4 Acknowledgements

We would like to thank all the participants for taking the time to prepare submissions and their help in getting them to run; in alphabetical order: Alex Fréchette, Chris Cameron, David Bergdoll, Fabio Biselli, François Gonard, Frank Hutter, Holger Hoos, Jacopo Mauro, Kevin Leyton-Brown, Marc Schoenauer, Marius Lindauer, Michèle Sebag, Roberto Amadini, Tong Liu, and Torsten Schaub. We thank Barry Hurley for setting up and maintaining the submission website and Luc De Raedt, Siegfried Nijssen, Benjamin Negrevergne, Behrouz Babaki, Bernd Bischl, and Marius Lindauer for feedback on the design of the challenge.