AUTOMATIC INSTANCE GENERATION FOR CLASSICAL PLANNING

ÁLVARO TORRALBA, JENDRIK SEIPP AND SILVAN SIEVERS
alto@cs.aau.dk, jendrik.seipp@liu.se, silvan.sievers@unibas.ch

EVALUATION IN CLASSICAL PLANNING

Usually, the evaluation in Planning follows the ICAPS/IPC way:

- Measure coverage.
- Time limit 5 or 30 minutes.
- Memory limit 2-8 GB.
- Use the benchmarks from the International Planning Competition.

Why? *Having a standard evaluation setting is beneficial:*

- *Reproducibility.*
- *Interpretability.*
- *Avoid hand picking results.*

Benchmarks are an important part of evaluation in Planning

But... The IPC benckmark set has some flaws

- Different number of instances per domain.
- Instance scaling: On some domains, IPC benchmarks do not show differences between planners even if they exist.

	IPC			
	#	L	D	O
Grid	5	**5**	**5**	**5**
Driverlog	20	**20**	**20**	**20**
Rovers	40	**40**	**40**	**40**
Snake	20	5	**15**	12
Total	85	70	**80**	77

Coverage of LAMA (L), DUAL-BFWS(D) and OLCFF (O)

AUTOSCALE'21: THE BENCHMARK SET

New set of benchmarks for Optimal and Satisficing Planning:

- Uniform number of instances (30 instances)
- Includes almost all IPC STRIPS domains
- Optimization based on IPC'11, IPC'14 AND IPC'18 planners →Useful to evaluate current and future planners.
- Example of Evaluation using Autoscale'14 (which didn't use IPC'18 planners):

	IPC				Autoscale			
	#	L	D	O	#	L	D	O
Grid	5	**5**	**5**	**5**	30	**17**	14	16
Driverlog	20	**20**	**20**	**20**	30	15	10	**25**
Rovers	40	**40**	**40**	**40**	30	**30**	23	28
Snake	20	5	**15**	12	30	6	**19**	16
Total	85	70	**80**	77	120	68	66	**85**

Coverage of LAMA (L), DUAL-BFWS(D) and OLCFF (O)

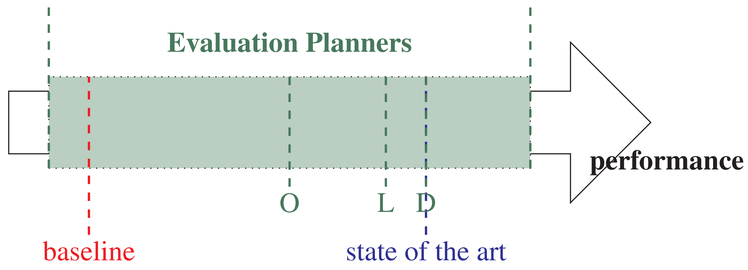
AUTOSCALE: AN AUTOMATIC TOOL TO SELECT INSTANCES

Principles:

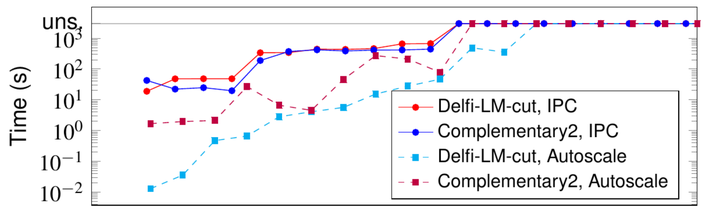
- Useful to evaluate current planners.
- Avoid bias.
- Keep the Spirit of the Domain.

Rules:

1. Agnostic to individual Planner Performance:
   Don't consider the individual results of all planners available for the optimization (only the best and the worst per instance).



2. Smooth Scaling: The instance set should have:
   - Easy instances
   - Hard instances
   - Scale smoothly



3. Parameter-based Selection: Avoid selecting the ramdom seed.
4. Sequence-based Selection: The parameter configurations can be organized in one or more sequences.
5. User Constraints: The domain designer specifies guidelines on which parameters to scale.

Optimization Process

- Generate candidate sequences with smooth scaling (using SMAC)
- Choose sub-sequences including easy, hard and diverse instances

C	S	I	time(s)
5	6	3	10.1
6	7	3	25.6
7	8	3	101.7
9	10	3	300
10	11	3	900
11	12	3	2700
13	14	3	8100

C	S	I	time(s)
1	3	2	1.8
1	4	2	2.2
1	5	2	2.9
1	6	2	4.5
1	7	2	8.3
1	8	2	26
1	9	2	120

C	S	I	time(s)
1	5	4	4.2
1	6	4	21
1	7	4	62
1	8	4	250
2	10	4	990
2	11	4	4000
2	12	4	16000

C	S	I	time(s)
1	3	5	2.8
1	4	5	3.7
1	5	5	6.1
1	6	5	16
1	7	5	62
2	9	5	200
2	10	5	660

HOW TO USE AUTOSCALE TO GENERATE INSTANCE SETS?

```
generator_command = "nomystery -l {locations}
-p {packages} -n {edgefactor} -m {edgeweight}
-c {constrainedness} -s {seed} -e 0"
parameters = [
  LinearParam("locations", lower_b=3, upper_b=10,
  lower_m=0.1, upper_m=1),
  LinearParam("packages",lower_b=2,upper_b=20,lower_m=1),
  ConstantParam("edgefactor", "1.5"),
  ConstantParam("edgeweight", "25"),
  EnumParam("constrainedness", [1.1, 1.5, 2.0])]
```

EXPERIMENTS

How evaluate the quality of a benchmark set?

- Coverage range:
  - Some instances are solved by all planners
  - No planner solves all instances
- Comparisons: pairs (X,Y) of planners with different coverage

Comparison between Autoscale'14 and IPC:

- Instance Selection: 6 planners up to IPC'14
- Evaluation: 8 planners from IPC'18

Domain	#IPC	OPT	AGL	Domain	#IPC	OPT	AGL
Barman	34/40	+12	+19	Nomystery	20	+10	+4
Blocksworld	35	+6	+26	Openstacks	70	-17	+25
Childsnack	20	+8	+1	Parking	40	-2	+5
Data-Network	20	-2	+2	Rovers	40	-4	+20
Depots	22	0	+25	Satellite	36	+5	+2
Driverlog	20	+5	+25	Scanalyzer	50	0	+8
Elevators	50	-3	+11	Snake	20	-1	0
Floortile	40	-3	+7	Storage	30	+6	+1
Grid	5	+7	+21	TPP	30	+2	+11
Gripper	20	0	+7	Transport	70	-8	+14
Hiking	20	+4	+3	Visitall	40	0	+17
Logistics	63	-3	+4	Woodworking	50	+5	+14
Miconic	150	0	0	Zenotravel	20	+4	+22

CONCLUSIONS

1. A tool to automatically select instances which:
   - Create useful benchmarks
   - Is based on sequences of parameters
   - Avoids bias with respect of the planners used
   - Keeps the spirit of the domain
2. **Autoscale'21**: New benchmark set: TRY IT OUT!

https://github.com/AI-Planning/autoscale
https://github.com/AI-Planning/autoscale-benchmarks